Stony Brook University

Complementarity in the Structure and Dynamics of Protein-DNA Search and Recognition: A Multiscale Modeling Study

A Dissertation Presented

by

Kevin Eduard Hauser

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

(Chemical Biology)

Stony Brook University

December 2016

Stony Brook University

The Graduate School

Kevin E. Hauser

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Carlos Simmerling – Dissertation Advisor
Professor, Department of Chemistry

Orlando Schärer – Chairperson of Defense
Professor, Department of Chemistry

Robert Rizzo – Committee Member
Associate Professor, Department of Applied Mathematics & Statistics

Miguel Garcia-Diaz – Dissertation Co-Advisor
Associate Professor, Department of Pharmacological Sciences

Evangelos Coutsias – External Member
Professor, Department of Applied Mathematics & Statistics

This dissertation is accepted by the Graduate School

Charles Taber

Dean of the Graduate School

Abstract of the Dissertation

Complementarity in the Structure and Dynamics of Protein-DNA Search and Recognition: A Multiscale Modeling Study

by

Kevin Eduard Hauser

Doctor of Philosophy

in

Chemistry

Stony Brook University

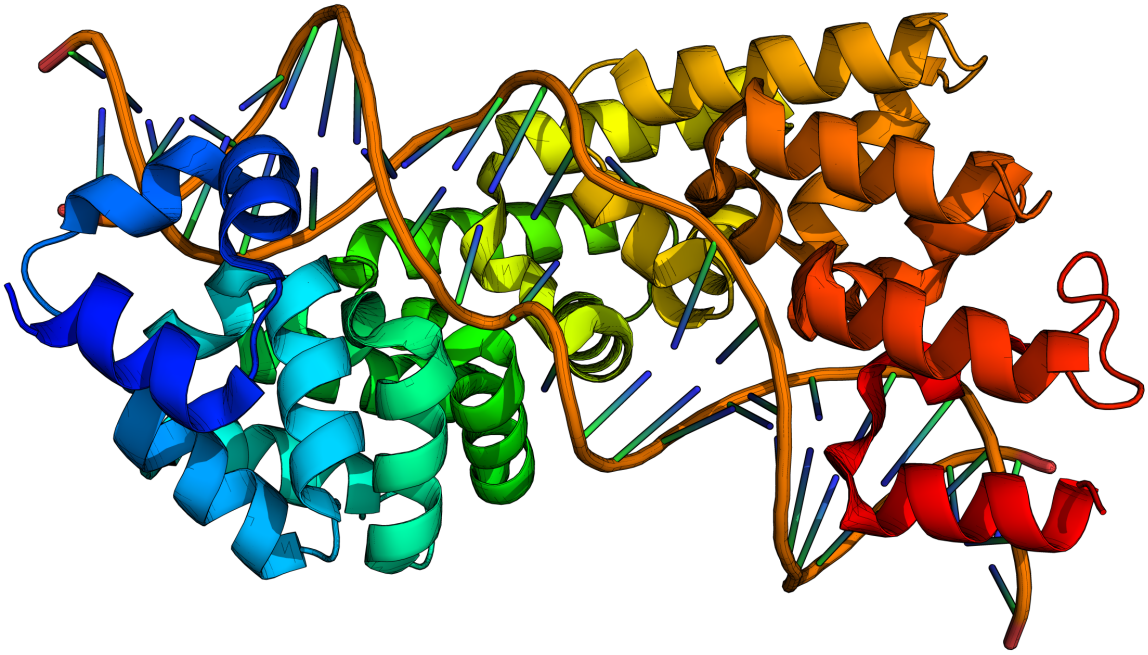
2016

Transcription factors (TF) interact with DNA to regulate gene expression by specifically binding one or two sequences out of millions of possible sites in a genome. Sequence specific binding (recognition) arises from two mechanisms: amino acid-nucleobase H-bonding (direct readout), and sequence-dependent DNA deformability and structure (indirect readout). These lead to tight, specific binding. TFs can also rapidly search for their target by attaching to any sequence and then sliding along DNA. This decreases the space through which a TF randomly walks to stumble on the target. Tight binding and rapid sliding are optimized because TFs can switch between different DNA-binding modes, a search mode and a recognition mode. Due to the disorder during search and the dynamics during recognition, the atomic level dynamics are too transient to see by experimental techniques. Therefore my goal was to develop an atomistic model of protein-DNA search and recognition using computer simulations. To meet this goal I studied a model TF, the human mitochondrial transcription termination factor-1 (MTERF1), which has a modular superhelical-architecture complementary to DNA. I used molecular dynamics (MD) simulations to characterize MTERF1 search and recognition, alongside other computational methods to generate my structural models. I was the first to observe spontaneous sliding of a TF on DNA. I found that flexibility between modules permitted contacts along the DNA footprint to shift independently. This implies that the net sliding barrier of MTERF1 is similar to the barrier for one module-DNA contact rather than the sum of all module-DNA contacts. Next I utilized enhanced sampling to drive MTERF1 over the recognition barrier(s). I found that the DNA deformed as MTERF1 made increasingly more direct readout contacts to DNA, with the global recognition process characterized by partnered DNA and MTERF1 helix unwinding. These results suggest that direct and indirect readout occur simultaneously during recognition as MTERF1 unwinds (deforms) the target DNA sequence. Helix unwinding motions were present in search mode, suggesting the dynamics that drive search and recognition are intrinsic to the MTERF1 architecture.

Dedication Page

To Johnny, and any kid who had a bad gene.

Frontispiece



A human transcription factor partners its superhelical geometry with a distorted DNA helix.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning
2D	2 dimension
3D	3 dimension
Å	Ångstrom (10^{-10} meters)
AAE	Average absolute error
AD	Alzheimer's Disease
AMBER	Assisted model building with energy refinement
ANM	Anisotropic network model
AU	Arbitrary units
bp	Base pair
CG	Coarse grained
COBALT	Constraint-based alignment tool
COM	Center of mass
CRISPR	Clustered regularly interspaced short palindromic repeat
CryoEM	Cryo-electron microscopy
CV	Coefficient of variation
DNA	Deoxyribonucleic acid
DBP	DNA binding protein
DBD	DNA binding domain
dA	Adenine (deoxyribose)
dC	Cytosine (deoxyribose)
dG	Guanine (deoxyribose)
dT	Thymine (deoxyribose)
EOM	Equation of motion
fs	femtosecond
HOMO	Highest occupied molecular orbital
K	Kelvin
LUMO	Lowest unoccupied molecular orbital
LJ	Lennard-Jones
LRR	Leucine-rich repeat
MC	Monte Carlo
MD	Molecular dynamics
MM	Molecular mechanics
MTERF1	Human mitochondrial transcription termination factor-1
NMA	Normal mode analysis
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
PDB	Protein data bank
PD	Parkinson's Disease
PMF	Potential of mean force
PNEB	Partial nudged elastic band
PPII	Poly-proline II
PPM	Parts per million
PPT	Points per turn
ps	Picosecond
RMS	Root mean square
RMSD	Root mean square distance
RMSIP	Root mean square inner product
TALEN	Transcription activation-like effector nuclease
TIP3P	Transferable intermolecular potential 3 points
TSS	Transcription start site
VDW	van der Waals
YR	pyrimidine (Y), purine (R) dinucleotide step
ZFN	Zinc finger nuclease

Acknowledgments

I have been very privileged to have many great mentors during this Ph.D. of mine. Dr. David Ferguson was instrumental in funneling me to Stony Brook. Dr. Nancy Goroff served as my intellectual reference state during my first few years at Stony Brook, and later. Dr. Daniel Raleigh showed me how mighty fine experiments are to perform, and ponder. Dr. Peter Tonge helped me think deeper about kinetics, and the non-equilibrium nature of biology. Dr. Nicole Sampson made sure my hypotheses were well designed. Dr. Elizabeth Boon introduced me to biochemistry. Dr. Isaac Carrico introduced me to artificial transcription factors. Dr. Gábor Balázsi gave me the opportunity to put some of my ideas to the test. Dr. Ken Dill let me test-drive his computers, shared jazzy morning coffees with me, had great parties and gave me many bits of advice over the years.

I thank my dear friend Dr. James A. Maier for more than I can write, and my family at the Center for Inclusive Education, who supported me at every twist and turn. I lovingly thank my family - Mãe, Hauser, Kerry, and Karly Ellen, the better KEH.

Professor Orlando Schärer, thank you for pointing me to the right place: DNA, and a Ph.D.

Professor Miguel Garcia-Diaz, thank you for your countless hours of patience while teaching me the basics of X-ray crystallography, termination assays, isothermal calorimetry experiments and mitochondrial genetics. Thanks for connecting me to biology.

Professor Evangelos Coutsias, thank you for helping me derive good ideas, for sharing the stars and for helping me calibrate a perspective. Thanks for being a great friend and mentor.

Professor Robert Rizzo, thank you for properly introducing me to molecular modeling, docking and scientific writing. My time with you Rob during 2009 and 2010 was essential to setting my Ph.D. trajectory. Thanks for helping me find my rhythm.

Professor Carlos Simmerling, thank you for changing my life.

I am grateful for the funding I have received: National Institutes of Health Ruth L. Kirschstein National Research Service Award [F31-GM101946]; Chemical Biology Training Program Fellowship [T32-GM092714]; National Science Foundation Louis Stokes Alliance for Minority Participation Bridges to the Doctorate Fellowship [HRD-0929353]; Alliance for Graduate Education and the Professoriate-Transformation Fellowship [HRD-1311318].

Chapter 1. Introduction

Structure begets function. In biology, this axiom is central because structure leads to emergent properties – the structures of enzymes, organelles, cells, organisms, groups and ecosystems beget higher-order properties we understand as a biological function. In chemistry, rich properties arise from combining the elements S, P, O, N, C and H and more[1a] in myriad combinations to form proteins, enzymes, lipids, sugars, RNA, DNA, water and ultimately a cell.

The goal of my dissertation was to model the biologically relevant structure and dynamics of protein-nucleic acid interactions. The aim of this chapter is to introduce the basic ideas of genetics and gene regulation (**Section 1.1**), how DNA-binding proteins diffuse to a target DNA sequence (**Section 1.2**), the geometry of helices in structural biology (**Section 1.3**), molecular dynamics simulations (**Section 1.4**) and elastic network models (**Section 1.5**).

[a] The following elements are also present in a human (and most organisms), ranked by the fraction of total human body weight: O (61%), C (23%), H (10%), N (2.6%), Ca (1.4%), P (1.1%), S (0.2%), K (0.2%), Na (0.14%), Cl (0.12%), Mg (270 parts per million, ppm), Si (260 ppm), Fe (60 ppm), F (37 ppm), Zn (33 ppm), Cu (1 ppm), Mn (0.2 ppm), Sn (0.2 ppm), I (0.2 ppm), Ni (0.1 ppm), Mo (0.1 ppm), V (0.1 ppm), Cr (0.03 ppm) and Co (0.02 ppm).

1.1. Transcription factors regulate gene expression

Genes encode proteins

A gene is transcribed into messenger RNA, which is translated into a protein[2]. Proteins are important because they are the doers of the cell: shuttling oxygen from our lungs to our cells so we can breathe, generating fuel to drive other proteins to catalyze chemical reactions and providing the structural scaffolding to support the shape of a cell. A protein is composed of a specific sequence of amino acids. All amino acids have the same backbone - three atoms bonded together in the same way. What differentiates the amino acids are their side chains.

The twenty naturally occurring amino acids that make up most proteins in most cells can have charged side chains (arginine, lysine, aspartate, glutamate), polar side chains (histidine, asparagine, glutamine, serine, threonine, tyrosine), or non-polar side chains (alanine, cysteine, phenylalanine, glycine, isoleucine, leucine, methionine, proline, valine, tryptophan). Oppositely charged side chains can form ionic interactions with each other; polar side chains can hydrogen bond (H-bond) to each other, to the backbone atoms of other amino acids, or with water; and non-polar side chains can blob together like an oily droplet. Because water molecules are polar, they tend to pull amino acids with polar side chains outwards while squeezing non-polar side chains inside a protein. H-bonds drive water molecules' pulling of polar side chains, while H-bonds indirectly drive water molecules' squeezing of non-polar side chains. Because water molecules can H-bond with other water molecules, but not non-polar side chains (or any non-polar molecule; consider oil), the space taken up by a non-polar side chain could have been taken up by a water molecule or polar side-chain that could participate in an H-bond. In addition to this

loss of H-bonding, the water molecules surrounding the non-polar side chain are robbed of their potential to switch between one more H-bond.

Robbing the freedom of a molecule to switch between similar states - H-bonding with a side chain, or the myriad neighboring water molecules - costs energy. This is entropy: the number of ways in which a molecule can adopt distinct geometries: more accessible states (disorder) has higher entropy, lowering the total free energy of the system. Ionic interactions, H-bonds, entropy and the energy of an amino acid backbone are the forces that drive a protein to adopt a structure uniquely defined by the sequence of amino acids composing the protein.

The structure of a protein determines its function. For example, a protein that binds to DNA must adopt a structure that complements the helical form of DNA. Because the structure of a protein is encoded by its amino acid sequence, with few exceptions, the function of a protein can effectively be encoded in this sequence. How can amino acid sequences be efficiently stored? Genes encode the amino acid sequence of proteins using the genetic code: three nucleotides (codon) represent one amino acid, and each of the twenty naturally occurring amino acids are encoded by different codons. A protein is encoded by a specific sequence of codons.

Different organisms have different genomes

A genome is important because it houses genes, along with short interspersed sequences of DNA that serve as keys to regulate the expression of those genes. A genome is composed of many genes because a cell requires many proteins to survive. Mitochondria and viruses have tens of genes, bacteria have thousands of genes and human cells have tens of thousands of genes[3], including mitochondrial genes. In general, the more complex an organisms is (bacteria versus

human), the more genes will be housed in the organism's genome. With this rise in genomic complexity comes the increased complexity of maintaining the integrity of the genome and expressing its genes.

Transcription factors toggle gene expression

A transcription factor binds to a specific DNA sequence, a regulatory site, to signal the expression of a gene. Transcription factors are functionally like a toggle switch, because their presence or absence at a regulatory site signals enzymes to begin or terminate copying the gene. These DNA-copying enzymes are polymerases. Saliently, there are two classes of polymerases: RNA polymerases that copy a gene into a messenger RNA destined for translation into a protein by the ribosome; DNA polymerases that copy the genome into a new, daughter genome. RNA-polymerases transcribe genes to make proteins (transcription), DNA polymerases replicate genomes so cells can duplicate and eventually pass on this genetic information to daughter cells following cell division (replication). Because cells must transcribe the right gene at the right time, transcription factors are of central importance to genomic integrity, metabolism and survival of a cell. To operate with such precision, transcription factors must bind with extremely high specificity one DNA sequence.

1.2. Diffusion of DNA-binding proteins to target DNA: search mode

One in ten genes in the human genome encodes a transcription factor (TF)[4], and once expressed, TFs direct the expression of other genes. All living organisms have TFs because these proteins regulate gene expression, a vital cellular process. TFs are defined by their ability to bind to a specific DNA sequence. The structure of a TF is important because it must be able to bind to DNA and form contacts in the grooves to recognize the target sequence. The dynamics of a TF is also important because it must be able to adapt itself to sequence-specific DNA structures, or to induce a specific DNA structural distortion, to recognize the target sequence.

TFs adapt conformation to switch function: to bind, search, or recognize DNA. To rapidly respond to stimuli, TFs must locate target DNA quickly. Remarkably, TFs can actually bind to their target sequences faster than 3D diffusion – they exceed the kinetic speed limit. The Smoluchowski equation predicts the maximum on-rate (defining the kinetic speed limit) of a TF to its target DNA sequence to be a rate that is ten-fold slower than the rates observed *in vivo*[5]. It is as if the TF is on a track from a random position in space that leads directly to the target. As it turns out, that track is the DNA. TFs can bind to a random sequence on the genomic molecule and slide along the negatively charged DNA backbone. Depending on the TF, it can also bind and unbind to hop along the sequence of DNA, and it can jump between two pieces of DNA that are near in space but are distal in sequence (a loop). Thus, 3-D diffusion and 1-D facilitated diffusion (sliding) likely drive target search[6-11].

So-called frustration[10] can arise during search when affinity for nonspecific DNA is too high. If a TF is too strongly bound to random DNA sequences, the probability that it can eventually find the target decreases because it spends too much time bound to the wrong sites. Because the differences between a target DNA sequence and random DNA sequences are subtle, the TF faces a paradox: if it binds the target too tightly, then it will bind all sequences more tightly and search too slowly; if it binds all sequences less tightly, then its affinity for the target will become too low.

To overcome this search-recognition paradox, a TF can adopt two distinct conformations - one a search mode that is optimized for rapid diffusion and sliding, and the other a recognition mode that is optimized for tight, sequence specific binding. Mathematical models predicted[10-11] and single-molecule experiments of p53 corroborated[12-13] that TFs indeed switch from a rapid search mode to a tight-binding recognition mode by changing conformation. In search mode, scanning is facilitated by fleeting, nonspecific binding with ~ 1 k_BT energy gaps and small barriers between neighboring noncognate sites[11]. Significant perturbation of the DNA structure is unlikely on such small energy and time scales. Thus, a TF should be able to weakly bind a random sequence of DNA, the canonical structure of which is presumably B-form[14-15].

Conformational change regulates recognition.

During recognition the TF can change conformation. In changing its conformation, the amino acids in the TF shift to an arrangement that optimizes specific contacts with the chemical groups present in the cognate DNA sequence (recognition mode). These sequence specific contacts represent the direct readout component of recognition. The recognition mode TF

conformation has a new energy with large differences in energy between neighboring DNA sites and high barriers between them (**Figure 1.1**). This exaggerates the energy landscape from a smooth landscape that can be easily explored (search mode) to a rocky landscape replete with deep traps and precipitous crags that is tough to traverse (recognition mode). By exaggerating the energy landscape in recognition mode, the difference in binding energy between the one specific and the many nonspecific sites is dramatically enhanced[11].

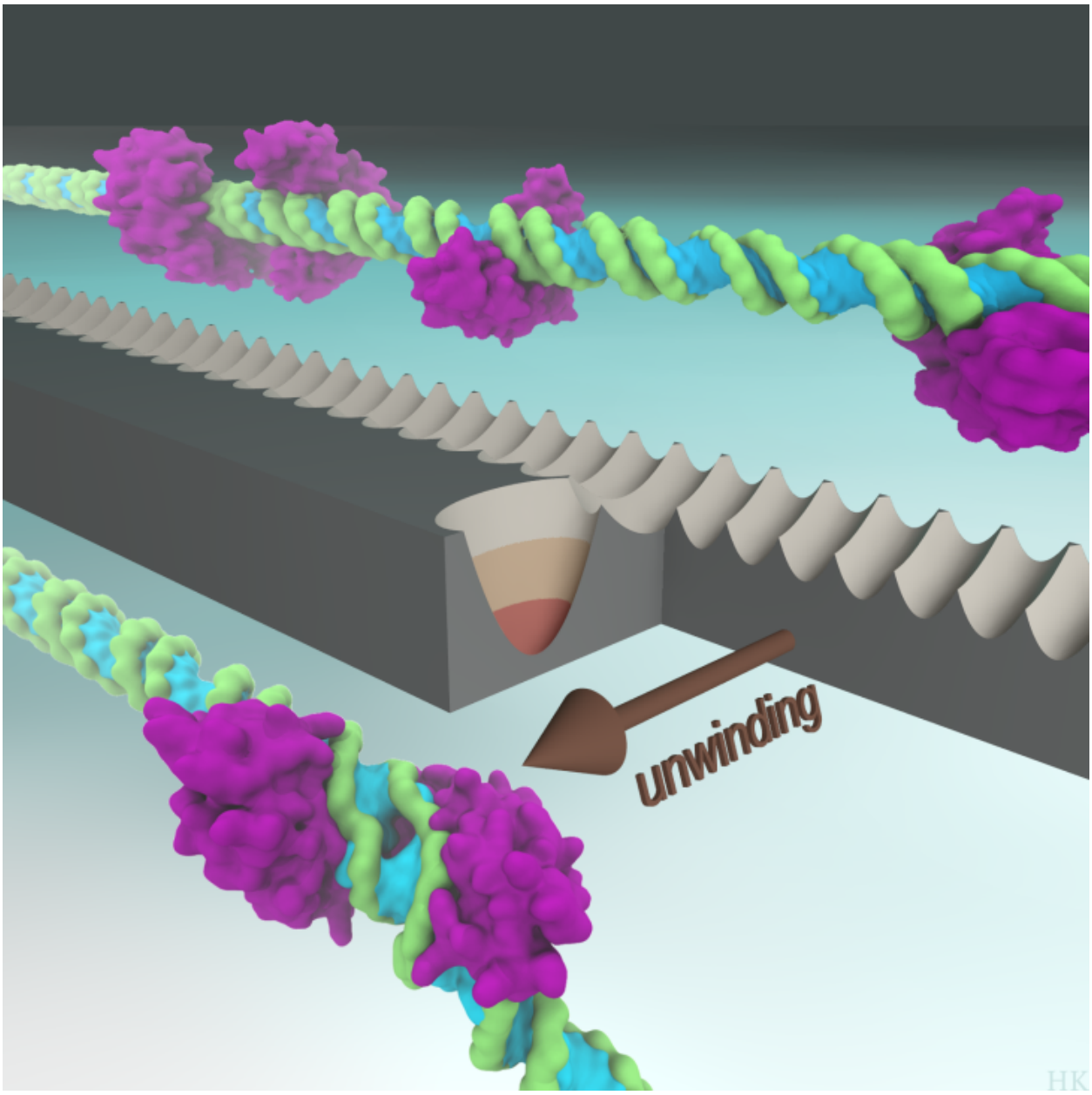


Figure 1.1. A TF can adjust its energy landscape during search by adopting a search mode conformation in which all places on the DNA have similar energies with small barriers between them. After a conformational switch, the TF exaggerates its energy landscape to enhance affinity for the target. DNA is shown in green and blue; green highlights the backbone of DNA, which is the same for all DNA sequences (the backbone is nonspecific); blue highlights the nucleobases of DNA, which is unique for a DNA sequence (the bases are specific). A model transcription factor is shown in purple. The energy landscape is shown in grey. In search mode, a TF can glide along the DNA backbone, experiencing only small bumps in the energy landscape as it transits from one sequence of DNA to the next. In recognition mode, a TF adopts an extended conformation (unwinding), switching into a portion of the energy landscape with a deep well (red bottomed-well in energy landscape). Once unwound, a TF must overcome a very large energy bump to unbind the DNA.

The kinetic aspect of recognition is analogous to enzyme inhibitors that exhibit long residence times following an induced fit conformational change in the protein[16-17]. Dynamics of the tightly bound TF can also induce DNA deformation, potentially giving rise to dynamic indirect readout via sequence-dependent deformability of DNA, or to shape readout (static indirect readout)[18-20]. Therefore, conformational changes in the TF and in the DNA during recognition are coupled dynamic processes that depend on atomistic intermolecular interactions—direct readout—and intramolecular interactions—indirect readout and the increase in the energy of the protein upon binding (strain). For example, NMR transverse relaxation rate measurements of the lac repressor DNA binding domain reveal that amino acids involved in direct readout in the recognition mode form nonspecific interactions with the phosphate backbone in the search mode[21]. The data suggest that conformational adaptation from search to recognition modes includes switching nonspecific contacts with the DNA backbone to specific TF-nucleobase interactions. TF-DNA binding and recognition is thus a function of the relative energies of the search and recognition metastates, which is determined by the thermodynamics and kinetics of TF and DNA conformational change.

The relative importance, however, of direct and indirect readout during the transition from search to recognition mode is poorly understood. Insight into the mechanism of conformational change, and thus of recognition, would be facilitated by high-resolution structural data for specific and nonspecific complexes. The lac repressor DNA-binding domain[21] and the enzymes BamHI[22], BstYI[23], and EcoRV[24] are prototypical DNA-binding proteins for which static structures of specific and putative nonspecific complexes have been experimentally characterized by NMR and X-ray crystallography. However, the lifetime of a true nonspecific complex is by definition fleeting[11]. To favor binding at a single nonspecific site requires

alterations to the DNA or protein, truncated constructs, or protein-DNA cross-links that stabilize the energy of an artificial nonspecific complex. In these altered complexes, usually only a few interactions have been modified and therefore a subset of the cognate recognition contacts may still be present – "hemispecific recognition"[23] - and the DNA is frequently shifted from B-form. For example, the structure of the human transcription factor MTERF1 was solved by X-ray crystallography for a putative nonspecific complex in which a subset of the recognition interactions was eliminated. The DNA conformation was unwound and kinked, however, and resembled that seen in the fully cognate complex[25]; the DNA conformation in putative nonspecific complexes of BamHI[26], BstYI[23], and EcoRV[27] enzymes also resemble that in the cognate complex. Consequently, it is unclear how accurately these altered complexes represent the actual structure during rapid search, outside the influence of methods used to redirect binding specificity and trap a unique noncognate structure. Moreover, static snapshots do not resolve dynamics. A complete mechanistic picture of how TFs regulate gene expression would involve a dynamic model of the ensemble of structures that correspond to search mode, as well as an atomistic description of the conformational and energetic changes that take place during the transition from nonspecific to specific complexes.

In this dissertation, I use a combination of experimental structural data and MD simulations to address the first element in this challenge and develop a dynamic model for nonspecific DNA binding, using the human mitochondrial transcription factor MTERF1 as a prime example. MD simulations provide the unique ability to control the structure of biomolecules such as MTERF1 and DNA, and calculate the energy of any subset of atoms in the system being simulated. In addition, MD simulations can be used to get the atomic-level details of transient structures that cannot be detected directly by experiments.

1.3. Helix geometry of biomolecules

The helix is a ubiquitous geometric form in structural biology. Local, secondary structure elements in proteins can adopt helical structures (α-helix, 3_{10}-helix, π-helix) to support global protein architectures. The architectures of proteins can adopt a range of helical forms. The triple helix of collagen is composed of three intertwined poly-proline II (PPII) secondary structure helices. Kinase receptors can adopt a superhelical architecture (tertiary structure) composed of covalently connected repeats that spiral around a common axis (helical axis)[28]. These particular repeats are composed primarily of non-polar amino acids, specifically leucine, which serve to stabilize the core structure of a repeat. Such leucine-rich repeats (LRR) are observed in a broad range of proteins, all of which adopt a similar superhelical structure as the kinase receptors: Toll-like receptors[29], which are involved in the innate immune response of cells; the ecto-domains of some G-protein coupled receptors (e.g. LGR4)[30], which are involved in intracellular signaling cascades; certain hormone receptors (e.g. DWARF14)[31]; and computationally designed, artificial LRR superhelices with bio-orthogonal functional properties[32]. It is essential to characterize the geometric properties of helices traced by atoms in a biomolecule because function is driven by structure and dynamics.

There are two major problems with characterizing the geometric parameters of biomolecular helices. Irregularities between the atoms that trace the helix make it difficult to fit the points directly to the parametric equations of a helix. Also, biomolecular helices frequently trace less than one helical turn. If the helical axis is known, methods from the field of high-energy particles physics can be used to characterize sub-one turn helices[33-37]. Without the

constraint of a defined helical axis, irregularities make the fitting problem challenging or impossible to solve.

In structural biology, empirical methods that utilize constraints are available that can characterize the geometric parameters of irregular helices, but are only applicable to the molecular fragments for which they were trained. Given the broad diversity of helices observed in structural biology, a general method that does not require empirical constraints - such as a prior knowledge of the helical parameters pitch, radius and axis - and is less affected by helical irregularities would enable the analysis of any biomolecular helical geometry. Such a method is needed to characterize the structures of superhelical proteins, such as nucleic acid binding proteins (NBPs) and modular superhelices[38].

Based on an extensive search of the particle-track fitting literature, I was able to find only one method that could provably converge to the global solution of helical parameters for a set of points that did not require prior knowledge of the helical axis. This method was the total least squares (TLS) approach developed by Yves Nievergelt[39]. The method operates in two steps. First, a set of points are fit to the surface of a cylinder by minimizing the sum of square distances between each point and the cylinder. This best-fit cylinder defines the helical axis and helix radius. In a subsequent fitting step, the pitch of the helix is determined by fitting the points to a straight line in the 2D surface of cylinder (the axial displacement along the cylindrical/helical axis, and the polar angle around this axis). Nievergelt used his method to characterize the helical parameters from a set of ten points tracing a 90° helix.

The assumption that a set of points tracing a real helix can be mapped to a cylindrical coordinate system is equivalent to the assumption that the points project a circle on a plane that

is radially sliced through the cylinder. This latter idea, that a cylindrical helix projects a circle on the plane whose unit normal vector is parallel to the helix axis (peering down the helical axis, the points look like a circle) greatly simplifies the mathematics involved in solving the parameters of a helix; total least-squares fitting for a 3D problem is reduced to linear least-squares fitting for a 2D problem. The advantage is not limited to a simplification of the mathematics: solving a 2D problem requires fewer data points than a 3D problem because there are fewer parameters. As will be discussed in greater detail in **Chapter 3**, additional advantages arise from simplifying Nievergelt's 3D method to a 2D method.

1.4. Molecular Dynamics

The goal of molecular dynamics (MD) simulations is to understand the changes in energy and structure of large biomolecules such as proteins and nucleic acids. Importantly, MD simulations are able to characterize the dynamics and energetics of biomolecules under biologically relevant conditions (temperature, solvent, pressure), on biologically relevant timescales (μs-ms). MD can complement high-resolution structural experiments like X-ray crystallography, cryo-electron microscopy (cryoEM) and NMR. In principle, MD simulations can resolve the complete ensemble of configurations accessible to a biomolecule, including transient configurations that are challenging to capture in experiments like NMR or cryoEM. If an MD simulation of a biomolecule is able to reproduce experimental observables (such as average structure), then one can be reasonably sure that the results of an MD simulation are representative of the dynamics of the single molecules that were present in the experiment. Before discussing the limitations of an MD simulation, let us briefly review the basics of an MD simulation and how it works.

1.4.1. The Calculus of classical mechanics

The following Newtonian (classical) mechanics is general to any system of particles. Here, the particles are atoms. In classical mechanics, the force **F** on an atom, is equal to the atom's mass, m, times the atom's acceleration, **a**:

$$\mathbf{F} = m\mathbf{a} \tag{1-1}$$

In addition, force is the negative gradient $\boldsymbol{\nabla}$ of the potential V:

$$\mathbf{F} = -\boldsymbol{\nabla}V \tag{1-2}$$

Thus, the force on an atom can be derived from its potential energy.

1.4.2. Molecular mechanics force field

In molecular mechanics (MM), the potential energy function is typically written as the sum of bonded terms (bonds, $V_{\mathbf{B}}$; angles, $V_{\mathbf{A}}$; and torsions, $V_{\mathbf{T}}$) and two non-bonded terms (Lennard-Jones, V_{LJ}; Coulomb electrostatics, V_C):

$$V_{MM} = V_{\mathbf{B}} + V_{\mathbf{A}} + V_{\mathbf{T}} + V_{LJ} + V_C \tag{1-3}$$

where the bonded terms are described by simple harmonic potential terms. **Figure 1.2** illustrates the geometric components to which these potential are applied.

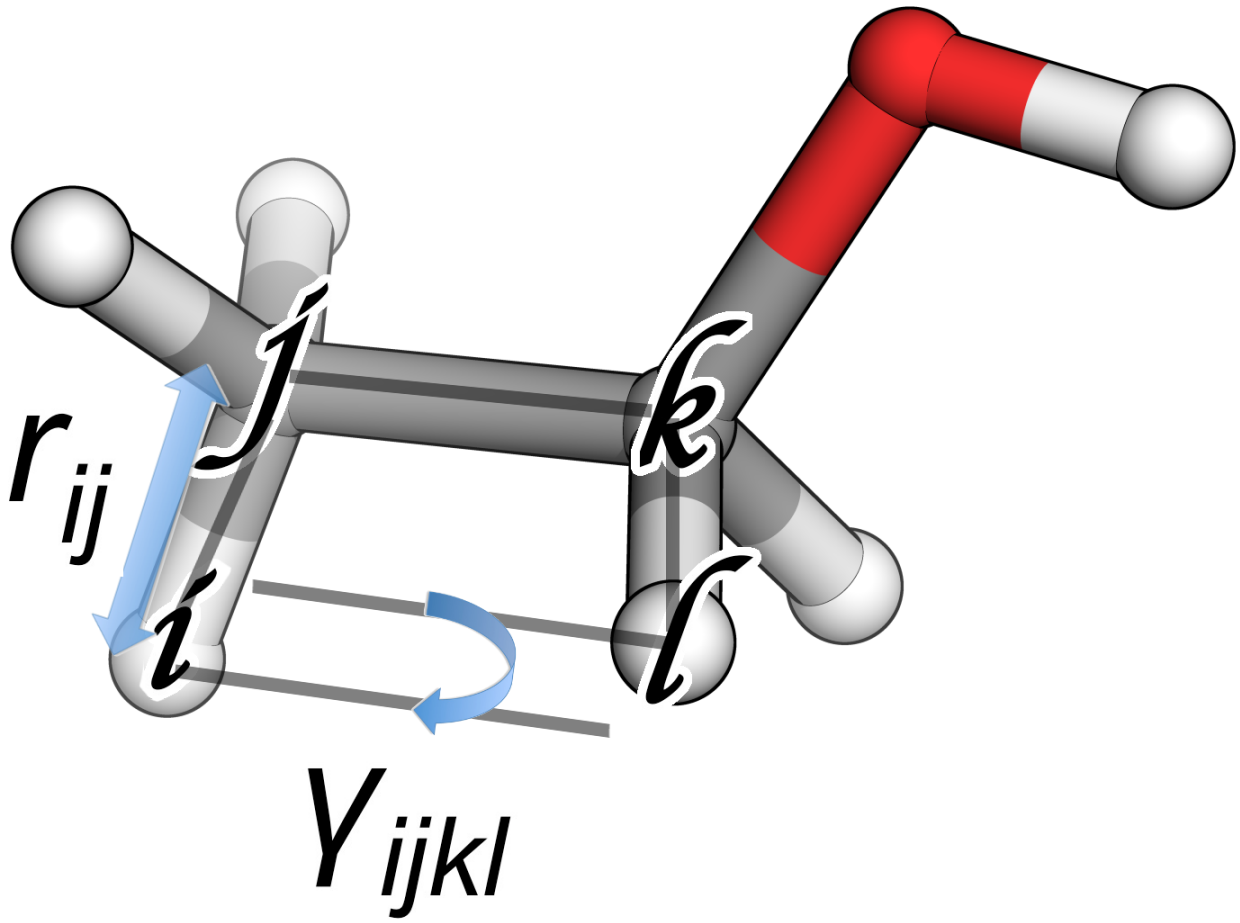


Figure 1.2. The structure of ethanol is shown as an example of the geometric components used by the potential energy function in equation (1-3). r_{ij} is the distance component between nucleic i and j; θ_{ijk} is the angle component between atoms i, j and k; and γ_{ijkl} is the torsion component between atoms i, j, k and l.

The potential energy of bonds is:

$$V_{\mathbf{B}} := \sum_{(i,j)\subset\mathbf{B}} \frac{k_{\mathbf{B}_{ij}}}{2}(r_{ij} - \bar{r}_{ij})^2 \tag{1-5}$$

where $k_{\mathbf{B}_{ij}}$ is the bond stretching force constant, r_{ij} is the instantaneous distance between atoms i and j, and $\bar{r}_{ij}$ is the reference distance between atoms i and j.

The potential energy of angles is:

$$V_{\mathbf{A}} := \sum_{(i,j,k)\subset\mathbf{A}} \frac{k_{\mathbf{A}_{ijk}}}{2}(\theta_{ijk} - \bar{\theta}_{ijk})^2 \tag{1-6}$$

where $k_{\mathbf{A}_{ijk}}$ is the angle bending force constant, θ_{ijk} is the instantaneous angle subtended by atoms i, j and k, and $\bar{\theta}_{ijk}$ is the equilibrium angle subtended by atoms i, j and k.

The potential energy of torsions angles is:

$$V_{\mathbf{T}} := \sum_{(i,j,k,l)\subset\mathbf{T}} \sum_{s=1}^{M} k_{\mathbf{T}_{ijkl,s}}(\cos(s\gamma_{ijkl} - \delta_s) + 1) \tag{1-7}$$

where $k_{\mathbf{T}_{ijkl,s}}$ is the torsion angle twisting force constant, γ_{ijkl} is the instantaneous torsion angle between atoms i, j, k and l, and δ_s is the phase shift. This function enforces planarity, mimicking

physical properties arising from the electronic structure of double bonds. The function also corrects for multipole and inductive effects. M is often limited to an index of 4.

The Lennard-Jones non-bonded potential energy term is:

$$V_{LJ} := \sum_{i=1}^{i-1} \sum_{j=1} 4\varepsilon \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) \tag{1-8}$$

where ε is the well depth of the favorable contact energy between atoms i and j, σ_{ij} determines the distance at which the Lennard-Jones energy is zero, and r_{ij} is the distance between atoms i and j. The second term, $(\sigma/r)^6$, models attractive dispersion forces; the first term, $(\sigma/r)^{12}$, is simply the square of the second and it models the repulsive forces between two atoms in close approach.

The Coulomb non-bonded potential energy term is:

$$V_C := \sum_{i=1}^{i-1} \sum_{j=1} \frac{e^2}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}} \tag{1-9}$$

where *e* is the elementary charge (**Table VI-1**), q_i and q_j are the charges of atoms i and j, ε_0 is the permittivity constant (**Table VI-1**), and r_{ij} is the distance between atoms i and j.

1.4.2.1. Additional terms used in contemporary molecular dynamics simulations

The equations described in the above section are very simple - they represent the bare components shared by contemporary atomistic MD simulation implementations (Coarse Grained MD simulations[41-42] may employ fewer terms than those discussed above). These additional terms include: (1) 1-4 scaling factors that adjust Lennard-Jones, equation (1-8), and electrostatic energies, equation (1-9), between atoms i and i+3 [43]; (2) Urey-Bradley 1-3 distance terms in addition to the angle term, equation (1-6), between atoms i and i+2 [44]; (3) empirical corrections to the peptide φ/ψ backbone torsion angles, equation (1-7), CMAP[45] and AMAP[46]; (4) atomic charge polarizability using the Drude oscillator model[47], which mimics the induction of dipoles with the delocalization of charges away from atom centers.

1.4.3. Classical mechanics equation of motion

Molecular dynamics simulations treat the atoms in a molecular system as point masses whose equation of motion (EOM) is described by classical mechanics. In a molecular dynamics simulation of three or more atoms[b], numerical solutions to the EOM are required. Let $\mathbf{r}_i$ be the position vector of atom i, and $\mathbf{r}_i'$ be its velocity (where the prime denotes $\mathbf{r}_i'$ being the derivative of $\mathbf{r}_i$, with respect to time t).

The velocity of the particle is equal to its momentum $\mathbf{p}_i$ divided by its mass m_i:

$$\mathbf{r}_i' = \mathbf{p}_i / m_i \qquad (1\text{-}10)$$

[b] The equations of motion for any system of three or more interacting particles is a general problem, the "many body problem".

The equation of motion for atom *i* is simply the change in its momentum with respect to time:

$$\mathbf{p}_i' = \mathbf{F}_i \tag{1-11}$$

The force $\mathbf{F}_i$ acting on atom i arises from its interactions with all the other atoms in the system. A truncated Taylor series expansion can then be used to calculate the position of the atom a short time in the future (say Δt) from its current position in 1D space (say, x) at time t:

$$x(t + \Delta t) = x(t) + x'(t)\,\Delta t + x''(t)\,\Delta t^2/2 \tag{1-12}$$

Equation (1-12) assumes constant atomic acceleration (x" = F/m) over Δt. To solve equation (1-12), one simply needs to know the position of the atom, x(t), the velocity of the atom x'(t), the acceleration of the atom x"(t) and the time over which forces are integrated, Δt. With appropriate approximation to higher-order terms in the Taylor series expansion, $x(t + \Delta t)$ can be calculated precisely. How accurately an algorithm approximates equation (1-12) can be tested easily, because Newton's third law states that the net force acting on all of the atoms in a system must be zero. Thus, energy must be conserved.

The integration time step, Δt, is a critical parameter in MD simulations because it determines how forces are propagated. Weak forces will lead to small accelerations of atoms; small atomic velocities mean that an atom will not travel far per unit time. Strong forces will impart large accelerations on atoms, launching them into high-velocity trajectories. At high velocity, the position of an atom can change quickly. Over a long integration time step, it is

conceivable that the forces acting on an atom can change as its position along the displacement vector changes. For example, the force vector of an atom may be directed towards another atom (as if often the case in condensed phase such as that found in a biological fluid within a cell). If the integration time step is very short (1 femtosecond, fs), then the atom will travel a short distance along the vector towards the second atom. If the integration time step is very long (10 fs), then the atom will travel a large distance that may end up within the second atom. If the integration time step is very short (1 femtosecond, fs), then the atom will travel short distances permitting smooth changes in forces from one integration to the next. If the integration time step is very long (10 fs), then the atom travels longer distances (with the force used by the integration that launched the atom's travel) over which it is very likely the forces acting on the atom should change (perhaps because the atom bumps into a second atom). Therefore, short integration time steps are utilized in atomistic MD simulations to prevent such catastrophic cases. These time steps typically are 1 or 2 fs. Velocity Verlet[48] and Leapfrog integrators are standard algorithms used to propagate the dynamics of atoms in an MD simulation because the MM force field is a continuous potential energy function and the interactions between all atoms are coupled; the forces acting on an atom in a system of three or more atoms in not constant. The Velocity Verlet finite difference integrator enables the simulation of trajectories that conserve energy and momentum, are time-reversible and can be calculated quickly.

1.5. Normal mode analysis & anisotropic network model

Normal mode analysis (NMA) is a powerful tool complementary to MD simulations because: (1) an exact energy landscape is obtained because of its analytical construction, versus numerical integration of discrete time steps towards an unknown equilibrium distribution in MD; (2) a landscape is found by a single mathematical operation - diagonalization of a Hessian matrix, versus what might be an (effectively) infinite number of integrations in MD; (3) entropy and heat capacity can be estimated because quantization of configuration space can be easily introduced, versus physical intuition-based choices of geometric degrees of freedom likely to be descriptive of the relevant phase subspace sampling. These advantages arise from the central assumption that the underlying energy landscape of the dynamics is described by a simple harmonic potential energy function around a minimum energy (equilibrium) structure. Despite this substantial simplification compared with the full MM energy function, the relative flexibility of atoms in biomolecules in an NMA calculation are comparable to MD simulations and X-ray crystallographic *B*-factors (which arise from thermal fluctuations of the biomolecule)[49].

1.5.1. Harmonic potential energy function

The following derivations detail how to change the system of coordinates describing the static geometry of a molecule to a new system of coordinates - normal modes - that lay along the vectors describing the dynamics of a molecule. Any atomic position is always described by a combination of Cartesian coordinates (i.e. x, y, z) whereas any atomic motion is always described by a combination of normal modes.

In Tirion's simplified model[50], the potential energy function of a molecule whose coordinates are defined by the generalized coordinates **q**:

$$\mathrm{E_P} = \frac{1}{2}\sum \mathrm{q}_i \mathrm{F}_{ij} \mathrm{q}_j = \frac{1}{2}\mathbf{q}^{\mathrm{T}}\mathbf{H}\mathbf{q} \tag{1-13}$$

where q_i and q_j are the coordinates of atoms *i* and *j*, H_{ij} is the force between atoms *i* and *j*, and **H** is the generalized force matrix:

$$\mathrm{H}_{i,j} = \left.\frac{\partial^2 \mathrm{E_P}}{\partial \mathrm{q}_i \partial \mathrm{q}_j}\right|_{\mathbf{q}=0} \tag{1-14}$$

The kinetic energy function is:

$$\mathrm{E_K} = \frac{1}{2}\dot{\mathbf{q}}^{\mathrm{T}}\mathbf{M}\dot{\mathbf{q}} \tag{1-15}$$

where $\dot{\mathbf{q}}$ is the vector of velocities, **M** is the "mass matrix", from which rotations and translations of the molecule are removed by the moving derivatives $\partial \mathbf{r}_l/\partial \mathrm{q}_i$ and $\partial \mathbf{r}_l/\partial \mathrm{q}_j$.

The elements of **M** are:

$$\mathrm{H}_{i,j} = \sum_{i=1}^{N} m_l \frac{\partial \mathbf{R}_l}{\partial \mathrm{q}_i} \cdot \frac{\partial \mathbf{R}_l}{\partial \mathrm{q}_j} \tag{1-16}$$

where m_l is the mass of atom *l*, and the summation runs over all *l* atoms in the molecule.

Since the law of conservation of energy demands that the total energy (kinetic plus potential) is constant, we can combine equations (1-13) and (1-15):

$$\mathbf{M}\ddot{\mathbf{r}} + \mathbf{H}\mathbf{r} = 0 \tag{1-17}$$

where $\ddot{\mathbf{r}}$ is the acceleration (second derivative) of the atoms with respect to time. Equation (1-17) can be solved by using the following coordinate transformation matrix:

$$\mathbf{r} = \mathbf{T}\mathbf{u} \tag{1-18}$$

where $\mathbf{T}$ is an orthogonal transformation matrix (i.e. $\mathbf{T}\mathbf{T}^{\mathbf{T}} = \mathbf{T}^{\mathbf{T}}\mathbf{T} = \mathbf{1}$) and $\mathbf{u}$ is a coordinate vector in a different 3D space (dimension 3*N*) from the Cartesian space of the coordinates $\mathbf{q}$. $\mathbf{u}$ has the following time dependence for each of its elements:

$$u_k = C_k \cos(\omega_k \tau + \phi_k) \tag{1-19}$$

where the parameters C_k, ω_k and ϕ_k define the position of an atom in space (u_1, u_2, u_3, ..., u_{3N}) in the new basis set $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_{3N}\}$, whereas the original Cartesian coordinates (r_1, r_2, r_3, ..., r_{3N}) along the basis set $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, ..., \mathbf{X}_{3N}\}$. The new coordinates u_k are the normal mode coordinates [51]. C_k is the amplitude at time τ=0 for the *k*th vibrational mode, ϕ_k is the phase at time τ=0 for the *k*th vibrational mode and ω_k is the angular frequency for the *k*th vibrational mode.

1.5.2. Normal modes

A mode defines the direction along which atoms fluctuating at the same frequency are in phase. Each mode has a unique frequency of vibration; for each frequency there is a mode. Thus, there are *k* modes, one for each frequency:

$$\Delta r_i = \sum_{k=1}^{3N} T_{ik}\, u_k = \sum_{k=1}^{3N} T_{ik}\, C_k \cos(\omega_k \tau + \phi_k) \tag{1-20}$$

where Δr_i is the fluctuation of atom *i* with respect to its position in the equilibrium structure of the molecule. T_{ik} is the element in the T matrix associated with the *k*th mode (frequency), for atoms *i* through 3*N*.

The equation of motion is an eigenvalue problem:

$$\mathbf{HA} = \mathbf{\Lambda MA} \tag{1-21}$$

A normalization condition, $\mathbf{A^T HA = I}$ ensures that the Hamiltonian of the system can be diagonalized by the eigenmodes.

The eigenfrequencies of the Hamiltonian, ω_i, are contained in the elements of the diagonal matrix $\mathbf{\Lambda}$:

$$\omega_i^2 = \Lambda_{ii} \tag{1-22}$$

In conclusion, calculating the normal modes of a molecule comes down to diagonalizing the Hessian matrix **H**.

A simple construction of normal modes

Let us consider a simple, two-dimensional harmonic potential energy landscape (**Figure 1.3**). The two coordinates defining the space containing the two-dimensional energy landscape are r_1 and r_2. Two additional arrows - e_1 and e_2 - lie along the special directions defined by the normal modes. The shape of the potential energy well illustrated in **Figure 1.3** predicts how a random motion in the landscape will become deflected (say, towards r_2) by the modes e_1 and e_2. Any motion in the landscape is a composition trajectory of displacement in e_1 and e_2. Modes are independent: the only motion that is not a composite of e_1 and e_2 is a motion that lies purely along one of the modes. Oscillating motions - composite or pure - have frequencies associated with them that are related to the curvature of the potential energy well.

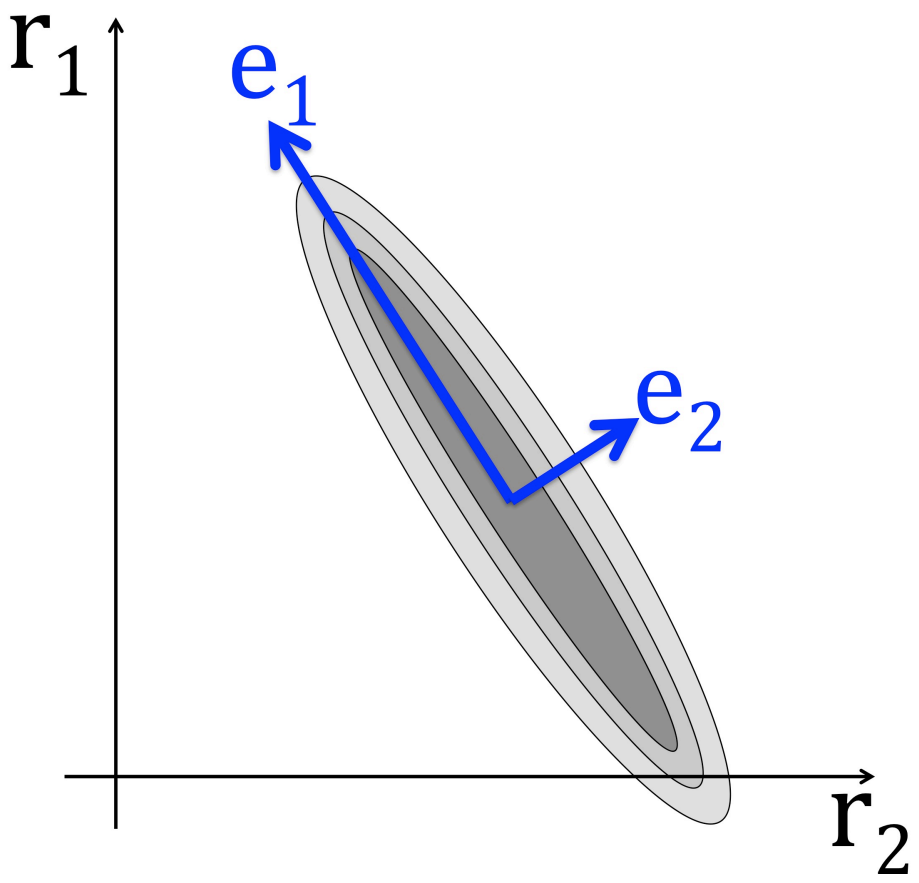


Figure 1.3. A two-dimensional harmonic potential energy landscape. The Cartesian coordinate axes r_1 and r_2 are shown as black arrows. The directions of the two normal modes lie along e_1 and e_2, blue arrows. Any motion in this landscape can be described by a linear combination of the components e_1 and e_2.

The harmonic potential is simple but reasonably accurate

In the pioneering work of Monique M. Tirion[50], it was shown that the normal modes of a large biomolecule subject to a Hookean potential energy function ($E \sim \frac{1}{2}\Delta\mathbf{r}^2$) rendered Cα-based root mean square (RMS) fluctuations of G-actin that were nearly indistinguishable from the RMS fluctuations obtained using Michael Levitt's L79 molecular mechanics potential energy function (equations (1-4) through (1-9)). B-factors are used during the refinement of the measured reflections from an X-ray diffraction experiment to account for thermal atomic vibrations and the different positions in different unit cells. Under certain conditions, B-factors are roughly related to the thermal vibrations of atoms about their mean position in space. Thus, dynamical calculations frequently compare simulated atomic fluctuations with crystallographic B-factors. For short MD simulations that fluctuate around a clear average geometry, or for ENM calculations, comparisons against B-factors usually enable qualitative assessments of calculated dynamics. MD simulations sampling conformational transitions can not be compared against B-factors; the average structure no longer has its intended meaning.

B-factors are related to RMS fluctuations[52]:

$$B_i = \frac{8\pi^2}{3}\langle(\Delta\mathbf{r}_i)^2\rangle = \frac{8\pi^2 k_{\mathrm{B}}\mathrm{T}}{\gamma}[\mathbf{\Gamma}^{-1}]_{ii} \tag{1-23}$$

where B_i is the magnitude of the *B*-factor for residue (Cα atom) i, $\Delta\mathbf{r}_i$ is the position fluctuation vector[c] for atom i, k_{B} is Boltzmann's constant (**Table VI-1,** see **Appendix VI**), T is temperature

[c] The atom's instantaneous position $\mathbf{r}_i$ with respect to the atom's equilibrium position, $\mathbf{r}^0_i$

in Kelvin, γ is the universal force constant applied to all interacting pairs of atoms. **Γ** is the Kirchhoff matrix constructed directly from the atomic coordinates of the protein:

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } \mathrm{R}_{ij} \leq \mathrm{R}_c \\ 0, & \text{if } i \neq j \text{ and } \mathrm{R}_{ij} > \mathrm{R}_c \\ -\sum_{j,j\neq i} \Gamma_{ij}, & \text{if } i = j \end{cases} \tag{1-24}$$

where R_{ij} is the distance between atoms *i* and *j*, and R_c is a cutoff distance that includes (Γ_{ij} = -1) or excludes (Γ_{ij} = 0) certain atoms in the molecular "network" of atoms from interacting with other atoms. R_c determines which pairs of atoms *i* and *j* are connected by springs.

Proteins with non-spherical (non-globular) shapes are unlikely to be modeled well by ANM. Conformational transitions such as protein folding can not be modeled by ANM. The magnitude and sense of the motions are arbitrary. In summary, ANM is simple and able to model the with reasonable accuracy, for certain biomolecular architectures, the shapes of the motions accessible to the ground-state geometry.

1.6. Overview of Projects: Chapter Summaries

In this dissertation, I sought to develop an atomistic model of the structural and dynamical principles governing sequence-specific protein-DNA binding, search and recognition. A mechanistic picture of binding, search and recognition is centrally important to understanding how genes are expressed and how artificial transcription factors used for genome editing bind to specific sequences. In **Chapter 2**, I review how the structural and dynamical properties of DNA are sequence-dependent and propose an idea that DNA has two codes: an information-genetic code and an energy-regulation code. The literature suggests that the energy code of DNA is "read" by regulatory proteins like transcription factors, and that the energy code predicts whether a regulatory will bind to a sequence better than traditional approaches that treat DNA as if it were a sequence of letters (information). To understand how regulatory proteins read the energy code, a tool was needed to quantify complementarity in the structure and dynamics of a protein and DNA. In **Chapter 3**, I developed a method based on geometry capable of accurately characterizing the complementarity of protein and DNA helical properties. In **Chapter 4**, I showed that the method is able to resolve subtle, previously unknown imperfections in helix complementarity between the genome editing targeting-protein TALE (transcription activator-like effector) and DNA. The results of the analyses provide a possible explanation for the observation that the specificity of TALE proteins for DNA depends on their construction and displays asymmetry with respect to the sequence being targeted (i.e. the TALE-DNA helix axis). How does protein-DNA helix complementarity change when a model DBP switches between specific and non-specific DNA-binding modes? In **Chapter 5**, I studied how a human transcription factor MTERF1 adapts its superhelical conformation when it switches from a

recognition mode to a search mode. MTERF1 spontaneously adapted from an extended conformation in recognition mode to a more compressed conformation in search mode. Does superhelix-DNA complementarity explain how MTERF1 slides on DNA in search mode? In **Chapter 6**, I found that MTERF1 slides by extending and compressing along DNA, stepping along a DNA sequence one contact at a time. Based on these results, I propose a new kinetic model of diffusion that can increase the rate of sliding despite retaining strong protein-DNA contacts. Overall, this dissertation presents a model of protein-DNA binding, search and recognition based on a simple principle: superhelix-DNA complementarity determines what functional mode a DBP is in and the dynamics of complementarity explain sliding.

1.6.1. The two codes of DNA

In the current era of genome sequencing, the time and resource costs of sequencing has plummeted to the point where it is almost routine to sequence whole genomes. Next-generation sequencing technology has produced an unprecedented sum of sequence information. Despite the richness of this information, the ability to predict functional elements such as transcription factor binding sites and nucleosome positions remains limited. Is the sequence of DNA letters enough to explain regulatory protein specificity? Here, a distinction between the information code of DNA and the energy code of DNA is made. The former is a lexicon encoding the genetic message while the latter is encoded by the subtle structure and dynamics of the double helix; direct readout decodes the letter code of DNA, indirect readout decodes the energy code of DNA. We review the energy code of DNA, specifically its ability to predict regulatory protein binding sites. Importantly, predictions of regulatory protein binding based on DNA structure and deformability were more successful than predictions based on DNA letters. Overall, the works

reviewed suggest that the energy code of DNA is essential for understanding the functions of genomes.

1.6.1. Characterization of biomolecular helices and their complementarity using geometric analysis

A general method is presented to characterize the helical properties of potentially irregular helices, such as those found in protein secondary and tertiary structures, and nucleic acids. The method was validated using artificial helices with varying numbers of points, points per helical turn, pitch and radius. The sensitivity of the method was validated by applying increasing amounts of random perturbation to the coordinates of these helices; 399,360 noisy helices were evaluated. In addition, the helical parameters of protein secondary-structure elements and nucleic acid helices were analyzed. Generally, at least seven points were required to recapitulate the parameters of a helix using our method. The method can also be used to calculate the helical parameters of nucleic acid-binding proteins, like TALE, enabling direct analysis of their helix complementarity to sequence-dependent DNA distortions.

1.6.2. Asymmetrically coupled structure specificity in protein-DNA complexes

Artificial nucleic acid-binding proteins are being designed to target arbitrary sequences of DNA to alter or rescue gene functions. Key to design success is highly specific, strong binding to one sequence; off-target nuclease activity by TALEN, CRISPR or ZFN could cause unforeseen, collateral damage. Therefore, understanding the driving forces of specificity is paramount to the future success of clinical gene therapies. Here, we uncover previously uncharacterized fractures

in helix complementarity between the superhelix of TALE proteins and the DNA to which they were specifically bound in sixteen extant X-ray crystal structures. These fractures were distributed asymmetrically along the DNA footprint, implying that previously reported affinity asymmetries of TALE proteins may arise from the subtle structural distortions in repeat-base partnerships.

1.6.3. A human transcription factor in search mode

Transcription factors (TF) can change shape to bind and recognize DNA, shifting the energy landscape from a weak binding, rapid search mode to a higher affinity recognition mode. However, the mechanism(s) driving this conformational change remains unresolved and in most cases high-resolution structures of the nonspecific complexes are unavailable. Here we investigate the conformational switch of the human mitochondrial transcription termination factor MTERF1, which has a modular, superhelical topology complementary to DNA. Our goal was to characterize the details of the nonspecific search mode to complement the crystal structure of the specific binding complex, providing a basis for understanding the recognition mechanism. In the specific complex, MTERF1 binds a significantly distorted and unwound DNA structure, exhibiting a protein conformation incompatible with binding to B-form DNA. In contrast, our simulations of apo MTERF1 revealed significant flexibility, sampling structures with superhelical pitch and radius complementary to the major groove of B-DNA. Docking these structures to B-DNA followed by unrestrained MD simulations led to a stable complex in which MTERF1 was observed to undergo spontaneous diffusion on the DNA. Overall, the data support an MTERF1-DNA binding and recognition mechanism driven by intrinsic dynamics of the MTERF1 superhelical topology.

1.6.4. Asynchronous shifts by asymmetrical modules bias how MTERF1 slides on DNA

DNA-binding proteins (DBP) can rapidly slide along DNA in search of a target sequence or mutation. Experiments have characterized the sliding rates, lengths and structure snapshots of a diversity of DBPs in search mode. The atomistic details of a sliding mechanism have eluded experimental structure characterization due to the fleeting nature of sliding. Our goal was to characterize the sliding mechanism of the human mitochondrial transcription factor MTERF1, a modular DBP with a superhelical tertiary structure that winds around DNA. Here, MTERF1 in search mode was studied using unrestrained, μs-timescale atomistic MD simulations. We found that MTERF1 modules established asymmetric contacts between the two DNA strands and along the DNA sequence. The contacts shifted asynchronously, suggesting that the modules diffused semi-autonomously. MTERF1 superhelical pitch – a metric of the relative orientation of modules – correlated with the overall position of the protein along the DNA. We propose that the sliding landscape is smoothed by asynchronous shifts of MTERF1 modules.

Chapter 2. The two codes of DNA

Acknowledgement

This Chapter constitutes a manuscript of a review article in preparation by myself, Alberto Perez, Miguel Garcia-Diaz and Carlos Simmerling. I conceived and wrote the manuscript, with edits and suggestions from the co-authors.

2.1. Introduction

A fundamental question in biology is how regulatory proteins bind to specific DNA sequences. In order to answer this question a full understanding of the energetics involved in binding is needed. Historically, direct readout has been used to understand recognition. Direct readout decodes the letters of DNA. In general, however, direct readout is frequently unable to explain specificity[53-57]. We can refine our understanding of recognition by taking into consideration indirect readout, which is the sequence-dependent structure and deformability of DNA. Direct and indirect readout are complementary sequence recognition mechanisms that read complementary sequence codes. The information code is recognized by direct readout, the energy code is recognized by indirect readout. These two codes of DNA can be used to understand protein-DNA binding specificity.

The goal of this section is to summarize recent advances in our understanding of indirect readout and its critical importance to protein-nucleic acid interactions and genetics. Quantitative models of indirect readout and the DNA energy code will become increasingly important as genome-editing and personalized medicine advance. In the future, these technologies should have a significant impact on genetic diseases and cancer. In this review, I will summarize the following topics as they relate to the aforementioned concepts: (1) DNA sequence information and energy, (2) direct and indirect readout, (3) experimental and computational models of the DNA energy code, (4) examples where the energy code predicts regulatory sequences better than traditional information-based methods, (5) DNA energy is constrained by evolution in humans, and (6) methylation switches DNA deformability.

2.2. Direct readout decodes DNA sequence letters

The traditional mechanism of direct readout was discovered through the early crystal structures of protein-DNA complexes[58-60], in which a few amino acids formed nucleotide-sequence specific contacts with the DNA. These interactions resembled the hydrogen bonding patterns observed in Watson-Crick base pairs. Therefore direct readout is a natural extension to protein-DNA interactions, in which amino acid letters can be paired with nucleotide letters. Information theory is ideally suited for decoding letters.

2.2.1. The letter representation of a DNA sequence is an information code

Direct readout can be described as a digital decoding of DNA sequence information by a regulatory protein: *Is this nucleotide an adenine or not*[61]*?* **Figure 2.1** illustrates the decoding process of a sequence of letters by a regulatory protein. The regulatory protein can bind to any DNA sequence ($\mathbf{\Omega_{seq}}$). For this example, the length of binding site is six letters. A combinatorial DNA sequence is drawn in **Figure 2.1a** that represents all possible sequence combinations (4^6 sequences) the protein might bind. For every sequence the protein binds and decodes (Ω_{read}), *information* is gained in proportion to the similarity between the decoded sequence and the target sequence[61] ($\Delta\Omega = \mathbf{\Omega_{seq}} - \Omega_{read}$). No information is gained when the protein binds the wrong sequence because the decoded sequence is nothing like the target sequence it expected (**Figure 2.1b**). Maximum information is gained when the protein binds the target because all six letters in the sequence the protein has evolved to recognize are present (**Figure 2.1c**). Despite the popularity of using information-visualization tools such as weblogos[62] to define sequence specificity, is this model of regulatory protein-DNA binding specificity detailed enough?

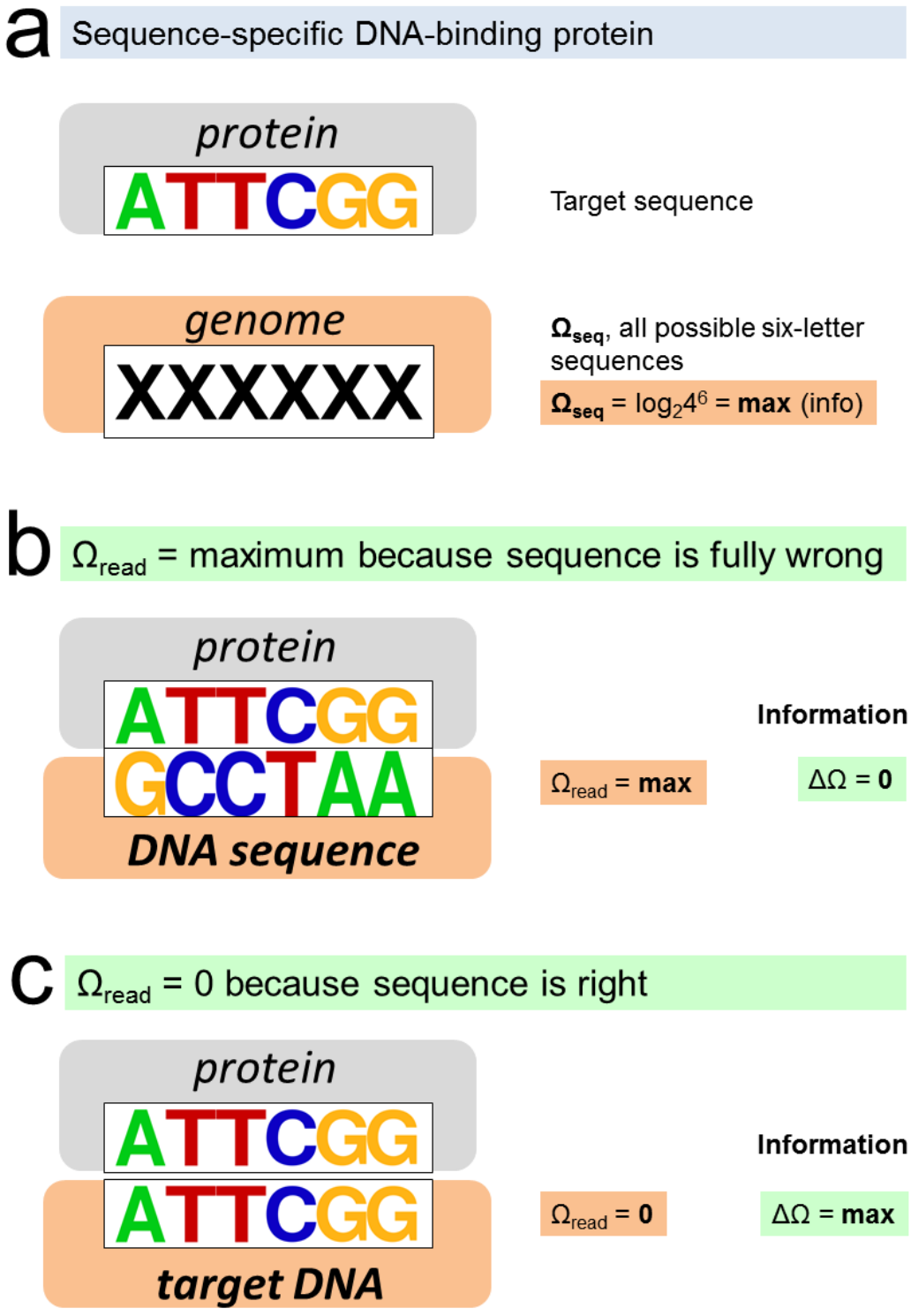


Figure 2.1 Direct readout described by information theory: A regulatory protein (grey) recognizes one specific DNA sequence. (**a**) The protein can bind any six-letter DNA sequence (X=A, T, C, G). The maximum information that can be gained ($\mathbf{\Omega_{seq}}$) depends on the number of possible six-letter sequences. (**b**) The protein binds a wrong sequence. Ω_{read} is maximum. No information is gained ($\Delta\Omega$=**0**). (**c**) The protein binds the target sequence, the Ω_{read} is 0. Maximum information is gained ($\Delta\Omega$=**max**).

Using information theory in this way assumes that a nucleotide in a sequence is independent of the identity of its neighbors[63] (sequence context). However, intrinsic biases are present in DNA that do prefer certain sequence contexts to others. For example, the positions of histones can be partially predicted by observing sequences where every tenth letter is an C followed by a A[64]; because CA (or TG on the complementary strand: 5'-CA-3':5'-TG-3') enables DNA to bend more easily than alternate dinucleotide steps[65] and DNA completes a helical spiral every tenth base, CA/TG steps permit DNA to adopt a geometry that can be more easily wrapped around a histone than other sequences. In some crystal structures of histone-DNA complexes, kinked segments of DNA that permit the DNA rod to wrap around the histone are always CA/TG[66]. Statistical analyses of experimental structures indicate that these CA/TG sequence patterns are highly enriched in histone-DNA complexes.[67-69] Therefore, a C followed by an A in a pattern that faces the minor groove of DNA towards the histone has been selected by evolution. The binding sites of hundreds of transcription factors reveal an abundance of sequence patterns. Why might evolution have selected these patterns? Why might proteins be expected to read these DNA "letters"?

2.2.2. Proteins read letters with direct readout

Direct readout is a useful and simple model of predicting whether a regulatory protein binds a sequence of DNA letters. The two most popular genome-editing reagents, clustered regularly interspaced short palindromic repeats (CRISPR) and transcription activator like effector nuclease (TALEN), utilize direct readout to confer sequence specificity. In CRISPR, direct readout is mediated by Watson-Crick hydrogen-bonding between the guide RNA and the target sequence[70], while TALEN proteins utilize amino acids to hydrogen bond to the target

sequence[71]. The TALEN family of transcription factors uses the following direct readout cipher[72]: Asp and Ile with A, His and Asp with C, and Asp and Gly with T. For example, the direct readout patterns for the TALEN protein antivirulence factor in the *Bs3* gene (AvrBs3) and its target DNA sequence in the UPA box is shown in **Table 2.1.** It is important to note this readout cipher is not deterministic because different amino acids can recognize the same nucleotide, and one amino acid can recognize multiple bases (e.g. Arg can read all four bases[73]). Structural constraints are assumed to have been met, in order that the amino acids and bases are arranged for direct readout. What are the structural constraints required for direct readout to occur?

Table 2.1. The AvrBs3 direct readout code[72].

Amino acids																	
His Asp	Asn Gly	Asn Ser	Asn Gly	Asn Ile	Asn Ile	Asn Ile	His Asp	His Asp	Asn Gly	Asn Ser	Asn Ser	His Asp	His Asp	His Asp	Asn Gly	His Asp	Asn Gly
Nucleobases																	
A	T	A	T	A	A	A	C	C	T	**X**	**X**	C	C	C	T	C	T

X=A, C, T or G.

Direct readout occurs via the DNA grooves[74], where nucleobase functional groups are exposed (**Figure 2.2a,b**). The floor of the major groove presents a unique chemical signature for each of the four canonical patterns of base pairs – A:T and T:A, G:C and C:G (**Figure 2.2c**). In contrast, in the minor groove of DNA, A:T and T:A are indistinguishable as are G:C and C:G (**Figure 2.2d**). The four possible base pairs present the same moiety at the first and third positions (H-bond acceptor). In the middle position, A:T and T:A display a non-polar H atom while G:C and C:G display an H-bond donor. Despite there being a degenerate binary code in the

minor groove (with only one moiety to check), proteins can bind the minor groove with high specificity (also using shape readout[75], see below). Rarely does direct readout confer specificity due to the paucity of hydrogen bonds between the protein and the nucleobases in the DNA sequence[70-71]. Even the recognition cipher of TALEN proteins depends on sequence context features that a direct readout cipher cannot predict[76]. To explain these phenomena, we now turn our attention to the DNA energy code.

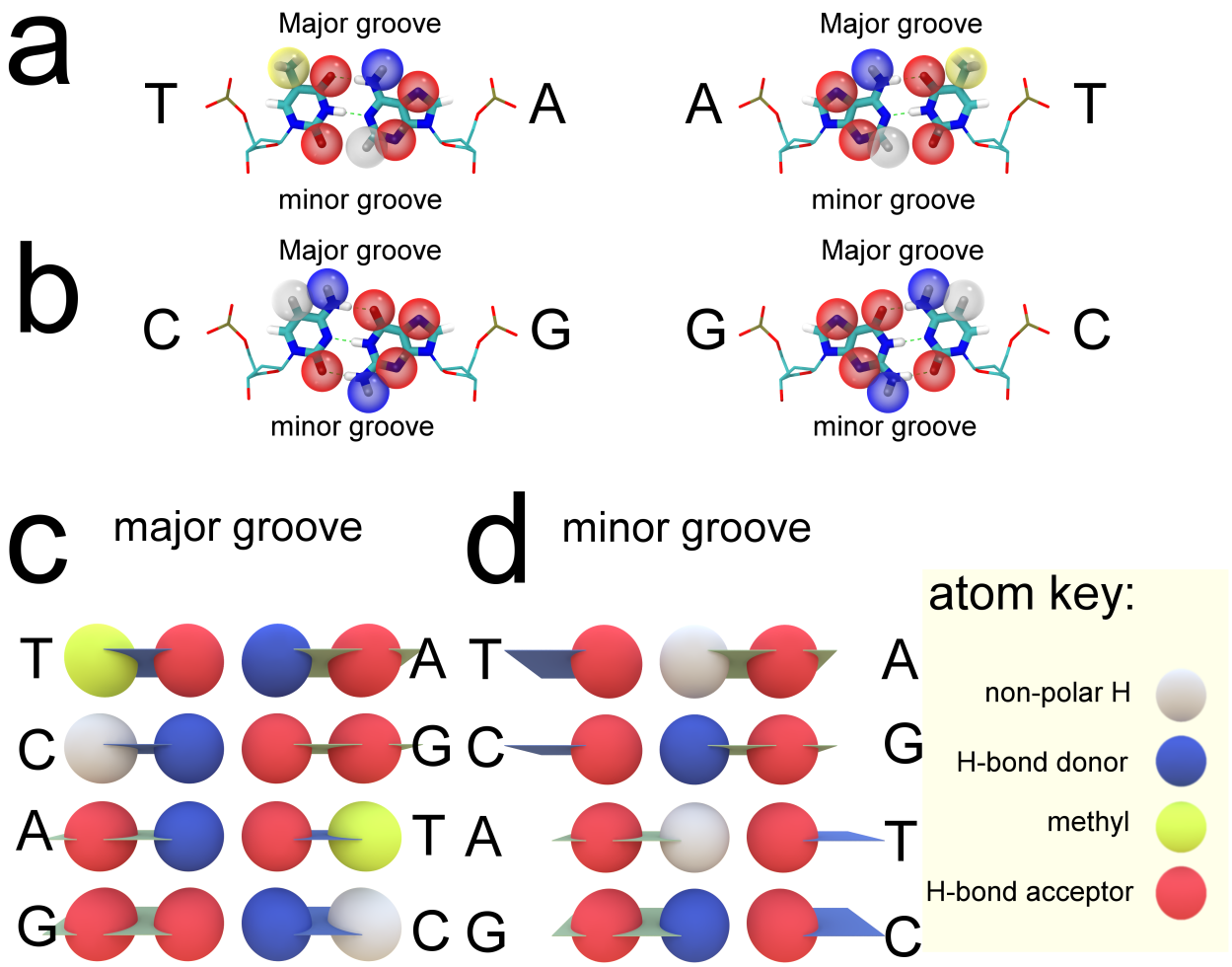


Figure 2.2. A digital DNA code. The four possible pairs (**a**) T:A and A:T, and (**b**) C:G and G:C display functional groups into the major and minor groove. (**c**) Three functional groups are present in the minor groove (hydrogen-bond donor, hydrogen-bond acceptor, and a non-polar hydrogen atom). A:T and T:A are indistinguishable as are G:C and C:G, and all four display an hydrogen-bond acceptor at the first and third position. (**d**) The floor of the major groove presents four moieties (a non-polar hydrogen atom, an hydrogen-bond donor, an hydrogen-bond acceptor, or a methyl group). Each of the four bp pair displays a unique pattern.

In previous sections, we highlighted examples where a letter-based model is unable to explain protein-DNA binding specificity. In the remainder of this review, we focus on the subtler properties of DNA that are also involved in driving a protein to read a DNA sequence. What are the properties of DNA sequences that can cause indirect readout? Might the genetic code of

sequence information be different from a second "code" that regulatory proteins "read" to bind a specific DNA sequence?

2.3. The energy code of DNA

Proteins can read the energy of a DNA sequence because the structure of DNA (double helix with two grooves), the flexibility of DNA (thermal vibrations) and the deformability of DNA (protein-induced conformational change) depends on the sequence of DNA. The term "energy code" describes the structure, flexibility, and deformability of DNA. Watson-Crick hydrogen bonding between the bases of the two strands forms the duplex, while dispersive forces drive nucleobase stacking that help stabilize the helical geometry of the DNA duplex[77-78]. Water molecules and ions condense in the grooves[79], screening the electrostatic repulsion between the negatively charged backbones of the DNA strands[80-81]. Fluctuations in the structure of water and ions around DNA can shift the preferences for stacking interactions in a sequence-dependent manner[82]. These are a few of the driving forces of DNA structure, flexibility and deformability. The essential geometric properties differentiating the sequence-dependent structures of DNA can be measured to better understand why regulatory proteins prefer to bind certain sequences of DNA.

2.3.1. Helicoidal parameters measure changes in DNA structure

Local DNA structure, flexibility and deformability can be characterized by the six bp-step helicoidal parameters[83]. For each bp-step, there are three translations and three rotations (**Figure 2.3**).

The translations are characterized by:

- **Rise**: displacement along the helical axis
- **Shift**: displacement in or out of the groove floors
- **Slide**: displacement from side-to-side between the strands

The three rotations are characterized by:

- **Roll**: rotation around the axis connecting the two strands
- **Tilt**: rotation around the axis pointing out of the grooves
- **Twist**: rotation around the helical axis

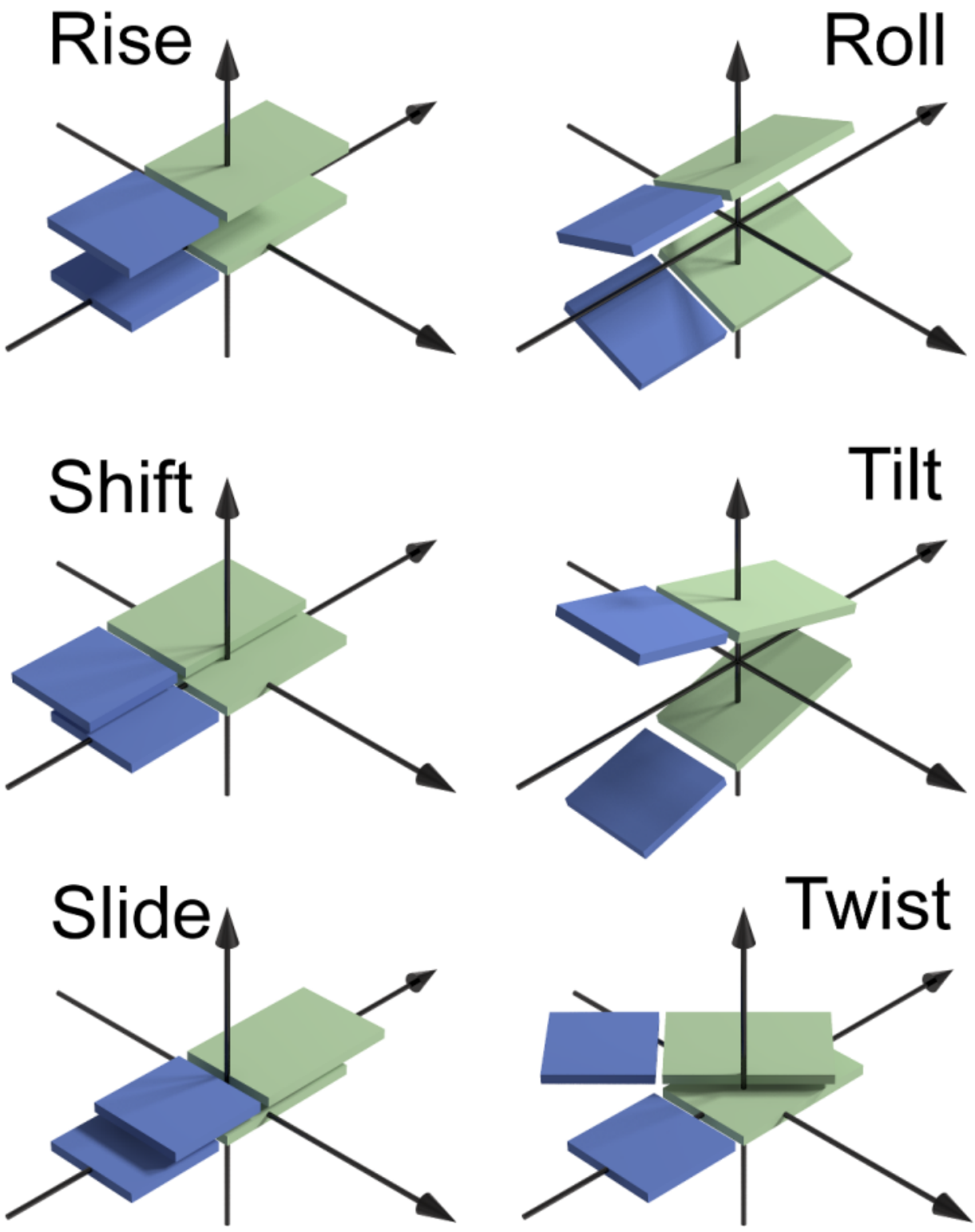


Figure 2.3. The bp-step helicoidal parameters simplify the geometry of DNA structure and flexibility. Green planes represent purines and blue planes represent pyrimidines. The upward pointing vector corresponds to the helical axis.

2.3.2. Proteins bind specific DNA structures

Energy landscapes connect the structure, flexibility and deformability of DNA to energy (which itself depends on sequence). DNA structures sampled due to thermal motions define the ground-state ensemble of the sequence, and flexibility is the degree of conformational diversity in the ensemble. DNA can also be driven into a deformed conformation (excited state) that is higher in energy than the ground state, but is still a local minimum in the energy landscape. Protein-DNA binding minimizes the free energy of the complex, and the unfavorable strain energy incurred by deforming DNA can be offset by favorable contributions to free energy from increased DNA entropy (flexibility) and protein-DNA contacts.[20] Entropy contributes favorably when DNA is deformed because ions are no longer immobilized in the grooves and the stacking interactions that rigidified the helical rod are weakened,[20] permitting DNA to freely flip and flop.

Shape readout[84] ("static indirect readout"[85]) can be used to explain how some proteins bind the ground state structure of a specific DNA sequence. Specificity arises because the protein binds a specific conformation of DNA that is stable for the target sequence, but would be energetically unstable for alternate sequences (**Figure 2.4**). Minute differences in the structure of non-target DNA sequences compared with the target sequence can distort the geometric partnership required to form direct readout H-bonds between the protein and the target sequence.[84]

Shape readout explains how a sequence-dependent structure also confers a unique electrostatic surface surrounding the DNA[75]. The pattern of the electrostatic potential painted on the surface of the DNA structure provides two complementary properties that a protein can specifically bind. For example, AT-rich sequences display narrowed minor grooves, which in

turn enhances the negative electrostatic potential in the groove[84]; the size of the groove is just-right for the positively charged guanidino group of an Arg amino acid to fit[84].

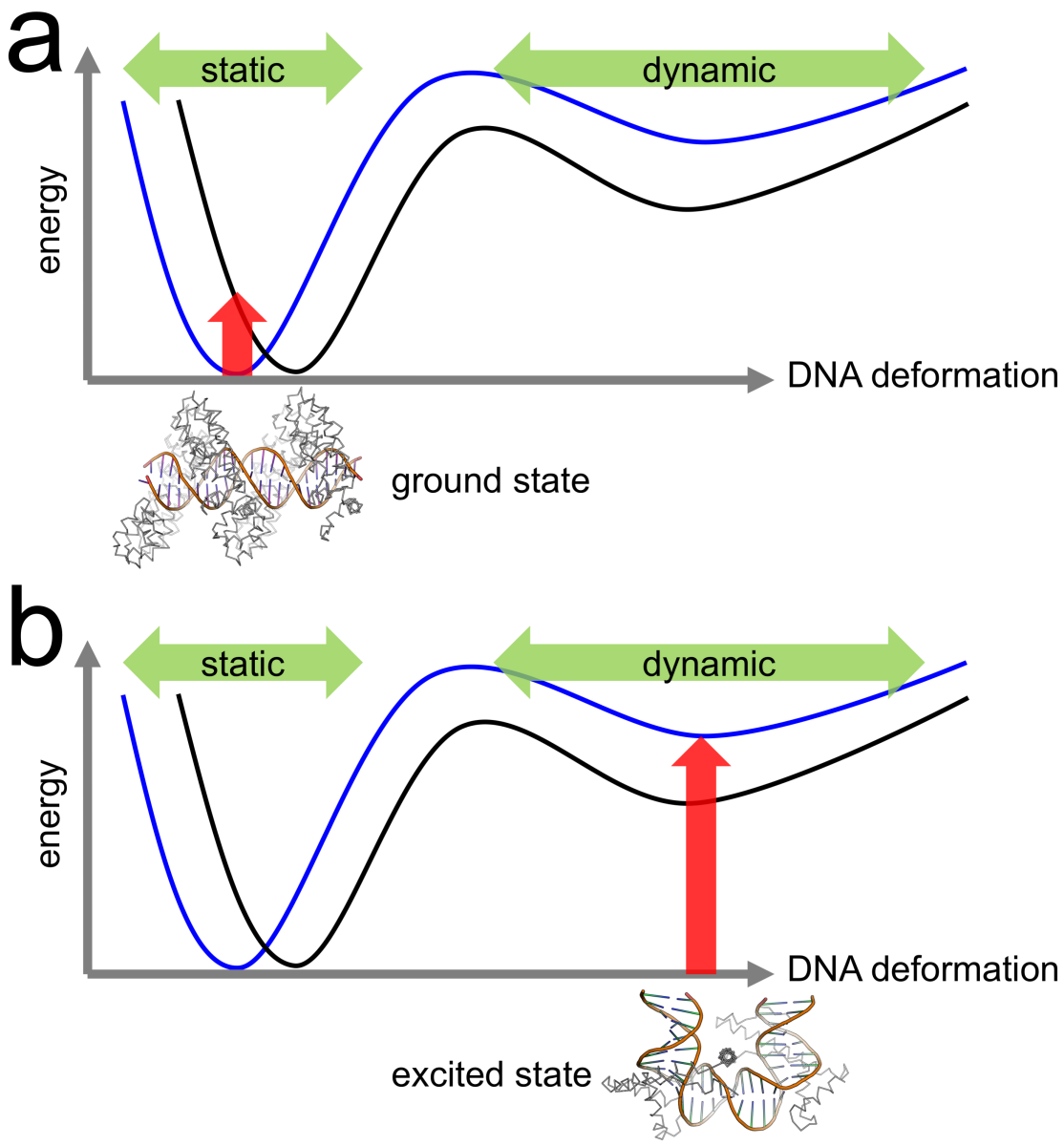


Figure 2.4. Static and dynamic indirect readout recognition mechanisms read the sequence-dependent energy of DNA sequences. (**a**) Specificity from static indirect (shape) readout arises from the difference in energy between the ground state structures of two DNA sequences (blue and black energy landscapes), given the specific structure bound by a protein. (**b**) Specificity from dynamic indirect readout arises from the difference in energy between the excited state structures of two DNA sequences that have been deformed.

Proteins that induce excited-state conformations of DNA may recognize the sequence using "dynamic indirect readout"[85]. Protein-DNA binding specificity in dynamic indirect readout arises when a particular DNA sequence is induced by the protein to adopt a specific conformation that is lower in energy than other DNA sequences in the same specific conformation[86-87] (**Figure 2.4**). Thus, proteins that bind excited-state conformations of DNA may utilize dynamic indirect readout. Direct readout, static indirect readout and dynamic indirect readout are not mutually exclusive sequence recognition mechanisms; a sequence specific protein-DNA complex may include all three. Dynamic indirect readout is the mechanism of recognition of deformed DNA structures that are not readily accessible in the absence of a protein inducing the DNA deformation[85]. Static and dynamic indirect readout are subtly coupled to direct readout because the geometry needed to form direct readout contacts depends on the ability of the protein and the DNA to adopt compatible conformations. Given the vast number of high-resolution experimental structures of DNA in the Protein Data Bank, can flexibility patterns be inferred if multiple structures are available for a sequence?

2.4. Quantitative Models of the DNA Energy Code

2.4.1. Experimental approaches to characterizing the DNA energy code

In seminal work, Olson et al. characterized the energy code of DNA by analyzing a database of experimentally elucidated structures of different DNA sequences.[88] It was assumed that the populations of the six helicoidal parameters (**Figure 2.3**) for each of the ten possible

dinucleotide sequences (two consecutive letters; e.g. AA) were normally distributed with one peak (unimodal). This permitted the energy code (landscape) to be calculated from the probability distributions of helicoidal parameters for dinucleotide sequences. The distributions were converted to energy using inverse harmonic analysis (the distribution of the length of a vibrating spring is directly proportional to the physical properties of the spring: the mean of the distribution is the equilibrium length, the variance of the distribution is proportional to the force constant). **Figure 2.5** visualizes the results of Olson et al.'s analysis (conformational volumes) for each of the 10 dinucleotide steps. Larger volumes sample a greater diversity of conformations given the same energy (paraboloids have the same height but different slopes). Olson et al could not calculate the energies for AC and GG because no experimental structures contained them.

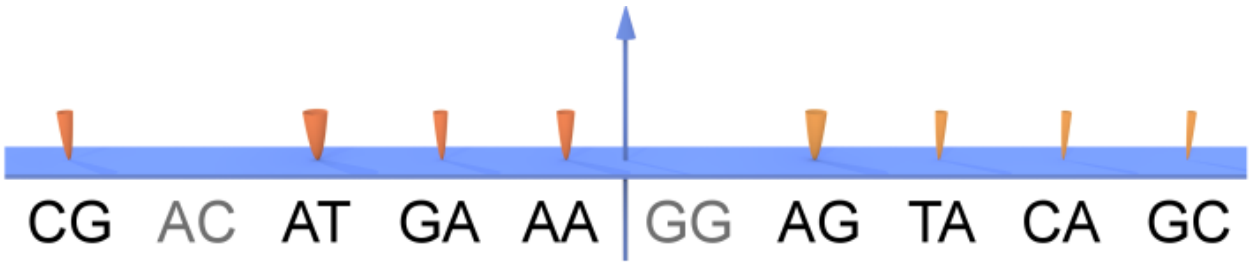


Figure 2.5. A ground-state energy code: conformational volumes of dinucleotide steps. The volumes of the paraboloids represent relative flexibilities. The shapes are based on Olson et al.'s data38. Two dinucleotide steps – AC and GG – were not reported because too few structures were available.

Two key experimental limitations faced by Olson et al two decades ago remain today: For each sequence combination (10, dinucleotide; 136 tetranucleotide) hundreds of structures are required to estimate the energy code; Inverse harmonic analysis assumes that DNA sequences have one stable structure, precluding the prediction of a sequence having metastable structures,

which may be important for protein-binding. A similar harmonic analysis based on NMR structures can be used,[89] but the combinatorial expansion beyond the dinucleotides (10 combinations; tetranucleotides have 136 combinations) imposes a severe limitation on most experimental approaches. The average shape of a DNA sequence can be characterized using hydroxyl radical cleavage assays[90], however the results do not readily divest information on flexibility - only static shape - nor is it straightforward to produce a structure model of DNA from the cleavage pattern results. An alternate, computational strategy to sample structure and flexibility of DNA is atomistic molecular dynamics (MD) simulations, the accuracy of which is beginning to become comparable to high-resolution experimental techniques like X-ray crystallography and NMR[91-94].

2.4.2. Atomistic molecular dynamics simulations fill gaps in the energy code left by experiment

The limitations discussed above left an important question open: How important is sequence context to flexibility of a dinucleotide step? That is, what is the energy code of the tetranucleotide sequences? To address this question, Pasi *et al*. mapped the ground-state DNA energy landscape (**Figure 2.4a**) using molecular dynamics (MD) simulations that generated 35 million conformations of the 136 unique tetranucleotide sequences (e.g. AAAA)[95]. The authors' analyses of the helicoidal parameter distributions of the tetranucleotides revealed a remarkable result. The flexibility of a dinucleotide could be affected almost as strongly by changing the flanking bases, as by changing the sequence of the dinucleotide itself. That is, sequence context can have almost as strong of an effect on DNA flexibility as sequence itself[95]. Pyrimidine-purine

(YR) steps preferentially populated low-twist negative-slide states while purine-purine (RR) steps preferentially populated high-twist negative-shift states[95]. Sequence context of the dinucleotides determined that YR and RR steps populated the canonical B-DNA state in twist, shift and slide - 33°, 0 Å and 0 Å, respectively[95]. These results show that the DNA energy code depends on sequences at least four nucleotides in length. Importantly, Pasi *et al.*'s results show that DNA letters are not enough to understand the flexibility of the simplest sequences (dinucleotides). In the absence of high-throughput high-resolution structural experiments, atomistic MD simulations will continue to play a critical role in studying the relationship between the energy code of DNA and the regulation of gene expression and genome maintenance.

2.5. The energy code predicts regulatory protein binding sites

2.5.1. Ground-state DNA energy landscapes are imprinted with protein binding patterns

Pederson *et al*. initially suggested that the structure and flexibility of DNA could play an important role in predicting the activity of promoters[96]. Due to the limited scope of this work, Goñi and coworkers followed up with larger datasets (more structures, more sequence)[97]. Goñi et al. developed a physics-based approach trained against the flexibility profiles inferred from atomistic MD simulations to identify transcription start sites (TSS).[97] To map the energy code of different DNA sequences, the method was trained on helicoidal parameter distributions obtained

from molecular dynamics simulations[97]. When their approach first appeared, the authors discovered multiple TSSs that had not yet been annotated in the reference dataset of promoter activities. These predictions were initially denoted as false positives[97]. To understand why Pro-Star exhibited the highest proportion of correct predictions compared to 11 information-based approaches[97], the expression experiments were repeated[98]. The new experiments – cap analysis gene expression (CAGE), luciferase assays and RNA-sequencing showed that in fact Pro-Star (http://mmb.pcb.ub.es/proStar/) was *correctly* predicting true promoters – the previously denoted false-positives[97] were verified as true-positives[98]. These results underscore the usefulness of energy-based approaches to understanding sequence data and using it to predict regulatory protein binding sites. Would such an approach also be able to predict the binding of nucleosomes, which induce significant DNA deformations?

2.5.2. Dynamic readout predicts nucleosome phasing

Due to the vast size of the nuclear genome, nucleosomes are essential for packaging DNA into the confines of the nucleus, and their presence on the DNA can abrogate regulatory protein binding or transmission of a transcribing polymerase past the core particle, thereby regulating gene expression[99]. Nucleosomes are composed of a quartet of histone homodimers (H2A, H2B, H3 and H4) with a squat-barrel shaped quaternary structure around which 147 bp of DNA wrap[66]. Maintaining the accessibility of regulatory sites and active genes can be partially achieved if the energy to wrap DNA around a histones is less favorable (decreasing the probability that histones will bind) than another sequence[100-101].

With remarkable accuracy, Minary and Levitt predicted the nucleosome occupancy profile for yeast chromosome 14, using molecular mechanics simulations to estimate the energy of forming these histone-DNA complexes[86]. Their protocol for characterizing dynamic indirect readout is summarized in **Figure 2.6**. The energy associated with the histone-DNA complex, the wrapped DNA (without the histone) and the linear DNA were compared to calculate the propensity of nucleosome formation for a given DNA sequence. The premise of the approach was that difference in the internal energy of different DNA sequences in the same conformation is proportional to the relative probability that those sequences will adopt them (**Figure 2.4b**); only the difference in energy of the two end states matters because the approach is concerned with thermodynamics rather than kinetics. Finally, Minary and Levitt compared the predicted nucleosome formation probability from the computed energies (via Boltzmann's equality) with in vivo and in vitro occupancies[86]. The computed nucleosome formation energies correlated remarkably well with the in vitro nucleosome occupancy profiles ($R^2 \sim 0.6$)[86]. Minary and Levitt went on to decompose the energy to uncover the forces that drove nucleosome formation. They found that the internal energy of the DNA when wrapped around the histone, not the energy of interacting with the histone, was the main driver of nucleosome formation[86]. This suggests that dynamic indirect readout – the deformation energy of DNA – dominates nucleosome formation propensity, and that direct readout (protein-DNA contacts) plays a relatively small role.

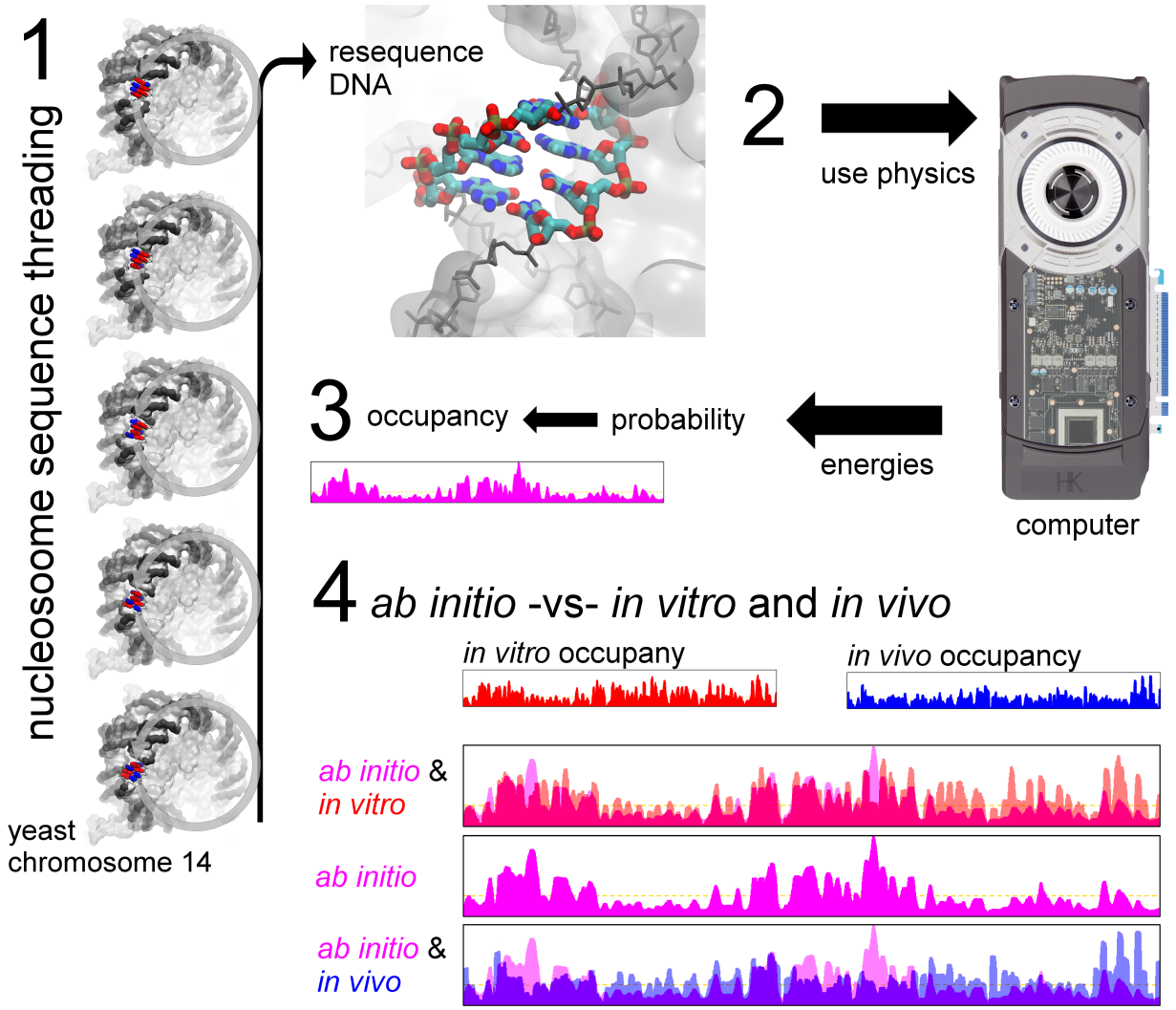


Figure 2.6. Dynamic readout predicts nucleosome formation. Minary and Levitt's approach for sequencing a DNA energy code: (1) 20,000 bp of yeast chromosome 14 was threaded over a nucleosome core particle (PDB ID: 1KX5[102]) by replacing atoms in nucleobases to artificially and rapidly mutate the DNA, (2) using a physics-based approach. (3) The relative energies (see text) were converted to relative probabilities[86] using Boltzmann's equality ($P_i/P_0=\ln\Omega_i$). (4) The predicted occupancy (pink) is overlaid on the observed in vitro (red) and in vivo (blue) nucleosome occupancy profile, for ~20,000 bp of yeast chromosome 14 [86]. Data (4) adapted from reference[86].

2.6. Methylation acts as a DNA flexibility switch

Methylation of C to 5-methyl-C (5C) can regulate gene expression. Minary and Levitt showed that methylation switches the propensity of some DNA sequences to form nucleosomes, in excellent agreement with the observed footprinting assays[86]. Their physics-based predictions (**Figure 2.6**) that methylation toggles the energy code of DNA is substantiated by MD simulations[103], Förster resonance energy transfer (FRET)[104-105], native gel mobility[104], atomic force microscopy[106], solid-state nanopores[107], and single-molecule studies[108]. Results from these diverse methods show that 5C, like the rest (A, T, C, G), can encode regulatory function by toggling the DNA energy code.

2.7. Summary of the two codes of DNA

Figure 2.7 summarizes the concepts presented in this section. The information code of DNA is a lexicon of letters, and the mechanism by which proteins read the letters of DNA is direct readout. The energy code of DNA can be broken down into two regimes, the ground-state and the excited-state energy landscapes. The ground-state energy landscape corresponds to the structure and flexibility of bare DNA. Proteins read the ground-state energy code of DNA via shape readout. Proteins read the excited-state energy code of DNA via dynamic indirect readout,

by inducing DNA structures the bare sequence would otherwise not adopt. Therefore, protein-mediated gene regulation relies on at least three mechanisms of readout to bind specific DNA sequences. Information and energy are degenerate because similar sequences of letters can have different energy landscapes while different sequences of letters can have similar energy landscapes.

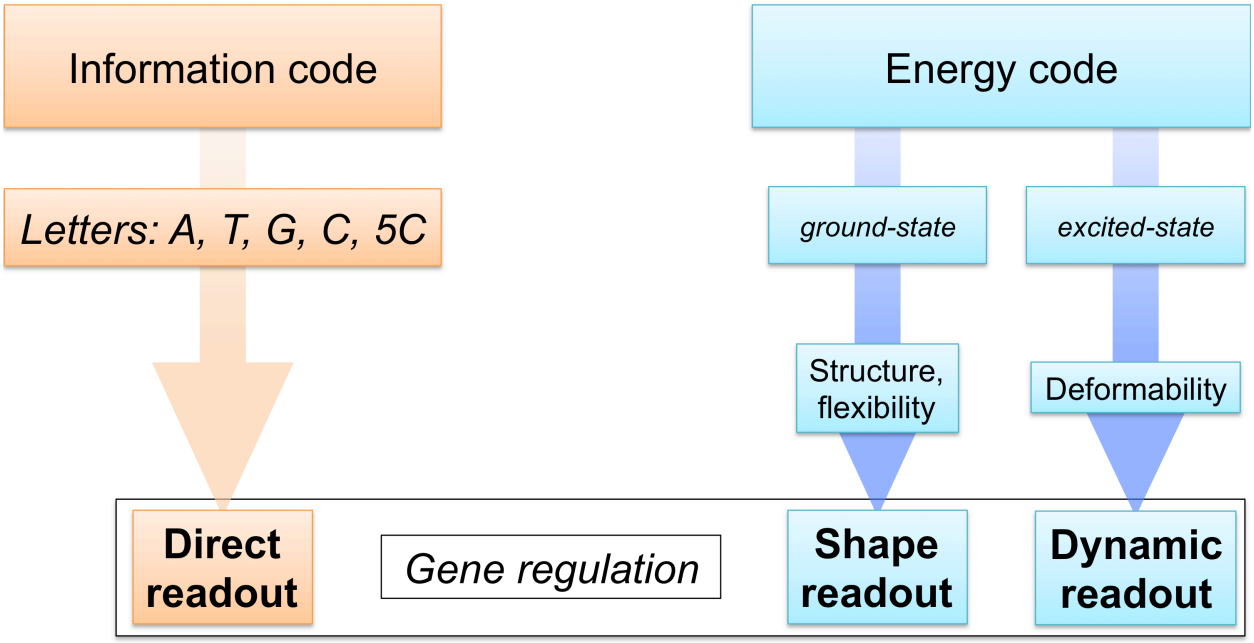


Figure 2.7. Summary of the two codes of DNA. "5C" is 5-methylcytosine.

2.8. Conclusion and perspective

An advanced understanding of the DNA energy code has the potential to help drive discoveries in genetics. Although it has been challenging to integrate biophysical techniques

directly into genome-scale models, advances in computing alongside advances in experimental approaches are beginning to enable genome-scale annotations of the DNA energy code[109].

The energy code of DNA depends on sequence context. It might therefore not be correct to utilize information theory alone to assess DNA function from patterns of letters, because information theory assumes that the observed frequencies of a letter (nucleotide) in a DNA sequence is fully independent of the rest of the letters in the sequence. This might help explain the challenges[110-111] faced by information-based methods.

Calculating the DNA energy code will help us understand the function(s) of a DNA sequence and how the function(s) are conserved. Sequence conservation analysis of transcription factor binding sites[112] within regulatory networks[113] can be analyzed in parallel for their letter and energy patterns, perhaps revealing a greater degree of similarity - or difference - between humans and other species. The energy code may also provide a means to calibrate high-throughput sequencing technologies with low sampling depths[86]. Information and energy provide distinct yet complementary pictures of sequence data that can be used to understand the functions of DNA. Computational methods like atomistic and coarse-grained molecular dynamics simulations will play an ever more important role in "sequencing" the DNA energy code, while providing training data for information-based methods.

Chapter 3. Characterization of biomolecular helices and their complementarity using geometric analysis

Acknowledgement

This Chapter constitutes a manuscript in preparation by myself, Yiqing Elissa He, Miguel Garcia-Diaz, Carlos Simmerling and Evangelos Coutsias. Professor Coutsias led me through the mathematics. I conceived and wrote the manuscript, with edits and suggestions from the co-authors.

3.1. Introduction

The helix is a ubiquitous geometric form in structural biology. It is essential to characterize the geometric properties of helices traced by atoms in a biomolecule because function is driven by structure and dynamics. There are two major problems with characterizing the geometric parameters of biomolecular helices. Irregularities between the atoms that trace the helix make it difficult to fit the points directly to the parametric equations of a helix. Also, biomolecular helices frequently trace less than one helical turn. Few methods exist that can derive accurate helical parameters from one-turn helices[114], and irregularities make the problem even more difficult.

In structural biology, empirical methods are available that can characterize the geometric parameters of irregular helices, but are only applicable to the molecular fragments for which they were trained. For example, DSSP[115] assumes a specific chemical connectivity associated with peptide secondary structures (e.g. integral H-bonding between amides). HELFIT[116] and HELANAL[117] assume Cα-Cα chain connectivity in peptides to define an internal coordinate system (i.e. interlinked torsions along a chain of nine sequential Cα). For nucleic acids, 3DNA[118] and Curves+[119] find a helical axis by RMSD-fitting a nucleobase pair to a reference base pair structure whose helical axis is pre-defined[120].

A general method that does not require empirical constraints - such as a prior knowledge of the helical axis, helix radius or chemical topology - and is less affected by helical irregularities would enable the analysis of any biomolecular helical geometry. Such a method is needed to characterize the structures of superhelical nucleic acid binding proteins (NBPs) and modular superhelices[38]. The only such method known to us was developed by Nievergelt, a total least squares (TLS) approach that can characterize irregular helices consisting of points comprising a helix that traces just 90°[39]; for context, a single amino acid in a helical secondary structure element (SSE) traces ~100°.

The method we introduce here is an extension of TLS into a 2D linear least squares form. The simplification from 3D to 2D leads to several advantages; mainly, requiring fewer points to achieve the same level of accuracy. Like TLS, our method provides the ability to accurately calculate, without constraints, the helical parameters of helices traced by irregularly spaced points. Our method can also be used to characterize the helices traced by superhelical NBPs along with the nucleic acids they bind. The resulting helical parameters for the protein and the nucleic acid are directly comparable, providing a novel tool to analyze complementarity of the

helical geometries involved in protein-nucleic acid binding. Such a method is likely to prove useful as the need to design highly specific genome editing enzymes[121] that recognize sequence-dependent DNA helix distortions continues to rise.

3.2. Theory

A cylindrical helix projects a circle on the *x*-*y* plane only if the *z*-axis is parallel to the helical axis. We deconstruct the problem into four parts. In part **3.2.1**, a general rotation matrix is derived. The algebra is simplified by assuming that the unit plane being rotated always passes through the origin of the coordinate systems (original and rotated). In part **3.2.2**, the rotation matrix is cast in spherical coordinates. In part **3.2.3**, the fitting problem is posed as a linear least squares problem and its solution described. In part **3.2.4**, we describe how the helical parameters are obtained from the optimized helix frame. The derived parameters represent the best helical curve through which the user-supplied points pass; the more points per helical turn, the smoother the helix will appear.

3.2.1. Rotate a helix in 3D, then project the helix on the *x*-*y* plane

The optimal helical axis must first be located in 3D space. The helical axis is defined as the normal vector to the plane onto which the helix projects a circle. To find this plane, a rotation matrix, **R**, is needed that relates the old coordinates of the points to rotated coordinates of the points, $(x,y,z) = \mathbf{R}^*(X,Y,Z)$. The original coordinates of the points are denoted (X,Y,Z) and the rotated coordinates are denoted (x,y,z).

A unit normal vector, $\hat{n}$, representing the plane ($aX + bY + cZ = 0$) containing the coordinates (X,Y,Z) is defined using direction cosines, $\hat{n} = (a,b,c)$. The basis vectors ($\hat{e}_X,\hat{e}_Y,\hat{e}_Z$) and ($\hat{e}_x,\hat{e}_y,\hat{e}_z$) represent the frame of the original coordinates (X,Y,Z) and the rotated coordinates (x,y,z), respectively. The unit vector $\hat{e}_z$ is chosen to be the unit normal vector of the plane (a,b,c):

$$\hat{e}_z = a\hat{e}_X + b\hat{e}_Y + c\hat{e}_Z \tag{3-1}$$

We define a right-handed coordinate system:

$$\hat{e}_X \cdot \hat{n} = a \tag{3-2}$$

The component of $\hat{e}_X$ that is parallel to $\hat{e}_z$ is

$$\hat{e}_{X\parallel} = a\hat{e}_z \tag{3-3}$$

and the perpendicular component $\hat{e}_{X\perp}$ is the rest of $\hat{e}_X$:

$$\hat{e}_{X\perp} = \hat{e}_X - a\hat{e}_z \tag{3-4}$$

Having defined $\hat{e}_z$ in equation (3-1), it can be substituted into equation (3-4):

$$\hat{e}_{X\perp} = (1 - a^2)\hat{e}_X - ab\hat{e}_Y + ac\hat{e}_Z \tag{3-5}$$

Because $\hat{e}_x$ is a unit vector equal to $\hat{e}_{X\perp}/|\hat{e}_{X\perp}|$, equation (3-5) can be rewritten as

$$\hat{e}_x = \left(\sqrt{1-a^2}\right)\hat{e}_X - \left(\frac{ab}{\sqrt{1-a^2}}\right)\hat{e}_Y - \left(\frac{ac}{\sqrt{1-a^2}}\right)\hat{e}_Z \tag{3-6}$$

The cross product of the vectors $\hat{e}_z$ and $\hat{e}_x$ yields the vector $\hat{e}_y$, the result of which can be conveniently written as the symbolic determinant

$$\hat{e}_y = \zeta \begin{vmatrix} \hat{e}_X & \hat{e}_Y & \hat{e}_Z \\ a & b & c \\ 1-a^2 & -ab & -ac \end{vmatrix} \tag{3-7}$$

where ζ is $\frac{1}{\sqrt{1-a^2}}$. The matrix **R** that rotates the frame housing the original coordinates (*X,Y,Z*) into a new frame housing coordinates (*x,y,z*) is

$$\mathbf{R} = \begin{pmatrix} 1/\zeta & 0 & a \\ -ab\zeta & c\zeta & b \\ -ac\zeta & -b\zeta & c \end{pmatrix} \tag{3-8}$$

We can now rotate (any) Cartesian coordinates (*X,Y,Z*) using the rotation matrix in equation (3-8) by changing the direction cosine angles of the unit plane $\hat{n}$ = (*a,b,c*).

3.2.2. Spherical coordinates are used to rotate the helix frame

The unit plane described above passes through the origin. Thus, the two angle components of a spherical coordinate system are sufficient to rotate a set of coordinates over all 3D orientations using equation (3-8). The radial spherical component, ρ, can be obsoleted because the plane always passes through the origin. Therefore, the rotation of a helix in 3D Cartesian space using direction cosines with only two angles is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/\zeta & 0 & a \\ -ab\zeta & c\zeta & b \\ -ac\zeta & -b\zeta & c \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{3-9}$$

where $a = \sin\phi\cos\theta$, $b = \sin\phi\sin\theta$ and $c = \cos\phi$, ϕ and θ are the two polar spherical components. As usual for spherical coordinates, ϕ is bounded by [0°,180°) and θ is bounded by [0°,360°).

3.2.3. A linear least squares problem is solved for the circle projected by the helix

For each rotation, the x and y coordinates of the rotated points are fit to a circle in the form of the linear least squares problem (**AX** = **B**)

$$\begin{pmatrix} 1 & 2x & 2y \end{pmatrix} \begin{pmatrix} k \\ x_0 \\ y_0 \end{pmatrix} = (x^2 + y^2) \tag{3-10}$$

The **A** and **B** matrices in equation (3-10) are dimension $3 \times N$ and $1 \times N$, respectively, where N is the number of points being fit. We use singular value decomposition (SVD) to calculate the best

estimate of the parameters in the **X** matrix in equation (3-10). The k parameter in the **X** matrix contains the radius of the circle, r,

$$r = \sqrt{x_0^2 + y_0^2 + k} \tag{3-11}$$

which is simply the radius of the helix. The point where the helical axis intercepts the plane (a,b,c) is the circle center, (x_0,y_0). The residual of the fitting, χ, is

$$\chi^2 = \sum_{i=1}^{N} (x - x_0)^2 + (y - y_0)^2 + \rho^2 - 2\rho\sqrt{(x - x_0)^2 + (y - y_0)^2} \tag{3-12}$$

A complete scan of spherical coordinate rotation angles ϕ and θ is performed, and the residual calculated for each discrete rotation. The rotation with the smallest residual (best fit) corresponds to the angles ϕ and θ of the optimal helix frame.

Figure 3.1 illustrates the rotation-projection method. The coordinates of the points tracing a suspected helix are projected onto the x-y plane. These points are then fit to a circle and the residual noted inside a program. The method continues to scan the range of (ϕ,θ) defined by the user, with spherical coordinate rotation scan-resolution also defined by the user. Once rotating and fitting across the full user-defined scan range is complete, the method determines which rotation yielded the projection that best fit a circle. Due to symmetry in the projections, a complete scan can be accomplished using ϕ in [0°,180°) and θ in [0°,180°), rather than the usual bounds of ϕ in [0°,180°) and θ in [0°,360°).

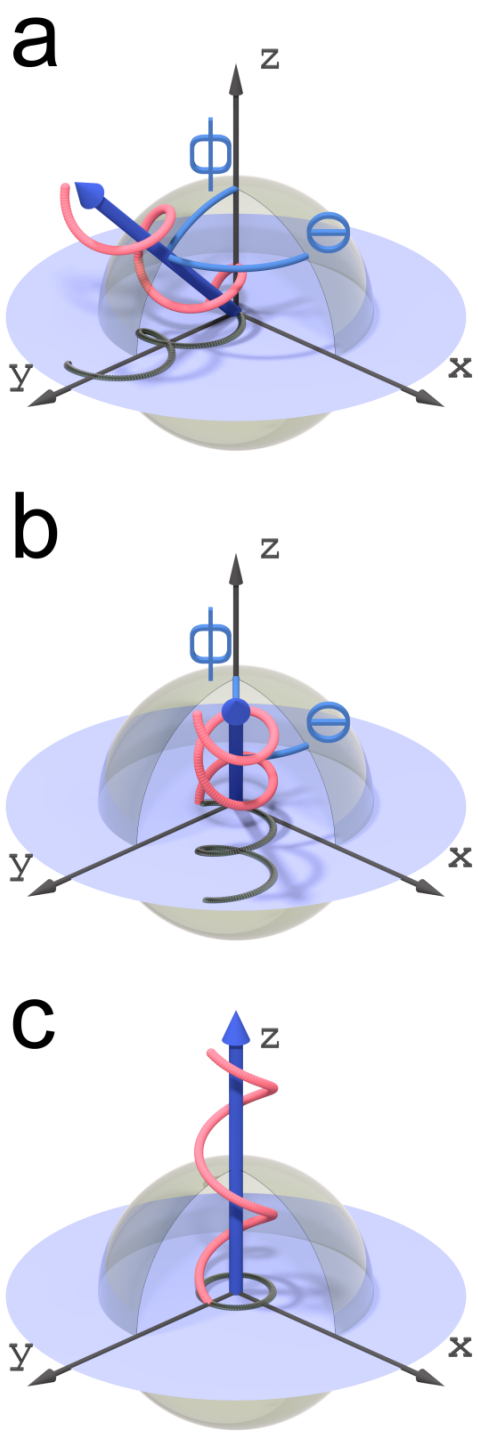


Figure 3.1. The frame of a helix is rotated using spherical coordinates to find the projection that best fits a circle. The points tracing a suspected helix (pink) in a frame (x,y,z) = (a,b,c); the points are projected (black) onto the x-y plane (blue disk) and are then fit to a circle using SVD. (**a**) A helix is projected on the x-y plane, with its frame defined by spherical coordinates (45°,90°). (**b**) The helix is rotated, its coordinates projected on x-y plane, and fit again; spherical coordinates (45°,45°). (**c**) After all rotations are complete, the rotation whose projection best fit a circle is the helix frame; spherical coordinates (0°,0°).

3.2.4. Helix pitch, twist and rise are calculated from the optimal helix frame

A helical curve is defined by three properties: radius, pitch and helix axis (**Figure 3.2**). A discrete set of points that trace a helix curve can be used to derive these three properties. A helical step describes the geometry between successive points tracing a helix. Twist is the radial angle subtended by a helical step. Rise is the axial displacement spanned by a step. The height of a discrete helix is the displacement between the last and first point along the helical axis. Helical sweep is the sum of the twists. If the helical sweep, Φ, is 360°, then the height of the helix is the pitch because pitch is the axial displacement of the helix per turn (360°). In a unit helix, the radius and pitch have equal magnitude. A helix whose pitch is greater than its radius is shown in **Figure 3.2a**, and a helix whose pitch is smaller than its radius is shown in **Figure 3.2b**.

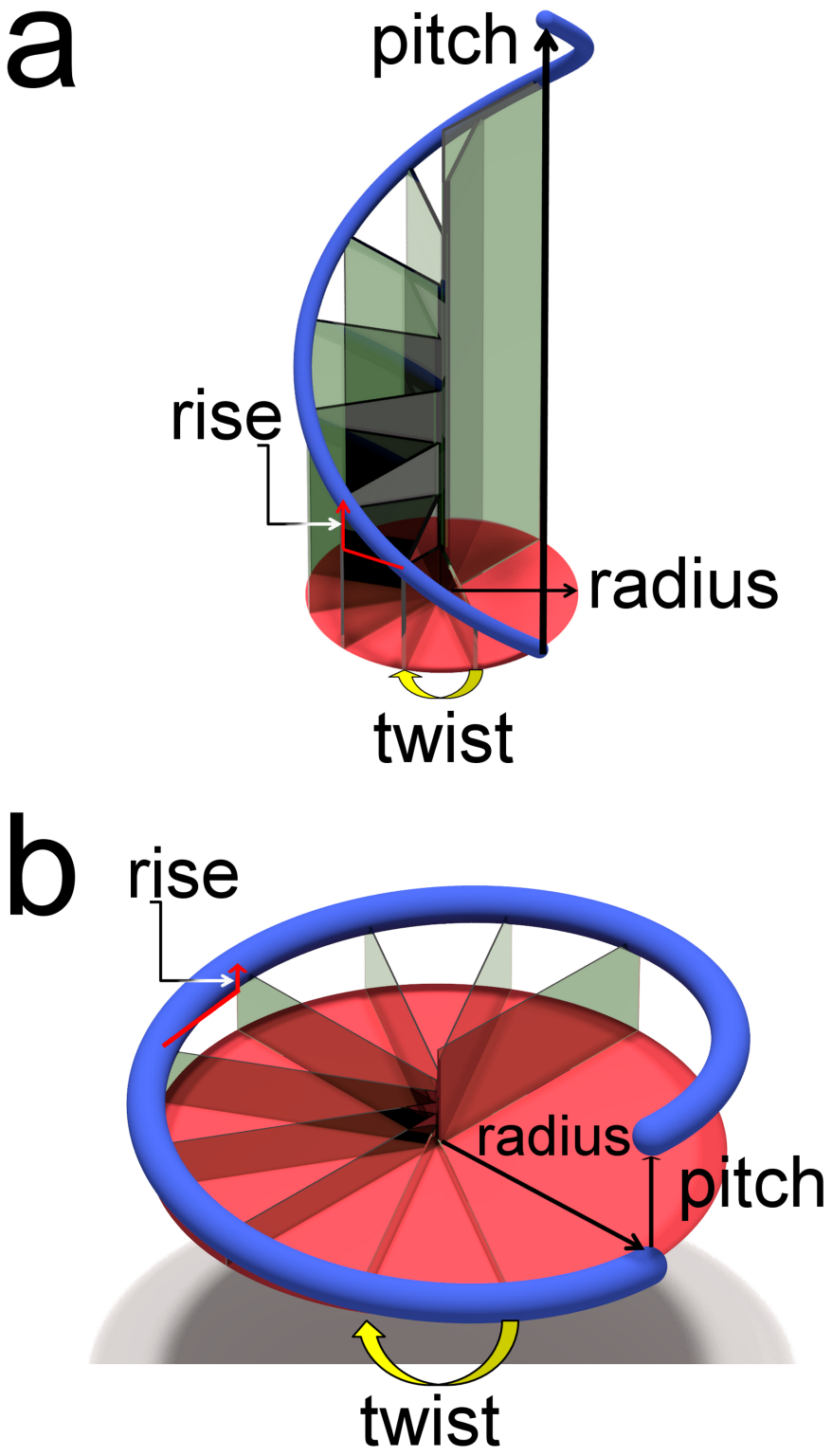


Figure 3.2. A helix and its parameters. A helix (blue) whose pitch is (**a**) greater than its radius or (**b**) smaller than its radius. The red disk represents the *x*-*y* plane. Vertical planes (grey) represent steps whose height is the *z*-component of the point and position is the *x*- and *y*-component of the point (i.e. $\text{radius}^2 = x^2 + y^2$). Twist is the angle between successive points. The increase in height between successive steps is rise. Pitch is the height of the helix for one (360°) revolution.

3.3. Methods

3.3.1. Nievergelt's helix

The Cartesian coordinates of Nievergelt's helix were obtained directly from the publication[39] (**Table 3.1**). A spherical coordinates scan over ϕ (from 0° to 180°) and θ (0° to 180°) was performed with 0.5° spherical coordinate scan resolution.

Table 3.1. Cartesian coordinates of Nievergelt's helix.

x	**y**	**z**
12	102	198
48	138	180
65	163	169
77	187	157
85	209	149
94	266	128
93	288	120
89	316	112
82	347	107
62	397	103

3.3.2. Generating artificial helices

The shape of a helix can be defined by the ratio of its radius and its pitch, *R/κ ratio*. Fractional values of *R/κ* represent thread-like (long and narrow) helices while values of *R/κ* larger than 1 represent ring-like helices (short and wide). The degree to which the helix resembles a polygon[122] (the tips of the translucent planes in **Figure 3.2**) or a smooth space curve (the blue curve in **Figure 3.2**) is determined by the number of points per helical turn, *PPT*. The number of turns in a helix with *N* total points is thus *N/PPT*.

For each helix shape R/κ, a matrix (represented as a shaded square in **Figure 3.3**) was generated by varying the number of points per helix, N, from 4 to 16, and varying points per turn, PPT, from 3 to 8. This generated 78 helices of the same shape, with varying length and granularity.

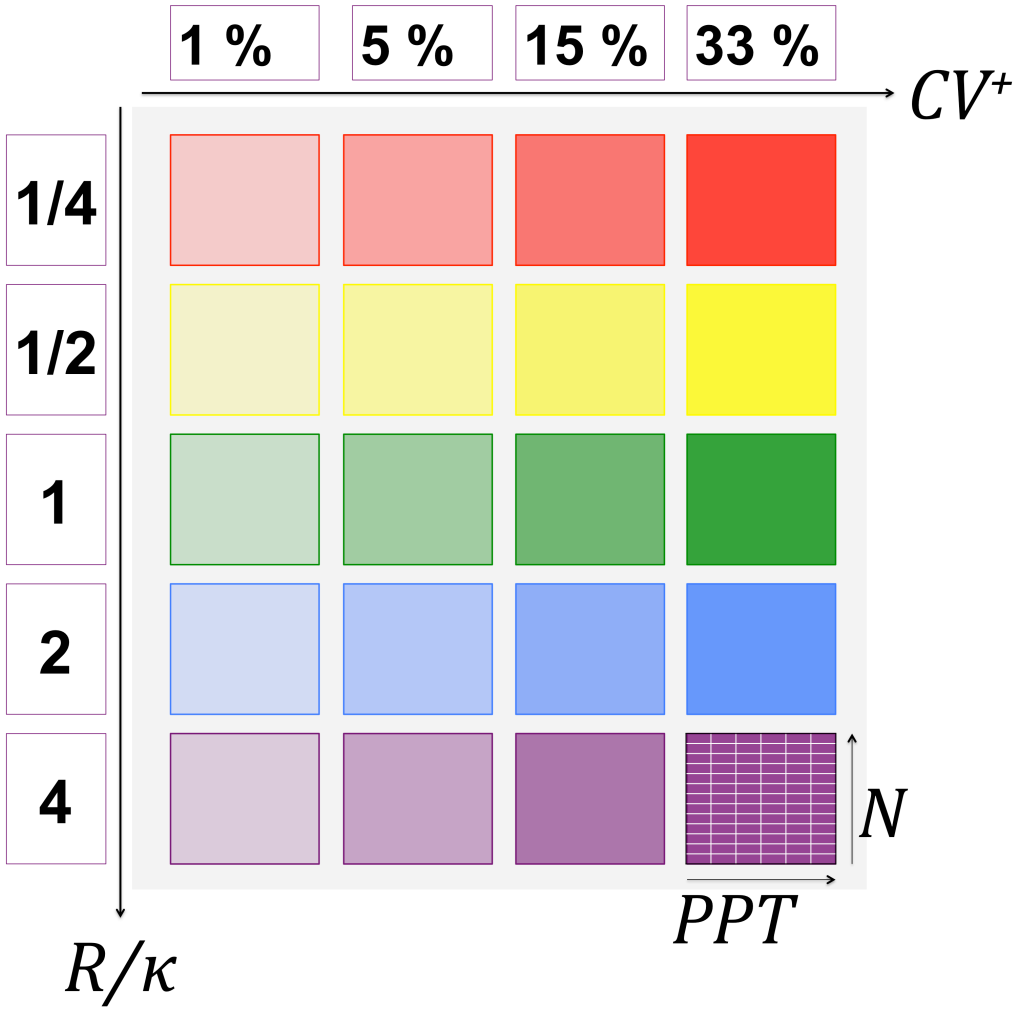


Figure 3.3. Overview of the test set of noisy helices, with varying shapes (R/κ), degrees of applied noise (CV^+), number of points per helix (N) and points per helical turn (PPT). Five R/κ shape ratios (rows of one color) and four CV^+-noise levels (columns with different colors) provided 20 shape-noise helix-matrices (each colored square depicts one helix-matrix). Each helix-matrix contains varying numbers of points per helical turn and total number of points in the helix. Each cell in this matrix (6 columns of PPT, 13 rows of N, 78 total cells) represents 256 different random perturbations of the helix. 399,360 total helices were evaluated below: 256 helices per cell, 78 cells per shape-noise matrix, 20 shape-noise matrices.

Helices were generated using the parametric equations:

$$x = R\sin\omega,\ y = \kappa\omega,\ z = R\cos\omega \tag{3-13}$$

where ω is the twist, which is 360°/*PPT*. The helical axis of these artificial helices coincide with the *y*-axis of the coordinate system. The spherical coordinates (ϕ,θ) of the helix frame (*a,b,c*) are (90°,0°). Helices were randomly oriented by rotating them using equation (3-9), with random values for ϕ bounded by [0, °180°) and θ bounded by [0°,360°).

Noisy helices generated by perturbing (x,y,z)

Noisy helices were generated by perturbing the Cartesian coordinates of each point in the artificial helices described above. Each component of each point was perturbed separately to mimic random (thermal) deviations. Noise levels were based on a coefficient of variation (CV^+),

$$CV^+ = \frac{\sigma}{\mu} \tag{3-14}$$

where σ denotes standard deviation and μ denotes the mean. Here, μ is simply the original Cartesian component of a point from the ideal helices described in the previous section. To generate perturbed coordinates subject to a level of noise prescribed by equation (3-14), values for the perturbed coordinate components were drawn from a normal distribution whose mean was μ and whose standard deviation was σ.

The superscript "+" in "CV^+" is used to distinguish the variation applied to perturb the Cartesian coordinates (CV^+) from the resulting variation in helical parameters (CV_{twist} and CV_{rise}) calculated for the noisy helices. The following CV^+ values were used: 0.01, 0.05, 0.15 and 0.33. This random perturbation was carried out 256 times for each cell in the *N vs. PPT* matrix for each shape *R/κ* and for each value of CV^+. Overall, 399,360 noisy helices were generated. **Figure 3.3** summarizes the set of noisy helices used in this study.

The accuracy of our results were quanitifed using percent average absolute error (%AAE). The absolute value of the difference between a calculated value and a reference value was divided by the reference value. If the value of a calculated parameter was identical to the reference value, the difference would be zero, and zero divided by the reference value would still be zero. Zero would then be multiplied by 100% to yield 0% AAE. The equation for %AAE is

$$\%\text{AAE} = 100\% \times \left|\frac{b_{cal} - b_{ref}}{b_{ref}}\right| \tag{3-15}$$

where b_{cal} represents the value of the calculated parameter and b_{ref} represents the value of the reference value.

3.3.3. Generating biomolecular helices

3.3.3.1. Generating helical peptide secondary structure elements

Three polypeptide α-helix elements were generated using the LEaP module in Amber[123]. Three slightly different α-helical geometries were used to test the sensitivity of the method. Poly-

alanine 32-mer peptides were generated by imposing backbone torsion angles of (-60°,-45°)[124] for ϕ and ψ (henceforth referred to as, $\alpha_{-60,-45}$) backbone torsion angles of (-57°,-47°)[125] for ϕ and ψ (henceforth referred to as, $\alpha_{-57,-47}$), and backbone torsion angles of (-60°,-40°) for ϕ and ψ (henceforth referred to as, $\alpha_{-60,-40}$).

One 3_{10}-helix peptide was generated using LEaP as above, except the backbone torsion angles for ϕ and ψ were (-49°,-27°)[125]. In addition, one π-helix was generated using LEaP as above, except the backbone torsion angles for ϕ and ψ were (-57°,-70°)[125]. Only the Cα atoms from the peptides were used in our fitting procedure.

The expected rise and twist between consecutive Cα atoms of each amino acid in the helical SSEs was calculated from the literature values[126] for pitch and residues per turn. Rise is pitch/residues per turn and twist is 360°/residues per turn.

3.3.3.2. Generating nucleic acid helices

Three nucleic acid duplexes were generated using 3DNA[118]. A-DNA was generated by imposing base pair (bp)-step twist and rise values of 32.7° and 2.548 Å, respectively. B-DNA was generated by imposing bp-step twist and rise values of 36.0° and 3.375 Å, respectively. A-RNA was generated by imposing bp-step twist and rise values of 32.7° and 2.812 Å, respectively. Curves+[119] was used to characterize the bp-step twist and rise using standard procedures[127]. The C1′ atoms of these nucleic acids were used in the fitting. The method has the capability to utilize both helices in dsDNA and dsRNA to find the optimal common helical axis. Both strands were used in our analyses of double-stranded nucleic acids.

3.3.3.3. Analyzing a superhelical protein-DNA complex

The atomic coordinates of a modular DNA-binding protein with a superhelical tertiary structure, BurrH[128], was obtained from the Protein Data Bank[129] (PDB ID: 4CJA[128]). Superhelical Cα atoms were selected using our previously described approach[85]. Briefly, amino acids tracking the DNA binding cleft were identified (**Table 3.2**), one from each module of the protein, and the points of the helix were represented by the Cα atoms of these amino acids. The C1′ atoms of both strands of the bound DNA were used to find the optimal DNA helical axis.

Table 3.2. Residues whose Cα atoms were used to define the BurrH superhelix.

Module	**Residue**
1	Q42
2	P75
3	S108
4	S141
5	P174
6	P207
7	P240
8	P273
9	P306
10	P339
11	P372
12	P405
13	P438
14	L471
15	L504
16	L537
17	R570
18	A603
19	A636
20	P669
21	P702
22	P734
23	P765

3.3.3.4. Using the test helices to estimate accuracy of an analysis

The number of points required to accurately characterize a helix using our method can be estimated by consulting the results of the analysis of the test helices. In the validation performed here, the R/κ ratio and *PPT* of helical secondary structure elements (proteins) and the helices of nucleic acids (single and double-stranded DNA and RNA) were known from the literature. The unknown parameter was the minimum number of points needed to accurately characterize these helices. If in future use of this method the R/κ and *PPT* is not known but the number of points is known, then the test helix results can be consulted to project the expected range of accuracy.

3.4. Results

3.4.1. The method requires fewer points than TLS to achieve the same accuracy

First, the method was compared directly to Nievergelt's related TLS[39] approach. The ten Cartesian coordinates published by Nievergelt trace an irregular helix whose pitch and radius were 600 and 195, respectively. We sought to determine the minimum number of points required by our method to reproduce the parameters obtained by Nievergelt's TLS approach. **Figure 3.4** shows the results of the helix-fitting using our approach. With all ten points, our method achieves 0.5% and 0.4% relative error in pitch and radius respectively. Less than 1% relative error in pitch and radius was achieved with seven points and < 5% relative error in pitch and radius with only six points. For comparison, previous work showed that HELFIT requires all ten points to achieve 9% relative error in radius, and 0% relative error in pitch[116].

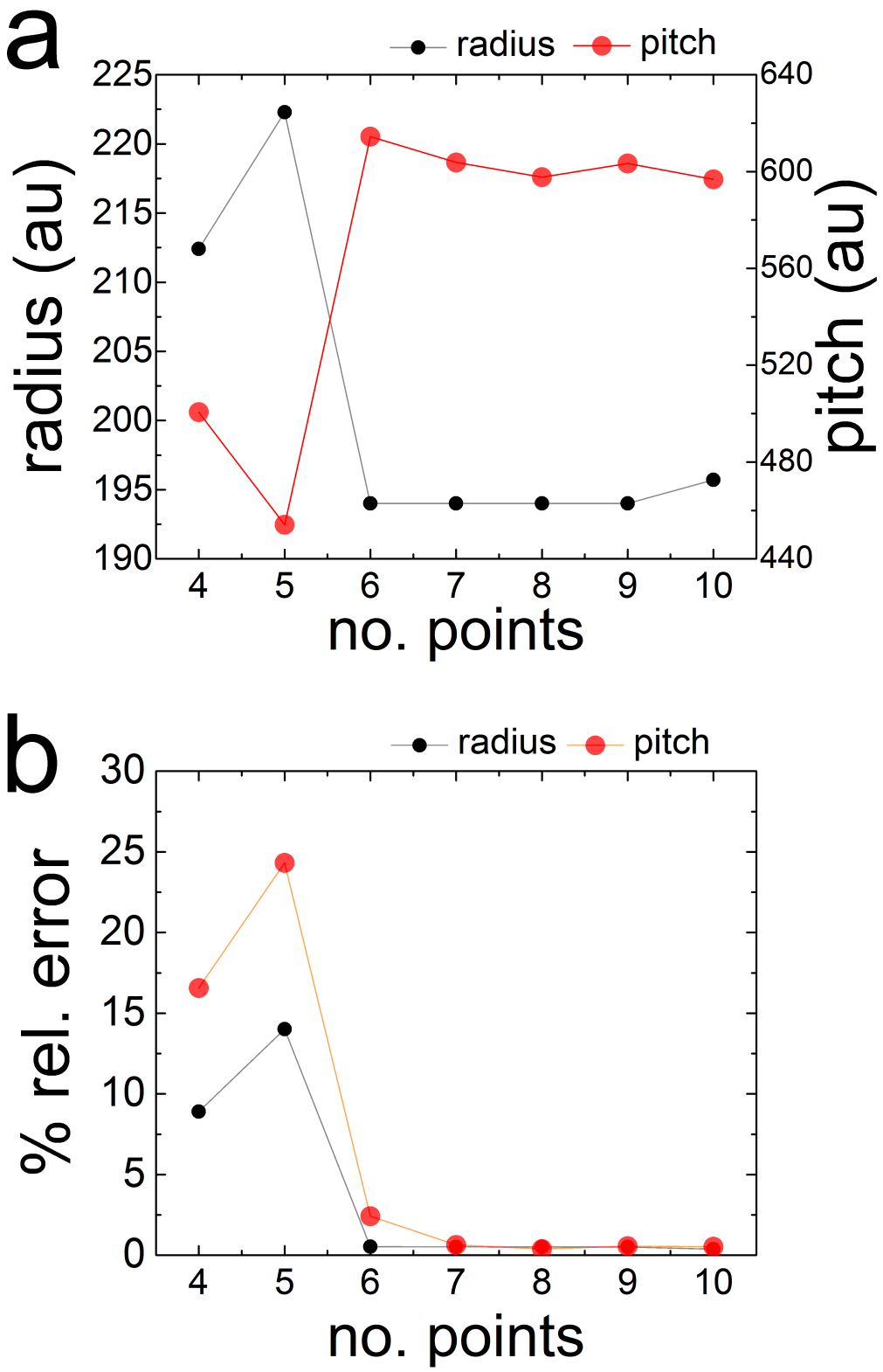


Figure 3.4. Helix parameters of Nievergelt's helix. (**a**) Helix pitch and radius, with increasing number of points. (**b**) Percent relative error, using Nievergelt's data as the reference, with increasing number of points.

3.4.2. Validation tests of ideal artificial helices

The search for a helical axis over a scan of spherical coordinate rotations (step-size in ϕ and θ) is the only component of the fitting that is under user control. Scan step-size determines how finely rotations of the helix frame are made, ostensibly increasing the accuracy by increasing the likelihood that the helix can be ideally projected. The user-defined fitting parameters ϕ-range and θ-range were not evaluated during this test because changing them would amount to applying fitting constraints. This would improve the accuracy of the method in an obvious manner; therefore there is no need to test situations in which the user already knows some bounds of the solution.

How does the scan resolution affect accuracy?

As the scan is made coarser, it becomes less likely that the helix can be properly aligned during the projection operation, resulting in potential inaccuracies in calculated helical parameters. As a control, an ideal helix with PPT = 6 points per turn, R/κ = 1 radius/pitch ratio and 12 points, N = 12, was characterized using different spherical coordinate step size (scan resolution) for the spherical coordinate scan. This helix was placed in 64 different random orientations prior to the spherical coordinates scan, and helical parameters were calculated for each orientation. Ideally, the scan is fine enough that the same parameters are calculated for all 64 orientations. The measured percent average absolute error (%AAE) indicates the inaccuracy expected solely from helical axis alignment errors due to the scan spacing (**Figure 3.5a**). As expected, the coarse scan resolutions (6° and 3°) did not recapitulate the input helix twist (60°) and rise (360°/6) as accurately as the finer scan resolutions because the structures could not be perfectly oriented. Scan resolutions of 1° grid steps or smaller resulted in good accuracy (AAE < 0.1%).

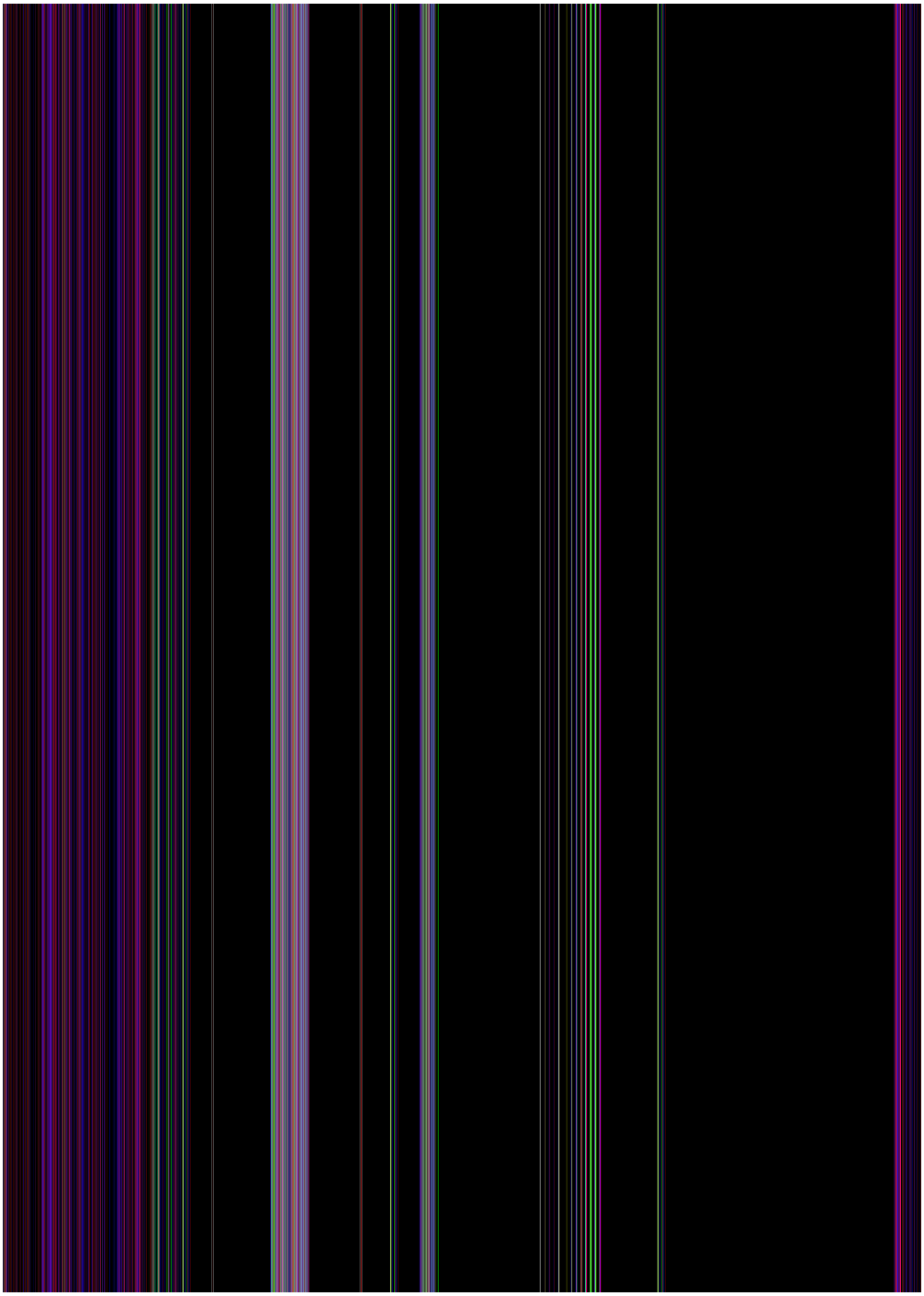


Figure 3.5. The effect of scan resolution on accuracy measured by percent AAE (%AAE) of helical parameters for a set of 64 randomly oriented copies of a helix. Smaller %AAE values indicate less sensitivity to random orientation. (**a**) An ideal helix characterized using six scan resolutions (grid steps) of 6°, 3°, 2°, 1°, 0.5° and 0.25° (X axis). The reference rise and twist were 1 and 60, respectively. The Y axes show the measured %AAE for rise and twist. Black and blue symbols represent the measured %AAE of rise and twist respectively. (**b**) The coordinates of the ideal helix shown in panel (**a**) with random Cartesian coordinate perturbations applied to mimic the structure of an irregular, noisy helix.

Next we tested whether the sensitivity for scan resolution was different for a typical non-ideal helix, as expected in biomolecular structures. A representative noisy helix was generated by perturbing the coordinates of the above regular helix by applying noise to the *x-, y-, z-* coordinates of each point in the helix. The reference pitch and radius for calculation of %AAE were obtained from the calculation using a fine scan with 0.25° scan resolution, which provided the best estimate of the otherwise unknown parameters. The helix was again placed in 64 random orientations. The results were comparable to those for the ideal helix, and increasing scan resolution displayed a sharp increase in accuracy (below 0.1% AAE) at 3° scan resolution, remaining below 0.1% AAE from scan resolutions of 1° scan resolution and below (**Figure 3.5b**). Overall, these results indicate that the accuracy of the method does not depend on scan resolutions below 1°, and performs comparably with noisy and ideal helices.

How sensitive are calculated helix parameters to random perturbations of Cartesian coordinates?

The goal of the following analysis was to determine the sensitivity of our method to uncertainty in the positions of the points tracing a helix. A broad validation was performed, in which all 78 helix geometries depicted schematically in **Figure 3.3** were tested. The approach validates the method across a diverse range of helical parameters and coordinate perturbations. These perturbations were applied to the Cartesian coordinates of the helix points rather than the helical parameters themselves (twist, rise, radius), which were affected indirectly by the perturbations. We chose to perturb the helical parameters indirectly because our goal was to determine the dependence of the derived helical parameters on noise in the coordinates to represent the mix of distortions seen in structures of biomolecular helices, or uncertainties in

selecting atoms to represent the helical superstructure. Noisy helices were generated by applying random perturbations to the *x*-, the *y*- and the *z*-coordinates of each point in the ideal helix subject to one of four CV^+'s (see **3.3 Methods**). Each helix-matrix (R/κ, CV^+) contained 78 distinct helix geometries, and each geometry contained 256 independent (perturbed) samples. A spherical coordinate scan resolution of 1° was used in the analysis. In the analysis below, "CV" is used to denote the coefficient of variation that was measured after using the method, whereas "CV^+" denotes the coefficient of variation that we applied to the data before using the method (see **3.3 Methods** for details).

We expected the test helices with the fewest points per helical turn (*PPT*) and the most points to be the least susceptible to noise because the helices would have more turns, and more data to fit. Multiple turns help distinguish the points as a helix (a circle on the projection plane); conversely, if the points trace less than one turn and are noisy (e.g. $N = 5$, $PPT = 8$) they might appear to trace a 2D arc with ambiguous helix axis. In addition, we expected that increasing levels of applied coordinate perturbations (CV^+) would introduce increasing uncertainties (CV measured) to the derived helical parameters.

Figure 3.6 reveals the dependence of derived helical twist (CV_{twist}) on the noisy helices with diverse geometries. Helices with $R/\kappa > 1$ are wide-short helices, $R/\kappa = 1$ are unit helices, and $R/\kappa < 1$ are narrow-tall helices (**Figure 3.2**). The points of the helices were perturbed by four increasing levels of noise applied to their Cartesian coordinates ($CV^+ = 0.01$, $CV^+ = 0.05$, $CV^+ = 0.15$, $CV^+ = 0.33$, see equation (3-14)). As expected, the results show that the helices with the fewest total points (small *N*) and the fewest turns (larger *PPT* at each *N*) were the most susceptible to noise (large CV_{twist}), and as expected the effect grew with increasing applied CV^+. The analysis also revealed the dependence of derived helical parameter sensitivity with R/κ. For

short helices with wide radii ($R/\kappa > 1$), the derived helical twist was less susceptible to noise than narrow-tall helix geometries. Conversely, the helical twist derived by the method was more sensitive to noise for extended helices with narrow radii. The method was sensitive to coordinate perturbations when $R/\kappa < 1$ because even for small applied CV^{+} of 5%, helices with fewer than nine points were incorrectly rotated from the expected helical axis (**Figure 3.6**). It is important to note that the test helices in **Figure 3.7** are very short – 4 to 16 points in total – representing the most challenging geometries expected in biomolecules. The %AAE for these the test helices also indicates that the shapes frequently observed in biomolecules ($R/\kappa < 1$) were also those that were most accurately calculated by our method (**Figure 3.8**).

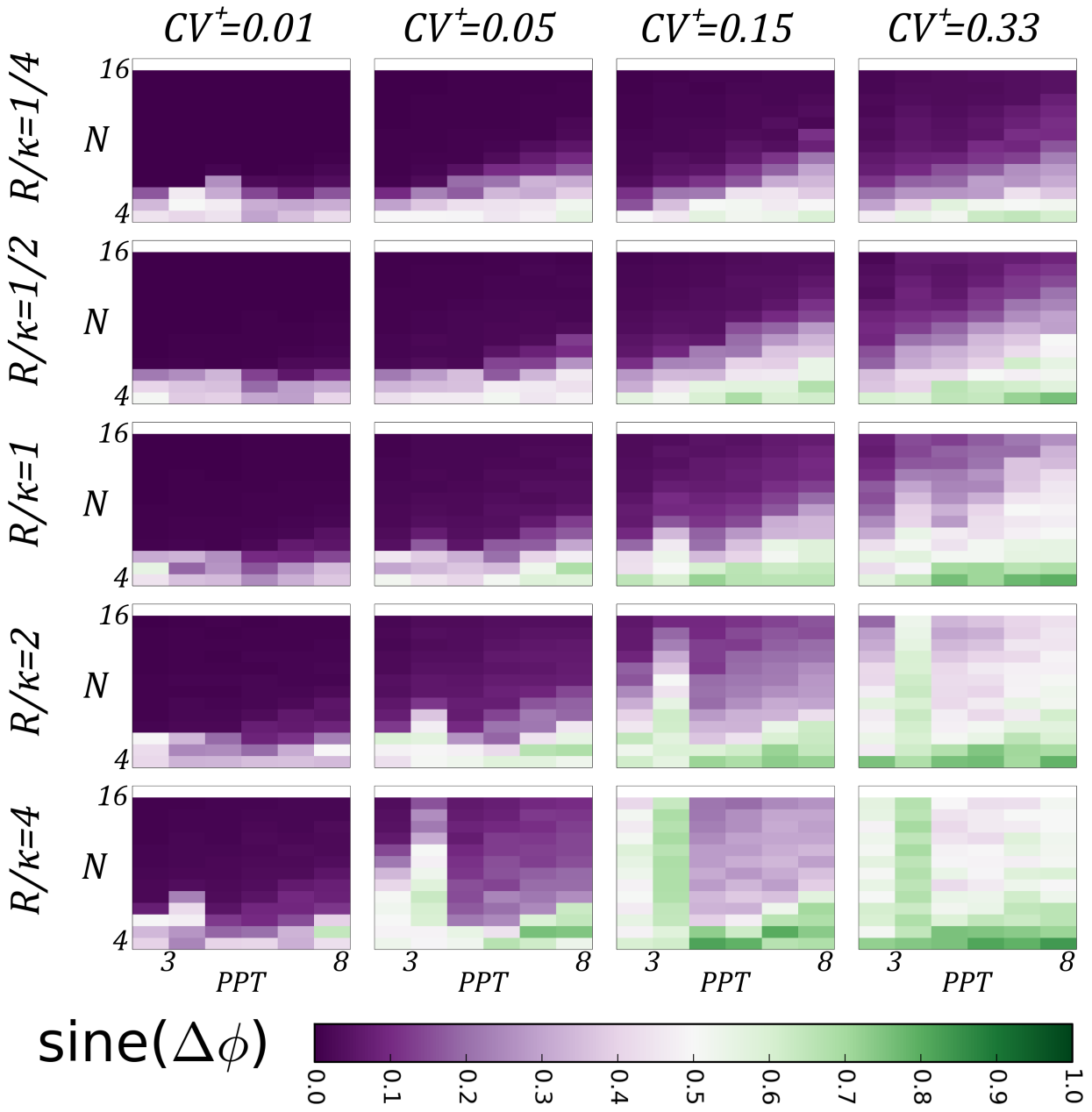


Figure 3.6. Sine of the average difference in the helical axes of the test helices and the ideal helices. Helix matrices are organized per **Figure 3.3**. We plot the sine of the difference in the average spherical coordinate φ-angle of the optimal helical axes for the 256 noisy helices in each cell and the spherical coordinate φ-angle of the optimal helical axis from the corresponding ideal helix If the difference is 0°, then sineΔφ is 0; if the difference is 45°, then sineΔφ is 0.5; if the difference is 90°, then sineΔφ is 1.

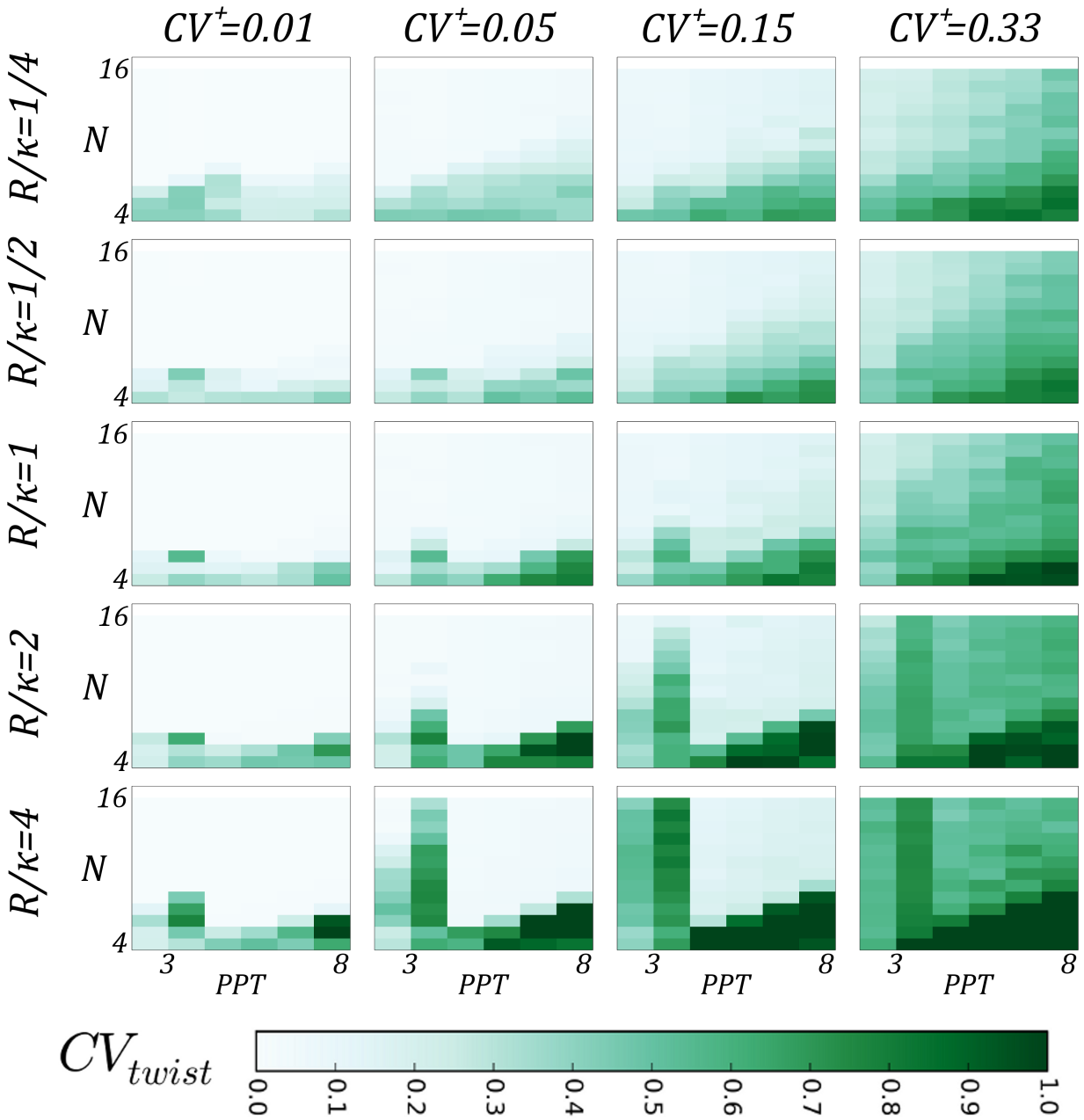


Figure 3.7. Twenty heat-maps of derived CV_{twist} for noisy helices with diverse geometries. Families of helices with five radius-pitch ratios, R/κ, are shown, each of which contains a matrix of helices with varying points/turn, *PPT*, and varying numbers of points, *N*. For a given *N*, increasing *PPT* decreases the number of turns present. For each shape geometry (colored cells, e.g. R/κ = 1/4, CV^{+} = 0.01 *N* = 16, *PPT* = 3, the top left-most cell), 256 irregular helices were generated subject to Cartesian coordinate perturbations with CV^{+}, 0.01, 0.05, 0.15 and 0.33. If a cell in a helix-matrix is light-colored (white), the measured CV_{twist} is low and the method precisely characterizes helix twist. However, if a cell is dark-colored (green), the measured CV_{twist} is large and the method does not precisely characterize helix twist; the method is susceptible to noise.

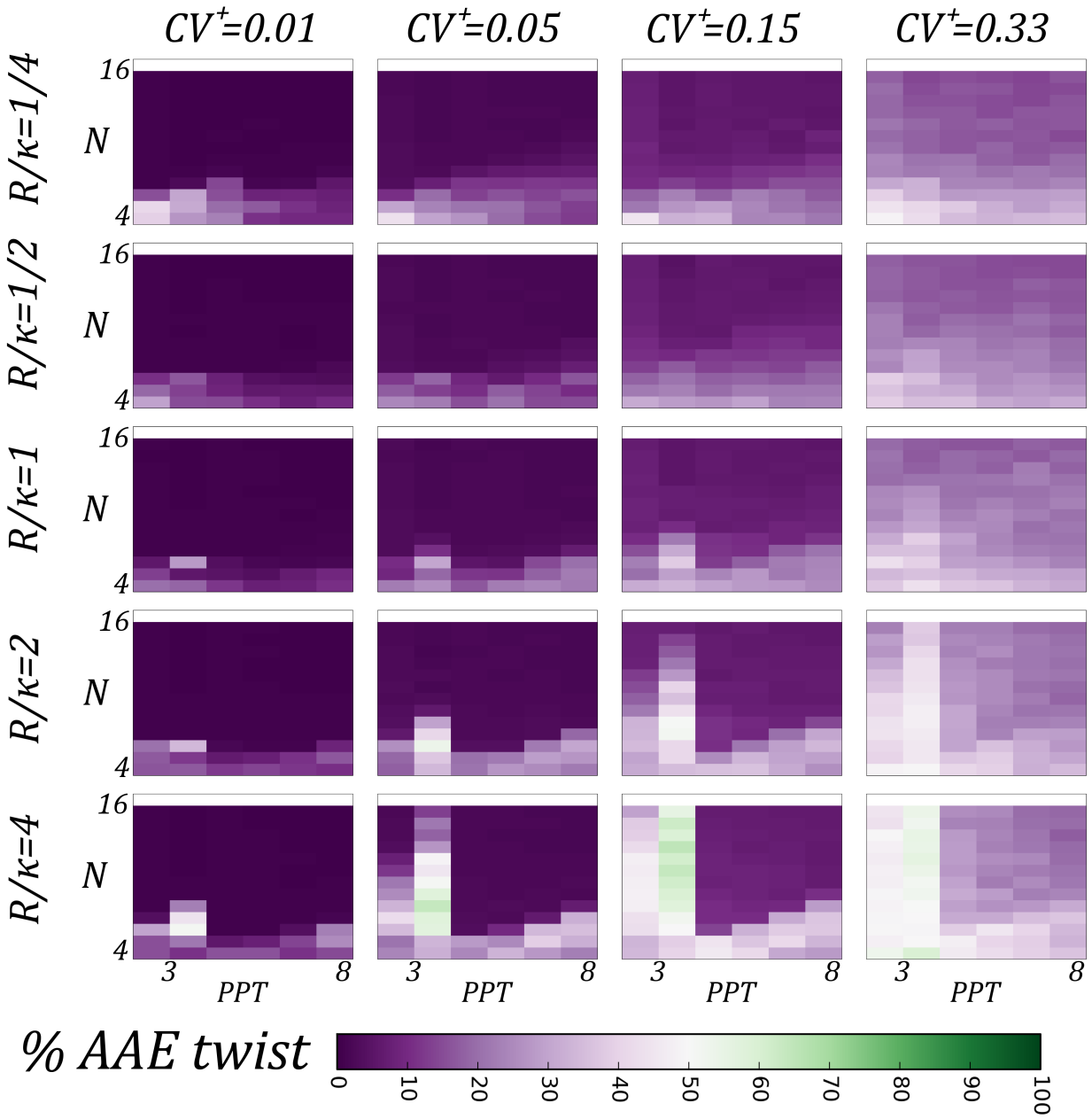


Figure 3.8. Percent AAE of derived helix twist for the 399,360 test helices. Helix matrices are organized per **Figure 3.3**. The plotted range of the percent average absolute error is 0 to 100.

Sensitivity of helical rise for regular helices with positional noise

Next we characterized how sensitive the derivation of the rise parameter was to CV^{+} noise in the (x,y,z) coordinates of the points tracing the test helices. Since rise and twist are

orthogonal parameters (the former depends only on $\hat{e}_z$ while twist depends on $\hat{e}_x$ and $\hat{e}_y$), we expected that rise might depend on helix shape (R/κ) in an opposite way as twist. Helices tracing less than one turn and helices with $N = 4$ points were still expected to be the most susceptible to noise because the former becomes degenerate with a noisy 2D arc, and the latter offers too few data distinguish noise (CV^+) from signal (the underlying helix). The sensitivity of helix rise for diverse geometries with increasing amounts of applied coordinate perturbations (CV^+) is shown in **Figure 3.9**. With the smallest applied CV ($CV^+ = 0.01$), the resulting helical rise matched our expectation of an R/κ-dependence opposite to that of twist. For the CV^+ values larger than 0.01, the results indicate that helix rise is more sensitive to noise than twist. CV_{rise} was at least 0.2 for all geometries, when CV^+ values of 0.05 and greater were applied. Like the results of the analysis above for twist, the percent AAE of derived rise for these the test helices indicates that the helices with $R/\kappa < 1$ were also those that were most accurately calculated by our method (**Figure 3.10**).

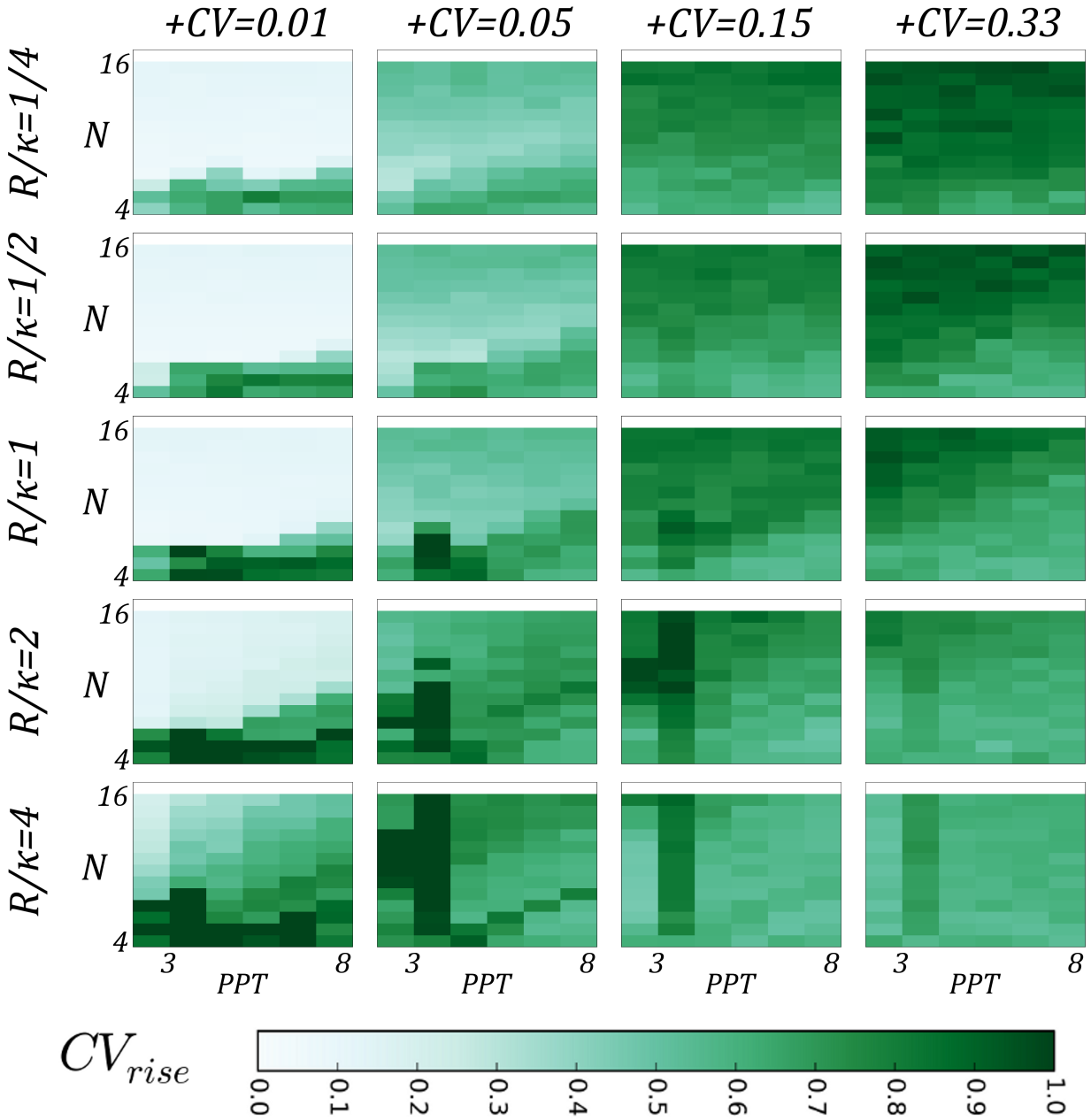


Figure 3.9. Twenty heat-maps of derived CV_{rise} for noisy helices with diverse geometries. The helices analyzed and the layout of the data are the same as in **Figure 3.7**.

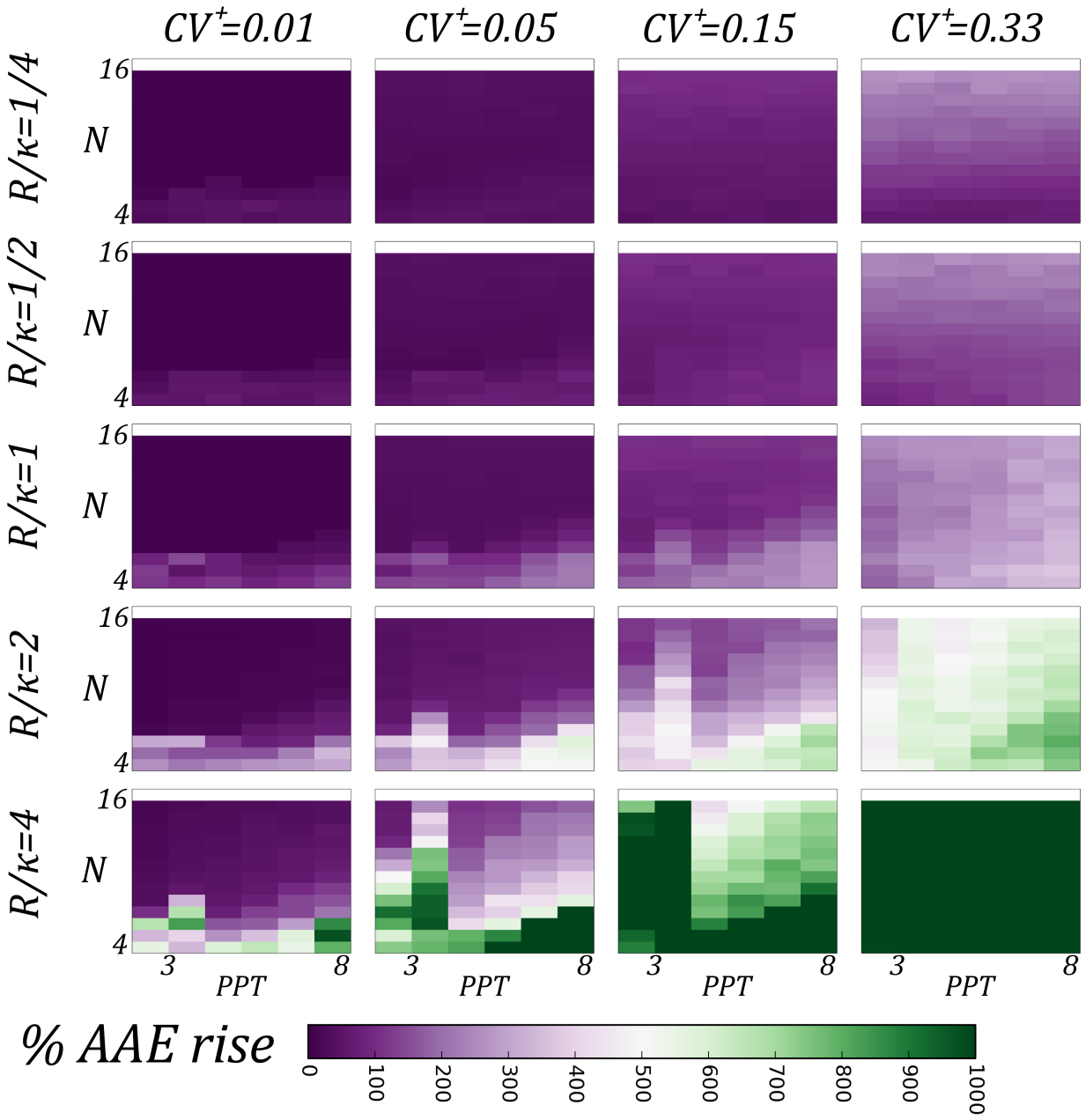


Figure 3.10. Percent AAE of derived helix rise for the 399,360 test helices. Helix matrices are organized per **Figure 3.3**. The plotted range of the percent average absolute error is 0 to 1000.

Summary of the analysis of test helices

Overall, twist was more precisely recapitulated than rise when characterizing noisy helices. Twist is a parameter of the helix that is fully contained within cylindrical slices (i.e. a

projection plane), thus twist is optimized directly by the method (*x*-, *y*-components) while rise is optimized indirectly. It is possible that twist is calculated more precisely than rise because twist is regularized during the fitting procedure. It is also possible that twist is more precisely characterized because twist contains two dimensions of information ($\hat{e}_x$ and $\hat{e}_y$) while rise contains only one ($\hat{e}_z$). The diverse test helices represent the limiting case because they possess the minimal geometric properties (number points and turns) required to unambiguously define a helix. The tests performed in this section represent the "worst case scenarios" potential users of the method might experience. The low *R/κ* ratio helices (1/2 and 1/4) led to the lowest %AAE for rise (**Figure 3.10**), twist (**Figure 3.8**) and most accurately derived helical axis orientations (**Figure 3.6**). These helix shapes are representative of protein secondary structure helices, nucleic acid helices and protein tertiary structure helices.

3.4.3. Testing helical secondary structure elements

α-helix secondary structures

The radius, pitch and *PPT* of an α-helix are 2.3 Å, 5.5 Å and 3.6,[126] respectively, and its *R/κ* is 0.42. Here, the Cα atoms from the amino acids were used as the points to carry out the fitting. Referencing **Figures 3.7** and **3.9** for results on ideal helices with comparable *PPT* and *R/κ* (consulting test helices with *PPT*=4 and *R/κ*=1/2), we expected ~7 points would be required to accurately define the helical parameters of an α-helix. **Figure 3.11** shows the results of our analysis of the helical parameters for α-helical secondary structure elements. Three slightly different ϕ/ψ backbone torsion angles within the α-helix region of the Ramachandran map were tested. The residual of the fitting for these three helical shapes rises linearly with the number of

Cα atoms included in the fit because each atom adds to the total residual (**Figure 3.12**). As expected, seven Cα atoms were required to accurately calculate the helical rise (1.5 Å, **Figure 3.11a**), twist (100°, **Figure 3.11b**) and radius (2.3 Å, **Figure 3.11c**) for these ideal structures. Two turns are likely required to characterize helical secondary structure elements because the points tracing these helices are highly polygonal (< 4 *PPT*).

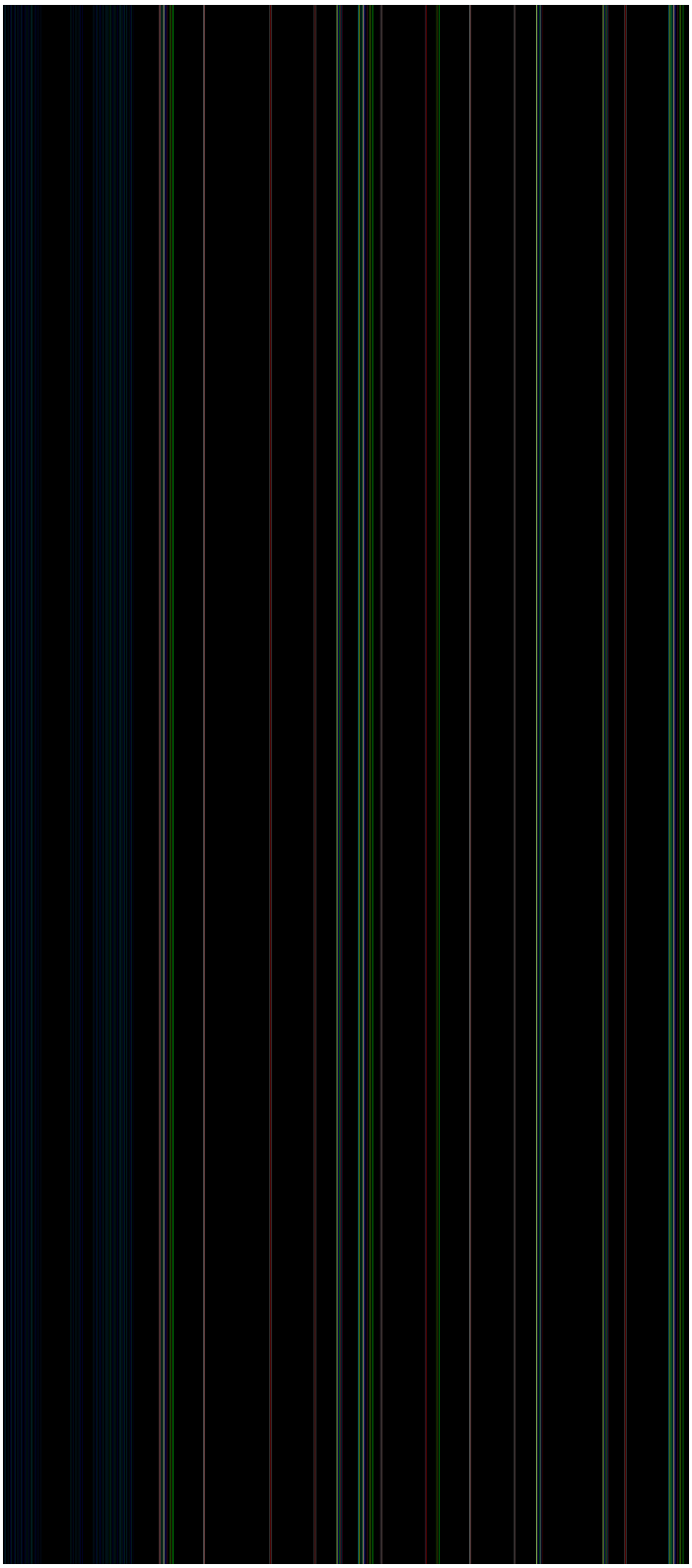


Figure 3.11. Helical parameters of α-helical secondary structure elements defined using three different pairs of φ/ψ backbone torsions (black, blue and red symbols represent $\alpha_{-57,-47}$, $\alpha_{-60,-40}$ and $\alpha_{-60,-45}$ respectively), and an increasing number of Cα atoms used in the fitting (X axis). Shown on the Y axes are the derived values for (**a**) rise; (**b**) twist; (**c**) radius.

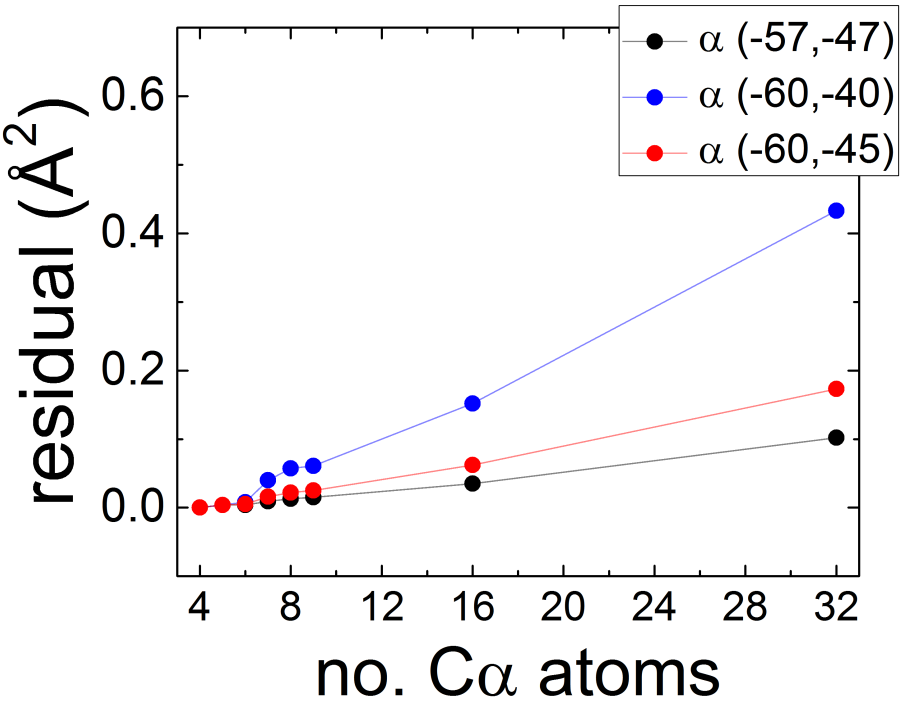


Figure 3.12. Fitting residual of three slightly different α-helical peptide secondary structure elements. Fitting residual was calculated using equation (3-12) and is plotted on the Y axis. The number of Cα atoms (one per amino acid) used in the fitting is plotted on the X axis. Residual rises with the number of Cα atoms used in the fitting because each atom contributes to the total deviation; residual rises linearly with number of atoms.

π-helix and 3_{10} secondary structures

The radius, pitch and *PPT* of a π-helix are 2.7 Å, 4.1 Å and 4.2,[126] respectively, and its R/κ is 0.66. From the results of **Figures 3.7** and **3.9** (consulting test helices with *PPT*=4 and R/κ=1/2), we expected ~8 points would be required to accurately calculate the helical parameters of a π-helix. The radius, pitch and *PPT* of a 3_{10}-helix is 1.9 Å, 5.8 Å[126] and 3.0 respectively and its R/κ was 0.33; (consulting test helices with *PPT*=6 and R/κ=1/4), we expected ~7 points would be required to calculate accurately the helical parameters of a 3_{10}-helix. **Figure 3.13** shows the results of the analysis of the helical parameters for π- and 3_{10}-helical secondary structure elements. Seven and nine Cα atoms were required to accurately calculate the helical rise (**Figure**

3.13a), twist (**Figure 3.13b**) and radius (**Figure 3.13c**) for π- and 3_{10}-helical elements, respectively. As with the α-helices, the residual of the fitting for these three helical shapes rises linearly with the number of Cα atoms included in the fit because each atom adds to the total residual (**Figure 3.14**). Overall, our method requires no more than nine atoms to achieve high-accuracy helix parameters for ideal protein secondary structure elements.

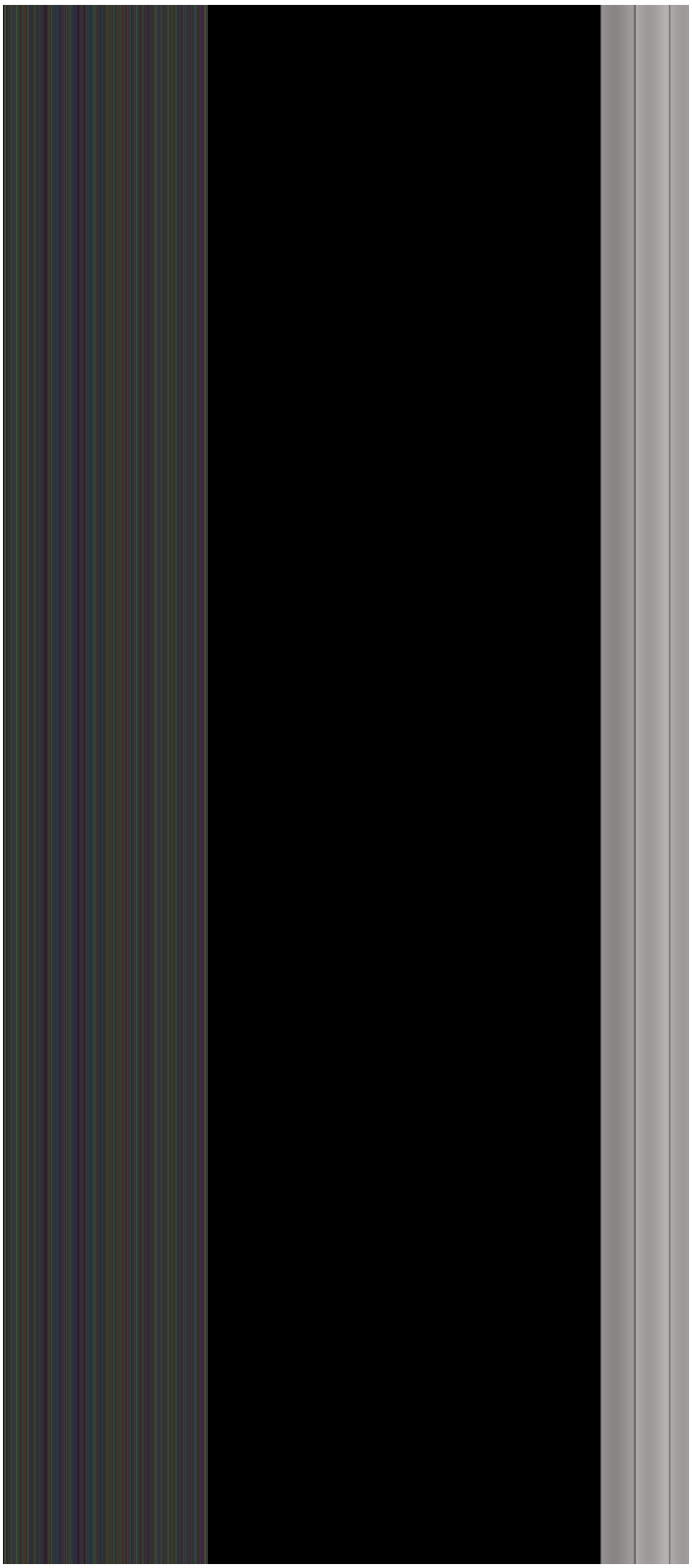


Figure 3.13. Helical parameters of π- and 3_{10}- helical secondary structure elements, black and blue symbols respectively, with an increasing number of Cα atoms used in the fitting (X axis). Shown on the Y axes are the derived values for (**a**) rise; (**b**) twist; (**c**) radius.

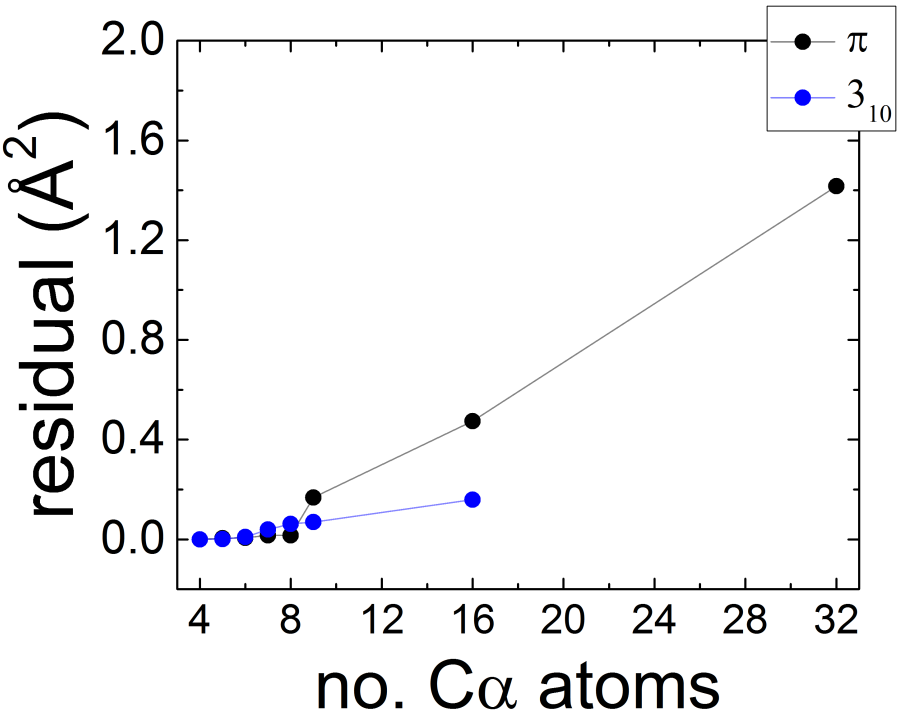


Figure 3.14. Fitting residual of π- and 3_{10}-helical peptide secondary structure elements. Fitting residual was calculated using equation (3-12) and is plotted on the Y axis. The number of Cα atoms (one per amino acid) used in the fitting is plotted on the X axis. Residual rises with the number of Cα atoms used in the fitting because each atom contributes to the total deviation; residual rises linearly with number of atoms.

3.4.4. Validation of nucleic acid helices: single- and double-stranded DNA and RNA

The method can calculate the helical parameters of single-stranded (ss) and double-stranded (ds) nucleic acids. However, 3DNA[118] and Curves+[119] can only define a helix if base pairs are present; the nucleic acid must be double stranded, precluding these methods from comparison with ours in the assessment single-stranded nucleic acid helices. It is also important to note that 3DNA and Curves+ require multiple atoms per nucleotide to define a helix frame, whereas our method requires at minimum only one. Here, we used C1′ atoms. **Table 3.3** compares dsA-DNA, dsB-DNA and dsA-RNA rise and twist parameters obtained using our

approach, Curves+[119] and 3DNA[118], which was used to generate the DNA structures. The overall agreement is excellent between the three methods, < 1% AAE for both parameters of all three nucleic acids, indicating that the method can be used to characterize helix parameters of nucleic acids.

Table 3.3. RNA and DNA rise and twist are accurately calculated compared with Curves+[131] and 3DNA[132].

Nucleic acid	**Helix property**	**Curves+**[131]	**3DNA**[132]	**Helios**
A-DNA	rise (Å)	2.55 ± 0.00	2.54_8	2.56 ± 0.02
	twist (°)	32.7 ± 0.0	32.7	32.6 ± 0.1
B-DNA	rise (Å)	3.37 ± 0.00	3.37_5	3.37 ± 0.02
	twist (°)	36.0 ± 0.0	36.0	36.1 ± 0.1
A-RNA	rise (Å)	2.81 ± 0.00	2.81_2	2.80 ± 0.02
	twist (°)	32.7 ± 0.0	32.7	32.5 ± 0.1

All three nucleic acid structures were built using 3DNA[132]. For A-DNA and A-RNA, 11-bp duplexes were used. For B-DNA, a 10-bp duplex was used. One extra digit is provided for rise parameters of 3DNA because such was the precision used to generate the coordinates.

How does the accuracy of the method depend on the number of bp (in dsDNA and dsRNA) or nucleotides (in ssDNA and ssRNA) used in the fitting? The accuracy of helix rise, radius and twist were expected to depend on the total number of C1′ atoms used in the fitting. The radius, pitch and *PPT* of B-DNA are 10 Å, 34 Å and 10,[133] respectively, and its R/κ is 0.29. The radius, pitch and *PPT* of A-DNA are ~11 Å, 28 Å and 11,[133] respectively, and its R/κ is 0.39; the radius, pitch and *PPT* of A-RNA are ~11 Å, 30 Å and 11,[133] respectively, and its R/κ is 0.37. One turn of a B-DNA double helix has ten base pairs (bp), with ten atoms per strand. Based on our analysis of the test helices (consulting test helices with *PPT*=11 and R/κ=1/4), we expected to obtain accurate parameters when ~8 atoms in total were used, i.e. four bp of dsB-

DNA. **Figure 3.15** shows the results of the analysis of dsA-DNA, dsB-DNA and dsA-RNA. Four bp (eight atoms) were needed to accurately calculate the helical rise (**Figure 3.15a**) and twist (**Figure 3.15b**) of all three double-stranded nucleic acids.

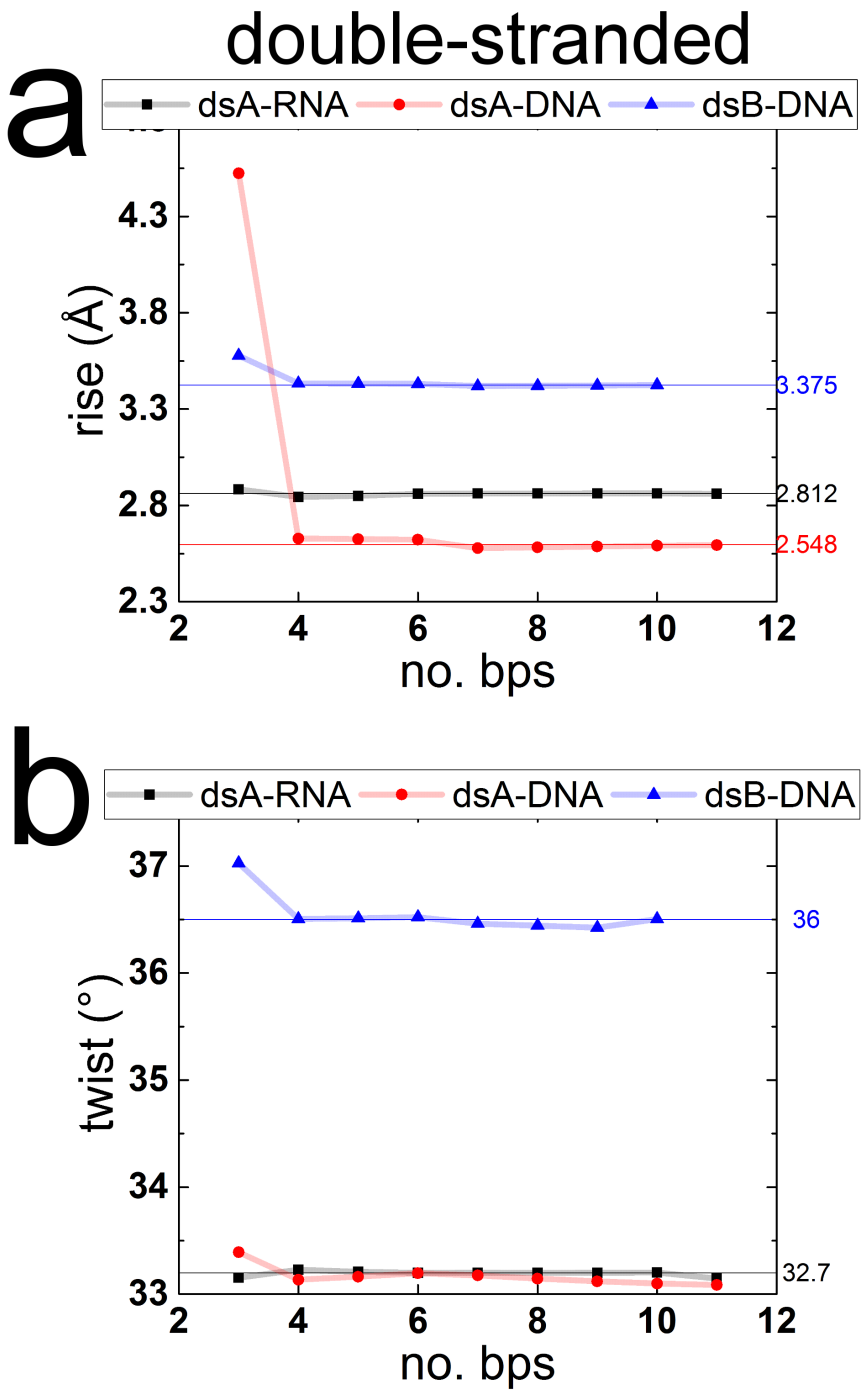


Figure 3.15. Helical parameters of double-stranded nucleic acids using atoms from both strands. The X axis shows the number of atoms used in the fitting. Results for dsA-RNA, dsA-DNA and dsB-DNA are shown as black, red and blue symbols respectively. (**a**) Helical rise, with horizontal lines showing the reference values[118] for dsA-RNA (black line), dsA-DNA (red line) and dsB-DNA (blue) line. (**b**) Helical twist, with horizontal lines showing the reference values[118] for dsA-RNA (black line), dsA-DNA (red line) and dsB-DNA (blue) line. Two atoms per bp were used in the fitting.

The helical properties of single-stranded nucleic acids were characterized next. As with dsB-DNA, eight atoms were required to accurately calculate the helical parameters for ssB-DNA (**Figure 3.16a,b**). However, only seven atoms were required to accurately calculate helical rise (**Figure 3.16a**) and helical twist (**Figure 3.16b**) for ssA-DNA and ssA-RNA. Eight atoms trace significantly less the one helical turn - 288° for ssB-DNA and 229° for ssA-DNA and ssA-RNA. Considering the peptide results in the previous section, it is possible that seven atoms represents the lower-limit required by the method to obtain reliable and accurate helix parameters of biomolecular helices. Seven points may be the critical value balanced between tracing more turns or a less coarse curve.

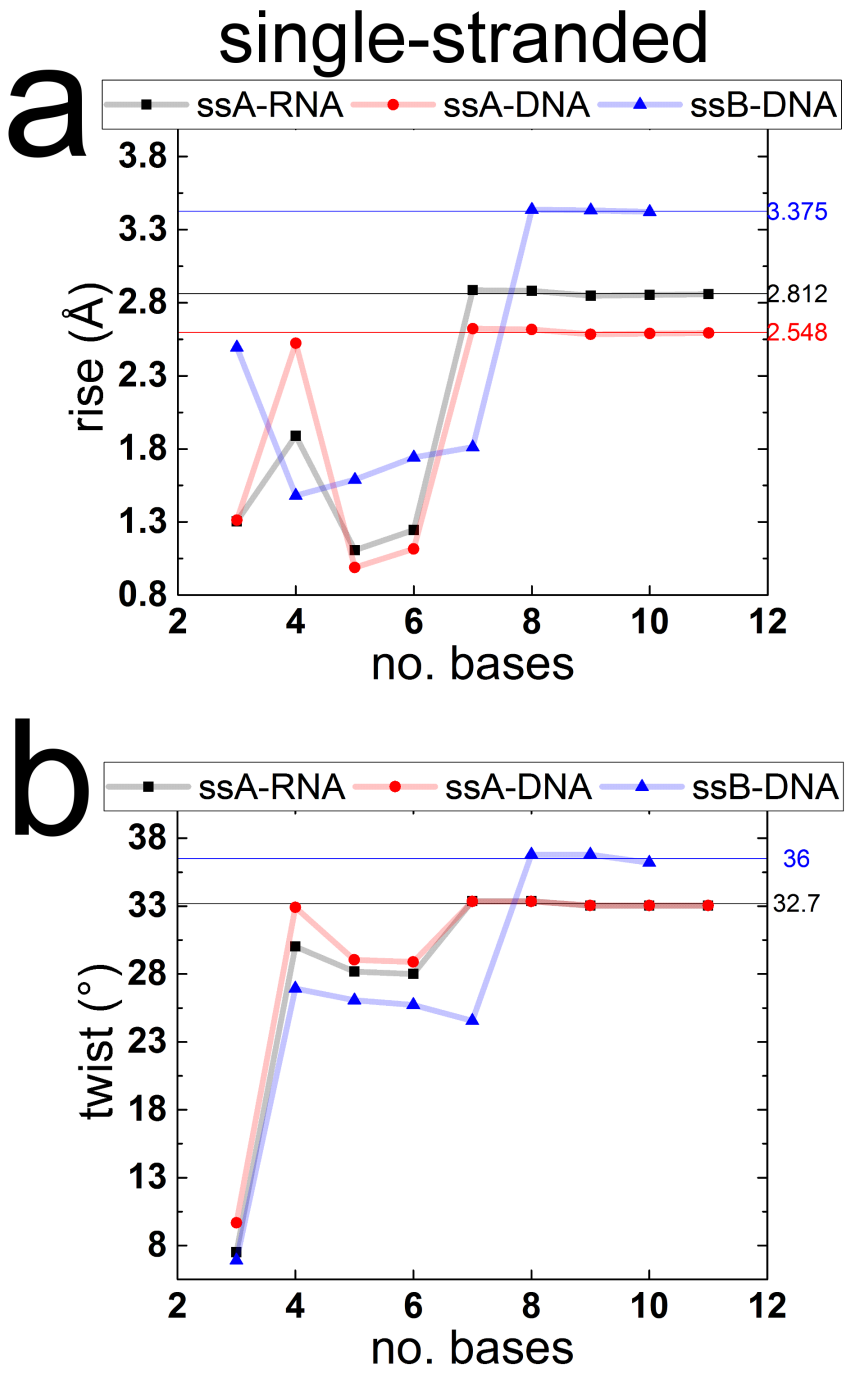


Figure 3.16. Helical parameters of single-stranded nucleic acids using atoms from both strands. The X axis shows the number of atoms used in the fitting. Results for ssA-RNA, ssA-DNA and ssB-DNA are shown as black, red and blue symbols respectively. (**a**) Helical rise, with horizontal lines showing the reference values[118] for ssA-RNA (black line), ssA-DNA (red line) and ssB-DNA (blue) line. (**b**) Helical twist, with horizontal lines showing the reference values[118] for ssA-RNA (black line), ssA-DNA (red line) and ssB-DNA (blue) line.

3.4.5. Characterizing superhelix protein tertiary-structure

Unlike other helix analysis approaches, our method can characterize the helical parameters of superhelical protein tertiary structures without empirical constraints. Based on a simpler version of the method presented here, we previously calculated[85] the helical parameters of a modular human transcription factor MTERF1[25], which is structurally homologous to TALE proteins[134]. Thus, we were motivated to determine the generality of the method and characterize an alternate modular superhelical protein, the TALE protein BurrH[128], along with the helical parameters of the nucleic acid to which the protein was bound. We expected that the helical parameters of the protein and the DNA would be similar because of the high degree of apparent structural complementarity (**Figure 3.17**). Points tracing the BurrH superhelix were defined as the Cα atoms of one amino acid from the same location in each module (**Table 3.2**). Without covalent connections between these atoms or a reference frame to define the superhelical axis, ours is the only approach capable of characterizing an irregular superhelical tertiary structure. The results of the analysis indicate that, as expected, the average rise and twist of the protein (3.28 Å ± 1.1 Å and 32.6° ± 2.4° respectively) were nearly identical to those calculated for the DNA (3.34 Å ± 0.4 Å and 31.8° ± 4.7° respectively). The radius of BurrH (20.5 Å) was larger than the radius of the DNA (6.6 Å) because the protein traces a wider helix that wraps around DNA. Deviations in the parameters reflect local variation of the steps between superhelical Cα atoms for steps between modules. The deviations in rise and twist for BurrH - 1.1 Å and 2.4° respectively - reflect the average helix irregularity between modules (steps). Superhelical repeat-step parameters, their variation and their complementarity to the bp-step parameters of DNA will be expanded upon elsewhere.

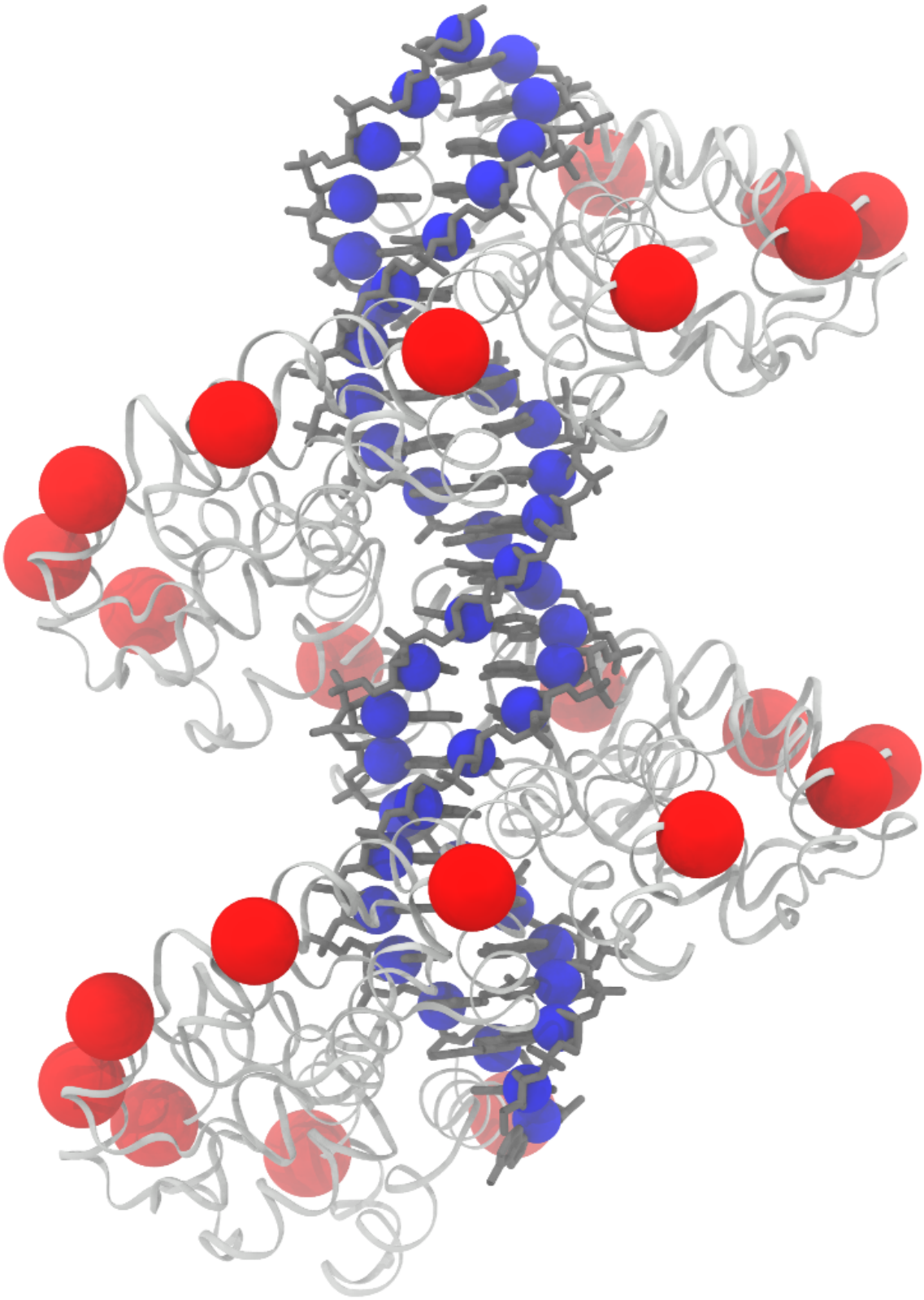


Figure 3.17. Helix complementarity in the BurrH-DNA complex (PDB ID: 4CJA[128]). Protein (light grey ribbons), superhelical Cα atoms (red spheres), DNA (dark grey sticks) and C1′ atoms (blue spheres).

3.5. Conclusion

We developed and tested a general method of helix-fitting that was geared towards characterizing geometries observed in structural biology applications. Validation tests were based on 399,360 test helices whose geometric parameters were representative of diverse biomolecules whose atoms might be perturbed from an ideal helix. The test helices were perturbed with known levels of noise to determine the sensitivity of the method. Helices frequently observed in structural biology applications were tested (peptide helical secondary structure elements, and nucleic acid single- and double-helices). The method was also used determine the helical complementarity of a TALE protein's superhelical tertiary structure and the DNA to which it was bound. Overall, the method introduced here is general, accurate and robust to noisy helical geometries. Based purely in geometry, our method can be used to characterize complementarity in protein-nucleic acid complexes, with potential applications in the design of genome editing reagents and biomaterials, astronomy and particle physics.

Chapter 4. Asymmetrically coupled structure specificity in protein-DNA complexes

Acknowledgement

This Chapter constitutes a manuscript in preparation by myself, Miguel Garcia-Diaz, Evangelos Coutsias and Carlos Simmerling. I conceived and wrote the manuscript, with edits and suggestions from the co-authors.

4.1. Introduction

A fundamental aspect of cellular vitality is the ability of transcription factors to bind certain DNA sequences with high specificity. Specificity enables transcriptions factors to precisely regulate the expression of genes so that a cell can respond to environmental, endogenous or developmental cues[135]. Designable specificity is key to programming artificial DNA-binding proteins that edit genetic material for biochemical research or biomedical therapy. There has been a surge of interest in clustered regularly interspaced short palindromic repeats (CRISPR)[136] and transcription activator-like effectors (TALE)[136], because the sequence specificity of these gene editing reagents can be arbitrarily reprogrammed. Although advances in achieving

increased specificity and affinity are being reported[136], the problem of identifying and curbing activity at non-target sites remains a significant challenge[136]. Simplified programmability of specificity makes TALE and CRISPR ideal test systems for studying the mechanistic details of protein-nucleic acid binding and recognition, which is in general not completely understood[137].

The structural work to optimize specificity has focused on the complementarity of direct hydrogen bonds (H-bonds) between CRISPR or TALE and a target DNA sequence[138-142]. Amino acid-nucleotide interactions are essential for specificity in both reagents: TALE amino acids interact directly with the target DNA[138, 140-142], whereas CRISPR interacts with the protospacer nucleic acid[143], which recognizes the target nucleic acid sequence. Structural complementarity between the protein architectures and the nucleic acids permits the appropriate arrangement of amino acid-nucleotide contacts for strong and specific binding.

The modular construction of TALE proteins provides a potentially simple recognition cipher of specific amino acid-base partners. This cipher is propped up by a superhelical tertiary structure that tracks the major groove of a target DNA sequence[142]. The superhelix arises from the symmetrical assembly of repeats orbiting a principal, helical axis (**Figure 4.1a**). TALE repeats are usually 34 amino acids[144], with the two amino acids at positions 33 and 34 responsible for sequence specific contacts with a nucleobase[138]. These two amino acids are the repeat variable diresidues (RVD); the first amino acid properly orients the second for a specific interaction with a base[141-142]. For example, the TALE protein dHax3[138] utilizes Asp in repeats 1 and 2 to H-bond with the amine group of the first and second dC base in the target DNA sequence (**Figure 4.1b**). dT is recognized by Gly (**Figure 4.1b**), which is uniquely able to accommodate the methyl of dT. Asn and Glu recognize dA via an H-bond to the acceptor N7 and donor N6 amine respectively, and Gln recognizes dG via an H-bond to the O6 acceptor[138].

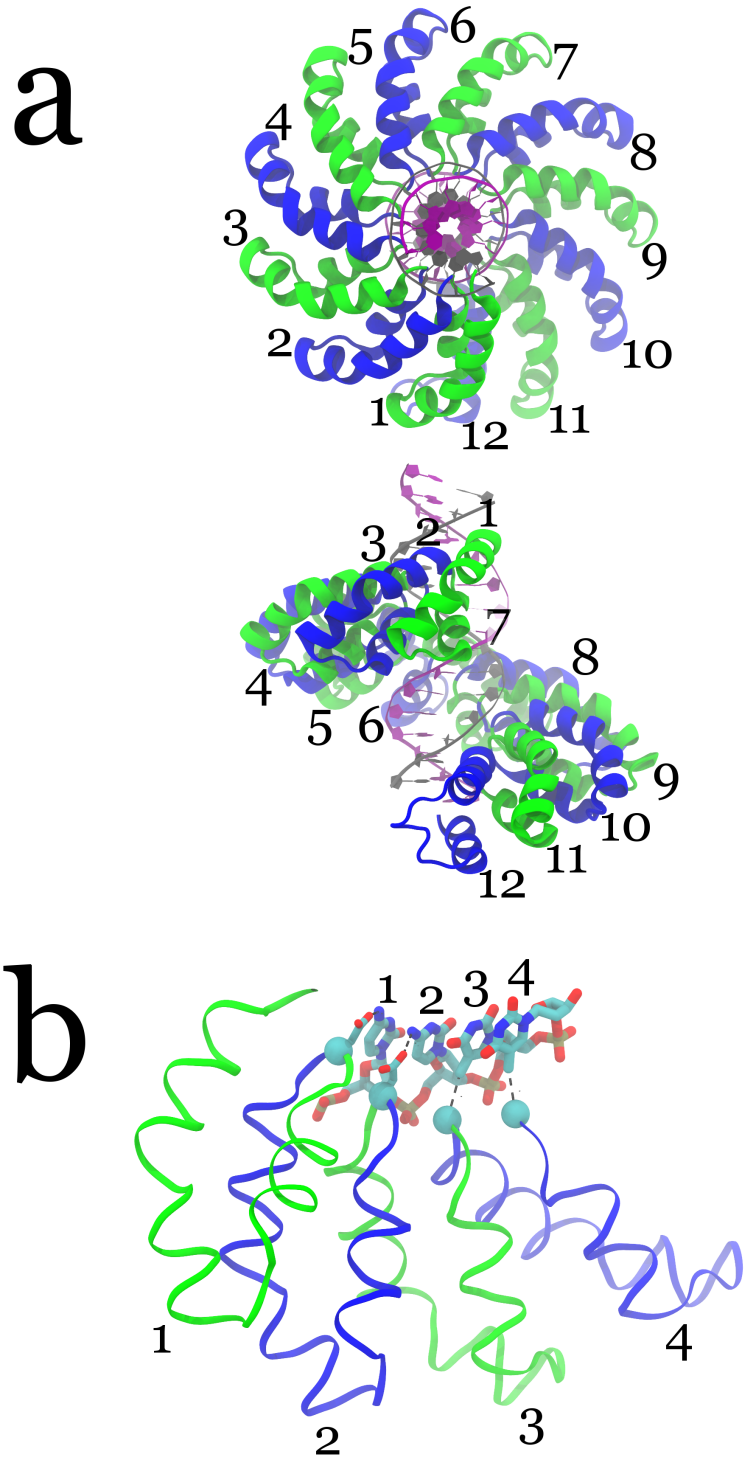


Figure 4.1. The TALE protein dHax3 bound to a 13 bp DNA sequence (PDB id: 4osh [138]). (**a**) Top: twelve 34 amino acid TALE repeats orbit the common superhelix-DNA helical axis. Bottom: the superhelical architecture tracks the major groove of DNA. Odd numbered repeats are colored green, even numbered repeats are colored blue; the sense and anti-sense DNA strands are colored grey and purple respectively. (**b**) Asp at RVD34 in repeats 1 and 2 contact dC and dC; Gly at RVD34 in repeats 3 and 4 contact dT and dT.

However, such a simple cipher might not be sufficient to realize the full potential of specificity, which might depend on small axial asymmetries of TALE repeats along the DNA sequence in which the identity of RVDs and their combination in a construct affect affinity and activity[145-148]. Molecular mechanics (MM) energy decomposition found that an RVD often establishes stronger interactions with the adjacent, 5' base than the base in which the RVD is in direct contact [145]. This asymmetry of interaction with the 5' base but not the 3' base may arise from the chemical structure of a nucleotide, which is asymmetric with respect to reflection through the base pair (bp) plane. The experimental and computational observations point to a model of TALE-DNA specificity that depends on local asymmetric RVD-DNA interactions, the composition of individual RVDs and the particular combination of repeats in a construct.

In general, specificity in protein-DNA complexes involves direct H-bonding between the protein and the DNA, and indirect readout which involves the sequence-dependent structure and deformability of DNA[137]. Direct and indirect readout is a structural partnership because the arrangement between the protein and the DNA supporting specific H-bonds depends on the structure of both molecules. Thus, subtle sequence-dependencies of DNA structure can affect the geometry of H-bond donors and acceptors to the protein. However, subtle structural distortions in the DNA bound by TALE proteins are assumed to be negligible and unimportant for specificity[138]. Consequently, the RVDs in TALE repeats are generally assumed to be freely interchangeable[71] despite the biochemical[145-148] and computational[145] observations to the contrary. This raises important questions for further optimization of specificity in TALE-DNA complexes. Is the complementarity between TALE repeats and the DNA independent from the construction of a repeat? Is repeat-base complementarity at one position coupled to other repeat-base partnerships in the rest of the complex?

We recently developed an approach to characterize structural complementarity between modular, superhelical proteins like TALE and the DNA to which they bind[85, 149]. The approach was also used to analyze MD simulations of human MTERF1 – a modular superhelix reminiscent of TALE[141] – switching from an extended conformation bound to its unwound, target DNA[25], to a compressed conformation that could bind to ideal B-form DNA (canonical DNA structure)[85]. We anticipated that application of the method may provide mechanistic insight into variations in the structures of highly specific genome editing reagent TALE.

Our goal was to determine whether the variations in the structural complementarity of one repeat-base step in the TALE protein dHax3 affected the complementarity of other repeat-base partners. We hypothesized that variations in the complementarity of an amino acid in one RVD in one dHax3 repeat would affect the complementarity of neighboring repeats-base partners. The similarity between the DNA and TALE helical geometries along the major groove of the target DNA sense strand were expected to decrease when an RVD in one repeat was mutated; not only for the one mutated repeat, but for all repeat-base partners along the sequence. Changes in complementarity would arise from packing adjustments imposed by different amino acid side chains. Interdependent TALE-DNA complementarity at the local repeat-base level would suggest that recognition involves structural coupling mediated by indirect readout. An immediate implication of indirect readout in TALE-DNA binding specificity would be the need to incorporate design principles beyond the current, modular approach.

4.2. Methods

4.2.1. Experimental dataset

Sixteen high-resolution crystal structures of the TALE protein dHax3 bound to DNA were obtained from the PDB (**Table 4.1**). Coordinates of the complexes were aligned with the C1' atoms of nucleotides 1 to 12 in the sense strand (see **Table 4.1**) utilizing PyMol[150]; dT0 was excluded from the analysis because it is not directly contacted by a repeat. The coordinates of these aligned C1' atoms were selected from the structure for subsequent helix analysis. The Deng *et al.* amino acid numbering scheme for TALE repeats was used (e.g. RVD amino acids are 33 and 34)[138]. The Cα atom of the first and 34th amino acid in each repeat, invariant "Gly1"[138] and "RVD34" respectively, was selected from the structures for subsequent helix analysis. Overall, three sets of coordinates for each dHax3-DNA complex were used to define three helices: the DNA helix (12 atoms), the Gly1 superhelix (12 atoms) and the RVD34 superhelix (12 atoms).

Table 4.1. Summary of X-ray crystal structures analyzed.

PDB ID	Resolution (Å)	DNA sequence	RVD sequence in repeat 7	Reference
		0 1 2 3 4 5 6 **7** 8 9 10 11 12		
4osh	2.20	T C C C T T T **A** T C T C T	Asn Ile (NI)	[138]
4osi	2.85	T C C C T T T **A** T C T C T	Asn Ile (NI)	[138]
4osj	2.79	T C C C T T T **A** T C T C T	Asn Asn (NN)	[138]
4osk	2.4	T C C C T T T **A** T C T C T	Asn Ser (NS)	[138]
4osl	2.45	T C C A A C T **G** C T A G A	Asn His (NH)	[138]
4osm	2.45	T C C A A C T **A** C T A G A	Asn His (NH)	[138]
4osq	2.26	T C C A A C T **G** C T A G A	Asn Arg (NR)	[138]
4osr	1.94	T C C A A C T **G** C T A G A	Asn Lys (NK)	[138]
4oss	2.4	T C C A A C T **G** C T A G A	Asn Gln (NQ)	[138]
4ost	2.0	T C C A A C T **A** C T A G A	Asn Cys (NC)	[138]
4osv	2.0	T C C A A C T **A** C T A G A	Asn Met (NM)	[138]
4osw	2.3	T C C A A C T **A** C T A G A	Asn Glu (NE)	[138]
4osz	2.61	T C C A A C T **A** C T A G A	Asn Pro (NP)	[138]
4ot0	2.49	T C C A A C T **A** C T A G A	Asn Thr (NT)	[138]
4ot3	1.94	T C C A A C T **A** C T A G A	Asn Leu (NL)	[138]
4oto	2.59	T C C A A C T **A** C T A G A	Asn Trp (NW)	[138]

DNA sequence letters in bold represent the base contacted by TALE repeat 7.

4.2.2. Helix analysis

Noisy points that roughly trace a cylindrical helix can be accurately characterized using our analytical method[149]. Its principle is simple: if a set of points represent a helix, then a circle is projected on the plane whose unit normal is parallel to the helical axis. Finding the helical axis involves finding this plane by isomorphically rotating the coordinates of the suspected helix in all possible 3D orientations using spherical coordinates. With the unit plane passing through the origin, only two spherical components are required to complete the full 3D search via latitude (φ) and longitude (θ)[149]. Characterizing the helical parameters for the set of DNA, Gly1 and RVD34 coordinates involved three steps. First, the sixteen complexes were aligned to a common frame by RMS-fitting the DNA in each complex to B-DNA whose helical axis rose along the z-axis. Second, a full spherical coordinates rotation search for the helical axis for each helix in each

dHax3-DNA complex was performed with 0.5° spherical coordinates grid-resolution; this operation verified that a unique helical axis was present for all of the helices traced by the invariant Gly1 Cα atoms (**Table 4.2**). The principal helical axes of Gly1 helices laid in the range φ [0°,15°] and θ [190°, 230°]. Thus, a subsequent fitting was performed in which the spherical coordinate rotations were constrained in the range φ [0°,15°] and θ [190°, 230°], with 0.5° spherical coordinates grid-resolution. This subsequent constrained helical axis optimization permitted a highly-controlled, direct comparison between the DNA, Gly1 and RVD34 coordinates for all sixteen dHax3-DNA complexes in the presence of substantial local helix distortions and irregularities. No other fitting parameters were required to characterize the helices. (The software is described in **Chapter 3** and **Appendix III**)

Table 4.2. Summary of helical axes from full spherical coordinates scan.

PDB	DNA			Gly1			RVD34		
	φ (°)	θ (°)	ω (°)	φ (°)	θ (°)	ω (°)	φ (°)	θ (°)	ω (°)
4osh	**49.0**	**176.5**	87.4	113.5	163.5	74.9	111.5	165.5	76.5
4osi	112.0	162.5	73.8	112.0	163.0	74.3	110.0	164.0	75.0
4osj	**124.5**	**78.0**	36.3	113.5	164.0	75.4	111.5	165.5	76.5
4osk	**125.0**	**88.0**	35.0	110.5	163.5	74.6	**59.0**	**175.5**	86.1
4osl	115.0	159.0	71.0	114.0	161.0	72.7	**141.5**	**177.5**	88.4
4osm	113.0	160.5	72.1	114.5	160.5	72.3	111.0	162.5	73.7
4osq	113.5	160.0	71.7	116.0	159.5	71.7	109.5	162.5	73.5
4osr	112.5	161.0	72.5	114.0	161.0	72.7	**139.5**	**178.5**	89.0
4oss	113.5	160.0	71.7	111.5	162.5	73.8	110.0	162.5	73.6
4ost	114.0	159.5	71.3	112.0	162.0	73.4	110.0	162.5	73.6
4osv	114.5	160.0	71.9	114.0	161.0	72.7	**139.0**	**178.0**	88.7
4osw	114.5	159.5	71.4	112.0	162.5	73.8	111.5	162.0	73.3
4osz	**123.5**	**77.5**	35.5	114.5	161.0	72.8	**137.5**	**178.5**	89.0
4ot0	113.0	160.5	72.1	112.0	162.0	73.4	110.0	162.5	73.6
4ot3	113.0	160.5	72.1	111.5	162.5	73.8	110.0	163.0	74.1
4oto	112.5	161.0	72.5	113.0	161.5	73.0	110.5	162.5	73.6

φ and θ denote the absolute helical axis orientation spherical coordinate components. ω denotes the angle between the helical axis and the y-axis. φ/θ coordinates in bold denote orientations that fall outside the range φ [109°,116°] and θ [159°, 166°].

Global helical parameters radius, pitch and sweep were obtained from the above described fitting. Pitch is the displacement per revolution along the helical axis. Helix radius is the radius of the projected circle. Helix sweep is the total angle subtended by the atoms (12 DNA C1', 12 Gly1 Cα or 12 RVD34 Cα) around the helical axis. Local base-step (DNA) or repeat-step (TALE) parameters rise were calculated as the vertical displacement between atoms in the helix; twist was calculated as the radial angle subtended by consecutive atoms in the helix. The details of the method are described in greater detail elsewhere[149].

4.3. Results

A TALE-DNA complex is highly symmetric around its principal axis, with the superhelix of the protein and the double helix of the DNA sharing a common helical axis[142]. Our previous structural characterization of the human transcription factor MTERF1 in search [85] and recognition mode[25] found that the superhelix of the protein and the double helix of the DNA shared a common helical axis as well, and that the superhelical pitch was precisely partnered with DNA. When MTERF1 binds to B-form DNA, the two molecules have a helical pitch of 34 Å [85]. Although TALE repeats are more homogenous in sequence and structure than the MTERF repeats[141-142], the high structural similarity to MTERF1 (DALI score 7.0)[142] suggests that varying RVD residues might cause local distortions in the superhelix that alter repeat-base complementarity in TALE-DNA complexes.

Biochemical experiments showed an asymmetrical affinity of TALE repeats along the DNA sequence (helical axis), with N-terminal repeats binding tighter than C-terminal repeats[145-148]. In the analysis that follows, sixteen previously published high-resolution crystal structures of dHax3-DNA complexes are analyzed to quantitatively assess the structural complementary of TALE and DNA. The array of crystal structures published by Deng *et al.* provides a rich resource to study the effect of point mutations on repeat-base complementarity[138]. The second (recognition) amino acid at the RVD, "RVD34" (**Figure 4.1b**), of repeat number seven (7 in **Figure 4.1a**) was varied in these structures: in four complexes bound to one DNA sequence (**Table 4.1:** 4osh to 4osk) and in twelve complexes bound to a second DNA sequence (**Table 4.1**: 4osl to 4oto). Thus, the effect of an individual repeat-base interaction on the overall structure could be analyzed. We found that repeat-base complementarity was coupled throughout the complex, at least when RVD34 in repeat seven is changed, suggesting that sequence specificity in TALE-DNA complexes involves a nuanced indirect recognition mechanism on top of the more commonly accepted one-to-one amino acid-base cipher.

4.3.1. Global helix complementarity in sixteen dHax3-DNA complexes

Extensive analyses of TALE-DNA complexes suggests that the protein and DNA geometries are highly complementary. However, the helical parameters of TALE proteins have yet to be analytically determined. What is the global helix complementarity of the sixteen dHax3-DNA complexes? To analytically characterize the global superhelical geometry of Deng *et al.*'s sixteen dHax3-DNA complexes, we used a new geometric approach[149]. We expected the superhelical pitch of RVD34 residues to match the pitch of the DNA bound by the protein; the radius of the protein should be slightly larger than the DNA because the protein wraps around the

DNA. **Table 4.3** summarizes the global helical parameters of the sixteen dHax3-DNA complexes studied here. Deng *et al.* estimated the superhelical pitch of dHax3 to be ~ 35 Å (for PDB ID: 3V6T [142]), which slightly underestimates the average value of 36.1 Å (RVD34) for the sixteen structures analyzed here. The radius of the helix traced by the Cα atoms of the RVD34 amino acids is always larger than the radius of the helix traced by the C1' atoms of the DNA. However, the difference is never larger than 1 Å. This indicates that the average positions of RVD34 amino acids are highly complementary to the nucleobases they contact. Amino acid side chains of RVD34 are accommodated in the space between the Cα and the nucleobase because the helices are offset in helical twist (**Figure 4.2**). The subtle variation in the pitch and sweep of the RVD34 and the DNA helices across the sixteen complexes suggests that the structural differences in these complexes might not be limited to the local changes expected when only one amino acid in one of the twelve TALE repeats is being varied. To characterize the potential distortions within the RVD34 and DNA helices required finer analysis of local steps between repeats and between bases.

Table 4.3. dHax3 and DNA helical parameters are globally matched.

		DNA			RVD34		
Base 7	Repeat 7	Radius (Å)	Pitch (Å)	Sweep (°)	Radius (Å)	Pitch (Å)	Sweep (°)
dA	NI	7.2	36.7	345	7.9	37.2	336.5
dA	NI	7.2	37.2	346.7	7.7	37.5	347.5
dA	NN	7.2	36.4	347.8	7.9	36.8	339
dA	NS	7	36.7	349.4	7.3	36.9	352.8
dG	NH	7.4	36.2	340.8	7.5	35.8	353.1
dA	NH	7.3	35.5	346.1	7.3	36	352.2
dG	NR	7.3	35.8	345.3	7.3	35.7	355.8
dG	NK	7.3	35.9	345	7.4	35.8	354.3
dG	NQ	7.4	35.9	343.4	7.3	35.6	355.9
dA	NC	7.3	36.4	342.9	7.3	35.6	357
dA	NM	7.4	36.3	340.5	7.4	36.1	351.3
dA	NE	7.3	35.9	344.9	7.5	35.8	353.1
dA	NP	7.3	34.8	346.6	7.6	36	344.6
dA	NT	7.3	35.5	346	7.3	35.5	356
dA	NL	7.4	35.9	344.1	7.3	35.8	355.9
dA	NW	7.3	35.5	348.6	7.4	35.6	355.1

PDB IDs: 4osh (dA-NI) to 4oto (dA-NW) from Deng *et al.*[138]. Tight-binding complexes are signified in yellow; moderate-binding complexes are signified in green. N: Asn; I: Ile; S: Ser; H: His; R: Arg; K: Lys; Q: Gln; C: Cys; M: Met; E: Glu; P: Pro; T: Thr; L: Leu; W: Trp.

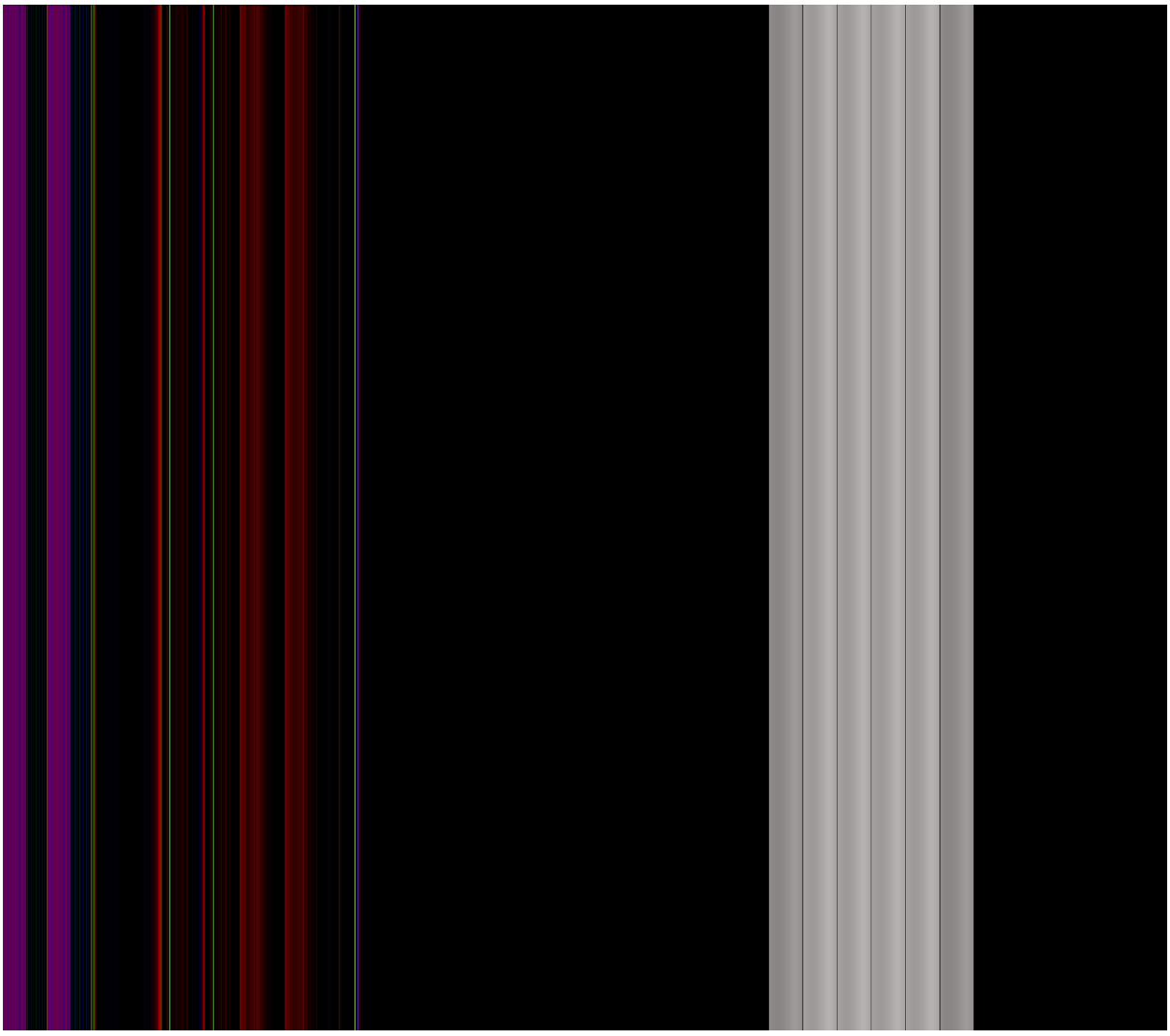


Figure 4.2. Overview of helix complementarity in sixteen dHax3-DNA complexes. For each comlex: PDB ID is presented on top left of each structure; RVD amino acid sequence identity of repeat number 7 and the identity of the contacted nucleotide. Bottom: Legend of atom color coding.

4.3.2. Are local deviations in RVD34 and DNA helices uniform?

Although the complexes are individually complementary with closely matched superhelix-DNA geometries, the structural differences between the sixteen different complexes indicates that dHax3 is sensitive to mutation of RVD34 at position 7. This observation raises the question of whether superhelix and DNA geometries are uniformly adapted, base-to-base and

repeat-to-repeat; or, whether superhelix and DNA geometries are asymmetrically adapted, distorted at some repeats and bases but not others. To asses how a mutation affects the helices in these complexes, we measured the local helix-step parameters rise and twist. Step parameters measure the displacement (rise) and rotation (twist) between consecutive atoms, relative to the helical axis. If the global structural changes arose from uniformly distributed local distortions, then all steps in the superhelix and the DNA should exhibit the same rise and twist. On the other hand if the global changes arose from a few local distortions, then helix adaptation would lead to variation in the rise and twist between superhelix and DNA steps. The likelihood that a superhelix and DNA randomly evolved geometries that were highly complementary seems negligible. More likely, helix complementarity may have evolved to tighten binding by reducing protein and DNA strain in the complex.

The first four complexes in **Table 4.1** are tight-binding, therefore the step parameters in these dHax3-DNA complexes should be uniform. **Figure 4.3** presents the dHax3 repeat-step and DNA base-step parameters rise and twist for four representative dHax3-DNA complexes including one of these strong-binding complexes. This control, tight-binding complex with the canonical RVD-base partner NI-dA exhibited uniform repeat and base step parameters (**Figure 4.3a**), as expected. The rise of the DNA (μ=3.2 Å, σ=0.2 Å) and the superhelix (μ=3.2 Å, σ=0.5 Å) were uniform across the steps; the twist of the DNA steps (μ=31.4°, σ=2.2°) and the superhelix steps (μ=30.6°, σ=2.9°) subtly undulated along the sequence, indicative of indirect readout via induced fit. For constructs with lower specificity, we expected local distortions in the steps of the superhelix at the position of the mutation. If dHax3 were perfectly modular and the combination of RVDs in a construct were fully independent, then distortions in the steps of the superhelix should be localized at the repeat housing the mutant. A weak-binding complex with

Trp present at RVD34 in repeat 7 exhibited a qualitative change in rise and twist of the superhelix (**Figure 4.3b**) compared with the control. The distortion propagates towards the N-terminus of the protein, but not the C-terminus. The moderate-binding complex with NH partnered with dG (**Figure 4.3c**) exhibits similar step distortions as the weak-binding complex (**Figure 4.3b**). However, the complementarity between the NW mutant and DNA differs from the complementarity between the moderate-binding NH mutant and DNA. Asymmetric propagation of helix distortions in the weak-binding (NW) and moderate-binding (NH) complexes may be matched differently to the distortions in the DNA. If the distortions in the rise and twist of the superhelix are matched by similar distortions in rise and twist of the DNA, then the complex could be expected to be tight-binding, despite the presence of distortions. To test that hypothesis, we calculated the difference in rise and twist of the superhelix and the DNA.

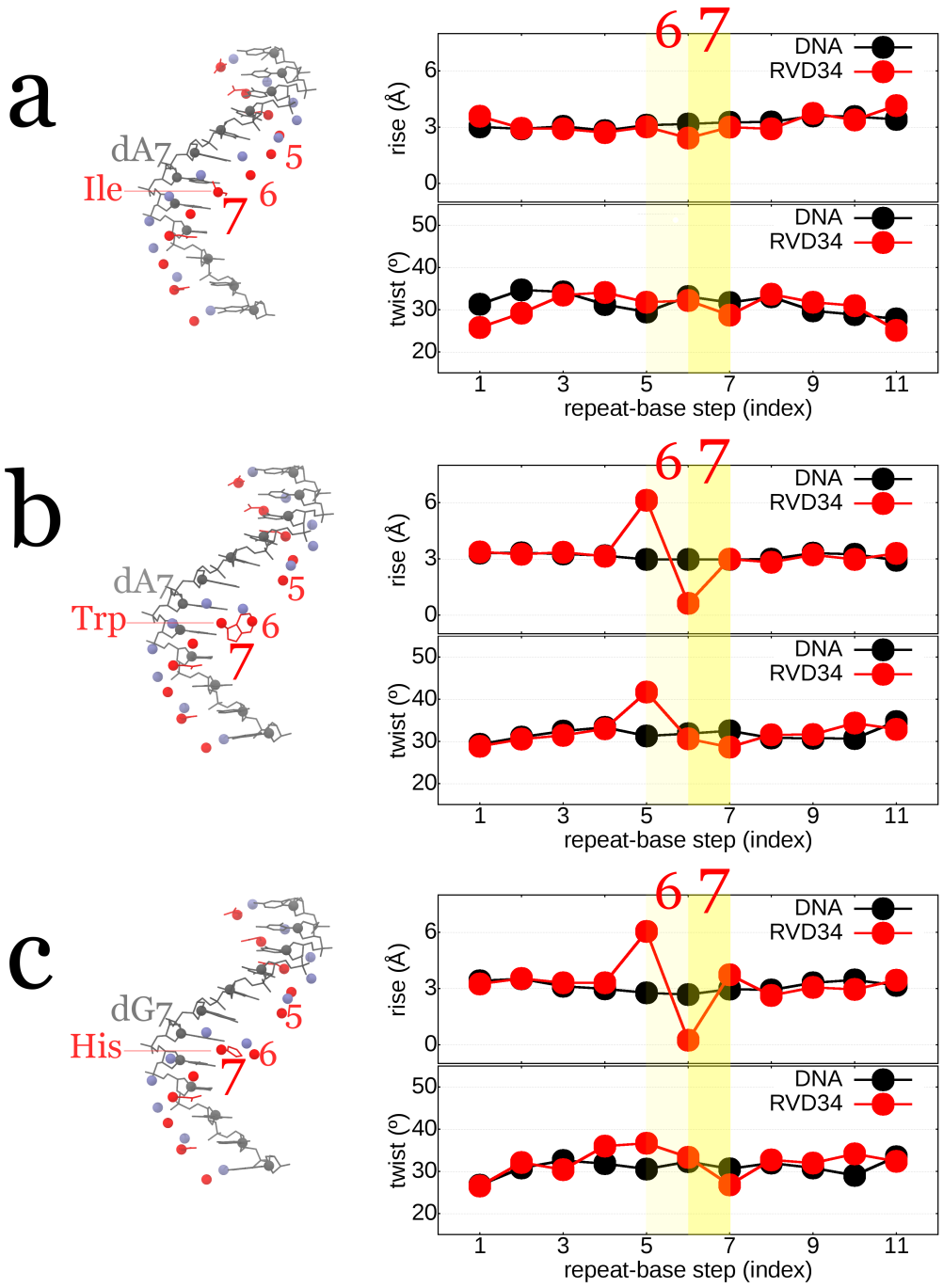


Figure 4.3. Local repeat-step and base-step parameters of representative dHax3-DNA complexes. The data plot the rise and the twist of RVD34 steps (red points) and the sense strand of DNA (black points). A vertical yellow band marks the position of dHax3 mutations: step 5 lies between the Cα at position 5 and 6; step 6 lies between the Cα in repeat 6 and 7. The DNA sense strand is shown as grey sticks, with C1' atoms as grey spheres; Cα atoms of RVD34 and Gly1 amino acids are shown as red and blue spheres respectively; amino acid side chains of RVD34 are shown as red sticks. (**a**) dHax3-NI with dA at repeat 7 of the DNA sequence (PDB ID: 4osh). (**b**) dHax3-NW with dA at repeat 7 of the DNA sequence (PDB ID: 4oto). (**c**) dHax3-NH dG at repeat 7 of the target sequence (PDB ID: 4osl).

4.3.3. How coupled are the helix complementarities of repeat-base partners?

We sought to determine the local complementarity between the superhelix and the DNA. If distortions in dHax3 are partnered with similar distortions in DNA, then the local helix step parameters should be similar and the difference between protein and DNA step parameters should be nearly zero. Changes in packing between repeats and between dHax3 and the DNA due to a mutation could cause the Cα atom of the mutant amino acid to shift without affecting the architecture of the protein. To resolve whether mutations caused a change in dHax3 architecture, rather than RVD34-base interactions, we defined a control superhelix defined by the Cα atoms of invariant Gly1, which do not contact DNA. Different helix complementarities between the RVD34-DNA helices and the Gly1-DNA helices would indicate that local recognition contacts do not affect global protein architecture. Similar helix complementarities would indicate that local protein-DNA contacts and the global dHax3 architecture are intertwined.

The tight-binding, control complex NI-dA displayed complementarity between the Gly1 superhelix and DNA (ΔGly1), and between the RVD34 superhelix and DNA (ΔRVD34) (**Figure 4.4a**). The differences in rise (Δrise) between repeats in the Gly1 superhelix and the bases in DNA were small (μ=0.1, σ=0.5) and the pattern of these small differences were similar to the pattern between repeats in the RVD34 superhelix and the bases in DNA (μ=0.0, σ=0.4). Twist was also complementary between the repeat-base steps in ΔGly1 and ΔRVD34 (**Figure 4.4a**), indicating that the superhelical architecture is structurally adapted to the target DNA sequence in a tight-binding dHax3-DNA complex. Interestingly, subtle variations are present in Δtwist across the dHax3-DNA footprint, ΔGly1 (μ=0.3, σ=2.2) and ΔRVD34 (μ=0.8, σ=3.0), with an undulating pattern of over- and under-twisting of the superhelix relative to the DNA.

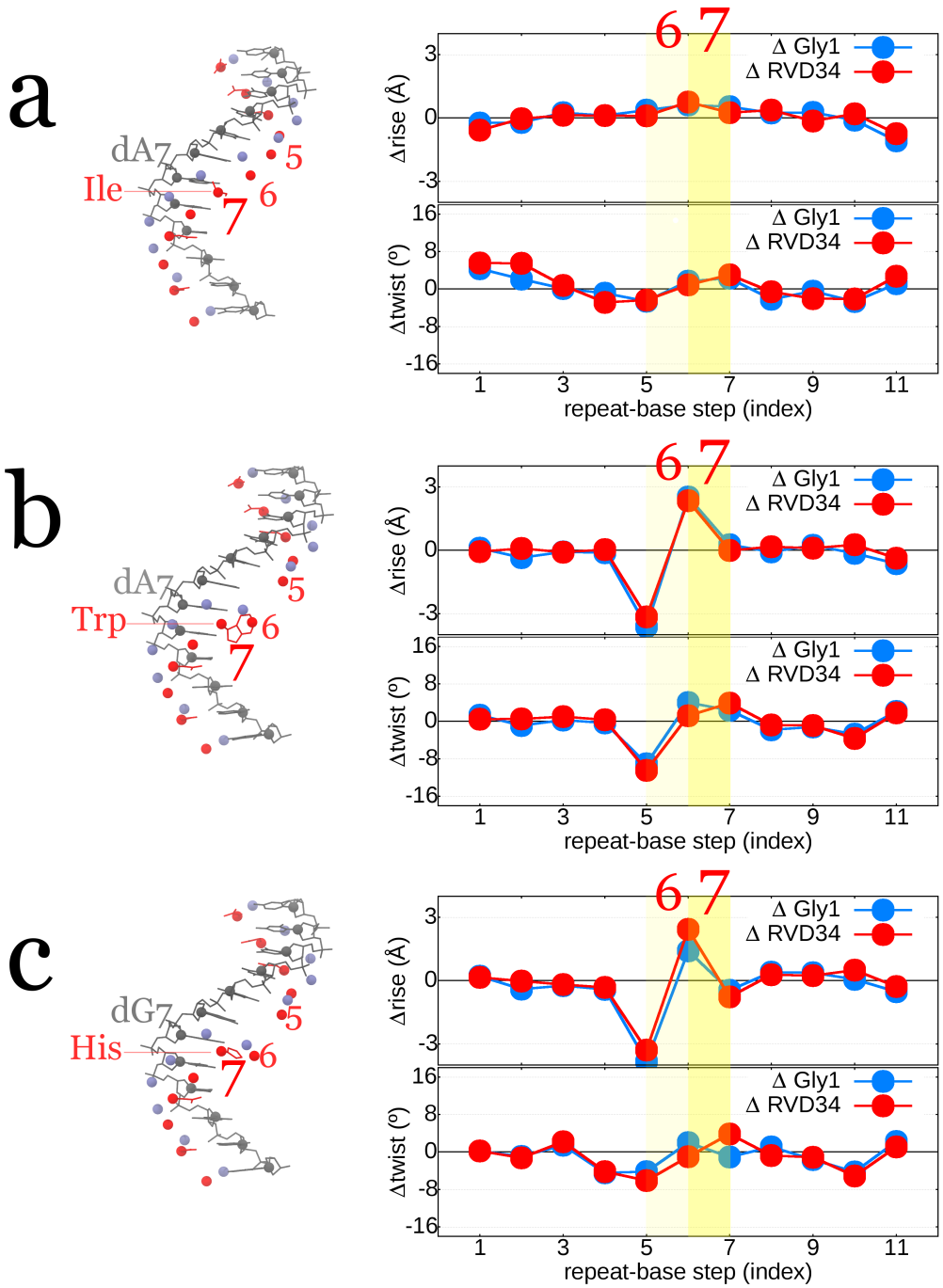


Figure 4.4. dHax3-DNA repeat-base helix complementarity. The data plot the difference in rise (Δrise) and twist (Δtwist) between Gly1 and DNA steps (blue points), and RVD34 and DNA steps (red points). A vertical yellow band marks the position of dHax3 mutation: step 5 lies between the Cα at repeat 5 and 6; step 6 lies between the Cα at repeat 6 and 7. The DNA sense strand is shown as grey sticks, with C1' atoms as grey spheres; Cα atoms of RVD34 and Gly1 amino acids are shown as red and blue spheres respectively; amino acid side chains of RVD34 are shown as red sticks. (**a**) dHax3-NI with dA at repeat 7 of the DNA sequence (PDB ID: 4osh). (**b**) dHax3-NW with dA at repeat 7 of the DNA sequence (PDB ID: 4oto). (**c**) dHax3-NH dG at repeat 7 of the target sequence (PDB ID: 4osl).

In the weak-binding complex Trp-dA, ΔGly1 and ΔRVD34 exhibit very similar patterns of rise and twist complementarity indicating that the mutation has altered the superhelical architecture of dHax3 along with its complementarity to DNA beyond the position expected, at the mutation (**Figure 4.4b**). Rise and twist complementarity are drastically altered at the two steps preceding the location of the mutation. Steps towards the C-terminus of repeat 7 are not altered relative to the control, including step 8 which involves repeat 7. The same asymmetric adaption of helix complementarity is observed for the moderate-binding His-dG complex (**Figure 4.4c**). In all sixteen complexes, the Gly1 and RVD34 superhelices display similar complementarities with DNA in rise (**Figure 4.5**) and twist (**Figure 4.6**), signifying a general trend in which sequence specific dHax3-DNA interactions are essential for adapting the global architecture of the superhelix. This indicates that mutation of RVD34 in repeat 7 alters the helix complementarity in an asymmetric manner, and the manner in which helix complementarity is affected also depends on the identity of the amino acid of RVD34 in repeat 7. These results provide further support for a subtle mechanism of sequence recognition by TALE proteins, mediated by asymmetric coupling (via induced fit) in the complementarity of individual repeat-base partners. This suggests that indirect readout is important for TALE specificity.

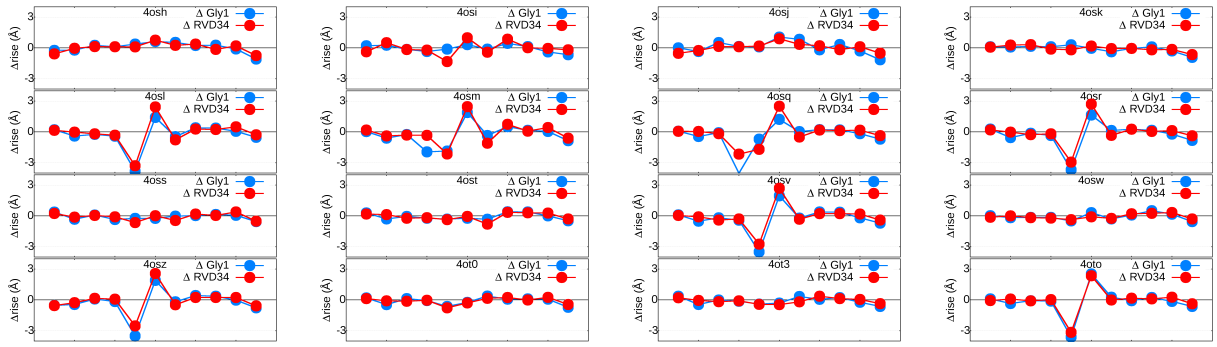


Figure 4.5. Helix rise complementarity in sixteen dHax3-DNA complexes. Helix complementarity of the control (Gly1) and recognition superhelices (RVD34) with the DNA (blue and red data respectively).

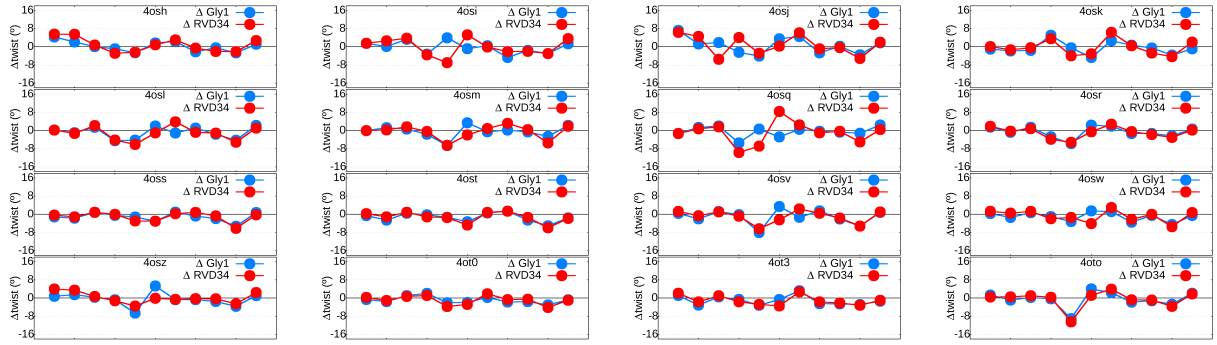


Figure 4.6. Helix twist complementarity in sixteen dHax3-DNA complexes. Helix complementarity of the control (Gly1) and recognition superhelices (RVD34) with the DNA (blue and red data respectively).

Despite the highly symmetric appearance of the model TALE protein dHax3 and its complex with DNA, we have found that local distortions across sixteen high-resolution crystal structures reveal asymmetry in the helix complementarity between TALE repeats and the nucleotides recognized by the protein. Mutations of RVD34 in repeat 7 caused distortions in the superhelix, which changed the complementarity to the DNA. These distortions propagated to repeats in the N-terminus of the protein, but not to repeats in the C-terminus. These results are consistent with biochemical observations that TALE proteins display N-/C-terminal repeat-asymmetrical affinity for DNA[145-148]. Asymmetrical affinity might be important for initial

binding, as TALE proteins rarely trace an open helix (< 360° sweep, i.e. < 10 repeats) that could permit DNA entry without considerable conformational change. It may be that an extended TALE superhelix contacts a few bp of DNA via the C-terminus, then twirls around the DNA, collapsing repeat-by-repeat into the major groove as the protein compresses into a geometry that is more complementary to DNA. Asymmetrical helix complementarity and affinity might also be important for nonspecific binding and target search, similar to the modular transcription factor Egr-1[152-153] or the superhelix MTERF1[85, 154]. A new view on the design of strong and specific binding between modular biomolecules such as proteins and nucleic acids might leverage asymmetry, rather than design against it[155], as a thermal trap to engineer energetic, geometric gradients needed for nanomechanical operations.

4.4. Conclusion

TALE-DNA specific complexes exhibit global and local repeat-base complementarity relative to their common helical axis. We hypothesized that a mutation of the repeat variable di-residue (RVD) involved in direct H-bonding with a base in the target sequence would affect the complementarity beyond the repeat harboring the mutation. To test our hypothesis, we analyzed sixteen high-resolution X-ray crystal structures of dHax3 and its variants. We found that RVD mutations not only affect local DNA complementarity, as expected, but the mutations affected the DNA complementarity of all repeats towards the N-terminus of the protein, but none of the repeats to the C-terminus of the protein. These results reveal a potential coupling between the

repeats in a TALE protein, suggesting that specificity might involve indirect readout and protein strain that could be considered for greater specificity designs.

Chapter 5. A human transcription factor in search mode

Acknowledgement

This Chapter constitutes the full text from a manuscript that published by Kevin Hauser, Bernard Essuman, Yiqing Elissa He, Evangelos Coutsias, Miguel Garcia-Diaz and Carlos Simmerling, in Nucleic Acids Research, Volume 44, Issue 1, pages 63-74, DOI: 10.1093/nar/gkv1091. Professor Simmerling, Professor Garcia-Diaz and I conceived the project; I wrote the manuscript, with edits and suggestions from the co-authors.

5.1. Introduction

One in ten genes in the human genome encodes a transcription factor (TF)[4], and once expressed, TFs direct the expression of other genes. TFs adapt conformation to switch function: to bind, search, or recognize DNA. To rapidly respond to stimulus, TFs must locate target DNA quickly. 3-D diffusion from solution directly onto the target DNA site, amongst an excess of nonspecific sites, predicts on-rates ten-fold slower than observed *in vivo*[5]. Thus, 3-D diffusion and 1-D facilitated diffusion (sliding) likely drive target search[6-11]. Frustration during 1-D diffusion can arise when affinity for nonspecific DNA is high. Theory predicted[10-11] and experiments on p53 corroborated[12-13] that TFs most likely switch from a rapid search mode to a

tight-binding recognition mode by changing conformation. In search mode, scanning is facilitated by fleeting nonspecific binding with ~ 1 k_BT energy gaps that reduce residence time on noncognate sites[11]. Significant perturbation of the DNA structure is unlikely on such small energy and time scales. Thus, a transcription factor should be able to weakly bind a random sequence of DNA, that is presumably B-form[14-15].

Conformational change regulates recognition.

During recognition the TF can conformation ally adapt to optimize specific contacts that directly recognize chemical groups present in the cognate sequence, shifting the free energy landscape to a regime with large energy gaps and high barriers between specific and nonspecific sites[11]. The kinetic aspect of recognition is analogous to enzyme inhibitors that exhibit long residence times following an induced fit conformational change in the protein[16, 156]. Dynamics of the tightly bound TF can also induce DNA deformation, potentially giving rise to dynamic indirect readout via sequence-dependent deformability of DNA, or to shape readout (static indirect readout)[18-20]. Therefore, conformational changes in the TF and in the DNA during recognition are coupled dynamic processes that depend on atomistic intermolecular interactions—direct readout—and intramolecular interactions—indirect readout and protein strain. For example, NMR transverse relaxation rate measurements of the lac repressor headpiece reveal that amino acids involved in direct readout in the recognition mode form nonspecific interactions with the phosphate backbone in the search mode[21]. The data suggest that conformational adaptation from search to recognition modes includes switching nonspecific contacts with the DNA backbone to specific TF-nucleobase interactions. TF-DNA binding and

recognition is thus a function of the relative energies of the search and recognition metastates, which is determined by the thermodynamics and kinetics of TF and DNA conformational change.

The relative importance, however, of direct and indirect readout during the transition from search to recognition mode is poorly understood. Insight into the mechanism of conformational change, and thus of recognition, would be facilitated by high-resolution structural data for specific and nonspecific complexes. The lac repressor headpiece[21] and the enzymes BamHI[22], BstYI[23] and EcoRV[24] are prototypical DNA-binding proteins for which static structures of specific and putative nonspecific complexes have been experimentally characterized. However, the lifetime of a true nonspecific complex is by definition fleeting[11]. To favor binding at a single nonspecific site requires alterations to the DNA or protein, truncated constructs, or protein-DNA cross-links that artificially stabilize the energy of the nonspecific complex. In these altered complexes, usually only a few interactions have been modified and therefore a subset of the cognate recognition contacts may still be present – "hemispecific recognition"[23] - and the DNA is frequently shifted from B-form. For example, the structure of the human transcription factor MTERF1 was solved for a putative nonspecific complex in which a subset of the recognition interactions was eliminated. The DNA conformation was deformed, however, and resembled that seen in the fully cognate complex[25]; the DNA conformation in putative nonspecific complexes of BamHI[26], BstYI[23] and EcoRV[27] enzymes also resemble that in the cognate complex. Consequently, it is unclear how accurately these altered complexes represent the actual structure during rapid search, outside the influence of methods used to redirect binding specificity and trap a unique noncognate structure. Moreover, static snapshots do not resolve dynamics. A complete mechanistic picture of how TFs regulate gene expression

would involve a dynamic model of the ensemble of structures that correspond to search mode, as well as an atomistic description of the conformational and energetic changes that take place during the transition from nonspecific to specific complexes. Here, we use a combination of experimental structural data and molecular dynamics simulations to address the first element in this challenge and develop a dynamic model for nonspecific DNA binding, using as a model system the human mitochondrial transcription factor MTERF1.

The MTERFs, Mitochondrial TERmination Factors, are vital human TFs[25]. MTERFs are involved in regulating gene expression in the mitochondria of eukaryotes and also the plastids of plants[157]. MTERF1 is the canonical mitochondrial transcription terminator, responsible for modulating the expression of mitochondrial DNA (*mterf*) genes[158] by preventing L-strand transcription interference[159] within the circular mtDNA. Mitochondrial dysfunction resulting from alterations in mitochondrial gene expression has been correlated with aging, cancer, diabetes, and neurological disorders like Parkinson's disease and Alzheimer's disease[160-161]. Since MTERF proteins play essential roles mediating gene expression in mitochondria and chloroplasts, further understanding the mechanism by which MTERF1 interacts with DNA will contribute to our understanding of organellar biology and the connection between bioenergy and disease. Furthermore, defects in MTERF1 binding have been previously associated with mitochondrial disease[25, 162-163].

MTERF1 has a modular tertiary structure topology. Modularity in TF tertiary structure is important for combinatorial discretization of binding site specificity and evolutionary stability[164], possibly explaining the abundance of organellar TFs that are modular[165]. The TAL effector is another superhelical TF whose modular structure eases the retargeting of specificity for genome editing[141]. Park et al. showed that it is possible to customize macromolecular topologies by

mixing and matching leucine-rich repeat modules[166]. Overall, modularity simplifies the challenge of characterizing the mechanism of protein-DNA search and recognition because modules can act as small and independent but linked proteins, thereby reducing the mechanical degrees of freedom likely to be important for functional dynamics.

The X-ray crystal structure of MTERF1 bound to the cognate termination sequence was resolved at 2.20 Å[25], revealing a superhelical topology[167] complementary to the bound DNA structure (**Figure 5.1**). The apparent architectural complementarity of MTERF1 and DNA simplifies the structure-dynamics-function relationship in MTERF1-DNA binding. MTERF1 is modular, composed of eight 33-residue *mterf* modules[25] that represent steps in the superhelix (**Figure 5.1A**). Each *mterf* module is composed of a triangular arrangement of 3 short helices stabilized by a hydrophobic core. Fewer packing interactions between motifs suggests that local changes in module-module stacking could give rise to global dynamics in superhelical pitch and radius. The macrodipole of the central helix in modules 5 through 8 align with the DNA phosphate backbone. Capping the central α-helices in all 8 modules are conserved proline residues, creating an S-loop that prevents a steric clash with the DNA.

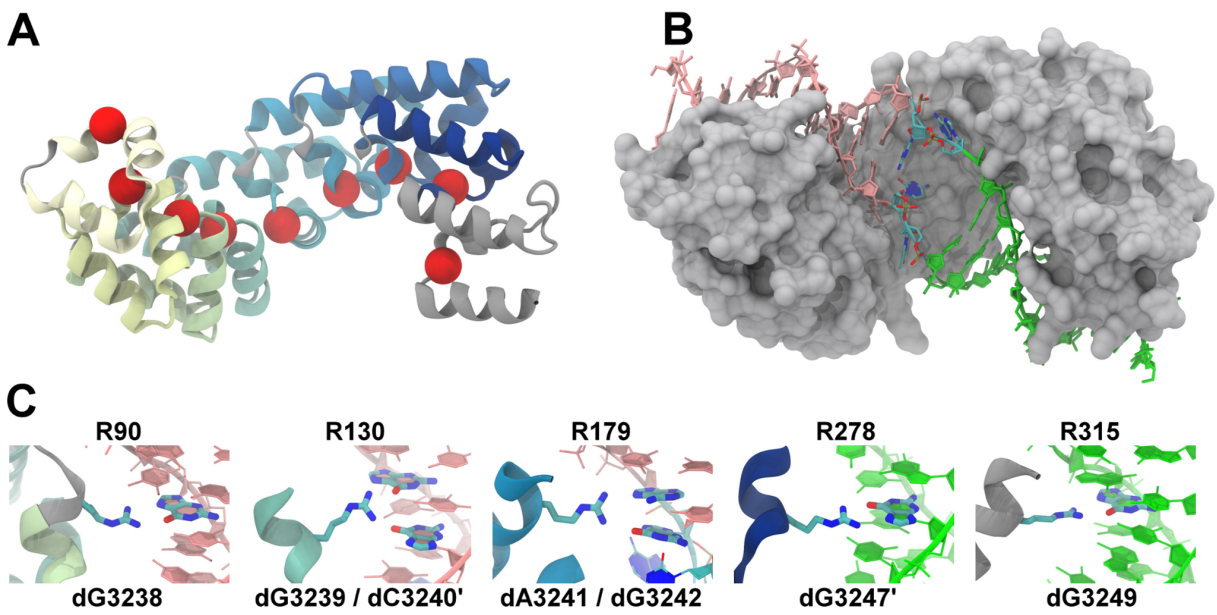


Figure 5.1. Human MTERF1 is a modular, superhelical TF that unwinds target DNA in recognition mode. (**A**) MTERF1 is modular, composed of 8 *mterf* modules (colored from yellow to blue). Intervening S-loops and the C-segment are grey. Superhelical residues are shown as red spheres. (**B**) The superhelical topology of MTERF1 (grey MSMS surface[168]) tracks the major groove of DNA. The bound DNA (displayed as sticks and ribbons) is unwound, which is focused on the central three base pairs (colored by element), while the N-site of the DNA (pink) and the C-site of the DNA (green) remain essentially undeformed. (**C**) MTERF1 forms direct readout interactions in the N-site and C-site: R90 forms a double H-bond with the N7 and O6 of dG3238 (light-strand, LS); R130 bridges a cross-strand dinucleotide step, H-bonding with O6 of dG3239 (LS) and dG3240' (heavy-strand, HS); R179 bridges a dinucleotide step on the HS, H-bonding to the N7 of dA3241 and O6 of dG3242; R278 double H-bonds with the N7 and O6 of dG3247' (LS); R315 double H-bonds with dG3249.

Unwinding of the bound DNA is dramatic (**Figure 5.1B**), providing structural support for a roadblock termination mechanism[25, 169]. The unwinding induced by MTERF1 is focused on the central three base pairs (**Figure 5.1B**), which are everted from the duplex and stabilized by hydrogen bonds and stacking interactions. On either side of the flipped bases the DNA is essentially B-form, but the helical axis is bent ~30º over the everted bases (**Figure 5.2**). Importantly, although the DNA is unwound, MTERF1 tracks the major groove across the full 22 bp footprint. The conserved proline residues within each motif line the major groove; tracing

their path outlines the superhelical topology of MTERF1 (**Figure 5.1A**) and the complementarity to the unwound DNA (**Figure 5.1B**). MTERF1 forms direct readout interactions with the bases in the B-form N-site and C-site segments of DNA, presumably stabilizing the intervening distortion in the duplex. (**Figure 5.1C**).

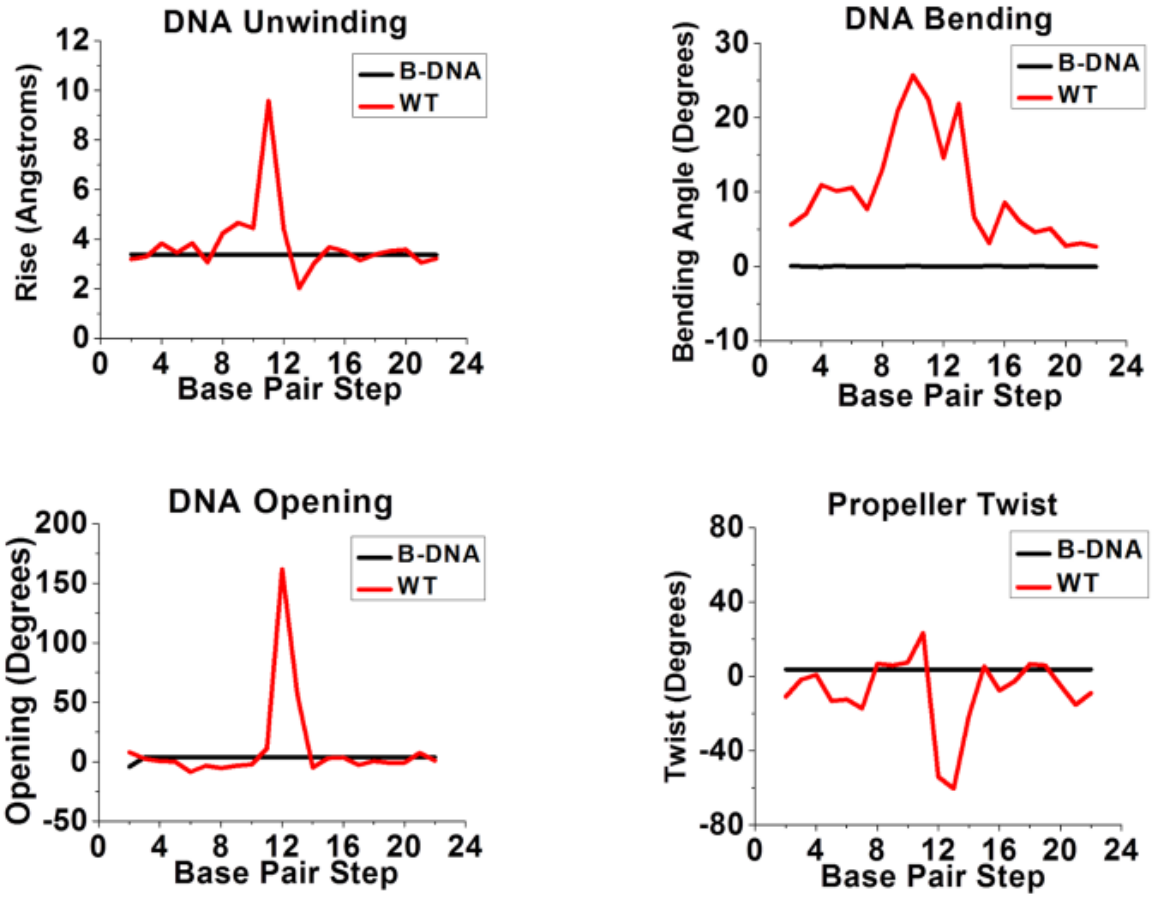


Figure 5.2. Structural analysis of DNA in the MTERF1-DNA specific complex[25]. Using Curves[170], the base pair step parameters rise distance (top left), bending angle (top right), opening angle (bottom left), and twist angle (bottom right) were calculated. For reference, the parameters for B-DNA are shown in each panel.

Representing a key knowledge gap in MTERF biology, neither the apo mode nor the search mode of MTERF1 has been structurally characterized. What might the conformations of apo MTERF1 and the MTERF1-DNA nonspecific complex be? MTERF1 binding to a transiently unwound DNA duplex seems unlikely since similar DNA deformation in EcoRI was estimated to cost ~100 kcal/mol in strain and entropy[15]. We explore two potentially more reasonable models in **Figure 5.3**. First, the conformation of holo and apo MTERF1 might be similar, as observed for BamHI[22], implying MTERF1 binds B-DNA and only the DNA changes conformation during recognition (**Figure 5.3, Model A**). However, we show below that the conformation of MTERF1 observed in the specific complex cannot accommodate B-form DNA without extensive steric clashes, as one might presume given the high level of DNA distortion observed in the complex. Alternatively, apo MTERF1 may be capable of adapting conformation to bind B-form DNA via induced fit[171], following a fly-casting binding mechanism[172], in which an unstructured tail increases the protein-DNA collision radius, or a gated binding mechanism[173], in which the protein oscillates to admit or deny ligand (DNA) entry into the binding pocket (**Figure 5.3, Model B**). Attempts to crystallize the protein in the absence of DNA were unsuccessful (Garcia-Diaz, unpublished data), supporting our hypothesis that apo MTERF1 is flexible or locally unstructured similar to p53[174], lac repressor headpiece[21] and the tails of SRY[175]. The difference between these models lies largely in the extent to which dynamics of MTERF1 plays a role in DNA binding and recognition.

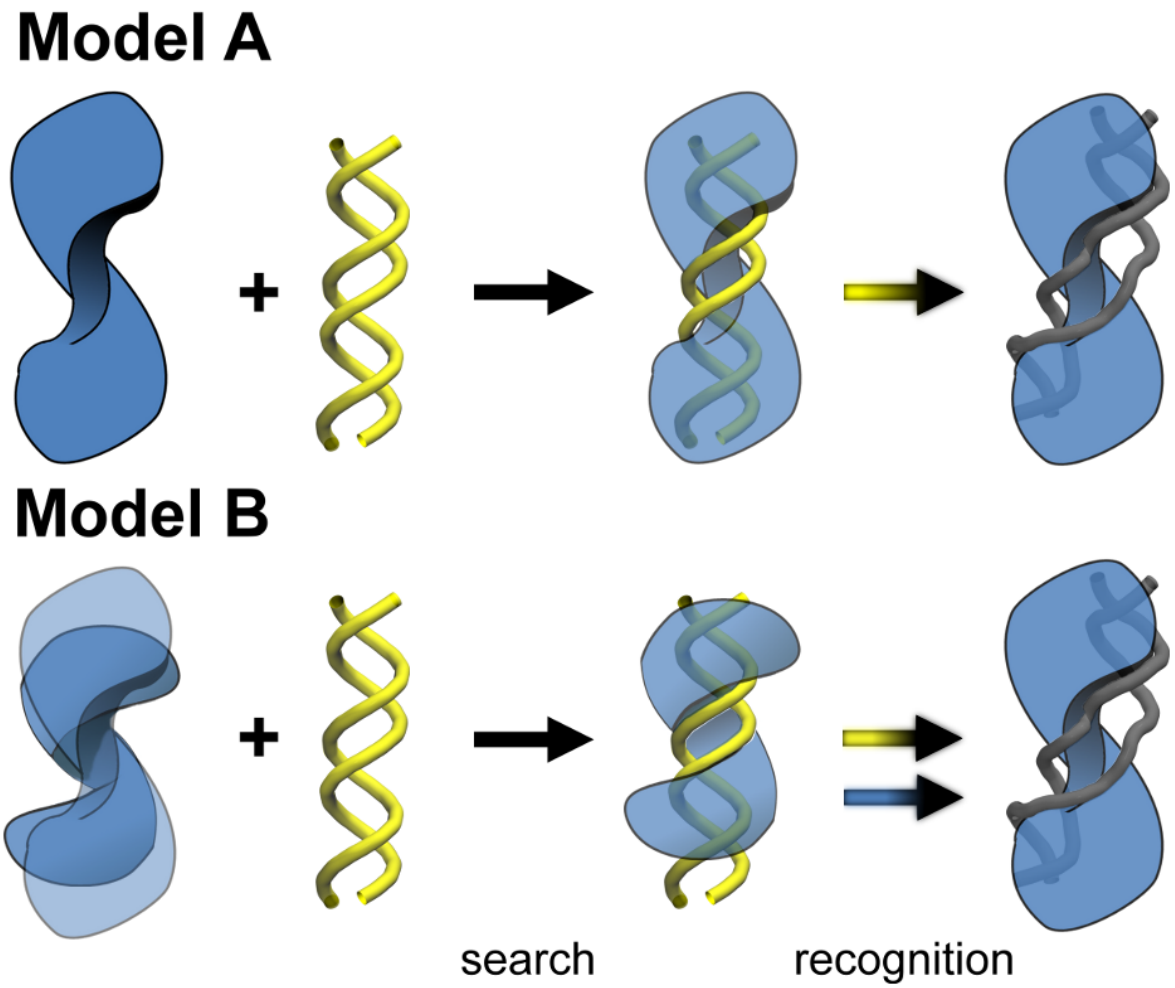


Figure 5.3. Potential MTERF1-DNA binding and recognition mechanisms. Pre-existing unwound DNA that MTERF1 can bind is not likely a viable mechanism (see text). We consider models with either singly or doubly induced fit. **Model A**: MTERF1 (blue) undergoes minimal conformational adaptation during binding and recognition, with the structure of apo MTERF1, MTERF1 in search mode (nonspecific complex) bound to B-DNA (yellow), and MTERF1 in recognition mode bound to unwound DNA (grey) all being similar in structure. During recognition, only the DNA undergoes conformational change (yellow arrow). **Model B**: apo MTERF1 is flexible, sampling a diverse ensemble of structures including those with a helical topology similar to B-DNA. During doubly induced fit recognition both MTERF1 and DNA undergo conformational change, blue and yellow arrows, respectively.

While p53, lac repressor headpiece, and SRY have been extensively studied in literature, how any TF undergoes a search to recognition conformational switch remains a gap in our knowledge. What might the conformational switch be for MTERF1? We hypothesize that

subsequent to sliding to the target, unpacking of the central *mterf* modules near the flipped bases might accompany DNA unwinding during recognition, allowing the superhelical pitch of the TF to adapt to, or perhaps drive, distortion in the curvature of the major groove during unwinding. Molecular simulations have been used in the past to study the role of flexibility in protein-DNA complex recognition[176-179] and inhibition[180]. Here, we report results of coarse grain elastic network model calculations as well as μs-timescale atomistic molecular dynamics (MD) simulations. Despite fundamental differences in the methods, both approaches support the same conclusion that the superhelical topology of MTERF1 is dynamic. The ensemble of structures obtained in MD samples a broad range of superhelical pitch and radius, including conformations matching the corresponding pitch and radius of B-DNA. Docking these low pitch apo MTERF1 structures to a B-DNA duplex resulted in a stable, dynamic complex in which MTERF1 shows 1D diffusion along the major groove of B-like DNA, providing an atomic resolution, dynamic model for a model TF searching DNA.

5.2. Methods

5.2.1. Helix analysis

Calculation of pitch and radius of MTERF1 used the Cartesian coordinates of the Cα atoms in positions that most closely track the major groove of DNA. The Cα atoms of the S-loop forming prolines, with two exceptions, defined the steps along the helix. First for motif 6, the Cα of A207 was used instead of P205 because the distance between the Cα atoms in motifs 5 and 6

and between modules 6 and 7 was significantly larger and smaller than other steps, respectively. The distinctive geometry of modules 5, 6, 7 that track the unwound central site of DNA is likely related to how MTERF1 unwinds DNA. Also, P205 of module 6 is in a GPG loop, the flexibility of which might potentially lead to local changes that could affect measurement of global dynamics. Second, the Cα of W311 was used in the C-segment (**Figure 5.1A**), which lacks a proline residue. The positions of the superhelical residues are shown in **Figure 5.1A**.

The Cα coordinates for the nine superhelical residues (**Figure 5.1A**) were projected onto a rotatable plane using equation (3-9) (**Chapter 3**) to find the plane that contained the best circle according to a linear least squares procedure solved by singular value decomposition (see **Appendix III**). The radius of the circle (helix radius) on the rotated plane was obtained directly from the fitting solution equation (3-10) (**Chapter 3**). With respect to the rotated plane and its frame, the sum of the angles swept between consecutive Cα atoms (8 angles between 9 helical steps) gave the helical sweep Φ. The superhelical pitch, κ, was then the distance between the first and last atom (e.g. $C\alpha_1$ and $C\alpha_9$) along the helical axis, Δz, multiplied by $\Phi/2\pi$.

5.2.1.1. Finding the helical parameters of apo MTERF1

An unconstrained fit of the superhelical residues was performed for the full ensemble of apo MTERF1 structures, after first RMS aligning the structures to the superhelical residues (using the recognition mode structure as the reference) to remove rotational and translational motions. Then the dependence of the helical parameters was analyzed, including fit residual, on the helical axis orientations as apo MTERF1 samples a breadth of conformations (**Figure 5.4**). A scatter plot of helical axis orientations shows the full range of θ is being sampled while a 90°

range in φ is being sampled, but clusters of distinct orientations are apparent (**Figure 5.4A**). To determine why different structures of apo MTERF1 lie along rather different helical axes, heat maps were plotted of the residual (**Figure 5.4B**), radius (**Figure 5.4C**), and pitch (**Figure 5.4D**). Configurations with low residual – φ in [50°,70°] and θ in [240°,300°] – lie in the same region of the grid where structures exhibit low radius. That same region of the grid also presents a range of superhelical pitch in which 30 Å is the minimum.

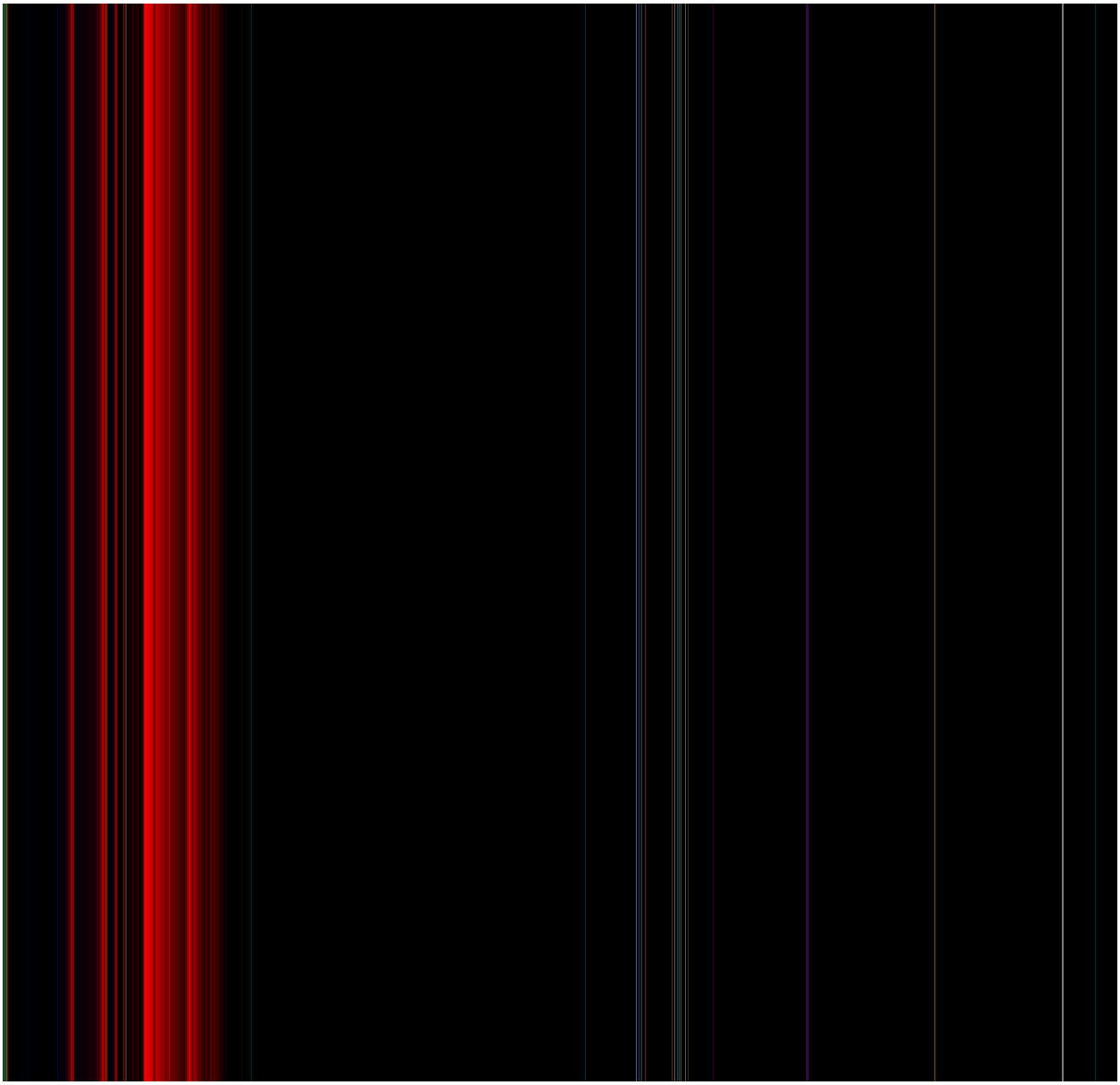


Figure 5.4. Calculating superhelical parameters of apo MTERF1. (**A**) Scatter plot of superhelical axes. Heat map of (**B**) residual, (**C**) radius, and (**D**) pitch. (**E**) Low pitch apo MTERF1. (**F**) High pitch apo MTERF1, the conformation corresponding to the structure found in the specific complex. (**G**) Very high pitch apo MTERF1. In (**E**) and (**F**) the unconstrained helical axis orientations were in φ in [50°,70°] and θ in [240°,300°] whereas the superhelical residues of the structure in (**G**) adopted an orientation in an alternate region of the map, the pitch and radius of which are shown in red. Data represents one of the 8 apo MTERF1 simulations.

Figure 5.4E shows a surface representation of an apo MTERF1 structure with low superhelical pitch – 35 Å. Two other conformations of MTERF1 with high pitch – 52 Å – and

very high pitch – 64 Å – are presented in **Figure 5.4F** and **5.4G**, respectively. The very high pitch structure (**Figure 5.4G**) is representative of conformations whose orientations led to artefactual helical parameters. The helical axis might not be the long axis of the protein, instead it was possible to be in a half-rotated orientation (the helix would be wider and more extended, with large pitch and radius but << 360° sweep) or full-rotated orientations (the helix would be disk-like, with << 30 Å pitch, >> 20 Å radius). The confined grid search of the same structures prevented these spurious helices by restricting helical orientations to [50°,70°] and θ in [240°,300°].

5.2.2. Anisotropic Network Model

Using ProDy[181], the anisotropic network model (ANM) modes were calculated using the crystallographic coordinates of MTERF1 Cα atoms (PDB: 3MVA)[25] with a distance weight of 2.5 [182]. A cutoff of 24 Å was selected because it gave the best correlation to B-factors (see **Table 5.1**). To display structures projected along the unit modes, a factor of 50 was used to arbitrarily scale up the displacements. To compare ANM and MD results, the overlap of the eigenvectors obtained from each method was calculated. The root mean square inner product (RMSIP)[183] was used to compare all pairs of ANM and MD eigenvectors. The eigenvectors defining each of the ANM modes and MD principal components (PC) would be parallel if they were identical and orthogonal if completely unrelated; the dot product of parallel vectors is zero if they are orthogonal and one if they are parallel.

Table 5.1. Comparison of ANM cutoff distances and correlation coefficients between the experimental B-factors of MTERF1[25] and the B-factors calculated from ANM.

ANM cutoff distance (Å)	Correlation coefficients between experimental and ANM B-factors
8	0.5778
10	0.6209
12	0.6217
15	0.6229
18	0.6311
21	0.6472
24	0.6619

5.2.3. Model building and parameter preparation

5.2.3.1. Specific MTERF1-DNA complex

Coordinates were obtained from the crystal structure of MTERF1 bound to DNA (PDB code 3MVA[25]). Density was missing for a disordered 19 residue N-terminal segment and the side chains on the first two N-terminal residues of the resolved chain (residues 20 and 21). The role of the disordered segment in binding and recognition was beyond the scope of this work (perhaps involved in signaling, part of the mitochondrial targeting sequence, etc.) and was removed from the model. The sidechains of residues 20 and 21 were added using Amber libraries[184]. 188 water O atoms were resolved and retained in our model building. Molprobity[185] was used to add H atoms to the model and check for N/Q/H flips; none were strongly favored over the original model. The complex was then encapsulated in a 96.3 Å truncated octahedron of explicit water providing a minimum 10 Å distance between any atom of the solute and any edge of the box. Explicit K+ and Cl- ions were added at random positions at least 6 Å from solute atoms and 4 Å from each other to achieve 0.2 M excess KCl concentration with additional K+ ions to neutralize the system. The force field parameters were ff99SB[186] for the protein, parmBSC0[187] for the

DNA, TIP3P[188] for the water, and TIP3P-specific ions[189]. The complete system contained 61042 atoms.

A sample program that automates the above procedure is provided in the **Appendix I.v.**

5.2.3.2. apo MTERF1

The procedure outlined above was repeated, except that the DNA was removed from the initial structure along with the crystallographic water. Initial simulations using an explicit solvent truncated octahedron with a 10 Å solvent buffer were found to be insufficient to enclose the protein during periods of large conformational change (data not shown). Thus a minimum distance of 18 Å between the protein and any edge of the box was used, yielding a final dimension of 111.9 Å and 109.5 degrees. Additional Cl- ions were added to neutralize the system, with 0.2 M excess K+ and Cl-. The same force field parameters were used. The complete system contained 98124 atoms.

5.2.3.3. Search mode MTERF1-DNA complex

The procedure used for the specific MTERF1-DNA complex was used for the search mode MTERF1-DNA complexes, except the initial coordinates were taken from the poses generated by docking (see below). The coordinates of B-DNA were generated using NAB[190] and the 22 base pair cognate sequence.

5.2.3.4. Determining the pitch of B-DNA

Table 5.2 summarizes the helical parameters of B-DNA calculated by Helios (**Chapter 3**). An upper limit of 42 Å B-DNA pitch was identified to be that which MTERF1 could bind in search mode. This value was arrived at by using two approaches. First, the published values of the base pair step parameter rise were reviewed (rise was multiplied by 10 since 10 bp/360°). Second, the helical pitch of B-DNA of MD simulations was measured.

Table 5.2. Summary of helical parameters calculated by our method for ideal B-DNA.

Atoms	Count	Residual (Å)	Radius (Å)	Pitch (Å)	Rise (Å)	Twist (°)
C1' (W)	22	0.023	5.85	33.80	3.38	36.00
C1' (C)	22	0.015	5.85	33.81	3.38	36.00

(W) indicates the Watson strand, (C) the Crick strand in the ABC[191] definition. The residual in our fitting procedure measures the deviation of the projected atomic coordinates from a perfect circle in units of Å. Our sweep parameter is analogous to twist (i.e. the sum of the twists for each step). An ideal DNA geometry with a base pair step rise of 3.38 Å and a step twist of 36.0 degrees was analyzed. As expected, the method reproduced the helical rise parameters of B-DNA built using NAB[192] (36° twist, 10 base pair per revolution multiplied by 3.38 Å = 33.8 Å). The method also reproduced twist. The radius of the major groove was defined as the radius of the helix traced by center of the C1' atoms.

A maximum value of rise in the central dinucleotide of CGCA/TGCG is 4.5 Å in nucleosome core particle crystal structures[193], which is similar to the average rise for all DNA sequences plus 3 standard deviations (4.4 Å) found from MD simulations[14] of the 136 tetranucleotide sequences. Pitch for B-DNA, which contains 10 nucleotides per helical turn, is then 44 Å (used to define the horizontal lines in **Figure 5.5** below). In a third example from the literature, Olson et al. reported average rise values of 3.36 Å and standard deviations of 0.25 Å[194]

of protein-DNA complexes (broken base pairs were omitted). The reported pitch for B-DNA was 41.1 Å. Together, the literature supports the assumptions that a B-DNA molecule that a protein would randomly encounter in solution would likely posses a ~42 Å pitch, or less.

Four independent MD simulations of B-DNA were performed to determine the range of helical pitch sampled under the same conditions that would be used in the nonspecific complex (except without the protein). The DNA was built, equilibrated, and production performed exactly as the search mode complexes. The helical parameters of the B-DNA in the MD trajectories were then analyzed using Helios (**Figure 5.5A and 5.5B**). Overall, the distributions of DNA pitch agree with the above conclusions. The average pitch was observed to be 35.8 Å for the HS and 33.3 Å for the LS, with 2.31 Å standard deviation in the HS and 3.35 Å standard deviation in the LS. Thus the HS accesses pitch with 42.7 Å pitch (average plus 3σ) and the LS, 43.4 Å.

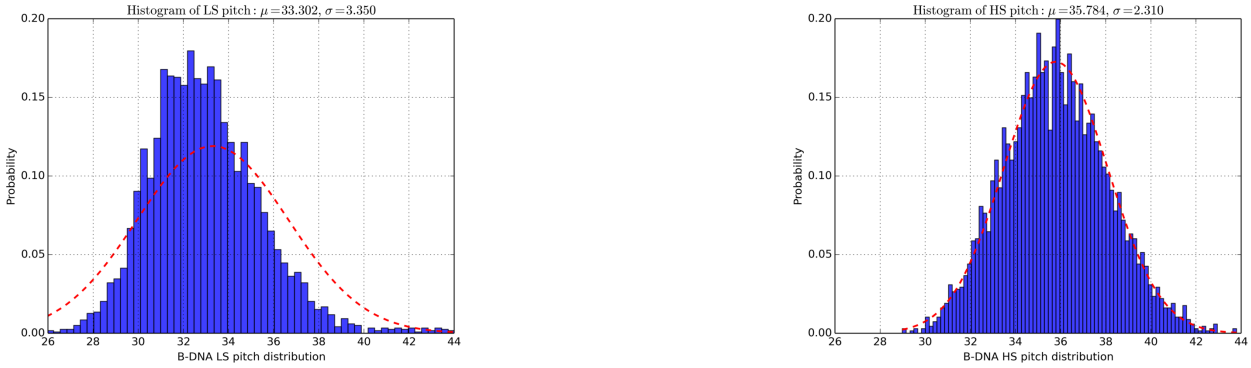


Figure 5.5. Control simulations of the 22 bp target sequence in a B-DNA geometry. (**A**) DNA light strand (LS) and (**B**) heavy strand (HS). Four independent 1 μs MD simulations were performed. Histogram used 100 bins.

5.2.3.5. Docking protocol and generating a nonspecific complex

apo MTERF1 structures were considered sufficiently complementary to B-DNA (**Table 5.2**) if superhelical pitch was less than 42 Å and radius between 9 Å and 16 Å (**Figure 5.5 and Section 5.2.3.4**). The 3 lowest pitch structures of apo MTERF1 from each of the eight independent simulations were used for docking; 3 simulations produced no low pitch structures. 15 apo MTERF1 structures were docked to B-DNA using DOT2.0[195], which has been previously been shown to be suitable for protein-DNA docking[196]. Briefly, REDUCE[197] parameters for heavy and polar hydrogen protein and DNA atoms were used, electrostatic potentials were calculated with APBS[198] and 0.200 M ionic strength, and electrostatic clamping was used to flatten pathological energies[195]. van der Waals energies were estimated by counting DNA atoms that were within an interaction region coating the protein; the inner surface of the region was defined by the MSMS[199] molecular surface with a 1.4 Å probe radius, and the outer surface was defined by expanding the protein van der Waals atomic radii by 3.0 Å[195]. Desolvation energies were not included in the calculations. 54,000 orientations of apo MTERF1 and B-DNA were evaluated for each of the low pitch protein structures. The protocol was validated by docking MTERF1 and DNA from crystallography, which reproduced the experimental complex (RMSD < 3 Å for the 7 highest ranked structures).

5.2.3.6. A geometric scoring function

The best-ranked (lowest energy) DOT2.0 poses were filtered by how well the protein tracked the major groove. The 30 best poses from each of the 15 docking calculations were filtered by how well MTERF1 tracked the major groove, with the correct polarity. First, major groove sites were defined geometrically (**Figure 5.6A**). For B-DNA, a vector connecting P

atoms on opposite strands separated by 5 nucleotides (the P of nucleotide 1 on one strand and the P of nucleotide 6 on the other strand) is nearly parallel to the helical axis with a length of 21 Å. The midpoint of the vector was defined as a major groove site (**Figure 5.6B**). The midpoint was ~10 Å from the P atoms and ~8 Å from nucleobase functional groups The distance between superhelical Cα atoms and major groove sites was measured.

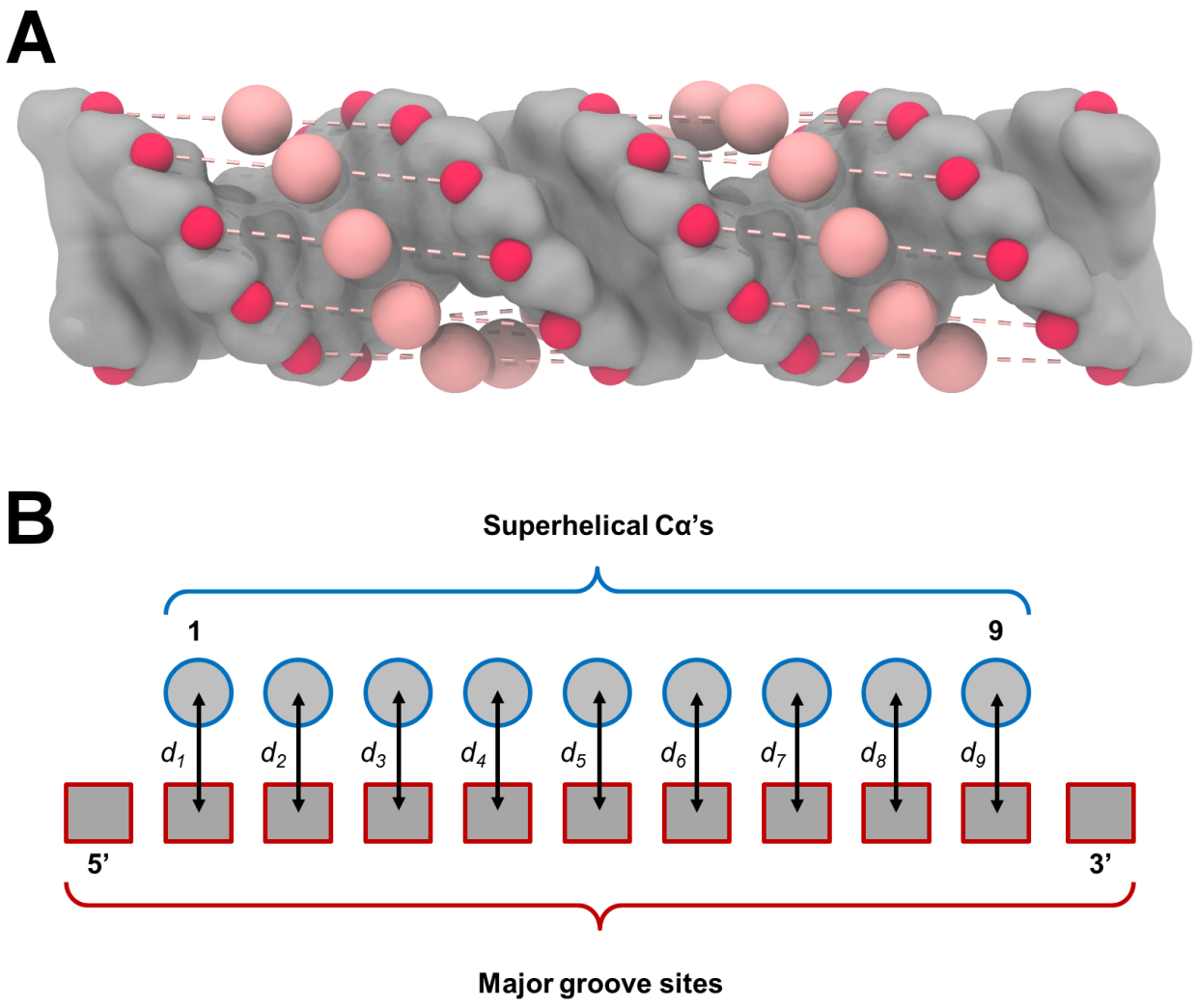


Figure 5.6. Method to geometrically measure how well MTERF1 tracks a major groove. (**A**) Major groove sites are the midpoints (pink spheres) between successive P atoms (dark red spheres) on opposite strands offset in sequence by 5, shown as pink dotted lines. (**B**) Major groove site-superhelical residue pairing scheme.

Each of the distances (d_1 through d_9 in **Figure 5.6**) was expected to be between 6 and 10 Å from a major groove site, based on inspection of the N- and C-site in the crystal structure and analysis of simulations with the specific MTERF1-DNA complex (~7 Å, see **Figure 5.7**). Poses were classified as nonspecific complexes when MTERF1 tracked the major groove and consecutive superhelical Cα atoms resided in consecutive major groove sites.

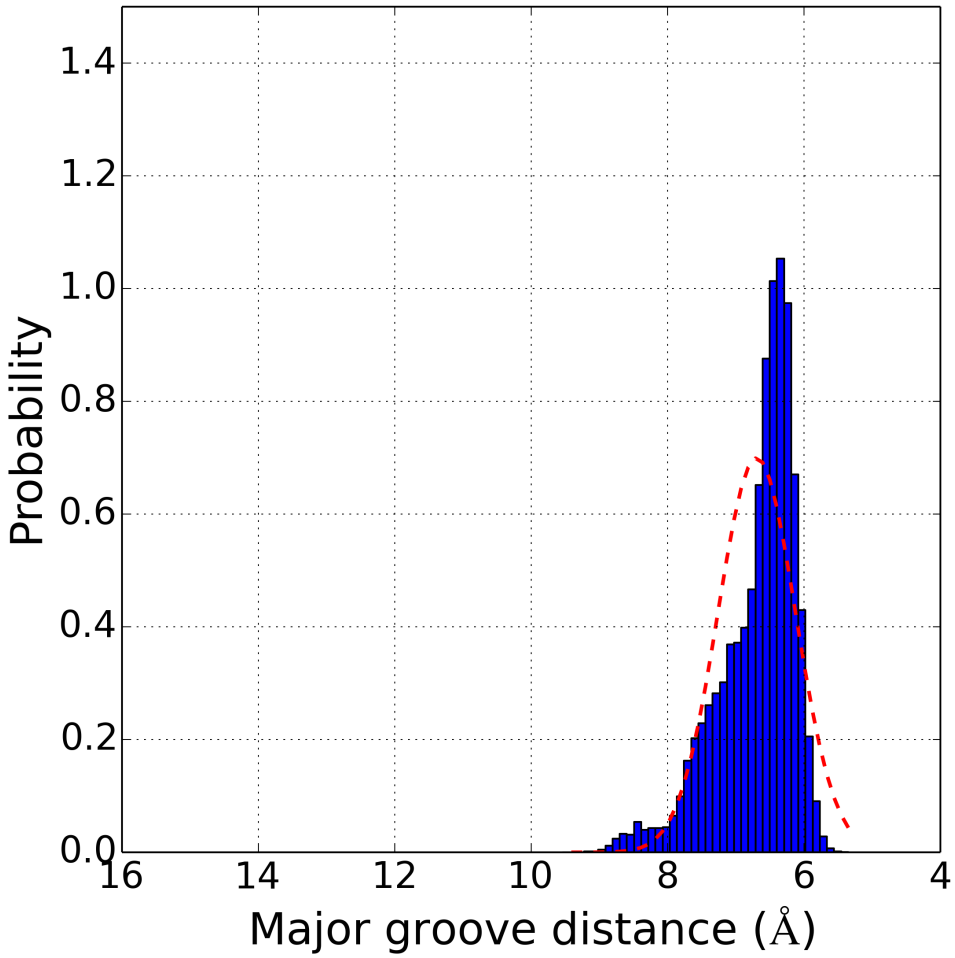


Figure 5.7. The expected value of a major groove-binding distance was established for MTERF1. The probability density was determined for the average major groove distance between each superhelical residue and its nearest major groove site for control simulations of the specific MTERF1-DNA complex. 40 bins were used (the integral of the bins is one).

When the docked pose visually appeared to be tracking the major groove (**Figure 5.8**), the average distance was < 11 Å. The pairing sequence of Cα superhelical atoms and major groove sites was set by the pair with the smallest separation for a particular pose. For example, if the shortest distance between a superhelical Cα and a major groove site was between the fourth Cα and the fifth major groove site (a pair), then the first Cα was automatically paired with the second major groove site, the second Cα paired with the third major groove site, and so on. The pairing rule caused large Cα-major groove site distances (>25 Å average) when MTERF1 was docked in the reverse polarity, or if MTERF1 crisscrossed the major and minor grooves (i.e. cross-threaded). Thus, the average distance reported how well MTERF1 tracked the major groove.

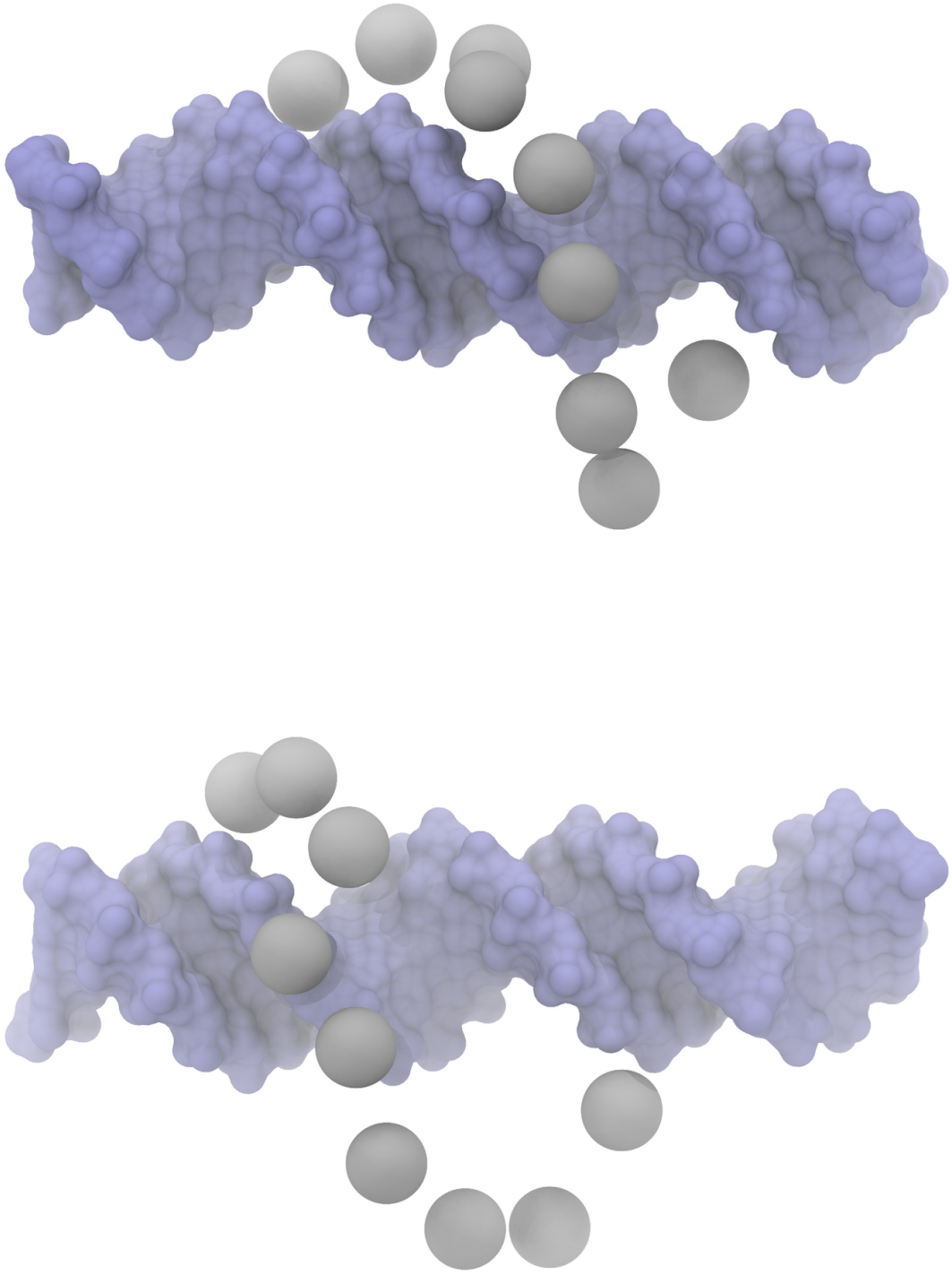


Figure 5.8. Quantifying the position of MTERF1 in the major groove of B-DNA. The upper and lower poses correspond to groove tracking distances of 9.55 Å and 14.03 Å, respectively. In the upper pose, the first superhelical residue was paired with the tenth major groove site, the second superhelical residue the ninth site and so on; the individual distances (d_1 through d_9) were 13.5 Å, 11.1 Å, 6.0 Å, 8.3 Å, 5.4 Å, 9.2 Å, 12.1 Å, 14.5 Å and 5.9 Å. For the lower pose, the first superhelical residue was paired with the fourteenth major groove site, the second superhelical residue with the thirteenth major groove site and so on; the individual distances were 11.2 Å, 11.6 Å, 10.6 Å, 13.5 Å, 13.6 Å, 17.0 Å, 16.8 Å, 15.9 Å and 16.3 Å (d_1 through d_9).

5.2.4. MD equilibration and production

5.2.4.1. MTERF1-DNA specific complex

The multi-stage equilibration procedure is outlined in **Table 5.3**. A one fs time step for dynamics and an 8.0 Å non-bonded direct space interaction cutoff were used, with PME [200] to calculate long-range electrostatic interactions across the periodic lattice containing the simulation cell. Initial minimization used the crystallographic structure as the reference, and subsequent stages used the final structure from the previous stage. Unless otherwise noted, the same force constant and ensemble were used in subsequent steps. Initially, all atoms added to the crystal structure were minimized while all atoms resolved by crystallography except crystallographic water (group A in **Table 5.3**) were restrained with a force constant of 100.0 kcal/mol/Å². The system was then heated in NVT from 100 K to 300 K linearly over 100 ps. Next, the density of the system was equilibrated at 300 K for 100 ps in NPT. With temperature and pressure equilibrated, MD continued at 300 K for 250 ps and the restraint force constant was decreased 10-fold. Since protein backbone atoms are often less susceptible to crystal packing forces, the restraint group was transitioned from all crystallographic protein atoms to only the protein and DNA backbone atoms in the subsequent stages (group B in **Table 5.3**). The system was minimized using a restraint force constant of 10.0 kcal/mol/Å² and otherwise identical conditions as the initial minimization. Stage 6 was 100 ps of NPT dynamics at 300 K. The next two stages were identical to stage 6, except the restraint force constant was decreased to 1.0 kcal/mol/Å² (stage 7) then 0.1 kcal/mol/Å² (stage 8). The ninth stage of equilibration was again identical to stage 6, except positional restraints were completely removed. Stage 9 was 1.25 ns of unrestrained NPT. Thereafter, the NVT ensemble was used for unrestrained production.

Independent trajectories constituted simply initializing dynamics with velocities drawn from a Maxwell-Boltzmann distribution in stage 2.

Table 5.3. Equilibration procedure for explicit solvent MD simulations.

Stage	Ref	EOM	Steps (x 10^3)	Temp (K)	Ensemble	Group	Force constant (kcal/molA^2)
1	xtal	min	10	-	-	A	100
2	1	MD	100	100/300	NVT	A	100
3	2	MD	100	300	NPT	A	100
4	3	MD	250	300	NPT	A	10
5	4	min	10	-	-	B	10
6	5	MD	100	300	NPT	B	10
7	6	MD	100	300	NPT	B	1
8	7	MD	100	300	NPT	B	0.1
9	-	MD	2500	300	NPT	-	0

Ref, reference coordinates. **EOM**, equation of motion: min, minimization; MD, molecular dynamics. **Ensemble**: NPT and NVT used a weak temperature coupling thermostat with isotropic position scaling. **Group**, atoms restrained to reference structure (**Ref**): **A** for apo MTERF1, all protein heavy atoms except the sidechains of residues 1 and 2, **B** for apo MTERF1, all protein backbone atoms – Cα, N, and C; **A** for holo MTERF1, all protein and DNA heavy atoms except, as in apo MTERF1, the side chains of residues 1 and 2, **B** for holo MTERF1, all protein and DNA backbone atoms – C1', C2', C3', C4', O4', C5', O3', O5', OP1, OP2, P; **A** for search mode MTERF1, superhelical Cα atoms and all DNA heavy atoms, **B** for search mode MTERF1, only DNA backbone atoms (MTERF1 fully unrestrained). **Force constant**, harmonic force constant for Cartesian restraints. For search mode, force constants were $1/10^{th}$ of those illustrated above, except for stage 8, which had no force applied. The Berendsen thermostat[201] was used for all stages of MD including production. Stage 2 and 3 used bath coupling constants of 0.1 ps; all subsequent stages of MD including production used 0.5 ps coupling constants.

5.2.4.2. apo MTERF1

Equilibrating apo MTERF1 followed the procedure above, except without the DNA present. Stage 9 was 2.25 ns of unrestrained NPT.

5.2.4.3. MTERF1-DNA search complex

The equilibration procedure above was used with only the following modifications. For each stage, one-tenth the restraint force constants were used because the nonspecific complexes generated from docking were expected to be less precise than a high resolution crystal structure. Also, the DOT2.0[195] energy function may have generated globally stable poses with locally unstable contacts that require flexible restraints to relax. The DNA restraint group was the same as the specific complex. In stages 1 through 4, only the Cα atoms of the superhelical residues were restrained. MTERF1 was not restrained during equilibration in stages 5 through 7. Stage 8, the final stage, was 250 ps of unrestrained NPT MD. Thereafter, production dynamics used the NVT ensemble. During production, each of the 12 search mode simulations switched to a 4 fs time step after ~3 μs of MD, since the H-mass repartitioning algorithm[202] in Amber became available.

5.3. Results

Exploring how TFs search DNA has been the focus of extensive experimental and theoretical research[203-205] yet many basic questions remain unanswered owing to the lack of structural data for unaltered nonspecific protein-DNA complexes. MD simulations, both coarse grained[206-207] and atomistic[208-209], have been able to provide some of the needed structural insight into the transient states (~μs) involved in search mode. Coarse-grained simulations lack atomistic resolution and internal flexibility to pinpoint specific interactions or DNA distortions that will be

needed for a high-resolution mechanism of search and recognition, and previous atomistic MD simulation studies have relied on biasing potentials to generate search mode models.

5.3.1. MTERF1 from crystallography clashes with B-DNA

5.3.1.1. Structure-based docking

As described above, we exclude models in which MTERF1 binds transiently predeformed DNA because its population would be much too low for efficient recognition. Thus we tested the next simplest model, in which MTERF1 in the recognition conformation binds to B-DNA (Model A in **Figure 5.3**). Since the N- and C-sites of DNA in the crystal structure were essentially B-form (**Figure 5.2**), we aligned the target cognate sequence in a B-form conformation to either site to generate potential search mode models. In contrast to MTERF1 and unwound DNA, large steric clashes occur between the molecular surfaces of MTERF1 and B-DNA (**Figure 5.9A**,**B**).

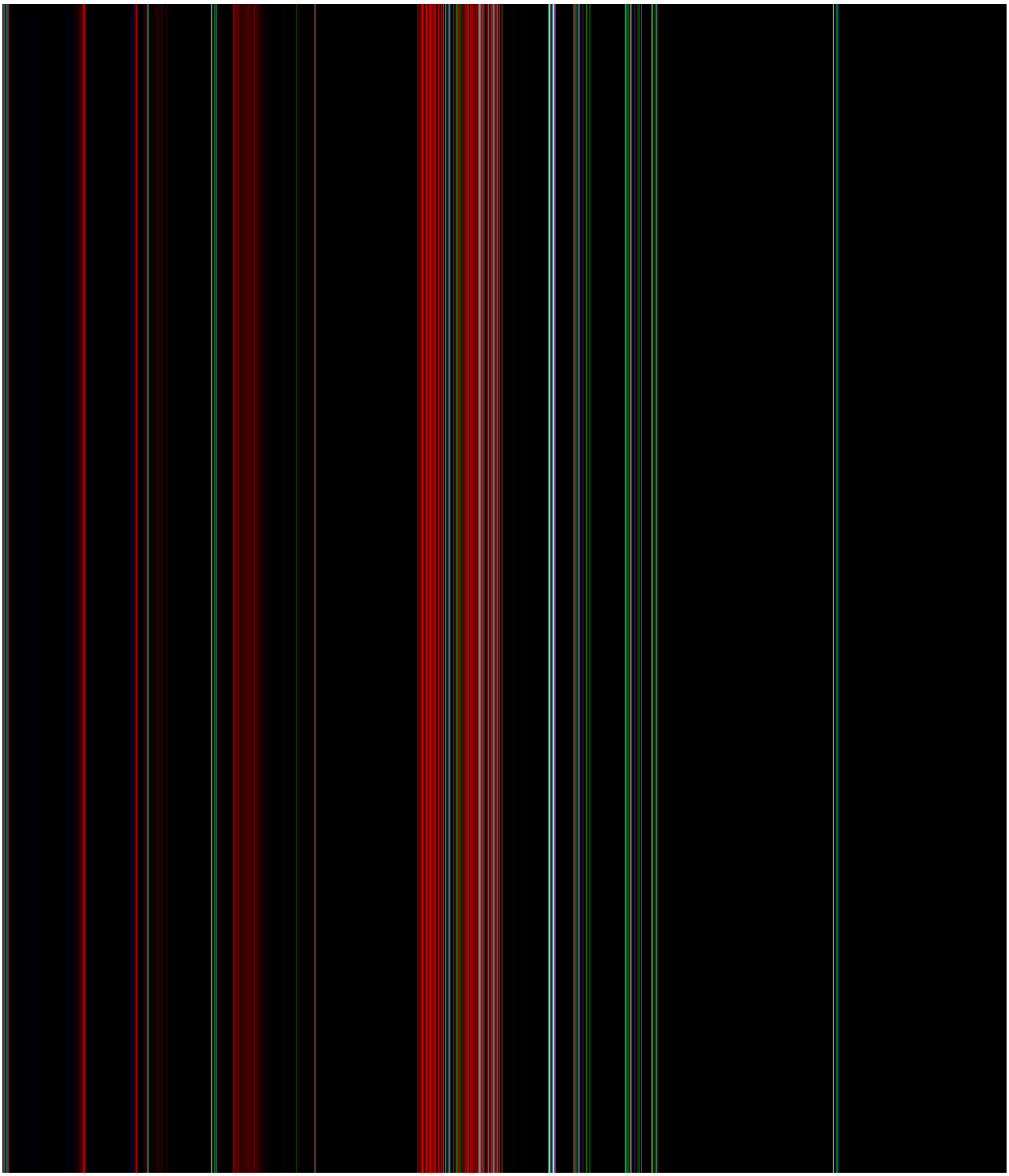


Figure 5.9. MTERF1 in the recognition mode conformation is too unwound to track the major groove of B-DNA. Surfaces were sliced to show incompatibility: the blue clipping plane appears purple where the DNA penetrates the protein. (**A**) Only minor steric clashes are present in the crystal structure of the recognition complex with unwound DNA. (**B**) Aligning the C-site P atoms of B-DNA to the corresponding region in the crystal led to large steric clashed between the N-site and the N-terminal domain of MTERF1. Alignment of the N-site resulted in similar clashes in the C-site.

5.3.1.2. Docking with DOT2.0

Using DOT2.0 to dock MTERF1 and B-DNA to find a more optimal threading of the two molecules, B-DNA passed through the binding cleft of the protein only when 10 steric clashes were permitted (**Figure 5.10**). All poses in which B-DNA was docked into the binding cleft of MTERF1 were tested for stability using our fully atomistic MD procedure (see **Section 5.2.4**), and found to be energetically unstable (>10^8 kcal/mol). Some poses appear reasonable (**Figure 5.10A** and **B**) and indeed MD energies were stable, though not structurally stable (RMSD > 7 Å after only 50 ns). Closer inspection of these poses revealed that the C-site of the DNA would clash if extended. To show these poses were not simply artifacts of using short DNA, the DNA in all the poses were lengthened and the complexes were subjected to the same equilibration procedure. As expected, the docked poses (with extended DNA) were not stable (>10^8 kcal/mol) due to van der Waals clashes. Overall, the structures were very high in energy and attempts to relieve the clashes using minimization and MD failed. The protein was unable to continuously track the major groove of B-DNA because the superhelical architecture of MTERF1 did not match that of that of the DNA. The result suggests that Model A might not be a reasonable paradigm for MTERF1 scanning DNA.

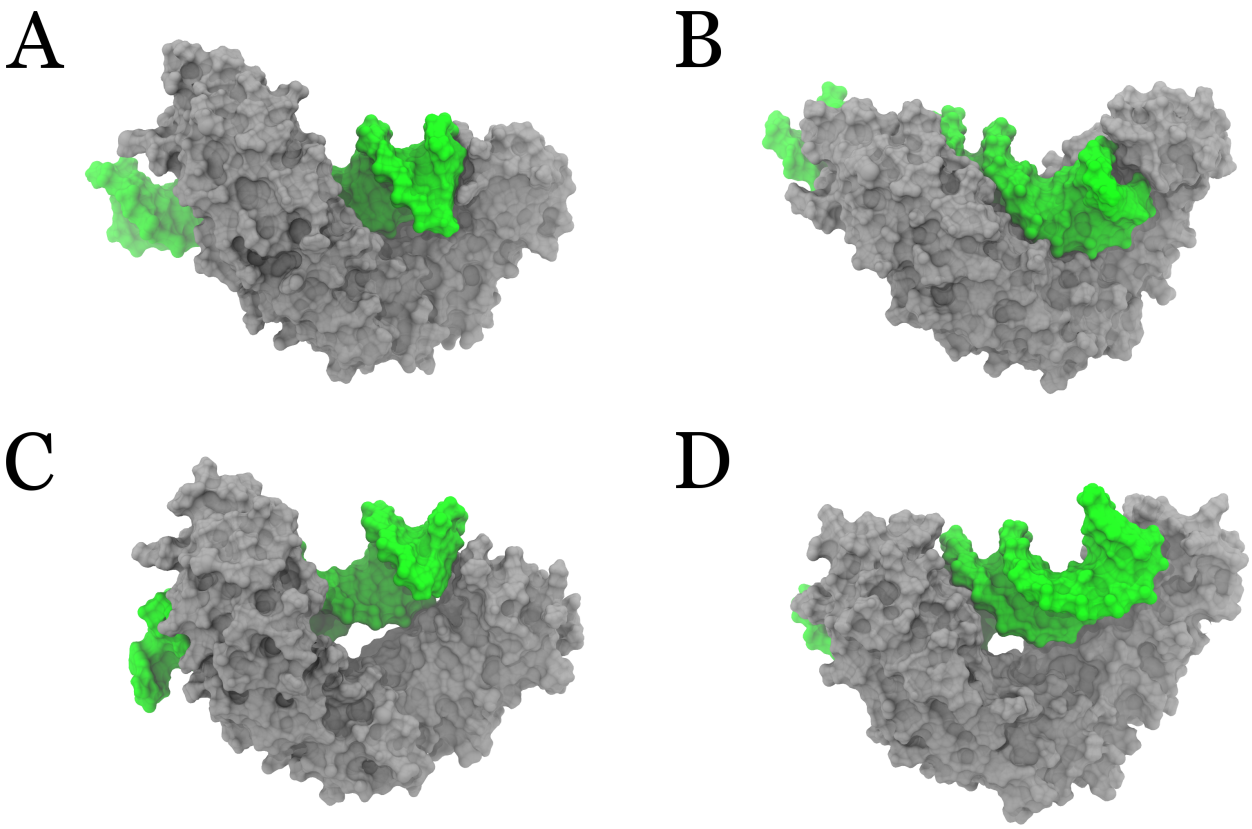


Figure 5.10. Docking B-DNA to MTERF1 from the recognition mode structure fails to produce poses in which the protein tracks the major groove. (**A**) One of two DOT2.0 docked poses in which MD equilibration energy did not result in high energy (>10^9 kcal/mol) using the exact procedure used to dock low-pitch apo MTERF1 to B-DNA and to dock MTERF1 from recognition mode to the corresponding unwound DNA from the crystal structure. (**B**) The second pose. (**C**) After 75 ns of unrestrained MD of the pose from (**A**), MTERF1 dissociates from the DNA (all atom RMSD > 7 Å) (**D**) After 50 ns of unrestrained MD of the pose from (**B**), MTERF1 dissociates from the DNA (all atom RMSD > 8 Å).

5.3.2. Intrinsic axial and radial motions of MTERF1

5.3.2.1. ANM reveals axial and radial motions in MTERF1

To track the major groove of B-DNA, MTERF1 must adopt an alternate conformation in search mode, corresponding to Model B in **Figure 5.3**. As the DNA helix in the recognition mode is unwound, we speculated that the MTERF1 superhelix in the recognition mode might

also be unwound (higher pitch) relative to the search mode. Thus we tested whether the MTERF1 topology possesses intrinsic motions that might lead to lower pitch conformations that better track a B-DNA major groove. To explore this hypothesis, the mechanical modes of the protein were calculated using an anisotropic network model (ANM). The results support the hypothesis. The global (lowest frequency) motions correspond to dynamics of the superhelical topology. To visualize superhelical motions, conformations were projected along the modes. Relative to the long axis of the protein, mode 1 was an axial motion and mode 2 was a radial motion (**Figure 5.11**). Importantly, dynamics along mode 1 might lead to a low pitch ensemble more compatible with tracking a B-DNA major groove.

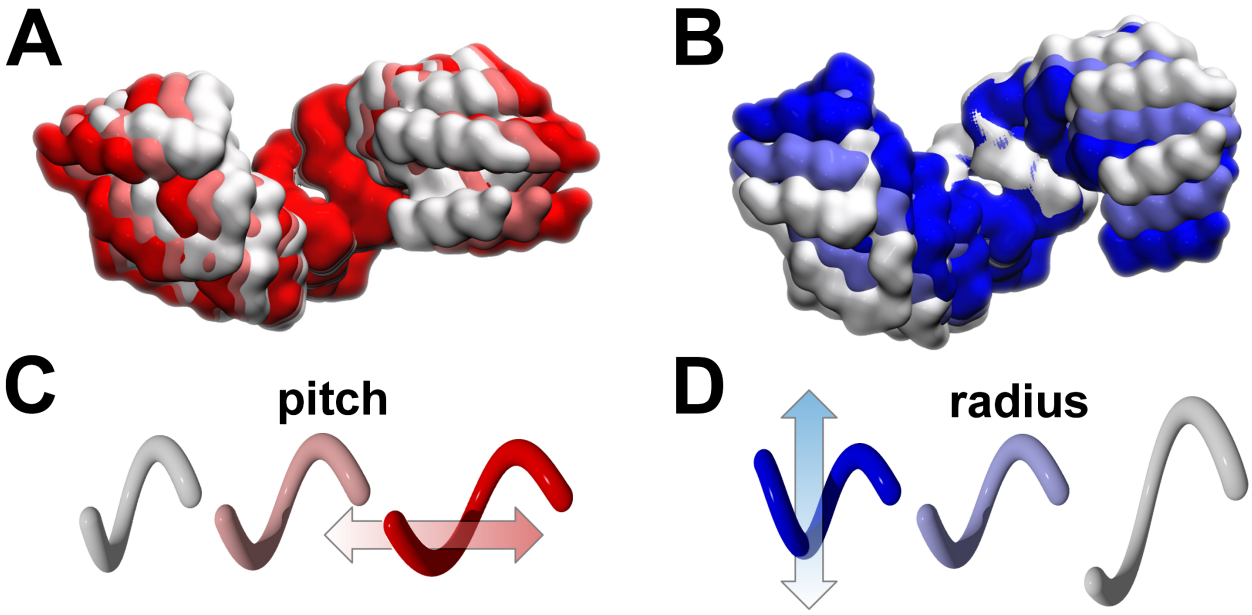


Figure 5.11. Lowest frequency modes of MTERF1 adapt superhelical pitch and radius and may permit binding to B-DNA. (**A**) The lowest frequency ANM mode of MTERF1 is an axial motion, white and red surfaces of Cα atoms denote positive and negative displacements, respectively. (**B**) The next lowest frequency ANM mode of MTERF1 is a radial motion, white and blue surfaces of Cα atoms are positive and negative displacements, respectively. (**C**) and (**D**) are cartoons of motions above, pitch and radius, respectively.

ANM cannot determine the magnitude and the sense (positive or negative) of the displacement. However, the direction of ANM modes can overlap with those of atomistic MD[210], suggesting ANM and MD are complementary methods for characterizing protein dynamics. Similar motions predicted by methods with different limitations would suggest that the model for the dynamics is less likely to be an artifact. Furthermore, atomistic MD may give more detailed insight not only into the types of dynamics encoded in the MTERF1 topology, but also quantify the ranges of pitch and radius that are sampled at ambient temperature, and whether these are compatible with binding a B-DNA duplex.

5.3.2.2. Control MD simulations of MTERF1 in recognition mode

To establish a baseline for analyzing the apo MTERF1 dynamics, four independent 1.5 μs control simulations of the MTERF1-DNA specific complex were performed. We expected small structural fluctuations around an average conformation similar to the crystal structure. To quantify similarity, the RMSD between the MD snapshots and the equilibrated crystal structure was calculated using cpptraj[211]. The evolution of RMSD in the control simulations (**Figure 5.12A,B)** shows that the conformation of MTERF1 throughout the simulations of the specific complex remained similar to that of the reference.

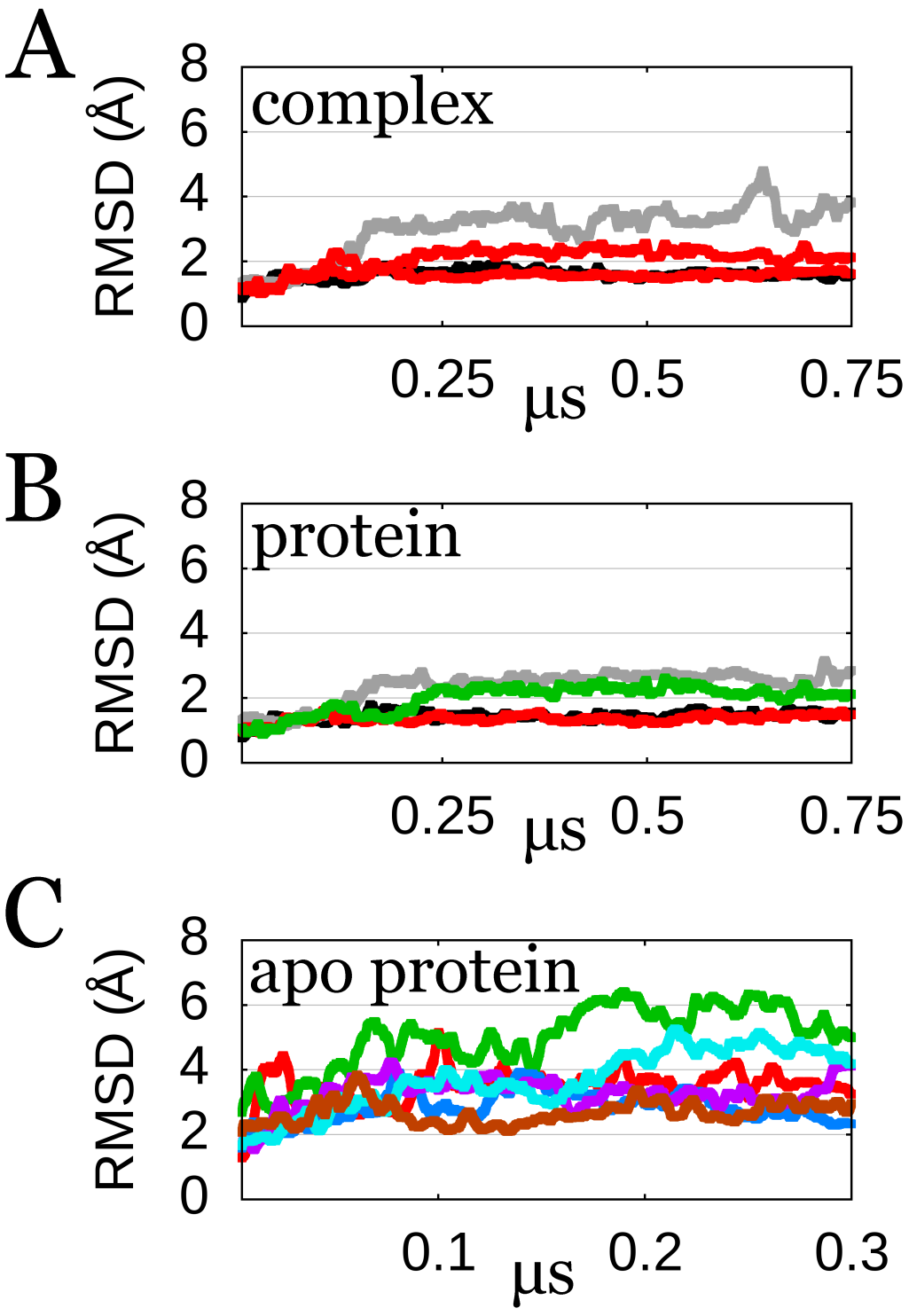


Figure 5.12. RMSD analysis of apo MTERF1 and holo MTERF1 (specific complex) unrestrained MD simulations. (**A**) RMSD of the specific MTERF1-DNA complex (crystal structure) protein backbone, excluding the first *mterf* motif and the C-tail, and the DNA C1' and P atoms, excluding the 3 terminal base pairs at each end. (**B**) RMSD of only the protein in the specific complex using the same atoms as in (**A**). (**C**) RMSD of MTERF1 in the unbound protein simulations using the same atoms as in (**A**) and (**B**).

5.3.2.3. Apo MTERF1 MD simulations

An ensemble of apo MTERF1 structures was generated by performing eight independent 0.3 μs MD simulations. In contrast to the control simulations of the specific complex, RMSD analysis of the apo MTERF1 simulations indicated that the protein undergoes significant conformational change with respect to the reference, in the absence of DNA (**Figure 5.12C**). To gain further insight into the nature of the changes in apo protein structure, the motions sampled in MD were compared to the global modes obtained from the ANM calculations (**Section 5.3.2.1**).

5.3.3. MD and ANM exhibit similar low frequency motions

To measure the similarity of ANM and MD motions, principal component analysis (PCA) was performed using the complete MD ensemble. The 10 eigenvectors with the lowest frequency for the ANM and the MD simulations show high similarity, indicated by an RMSIP [183] of 0.77 (**Table 5.4**). The similarity indicates that the global dynamics sampled in the atomistic MD simulations also correspond to changes in the superhelical pitch and radius of MTERF1, as was suggested by ANM (**Figure 5.11**). Observation of similar dynamics in the two different computational approaches also suggests that the results are less likely to be an artifact of a specific model. We next analyzed the range of fluctuations in these measures to determine whether these dynamics could result in structural excursions that would remodel the apo MTERF1 binding site to accommodate B-DNA without steric clashes.

Table 5.4. ANM and MD eigenvector RMSIP similarity analysis.

		MD (PC)			
		1 to 2	**1 to 3**	**1 to 4**	**1 to 10**
ANM (mode)	**1 to 2**	0.776			
	1 to 3		0.879		
	1 to 4			0.856	
	1 to 10				0.769

Root mean square inner product (RMSIP)[212] of all top ten eigenvectors (1 to 10), the top four eigenvectors (1 to 4), the top three (1 to 3), and the top two (1 to 2). RMSIP provides a global similarity comparison of eigenvector overlaps, accounting for the possibility that corresponding MD and ANM eigenvectors are not in the same order.

5.3.4. Quantifying MTERF1 superhelical motions using a general gauge of helical parameters

We hypothesized above that compatibility of MTERF1 and B-DNA would encompass similarity in the global helical pitch. The challenge was that no gauge of global helical pitch and radius exists for proteins, while DNA is naturally defined by helical coordinates. The pitch of DNA depends on the rise between each base pair step – the displacement along the helical axis – and the twist of the step - the rotation of the base pair plane about the helical axis. These parameters depend on a well-defined helical frame, which is well established for DNA[83] but has not been described for proteins. We thus developed a new approach of defining a helical frame to quantify MTERF1-DNA complementarity. The helical architecture of MTERF1 arises from its modular architecture and we cast our new helical reference frame with the assumption that a helical axis exists for MTERF1 and, importantly, that the axis is normal to the plane onto which the protein projects the best circle (see **Section 5.2.1** and **Chapter 3**). A set of proline residues (with two exceptions, see **Section 5.2.1**) were identified that occupy comparable positions within each motif, referred to as the superhelical residues (**Figure 5.1A**). As the superhelical residues

track the major groove, the radius of the resulting helix defined with these residues is expected to closely match that of the bound DNA.

5.3.5. apo MTERF1 spontaneously adopts structures with the same pitch as an average B-DNA

We carried out helical analysis on the MD simulations for apo and holo (specific complex) MTERF1. The superhelical dynamics of apo MTERF1 are strikingly different from holo MTERF1, with the ensemble sampling a much broader range of pitch and radius for the apo protein (helical parameters along with representative structures are shown in **Figures 5.13A** and **5.13B**). Comparing the standard deviations of the superhelical parameters for the two ensembles indicates that apo MTERF1 superhelical radius and pitch are roughly one order of magnitude more diverse than holo-sp (**Figure 5.14**). Interestingly, the broad range of superhelical radius values sampled by the apo protein has a lower bound of ~ 7 Å, the radius of a B-DNA major groove (**Table 5.2**), suggesting that although the type of motion is encoded in the topology, the protein may lack selective pressure to increase flexibility beyond that required for function. The ensemble sampled by the apo MTERF1 simulations also exhibits structures with superhelical pitch similar to that of B-DNA (**Figure 5.13B**) while, as expected, MTERF1 in the control simulation remains much higher than B-DNA (**Figure 5.13A**).

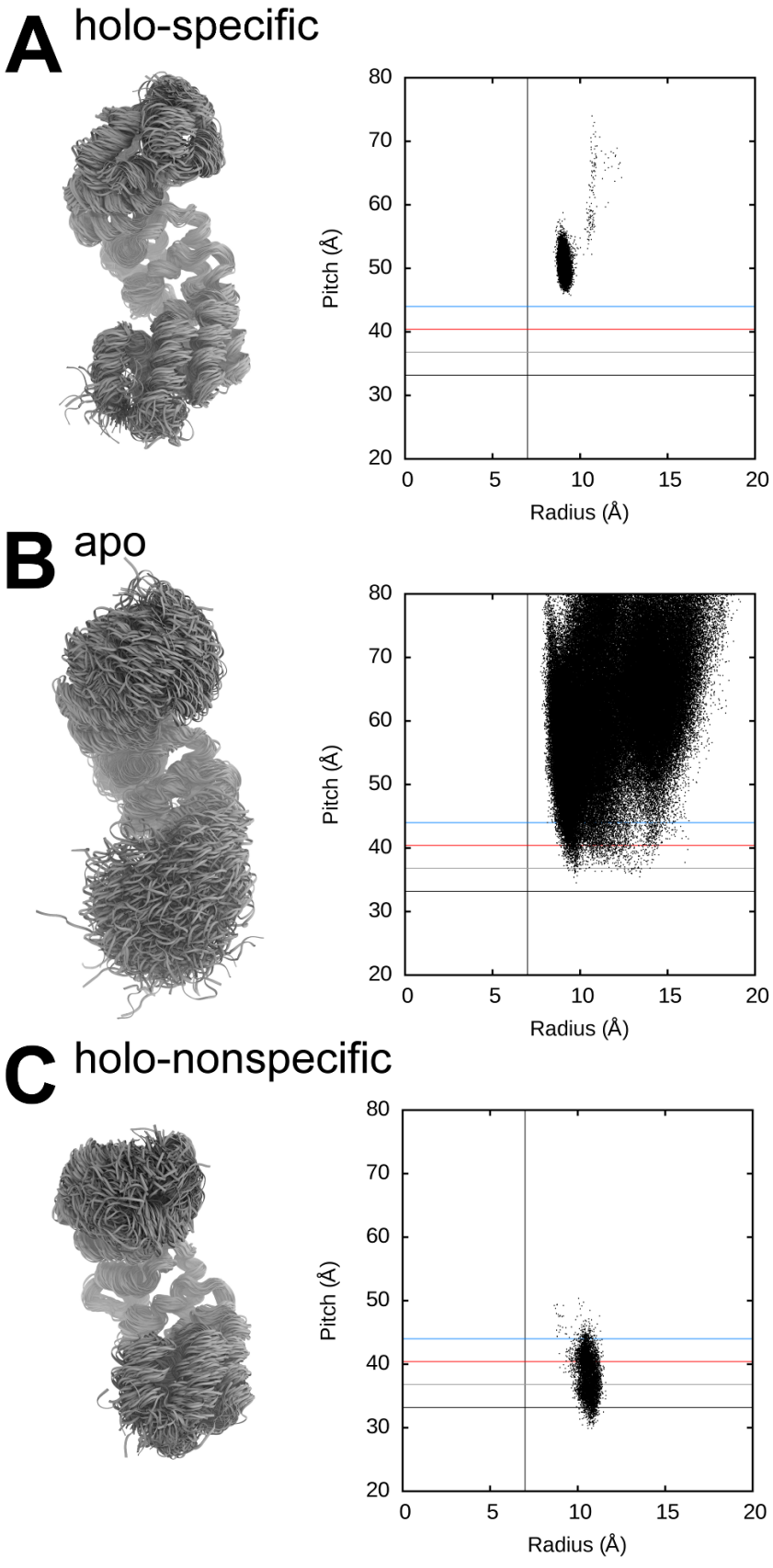

A holo-specific
B apo
C holo-nonspecific
Pitch (Å)
Radius (Å)
80
70
60
50
40
30
20
0
5
10
15
20

Figure 5.13. A switch in the MTERF1 superhelical architecture. (**A**) In recognition mode, unbiased MD simulations show that MTERF1 populates a high pitch state consistent being bound to unwound (high pitch) DNA. The DNA is not shown for clarity. To show the expected range of B-DNA pitch, horizontal lines mark the average structure of B-DNA (black) plus one, two, and three standard deviations (grey, red, blue, respectively). A vertical guide is placed at 7 Å to represent B-DNA radius. (**B**) In the absence of DNA, apo MTERF1 samples a huge range of superhelical conformations, extending into the range compatible with B-DNA. (**C**) In search mode, the superhelical dynamics of MTERF1 are suppressed by B-DNA with a much narrower distribution of both pitch and radius. Compared with holo-specific, the small increase in radius of holo-nonspecific is likely caused by the decrease in pitch. Snapshots of MTERF1 were selected evenly from concatenated trajectories of the respective ensembles; the N-terminus is toward the top.

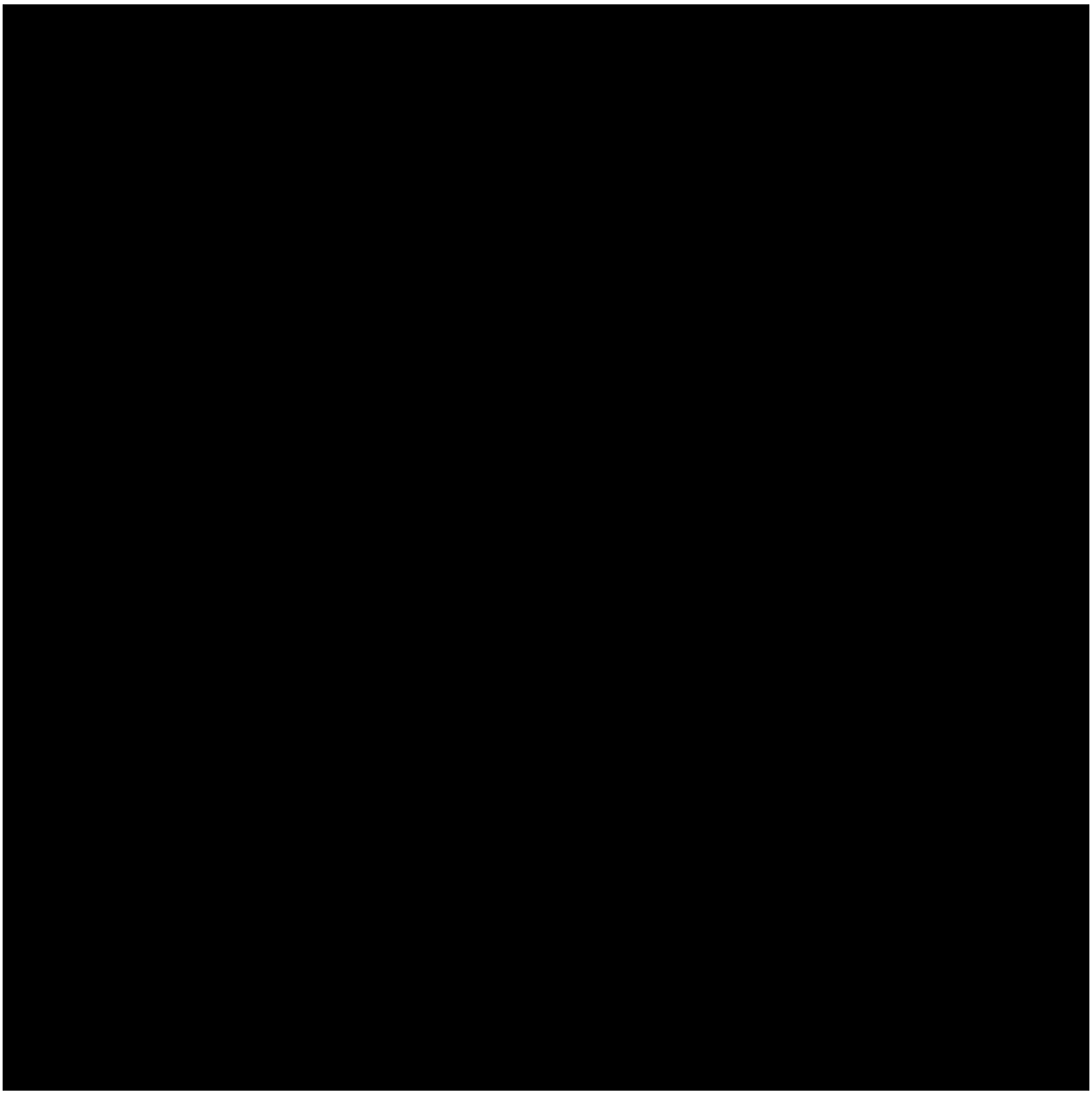


Figure 5.14. Histograms of helical parameters of holo and apo MTERF1 MD ensembles. Probabilities of holo MTERF1 pitch (top left), holo MTERF1 radius (top right), apo MTERF1 pitch (bottom left), and apo MTERF1 radius (bottom right).

5.3.6. Are low pitch apo MTERF1 structures compatible with a B-DNA major groove?

Although the helical analysis suggests that apo MTERF1 spontaneously adopts structures with superhelical pitch and radius compatible with B-DNA, these global measures of structure cannot confirm that the structure complementarity is sufficient to avoid the steric clashes that were obtained when the crystal conformation was docked to B-DNA (**Figure 5.3**). We therefore repeated the docking procedure using low pitch apo MTERF1 structures along with canonical B-DNA, and subsequently performed MD to relax the docked complexes and determine if they provide reasonable and stable models of the nonspecific complex. As a control, we also separated and then re-docked the DNA and protein structures from the crystal structure of the recognition complex; this control successfully recapitulated the crystal structure and MD simulations of the resulting complexes were stable. We thus proceeded with docking the low pitch apo structures to B-DNA.

5.3.7. Docking and scoring low pitch apo MTERF1 and B-DNA

To obtain a diverse set of docking poses mimicking productive nonspecific complexes, we docked to B-DNA the lowest-pitch protein structures from the five apo MTERF1 simulations that sampled conformations with superhelical pitch < 42 Å. This pitch cutoff was selected since it represents a statistically significant population of B-DNA structures[14, 18, 193] and thus is likely compatible with the major groove of B-DNA (**Section 5.2.3.4**.). Apo structures in this range also have radii larger than 9 Å (**Figure 5.13B**), suggesting that inward facing sidechains should fit over the major groove of B-DNA (5.7 Å **Table 5.2**).

We independently docked 14 low pitch apo MTERF1 structures to B-DNA. In each of the 14 calculations, the energy of 54,000 poses was evaluated using the DOT2.0 energy function (see Methods) and only the 30 lowest energy poses were retained. Next, productive poses in which MTERF1 tracked the major groove were filtered from poses that did not by measuring the distance between superhelical residues and major groove sites (see **Section 5.2.3.6.**). We considered acceptable values to range from a lower limit of ~7 Å (obtained from the specific complex **Figure 5.7**) up to 11 Å, beyond which poses did not visually appear to closely track the major groove (see **Figure 5.8** for examples). Thirteen poses fell within this range, and after culling one due to a steric clash, 12 poses were retained for further analysis (**Figure 5.15**).

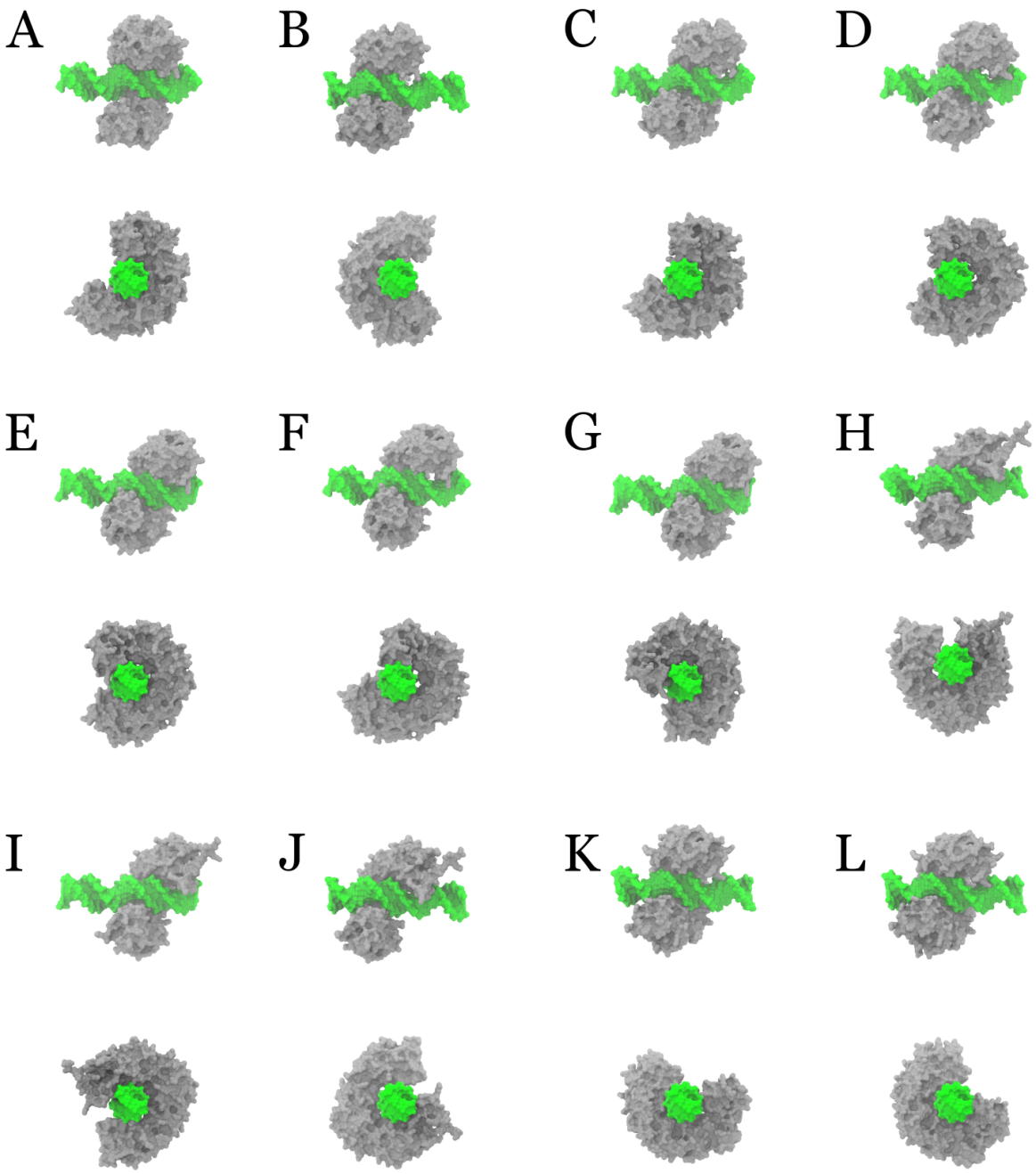


Figure 5.15. The 12 models of the nonspecific complex that best track the major groove obtained by docking low-pitch apo MTERF1 structures to B-DNA. The protein conformations that were used to generate the pose has the following helical parameters (pitch, radius, sweep): for (**A**) and (**B**): 41.1 Å, 9.8 Å, 372°; for (**C**): 41.6 Å, 10.1 Å, 368°; for (**D**), (**E**), (**F**) and (**G**): 40.3 Å, 11.4 Å, 340°; for (**H**) and (**I**): 39.8 Å, 16.2 Å, 276°; for (**J**): 37.8 Å, 16.0 Å, 287°; for (**K**) and (**L**): 34.5 Å, 9.8 Å, 389°.

5.3.8. Relaxing the docked poses using MD

To optimize and relax the docked complexes, and establish the stability of the search mode model, we simulated the productive poses using MD. The 12 docked poses were equilibrated and subjected to 3 μs of unrestrained MD. The complementarity of the protein-DNA interface increased during the simulations, as measured by shared surface area (**Figure 5.16**). Consistent with a weak binding model of search mode, the shared surface areas were less than that measured during MD of the specific complex (**Figure 5.17**). Despite looser binding, MTERF1 remained in contact with the major groove, indicated by stable time courses for the protein-major groove distance (**Figure 5.16**). The superhelical pitch and radius of MTERF1 from a representative simulation (**Figure 5.13C**) samples low pitch and radius metastates, the distributions of which are much narrower than apo MTERF1 (**Figure 5.13B**). This suggests the protein occupies a metastable conformational state complementary to B-DNA in the nonspecific complex. Overall, the observation of stable docked complexes with increased complementarity supports our hypothesis that the extensive superhelical dynamics of apo MTERF1 allow it to sample low pitch structures that are compatible with binding to a B-form DNA duplex.

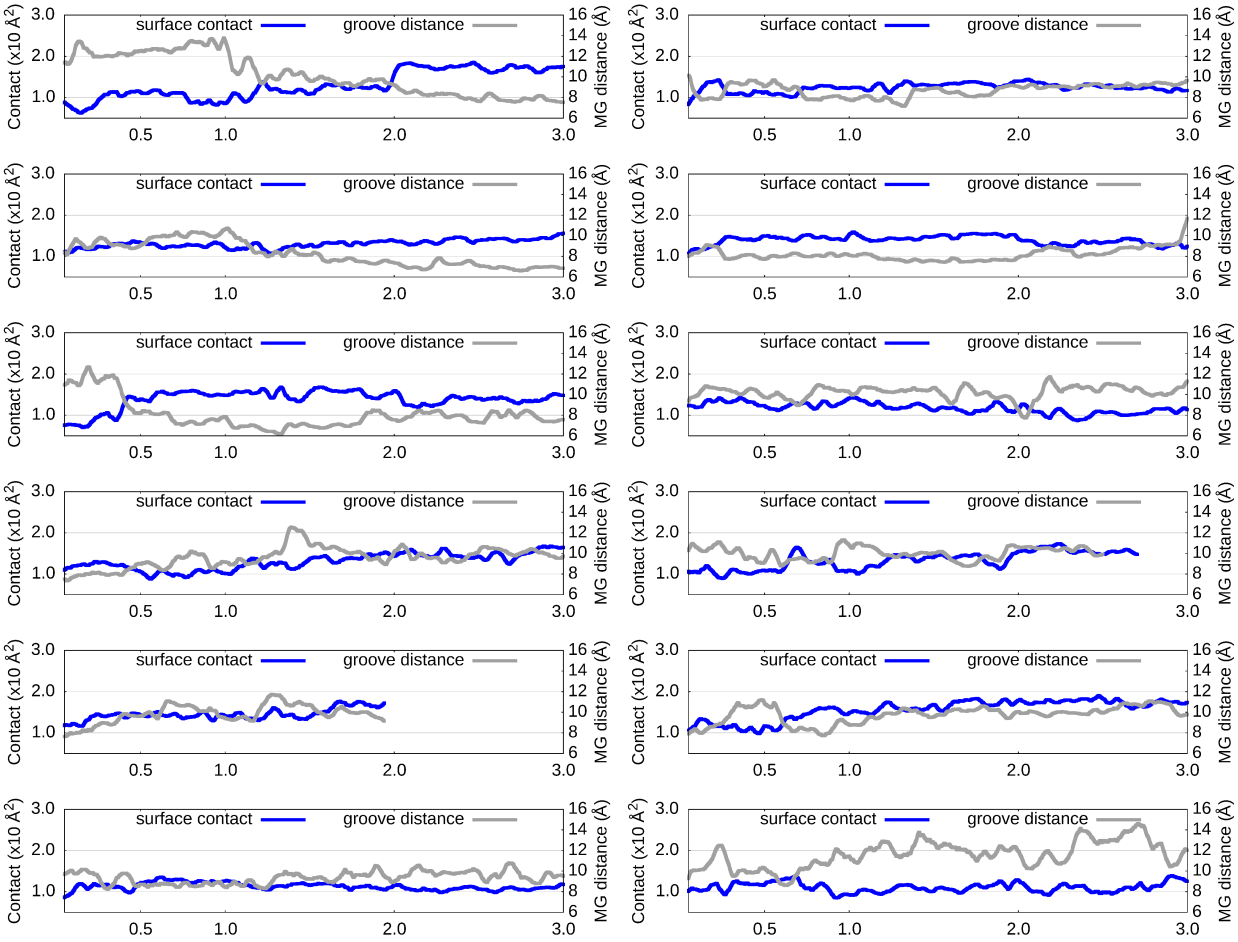


Figure 5.16. The shared surface area of MTERF1 and DNA in the 12 distinct search mode complexes. The abscissa is time, in units of μs.

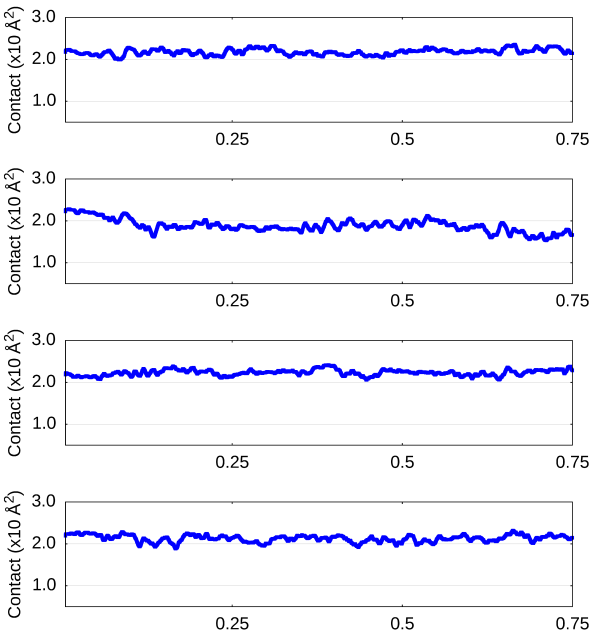


Figure 5.17. The shared surface area of MTERF1 and DNA in the recognition complex for each of the four independent simulations. The abscissa is time, in units of μs.

To gain more insight into the conformational changes that accompany the increasing surface complementarity of MTERF1 and DNA, we evaluated the RMSD of the protein, the DNA and the complex for a representative simulation (**Figure 5.18A**) along with the interface analysis discussed above (**Figure 5.18B**). Using a reference snapshot taken after 20 ns of MD (to account for initial relaxation), the DNA and protein structures were stable with RMSD remaining near 2 Å and 3 Å, respectively, during the entire MD run. This suggests that the increased structural complementarity involved relatively small changes to the protein and DNA structure, confirming our hypothesis that low pitch MTERF1 structures could accommodate B-DNA.

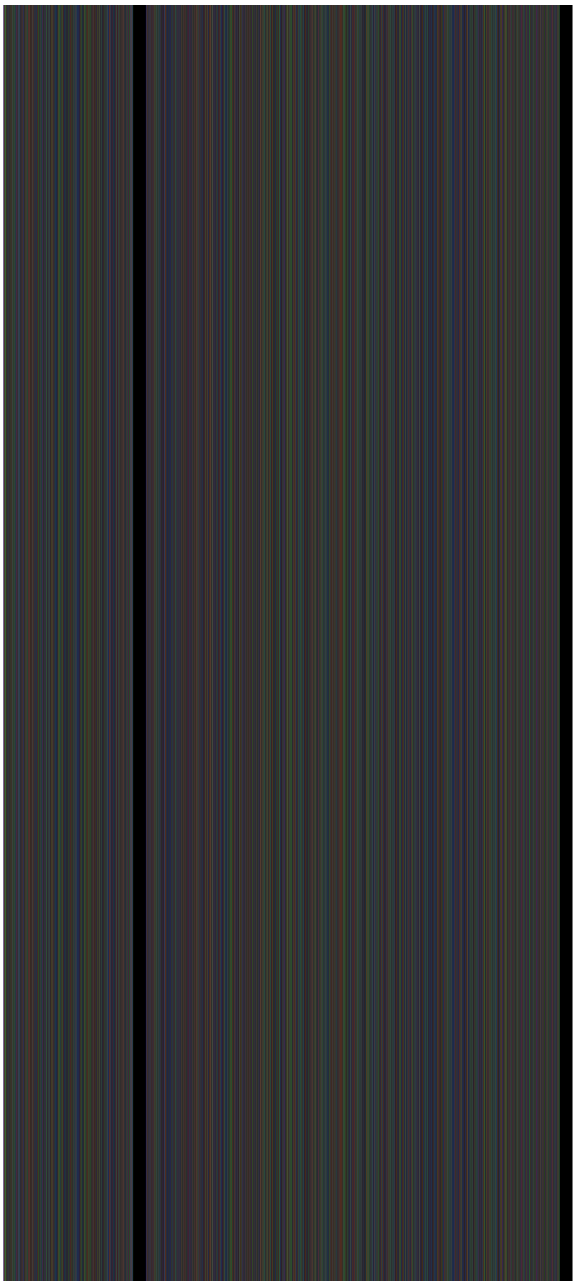


Figure 5.18. MTERF1 in search mode. (**A**) Structural stability of the search mode complex as measured by the backbone RMSD of the DNA (grey), MTERF1 (red) and the MTERF1-DNA complex (black), the last two of which were aligned to the protein; all used the structure at 0.5 μs as the reference, to account for docked pose relaxation. While the protein RMSD remains stable, that of the complex steadily rises. (**B**) Time dependence of the contact surface area shared by MTERF1 and the DNA (blue) and the major groove distance (grey). (**C**) Distance between the centers-of-mass of protein and DNA; increasing values with time suggest change in the location of the protein on the DNA. Data shown are averaged with a 50 ns sliding window.

Calculation of the RMSD values using the entire complex resulted in relatively low values during the first microsecond of the simulation, consistent with the argument that the docked poses were stable following modest relaxation (**Figure 5.18B**). For the final 2 μs, however, the RMSD of the complex drifts to higher values, eventually reaching 5 Å. The stable contact surface area suggests that the high RMSD value does not correspond to dissociation of the docked complex. We therefore investigated the possibility that the high RMSD values might arise from functionally relevant dynamics of MTERF1 in search mode.

5.3.9. MTERF1 undergoes 1D sliding during microsecond MD

Single-molecule fluorescence of eight different proteins sliding on DNA[213] suggests an accurate model of MTERF1 in search mode might exhibit spontaneous sliding along the major groove on the μs-timescale. To measure sliding, we define the translocation distance as the distance between the center of mass (COM) of the superhelical residues and the COM of the DNA (**Figure 5.18C** and **Figure 5.19**); this approximates the location of MTERF1 on the DNA since the contact surface indicates that the protein remains in the major groove. As shown in **Figure 5.18C**, this distance increases with time from an initial value of 8 Å to 17 Å, corresponding to sliding of 3 bp since the rise along each bp step is ~3 Å. During the first 0.5 μs, the translocation distance changes from an initial value of 8 Å to 11 Å, which may also correspond to initial relaxation of the docked protein into the major groove. After 1 μs, this distance increases again, to 14 Å; after the second μs, the distance becomes 17 Å, suggesting an approximate sliding rate of ~1 bp/μs that is consistent with experimental measurements on other protein-DNA complexes[213]. After 3 μs, the protein reaches the end of the duplex that was used for the docking simulations. To visually confirm sliding, we examined snapshots of the search

mode complex that were fit to the DNA (**Figure 5.19**). In these complexes, MTERF1 can be seen to diffuse along the DNA in the major groove. We conclude that the docking of low pitch apo MTERF1 to B-DNA leads to a dynamic model of MTERF1 in search mode.

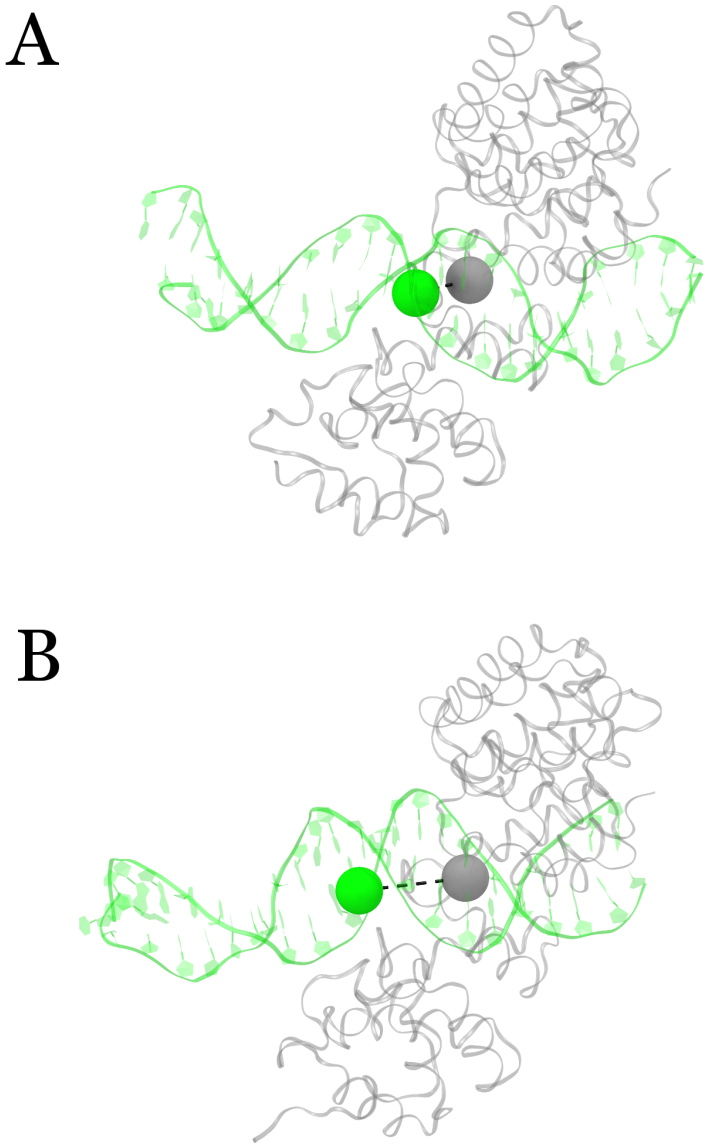


Figure 5.19. Visualization of a sliding distance metric. The protein (grey) and DNA (green) are shown as translucent cartoons to highlight the centers-of-mass. The center-of-mass (COM) of the superhelical residues is shown as a grey sphere. The COM of the C1' atoms in the DNA, excluding 4 bases at each end of the duplex to prevent end-fraying artifacts, is shown as a green sphere. (**A**) The equilibrated snapshot (0.5 μs) of the search mode. (**B**) A snapshot of the search mode (1.4 μs) that shows increased distance between the COM of the protein and the DNA.

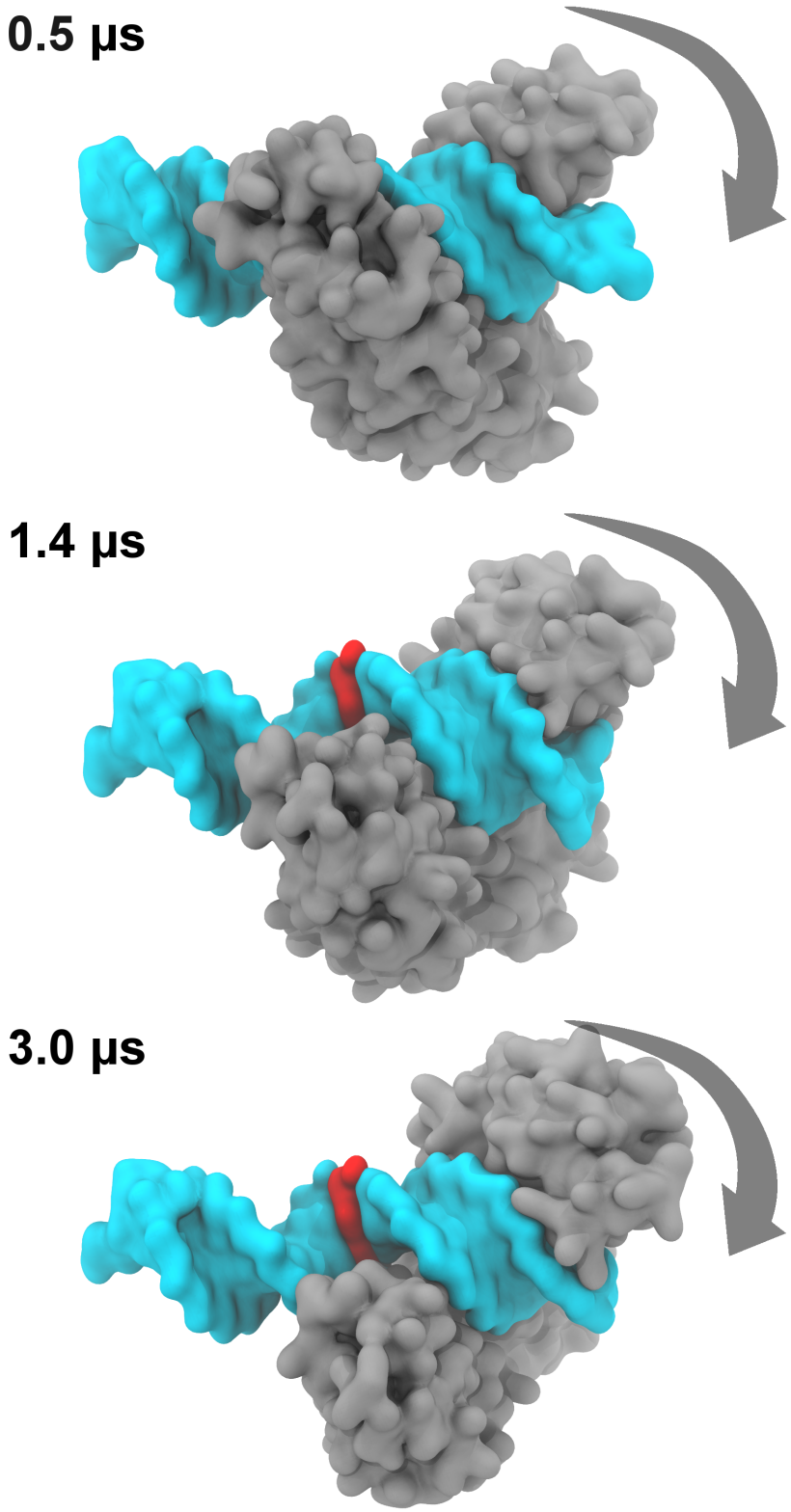


Figure 5.19. Snapshots of the complex at different time points. MTERF is shown in grey and is shown DNA in aqua, with the central bp colored red to visually highlight MTERF1 translocation along the major groove of B-DNA. Snapshots are RMS aligned to only the DNA backbone so that MTERF1 is seen to move (rightward) with respect to the DNA frame (indicated by arrows).

5.3.10. A model of the MTERF1-DNA search and recognition mechanism

The simulations described here provide a model of the dynamic MTERF1 nonspecific complex, supporting a hypothesis for the search and recognition mechanism. Based on our observation that the K-rich C-tail of MTERF1 is often unstructured, condensation of the protein onto DNA may be driven by a fly-casting mechanism[172]. It is possible that MTERF1 may follow a hybrid conformational selection-induced fit mechanism[214], in which a more unstructured MTERF1 folds in the proximity of the DNA. As MTERF1 collides with DNA, the intrinsic protein motions open the binding cleft to permit productive binding of B-DNA in a manner consistent with gated binding[173]. Once productively bound – the search mode described above – the MTERF1 superhelix is confined to helical motions compatible with the low pitch and low amplitude helical motions of B-DNA (**Figure 5.13C**). The precise motions and the degree to which they are coupled likely depend on sequence[18, 20], indirect readout, and shape readout[112, 215]. Generally, small barriers to sliding separate fleeting nonspecific complexes whose energy gaps are small (**Figure 5.19**). The intrinsic superhelical motions (**Figure 5.11**) not only allow the apo protein to bind DNA, but may also be a crucial factor in the ability of MTERF1 to unwind the target DNA, in which sequence and structure-dependent polarization may be key[216-217]. Contact with the target sequence switches the protein into recognition mode, accompanied by a conformational switch from low pitch (**Figure 5.13C**) to high pitch (**Figure 5.13A**) that drives DNA unwinding and base-flipping (**Figure 5.1**), likely with energetic compensation between DNA strain and formation of specific recognition contacts. The conformational switch is likely fast to allow efficient search and recognition, but the energy gaps are enlarged[218], leading to tight binding and a kinetic roadblock mechanism. A cartoon of the putative search and recognition landscape is shown in **Figure 5.20**. Thus, we propose MTERF1 follows an allosterically

modulated gated search and recognition mechanism in which the amplitude of the superhelical pitching motion is attenuated by direct and indirect readout. Future work will build on the model presented here to explore subsequent steps in specific recognition, with possible implications for genome editing reagent design[219-220].

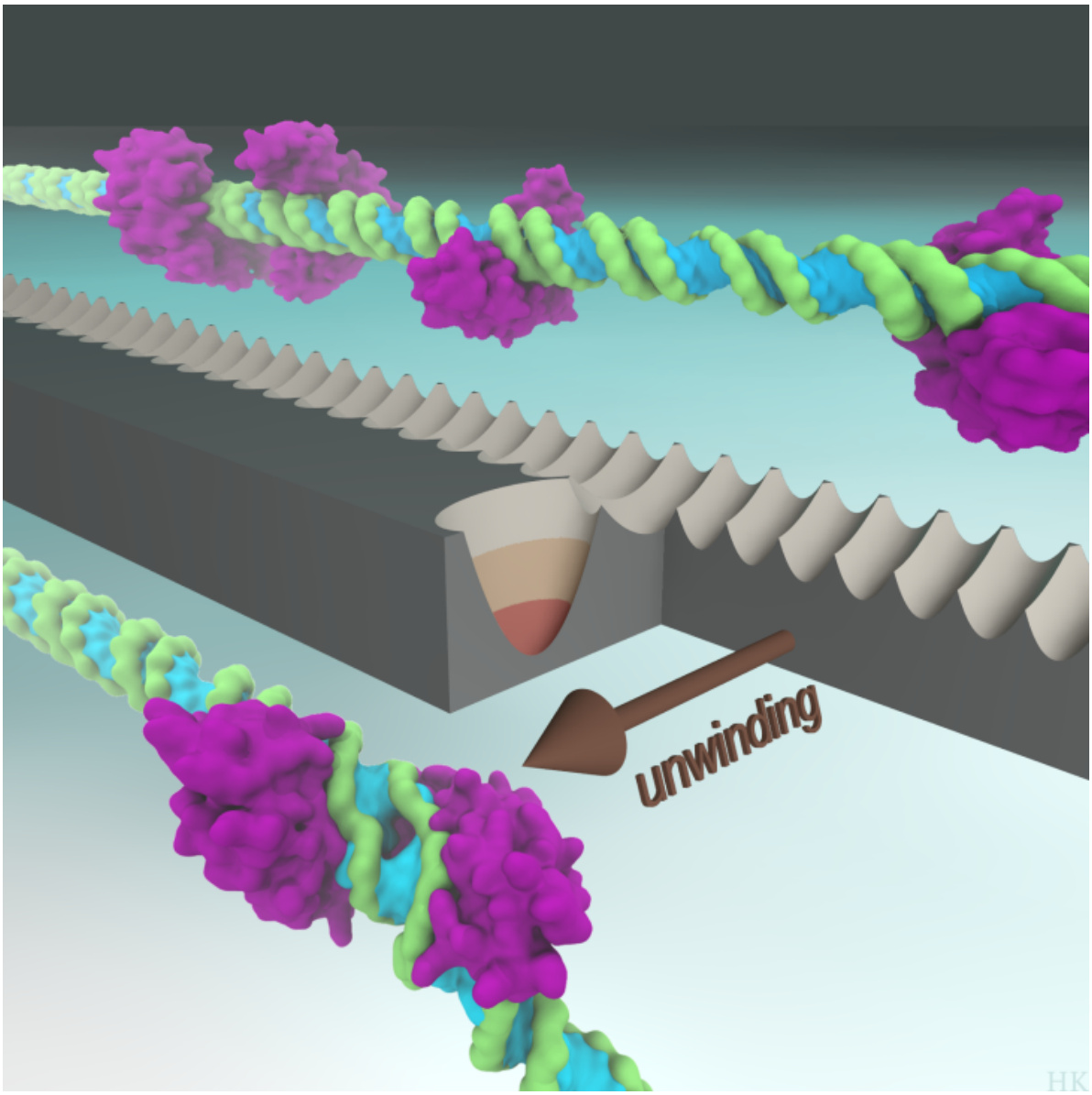


Figure 5.20. A model of the MTERF1 search and recognition landscape based on protein intrinsic superhelical motions. Helical motions drive translocation, and presumably for all but the target sequence, these motions are modulated so that the protein cannot fully unwind DNA before sliding on to the next site. At the target, the height of the unwinding barrier is sufficiently low for the protein to switch into recognition mode and unwind DNA rather than sliding to the next site.

5.4. Conclusions

We proposed several potential models for the nonspecific complex of transcription factors bound to B-DNA, using MTERF1 as a model TF. Our analysis indicated that conformational change of the TF was required, and MD simulations provided a model for the dynamic ensemble of apo MTERF1 structures. Analysis of the intrinsic motions indicated that dynamics of the superhelical topology characterize the changes during binding, and perhaps also during search and sequence recognition. Docked complexes provided reasonable models for the nonspecific complex, as indicated by low RMSD values, high surface area complementarity, and, on longer timescales, 1D diffusion (sliding) of the protein along the DNA major groove. The resulting dynamic model for this transcription factor carrying out nonspecific binding and search provides a view of protein-DNA recognition complementary to that obtained from a wealth of crystal structures of stable recognition complexes.

Chapter 6. Asynchronous shifts of protein-DNA contacts modulate sliding

Acknowledgement

This Chapter constitutes a manuscript in preparation by myself, Benjamin Schiffer, Angela Migues, Evangelos Coutsias, Miguel Garcia-Diaz and Carlos Simmerling. I conceived and wrote the manuscript, with edits and suggestions from the co-authors.

6.1. Introduction

Transcription factors (TF) regulate gene expression by specifically binding one or a few target sequences of DNA. Prior to binding the target, a TF diffuses by hopping along a continuous DNA sequence[6, 221], jumping between distal sequences that are near in space[222-223] and sliding in a spiral along DNA[204, 224-230]. Sliding can speed rather than slow target search, if a TF rapidly switches its conformation from a weak-binding search mode to a tight-binding recognition mode[203, 231-236]. Although atomic-level structures of search mode have been experimentally characterized for BstYI[23], BamHI[26], EcoRV[27] and lacI[237], these structures are static snapshots of a dynamic state and their relevance to the highly dynamic nature of search

mode may be limited[85]. However, these structures suggest that the protein and DNA, along with their interactions, are significantly different from the specific, recognition mode complex.

In search mode, the contacts between a DNA-binding protein (DBP) and the DNA backbone can be fleeting. Contacts qualitatively differ from recognition mode because amino acid-nucleotide contacts are short-lived in search mode[152-153, 237-239], but are highly stable in recognition mode. Transverse paramagnetic relaxation NMR on the HoxD9 homeodomain was consistent with a model in which the protein occupied all open DNA sites with equal probability[238], suggesting that the short-lived contacts may be due to amino acids shifting between phosphate groups on the DNA backbone. Contacts can shift as a DBP slides because a new register of protein-DNA interactions is formed at the next DNA site. Atomistic MD simulations of the lacI DNA-binding domain support this model: amino acids remain in contact with DNA but rapidly shift (ns-μs) between adjacent nucleotides[209]. Additional atomistic simulations on the lacI repressor (DNA-binding domain and activation domain) used umbrella sampling to estimate a sliding pathway, initiated from conformations that were generated by placing the protein in a helical path along the DNA groove[208]. These simulations were consistent with those of the lacI DNA-binding domain[209], suggesting that amino acids are labile in their interactions with DNA when sliding +1/-1 bp, with rate-limiting barriers estimated to be ~3 k_BT [208]. The computational expense of atomistic MD impeded these two sets of atomistic MD simulations from obtaining single trajectories beyond the μs-timescale likely to be important for sliding[213].

The ability of coarse-grained (CG) simulations to readily obtain trajectories on the μs-timescale has enabled them to characterize the general dynamics of how a variety of DBPs conduct search[240]. CG simulations have studied search using DBPs with multiple domains

(modules), which were known from biochemical experiments to exhibit asymmetric flexibility and asymmetric DNA-binding affinity[152, 241-243]. Modules with higher affinity for DNA tether lower affinity modules near the DNA, increasing the effective volume of the DBP[244]. This in turn increases the collision radius of the DBP and the DNA, increasing the rate of binding to a new DNA site by increasing the probability that the flexible module makes new DNA contacts. Critical to binding rate acceleration by asymmetric module dynamics is the flexibility of the linkers connecting them. Intermodule flexibility[245] should qualitatively change the energy landscape of search because the DBP behaves as if it were several independent modules rather than one rigid protein. That is, the energy barrier to break contact from one site by one module in a large modular protein with flexible linkers would be similar to the total energy barrier for the whole protein to slide +1/-1 bp. A rigid single-domain protein with the same number of contacts must break them all at once, leading to a large effective sliding barrier. Although CG simulations are ideally suited for studying the global dynamics of search mode, atomistic MD simulations are needed to resolve the possible change in coupling of intermodule dynamics during sliding.

In previous work, we used atomistic MD simulations on μs-timescales to model how the human mitochondrial transcription factor MTERF1 switched between search and recognition mode[85]. We found that MTERF1 adapts its superhelical tertiary structure to geometrically partner with DNA. For instance, MTERF1 in recognition mode adopts an extended superhelical conformation (high pitch) complementary to the unwound conformation (high pitch) of its target sequence[25] (**Figure 6.1a**). Whereas in search mode, our MD simulations predicted that MTERF1 adopts a compressed (low pitch) conformation that was compatible with the conformation of nonspecific B-DNA[85] (**Figure 6.1b**). MTERF1 spontaneously slid on DNA in one of our simulations, suggesting that sliding was assisted by thermal fluctuations of nonspecific

MTERF1-DNA interactions and implying that the net barrier to slide +1/-1 bp was ~ 1 k_BT. This suggests a smaller barrier to sliding compared with the barrier estimated by umbrella sampling simulations on lacI[208] discussed above (~ 3 k_BT).

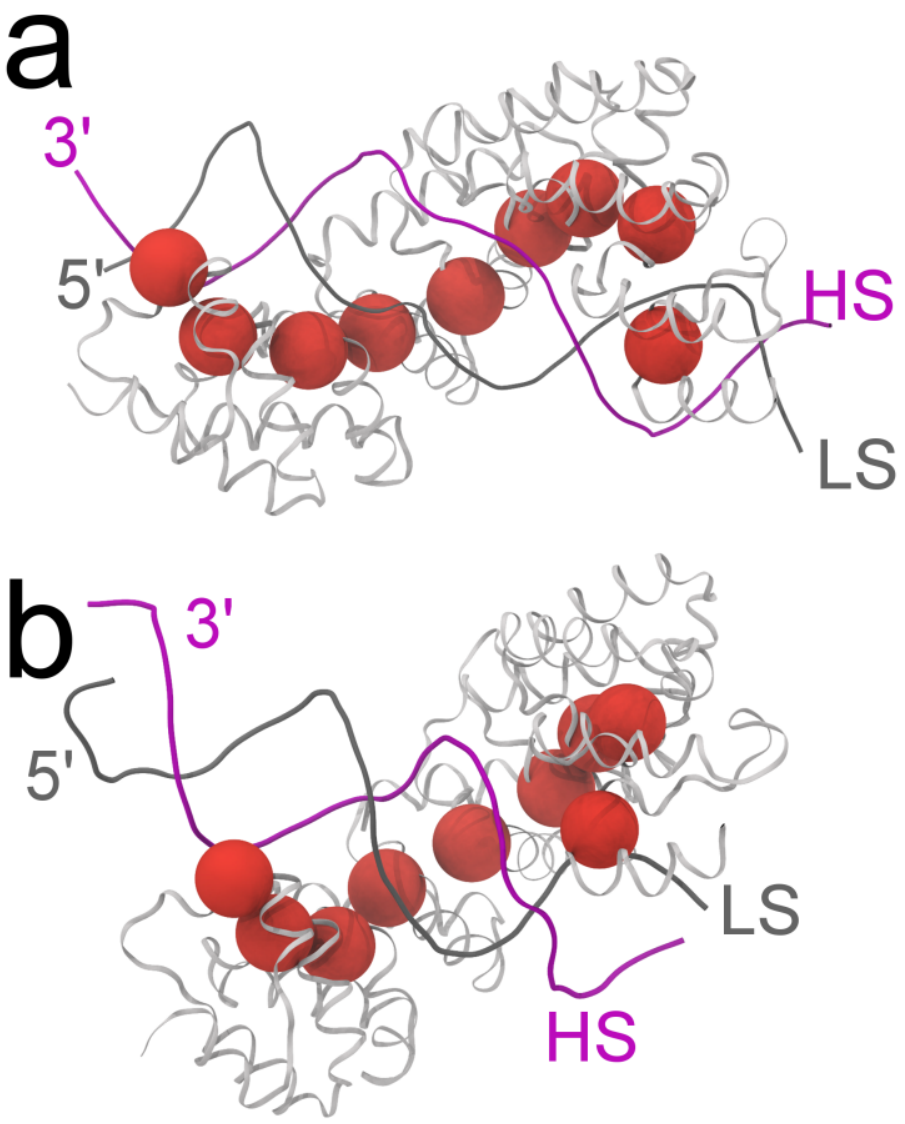


Figure 6.1. MTERF1 is a modular transcription factor with a superhelical tertiary structure that tracks the major groove of DNA. (**a**) In recognition mode the helix traced by superhelical Cα atoms (red spheres) adopts a high pitch conformation that complements the high pitch (unwound) conformation of the target DNA sequence (PDB ID: 3MVA[25]). (**b**) In search mode, the superhelix adopts a low pitch conformation that complements a B-like conformation of DNA representative of a nonspecific sequence[85]. MTERF1 is shown as white ribbons. The light-strand (LS) and heavy-strand (HS) of DNA are shown as grey and purple ribbons respectively.

Despite the breadth of experimental and computational investigations of DBP sliding, the mechanistic details remain unclear. NMR experiments and CG simulations suggest that the modules of Egr-1[152-153] and p53[235, 243] can have asymmetric DNA-binding affinities and flexibilities, and atomistic MD simulations suggest that the DNA-binding domains of lacI are stable but their relative orientation is flexible[208-209]. Our previous atomistic MD simulations suggest that the nine ~33 amino acid modules that comprise MTERF1 are similarly stable, but their orientation is highly dynamic[85]. As proposed by Yakubovskaya *et al.*, intermodule flexibility permits MTERF1 to adopt different conformations in the unbound state despite retaining the structure of individual modules[25].

What are the intramodule atomistic dynamics of thermally assisted MTERF1 sliding? Here, we hypothesized that nonspecific contacts between modules and the DNA backbone could independently (asynchronously) shift from one phosphate group to the next. It was expected that asynchronous shifts of nonspecific MTERF1-DNA interactions along the DNA sequence would appear as oscillations in the superhelical architecture (pitch), which would correlate with changes in the position of MTERF1 on DNA. Alternatively, module shifts could be highly coupled and MTERF1 superhelical pitch would be constant during sliding. **Figure 6.2** illustrates these potential models. In **Model a**, shifts in the contacts between modules and DNA are synchronized, requiring all nonspecific contacts to simultaneously shift to a new register. To maintain a high sliding rate, each contact must be weak such that their sum remains below a few k_BT. In **Model b**, modules asynchronously shift on DNA, permitting nonspecific contacts to break and form independently. Flexibility between modules uncouples the strength of the individual contacts from the net sliding rate. If **Model a** is true, superhelical pitch should be

constant. If **Model b** is true, superhelical pitch should fluctuate with an amplitude of ~1 bp, and correlate with sliding.

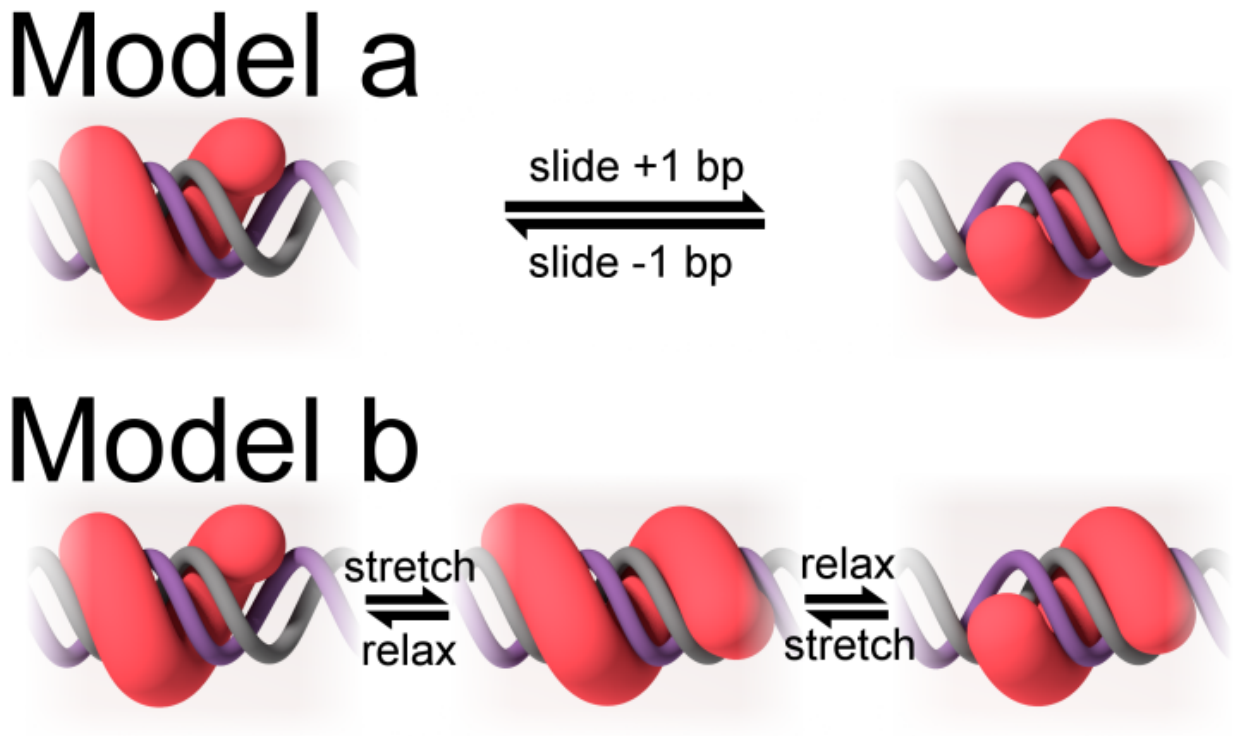


Figure 6.2. Two potential mechanisms of how MTERF1 slides. **Model a**, modules spiral in coordination by synchronously translating along the helical axis, forward (+1 bp) or reverse (-1 bp), and by rotating around the helical axis. **Model b**, modules spiral by asynchronously translating along and rotating around the helical axis. MTERF1 is shown in red and the DNA strands are shown in grey and purple.

Model a and **Model b** differ by the degree to which module shifts are coupled during MTERF1 sliding. To study the mechanism of MTERF1 sliding and determine which model better explains our observations, we followed our previous approach[85] utilizing unrestrained atomistic MD. While sliding in three new 7 μs simulations, MTERF1 remained nonspecifically bound to DNA. Key nonspecific contacts were asymmetrically arranged along the DNA-binding

footprint. Supporting **Model b**, we discovered that these nonspecific contacts asynchronously shifted as MTERF1 slid on DNA. MTERF1 superhelical pitch fluctuated with an average amplitude of 1 bp, and the fluctuations correlated with the discrete steps taken by MTERF1 as it slid. The results support a highly dynamic model of MTERF1 in search mode, characterized by loose coupling between its asymmetrically arranged nonspecific DNA-contacts. A new model of sliding emerges: Asynchronous shifts of protein-DNA contacts smooth the sliding landscape by decoupling the strength of individual amino acid-nucleotide interactions from the net sliding rate.

6.2. Methods

6.2.1. Molecular dynamics simulations

The MTERF1-DNA search mode complex was prepared using previously described procedures[85], using a truncated octahedron of explicit water with dimensions suffice to ensure all solute atoms were at least 10 Å from the edge of the cell. Explicit K+ and Cl- ions were added in 0.2 M excess by randomly replacing water molecules, such that ions were > 4 Å from the solute and > 6 Å from each other. The Amber ff94 force field[246] was utilized with the ff99SB[186] peptide φ/ψ and φ'/ψ' backbone torsion modifications and the parmBSC0[187] DNA α/γ backbone torsion adjustments; the TIP3P[188] parameters were used for the explicit water molecules and the Joung & Cheatham monovalent ion parameters[189] used for the explicit ions.

We used the Berendsen thermostat and barostat[201] to regulate temperature and pressure. Electrostatic interactions across the lattice of the periodic simulation cell were truncated using

particle mesh Ewald summation[200]. An 8 Å non-bonded interaction cutoff was used. Initial equilibration was performed using a nine stage protocol as previously described[85]. The 0.5 μs equilibration and 6.5 μs production simulations used constant volume. H-masses were partitioned to permit 4 fs integration[202].

6.2.2. Structural analyses

Hydrogen bonds (H-bonds) between the protein and the DNA were counted if the distance separating the donor (X-H) and acceptor atoms (Y) was ≤ 3.0 Å, and if the angle subtended by the X-H:Y interaction was > 135°. The MTERF1-DNA interface surface area was calculated using the LCPO method[247] as implemented in cpptraj[211], by subtracting the surface area of the MTERF1-DNA complex from the sum of the MTERF1 and DNA surface areas. The average difference between the Cartesian coordinates of the atoms in a reference structure and MTERF1, DNA and the MTERF1-DNA complex was calculated using the RMSD algorithm implemented in cpptraj[211]. The equilibrated structure was used as the reference (time point 0.5 μs), using the Cα atoms of motifs 2 through 8 and the C1' atoms of the DNA excluding the termini, per our protocol[85]. For each of the three calculations (MTERF1, DNA and complex), RMS-fitting was used to remove rotational and translational displacements from each MD frame; this RMSD reports the change in internal structure of the molecule. The RMSD of the superhelical Cα relative to the DNA was calculated by RMS-fitting the trajectories to the DNA, followed by calculation of superhelical Cα RMSD (without refitting); this RMSD reports the change in orientation of MTERF1 with respect to the DNA.

The position of MTERF1 modules on the DNA sequence were defined by determining to which P atom of a nucleotide the Cζ atom of an Arg residue was nearest (**Table 6.1**). These Arg amino acids formed interdigitated contacts with both DNA strands: Arg92-$P_{LS,j}$, Arg127-$P_{LS,j-1}$, Arg162-$P_{LS,j-2}$, Arg195-$P_{LS,j-3}$; and Arg99-$P_{HS,i}$, Arg134-$P_{HS,i+1}$, Arg169-$P_{HS,i+2}$, Arg202-$P_{HS,i+3}$. The indices run 5' to 3'. The HS increments in reverse of the LS because the strands are complementary. Nucleotide sequence indices used to monitor the position of modules on DNA correspond to the numbering of the human mitochondrial genome sequence. The amide H atoms of Ser355 and Lys385 that H-bonded to the DNA

Table 6.1. Residues and atoms that donate H-bonds in nonspecific contacts with DNA.

Arg92, motif 1	Cζ
Arg99, motif 1	Cζ
Arg127, motif 2	Cζ
Arg134, motif 2	Cζ
Arg162, motif 3	Cζ
Arg169, motif 3	Cζ
Arg195, motif 4	Cζ
Arg202, motif 4	Cζ
Ser355, C-segment	H (backbone)
Lys385, C-tail	H (backbone)

Residues in bold are involved in recognition (see Yakubovskaya et al.[25]).

MTERF1 superhelical pitch and the helical pitch of the two DNA strands were calculated using previously described procedures[85]. In short, a helical axis was defined as the normal vector

to the plane that a helix's atoms (superhelical Cα atoms or DNA C1' atoms) projected a circle. Pitch was calculated by dividing the total height of the helix by the total angle swept by the atoms in the projection plane. Helix radius was the radius of the projected circle. A 0.5° grid-step resolution over spherical coordinates was used to find the helical axis.

Data were smoothed using rolling averages with a period of 25 ns. The time-dependent correlation between MTERF1 and the DNA helical pitching motions was calculated using Pearson's R^2 in non-overlapping blocks of 50 ns.

6.3. Results

NMR and coarse-grained MD simulations suggest that modular DBPs can exhibit asymmetry in the relative flexibility and the relative DNA-binding affinity of the modules. Structural asymmetry can create a gradient that drives diffusion[248]. The varying size and amino acid composition of the MTERF1 modules implies that it may exhibit asymmetric DNA-binding affinity and flexibility like that observed in Egr-1[152-153], Pax6[245] and p53[243]. Asymmetrical flexibility implies that intermodule dynamics are independent. If nonspecific contacts between MTERF1 modules synchronously shift along DNA, then MTERF1 sliding follows the mechanism described by **Model a** (**Figure 6.2**). On the other hand, if nonspecific contacts between MTERF1 modules synchronously shift along DNA, then MTERF1 sliding follows the mechanism described by **Model b**. To determine if MTERF1 modules shift synchronously (**Model a**) or asynchronously (**Model b**), we used our previous approach[85] and utilized

unrestrained atomistic MD simulations of low pitch MTERF1 structures bound to B-form DNA. Here we present an analysis of three, unrestrained 7 μs trajectories in which MTERF1 slides.

6.3.1. The superhelix and DNA are stable, but change with respect to each other

For the three 7 μs trajectories, the RMSD of MTERF1 is stable and the RMSD of the DNA is stable, indicating that their overall structural features are maintained (**Figure 6.3a**). However, the RMSD of the complex steadily rises, suggesting that the structure of the complex is changing (**Figure 6.3a**). If the RMSD of the complex was rising because MTERF1 was disassociating from the DNA, then the number of H-bonds should decrease and the surface area of the interface should also decrease. The number of H-bonds did not decrease (**Figure 6.3b**). Similarly, the interface surface area (iSA) between MTERF1 and DNA remained stable (**Figure 6.3b**). Because the RMSD of MTERF1 and the DNA were stable, disassociation is not causing the change in RMSD of the complex. If the structure of the complex was changing because MTERF1 was sliding on the DNA, then the relative position of the superhelix with respect to the DNA should change. The change in position of MTERF1 relative to the DNA was measured by first RMS-fitting the trajectories to the DNA, then calculating the RMSD of the superhelical Cα without further fitting (retaining rotations and translations relative to DNA). **Figure 6.3c** shows that the RMSD of the superhelical Cα relative to the DNA increased during the simulations, often as discrete steps. The decrease in iSA and number of H-bonds at ~ 5 μs in run 3 might reflect a hopping event because C-terminal domain of MTERF1 became unstructured (**Figure 6.4**); as the dynamics were not strictly sliding, these potential hopping dynamics were beyond the scope of this work. Overall, the results in **Figure 6.3** indicate that the structure of MTERF1 and the DNA are stable, the MTERF1-DNA interface is stable and MTERF1 is sliding. Given these

observations, we hypothesized that MTERF1 was sliding while remaining in contact with the DNA via nonspecific contacts to the DNA backbone. Thus, we visually inspected the trajectories to identify particular amino acids involved in dynamic nonspecific contacts with the DNA.

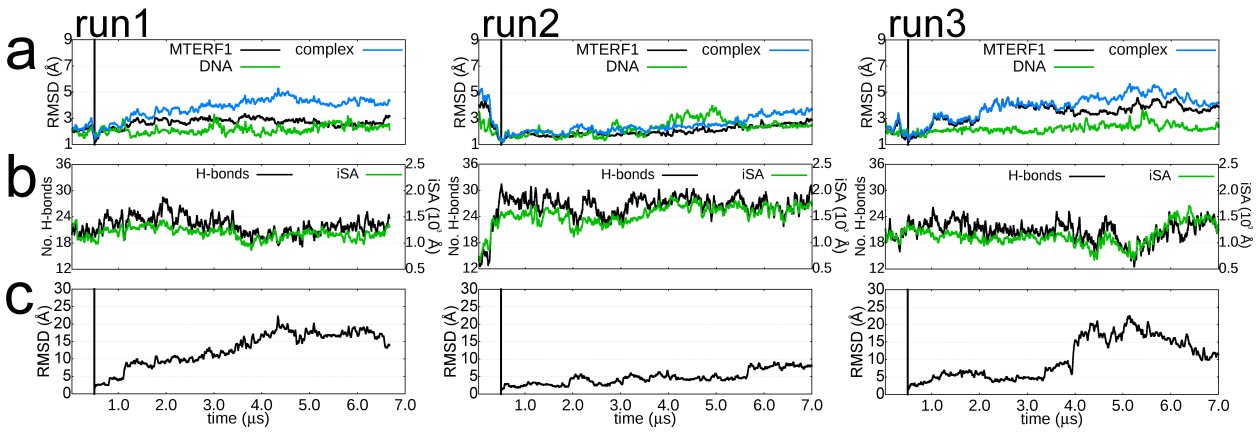


Figure 6.3. General dynamics of three MTERF1-DNA sliding trajectories. (**a**) RMSD of MTERF1 (black), DNA (green) and the complex (blue) RMS-aligned to themselves to remove rotational and translational displacements. The structure at 0.5 μs was used as the reference (vertical line). (**b**) Number of H-bonds between MTERF1 and the DNA (black) is plotted using the left y-axis, and the MTERF1-DNA interface surface area, iSA (green), is plotted using the right y-axis. (**c**) RMSD of the superhelical Cα RMS-aligned to DNA.

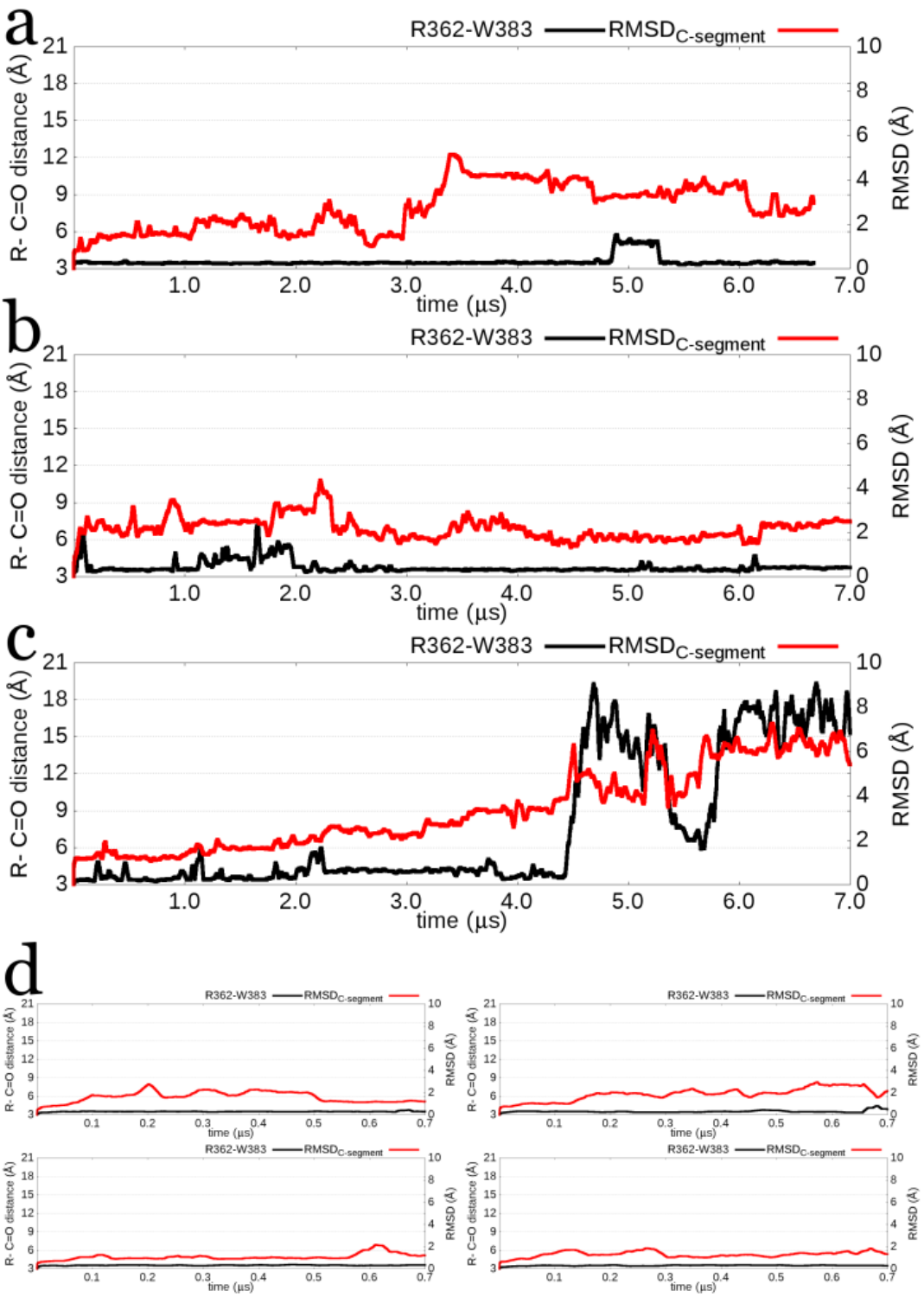


Figure 6.4. Structural analysis of the C-segment of MTERF1, which includes the Lys-rich C-tail. Panels (**a**), (**b**) and (**c**) correspond to simulations 1, 2 and 3, respectively. The contact between the guanidino group of Arg362 and the backbone carbonyl of Trp383 remains through the course of simulations 1 and 2, with only transient disassociations. At 4.4 μs in simulation 3, however, the distance grew rapidly as the C-segment became unstructured. A rolling average was used to smooth the data (0.025 μs block size).

6.3.2. MTERF1 establishes asymmetric, modular contacts with DNA

How does MTERF1 contact the DNA in search mode? We anticipated a modular arrangement of nonspecific MTERF1-DNA contacts because of the modular architecture of the protein. To identify a particular set of amino acids that shifted along the DNA backbone, trajectories were visualized. We found MTERF1 modules 1, 2, 3 and 4 each displayed a pair of Arg sidechains that contacted the DNA backbone (**Figure 6.5a**). Arg92, Arg127, Arg162 and Arg195 contact the LS of DNA (henceforth, "Arg-LS contacts"), and Arg99, Arg134, Arg169 and Arg202 contact the HS of DNA (henceforth, "Arg-HS contacts"). The C-terminal domain of MTERF1 displayed two additional amino acids that contacted the HS of DNA in the three trajectories: Ser355 and Lys385 (henceforth, "SK-HS contacts"). The arrangement of the nonspecific contacts mediated by these ten amino acids led to asymmetry along the DNA sequence and between the two DNA strands. Eight contacts versus two contacts were formed along the DNA-binding footprint (axial asymmetry, **Figure 6.5a**) and four contacts versus six contacts were formed with the two DNA strands (radial asymmetry, **Figure 6.5a**). Despite their asymmetric arrangement, all ten amino acids could contact the DNA backbone simultaneously (**Figure 6.5b**). Unlike the contacts in the N-terminal domain utilizing Arg side chains, the contacts in the C-terminal domain were mediated by the backbone amides of Ser355 and Lys385. The eight Arg amino acids are highly conserved (**Figure 6.6**), as expected; Ser355 and Lys385 are not (**Figure 6.6**), because any amino acid should be capable of forming a backbone amide-mediated contact. As MTERF1 slides, how do these ten amino acids shift between nucleotides?

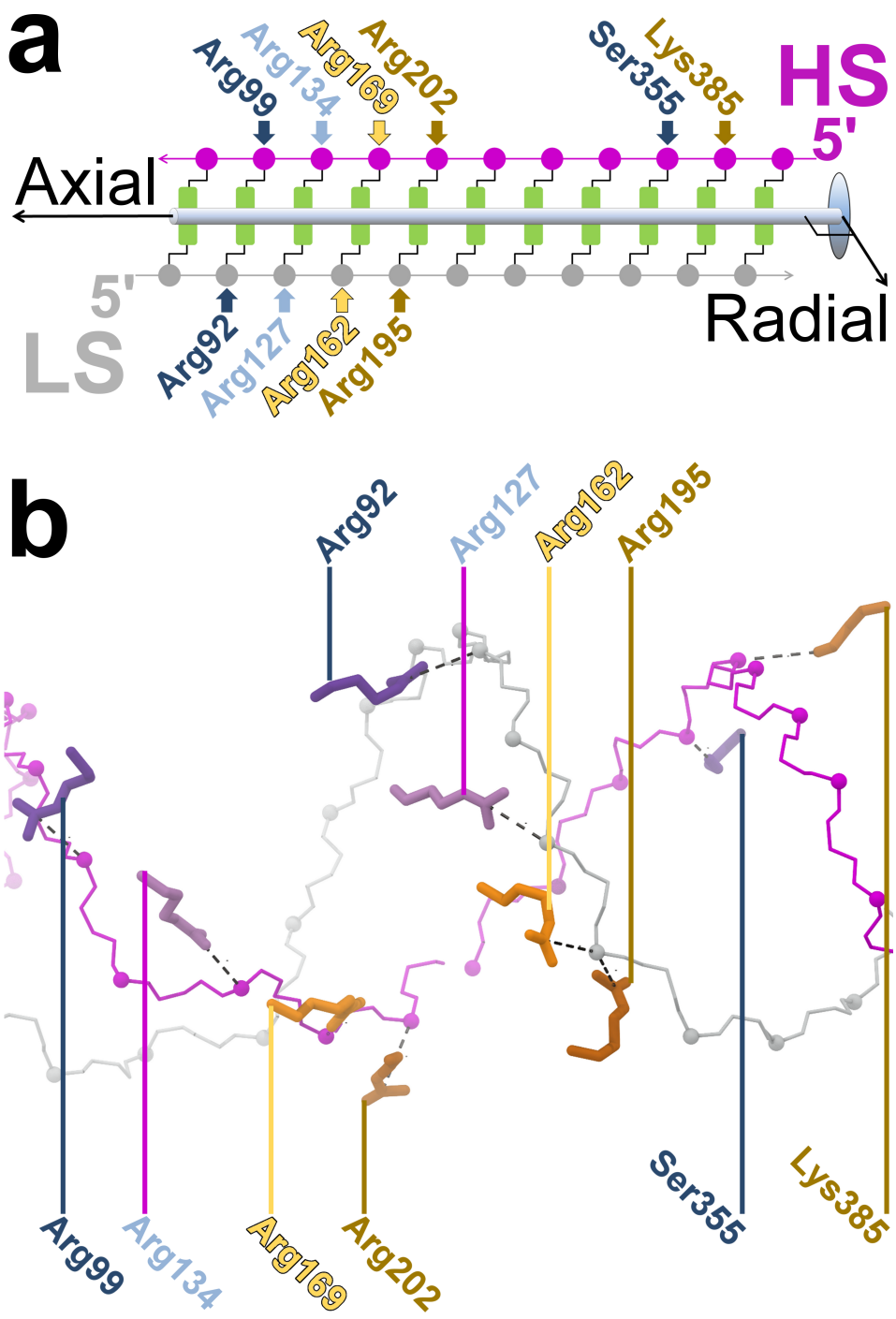


Figure 6.5. Asymmetric, modular MTERF1-DNA contacts in search mode. (**a**) Schematic of the ten modular MTERF1-DNA interactions. Arg99, Arg134, Arg169, Arg202, Ser355 and Lys385 H-bond to the phosphate groups (purple circles) on the HS of DNA. Arg92, Arg127, Arg162 and Arg195 H-bond to the phosphate groups (grey circles) on the LS of DNA. Axial asymmetry arises from an imbalance in the number of H-bonds relative to the helix axis; radial asymmetry arises from an imbalance in the number of H-bonds perpendicular to the helix axis. (**b**) Representative snapshot of modular contacts. Purple and grey spheres represent P atoms on the HS and LS of DNA respectively.

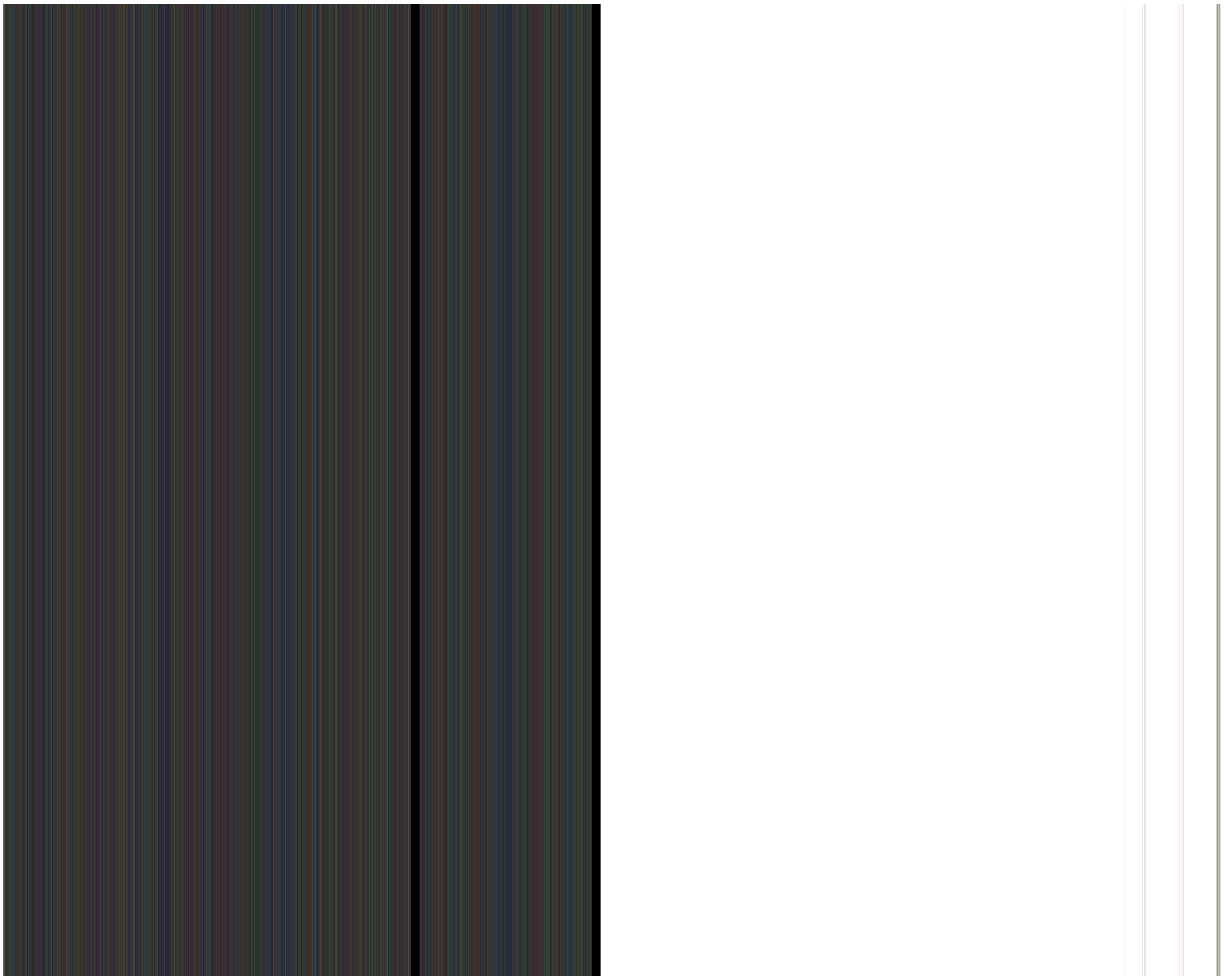


Figure 6.6. Human MTERF1 sequence alignment. Eight species were aligned to the MTERF1 protein sequence: *hs*, homo sapiens; *fc*, felis catus; *ss*, sus scrofa; *ec*, equus caballus; *bt*, bos taurus; *mm*, mus musculus; *rn*, rattus norvegicus; *oa*, ornithorhynchus anatinus; *dr*, danio rerio. Accession codes are provided; "XP" refer to sequences predicted to be MTERF1. Red dots highlight amino acids that do note conserve sequence identity. For Ser355 and Lys385, open circles are used to highlight amino acids that are not conserved in other species. Sequence alignments were performed using constraint based alignment tool (COBALT).

6.3.3. Asymmetric contacts shift asynchronously

NMR relaxation experiments [238] and MD simulations [85] suggest that amino acids can shift their H-bonding between adjacent phosphate groups on a DNA backbone. However, it is unclear whether sliding involves concerted shifts (**Model a**) or asynchronous shifts (**Model b**) of these nonspecific amino acid-nucleotide H-bonds. To determine whether MTERF1 follows **Model a** or

Model b, we monitored the position along the DNA sequence of the ten, modular MTERF1-DNA contacts described in **Figure 6.5**. In run 1, the Arg-LS contacts shift at 1.8 μs and 3.8 μs (**Figure 6.7a**), the Arg-HS contacts shift at 1.2 μs and 3.2 μs (**Figure 6.7b**), and the SK-HS shift at 0.8 μs, 1.5 μs, 2.8 μs and 4.3 μs (**Figure 6.7c**). In runs 2 and 3, the Arg-HS (**Figure 6.7a**), Arg-LS (**Figure 6.7b**) and SK-HS contacts (**Figure 6.7c**) shift at different time points as in run 1. Arg-HS, Arg-LS and SK-HS contacts shift asynchronously during the three sliding trajectories. This potentially suggests that radial contact-asymmetry between the two DNA strands contributes to asynchronous shifts (**Figure 6.7a** versus **Figure 6.7b,c**) and that axial contact-asymmetry along the DNA footprint also contributes to asynchronous shifts (**Figure 6.7a,b** versus **Figure 6.7c**). Asynchronicity also appears between adjacent contacts on one DNA strand: the Arg-HS contacts in run 1 shift independently because Arg134 and Arg202 shift (1.2 μs) without Arg99 and Arg169; the SK-HS contacts in run 1 shift independently because Lys385 shifts (1.5 μs) without Ser355, then Ser355 shifts (4.2 μs) without Lys385. These results indicate that individual amino acid-nucleotide contacts are highly dynamic and weakly coupled, permitting contacts to shift semi-autonomously as MTERF1 slides. Since the superhelix of MTERF1 is defined by the relative orientation of its modules, autonomous shifts of MTERF1 modules along the DNA sequence (helical axis) should lead to changes in superhelical pitch.

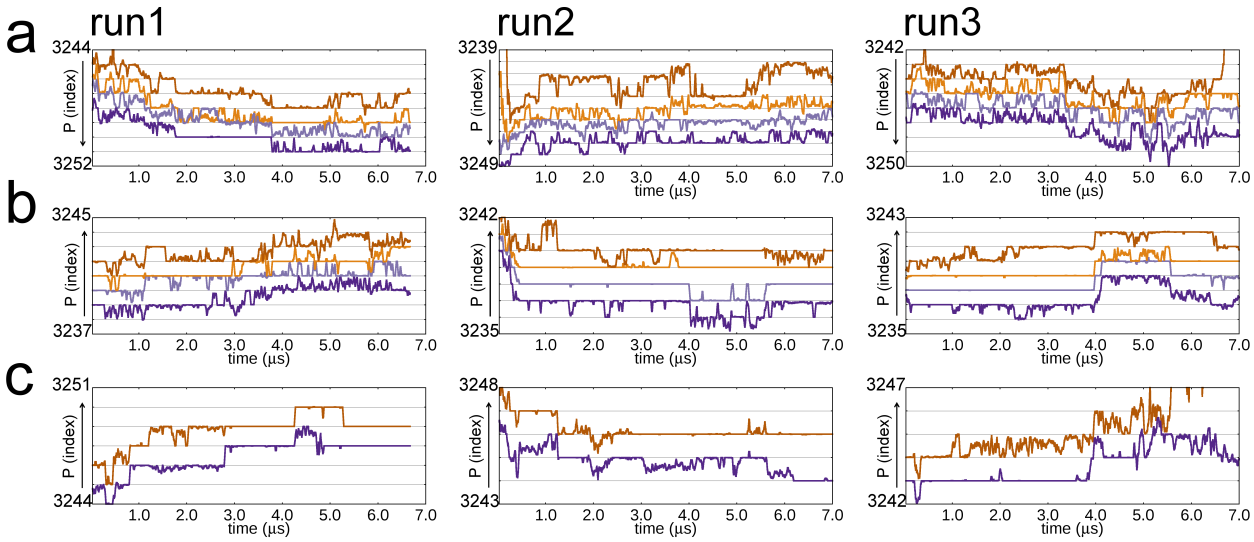


Figure 6.7. MTERF1 modules shift asynchronously along the DNA backbone. (**a**) Time course of nonspecific contacts to the LS made by Arg92 (dark purple), Arg127 (light purple), Arg162 (light orange) and Arg195 (dark orange). The y-axis depicts the 5' to 3' sequence index of the P atom in a nucleotide to which an amino acid is H-bonded. (**b**) Time course of nonspecific contacts to the HS made by Arg99 (dark purple), Arg134 (light purple), Arg169 (light orange) and Arg202 (dark orange). The y-axis is depicted as in (**a**), but for the HS, plotted 5' to 3'. (**c**) Time courses of nonspecific contacts to the HS made by Ser355 (dark purple) and Lys385 (dark orange). The y-axis is depicted as in (**b**).

6.3.4. Oscillations in MTERF1 superhelical pitch correlate with sliding

A change in the relative orientation of the MTERF1 modules can be characterized by changes in MTERF1 superhelical pitch [85]. Therefore, asynchronous shifts in the MTERF1 modules will lead to fluctuations in MTERF1 superhelical pitch corresponding to time points when asynchronous shifts occur. Thus, changes in MTERF1 pitch should correlate with MTERF1 sliding. If the dynamics of MTERF1 sliding are similar to the dynamics of the shifts for individual modules, then the position of MTERF1 on the DNA sequence should exhibit steps corresponding to one bp (~4 Å). To monitor MTERF1 sliding, the position of the center of mass of the superhelical Cα along the DNA helical axis was measured. MTERF1 slides in discrete

steps on the ns-timescale, followed by extended stationary periods on the μs-timescale (**Figure 6.8a**). As expected, MTERF1 slides along DNA in discrete steps of ~4 Å. Since MTERF1 modules shift along the DNA asynchronously, and MTERF1 steps along the DNA in discrete steps, the superhelical pitch of MTERF1 should exhibit fluctuations of ~4 Å. We measured MTERF1 superhelical pitch in the three sliding trajectories and calculated the standard deviation of the pitch for each trajectory. It was assumed that the standard deviation was representative of the fluctuations in MTERF1 superhelical pitch, and thus would be representative of the amplitude of its fluctuation. Excluding the equilibration period (and the hopping switch in run 3), the standard deviation of MTERF1 superhelical pitch in all three simulations was 3 Å (**Table 6.2**), which is precisely the fluctuation required for MTERF1 to step one bp along a DNA sequence. If these superhelical oscillations corresponded to the sliding steps above, the two datasets should be correlated. Alternatively, if sliding involved purely random MTERF1 module fluctuations (random changes in superhelical pitch), then pitching and sliding should be uncorrelated.

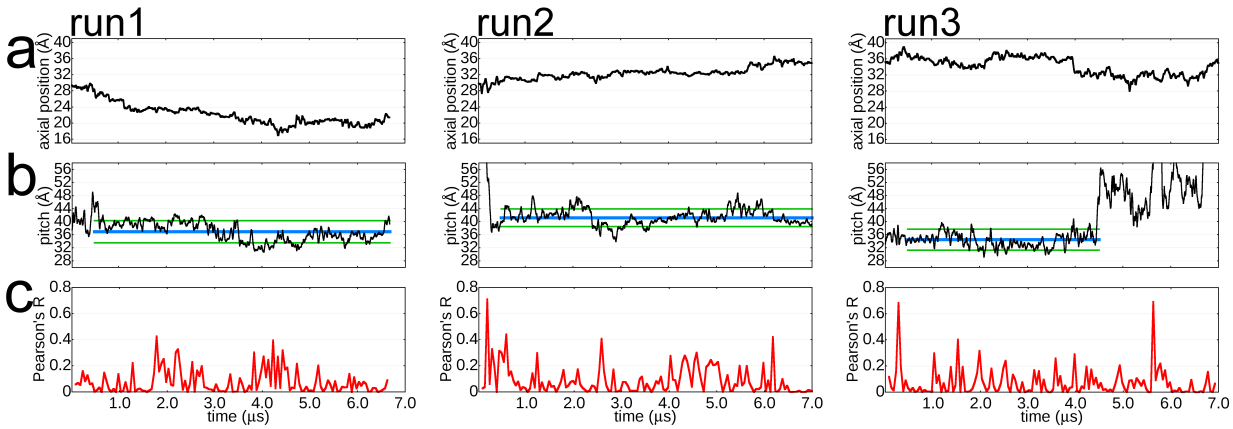


Figure 6.8. MTERF1 superhelical pitch undulates when MTERF1 takes a step along DNA. (**a**) Axial position of the COM of the superhelical Cα along the DNA helical axis in three simulations: run 1 (steps at 0.8 μs, 1.2 μs, 3.8 μs, 4.3 μs); run 2 (steps at 0.7 μs, 1.2 μs, 2.7 μs, 5.5 μs); run 3 (steps at 1.1 μs, 2.3 μs, 3.9 μs). (**b**) Superhelical pitch of MTERF1 in three simulations (run1, run2, run3). (**c**) Pearson's R^2 of MTERF1 superhelical pitch and the position of MTERF1 along the DNA; time-dependent R^2 values were calculated from non-overlapping 50 ns blocks, for three simulations (run1, run2, run3).

Table 6.2. Means and standard deviations of superhelical pitch, radius and sweep.

Run	<pitch>	σ, pitch	<radius>	σ, radius	<sweep>	σ, sweep
1	36.9	3.4	13.2	0.6	303	13.2
2	41.2	2.7	13.7	0.5	296	9.7
3	34.5	3.2	16.2	0.8	252	14.0

Run 1, 2, 3: equilibration 0.5 μs of data excluded. Run 3: data after 4.5 μs was excluded due to unbinding. <> signifies the mean; σ signifies standard deviation.

Figure 6.8c presents the correlation between the axial position of MTERF1 and its superhelical pitch, calculated using Pearson's R^2 over non-overlapping 50 ns slices of the trajectories. Peaks of correlation are present that do not correspond to the 4 Å steps in **Figure 6.8a**, however when steps do occur, a peak is present in the correlation. This indicates that the

intrinsic superhelical pitching motion of MTERF1 is constantly fluctuating with the amplitude needed to step one bp along a DNA sequence. Therefore, MTERF1 sliding is consistent with **Model b** because MTERF1 sliding involves asynchronous shifts of module-nucleotide interactions that are weakly coupled, fluctuating in concert when the protein takes a step to slide.

We previously showed using CG and atomistic MD simulations that the principal motion of MTERF1, based solely on its superhelical architecture, is to adapt superhelical pitch [85]. Here, MTERF1 appears to harness thermal fluctuations when it slides because its intrinsic superhelical pitching motion correlates with stepping events. In addition, the step-wise global sliding dynamics of MTERF1 are similar to the local shift dynamics of its modules, suggesting that MTERF1 behaves as a system of semi-autonomous modules, rather than a single rigid protein. Asynchronous sliding dynamics of MTERF1 should affect the search energy landscape.

Figure 6.9 illustrates a potential model of MTERF1 sliding under two dynamical regimes: synchronous shifts and asynchronous shifts, corresponding to **Model a** and **Model b** respectively. In **Figure 6.9a**, each protein-DNA contact resides in a sinusoidal energy landscape of amplitude ε. The position of each contact is highly coupled to the rest because the protein is rigid. Thus, modules shift synchronously. To slide, the protein must simultaneously break all three contacts, leading to a net sliding barrier of 3ε. **Figure 6.9b** illustrates a sliding landscape when modules asynchronously shift (**Model b**). If the coupling strength between the modules is low, one module at a time can shift. Therefore the net barrier to sliding is ~ε. Interestingly, if both models are assumed to have the same overall rate of sliding, then the barriers in **Model a** must be 1/3 the value of those in **Model b**. Conversely, the barriers of individual contacts in **Model b** could be 3-fold higher than **Model a**, potentially permitting enhanced interrogation without slowing the overall rate of sliding. **Model b** is consistent with copious experimental

observations of DBPs exhibiting high sliding rates despite retaining high sequence specificity. Our previous simulations suggest that the sliding landscape can be smoothed by a conformational switch between search and recognition mode [85], which involved adaptation of MTERF1 superhelical pitch. Here, we refine that model by discovering that MTERF1 can further smooth its sliding landscape by decoupling the dynamics of individual amino acid-nucleotide interactions, driven by the same superhelical pitching motion. Thusly uncoupled, MTERF1 might retain the strength of individual contacts without increasing its net sliding rate. The model may be applicable to other DBPs with superhelical architectures such as TALE.

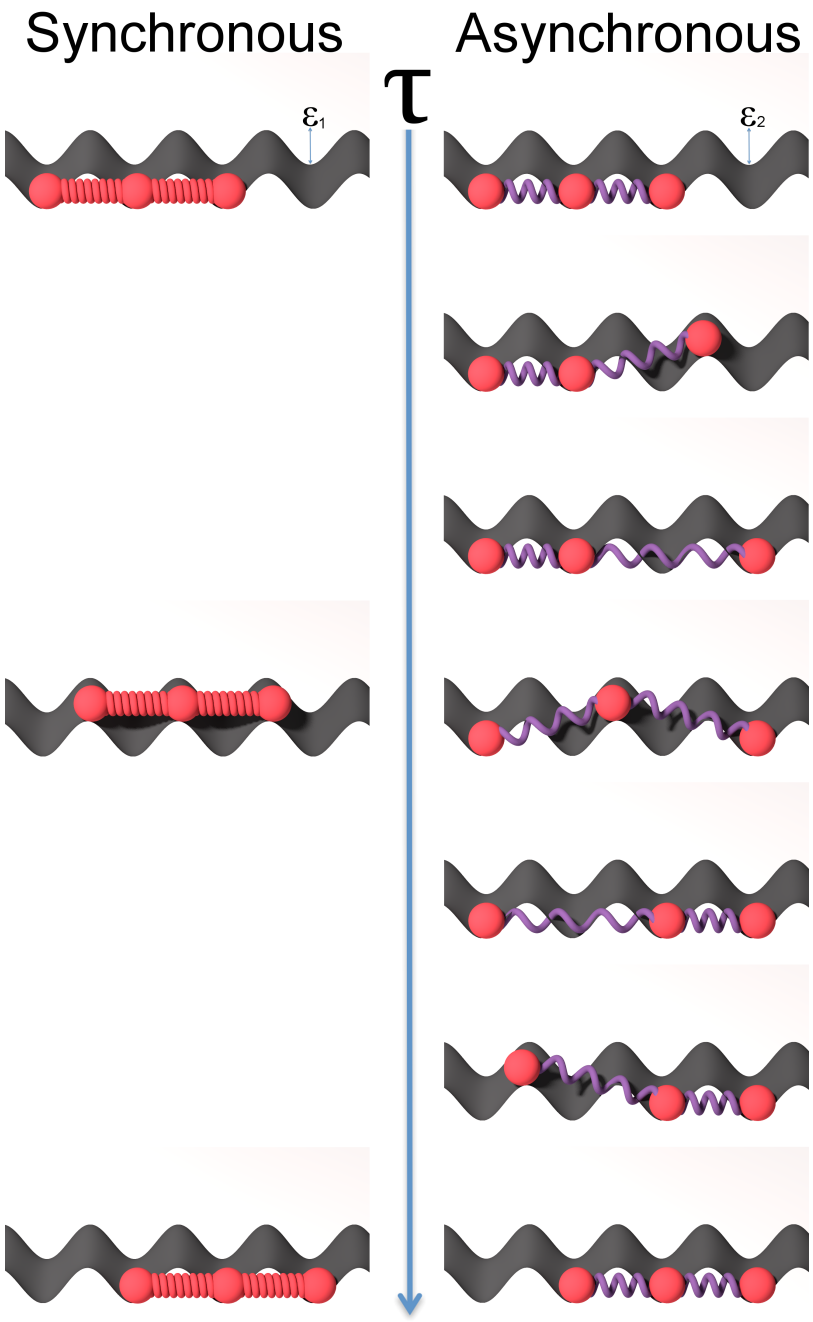


Figure 6.9. Differential coupling between modules adapts sliding speed: synchronous sliding (**Model a**) and asynchronous sliding (**Model b**). Modules are depicted as red spheres connected by springs. Each contact energy, ε, is represented by a well in the energy landscape (grey sinusoidal ribbon) arising from thermally oscillating protein and DNA. **τ** registers time. For synchronous sliding, the modules are perfectly coupled, represented by stiff (red) spring connecting each module. To slide, all three modules must simultaneously break, requiring 3ε of energy. For asynchronous sliding, the modules are weakly coupled, represented by soft (purple) springs connecting each module. Again the energy barrier to shift each contact is ε, except due to weak coupling (<< ε) the total energy barrier for sliding is ~ ε, because each module shifts semi-autonomously.

6.4. Conclusions

A modular transcription factor sliding on DNA was simulated in three 7 μs, unrestrained atomistic MD trajectories. With no external gradient applied, MTERF1 spontaneously slid multiple base pairs. MTERF1 modules lent radial and axial asymmetric contacts to the two DNA strands and along the DNA sequence respectively. These asymmetric contacts shifted asynchronously during sliding. MTERF1 slid in steps (ns-timescale), followed by stationary periods (μs-timescale). The superhelical pitch of MTERF1 – a global metric of the relative position of the modules – correlated with MTERF1 moving along DNA, suggesting that thermal fluctuations are harnessed for sliding. These loosely coupled contacts and modularity in MTERF1 structure may smooth the sliding landscape while maintaining the ability to carefully interrogate bases. It may be of interest to the design of genome editing reagents, high-affinity antibodies and all-new biomaterials to study the role of modularity and asymmetry in binding dynamics.

Chapter 7. Summary and overall impact

The goal of my dissertation was to develop a universal model of protein-DNA binding, search and recognition by characterizing protein conformational change in the natural geometric frame for this problem: A helix. A universal model of protein-DNA binding, search and recognition is essential for understanding gene regulation, the origins of genetic disease and the evolution of gene sequences, and for the design of highly specific genome editing enzymes. My dissertation has four parts: First, I introduce the functional importance of DNA structure (**Chapter 2**); Second, I develop a mathematical tool to study the superhelix geometry of proteins that can bind DNA (**Chapter 3**); Third, I apply the new helix method to study specificity in protein-DNA complexes (**Chapter 4**) and the switch between specific and non-specific binding (**Chapter 5**); Fourth, I propose a mechanism of search based on the semi-autonomous dynamics of the parts of the protein comprising the superhelix (**Chapter 6**).

In **Chapter 2**, the two codes of DNA were reviewed. The genetic code is based on DNA sequence information. While information can be used to describe a variety of phenomena (including energy and structure), it is traditionally used in genetics to study the sequence of DNA letters contained with a sequence. Proteins read sequence information utilizing the mechanism of direct readout. A second code of DNA is based on the energy of a DNA sequence, which governs its structure, flexibility and deformability. This is the energy code of DNA. Regulatory proteins like transcription factors read the energy code using the mechanism of indirect readout. The DNA energy code is important because it can be more highly conserved than sequence information, e.g. nucleotides change (not conserved) but shape does not (conserved). Whether a

regulatory protein binds a specific sequence can also be better predicted by the DNA energy code than by the DNA information code. This Chapter represents a step towards a new facet of genetics in which evolutionary driving forces might be quantified by measuring relative changes in thermodynamic stability of DNA sequences, more accurate models of gene regulation and genetic disease, and the development of new research areas including physical genetics.

Nucleic acid-binding proteins can have a superhelical tertiary structure complementary to the helical curve of DNA. In **Chapter 3**, I developed a novel approach to calculate the helical parameters of nucleic acid-binding protein tertiary structures. The method is based on a simple idea – a helix appears to be a circle when viewed down its helical axis. This idea can be posed mathematically, as finding the plane onto which a helix projects a circle. The helical axis is the normal vector that rises from the plane at the point of the projected circle's center. As with other helix analysis methods, finding the helical axis is the most important step in defining a helix; calculating the remaining parameters of the helix (pitch and radius) is trivial if the helical axis is known. Finding the helical axis using projections simplifies the fitting problem from 3D to 2D. This simplification affords an immediate advantage because solving a least squares problem for a circle is much simpler than solving a total least squares problem for a cylinder. The second strength of the method was its ability to rotate a 3D coordinate system using only 2 parameters. The method can accurately characterize the helical parameters of diverse geometries - from helical secondary structures, to double- and single-stranded nucleic acids, to superhelical tertiary structures. In addition, we performed a broad analysis of 399,360 helices generated with varying shapes, sizes and applied coordinate perturbations. These diverse helices with noisy coordinates mimic the geometries and potential uncertainties found in a biomolecular helix. Future work might explore the ability to automatically select superhelical residues in protein tertiary

structures, to find helical secondary structure elements in proteins and implement high-order optimization procedures to analytically fit the helical axis. It might prove helpful to consider fitting constraints – such as imposing constraints on reasonable orientations along which a helix axis might lay (e.g. within 45° of the vector connecting the Cα or P atoms of the ith and i+6th or i+9th atom, respectively, in a chain, or constraining solutions based on expected radius or pitch). It might be possible to use the information contained in the singular values obtained from the fitting to estimate the parameters of axis-asymmetric helices (e.g. bent or cone-shaped). The mathematics developed in this Chapter can be applied to geometries beyond the cylindrical helix; as long as the geometry contains an axis of symmetry (e.g. Cn, **Table VI-2** see **Appendix VI**), our 2D system can be used to define the 3D parameters for the system of points given a known functional form.

In **Chapter 4**, I used the helix analysis method discussed above to characterize helix complementarity in protein-nucleic acid structures. Understanding helix complementarity in protein-nucleic acid structures is important for developing models of structure-function relationships, and characterizing the subtle distortions that might be important for binding specificity. Because indirect readout is essential for protein-DNA binding specificity (**Chapter 3**), the ability to characterize the distortions in the protein that complement the distortions in the DNA is critical to determining mechanisms of sequence-specific binding. Beyond the biological consequences of advanced mechanisms of protein-nucleic acid binding specificity, like understanding gene expression, understanding specificity in protein-DNA and protein-RNA binding is central to genome editing reagents such as transcription activator like effector-nucleases (TALEN) and clustered regularly interspaced short palindromic repeats (CRISPR). The major result of this section was that the helical geometries of proteins and nucleic acids were

irregular but complementary. This was an important result because no previous analysis has been able to show that superhelical nucleic acid binding proteins are irregular, and that the irregularities complement those in the nucleic acid to which they were bound. Future work might look to utilize, rather than design-around, asymmetrical protein-DNA interactions with asymmetrical affinities. Such a design principle may also be applicable in drug-discovery, but with the geometry of the functional reference vector adapted from a helical axis to an alternate principal axis aligned along the substrate reaction mode (e.g. a unit normal of an SN2 reaction).

In **Chapter 5**, I presented a model of the human mitochondrial transcription termination factor MTERF1 in search mode. Developing a model of MTERF1 in search mode is important for understanding MTERF1 function and mitochondrial genetics. MTERF1 is the canonical mitochondrial transcriptional regulator; however, its precise mechanism of action is hotly debated. Thus, it was essential to understand how MTERF1 binds and recognizes DNA in order to better understand its function. MTERF1 is an exemplary DNA-binding protein. Previous X-ray crystallography studies revealed that MTERF1 has a modular architecture with a superhelical tertiary structure, it utilizes direct readout and it binds the major groove of kinked and unwound DNA. The key property of MTERF1 that made it an exemplary DBP is its superhelical architecture: the helical spiral of DNA is precisely matched by the superhelical spiral of MTERF1. In recognition mode, the partnered spirals of MTERF1 and DNA are unwound. In search mode, MTERF1 should be able to bind nonspecific DNA, the canonical structure of which is B-DNA. Compared with the unwound DNA in the specific complex, B-DNA has a low helical pitch. Thus MTERF1 should adopt a superhelical conformation with the same low pitch as B-DNA. To test this hypothesis, I used MD simulations to generate an ensemble of unbound MTERF1 conformations. I measured MTERF1 superhelical pitch and found that the protein

spontaneously adopted structures with the same low pitch as B-DNA. To generate nonspecific MTERF1-DNA complexes, I docked these low pitch structures to B-DNA. Unlike the conformation of MTERF1 in recognition mode, these low pitch MTERF1 structures were able to bind the major groove of B-DNA. To show that these nonspecific complexes were stable, I ran extensive atomistic MD equilibration on each of the twelve productive poses. Without external biasing potentials applied in to the MD simulations, MTERF1 began spontaneously sliding on DNA. Future work might utilize the straightforward techniques used here to study the search mode of other DBPs, such as the transcription activator like effectors that are structurally homologous to MTERF1. In addition, future work might compare the ability of MTERF1 to adapt superhelical conformation when key residues are mutated, such as the Pro residues used to define the superhelix are mutated. Here, I showed that a simple model of binding based on helix geometry was sufficient to characterize, for the first time, the conformational switch from search to recognition mode. The model should be applicable to any DBP, because DNA is always a participant. A similar approach could be used in the development of binding models in for substrates other than DNA; all that is required is a functional reference vector describing the principle mode of action to cast the appropriate geometric frame.

In **Chapter 6**, the mechanistic details of MTERF1 sliding along DNA were studied. Characterizing the mechanism of sliding is important because DBPs must bind the target sequence to operate, and sliding accelerates the binding rate. In addition, sliding involves nonspecific protein-DNA interactions and is highly dynamic because the protein is moving rapidly. These factors – nonspecific binding and metastability – pose a significant challenge because the dynamics are too fast for atomic-level experiments but have been too slow for atomic-level MD simulations; high resolution experiments like X-ray crystallography and NMR

are ideal tools for the characterization of atomistic structures, but not the atomistic dynamics characteristic of search mode. To overcome this challenge, I utilized unrestrained MD simulations on timescales appropriate for sliding (~ 10 μs). These simulations represent an order-of-magnitude greater sampling than previous studies by other groups, and two-fold greater sampling than my previous work (**Chapter 5**). The ability of MTERF1 to form a nonspecific complex with B-DNA was directly related to the degree of structural similarity between the geometry of the protein and DNA. In recognition mode, MTERF1 adopted a high helical pitch conformation complementary to the unwound (high pitch) target DNA. In search mode, MTERF1 adopted a low helical pitch conformation complementary to canonical B-DNA (low pitch). Because a change in helical pitch changes the length of the MTERF1 – in search mode a change in pitch amounts to extending the footprint on the DNA – I hypothesized that sliding involved oscillations in superhelical pitch, extending and contracting the length of the footprint on the DNA so that MTERF1 could mechanically move along DNA. To test the hypothesis, I analyzed the structure and dynamics of my three new simulations in which MTERF1 was sliding multiple base pairs. First I measured the movement of MTERF1 along DNA at the level of individual amino acid-nucleotide interactions. In doing so, I found a novel set of nonspecific MTERF1-DNA contacts involving eight Arg residues – two from each of the four N-terminal modules. One Arg from each of the four pairs contacted one DNA strand while the second Arg from each of the four pairs contacted the other DNA strand. Sliding was defined as the shift of these Arg residues along the DNA backbone, from one phosphate group to the next along the sequence. The modular nature of these interactions permitted the contacts to shift along the DNA sequence in a semi-autonomous fashion. This is an important result. By taking small steps, the energy barrier to slide the whole protein from one register of DNA contacts to the next is roughly

the energy barrier to break a single contact. Contrast this scenario with the alternate case in which all contacts break and form simultaneously: to slide the whole protein from one register of DNA contacts to the next is the *sum* of the energy barriers of all contacts. A key implication of the model proposed in **Chapter 6** is that individual modules or domains of a DBP in search mode can interrogate a subset of nucleotides in the sequence while maintaining an overall rapid sliding regime. Future work might look to utilize this asymmetrical design principle to develop materials and therapeutics with switchable mechanical-functional properties such as toggling between tight-binding and fast-diffusion (nanomachines or genome editing reagents), toxicity and bio-availability (protein-based therapeutics or ligand-sequestration until target-bound), or compound encapsulation and release (drug-delivery or pollutant sequestration).

My dissertation began with an introduction of the functional importance of sequence-dependent DNA structure (**Chapter 2**). Next, I developed a geometric tool to characterize protein structure in the frame of a helix (**Chapter 3**). I studied the relation between specificity and protein-DNA helix complementarity (**Chapter 4**) and I characterized the switch in protein-DNA helix complementarity from specific and non-specific (search) binding modes (**Chapter 5**). Last, I studied how the autonomous dynamics of protein modules in search mode smooth's the energy landscape (**Chapter 6**). A mathematical framework was developed that simplified a 3D geometry into a 2D space without losing information about the original 3D object. A simple binding model was developed by casting the functional modes of the protein in the frame of the substrate; as protein function often revolves around acting on a substrate, defining a functional reference vector on the substrate rather than the protein, a potentially universal mode of reaction may have been found. Lastly, I found that structural discretization by a modular protein

architecture permits the simplification of functionally relevant dynamics in its native energy landscape.

Appendix I. General MD simulation protocols

The goal of **Appendix I** is to provide sample protocols that can be used (or easily modified) to prepare the parameters and coordinates of a molecular model for Amber MD simulations (**Appendix I.i**), to equilibrate the model (**Appendix I.ii**), and to impose distance restraints (**Appendix I.iii**).

I.i. Prepare Amber parameter and coordinate files using LEaP

This script automates the procedure of generating explicit solvent systems, with explicit ions, with four explicit solvent box sizes (first **for loop**) and four ion concentrations (second **for loop**), and randomizes the positions of those ions.

```
#!/bin/bash
for box in $(echo 8 12 16 20); do
for salt in $(echo 0.1 0.2 0.4 0.8); do
mkdir -p ../tip3p_parms
mkdir -p ../tip3p_rsts
mkdir -p ../tip3p_box
run=run0
#<<'comment2'
#---first build a box to estimate its size primum...
cat > leap.in << EOF
source leaprc.ff14SB
mol = loadpdb ./1kis.prep.pdb
loadoff ./ions08.lib
mods = loadamberparams ./frcmod.ionsjc_tip3p
set default PBradii mbondi3
solvateOCT mol TIP3PBOX $box
addionsRand mol K+ 0 6.0
saveamberparm mol ../tip3p_box/tT.salt_$salt.box_$box.parm7 ../tip3p_box/tT.salt_$salt.box_$box.rst7
quit
EOF
tleap -f leap.in
#comment2
#---now estimate (quite well) how many ions we need to get a conc..
dim=`tail -n 1 ../tip3p_box/tT.salt_$salt.box_$box.rst7`
./Toms_skrip.pl $salt $dim > temp
mol=`grep MOLAR temp|awk '{print $3}'`
echo mole: $mol
vol=`grep Volum temp|awk '{print $3}'`
echo volume: $vol
kion=`grep molec temp|awk '{print int($1)}'`
ion=`echo $kion + $kion + 30|bc -l` #two ions per kion, PLUS 30 neutralizing...
echo ions: $ion
#<<'comment1'
#---NOW add them ions...
cat > leap.in << EOF
source leaprc.ff14SB
mol = loadpdb ./1kis.prep.pdb
loadoff ./ions08.lib
mods = loadamberparams ./frcmod.ionsjc_tip3p
set default PBradii mbondi3
solvateOCT mol TIP3PBOX $box
addionsRand mol K+ 0 6.0
`for ((i=1;i<$ion+1;i++));  do echo addionsRand mol K+ 1 Cl- 1 6.0; done`
saveamberparm mol ../tip3p_parms/tT.salt_$salt.box_$box.parm7 ../tip3p_rsts/tT.salt_$salt.box_$box.rst7
quit
EOF
tleap -f leap.in
#comment1
#---randomize IONS and redistribute IONS far away from SOLUTE and other IONS....
echo Randomizing ions in directory search.$run.$frame.$dna
R=1-32
cpptraj -p ../tip3p_parms/tT.salt_$salt.box_$box.parm7 << EOF
trajin ../tip3p_rsts/tT.salt_$salt.box_$box.rst7
```

```
trajout ../tip3p_rsts/tT.salt_$salt.box_$box.rand.rst7 restart
prnlev 3
randomizeions @K+,Cl- around :$R by 6.0 overlap 4.0 noimage
EOF
#---
done
done
```

I.ii. Generate Amber MD input parameter for equilibration with distance restraints

This bash script automatically sets up Amber "mdin" files for use in the simulation of explicit ions in Amber GB MD simulations (see **Section 6.2**). The first **for loop** passes over box sizes - 8 Å, 12 Å, 16 Å and 20 Å - and the second **for loop** passes over four salt concentrations - 0.1 M, 0.2 M, 0.4 M and 0.8 M. The third line obtains the explicit solvent box size (length, width, height in Å of the box, and the 3 angles of the truncated octahedron α, β and γ in degrees of the box), which is passed as input to Tom Cheatham III's perl script (see below) which calculates the number of explicit ions required to achieve a desired concentration for a box size.

For this particular example (Tar-Tar* RNA kissing loop), the solute has 32 nucleotides; thus, the **restraintmask** in the **mdin** files below use ":1-32".

```
#!/bin/bash
for box in $(echo 8 12 16 20); do
for salt in $(echo 0.1 0.2 0.4 0.8); do
dim=`tail -n 1 ../tip3p_rsts/tT.salt_$salt.box_$box.rand.rst7`
./Toms_skrip.pl $salt $dim > temp.$salt.$box
mol=`grep MOLAR temp.$salt.$box|awk '{print $3}'`
#echo mole: $mol
vol=`grep Volum temp.$salt.$box|awk '{print $3}'`
#echo volume: $vol
kion=`grep molec temp.$salt.$box|awk '{print int($1)}'`
echo kion: $kion
ion=`echo $kion + $kion + 30|bc -l` #two ions per kion, PLUS 30 neutralizing...
echo tT.salt_$ion.box_$box.gas.pdb
echo ions: $ion
echo -------
mkdir -p ../mdins/mdins.salt_$ion.box_$box
run=run0
cat > ../mdins/mdins.salt_$ion.box_$box/1min.in << EOF
kTest
&cntrl
       imin = 1, maxcyc=1000,
       ntx = 1,
       ntwr = 100, ntpr = 100,
       cut = 999.0,
       ntb=0, igb = 8,  saltcon = 0.0,
       ntr=1, restraint_wt=10, restraintmask = ":1-32 & !@H=",
       nmropt=1,
       ntxo=2,ioutfm=1,
/
&wt TYPE="DUMPFREQ", istep1=1000 /
/
&wt type='END', /
    DISANG=../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
    DUMPAVE=ion_shell_dist.salt_$ion.box_$box.out
    LISTIN=POUT
    LISTOUT=POUT
EOF
cat > ../mdins/mdins.salt_$ion.box_$box/2md.in << EOF
kTest
&cntrl
       imin = 0, nstlim = 500000, dt = 0.001,
       ntx = 1, irest=0,
       ntt=3, tempi = 100.0, temp0 = 300.0, gamma_ln=1.,
       ntc = 2, ntf = 2,
       ntwx =  500, ntwe = 0, ntwr = 500, ntpr = 500,
       cut = 999.0,
       ntb=0, igb = 8,  saltcon = 0.0,
       ntr=1, restraint_wt=10, restraintmask = ":1-32 & !@H=",
       nmropt=1,
       ntxo=2,ioutfm=1,
/
&wt TYPE="DUMPFREQ", istep1=1000 /
/
&wt type='END', /
    DISANG=../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
    DUMPAVE=ion_shell_dist.salt_$ion.box_$box.out
    LISTIN=POUT
```

```
    LISTOUT=POUT
/
&wt
   TYPE="TEMP0",istep1=0, istep2=250000,
   value1=100., value2=300.,
/
&wt
   TYPE="TEMP0",istep1=250001, istep2=500000,
   value1=300., value2=300.,
/
&wt
   TYPE="END",
/
EOF
cat > ../mdins/mdins.salt_$ion.box_$box/3min.in << EOF
kTest
&cntrl
       imin = 1, maxcyc=1000,
       ntx = 1,
       ntwr = 100, ntpr = 100,
       cut = 999.0,
       ntb=0, igb = 8,  saltcon = 0.0,
       ntr=1, restraint_wt=10,
       restraintmask=":1-32 & @CA,N,C,N1,C6,C5,N7,C8,N9,C4,N3,C2,O5',C5',C4',C3',O3',O2P,O1P,P,O4',C2',C1'",
       nmropt=1,
       ntxo=2,ioutfm=1,
/
&wt TYPE="DUMPFREQ", istep1=1000 /
/
&wt type='END', /
    DISANG=../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
    DUMPAVE=ion_shell_dist.salt_$ion.box_$box.out
    LISTIN=POUT
    LISTOUT=POUT
EOF
cat > ../mdins/mdins.salt_$ion.box_$box/4md.in << EOF
kTest
&cntrl
       imin = 0, nstlim = 500000, dt = 0.001,
       ntx = 1, irest=0,
       ntt=3, tempi = 100.0, temp0 = 300.0, gamma_ln=1.,
       ntc = 2, ntf = 2,
       ntwx =  500, ntwe = 0, ntwr = 500, ntpr = 500,
       cut = 999.0,
       ntb=0, igb = 8,  saltcon = 0.0,
       ntr=1, restraint_wt=10.,
       restraintmask=":1-32 & @CA,N,C,N1,C6,C5,N7,C8,N9,C4,N3,C2,O5',C5',C4',C3',O3',O2P,O1P,P,O4',C2',C1'",
       nmropt=1,
       ntxo=2,ioutfm=1,
/
&wt TYPE="DUMPFREQ", istep1=1000 /
/
&wt type='END', /
    DISANG=../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
    DUMPAVE=ion_shell_dist.salt_$ion.box_$box.out
    LISTIN=POUT
    LISTOUT=POUT
/
&wt
   TYPE="TEMP0",istep1=0, istep2=250000,
   value1=100., value2=300.,
/
&wt
   TYPE="TEMP0",istep1=250001, istep2=500000,
   value1=300., value2=300.,
/
&wt
   TYPE="END",
/
EOF
cat > ../mdins/mdins.salt_$ion.box_$box/5md.in << EOF
kTest
&cntrl
       imin = 0, nstlim = 500000, dt = 0.001,
       ntx = 5, irest=1,
       ntt=3, temp0 = 300.0, gamma_ln=1.,
       ntc = 2, ntf = 2,
       ntwx =  500, ntwe = 0, ntwr = 500, ntpr = 500,
       cut = 999.0,
```

```
       ntb=0, igb = 8,  saltcon = 0.0,
       ntr=1, restraint_wt=1.,
       restraintmask=":1-32 & @CA,N,C,N1,C6,C5,N7,C8,N9,C4,N3,C2,O5',C5',C4',C3',O3',O2P,O1P,P,O4',C2',C1'",
       nmropt=1,
       ntxo=2,ioutfm=1,
/
&wt TYPE="DUMPFREQ", istep1=1000 /
/
&wt type='END', /
    DISANG=../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
    DUMPAVE=ion_shell_dist.salt_$ion.box_$box.out
    LISTIN=POUT
    LISTOUT=POUT
EOF
cat > ../mdins/mdins.salt_$ion.box_$box/6md.in << EOF
kTest
&cntrl
       imin = 0, nstlim = 500000, dt = 0.001,
       ntx = 5, irest=1,
       ntt=3, temp0 = 300.0, gamma_ln=1.,
       ntc = 2, ntf = 2,
       ntwx =  500, ntwe = 0, ntwr = 500, ntpr = 500,
       cut = 999.0,
       ntb=0, igb = 8,  saltcon = 0.0,
       ntr=1, restraint_wt=0.1,
       restraintmask=":1-32 & @CA,N,C,N1,C6,C5,N7,C8,N9,C4,N3,C2,O5',C5',C4',C3',O3',O2P,O1P,P,O4',C2',C1'",
       nmropt=1,
       ntxo=2,ioutfm=1,
/
&wt TYPE="DUMPFREQ", istep1=1000 /
/
&wt type='END', /
    DISANG=../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
    DUMPAVE=ion_shell_dist.salt_$ion.box_$box.out
    LISTIN=POUT
    LISTOUT=POUT
EOF
cat > ../mdins/mdins.salt_$ion.box_$box/7md.in << EOF
kTest
&cntrl
       imin = 0, nstlim = 5000000, dt = 0.002,
       ntx = 5, irest=1,
       ntt=3, temp0 = 300.0, gamma_ln=1.,
       ntc = 2, ntf = 2,
       ntwx =  2500, ntwe = 0, ntwr = 2500, ntpr = 2500,
       cut = 999.0,
       ntb=0, igb = 8,  saltcon = 0.0,
       ntr=0, nscm=0,
       nmropt=1,
       ntxo=2,ioutfm=1,
/
&wt TYPE="DUMPFREQ", istep1=1000 /
/
&wt type='END', /
    DISANG=../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
    DUMPAVE=ion_shell_dist.salt_$ion.box_$box.out
    LISTIN=POUT
    LISTOUT=POUT
EOF
cat > ../mdins/mdins.salt_$ion.box_$box/prod.in << EOF
kTest
&cntrl
       imin = 0, nstlim = 25000000, dt = 0.002,
       ntx = 5, irest=1,
       ntt=3, temp0 = 300.0, gamma_ln=1.,
       ntc = 2, ntf = 2,
       ntwx =  2500, ntwe = 0, ntwr = 2500, ntpr = 2500,
       cut = 999.0,
       ntb=0, igb = 8,  saltcon = 0.0,
       ntr=0, nscm=0,
       nmropt=1,
       ntxo=2,ioutfm=1,
/
&wt TYPE="DUMPFREQ", istep1=1000 /
/
&wt type='END', /
    DISANG=../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
    DUMPAVE=ion_shell_dist.salt_$ion.box_$box.out
    LISTIN=POUT
```

```
    LISTOUT=POUT
EOF
done
done
```

I.iii. Generate distance restraints for the above GB-ions MD

This script generates the distance restraint parameters used in the above GB-ions MD simulations.

```
#!/bin/bash
############ assuming you've already run the run_leap script...
##
mkdir -p ../ion_DistRst
for box in $(echo 8 12 16 20); do
for salt in $(echo 0.1 0.2 0.4 0.8); do
#====
dim=`tail -n 1 ../tip3p_rsts/tT.salt_$salt.box_$box.rst7`
./Toms_skrip.pl $salt $dim > temp
mol=`grep MOLAR temp|awk '{print $3}'`
vol=`grep Volum temp|awk '{print $3}'`
kion=`grep molec temp|awk '{print int($1)}'`
ion=`echo $kion + $kion + 30|bc -l`
echo ions: $ion
#echo mol $mol
#echo vol $vol
#echo ion $ion
# solve for radius...
pi43=`echo "4*3.1415926535/3"|bc -l`
rad=`echo "e(l($pi43 * $vol)/3)"|bc -l` #this is how to cube-root in bash...
#
################ now generate Amber rst files...
#provide a PDB file of JUST the ATOMS you want to be used to define your COM
kpdb=kTarTar.P_atom.pdb
#how many solute ATOMS in COM mask?
resid=`grep ATOM $kpdb|wc -l`
echo ATOM= $resid
#resid of first ion
first=1029 # the last atom of solute is 1028, see kTarTar.AllAtom.pdb
#resid of last ion
last_ion=`echo "$first + $ion"|bc`
echo first_ion: $first
echo last_ion:  $last_ion
#distance where parabolic restraint starts (near actual sphere cut)
firstcut=`echo $rad/2|bc -l`
#distance where parabola rises maximally
lastcut=$rad
#strength (slope) of the droplet cutoff restraint
cutforce=1000
#-------------------------------------------------------------------------------
last=$((last_ion-1))
#===============================================================================
# quick digression, need to generate iat for protein, since need COM...
ksol=($(head -n $resid $kpdb|awk '{print $2}'))
#echo ${ksol[0]}
#array lengths, for looping
p1=`echo ${#ksol[@]}`
j=0
for ((i=0;i<$p1;i++)); do
j=$((j+1))
igr1[$i]=`echo IGR1'('$j')'=${ksol[$i]},`
done
#echo ${igr1[@]}
kIGRs=`echo ${igr1[@]}`
#===============================================================================

# ok, now actually generate the Amber rst file....
rm ../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst
for ion_idx in $(seq $first $last); do
fkion=$((ion_idx+1))
cat >> ../ion_DistRst/ion_shell_dist.salt_$ion.box_$box.rst << EOF
&rst
iat=-1,$fkion, r1=0, r2=$firstcut, r3=$firstcut, r4=$lastcut, rk2=$cutforce, rk3=$cutforce,
$kIGRs
/
EOF
done
echo "---"
#====
done
echo "-------"
done
```

```
#and clean up that last comma
sed -i 's/1027,/1027/g' ../ion_DistRst/ion_shell_dist.salt_*.box_*.rst
```

I.iii.i. Calculate the number of explicit ions required to achieve a desired concentration

This perl script was written by Thomas E. Cheatham III. In the automation scripts above, the name of this script must be "**Toms_skrip.pl**".

```
#!/usr/bin/perl
$molar = $ARGV[0];
$box_x = $ARGV[1];
$box_y = $ARGV[2];
$box_z = $ARGV[3];
$box_a = $ARGV[4];
$box_b = $ARGV[5];
$box_g = $ARGV[6];
if ($box_x == 0) {
   printf("Usage: molarity box-x box-y box-z alpha beta gamma\n");
   die;
}
if ($box_y == 0) {
   $box_y = $box_x;
}
if ($box_z == 0) {
   $box_z = $box_x;
}
if ($box_a == 0) {
   $box_a = 90.0;
}
if ($box_b == 0) {
   $box_b = $box_a;
}
if ($box_g == 0) {
   $box_g = $box_a;
}
$torad = 2 * 3.141592654 / 360.0;
$rad_a = $box_a * $torad;
$rad_b = $box_b * $torad;
$rad_g = $box_g * $torad;
$angles = 1 - cos($rad_a)*cos($rad_a) - cos($rad_b)*cos($rad_b) - cos($rad_g)*cos($rad_g);
$angles += 2 * cos($rad_a)*cos($rad_b)*cos($rad_g);
$angles = sqrt($angles);
$volume = $box_x * $box_y * $box_z * $angles;
$molecules = 6.022 * $volume * $molar / 10000;
printf("  MOLARITY = %8.3f\n", $molar);
printf("  Box size = %8.3f %8.3f %8.3f %8.3f %8.3f %8.3f\n", $box_x,
$box_y, $box_z, $box_a, $box_b, $box_g);
printf("  Volume   = %8.3f\n", $volume);
printf("\n  %8.3f molecules are necessary to make a molarity of %6.2fM\n\n", $molecules, $molar);
```

I.iv. Multi-stage equilibration procedure: job submission to Slurm

This bash script automates the submission of multiple equilibration jobs to a cluster that utilizes the Slurm queuing system. Notice that the minimization stages below use the CPU version of Amber's pmemd, whereas the MD stages use the GPU version of Amber's pmemd.

```
#!/bin/bash
pmemdCPU=/opt/amber/bin/pmemd
pmemd=/opt/amber/bin/pmemd.cuda
for j in $(cat Dock.list); do
mkdir $j
cd $j
########################################
sbatch << EOF
#!/bin/bash
#SBATCH --gres=gpu:1
#SBATCH --job-name pAT.$j
#SBATCH --partition=all
$pmemdCPU  -O  -i  ../MDINS/1min.in  -p  ../000.build/LEaP/final-crds/$j.parm7  -c  ../000.build/LEaP/final-crds/$j.rand.rst7 -ref ../000.build/LEaP/final-crds/$j.rand.rst7 -o ./1min.out -x ./1min.x -inf ./1min.info -r ./1min.r
$pmemd -O -i ../MDINS/2Md.in -p ../000.build/LEaP/final-crds/$j.parm7 -c ./1min.r -ref ./1min.r -o ./2Md.out -x ./2Md.x -inf ./2Md.info -r ./2Md.r
$pmemd -O -i ../MDINS/3Md.in -p ../000.build/LEaP/final-crds/$j.parm7 -c ./2Md.r -ref ./2Md.r -o ./3Md.out -x ./3Md.x -inf ./3Md.info -r ./3Md.r
$pmemd -O -i ../MDINS/4Md.in -p ../000.build/LEaP/final-crds/$j.parm7 -c ./3Md.r -ref ./3Md.r -o ./4Md.out -x ./4Md.x -inf ./4Md.info -r ./4Md.r
$pmemd -O -i ../MDINS/5min.in -p ../000.build/LEaP/final-crds/$j.parm7 -c ./4Md.r -ref ./4Md.r -o ./5min.out -x ./5min.x -inf ./5min.info -r ./5min.r
$pmemd -O -i ../MDINS/6Md.in -p ../000.build/LEaP/final-crds/$j.parm7 -c ./5min.r -ref ./5min.r -o ./6Md.out -x ./6Md.x -inf ./6Md.info -r ./6Md.r
$pmemd -O -i ../MDINS/7Md.in -p ../000.build/LEaP/final-crds/$j.parm7 -c ./6Md.r -ref ./6Md.r -o ./7Md.out -x ./7Md.x -inf ./7Md.info -r ./7Md.r
$pmemd -O -i ../MDINS/8Md.in -p ../000.build/LEaP/final-crds/$j.parm7 -c ./7Md.r -ref ./6Md.r -o ./8Md.out -x ./8Md.x -inf ./8Md.info -r ./8Md.r
$pmemd  -O  -i  ../MDINS/9Md.NVT.in  -p  ../000.build/LEaP/final-crds/$j.parm7  -c  ./8Md.r  -o  ./9Md.NVT.out  -x ./9Md.NVT.x -inf ./9Md.NVT.info -r ./9Md.NVT.r
EOF
#######
cd ../
####
done
```

I.v. Prepare explicitly solvated molecular models from multiple conformations

The purpose of this protocol is to generate models for four different conformations of a biomolecule, in an explicit solvent, such that the final models all have the same number of water molecules. The key idea is that no model is solvated by an explicit solvent box with a buffer size smaller than 8 Å (the value can be changed in the BASH source file **tleap1.sh**). Importantly, the shell scripts that follow below must be placed in a directory **BASH_SOURCE**, which must reside in the current working directory.

```
!/usr/bin/python
import subprocess; import os.path
from Bio import pairwise2; from Bio.pairwise2 import format_alignment; from Bio.Seq import Seq; from
Bio.Alphabet import IUPAC
from Bio.PDB.Polypeptide import three_to_one
from Bio import AlignIO
#
#  MAKE SURE THAT --- BASH_SOURCE --- IS IN THE CURRENT WORKING DIRECTORY...
#
verbose='0'  #==1, print intermediate info; ==0, carry-on silently...
printalign='!yes'
protein='../T0759/'
server1='FALCON_EnvFold_TS1' #'3D-Jigsaw-V5_1_TS1'
server2='FALCON_EnvFold_TS2' #'3D-Jigsaw-V5_1_TS2'
server3='FALCON_EnvFold_TS3' #'3D-Jigsaw-V5_1_TS3'
server4='FALCON_EnvFold_TS4' #'3D-Jigsaw-V5_1_TS4'
serverX=[server1,server2,server3,server4]
#
#  Stage 0: Check server structures for residue congruence (all models have same number of residues):
#           and align them...
#
numWat,info1,info2,info3,info4=[],[],[],[],[]
with open( os.path.join(protein, server1) ) as input:
  for line in input:
    if " CA " in line:
      i1 = [ item.strip() for item in line.split() ]
      info1.append( three_to_one(i1[3]) )
with open( os.path.join(protein, server2) ) as input:
  for line in input:
    if " CA " in line:
      i2 = [ item.strip() for item in line.split() ]
      info2.append( three_to_one(i2[3]) )
with open( os.path.join(protein, server3) ) as input:
  for line in input:
    if " CA " in line:
      i3 = [ item.strip() for item in line.split() ]
      info3.append( three_to_one(i3[3]) )
with open( os.path.join(protein, server4) ) as input:
  for line in input:
    if " CA " in line:
      i4 = [ item.strip() for item in line.split() ]
      info4.append( three_to_one(i4[3]) )
if printalign=='yes':
  print '-----------------------------------------------------------------'
  print '-------- CHECK FOR MISSING RESIDUES: PAIRWISE2 ALIGN ------------'
  print '-----------------------------------------------------------------'
# align the protein sequences to check for missing stuff:
align1= pairwise2.align.globalxx( ''.join(info1), ''.join(info2) )
for CA in pairwise2.align.globalxx( ''.join(info1), ''.join(info2) ):
  if printalign=='yes':
    print 'Alignment: model 1, model 2:'
    print(format_alignment(*CA))
    print '================================================================='
# align the protein sequences to check for missing stuff:
align2= pairwise2.align.globalxx( ''.join(info1), ''.join(info3) )
for CA in pairwise2.align.globalxx( ''.join(info1), ''.join(info3) ):
  if printalign=='yes':
    print 'Alignment: model 1, model 3:'
    print(format_alignment(*CA))
    print '================================================================='
# align the protein sequences to check for missing stuff:
pairwise2.align.globalxx( ''.join(info1), ''.join(info4) )
for CA in pairwise2.align.globalxx( ''.join(info1), ''.join(info4) ):
  if printalign=='yes':
    print 'Alignment: model 1, model 4:'
```

```
    print(format_alignment(*CA))
    print '============================================================'
# align the protein sequences to check for missing stuff:
pairwise2.align.globalxx( ''.join(info2), ''.join(info3) )
for CA in pairwise2.align.globalxx( ''.join(info2), ''.join(info3) ):
  if printalign=='yes':
    print 'Alignment: model 2, model 3:'
    print(format_alignment(*CA))
    print '============================================================'
# align the protein sequences to check for missing stuff:
pairwise2.align.globalxx( ''.join(info2), ''.join(info4) )
for CA in pairwise2.align.globalxx( ''.join(info2), ''.join(info4) ):
  if printalign=='yes':
    print 'Alignment: model 2, model 4:'
    print(format_alignment(*CA))
    print '============================================================'
# align the protein sequences to check for missing stuff:
pairwise2.align.globalxx( ''.join(info3), ''.join(info4) )
for CA in pairwise2.align.globalxx( ''.join(info3), ''.join(info4) ):
  if printalign=='yes':
    print 'Alignment: model 3, model 4:'
    print(format_alignment(*CA))
#
#  Stage 1: tLEaP
#
if os.path.isfile("./leap.log"):
  os.remove("./leap.log")
subprocess.call(["./BASH_SOURCE/tleap1.sh", protein, server1, server2, server3, server4])
#
#  Stage 2: figure out which model has the MOST water given 8 \AA buffer
#            then add a bunch more WAT (say 20 \aa buffer) to the rest
#              then strip XS water from them to MATCH the first model (the one with MOST)
#
w1=[]
models=[ 'model1', 'model2', 'model3', 'model4' ]; kmodels=models
parms=[ 'temp1/model1.parm7', 'temp1/model2.parm7', 'temp1/model3.parm7', 'temp1/model4.parm7' ]
with open( 'leap.log' ) as input:
  for line in input:
    if "\tWAT\t" in line:
      j1 = [ item.strip() for item in line.split() ]
      w1.append( j1[1] )
  #now figure out which is MOST:
  modelMOST = w1.index(max(w1))
  if verbose=='yes':
    print 'max wat: %s' % max(w1), 'Model: ', w1.index(max(w1))+1
#
#  Stage 3: tLEaP  re-run the 3 smaller models with HUGE (20 \AA) box sizes
#
print modelMOST
del serverX[int(modelMOST)]
del models[int(modelMOST)]
for s, m in zip(serverX, models):
  subprocess.call(["./BASH_SOURCE/tleap3.sh", protein, s, m ])
#
#  Stage 4: CPPtraj  strip XS water from what we made in stage 3
#
#Figure out how many atoms we need to strip:
with open ( parms[int(modelMOST)] ) as f:
  for i, line in enumerate(f):
    if i == 6:
      l1 = [ item.strip() for item in line.split() ]
numstrip=int(l1[0])+1; print numstrip
#now do the stripping...
for m in models:
  print m, numstrip
  subprocess.call(["./BASH_SOURCE/cpptraj4.sh", m, str(numstrip)])
#finally, copy MOST over to 000Models/ where all final parm7/rst7 files at
subprocess.call(["./BASH_SOURCE/CopyMost.sh", kmodels[modelMOST]])
print 'all models have been built and are ready to be equilibrated!'
##########################
#############
########
#####
###
##
#
#
```

The following shell scripts must all be placed in a directory named **BASH_SOURCE**, and the directory must reside in the current working directory of the above protocol.

I.v.i. BASH_SOURCE/tleap1.sh

```
!/bin/bash
#
#  Stage 1: Build ALL systems with 8 \AA buffer to determine which to "over-buffer" in Stage 2...
#
mkdir -p temp1
protein=$1
server1=$2
server2=$3
server3=$4
server4=$5
cat > leap1.in << EOF
source leaprc.ff14SB
mol1=loadPDB $protein/$server1
mol2=loadPDB $protein/$server2
mol3=loadPDB $protein/$server3
mol4=loadPDB $protein/$server4
solvateOCT mol1 TIP3PBOX 8
solvateOCT mol2 TIP3PBOX 8
solvateOCT mol3 TIP3PBOX 8
solvateOCT mol4 TIP3PBOX 8
saveAmberparm mol1 temp1/model1.parm7 temp1/model1.rst7
saveAmberparm mol2 temp1/model2.parm7 temp1/model2.rst7
saveAmberparm mol3 temp1/model3.parm7 temp1/model3.rst7
saveAmberparm mol4 temp1/model4.parm7 temp1/model4.rst7
quit
EOF
tleap -f leap1.in
```

I.v.ii. BASH_SOURCE/getWat2.sh

```
#!/bin/bash
# figure out how many water molecules there are...
grep -B4 "no restraints" $1 |grep WAT|awk '{print $2}'
```

I.v.iii. BASH_SOURCE/tleap3.sh

```
#!/bin/bash
#
#  Stage 3: Build HUGE box so 3 smaller models have more WAT than the model with MOST wat (given 8 \AA buffer)
#
mkdir -p temp2
protein=$1
server1=$2
model=$3
cat > leap1.in << EOF
source leaprc.ff14SB
mol1=loadPDB $protein/$server1
solvateOCT mol1 TIP3PBOX 20
saveAmberparm mol1 temp2/$model.parm7 temp2/$model.rst7
quit
EOF
tleap -f leap1.in
```

I.v.iv. BASH_SOURCE/cpptraj4.sh

```
#!/bin/bash

#
#  Stage 4: Strip the 3 models of their XS WAT so they have EXACT SAME numWAT as the model with the MOST
#
mkdir -p 000Models
model=$1
numstrip=$2
cpptraj << EOF
parm temp2/$model.parm7
trajin temp2/$model.rst7
strip @$numstrip-999999
trajout 000Models/$model.rightWAT.rst7 restart
EOF
cpptraj << EOF
parm temp2/$model.parm7
parmstrip @$numstrip-999999
parmwrite out 000Models/$model.rightWAT.parm7
EOF
```

I.v.v. BASH_SOURCE/CopyMost.sh

```
!/bin/bash
# move files around
most=$1
cp temp1/$most.parm7 000Models/
cp temp1/$most.rst7 000Models/
```

Appendix II. Protocols for PNEB-MMPBSA

II.i. Explicit solvent equilibration & mult-stage PNEB simulations

This is an advanced protocol because it is designed to fully automate the preparation of six PNEB simulations between four end-points: (1) generate an organized structure of working directories; (2) prepare a nine-stage equilibration execution file for a SLURM queuing system (per **Table 3.3**); (3) prepare ten-stage PNEB groupfiles (six for the six paths); (4) generate the Amber MD equilibration input files (per **Table 3.3**, where group A is "CA,N,C" and group B is "CA" because in this example server models from CASP11 were being optimized); (5) generate Amber PNEB MD input files.

```
#!/usr/bin/python
from __future__ import print_function
import os
# After having optimized the NEB path to ZERO Kelvin, do you want to warm it back up to 300K (for umbrella
sampling or something)?
warmBackUp='yes'
numBeads='16'
protein='T0759'
#
#  Stage 1: Write EQUILIBRATION stage list (ADD your AMBERHOME and SGE/SLURM flags)
#
modelist=[ 'model1', 'model2', 'model3', 'model4' ]
pathlist=[ '1st2nd', '1st3rd', '1st4th', '2nd3rd', '2nd4th', '3rd4th' ]
direlist=[ '001--1st2nd',  '002--1st3rd', '003--1st4th', '004--2nd3rd', '005--2nd4th', '006--3rd4th']

if not os.path.isdir('./eQUILIB-'+protein):
  os.makedirs('./eQUILIB-'+protein)
for mod in modelist:
  if not os.path.isdir('./eQUILIB-'+protein+'/'+mod):
    os.makedirs('./eQUILIB-'+protein+'/'+mod)
  try:
    os.remove('run_equilibration.'+mod+'.sh')
  except OSError:
    pass
  with open('run_equilibration.'+mod+'.sh', 'a') as f:
    f.write("#!/bin/bash\npmemdC=/opt/amber/bin/pmemd\npmemdG=/opt/amber/bin/pmemd.cuda\ncd   "+mod+"\n\nsbatch
<<      EOF\n#!/bin/bash\n#SBATCH      --gres=gpu:1\n#SBATCH      --job-name      CASP11."+mod+"\n#SBATCH      --
partition=all\n\n$pmemdC      -O      -i      ../mdins/1min.in      -p      ../../000Models/"+mod+"*parm7      -c
../../000Models/"+mod+"*rst7 -ref ../../000Models/"+mod+"*rst7 -o ./1min.out -x ./1min.x -inf ./1min.info -r
./1min.rst7\n$pmemdG -O -i ../mdins/2Md.in -p ../../000Models/"+mod+"*parm7 -c ./1min.rst7 -ref ./1min.rst7 -o
./2Md.out    -x    ./2Md.x    -inf    ./2Md.info    -r    ./2Md.rst7\n$pmemdG    -O    -i    ../mdins/3Md.in    -p
../../000Models/"+mod+"*parm7  -c  ./2Md.rst7  -ref  ./2Md.rst7  -o  ./3Md.out  -x  ./3Md.x  -inf  ./3Md.info  -r
./3Md.rst7\n$pmemdG -O -i ../mdins/4Md.in -p ../../000Models/"+mod+"*parm7 -c ./3Md.rst7 -ref ./3Md.rst7 -o
./4Md.out    -x    ./4Md.x    -inf    ./4Md.info    -r    ./4Md.rst7\n$pmemdC    -O    -i    ../mdins/5min.in    -p
../../000Models/"+mod+"*parm7  -c  ./4Md.rst7  -ref  ./3Md.rst7  -o  ./5min.out  -x  ./5min.x  -inf  ./5min.info  -r
./5min.rst7\n$pmemdG -O -i ../mdins/6Md.in -p ../../000Models/"+mod+"*parm7 -c ./5min.rst7 -ref ./5min.rst7 -o
./6Md.out    -x    ./6Md.x    -inf    ./6Md.info    -r    ./6Md.rst7\n$pmemdG    -O    -i    ../mdins/7Md.in    -p
../../000Models/"+mod+"*parm7  -c  ./6Md.rst7  -ref  ./6Md.rst7  -o  ./7Md.out  -x  ./7Md.x  -inf  ./7Md.info  -r
./7Md.rst7\n$pmemdG -O -i ../mdins/8Md.in -p ../../000Models/"+mod+"*parm7 -c ./7Md.rst7 -ref ./6Md.rst7 -o
./8Md.out    -x    ./8Md.x    -inf    ./8Md.info    -r    ./8Md.rst7\n$pmemdG    -O    -i    ../mdins/9Md.in    -p
../../000Models/"+mod+"*parm7 -c ./8Md.rst7 -o ./9Md.out -x ./9Md.x -inf ./9Md.info -r ./9Md.rst7\nEOF\ncd
../\n")
#
#  Stage 2: Write pNEB stage list AND AND groupfile
#
#first, let-s do the groupfiles...
if not os.path.isdir('./001--1st2nd'):
 os.makedirs('./001--1st2nd')
```

```
if not os.path.isdir('./002--1st3rd'):
 os.makedirs('./002--1st3rd')
if not os.path.isdir('./003--1st4th'):
 os.makedirs('./003--1st4th')
if not os.path.isdir('./004--2nd3rd'):
 os.makedirs('./004--2nd3rd')
if not os.path.isdir('./005--2nd4th'):
 os.makedirs('./005--2nd4th')
if not os.path.isdir('./006--3rd4th'):
 os.makedirs('./006--3rd4th')
for dire in direlist:
  if dire=='001--1st2nd': modelA='model1'; modelB='model2'
  if dire=='002--1st3rd': modelA='model1'; modelB='model3'
  if dire=='003--1st4th': modelA='model1'; modelB='model4'
  if dire=='004--2nd3rd': modelA='model2'; modelB='model3'
  if dire=='005--2nd4th': modelA='model2'; modelB='model4'
  if dire=='006--3rd4th': modelA='model3'; modelB='model4'
  # 1-f32t300:
  try:
    os.remove('./'+dire+'/groupfile.1-f32t300')
  except OSError:
    pass
  for bea in xrange( 1, int(numBeads)/2 ):
   bead=str(bea)
   with open('./'+dire+'/groupfile.1-f32t300', 'a') as f:
    f.write('-O -p parm -c ../eQUILIB-T0759/'+modelA+'/4md.rst7   -i ../nebins/1-f32t300.in   -x 1.'+bead+'.nc
-o 1.'+bead+'.out   -inf 1.'+bead+'.info   -r 1.'+bead+'.rst7\n')
  for bea in xrange( int(numBeads)/2+1, int(numBeads)+1 ):
   bead=str(bea)
   with open('./'+dire+'/groupfile.1-f32t300', 'a') as f:
    f.write('-O -p parm -c ../eQUILIB-T0759/'+modelB+'/4md.rst7   -i ../nebins/1-f32t300.in   -x 1.'+bead+'.nc
-o 1.'+bead+'.out   -inf 1.'+bead+'.info   -r 1.'+bead+'.rst7\n')
  #2-f32t300-400:
  try:
    os.remove('./'+dire+'/groupfile.2-f32t300-400')
  except OSError:
    pass
  for bea in xrange( 1, int(numBeads)+1 ):
   bead=str(bea)
   with open('./'+dire+'/groupfile.2-f32t300-400', 'a') as f:
    f.write('-O -p parm -c 1.'+bead+'.rst7 -i ../nebins/2-f32t300-400.in -x 2.'+bead+'.nc -o 2.'+bead+'.out -
inf 2.'+bead+'.info -r 2.'+bead+'.rst7\n')
  #3-f32t400-400:
  try:
    os.remove('./'+dire+'/groupfile.3-f32t400-400')
  except OSError:
    pass
  for bea in xrange( 1, int(numBeads)+1 ):
   bead=str(bea)
   with open('./'+dire+'/groupfile.3-f32t400-400', 'a') as f:
    f.write('-O -p parm -c 2.'+bead+'.rst7 -i ../nebins/3-f32t400-400.in -x 3.'+bead+'.nc -o 3.'+bead+'.out -
inf 3.'+bead+'.info -r 3.'+bead+'.rst7\n')
  #4-f32t400-500:
  try:
    os.remove('./'+dire+'/groupfile.4-f32t400-500')
  except OSError:
    pass
  for bea in xrange( 1, int(numBeads)+1 ):
   bead=str(bea)
   with open('./'+dire+'/groupfile.4-f32t400-500', 'a') as f:
    f.write('-O -p parm -c 3.'+bead+'.rst7 -i ../nebins/4-f32t400-500.in -x 4.'+bead+'.nc -o 4.'+bead+'.out -
inf 4.'+bead+'.info -r 4.'+bead+'.rst7\n')
  #5-f32t500-500:
  try:
    os.remove('./'+dire+'/groupfile.5-f32t500-500')
  except OSError:
    pass
  for bea in xrange( 1, int(numBeads)+1 ):
   bead=str(bea)
   with open('./'+dire+'/groupfile.5-f32t500-500', 'a') as f:
    f.write('-O -p parm -c 4.'+bead+'.rst7 -i ../nebins/5-f32t500-500.in -x 5.'+bead+'.nc -o 5.'+bead+'.out -
inf 5.'+bead+'.info -r 5.'+bead+'.rst7\n')
  #6-f32t500-300:
  try:
    os.remove('./'+dire+'/groupfile.6-f32t500-300')
  except OSError:
    pass
  for bea in xrange( 1, int(numBeads)+1 ):
   bead=str(bea)
```

```
   with open('./'+dire+'/groupfile.6-f32t500-300', 'a') as f:
    f.write('-O -p parm -c 5.'+bead+'.rst7 -i ../nebins/6-f32t500-300.in -x 6.'+bead+'.nc -o 6.'+bead+'.out -
inf 6.'+bead+'.info -r 6.'+bead+'.rst7\n')
  #7-f32t300-0K:
  try:
    os.remove('./'+dire+'/groupfile.7-f32t300-0K')
  except OSError:
    pass
  for bea in xrange( 1, int(numBeads)+1 ):
   bead=str(bea)
   with open('./'+dire+'/groupfile.7-f32t300-0K', 'a') as f:
    f.write('-O -p parm -c 6.'+bead+'.rst7 -i ../nebins/7-f32t300-0K.in -x 7.'+bead+'.nc -o 7.'+bead+'.out -inf
7.'+bead+'.info -r 7.'+bead+'.rst7\n')
  #8-f32t0-0K:
  try:
    os.remove('./'+dire+'/groupfile.8-f32t0-0K')
  except OSError:
    pass
  for bea in xrange( 1, int(numBeads)+1 ):
   bead=str(bea)
   with open('./'+dire+'/groupfile.8-f32t0-0K', 'a') as f:
    f.write('-O -p parm -c 7.'+bead+'.rst7 -i ../nebins/8-f32t0-0K.in -x 8.'+bead+'.nc -o 8.'+bead+'.out -inf
8.'+bead+'.info -r 8.'+bead+'.rst7\n')

  if warmBackUp=='yes':
   #9-f32t0-300K:
   try:
     os.remove('./'+dire+'/groupfile.9-f32t0-300K')
   except OSError:
     pass
   for bea in xrange( 1, int(numBeads)+1 ):
    bead=str(bea)
    with open('./'+dire+'/groupfile.9-f32t0-300K', 'a') as f:
     f.write('-O -p parm -c 8.'+bead+'.rst7 -i ../nebins/9-f32t0-300K.in -x 9.'+bead+'.nc -o 9.'+bead+'.out -
inf 9.'+bead+'.info -r 9.'+bead+'.rst7\n')
   #010-f32t300-300K:
   try:
     os.remove('./'+dire+'/groupfile.010-f32t300-300K')
   except OSError:
     pass
   for bea in xrange( 1, int(numBeads)+1 ):
    bead=str(bea)
    with open('./'+dire+'/groupfile.010-f32t300-300K', 'a') as f:
     f.write('-O  -p  parm  -c  9.'+bead+'.rst7  -i  ../nebins/010-f32t300-300K.in  -x  010.'+bead+'.nc  -o
010.'+bead+'.out -inf 010.'+bead+'.info -r 010.'+bead+'.rst7\n')
#
#  Stage 3: Write EQUILIBRATION mdin files...
#
if not os.path.isdir('./eqins'):
  os.makedirs('./eqins')
#1min.in
with open('eqins/1min.in', 'w') as f:
 f.write("Title: 1min.in\n &cntrl\n imin = 1, maxcyc = 10000, ntx = 1,\n ntwx =  50, ntwe = 0, ntwr = 500, ntpr
= 50,\n ntc = 1, ntf = 1, ntb = 1, ntp = 0, cut = 8.0,\n ntr=1,restraintmask='@CA,N,C',\n restraint_wt =
100.,\n/\n")
#2mdheat.in
with open('eqins/2md.in', 'w') as f:
 f.write("Title: 2mdheat.in\n&cntrl\n imin = 0, nstlim = 100000, dt = 0.001,\n irest = 0, ntx = 1, ig = -1,\n
tempi = 100.0, temp0 = 300.0,\n ntc = 2, ntf = 2, tol = 0.00001,\n tautp = 0.1, taup = 0.1,\n ntwx = 1000, ntwe
= 0, ntwr = 1000, ntpr = 1000,\n cut = 8.0, iwrap = 1, ioutfm=1,\n ntt =1, ntb = 1, ntp = 0,\n nscm = 0,\n
ntr=1,restraintmask='@CA,N,C',\n  restraint_wt  =  100.,\n  nmropt=1,\n/\n&wt\n  TYPE='TEMP0',  istep1=0,
istep2=100000,\n value1=100., value2=300.,\n/\n&wt\n TYPE='END',\n/\n/\n")
#3md.in
with open('eqins/3md.in', 'w') as f:
 f.write("Title: 3md.in\n&cntrl\n imin = 0, nstlim = 100000, dt = 0.001,\n irest = 1, ntx = 5, ig = -1,\n tempi
= 300.0, temp0 = 300.0,\n ntc = 2, ntf = 2, tol = 0.00001,\n tautp = 0.1, taup = 0.1,\n ntwx = 1000, ntwe = 0,
ntwr = 1000, ntpr = 1000,\n cut = 8.0, iwrap = 1, ioutfm=1,\n ntt =1, ntb = 2, ntp = 1,\n nscm = 500,\n ntr=1,
restraintmask='@CA,N,C',\n restraint_wt=100.\n/\n")
#4md.in
with open('eqins/4md.in', 'w') as f:
 f.write("Title: 4md.in\n&cntrl\n imin = 0, nstlim = 250000, dt = 0.001,\n irest = 1, ntx = 5, ig = -1,\n tempi
= 300.0, temp0 = 300.0,\n ntc = 2, ntf = 2, tol = 0.00001,\n tautp = 0.5, taup = 0.5,\n ntwx = 1000, ntwe = 0,
ntwr = 1000, ntpr = 1000,\n cut = 8.0, iwrap = 1, ioutfm=1,\n ntt =1, ntb = 2, ntp = 1,\n nscm = 0,\n
ntr=1,restraintmask='@CA,N,C',\n restraint_wt=10.\n/\n")
#5min.in
with open('eqins/5min.in', 'w') as f:
 f.write("Title: 5min.in\n&cntrl\n imin = 1, maxcyc = 10000,\n ntx = 1, \n ntwx =  50, ntwe = 0, ntwr = 500,
ntpr = 50,\n ntc = 1, ntf = 1, ntb = 1, ntp = 0,\n cut = 8.0,\n ntr=1, restraintmask='@CA',\n
restraint_wt=10.\n/\n")
```

```
#6md.in
with open('eqins/6md.in', 'w') as f:
 f.write("Title: 6md.in\n&cntrl\n imin = 0, nstlim = 100000, dt = 0.001,\n irest = 0, ntx = 1, ig = -1,\n tempi
= 300.0, temp0 = 300.0,\n ntc = 2, ntf = 2, tol = 0.00001,\n tautp = 0.5, taup = 0.5,\n ntwx = 1000, ntwe = 0,
ntwr = 1000, ntpr = 1000,\n cut = 8.0, iwrap = 1, ioutfm=1,\n ntt =1, ntb = 2, ntp = 1,\n nscm = 0,\n ntr=1,
restraintmask='@CA', restraint_wt=10.\n/\n")
#7md.in
with open('eqins/7md.in', 'w') as f:
 f.write("Title: 7md.in\n&cntrl\n imin = 0, nstlim = 100000, dt = 0.001,\n irest = 1, ntx = 5, ig = -1,\n tempi
= 300.0, temp0 = 300.0,\n ntc = 2, ntf = 2, tol = 0.00001,\n tautp = 0.5, taup = 0.5,\n ntwx = 1000, ntwe = 0,
ntwr = 1000, ntpr = 1000,\n cut = 8.0, iwrap = 1, ioutfm=1,\n ntt =1, ntb = 2, ntp = 1,\n nscm = 0,\n ntr=1,
restraintmask='@CA', restraint_wt=1.\n/\n")
#8md.in
with open('eqins/8md.in', 'w') as f:
 f.write("Title: 8md.in\n&cntrl\n imin = 0, nstlim = 100000, dt = 0.001,\n irest = 1, ntx = 5, ig = -1,\n tempi
= 300.0, temp0 = 300.0,\n ntc = 2, ntf = 2, tol = 0.00001,\n tautp = 0.5, taup = 0.5,\n ntwx = 1000, ntwe = 0,
ntwr = 1000, ntpr = 1000,\n cut = 8.0, iwrap = 1, ioutfm=1,\n ntt =1, ntb = 2, ntp = 1,\n nscm = 0,\n ntr=1,
restraintmask='@CA', restraint_wt=0.1\n/\n")
#9md.in
with open('eqins/9md.in', 'w') as f:
 f.write("Title: 9md.in\n&cntrl\n imin = 0, nstlim = 250000, dt = 0.001,\n irest = 1, ntx = 5, ig = -1,\n tempi
= 300.0, temp0 = 300.0,\n ntc = 2, ntf = 2, tol = 0.00001,\n tautp = 0.5, taup = 0.5,\n ntwx = 1000, ntwe = 0,
ntwr = 1000, ntpr = 1000,\n cut = 8.0, iwrap = 1, ioutfm=1,\n ntt =1, ntb = 2, ntp = 1,\n nscm = 500,\n/\n")
#
#  Stage 4: write pNEB mdins...
#
if not os.path.isdir('./nebins'):
  os.makedirs('./nebins')
#1-f32t300.in
with open('nebins/1-f32t300.in', 'w') as f:
 f.write("khCASP\n&cntrl\n   nstlim=40000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 0, ntx =1,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=300.0, temp0=300.0,\n   tgtfitmask='@CA',
tgtrmsmask='@CA',\n   ineb = 1, skmin = 32.0, skmax = 32.0,\n   tmode=1,\n /\n")
#2-f32t300-400.in
with open('nebins/2-f32t300-400.in', 'w') as f:
 f.write("khCASP\n&cntrl\n   nstlim=100000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=300.0, temp0=300.0,\n   tgtfitmask='@CA',\n
tgtrmsmask='@CA',\n    ineb = 1, skmin = 32.0, skmax = 32.0,\n    tmode=1,\n    nmropt=1,\n   /\n   &wt
type='TEMP0', istep1=0,istep2=100000,\n    value1=300.0, value2=400.0\n /\n &wt type='END'\n /\n/\n")
#3-f32t400-400.in
with open('nebins/3-f32t400-400.in', 'w') as f:
 f.write("khCASP\n&cntrl\n  nstlim=100000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=400.0, temp0=400.0,\n   tgtfitmask='@CA',
tgtrmsmask='@CA',\n   ineb = 1, skmin = 32.0, skmax = 32.0,\n   tmode=1,\n /\n")
#4-f32t400-500.in
with open('nebins/4-f32t400-500.in', 'w') as f:
 f.write("khCASP\n&cntrl\n  nstlim=100000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=300.0, temp0=300.0,\n   tgtfitmask='@CA',\n
tgtrmsmask='@CA',\n    ineb = 1, skmin = 32.0, skmax = 32.0,\n    tmode=1,\n    nmropt=1,\n   /\n   &wt
type='TEMP0', istep1=0,istep2=100000,\n    value1=400.0, value2=500.0\n /\n &wt type='END'\n /\n/\n")
#5-f32t500-500.in
with open('nebins/5-f32t500-500.in', 'w') as f:
 f.write("khCASP\n&cntrl\n  nstlim=100000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=500.0, temp0=500.0,\n   tgtfitmask='@CA',
tgtrmsmask='@CA',\n   ineb = 1, skmin = 32.0, skmax = 32.0,\n   tmode=1,\n /\n")
#6-f32t500-300.in
with open('nebins/6-f32t500-300.in', 'w') as f:
 f.write("khCASP\n&cntrl\n  nstlim=100000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=300.0, temp0=300.0,\n   tgtfitmask='@CA',\n
tgtrmsmask='@CA',\n    ineb = 1, skmin = 32.0, skmax = 32.0,\n    tmode=1,\n    nmropt=1,\n   /\n   &wt
type='TEMP0', istep1=0,istep2=100000,\n    value1=500.0, value2=300.0\n /\n &wt type='END'\n /\n/\n")
#7-f32t300-0K.in
with open('nebins/7-f32t300-0K.in', 'w') as f:
 f.write("khCASP\n&cntrl\n   nstlim=120000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=300.0, temp0=300.0,\n   tgtfitmask='@CA',\n
tgtrmsmask='@CA',\n    ineb = 1, skmin = 32.0, skmax = 32.0,\n    tmode=1,\n    nmropt=1,\n   /\n   &wt
type='TEMP0',   istep1=0,istep2=10000,\n          value1=300.0,   value2=250.0\n     /\n     &wt   type='TEMP0',
istep1=10001,istep2=20000,\n              value1=250.0,    value2=250.0\n       /\n       &wt    type='TEMP0',
istep1=20001,istep2=30000,\n              value1=250.0,    value2=200.0\n       /\n       &wt    type='TEMP0',
istep1=30001,istep2=40000,\n              value1=200.0,    value2=200.0\n       /\n       &wt    type='TEMP0',
istep1=40001,istep2=50000,\n              value1=200.0,    value2=150.0\n       /\n       &wt    type='TEMP0',
istep1=50001,istep2=60000,\n              value1=150.0,    value2=150.0\n       /\n       &wt    type='TEMP0',
```

```
istep1=60001,istep2=70000,\n               value1=150.0,    value2=100.0\n       /\n       &wt    type='TEMP0',
istep1=70001,istep2=80000,\n               value1=100.0,    value2=100.0\n       /\n       &wt    type='TEMP0',
istep1=80001,istep2=90000,\n               value1=100.0,    value2=50.0\n       /\n       &wt    type='TEMP0',
istep1=90001,istep2=100000,\n               value1=50.0,    value2=50.0\n       /\n       &wt    type='TEMP0',
istep1=100001,istep2=110000,\n               value1=50.0,    value2=0.0\n       /\n       &wt    type='TEMP0',
istep1=110001,istep2=120000,\n    value1=0.0, value2=0.0\n  /\n  &wt type='END'\n  /\n/\n")
#8-f32t0-0K.in
with open('nebins/8-f32t0-0K.in', 'w') as f:
 f.write("khCASP\n&cntrl\n  nstlim=200000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n      cut=8.0,  iwrap=1,\n      ioutfm=1,ntxo=2,\n      temp0=0.0,\n      tgtfitmask='@CA',\n
tgtrmsmask='@CA',\n   ineb = 1, skmin = 32.0, skmax = 32.0,\n   tmode=1,\n   vv=1,vfac=0.1\n/\n")
#9-f32t0-300K.in
if warmBackUp=='yes':
 with open('nebins/9-f32t0-300K.in', 'w') as f:
  f.write("khCASP\n&cntrl\n  nstlim=100000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=0.0, temp0=300.0,\n   tgtfitmask='@CA',\n
tgtrmsmask='@CA',\n    ineb = 1, skmin = 32.0, skmax = 32.0,\n    tmode=1,\n    nmropt=1,\n   /\n   &wt
type='TEMP0', istep1=0,istep2=100000,\n    value1=0.0, value2=300.0\n /\n &wt type='END'\n /\n/\n")
#010-f32t300-300K.in
 with open('nebins/010-f32t300-300K.in', 'w') as f:
  f.write("khCASP\n&cntrl\n  nstlim=100000,\n   dt = 0.0005,\n   ig=-1,\n   imin = 0,\n   irest= 1, ntx =5,\n
ntc=2, ntf=2,\n    ntpr=2000, ntwx=2000,\n    ntt = 3,\n    ntb=2, ntp=1,\n    taup = 0.1, tautp = 0.1,\n
gamma_ln=30.0,\n   cut=8.0, iwrap=1,\n   ioutfm=1,ntxo=2,\n   tempi=300.0, temp0=300.0,\n   tgtfitmask='@CA',
tgtrmsmask='@CA',\n   ineb = 1, skmin = 32.0, skmax = 32.0,\n   tmode=1,\n /\n")
########################################
############################
############
########
#####
###
##
#
#
```

II.ii. Estimating PNEB bead free energies

II.ii.i. Perform MMPBSA analysis of PNEB beads

The goal of this protocol is to estimate the free energy of PNEB beads that have been fully optimized. MMPBSA is used to calculate the solvation free energy of each bead (**Protocol I.vii.ii.**).

```
#!/bin/bash
#
# PNEB-MMPBSA
# calculate the total energy of the conformation, bead for bead, path for path:
#
#set sander:
sander=/home/hk/programs/amber14/binOrig/sander
#set parm7
parm=../ala2.gas.2015.parm7
mkdir -p output_PB
for ((i=0;i<28;i++)); do
ref=`printf "%03d" $i`
#  gen mmpbsa input file
#
#  JSwails says: use inp=2, radiopt=0
#   so radii taken from top file rather
#   than what is hard-coded in sander..
#   http://archive.ambermd.org/201208/0074.html
#
########################################
cat > Idecomp.in << EOF
 test of pbsa
 &cntrl
   ntx=1, imin=1, ipb=1, ntb=0, inp=0
 /
 &pb
   npbverb=0, istrng=200, iprob=2.0, epsout=80.0, epsin=1.0, space=0.5,
   accept=1e-3, dprob=1.5, radiopt=0, fillratio=2, bcopt=6, smoothopt=2, nfocus=1,
   eneopt=2, cutnb=0, maxitn=10000, arcres=0.0625, frcopt=1
 /
EOF
########################################
#
#  run mmpbsa
#
########################################
$sander -O \
-i Idecomp.in \
-p $parm \
-c ../trajs/neb.$ref.rst7 \
-o output_PB/mmpbsa.$ref.out \
-y ../trajs/neb.$ref.nc \
-ref ../trajs/neb.$ref.rst \
-inf output_PB/mmpbsa.$ref.info \
-r output_PB/mmpbsa.$ref.rst7
########################################
done
```

Appendix III. Sample MD simulation analysis protocols

The goal of this section is to provide sample protocols that can be used (or easily modified) to perform some non-standard analyses utilized in the dissertation.

III.i. Automate the analysis of ion distributions

This script utilizes the radial distribution analysis algorithm implemented in cpptraj to characterize the population of ions as a function of the radial distance from the RNA kissing loop.

```
#!/bin/bash
cutforce1=1000
cutforce=0_and_1000
mkdir -p grid
mkdir -p radial
for box in $(echo 8 12 16 20); do
for salt in $(echo 0.1 0.2 0.4 0.8); do
krad=`head ../ion_DistRst/ion_shell_dist.salt_$salt.box_$box.fc_$cutforce1.rst| grep "iat=-1,1029"|awk '{print
$4}'|sed  -e 's/r3=//g' -e 's/,//g'| awk '{printf "%.4f", $1}'`
rad=`echo $krad+$krad|bc -l`
parm=../gb_parms/tT.salt_$salt.box_$box.parm7
rst=../gb_rsts/tT.salt_$salt.box_$box.rst7
#cpptraj << EOF
/opt/amber/bin/cpptraj << EOF
parm $parm
reference $rst
trajin ../salt_$salt.box_$box.fc_$cutforce1/7md.x
trajin ../salt_$salt.box_$box.fc_$cutforce1/8md.x
radial radial/radial.K.salt_$salt.box_$box.dat 1 $rad @K+ :1-32 center1
radial radial/radial.C.salt_$salt.box_$box.dat 1 $rad @Cl- :1-32 center1
radial radial/radial.K.salt_$salt.box_$box.ctr2.dat 1 $rad @K+ :1-32 center2
radial radial/radial.C.salt_$salt.box_$box.ctr2.dat 1 $rad @Cl- :1-32 center2
EOF
# - - - - - -
done
done
```

Appendix IV. Software documentation: Helios

Helios: A method to characterize helix geometry

Freely available via:

https://github.com/kehauser/heliosv1

IV.i. Advantages and disadvantages of Helios

Advantages of the method

The method is accurate, general, and appropriate for irregular helices. Because a grid search is used rather than a residual-minimization algorithm (such as L-BFGS), the algorithm is robust against non-convex solution surfaces.

Limitations of the method

The current method is limited to cylindrical helices and does not guarantee the best solution if a global optimum does exist (because we use a grid search). The user must supply points of the helix.

IV.ii. Details of singular value decomposition used by Helios

IV.ii.1. Using singular value analysis to obtain singular values

We used the singular value analysis (SVA) algorithm within Lawson and Hanson's' Fortran90 library of least squares solvers[249] to solve our linear least squares problem of the form **AX**=**B**. The singular value decomposition of the **A** matrix (M by N dimension) was computed from:

$$\mathbf{A}=\mathbf{USV}' \tag{IV-1}$$

where **U** is M by M orthogonal, **S** is M by N diagonal with the diagonal terms being nonnegative and ordered from large to small, and **V'** is N by N orthogonal. These matrices must also satisfy the condition that:

$$\mathbf{S}=\mathbf{U}'\mathbf{AV} \tag{IV-2}$$

The product matrix **G** is a multiplication of the above **U'** unitary matrix and the **B** matrix:

$$\mathbf{G}=\mathbf{U}'\mathbf{B} \tag{IV-3}$$

To obtain the best estimate of the solution for our linear least squares problem, the **X** matrix of solutions is found by multiplying the ith column of **V**, the ith singular value and the ith row of **G**, where i is an index from 1 to 3 (the pseudo-rank of **A**).

Finally, to obtain the values of ρ, x_0, and y_0 estimated by SVD, given any rotation of $\hat{n}$ (θ and ϕ):

$$v_j * \sigma_j^{-1} * g_j \tag{IV-4}$$

where $j = 1$ returns the radius, ρ, $j = 2$ returns x_0, and $j = 3$ returns y_0; v_j is the jth column of **V**, σ_j is the jth singular value in **S**, and g_j is the jth row of **G**.

IV.ii.2. Modifications made to Lawson & Hanson's SVD library

The output formatting of the **liblawson.f90** library was modified to return only the actual singular values. The source code of this library was obtained from:

https://people.sc.fsu.edu/~jburkardt/f_src/lawson/lawson.html

The following lines in the library file were modified:

- Removed 3060 to 3065; mute subaccuracy warning from the QR bidiagonal matrix singular value decomposition algorithm. Subaccuracies are reflected in the resulting singular values so that the residuals being calculated in the main program would be very large.
- Modified 2596; return only the solutions of singular value analysis.
- Removed 2616 to 2623; return only the r-norms.
- Removed 1249, 1276, 1278, 1285, 1288, 1293, 1295 , 2589, 2591; mute formatted printing.

IV.iii. A brief usage guide for Helios

Helios options:	User operation:
	(X,Y,Z) coordinates of object
`coord_type = 0 (x y z), 1 (PDB), 2 (Amber)`	Set file format
	single- or double-helix (e.g. DNA)
`dsDNA = 0 (single), 1 (double)`	Set helix type
	domain of the grid in degrees
`grid_phi_beg = [0, 180); grid_phi_end = [0, 180)` `grid_theta_beg = [0, 360); grid_theta_end = [0, 360)`	Search for the helix axis orientation
	resolution of the grid
`num_grid = 90 (90 in ϕ and 90 in θ)`	Set number of grid points in ϕ | θ
	atom names in a PDB file
`helix_atom_names = P (or any len=3 character)`	Select the atom names, if using PDB
	print the helix step twist and rise
`print_step = 0 (do not print); 1 (print)`	Toggle the printing option
	output file name
`helixout_name = helios.dat (default)`	Set the name of the output file

Additional Helios options:	User operation:
	output angles in degrees or radians
`oradian = 0 (print in degrees); 1 (print in radians)`	Select the angular space
	simplify output file style
`print_to_plot = 0 (verbose output); 1 (simple output)`	Select simpler output file formatting
	print singular values from SVA
`print_sing = 0 (do not print); 1 (print)`	Select more fitting information
	fitting results for all grid points
`all_helix_out = 0 (do not print); 1 (print)`	Select verbose grid information
	helix coordinates in the new frame
`opt_axis_out = 0 (do not print); 1 (print)`	Print the helix in its optimal frame
`opt_Axis_out_name = Helix_along_z_now.pdb`	Set file name. PDB format

Appendix V. Supplementary discussion

V.i. Vibrational spectroscopy

The purpose of this section is to provide a brief introduction to the theory of vibrational spectroscopy, including reference to molecular group theory. Vibrational spectroscopy is important for appreciating the chemically intuitive construction of molecular mechanics force fields, as the bonded terms – bonds and angles – correspond to the chemical models derived directly from vibrational spectroscopy (also referred to as infrared spectroscopy).

At room temperature, the covalent bonds between atoms vibrate like springs. A spring is a harmonic oscillator. Modeling the vibrations of covalent bonds between atoms as harmonic oscillators is a good assumption, if a bond is strong and can be assumed to be always present. The energy of a harmonic oscillator tends to infinity as it is stretched; a harmonic oscillator cannot model bond breaking. The harmonic oscillator approximation assumes that the ground state of a molecule's electronic structure is the minimum energy (equilibrium) geometry around which vibrations oscillate (normal modes).

Let us assume for a moment that we have a molecule with strong covalent bonds that do no break under the conditions in which the experiment is being performed. For the simplest molecule composed of two atoms, the bond stretching frequency, $\tilde{\nu}$, is:

$$\tilde{\nu} = \frac{1}{2\pi c}\sqrt{\frac{f(m_1+m_2)}{m_1+m_2}} \qquad \text{(V-1)}$$

where c is the velocity of light (**Table VI-1, Appendix VI**), m_1 and m_2 are the masses of atoms 1 and 2 respectively, and f is a spring force constant proportional to the strength of the bond.

Molecular symmetry (**Table VI-3**) defines the combination of vibrational modes that are available for exercise (properly referred to as excitation; or less properly referred to as wiggling) upon absorption of radiant energy with resonant frequency to a molecule's vibrational mode(s). The absorption of radiant energy with resonant frequency to the normal modes of a molecule pumps the molecule into vibrational excited states. Relaxation from a vibrational excited state back to the ground state is always accompanied by release of electromagnetic energy (phonon) corresponding to the exact frequency of the radiant energy that was absorbed. A wide band of frequencies can be shone onto the molecule, and only the resonant frequencies will be absorbed. A precise absorption (and emission) spectrum will be detected behind the sample (absorption and emission spectroscopy, infrared spectroscopy) or at a right angle to the sample (Raman spectroscopy, ultraviolet/visible spectroscopy). The frequencies of light that are absorbed or emitted by the sample are often sufficient to characterize the composition of the molecule and, in the case of simple molecules, its chemical structure. Therefore, vibrational spectroscopy is a simple but fundamental experimental tool to accurately study the basic physical and dynamical properties of molecules. Vibrational spectroscopy is a natural starting point for the development of molecular models: The first force fields were parameterized from vibrational spectroscopy.[250-252]

A basic force field is the combination of the terms in equations (1-5) through (1-9) along with the individual parameters for them. Each hybridization state of the elements C, N and O (and other elements depending on the force field) has its own set of parameters. For each of these parameters sets for a particular hydration state of an element, the individual parameters depends

on the bonded atom types: for example, an –H bond versus an –OH bond. Clearly, the bond stretching frequency ($k_{\mathbf{B}_{C_{sp3}\text{-H}}}$) of the C-H bond in methyl (2917 cm^{-1}, asymmetric A_1)[253] will be different from the stretching frequency ($k_{\mathbf{B}_{C_{sp3}\text{-OH}}}$) of the C-OH bond in methanol (3681 cm^{-1})[253] despite the hybridization state of the C atom being the same (sp^3). Bond stretching frequencies are directly related to the bond force constants (equation (V-1)).

Appendix VI. Useful physical constants and relations

Table VI-1. Table of physical constants from the NIST.

Property	**Constant**	**Value**
Avogradro's number	N_A	6.022 140 857 × 10^{23} mol^{-1}
Boltzmann's constant	k_B	1.380 648 52 × 10^{-23} m^2 kg s^{-2} K^{-1}
Boltzmann's constant	k_B	1.380 648 52 × 10^{-23} J K^{-1}
Velocity of light, in vacuum	c	299 792 458 m s^{-1}
Permittivity, in vacuum	ε_0	8.854 187 817 × 10^{-12} F m^{-1}
Energy, Calorie	kcal	1.048 54 × 10^{13} Hz
Energy, Joule	J	4.184 cal
Planck's constant	h	6.626 070 040 × 10^{-34} J s
Planck's reduced constant	ħ	1.054 571 800 × 10^{-34} J s
Impedence of vacuum	Z_0	376.730 313 461 Ω
Magnetic constant	μ_0	12.566 370 614 N A^{-2}
Atomic mass constant	m_u	1.660 539 040 × 10^{-34} kg
Faraday's constant	F	96 485.332 89 C mol^{-1}
Gas constant	R	8.314 4598 J K^{-1} mol^{-1}
Molar volume, ideal gas	V_m	22.710 947 × 10^{-3} m^3 mol^{-1}
Sackur-Tetrode constant (1K, 100 kPa)	S_0/R	-1.151 7084 Dimensionless
Stefan-Boltzmann constant	σ	5.670 367 × 10^{-3} W m^{-2} K^{-4}
Wien frequency displacement	b'	5.878 9238 × 10^{10} Hz K^{-1}
Bohr radius	a_{0m}	0.529 177 210 × 10^{-10} m
Electron mass	m_e	9.109 383 56 × 10^{-3} kg
Elementary charge	e	1.602 176 6208× 10^{-19} C
Fine structure constant	α♪	0.08 5424 55 Dimensionless

♪Denotes the fine structure constant, α, whose value is dimensionless. The literature often cites the square of the inverse of the value in **Table VI-1**. The fine structure constant is derived:

$$\alpha = \frac{e^2}{\hbar c^2 4\pi\varepsilon_0}$$

where e is the elementary charge, ħ is Planck's reduced constant, c is the speed of light in a vacuum and ε_0 is the permittivity constant in a vacuum.

Table VI-2. Molecular symmetry elements and operations.

Symmetry element	**Symbol**	**Operation**
Identity	E	Nothing changes
n-fold proper	C_n	Axial rotation by $2\pi/n$
Mirror plane	σ	Reflection through plane
Inversion center	i	Inversion through point
n-fold improper	S_n	Axial rotation by $2\pi/n$, then reflection through norm-plane

Table VI-3. Table of fundamental relations.

Property	Components	Name	Relation
Energy	h, Planck's constant ν, frequency	Planck-Einstein	$E=h\nu$
Energy	m, mass *c*, speed of light	Einstein's speed limit	$E=mc^2$
Energy	ħ, Planck's reduced constant ω, angular frequency	Planck's constant	$E=\hbar\omega$
Energy	p, momentum *c*, speed of light	Relativistic momentum	$E=pc$
Energy	λ, wavelength p, momentum	de Broglie relation	$h=\lambda p$
Force	m, mass a, acceleration	Gram	$F=ma$
Energy	m, mass v, velocity	Meters per second	$E=mv$

Table VI-4. Energy unit conversions.

	Hartree	eV	cm^{-1}	kcal/mol*	°K	J
Hartree	1	27.2107	219474.63	627.503	315777	43.60×10^{-19}
eV	0.0367502	1	8065.73	23.0609	11604.9	1.60210×10^{-19}
cm^{-1}	4.55633×10^{-6}	1.23981×10^{-4}	1	0.00285911	1.42879	1.98630×10^{-23}
kcal/mol*	0.00159362	0.043634	349.757	1	503.228	6.95×10^{-21}
°K	0.00000316678	0.0000861705	0.00198717	0.00831435	1	1.38054×10^{-23}
J	2.294×10^{17}	6.24181×10^{18}	5.03445×10^{22}	1.44×10^{20}	7.24354×10^{22}	1

References cited

1. Emsley, J. U. h. b. g. c. p. b. i. Y. X. O., *Nature's Building Blocks: An A-Z Guide to the Elements*. Oxford University Press: 2001.
2. Crick, F., Central dogma of molecular biology. *Nature* **1970,** *227* (5258), 561-3.
3. Lynch, M.; Walsh, B., *The origins of genome architecture*. Sinauer Associates Sunderland: 2007; Vol. 98.
4. Vaquerizas, J. M.; Kummerfeld, S. K.; Teichmann, S. A., et al., A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **2009,** *10* (4), 252-63.
5. Elf, J.; Li, G. W.; Xie, X. S., Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **2007,** *316* (5828), 1191-1194.
6. Berg, O. G.; Winter, R. B.; von Hippel, P. H., Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **1981,** *20* (24), 6929-48.
7. Fazio, T. A.; Visnapuu, M.; Greene, E. C., et al., Fabrication of nanoscale "curtain rods" for DNA curtains using nanoimprint lithography. *J Vac Sci Technol B* **2009,** *27* (6), 3095-3098.
8. Halford, S. E., An end to 40 years of mistakes in DNA-protein association kinetics? *Biochem Soc Trans* **2009,** *37* (Pt 2), 343-8.
9. Zhou, H. X., Rapid search for specific sites on DNA through conformational switch of nonspecifically bound proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2011,** *108* (21), 8651-6.
10. Marcovitz, A.; Levy, Y., Weak frustration regulates sliding and binding kinetics on rugged protein-DNA landscapes. *J Phys Chem B* **2013,** *117* (42), 13005-14.
11. Slutsky, M.; Kardar, M.; Mirny, L. A., Diffusion in correlated random potentials, with applications to DNA. *Phys Rev E* **2004,** *69* (6).
12. Melero, R.; Rajagopalan, S.; Lazaro, M., et al., Electron microscopy studies on the quaternary structure of p53 reveal different binding modes for p53 tetramers in complex with DNA. *Proc. Natl. Acad. Sci. U.S.A.* **2011,** *108* (2), 557-562.
13. Leith, J. S.; Tafvizi, A.; Huang, F., et al., Sequence-dependent sliding kinetics of p53. *Nucleic Acids Res.* **2012,** *109* (41), 16552-7.
14. Pasi, M.; Maddocks, J. H.; Beveridge, D., et al., muABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **2014,** *42* (19), 12272-83.
15. Jayaram, B.; McConnell, K. J.; Dixit, S. B., et al., Free Energy Analysis of Protein–DNA Binding: The EcoRI Endonuclease–DNA Complex. *J Comput Phys* **1999,** *151* (1), 333-357.
16. Copeland, R. A.; Pompliano, D. L.; Meek, T. D., Drug-target residence time and its implications for lead optimization. *Nature reviews. Drug discovery* **2006,** *5* (9), 730-9.
17. Yang, L.; Zhou, T.; Dror, I., et al., TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res* **2014,** *42* (Database issue), D148-55.
18. Olson, W. K.; Gorin, A. A.; Lu, X. J., et al., DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95* (19), 11163-11168.
19. Rohs, R.; Jin, X.; West, S. M., et al., Origins of specificity in protein-DNA recognition. *Annual review of biochemistry* **2010,** *79*, 233-69.

20. Jen-Jacobson, L.; Engler, L. E.; Jacobson, L. A., Structural and thermodynamic strategies for site-specific DNA binding proteins. *Structure* **2000,** *8* (10), 1015-23.
21. Kalodimos, C. G.; Biris, N.; Bonvin, A. M. J. J., et al., Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **2004,** *305* (5682), 386-389.
22. Viadiu, H.; Aggarwal, A. K., Structure of BamHI bound to nonspecific DNA: a model for DNA sliding. *Molecular cell* **2000,** *5* (5), 889-95.
23. Townson, S. A.; Samuelson, J. C.; Bao, Y., et al., BstYI bound to noncognate DNA reveals a "hemispecific" complex: implications for DNA scanning. *Structure* **2007,** *15* (4), 449-59.
24. Winkler, F. K.; Banner, D. W.; Oefner, C., et al., The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *The EMBO journal* **1993,** *12* (5), 1781-95.
25. Yakubovskaya, E.; Mejia, E.; Byrnes, J., et al., Helix unwinding and base flipping enable human MTERF1 to terminate mitochondrial transcription. *Cell* **2010,** *141* (6), 982-93.
26. Viadiu, H.; Aggarwal, A. K., Structure of BamHI bound to nonspecific DNA: a model for DNA sliding. *Molecular cell* **2000,** *5* (5), 889-95.
27. Winkler, F. K.; Banner, D. W.; Oefner, C., et al., The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *The EMBO journal* **1993,** *12* (5), 1781-95.
28. Morita, J.; Kato, K.; Nakane, T., et al., Crystal structure of the plant receptor-like kinase TDR in complex with the TDIF peptide. *Nat Commun* **2016,** *7*, 12383.
29. Song, W.; Wang, J.; Han, Z., et al., Structural basis for specific recognition of single-stranded RNA by Toll-like receptor 13. *Nat Struct Mol Biol* **2015,** *22* (10), 782-7.
30. Wang, D.; Huang, B.; Zhang, S., et al., Structural basis for R-spondin recognition by LGR4/5/6 receptors. *Genes Dev* **2013,** *27* (12), 1339-44.
31. Yao, R.; Ming, Z.; Yan, L., et al., DWARF14 is a non-canonical hormone receptor for strigolactone. *Nature* **2016,** *536* (7617), 469-73.
32. Park, K.; Shen, B. W.; Parmeggiani, F., et al., Control of repeat-protein curvature by computational protein design. *Nat Struct Mol Biol* **2015,** *22* (2), 167-74.
33. Altherr, T.; Mendes, R. V.; Seixas, J., A new method for fast track recognition at present and future colliders. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1993,** *332* (1), 284-287.
34. Frühwirth, R.; Lichtenwagner, P.; Regler, M., et al., The DELPHI forward track fit Track fitting with outlier rejection. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1993,** *334* (2), 528-536.
35. Gyulassy, M.; Harlander, M., Elastic tracking and neural network algorithms for complex pattern recognition. *Computer Physics Communications* **1991,** *66* (1), 31-46.
36. Lindström, M., Track reconstruction in the ATLAS detector using elastic arms. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1995,** *357* (1), 129-149.
37. Ohlsson, M., Extensions and explorations of the elastic arms algorithm. *Computer Physics Communications* **1993,** *77* (1), 19-32.
38. Doyle, L.; Hallinan, J.; Bolduc, J., et al., Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **2015,** *528* (7583), 585-588.
39. Nievergelt, Y., Fitting helices to data by total least squares. *Comput Aided Geom D* **1997,** *14* (8), 707-718.

40. Michels, D. L.; Desbrun, M., A semi-analytical approach to molecular dynamics. *Journal of Computational Physics* **2015,** *303*, 336-354.
41. Levitt, M., A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* **1976,** *104* (1), 59-107.
42. Levitt, M.; Warshel, A., Computer simulation of protein folding. *Nature* **1975,** *253* (5494), 694-8.
43. D.A. Case, R. M. B., W. Botello-Smith, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman, AMBER 2016. *University of California San Francisco* **2016**.
44. MacKerell, A. D.; Bashford, D.; Bellott, M., et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **1998,** *102* (18), 3586-616.
45. MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., 3rd, Improved treatment of the protein backbone in empirical force fields. *J Am Chem Soc* **2004,** *126* (3), 698-9.
46. Perez, A.; MacCallum, J. L.; Brini, E., et al., Grid-based backbone correction to the ff12SB protein force field for implicit-solvent simulations. *Journal of chemical theory and computation* **2015,** *11* (10), 4770-9.
47. Lemkul, J. A.; Huang, J.; Roux, B., et al., An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem Rev* **2016,** *116* (9), 4983-5013.
48. Swope, W. C.; Andersen, H. C.; Berens, P. H., et al., A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics* **1982,** *76* (1), 637-649.
49. Brooks, B.; Karplus, M., Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* **1983,** *80* (21), 6571-5.
50. Tirion, M. M., Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys Rev Lett* **1996,** *77* (9), 1905-1908.
51. Wilson Jr, E. B.; Decius, J. G.; Cross, P. G., et al., Molecular vibrations. *American Journal of Physics* **1955,** *23* (8), 550-550 %@ 0002-9505.
52. Ashcroft, N. W.; Mermin, N. D., Introduction to Solid State Physics. *Saunders, Philadelphia* **1976**.
53. Orth, P.; Schnappinger, D.; Hillen, W., et al., Structural basis of gene regulation by the tetracycline inducible Tet repressor-operator system. *Nat Struct Biol* **2000,** *7* (3), 215-9.
54. Cordeiro, T. N.; Schmidt, H.; Madrid, C., et al., Indirect DNA readout by an H-NS related protein: structure of the DNA complex of the C-terminal domain of Ler. *PLoS Pathog* **2011,** *7* (11), e1002380.
55. Little, E. J.; Babic, A. C.; Horton, N. C., Early interrogation and recognition of DNA sequence by indirect readout. *Structure* **2008,** *16* (12), 1828-37.
56. Watkins, D.; Hsiao, C.; Woods, K. K., et al., P22 c2 repressor-operator complex: mechanisms of direct and indirect readout. *Biochemistry* **2008,** *47* (8), 2325-38.
57. Thyme, S. B.; Jarjour, J.; Takeuchi, R., et al., Exploitation of binding energy for catalysis and design. *Nature* **2009,** *461* (7268), 1300-4.

58. Otwinowski, Z.; Schevitz, R. W.; Zhang, R. G., et al., Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* **1988,** *335* (6188), 321-9.
59. Schultz, S. C.; Shields, G. C.; Steitz, T. A., Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science* **1991,** *253* (5023), 1001-7.
60. Parkinson, G.; Wilson, C.; Gunasekera, A., et al., Structure of the CAP-DNA complex at 2.5 angstroms resolution: a complete picture of the protein-DNA interface. *J Mol Biol* **1996,** *260* (3), 395-408.
61. Schneider, T. D., A brief review of molecular information theory. *Nano Commun Netw* **2010,** *1* (3), 173-180.
62. Crooks, G. E.; Hon, G.; Chandonia, J. M., et al., WebLogo: a sequence logo generator. *Genome Res* **2004,** *14* (6), 1188-90.
63. Stormo, G. D., DNA binding sites: representation and discovery. *Bioinformatics* **2000,** *16* (1), 16-23.
64. Thastrom, A.; Lowary, P. T.; Widlund, H. R., et al., Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J Mol Biol* **1999,** *288* (2), 213-29.
65. Balasubramanian, S.; Xu, F.; Olson, W. K., DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences. *Biophys J* **2009,** *96* (6), 2245-60.
66. Richmond, T. J.; Davey, C. A., The structure of DNA in the nucleosome core. *Nature* **2003,** *423* (6936), 145-50.
67. Tolstorukov, M. Y.; Colasanti, A. V.; McCandlish, D. M., et al., A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J Mol Biol* **2007,** *371* (3), 725-38.
68. Bishop, T. C., Geometry of the nucleosomal DNA superhelix. *Biophys J* **2008,** *95* (3), 1007-17.
69. Marathe, A.; Bansal, M., An ensemble of B-DNA dinucleotide geometries lead to characteristic nucleosomal DNA structure and provide plasticity required for gene expression. *BMC Struct Biol* **2011,** *11*, 1.
70. Doench, J. G.; Fusi, N.; Sullender, M., et al., Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **2016,** *34* (2), 184-91.
71. Bogdanove, A. J.; Voytas, D. F., TAL effectors: customizable proteins for DNA targeting. *Science* **2011,** *333* (6051), 1843-6.
72. Boch, J.; Scholze, H.; Schornack, S., et al., Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **2009,** *326* (5959), 1509-12.
73. 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* (13), 2860-74.
74. Metzler, R.; van den Brock, B.; Wuite, G. J. L., et al., Role of DNA Conformations in Gene Regulation. *Biophysics of DNA- Protein Interactions: From Single Molecules to Biological Systems* **2011**, 69-84.
75. Rohs, R.; Jin, X.; West, S. M., et al., Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry* **2010,** *79*, 233-69.
76. Rogers, J. M.; Barrera, L. A.; Reyon, D., et al., Context influences on TALE-DNA binding revealed by quantitative profiling. *Nature Communications* **2015,** *6*.

77. Mills, J. B.; Hagerman, P. J., Origin of the intrinsic rigidity of DNA. *Nucleic Acids Res* **2004,** *32* (13), 4055-9.
78. Luo, R.; Gilson, H. S.; Potter, M. J., et al., The physical basis of nucleic acid base stacking in water. *Biophys J* **2001,** *80* (1), 140-8.
79. Pasi, M.; Maddocks, J. H.; Lavery, R., Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res* **2015,** *43* (4), 2412-23.
80. Range, K.; Mayaan, E.; Maher, L. J., et al., The contribution of phosphate-phosphate repulsions to the free energy of DNA bending. *Nucleic Acids Res* **2005,** *33* (4), 1257-1268.
81. Olson, W. K.; Zhurkin, V. B., Twenty years of DNA bending. *Biological Structure and Dynamics, Vol 2* **1996**, 341-370.
82. Lavery, R.; Maddocks, J. H.; Pasi, M., et al., Analyzing ion distributions around DNA. *Nucleic Acids Res* **2014,** *42* (12), 8138-49.
83. Babcock, M. S.; Pednault, E. P.; Olson, W. K., Nucleic acid structure analysis: mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *J Mol Biol* **1994,** *237* (1), 125-156.
84. Rohs, R.; West, S. M.; Sosinsky, A., et al., The role of DNA shape in protein-DNA recognition. *Nature* **2009,** *461* (7268), 1248-53.
85. Hauser, K.; Essuman, B.; He, Y., et al., A human transcription factor in search mode. *Nucleic Acids Res* **2016,** *44* (1), 63-74.
86. Minary, P.; Levitt, M., Training-free atomistic prediction of nucleosome occupancy. *Proc Natl Acad Sci U S A* **2014,** *111* (17), 6293-8.
87. Paillard, G.; Lavery, R., Analyzing protein-DNA recognition mechanisms. *Structure* **2004,** *12* (1), 113-22.
88. Olson, W. K.; Gorin, A. A.; Lu, X. J., et al., DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proceedings of the National Academy of Sciences* **1998,** *95* (19), 11163-11168.
89. Heddi, B.; Oguey, C.; Lavelle, C., et al., Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res* **2010,** *38* (3), 1034-47.
90. Parker, S. C.; Hansen, L.; Abaan, H. O., et al., Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **2009,** *324* (5925), 389-92.
91. Savelyev, A.; MacKerell, A. D., Jr., All-atom polarizable force field for DNA based on the classical Drude oscillator model. *J Comput Chem* **2014,** *35* (16), 1219-39.
92. Savelyev, A.; MacKerell, A. D., Jr., Differential Deformability of the DNA Minor Groove and Altered BI/BII Backbone Conformational Equilibrium by the Monovalent Ions Li(+), Na(+), K(+), and Rb(+) via Water-Mediated Hydrogen Bonding. *J Chem Theory Comput* **2015,** *11* (9), 4473-85.
93. Galindo-Murillo, R.; Robertson, J. C.; Zgarbova, M., et al., Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* **2016,** *12* (8), 4114-27.
94. Ivani, I.; Dans, P. D.; Noy, A., et al., Parmbsc1: a refined force field for DNA simulations. *Nat Methods* **2016,** *13* (1), 55-8.
95. Pasi, M.; Maddocks, J. H.; Beveridge, D., et al., muABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* **2014,** *42* (19), 12272-83.
96. Pedersen, A. G.; Baldi, P.; Chauvin, Y., et al., DNA structure in human RNA polymerase II promoters. *J Mol Biol* **1998,** *281* (4), 663-73.

97. Goni, J. R.; Perez, A.; Torrents, D., et al., Determining promoter location based on DNA structure first-principles calculations. *Genome Biol* **2007,** *8* (12), R263.
98. Duran, E.; Djebali, S.; Gonzalez, S., et al., Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. *Nucleic Acids Res* **2013,** *41* (15), 7220-30.
99. Kornberg, R. D.; Lorch, Y., Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **1999,** *98* (3), 285-94.
100. Geggier, S.; Vologodskii, A., Sequence dependence of DNA bending rigidity. *Proc Natl Acad Sci U S A* **2010,** *107* (35), 15421-6.
101. Xu, F.; Olson, W. K., DNA architecture, deformability, and nucleosome positioning. *J Biomol Struct Dyn* **2010,** *27* (6), 725-39.
102. Davey, C. A.; Sargent, D. F.; Luger, K., et al., Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J Mol Biol* **2002,** *319* (5), 1097-113.
103. Portella, G.; Battistini, F.; Orozco, M., Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations. *PLoS Comput Biol* **2013,** *9* (11), e1003354.
104. Choy, J. S.; Wei, S.; Lee, J. Y., et al., DNA methylation increases nucleosome compaction and rigidity. *J Am Chem Soc* **2010,** *132* (6), 1782-3.
105. Jimenez-Useche, I.; Yuan, C., The effect of DNA CpG methylation on the dynamic conformation of a nucleosome. *Biophys J* **2012,** *103* (12), 2502-12.
106. Zhu, R.; Howorka, S.; Proll, J., et al., Nanomechanical recognition measurements of individual DNA molecules reveal epigenetic methylation patterns. *Nat Nanotechnol* **2010,** *5* (11), 788-91.
107. Shim, J.; Humphreys, G. I.; Venkatesan, B. M., et al., Detection and quantification of methylation in DNA using solid-state nanopores. *Sci Rep* **2013,** *3*, 1389.
108. Lee, J. Y.; Lee, T. H., Effects of DNA methylation on the structure of nucleosomes. *J Am Chem Soc* **2012,** *134* (1), 173-5.
109. Hadzic, T.; Park, D.; Abruzzi, K. C., et al., Genome-wide features of neuroendocrine regulation in Drosophila by the basic helix-loop-helix transcription factor DIMMED. *Nucleic Acids Res* **2015,** *43* (4), 2199-215.
110. Hawkins, J.; Grant, C.; Noble, W. S., et al., Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics* **2009,** *25* (12), i339-47.
111. Halfon, M. S.; Zhu, Q.; Brennan, E. R., et al., Erroneous attribution of relevant transcription factor binding sites despite successful prediction of cis-regulatory modules. *BMC Genomics* **2011,** *12*, 578.
112. Zhou, T.; Shen, N.; Yang, L., et al., Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A* **2015,** *112* (15), 4654-9.
113. Yue, F.; Cheng, Y.; Breschi, A., et al., A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **2014,** *515* (7527), 355-+.
114. Christopher, N. A.; Swanson, R.; Baldwin, T. O., Algorithms for finding the axis of a helix: Fast rotational and parametric least-squares methods. *Computers & chemistry* **1996,** *20* (3), 339-345.
115. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-2637.
116. Enkhbayar, P.; Damdinsuren, S.; Osaki, M., et al., HELFIT: Helix fitting by a total least squares method. *Comput. Biol. Chem.* **2008,** *32* (4), 307-10.

117. Bansal, M.; Kumar, S.; Velavan, R., HELANAL: a program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.* **2000,** *17* (5), 811-9.
118. Zheng, G.; Lu, X. J.; Olson, W. K., Web 3DNA--a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.* **2009,** *37* (Web Server issue), W240-6.
119. Blanchet, C.; Pasi, M.; Zakrzewska, K., et al., CURVES plus web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.* **2011,** *39*, W68-W73.
120. Babcock, M. S.; Pednault, E. P.; Olson, W. K., Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *J. Mol. Biol.* **1994,** *237* (1), 125-56.
121. Porteus, M., Genome Editing: A New Approach to Human Therapeutics. *Annu. Rev. Pharmacol. Toxicol.* **2016,** *56*, 163-90.
122. Whitworth, W. A., The Regular Polygon in Space. In *The Oxford, Cambridge and Dublin Messenger of Mathematics*, W. Allen Whitworth, C. T., R. Pendlebury and J. W. L. Glaiser Ed. Macmillan & Co.: Cambridge: Trinity Street, Corner of Green Street, 1875; Vol. 4, pp 88-89.
123. Case, D.; Babin, V.; Berryman, J., et al., Amber 14. **2014**.
124. Maier, J. A.; Martinez, C.; Kasavajhala, K., et al., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015,** *11* (8), 3696-713.
125. Armen, R.; Alonso, D. O.; Daggett, V., The role of α-, 310-, and π-helix in helix→ coil transitions. *Protein Science* **2003,** *12* (6), 1145-1157.
126. Guo, Z. Y.; Kraka, E.; Cremer, D., Description of local and global shape properties of protein helices. *J. Mol. Model.* **2013,** *19* (7), 2901-2911.
127. Pasi, M.; Maddocks, J. H.; Beveridge, D., et al., μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **2014,** *42* (19), 12272-12283.
128. Stella, S.; Molina, R.; Lopez-Mendez, B., et al., BuD, a helix-loop-helix DNA-binding domain for genome modification. *Acta Crystallographica Section D-Biological Crystallography* **2014,** *70*, 2042-2052.
129. Berman, H. M.; Westbrook, J.; Feng, Z., et al., The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28* (1), 235-242.
130. Hauser, K.; Essuman, B.; He, Y., et al., A human transcription factor in search mode. *Nucleic Acids Res.* **2016,** *44* (1), 63-74.
131. Lavery, R.; Moakher, M.; Maddocks, J. H., et al., Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* **2009,** *37* (17), 5917-29.
132. Zheng, G.; Lu, X.-J.; Olson, W. K., Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res* **2009**, gkp358.
133. Saenger, W., *Principles of nucleic acid structure*. Springer Science: 1984.
134. Mak, A. N.; Bradley, P.; Cernadas, R. A., et al., The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **2012,** *335* (6069), 716-9.
135. Kuriyan, J.; Konforti, B.; Wemmer, D., *The molecules of life: Physical and chemical principles*. Garland Science: 2012.

136. Maeder, M. L.; Gersbach, C. A., Genome-editing Technologies for Gene and Cell Therapy. *Molecular therapy : the journal of the American Society of Gene Therapy* **2016,** *24* (3), 430-46.
137. Hauser, K.; Perez, A.; Garcia-Diaz, M., et al., The Two Codes of DNA. *in preparation*.
138. Deng, D.; Yan, C.; Wu, J., et al., Revisiting the TALE repeat. *Protein Cell* **2014,** *5* (4), 297-306.
139. Streubel, J.; Blucher, C.; Landgraf, A., et al., TAL effector RVD specificities and efficiencies. *Nature biotechnology* **2012,** *30* (7), 593-5.
140. Stella, S.; Molina, R.; Bertonatti, C., et al., Expression, purification, crystallization and preliminary X-ray diffraction analysis of the novel modular DNA-binding protein BurrH in its apo form and in complex with its target DNA. *Acta Crystallogr F Struct Biol Commun* **2014,** *70* (Pt 1), 87-91.
141. Mak, A. N.; Bradley, P.; Bogdanove, A. J., et al., TAL effectors: function, structure, engineering and applications. *Curr Opin Struct Biol* **2013,** *23* (1), 93-9.
142. Deng, D.; Yan, C.; Pan, X., et al., Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **2012,** *335* (6069), 720-3.
143. Nunez, J. K.; Harrington, L. B.; Kranzusch, P. J., et al., Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature* **2015,** *527* (7579), 535-8.
144. Kay, S.; Boch, J.; Bonas, U., Characterization of AvrBs3-like effectors from a Brassicaceae pathogen reveals virulence and avirulence activities and a protein with a novel repeat architecture. *Molecular plant-microbe interactions : MPMI* **2005,** *18* (8), 838-48.
145. Wicky, B. I.; Stenta, M.; Dal Peraro, M., TAL effectors specificity stems from negative discrimination. *Plos One* **2013,** *8* (11), e80261.
146. Meckler, J. F.; Bhakta, M. S.; Kim, M. S., et al., Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic acids research* **2013,** *41* (7), 4118-28.
147. Juillerat, A.; Dubois, G.; Valton, J., et al., Comprehensive analysis of the specificity of transcription activator-like effector nucleases. *Nucleic acids research* **2014,** *42* (8), 5390-402.
148. Doyle, E. L.; Hummel, A. W.; Demorest, Z. L., et al., TAL effector specificity for base 0 of the DNA target is altered in a complex, effector- and assay-dependent manner by substitutions for the tryptophan in cryptic repeat -1. *PloS one* **2013,** *8* (12), e82120.
149. Hauser, K.; Garcia-Diaz, M.; Simmerling, C., et al., Characterizing helix complementarity using geometric analysis. *in preparation*.
150. DeLano, W. L., The PyMOL molecular graphics system. **2002**.
151. Hauser, K.; Garcia Diaz, M.; Simmerling, C., et al., Characterization of biomolecular helices and their complementarity using geometric analysis. *in preparation*.
152. Zandarashvili, L.; Esadze, A.; Vuzman, D., et al., Balancing between affinity and speed in target DNA search by zinc-finger proteins via modulation of dynamic conformational ensemble. *Proc Natl Acad Sci U S A* **2015,** *112* (37), E5142-9.
153. Zandarashvili, L.; Vuzman, D.; Esadze, A., et al., Asymmetrical roles of zinc fingers in dynamic DNA-scanning process by the inducible transcription factor Egr-1. *Proc Natl Acad Sci U S A* **2012,** *109* (26), E1724-32.
154. Hauser, K.; Schiffer, B.; Coutsias, E., et al., Asynchronous shifts by asymmetrical modules bias how MTERF1 slides on DNA. *in preparation*.
155. Boyken, S. E.; Chen, Z.; Groves, B., et al., De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **2016,** *352* (6286), 680-7.

156. Li, H. J.; Lai, C. T.; Pan, P., et al., A Structural and Energetic Model for the Slow-Onset Inhibition of the Mycobacterium tuberculosis Enoyl-ACP Reductase InhA. *Acs Chem Biol* **2014,** *9* (4), 986-993.
157. Kleine, T.; Leister, D., Emerging functions of mammalian and plant mTERFs. *Biochimica et biophysica acta* **2015**.
158. Roberti, M.; Polosa, P. L.; Bruni, F., et al., The MTERF family proteins: mitochondrial transcription regulators and beyond. *Biochim Biophys Acta* **2009,** *1787* (5), 303-11.
159. Terzioglu, M.; Ruzzenente, B.; Harmel, J., et al., MTERF1 binds mtDNA to prevent transcriptional interference at the light-strand promoter but is dispensable for rRNA gene transcription regulation. *Cell metabolism* **2013,** *17* (4), 618-26.
160. Wallace, D. C., A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: A dawn for evolutionary medicine. *Annu Rev Genet* **2005,** *39*, 359-407.
161. Larsson, N. G., Somatic Mitochondrial DNA Mutations in Mammalian Aging. *Annual review of biochemistry* **2010,** *79*, 683-706.
162. Helm, M.; Florentz, C.; Chomyn, A., et al., Search for differences in post-transcriptional modification patterns of mitochondrial DNA-encoded wild-type and mutant human tRNA(Lys) and tRNA(Leu(UUR)). *Nucleic Acids Research* **1999,** *27* (3), 756-763.
163. Chomyn, A.; Martinuzzi, A.; Yoneda, M., et al., Melas Mutation in Mtdna Binding-Site for Transcription Termination Factor Causes Defects in Protein-Synthesis and in Respiration but No Change in Levels of Upstream and Downstream Mature Transcripts. *Proc Natl Acad Sci U S A* **1992,** *89* (10), 4221-4225.
164. Frankel, A. D.; Kim, P. S., Modular structure of transcription factors: implications for gene regulation. *Cell* **1991,** *65* (5), 717-9.
165. Hammani, K.; Bonnard, G.; Bouchoucha, A., et al., Helical repeats modular proteins are major players for organelle gene expression. *Biochimie* **2014,** *100*, 141-50.
166. Park, K.; Shen, B. W.; Parmeggiani, F., et al., Control of repeat-protein curvature by computational protein design. *Nat Struct Mol Biol* **2015**.
167. Rubinson, E. H.; Eichman, B. F., Nucleic acid recognition by tandem helical repeats. *Curr Opin Struct Biol* **2012,** *22* (1), 101-9.
168. Sanner, M. F.; Olson, A. J.; Spehner, J.-C., Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **1996,** *38* (3), 305-320.
169. He, B.; Zalkin, H., Repression of Escherichia coli purB is by a transcriptional roadblock mechanism. *Journal of bacteriology* **1992,** *174* (22), 7121-7.
170. Lavery, R.; Moakher, M.; Maddocks, J. H., et al., Conformational analysis of nucleic acids revisited: Curves. *Nucleic Acids Res* **2009,** *37* (17), 5917-5929.
171. Johnson, K. A., Role of induced fit in enzyme specificity: a molecular forward/reverse switch. *J Biol Chem* **2008,** *283* (39), 26297-301.
172. Levy, Y.; Onuchic, J. N.; Wolynes, P. G., Fly-casting in protein-DNA binding: frustration between protein folding and electrostatics facilitates target recognition. *J Am Chem Soc* **2007,** *129* (4), 738-9.
173. Mccammon, J. A.; Northrup, S. H., Gated Binding of Ligands to Proteins. *Nature* **1981,** *293* (5830), 316-317.
174. Xue, B.; Brown, C. J.; Dunker, A. K., et al., Intrinsically disordered regions of p53 family are highly diversified in evolution. *Biochim Biophys Acta* **2013,** *1834* (4), 725-38.

175. Bouvier, B.; Lavery, R., A Free Energy Pathway for the Interaction of the SRY Protein with Its Binding Site on DNA from Atomistic Simulations. *J Am Chem Soc* **2009,** *131* (29), 9864-+.
176. Jain, V.; Hilton, B.; Lin, B., et al., Unusual sequence effects on nucleotide excision repair of arylamine lesions: DNA bending/distortion as a primary recognition factor. *Nucleic Acids Res* **2013,** *41* (2), 869-80.
177. Kuznetsov, N. A.; Bergonzo, C.; Campbell, A. J., et al., Active destabilization of base pairs by a DNA glycosylase wedge initiates damage recognition. *Nucleic Acids Res* **2015,** *43* (1), 272-81.
178. Mu, H.; Geacintov, N. E.; Zhang, Y., et al., Recognition of Damaged DNA for Nucleotide Excision Repair: A Correlated Motion Mechanism with a Mismatched cis-syn Thymine Dimer Lesion. *Biochemistry* **2015,** *54* (34), 5263-7.
179. Yang, L.; Beard, W. A.; Wilson, S. H., et al., Polymerase beta simulations suggest that Arg258 rotation is a slow step rather than large subdomain motions per se. *J Mol Biol* **2002,** *317* (5), 651-71.
180. Zhong, S.; Chen, X.; Zhu, X., et al., Identification and validation of human DNA ligase inhibitors using computer-aided drug design. *J Med Chem* **2008,** *51* (15), 4553-62.
181. Bakan, A.; Meireles, L. M.; Bahar, I., ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* **2011,** *27* (11), 1575-1577.
182. Eyal, E.; Yang, L. W.; Bahar, I., Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics* **2006,** *22* (21), 2619-27.
183. Skjaerven, L.; Martinez, A.; Reuter, N., Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins-Structure Function and Bioinformatics* **2011,** *79* (1), 232-43.
184. D.A. Case, V. B., J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman *Amber 14*, University of California, San Francisco, 2014.
185. Chen, V. B.; Arendall, W. B., 3rd; Headd, J. J., et al., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **2010,** *66* (Pt 1), 12-21.
186. Hornak, V.; Abel, R.; Okur, A., et al., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics* **2006,** *65* (3), 712-25.
187. Perez, A.; Marchan, I.; Svozil, D., et al., Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* **2007,** *92* (11), 3817-29.
188. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D., et al., Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **1983,** *79* (2), 926-935.
189. Joung, I. S.; Cheatham, T. E., 3rd, Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B* **2008,** *112* (30), 9020-41.
190. Macke, T. J.; Case, D. A., Modeling Unusual Nucleic Acid Structures. In *Acs Sym Ser*, American Chemical Society: 1997; Vol. 682, pp 379-393.

191. Pasi, M.; Maddocks, J. H.; Beveridge, D., et al., μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Research* **2014**.
192. Macke, T. J.; Case, D. A., Modeling unusual nucleic acid structures. *Acs Sym Ser* **1998,** *682*, 379-393.
193. Pedone, F.; Santoni, D., Sequence-dependent DNA helical rise and nucleosome stability. *BMC molecular biology* **2009,** *10*, 105.
194. Olson, W. K.; Gorin, A. A.; Lu, X. J., et al., DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* **1998,** *95* (19), 11163-8.
195. Roberts, V. A.; Pique, M. E.; Ten Eyck, L. F., et al., Predicting protein-DNA interactions by full search computational docking. *Proteins-Structure Function and Bioinformatics* **2013,** *81* (12), 2106-2118.
196. Roberts, V. A.; Case, D. A.; Tsui, V., Predicting interactions of winged-helix transcription factors with DNA. *Proteins-Structure Function and Bioinformatics* **2004,** *57* (1), 172-187.
197. Word, J. M.; Lovell, S. C.; LaBean, T. H., et al., Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol* **1999,** *285* (4), 1711-33.
198. Baker, N. A.; Sept, D.; Joseph, S., et al., Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **2001,** *98* (18), 10037-41.
199. Sanner, M. F.; Olson, A. J.; Spehner, J. C., Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996,** *38* (3), 305-20.
200. York, D. M.; Wlodawer, A.; Pedersen, L. G., et al., Atomic-level accuracy in simulations of large protein crystals. *Proc Natl Acad Sci U S A* **1994,** *91* (18), 8715-8.
201. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F., et al., Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **1984,** *81* (8), 3684-3690 %@ 0021-9606.
202. Hopkins, C. W.; Le Grand, S.; Walker, R. C., et al., Long Time Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* **2015**.
203. Mirny, L.; Slutsky, M.; Wunderlich, Z., et al., How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J Phys a-Math Theor* **2009,** *42* (43).
204. Tafvizi, A.; Mirny, L. A.; van Oijen, A. M., Dancing on DNA: Kinetic Aspects of Search Processes on DNA. *Chemphyschem* **2011,** *12* (8), 1481-1489.
205. Mahmutovic, A.; Berg, O. G.; Elf, J., What matters for lac repressor search in vivo-sliding, hopping, intersegment transfer, crowding on DNA or recognition? *Nucleic Acids Res* **2015**.
206. Bhattacherjee, A.; Levy, Y., Search by proteins for their DNA target site: 1. The effect of DNA conformation on protein sliding. *Nucleic Acids Research* **2014,** *42* (20), 12404-12414.
207. Bhattacherjee, A.; Levy, Y., Search by proteins for their DNA target site: 2. The effect of DNA conformation on the dynamics of multidomain proteins. *Nucleic Acids Research* **2014,** *42* (20), 12415-12424.
208. Marklund, E. G.; Mahmutovic, A.; Berg, O. G., et al., Transcription-factor binding and sliding on DNA studied using micro- and macroscopic models. *Proc Natl Acad Sci U S A* **2013,** *110* (49), 19796-19801.

209. Furini, S.; Barbini, P.; Domene, C., DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. *Nucleic Acids Res* **2013,** *41* (7), 3963-3972.
210. Gur, M.; Zomot, E.; Bahar, I., Global motions exhibited by proteins in micro- to milliseconds simulations concur with anisotropic network model predictions. *The Journal of chemical physics* **2013,** *139* (12), 121912.
211. Roe, D. R.; Cheatham, T. E., 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013,** *9* (7), 3084-95.
212. Skjaerven, L.; Martinez, A.; Reuter, N., Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins* **2011,** *79* (1), 232-43.
213. Blainey, P. C.; Luo, G. B.; Kou, S. C., et al., Nonspecifically bound proteins spin while diffusing along DNA. *Nature structural & molecular biology* **2009,** *16* (12), 1224-U34.
214. Naganathan, A. N.; Orozco, M., The conformational landscape of an intrinsically disordered DNA-binding domain of a transcription regulator. *J Phys Chem B* **2013,** *117* (44), 13842-50.
215. Abe, N.; Dror, I.; Yang, L., et al., Deconvolving the recognition of DNA shape from sequence. *Cell* **2015,** *161* (2), 307-18.
216. Bouvier, B.; Zakrzewska, K.; Lavery, R., Protein-DNA recognition triggered by a DNA conformational switch. *Angew Chem Int Ed Engl* **2011,** *50* (29), 6516-8.
217. Lemkul, J. A.; Savelyev, A.; MacKerell, A. D., Jr., Induced Polarization Influences the Fundamental Forces in DNA Base Flipping. *The journal of physical chemistry letters* **2014,** *5* (12), 2077-2083.
218. Fleishman, S. J.; Baker, D., Role of the biomolecular energy gap in protein design, structure, and evolution. *Cell* **2012,** *149* (2), 262-73.
219. Thyme, S. B.; Song, Y.; Brunette, T. J., et al., Massively parallel determination and modeling of endonuclease substrate specificity. *Nucleic Acids Research* **2014,** *42* (22), 13839-52.
220. Thyme, S. B.; Baker, D.; Bradley, P., Improved modeling of side-chain--base interactions and plasticity in protein--DNA interface design. *J Mol Biol* **2012,** *419* (3-4), 255-74.
221. Berg, O. G.; Blomberg, C., Association kinetics with coupled diffusion III. Ionic-strength dependence of the lac repressor-operator association. *Biophysical chemistry* **1978,** *8* (4), 271-80.
222. van den Broek, B.; Lomholt, M. A.; Kalisch, S. M., et al., How DNA coiling enhances target localization by proteins. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** *105* (41), 15738-42.
223. Lomholt, M. A.; van den Broek, B.; Kalisch, S. M., et al., Facilitated diffusion with DNA coiling. *Proceedings of the National Academy of Sciences of the United States of America* **2009,** *106* (20), 8204-8.
224. Halford, S. E.; Marko, J. F., How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res* **2004,** *32* (10), 3040-3052.
225. Givaty, O.; Levy, Y., Protein Sliding along DNA: Dynamics and Structural Characterization. *J Mol Biol* **2009,** *385* (4), 1087-1097.
226. Vuzman, D.; Azia, A.; Levy, Y., Searching DNA via a "Monkey Bar" Mechanism: The Significance of Disordered Tails. *J Mol Biol* **2010,** *396* (3), 674-684.
227. Vuzman, D.; Polonsky, M.; Levy, Y., Facilitated DNA Search by Multidomain Transcription Factors: Cross Talk via a Flexible Linker. *Biophys J* **2010,** *99* (4), 1202-1211.
228. Kolomeisky, A. B., Physics of protein-DNA interactions: mechanisms of facilitated target search. *Phys Chem Chem Phys* **2011,** *13* (6), 2088-2095.

229. Redding, S.; Greene, E. C., How do proteins locate specific targets in DNA? *Chemical physics letters* **2013,** *570*.
230. Ryu, K. S.; Tugarinov, V.; Clore, G. M., Probing the rate-limiting step for intramolecular transfer of a transcription factor between specific sites on the same DNA molecule by (15)Nz-exchange NMR spectroscopy. *Journal of the American Chemical Society* **2014,** *136* (41), 14369-72.
231. Mahmutovic, A.; Berg, O. G.; Elf, J., What matters for lac repressor search in vivo--sliding, hopping, intersegment transfer, crowding on DNA or recognition? *Nucleic Acids Res* **2015,** *43* (7), 3454-64.
232. Winter, R. B.; Berg, O. G.; von Hippel, P. H., Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry* **1981,** *20* (24), 6961-77.
233. Zhou, H. X., Rapid search for specific sites on DNA through conformational switch of nonspecifically bound proteins. *Proceedings of the National Academy of Sciences of the United States of America* **2011,** *108* (21), 8651-6.
234. Tafvizi, A.; Huang, F.; Leith, J. S., et al., Tumor suppressor p53 slides on DNA with low friction and high stability. *Biophys J* **2008,** *95* (1), L01-3.
235. Tafvizi, A.; Huang, F.; Fersht, A. R., et al., A single-molecule characterization of p53 search on DNA. *Proc Natl Acad Sci U S A* **2011,** *108* (2), 563-8.
236. Leith, J. S.; Tafvizi, A.; Huang, F., et al., Sequence-dependent sliding kinetics of p53. *Proceedings of the National Academy of Sciences of the United States of America* **2012,** *109* (41), 16552-7.
237. Kalodimos, C. G.; Biris, N.; Bonvin, A. M., et al., Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **2004,** *305* (5682), 386-9.
238. Iwahara, J.; Zweckstetter, M.; Clore, G. M., NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (41), 15062-7.
239. Loth, K.; Gnida, M.; Romanuka, J., et al., Sliding and target location of DNA-binding proteins: an NMR view of the lac repressor system. *Journal of biomolecular NMR* **2013,** *56* (1), 41-9.
240. Vuzman, D.; Levy, Y., The "Monkey-Bar" Mechanism for Searching for the DNA Target Site: The Molecular Determinants. *Isr J Chem* **2014,** *54* (8-9), 1374-1381.
241. Vuzman, D.; Levy, Y., DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proc Natl Acad Sci U S A* **2010,** *107* (49), 21004-9.
242. Vuzman, D.; Azia, A.; Levy, Y., Searching DNA via a "Monkey Bar" mechanism: the significance of disordered tails. *J Mol Biol* **2010,** *396* (3), 674-84.
243. Khazanov, N.; Levy, Y., Sliding of p53 along DNA can be modulated by its oligomeric state and by cross-talks between its constituent domains. *J Mol Biol* **2011,** *408* (2), 335-55.
244. Zhou, H. X., The affinity-enhancing roles of flexible linkers in two-domain DNA-binding proteins. *Biochemistry* **2001,** *40* (50), 15069-73.
245. Vuzman, D.; Polonsky, M.; Levy, Y., Facilitated DNA search by multidomain transcription factors: cross talk via a flexible linker. *Biophys J* **2010,** *99* (4), 1202-11.
246. Cornell, W. D.; Cieplak, P.; Bayly, C. I., et al., A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *Journal of the American Chemical Society* **1995,** *117* (19), 5179-5197.

247. Weiser, J.; Shenkin, P. S.; Still, W. C., Approximate solvent-accessible surface areas from tetrahedrally directed neighbor densities. *Biopolymers* **1999,** *50* (4), 373-80.
248. Slater, G. W.; Guo, H. L.; Nixon, G. I., Bidirectional transport of polyelectrolytes using self-modulating entropic ratchets. *Phys Rev Lett* **1997,** *78* (6), 1170-1173.
249. Lawson, C. L. H., Richard J, *Solving least squares problems*. SIAM: Philadelphia, 1995.
250. Lifson, S.; Warshel, A., Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *The Journal of Chemical Physics* **1968,** *49* (11), 5116-5129.
251. Allinger, N. L., Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *Journal of the American Chemical Society* **1977,** *99* (25), 8127-8134.
252. Lii, J. H.; Allinger, N. L., Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics. *Journal of the American Chemical Society* **1989,** *111* (23), 8566-8575.
253. Shimanouchi, T., Tables of molecular vibrational frequencies. Consolidated volume II. *Journal of physical and chemical reference data* **1977,** *6* (3), 993-1102.

fin.