Atomistic simulation of interactions between DNA and integration host factor

George Daniel Watson

Doctor of Philosophy

University of York
Physics
January 2021

To Maria

Abstract

The prokaryotic genome is structured by many proteins known collectively as nucleoid-associated proteins (NAPs), which have many functions relevant to gene regulation; among them is integration host factor (IHF), a DNA-binding protein which induces some of the sharpest DNA bends found in nature. This thesis presents the results of advanced all-atom molecular dynamics simulations of the IHF–DNA complex.

These simulations confirm previous observations that IHF bends DNA in multiple states with distinct bend angles, finding three states with varying sequence specificity, and provide for the first time their structures in atomistic detail. These include an "associated" state corresponding to a DNA bend of 66° and a "half-wrapped" state with a 115° bend angle, in addition to the previously known "fully wrapped" (157°) state, and agree with data from complementary atomic force microscopy (AFM) experiments. These states differ primarily in the position of the DNA "arms" on each side of the protein's binding site; by performing advanced simulations using a modified potential to improve sampling of the conformation space and applying the weighted histogram analysis method, the free-energy landscape for binding of each arm is obtained and a remarkable asymmetry and interdependence are observed, explaining the three binding modes.

This technique also reveals that the bridging of two pieces of DNA by IHF is highly energetically favourable, explaining the formation of large DNA–IHF clusters and the role of IHF in stabilising biofilms associated with bacterial infections.

Simulations of DNA minicircles illustrate the complex interplay between DNA topology and DNA bending, finding that supercoiling—a change in the number of turns in the double helix—changes the distribution of IHF binding states and bend angles, while IHF binding consistently controls the minicircle conformation, always being positioned at the apex of plectonemes.

Contents

List of Tables

List of Figures

Acknowledgements

> If I have seen further it is by standing on the sholders [*sic*] of Giants.
>
> – Sir Isaac Newton*

Science is, by nature, a collaborative endeavour, and this work could not have been achieved without my mentors, collaborators, and friends. In recognition of this, I would like to thank first and foremost Dr Agnes Noy and Prof. Mark Leake for their supervision, support, and guidance over the last three years. Similar thanks must be extended to the entire Physics of Life Group, which comprises exclusively wonderful people with whom I am truly fortunate to have had the opportunity to work. Particular emphasis should be placed on the contributions of Dr Sam Yoshua, without whom this work could not have taken the form it has; and my various office-mates, whose presence made my work much easier in myriad small ways. To Drs Robert Greenall, who kindled my love of molecular simulation, and Neville Yee, who taught me most of what I know, I am indebted.

On a more personal level, I must express my deepest appreciation of and gratitude for my fiancée, Maria Hyde, who has stood by me and supported me unfalteringly for as long as I have had the privilege to know her, and without whom I would not be where I am today; and to my family, whose love and support have been of incalculable value over the course of my entire life. I wish to thank also my closest friends—the quiz team that hasn't quizzed for a while and the book club that hasn't read a book for a while—for the many contributions they have made to making the world feel like a better place to be.

This work has been generously supported by grants from the Engineering and Physical Sciences Research Council. Calculations were performed on Archer, JADE, the Cambridge Tier-2 computing cluster, and the local facilities at the University of York (Viking and YARCC).

This thesis is typeset in Palatino and AMS Euler using LaTeX.

---

* Newton I 1675 Letter to Hooke R `https://discover.hsp.org/Record/dc-9792` (Appropriately, this quote was not an original innovation of Newton's; he stood a great many shoulders by expressing a concept that can be traced to the 12th Century scholar Bernard of Chartres, who spoke of *nanos gigantum humeris insidentes*—dwarfs mounted on the shoulders of giants.)

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

I performed all of the simulations and analysis of simulation data presented herein (except when discussed as an example of prior work in the field and acknowledged with a reference) independently, albeit with support and guidance from my supervisor Agnes Noy.

All experimental results using atomic force microscopy were obtained by collaborator Sam Yoshua; it is necessary that some of this work be discussed in this thesis as the simulations and experiments were developed and performed in parallel to complement and enhance one another.

Publications

The following publications arose from this project.

[1] Yoshua S B, Watson G D, Howard J A L, Velasco-Berrelleza V, Leake M C and Noy A 2021 "A nucleoid-associated protein bends and bridges DNA in a multiplicity of topological states with varying specificity" *Nucleic Acids Res.* doi:10.1101/2020.04.17.047076 (Under review)

[2] Watson G D and Noy A 2021 "Structural interplay between DNA supercoiling and protein-induced DNA bending" (In preparation)

Chapter 1

Introduction and background

This thesis describes a series of simulations modelling the interactions between DNA and bacterial nucleoid-associated proteins. This chapter provides the necessary biological background, including nucleic acids and nucleoid-associated proteins; reviews the existing literature in the field; and discusses the motivation for this work. In chapter 2, the simulation methodology is discussed, including an overview of molecular dynamics simulation as well as anlysis techniques and advanced sampling methods that were of utility in this project.

Chapter 3 describes a series of simulations of DNA binding to the protein IHF, which were developed alongside experimental work, the relevant parts of which are provided in appendix 3. This work provides new evidence that IHF can bind to DNA in any of three distinct modes, provides structures for these states, and illustrates a constructed free-energy landscape obtained through advanced MD simulations. In chapter 4, simulations of supercoiled DNA minicircles—covalently closed circular pieces of DNA with applied torsion—are described, revealing for the first time the effect of DNA topology on the binding of IHF and the effect of IHF binding on the topology of DNA.

Chapter 5 concludes by discussing the results described in the preceding chapters, evaluating the success of the project, and proposing promising areas for further study.

1.1 Nucleic acids

Nucleic acids are self-replicating biological molecules that encode the genetic information in all known living organisms. This genetic information has many purposes within biological cells, but foremost among them is the production of proteins through a process so important it was named the "central dogma" of molecular biology by Francis Crick in 1958 [1]. A nucleic acid is a polymer composed of a sequence of nucleotides, each of which comprises a pentose sugar, a negatively-charged phosphate group, and a nitrogenous base. In ribonucleic

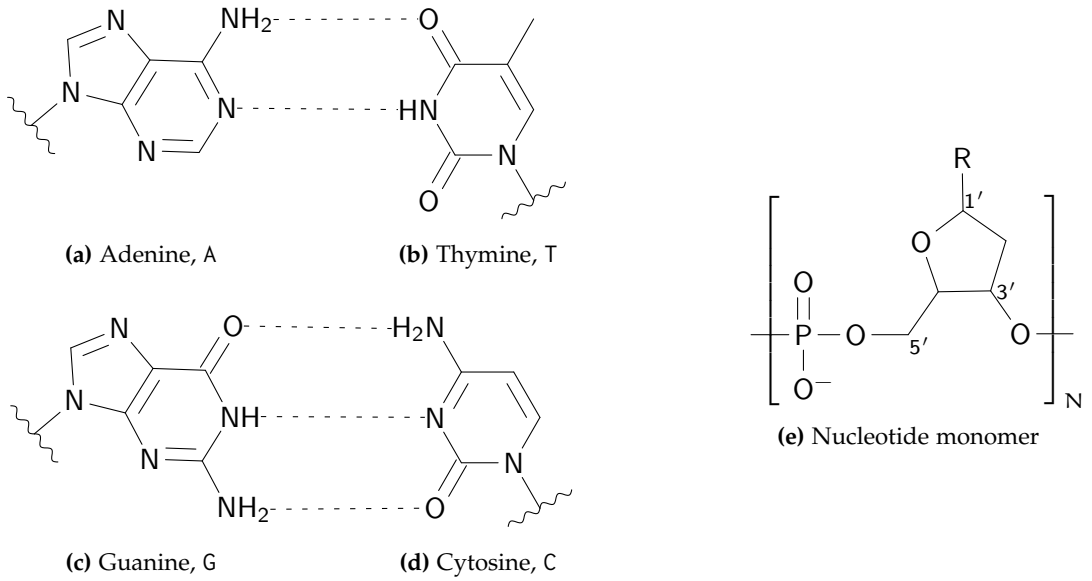


Figure 1: The DNA nucleobases consist primarily of one (pyrimidine derivatives, right) or two (purine derivatives, left) planar aromatic rings. The hydrogen bonds, indicated by dotted lines, are restricted to those shown here by the positions of acceptor and donor atoms; adenine (a) and thymine (b) each have two hydrogen bonding sites, while guanine (c) and cytosine (d) have three. Wavy lines indicate the site of attachment to the 1′ carbon atom of the deoxyribose sugar by a glycosidic bond. This corresponds to the label R in panel e, which shows the structure of the rest of a monomer nucleotide. In a nucleic acid polymer, the oxygen atom bound to the 3′ carbon in the sugar forms part of the phosphate group bound to the 5′ carbon of the neighbouring nucleotide, as represented here by the surrounding brackets.

acid (RNA) the sugar is ribose, $C_5H_{10}O_5$; in deoxyribonucleic acid (DNA) it is deoxyribose, $C_5H_{10}O_4$, which is derived from ribose by the loss of an oxygen atom.

In both cases, the nucleic acid polymer contains only four distinct monomers, which differ only in their nitrogenous base. In DNA these are adenine (A), cytosine (C), guanine (G), and thymine (T), all of which are illustrated in figure 1, while in RNA thymine is replaced by uracil (U). These are usually classified based on the parent compound from which they are derived, which also acts as a convenient proxy for size; thymine, cytosine, and uracil are derived from monocyclic pyrimidine, while adenine and guanine are derived from bicyclic purine. All of the nucleobases are aromatic, meaning they consist of planar rings of resonance bonds; such a structure consists of an alternating cycle of single and double covalent bonds, with the electrons in the molecular π system effectively delocalised. The resonance between different structures with different arrangements of single and double bonds results in a ring of bonds of effective order 1.5, leading to increased chemical stability and ensuring that the structures remain planar.

Another grouping of more significance is in terms of the number of hydrogen bonds each nucleobase is able to form, and the arrangement of their hydrogen bond acceptor and donor atoms. While adenine and thymine (and uracil) can form two hydrogen bonds, guanine and cytosine are able to form three, and the acceptor and donor locations are reversed (figure 1). In this way, base pairing is restricted to A···T and G···C; these pairs each consist of a purine-derived and a pyrimidine-derived base, resulting in base pairs of roughly consistent size. The A···T double hydrogen bond is typically weaker than the G···C triple hydrogen bond, allowing AT-rich sequences of DNA to be broken apart more easily (although base stacking interactions actually contribute more to this effect than does base paring [2]). While RNA is typically single-stranded, utilising this base pairing in storage only to form simple structures such as hairpins [3], DNA is almost always packaged into the famous double helix described by Watson* and Crick in 1953 [4], with two complementary strands coiled around a common axis to maintain the stability of the molecule and protect the genome from damage. The planarity of the base pairs allows them to stack neatly above one another, stabilised by intra-base-pair stacking interactions [2].

This simple base-pairing restriction is perhaps the most powerful tool in the biochemical arsenal, as it fundamentally underlies the processes of replication and transcription. DNA can, for example, replicate itself by simply unwinding to expose the hydrogen bonding sites of the bases on both strands; in time, complementary nucleotides will find their way to these sites, eventually resulting in the formation of two double helices, each nearly identical to the original. The slight imperfections in this process, which result in mutations of the genetic code, are a double-edged sword, being among the main drivers of both evolution and life-threatening diseases such as cancer [5].

The process of decoding the genome in order to construct proteins similarly begins with DNA unwinding, but the coding strand in this case pairs with ribonucleotides, transcribing the genetic code into messenger RNA [6]. Each three-base-pair (3 bp) "codon" in this messenger RNA goes on to pair with a piece of amino-acid-carrying transfer RNA in the ribosome, allowing translation of the genetic code into a polypeptide sequence and resulting in the formation of a protein [7].

In reality, these processes do not occur spontaneously—many important proteins are involved in regulating, initiating, and facilitating reproduction, transcription, and translation. In fact, most DNA does not directly code for proteins [8], instead being involved in complex gene-regulation processes [9]. For example, many DNA sequences exist primarily to mark the start of genes, providing a binding site for

* No relation

the molecular machinery associated with transcription or for regulatory proteins or RNA molecules that up- or down-regulate gene expression by preventing transcription enzymes from binding nearby or changing the local shape of DNA to facilitate it [10]. This promoter–operator model—in which genes are located in "operons" alongside a promoter region, to which transcription enzymes bind to initiate transcription, and an operator, to which repressor proteins bind to down-regulate transcription—is common in prokaryotes (such as bacteria and archæa).

Perhaps the most well-known example of such an operon is the *lac* operon in *Escherichia coli*. Like most bacteria, *E. coli*'s preferred energy source is the sugar glucose. When glucose is plentiful, producing enzymes to digest lactose would be wasteful, but in a glucose-poor (but lactose-rich) environment *E. coli* must either digest lactose or die. In the absence of lactose, the *lac* repressor binds to the operator region just upstream of the gene that codes for the lactose-metabolising enzyme, preventing transcription; when lactose binds to the repressor, allosteric effects cause the protein to release from the DNA, allowing transcription. Actually initiating transcription, however, further requires the catabolite activator protein, which is similarly suppressed by glucose. Thus, lactose metabolism begins if and only if lactose is present but glucose is not [11, 12].

Similar regulatory mechanisms exist for a great many genes, but most are less well understood. Operon functions are known to depend upon local DNA structure [13, 14], which provides an important but poorly understood regulatory mechanism.

1.1.1 DNA structure

The DNA double helix has an interesting geometry, which can only be described by taking into account a large number of parameters, each of which can take on a range of values for each base pair or base-pair step [15]. Broadly, relaxed DNA *in vivo* can be divided into two main forms, termed A-DNA and B-DNA, both of which form right-handed helices, although unusual conformations such as left-handed Z-DNA may form in certain circumstances [16]. Of these, B-DNA is the most common in nature and represents relaxed DNA under typical conditions. Some of the more significant parameters are listed for A-DNA and B-DNA in table 1. The rotation of each strand about the helix axis over a base-pair step is sometimes referred to as twist; for reasons that will become apparent in section 1.1.2, it will herein be referred to as rotation.

The two strands of the double helix are not directly opposite each other, so the "grooves" formed by the spaces between the strands are of unequal size. In B-DNA, when measured parallel to the helix axis, the major groove has a width of 22 Å, while the minor groove is only 12 Å wide [17]. The bases are significantly more

Table 1: Structural parameters of A-DNA and B-DNA

Parameter	A-DNA	B-DNA
Base pairs per turn	11	10.5
Rotation / $^\circ\,\mathrm{bp}^{-1}$	32.7	34.3
Rise along axis / $\mathrm{Å\,bp}^{-1}$	2.3	3.32
Diameter / Å	23	20

accessible in the major groove, so this serves as the primary binding site for many DNA-binding proteins, although minor-groove binding is important to many small ligands such as dyes.

1.1.2 DNA supercoiling

Unwinding the DNA double helix in order to make use of the genetic information it encodes necessarily entails applying some torsion. The application of torsion about the helical axis of DNA is known as supercoiling; applying torsion that would tend to increase the number of turns the DNA strands make about this axis is known as positive supercoiling, while torsion that would tend to remove turns is referred to as negative supercoiling [18]. DNA *in vivo* is almost never torsionally relaxed, with the genomes of both prokaryotes and eukaryotes (such as plants, animals, and fungi) generally maintained in a negatively supercoiled state since this reduces the helical energy of DNA, making it easier to separate the strands [19]; topoisomerase proteins actively maintain this state by cleaving DNA [20]. Some extremophiles, conversely, have a positively supercoiled genome to prevent denaturation of DNA by extremely acidic or high-temperature environments [21]. Dynamic changes to the level of supercoiling are introduced by important metabolic processes such as transcription [18].

DNA topology plays a critical role in the regulation of metabolic processes [18, 22]. For example, the epigenetic switch responsible for switching between a dormant, lysogenic prophage state and lytic reproduction in bacteriophage λ is significantly more efficient and cooperative when the relevant DNA is supercoiled [23]. Supercoiling-dependent modulation of the *lac* operon is also observed, implying that supercoiling could act as a signalling pathway for gene regulation and that the *lac* operon may be sensitive to small changes in local DNA topology; the mechanism of the operon requires the formation of DNA loops, which have two supercoiling-dependent conformations [12, 14].

Molecular dynamics simulations of supercoiled DNA have demonstrated additional ligand–DNA contacts that would not be possible in torsionally relaxed DNA, and associated interference between ligands bound to distal sites [24, 25], and

experimental results suggest that structures associated with supercoiled DNA are likely to form directly upstream of the start sites of gene transcription, suggesting that topology plays a role in marking the start of genes [26].

Understanding the important and diverse effects of DNA supercoiling, and the factors that control it, will have a significant effect on fields including gene therapy [27] and biotechnology. The ability to predict and control genome topology could allow the manufacture of bespoke genetic switches to control gene expression. Effective utilisation of DNA supercoiling may allow regulatory signals to be transmitted across long distances along the double helix, and distal ligands can be made to interact, opening up new ways to control gene expression without modifying the nucleotide sequence.

1.1.2.1 Linking number

Within a bounded topological domain—that is, within a region of DNA with fixed ends beyond which torsion cannot be transferred—the overall degree of supercoiling is fixed. An obvious example of a topologically constrained system is covalently closed circular DNA (ccDNA) such as a plasmid or a typical bacterial genome; a DNA minicircle is a smaller piece of ccDNA, with a length of up to around a thousand base pairs. In such a system, no global relaxation of torsion is possible as this would require that covalent bonds be broken and reformed, or else that the strands pass through one another. This is encapsulated by a mathematical quantity known as the linking number, Lk, which is provably invariant for any two intertwined closed curves—such as the two strands of ccDNA—in three-dimensional space [28]. The value of Lk is an integer that may be positive or negative, depending on the handedness of the crossings; for DNA, Lk is defined so that its value for torsionally relaxed B-DNA is positive. The notation ΔLk is typically used to denote the change in the linking number of a piece of DNA relative to that which it would have if torsionally relaxed (denoted Lk_0).

The linking number is, intuitively, the number of times each curve winds around the other. Formally, a pair of closed curves is homotopic to any other pair of closed curves with the same linking number. The value of Lk can be determined by continuously deforming the curves to lie in a plane, allowing them to pass through themselves but not each other; the curves are then assigned a consistent directionality (clockwise or anticlockwise) and the crossings are categorised according to the right-hand rule, with conforming crossings counted as positive and nonconforming (left-handed) crossings counted as negative. The linking number is then

$$Lk = \frac{1}{2}(N_+ - N_-)\,, \tag{1}$$

where N_+ and N_- are the numbers of positive and negative crossings, respectively.

Actually performing the correct series of continuous deformations in order to obtain a standard link diagram is impractical, and devising a robust and universal definition of a crossing in three dimensions is more difficult than it might appear. In order to calculate the linking number for an arbitrary link, it is helpful to map the system onto a two-dimensional surface, so that crossings are simply the points where the projected curves meet. For a closed curve γ, the Cartesian coordinates, $\mathbf{r} \in \mathbb{R}^3$, of each point along γ can be expressed as a function of the arc length, s, a scalar (in the domain S^1) representing the distance travelled along the curve. Consider the pair of closed curves

$$\gamma_1 = \{\mathbf{r}_1(s_1) : s_1 \in S^1\} \quad \text{and} \tag{2a}$$
$$\gamma_2 = \{\mathbf{r}_2(s_2) : s_2 \in S^1\}. \tag{2b}$$

To these curves one can apply the Gauß map, $\mathbf{\Gamma}$, which maps each point (s_1, s_2) in the link to the surface of a unit sphere. The Gauß map is

$$\mathbf{\Gamma}(s_1, s_2) = \frac{\mathbf{r}_1(s_1) - \mathbf{r}_2(s_2)}{|\mathbf{r}_1(s_1) - \mathbf{r}_2(s_2)|}\,; \tag{3}$$

for a point $\boldsymbol{\nu}$ on the unit sphere corresponding to a crossing where γ_1 passes over γ_2, a projection of the link into the plane perpendicular to $\boldsymbol{\nu}$ produces a well defined link diagram. All crossings map to $\boldsymbol{\nu}$, and the neighbourhood of each crossing maps to the neighbourhood of $\boldsymbol{\nu}$ with its orientation preserved or reversed depending on the sign of the crossing. Thus, the linking number is the number of times the Gauß map covers the point $\boldsymbol{\nu}$ with positive orientation minus the number of times the Gauß map covers the point $\boldsymbol{\nu}$ with negative orientation. This is equal to the signed number of times the Gauß map covers the unit sphere, or the signed area of the Gauß map divided by the surface area of the unit sphere.* The signed area of the Gauß map can be calculated by integrating over the entire domain of the contours γ_1 and γ_2, and the linking number can in this way be obtained. This is the Gauß linking integral [28, 29],

$$\begin{aligned} \mathrm{Lk}(\gamma_1, \gamma_2) &= \frac{1}{4\pi} \oint_{\gamma_1} \oint_{\gamma_2} \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|^3} \cdot (\mathrm{d}\mathbf{r}_1 \times \mathrm{d}\mathbf{r}_2) \\ &= \frac{1}{4\pi} \int_{S^1 \times S^1} \frac{\mathbf{r}_1(s_1) - \mathbf{r}_2(s_2)}{|\mathbf{r}_1(s_1) - \mathbf{r}_2(s_2)|^3} \cdot \left(\frac{\mathrm{d}\mathbf{r}_1}{\mathrm{d}s_1} \times \frac{\mathrm{d}\mathbf{r}_2}{\mathrm{d}s_2} \right) \mathrm{d}s_1 \, \mathrm{d}s_2 \,. \end{aligned} \tag{4}$$

1.1.2.2 Twist and writhe

The linking number can be usefully partitioned into two quantities: twist (Tw), the number of times the strands coil about the helix axis; and writhe (Wr), the number of times the helix axis crosses itself.† While Lk is always an integer, Tw and Wr

* The proof of this requires relatively complex topology and some understanding of manifold theory, but is widely accepted by mathematicians and is here left as an exercise for the reader.

† These definitions are, of course, correct only for a double helix like DNA; the general definitions of these quantities in ribbon theory are somewhat more complex and of little importance here.

can take on non-integer values for three-dimensional curves; non-integer twist corresponds to a partial turn about the helix axis, while non-integer writhe occurs when the helix axis crosses itself when viewed from some angles but not others. The Călugăreanu–White–Fuller theorem [30–32] states that

$$\mathrm{Lk} = \mathrm{Tw} + \mathrm{Wr}\,. \tag{5}$$

In a torsionally relaxed piece of DNA, Lk is typically partitioned entirely into twist, but as $|\Delta \mathrm{Lk}|$ increases, it becomes increasingly favourable for the helix axis to coil about itself, partitioning some of the torsion into writhe in return for relaxing some twist. A piece of DNA with non-zero writhe is a plectoneme. The same effect can be observed in the macroscopic world by twisting any coiled cable or piece of string—while small amounts of applied torsion can be absorbed by twisting about the axis of rotation, writhed structures will spring into being after some number of turns.* A number of factors affect the balance between twist and writhe, including the nucleotide sequence, the surrounding solvent environment, and the presence of structural influences such as DNA-binding proteins. There is currently no widely accepted model that can reliably predict the effect of the complex interactions between these factors.

Attempts to calculate writhe for an arbitrary curve suffer from similar problems to those encountered when considering the linking number. Intuitively, writhe is roughly akin to the linking number of a curve with itself, although there exists an important distinction: continuous deformations may change the writhe of a system but not its linking number. This leads to the significant conclusion that the writhe of a system depends on the angle from which it is viewed, or the plane onto which it is projected. The overall writhe is thus defined as the average of the apparent writhes from all angles. The writhe of the closed curve

$$\gamma = \{\mathbf{r}(s) : s \in S^1\} \tag{6}$$

can be calculated as [29]

$$\mathrm{Wr}(\gamma) = \frac{1}{4\pi} \oint_\gamma \oint_\gamma \frac{\mathbf{r}(s_1) - \mathbf{r}(s_2)}{|\mathbf{r}(s_1) - \mathbf{r}(s_2)|^3} \cdot [\mathrm{d}\mathbf{r}(s_1) \times \mathrm{d}\mathbf{r}(s_2)]\,, \tag{7}$$

which is very similar to equation 4, except that s_1 and s_2 are points along the single curve γ, rather than along two distinct curves γ_1 and γ_2.

The twist is trivially $\mathrm{Tw} = \mathrm{Lk} - \mathrm{Wr}$, but can also be calculated for a double helix as the number of times each strand twists around the helix axis, and an integral similar to those described above can be formed accordingly. The simple nature of a helical system, in contrast to an arbitrary link, allows this integral to be simplified

* In fact, the word *plectoneme* is derived from the Greek πλεκτό νῆμα (*plektó nêma*), meaning "twisted thread".

considerably, so the twist of the curve γ about its axis (the local direction of which is $\hat{\mathbf{n}}(s)$) is given by [29]

$$\mathrm{Tw}(\gamma) = \frac{1}{2\pi}\oint_{\gamma}\left(\frac{d\mathbf{r}}{ds}\times\hat{\mathbf{n}}(s)\right)\cdot\frac{d\hat{\mathbf{n}}}{ds}\,ds\,. \tag{8}$$

For the DNA double helix, the contours γ_1 and γ_2 are the backbones of the two DNA strands and the set $\{\mathbf{n}(s) : s \in S^1\}$ is the helix axis (or "molecular contour"), the precise computation of which will be discussed in chapter 2.

1.1.2.3 Plectoneme structure

The linking number reflects the global topology of a piece of DNA, but the additional twist need not be spread evenly throughout the entire topological domain. Torsion may pass along the DNA helix as a wave, or become concentrated in a particular area due to some combination of sequence and structural factors. A useful quantity for the investigation of local supercoiling is the superhelical density, σ; the global superhelical density is defined as

$$\sigma = \frac{\Delta\mathrm{Lk}}{\mathrm{Lk}_0}\,. \tag{9}$$

This value, also referred to as the specific linking difference, represents the change in the linking number relative to its relaxed value, so does not depend on the length of the DNA segment under consideration; the local superhelical density can be estimated for each point along the helix axis, and the mean value of this quantity over the whole contour should be equal to the global value.

There is a great deal of structural diversity within the conformational space of plectonemes even for short, rigid pieces of DNA [33]. Generally,

$$0 \leqslant \frac{\mathrm{Wr}}{\Delta\mathrm{Lk}} \leqslant 1\,, \tag{10}$$

so the helix axis can cross itself up to $|\Delta\mathrm{Lk}|$ times.* For small $|\Delta\mathrm{Lk}|$, this results in simple plectonemes; the simplest structure with $|\mathrm{Wr}| = 1$ is a simple figure-8, for example. As the writhe—and hence the number of crossing points—increases, the structures become more complex and, relative to an open circle, more compact. The compaction of DNA is enhanced by applying supercoiling in either direction, with highly compact rods formed for $|\Delta\mathrm{Lk}| \gg 0$ [33].

Twist and writhe can only absorb so much torsion, however, without disrupting DNA base pairing. It is frequently energetically favourable to form defects such as kinks—which occur when stacked base pairs become unstacked, resulting in a sharp bend—and bubbles—in which the base pairs in a region of DNA break

* Mathematically, there is nothing preventing $\Delta\mathrm{Tw}$ and Wr from having opposite signs, allowing values outside this range, but this would be highly unusual in reality.

apart. It is thought that DNA can reversibly transition between smooth bends, kinks, structures with individual flipped bases, and denatured bubbles [33]. These structures perform useful biological roles but are difficult to model and to predict. Nicking DNA by creating a break in the sugar–phosphate backbone allows the stored torsion to be released and relaxes supercoiling [33].

While the mathematical definitions of the topological properties discussed above are formulated for a pair of closed curves, this is not the only way in which DNA can be torsionally constrained. Bound proteins, particularly those that severely disrupt the local structure or bridge distal DNA sites, can divide DNA into distinct, linear topological domains with fixed linking numbers [34]. This provides a biological tool to control the formation of plectonemes, and determining the extent to which proteins divide DNA into topological domains is an important problem in the physics of life.

1.2 Nucleoid-associated proteins

Genome structure and gene expression are also regulated by bending of the double helix. This is vital for compaction of the genome—the 4.6 Mbp genome of *E. coli* [35] has a contour length of around 1.5 mm (assuming a rise of 3.32 Å), 750 times longer than the cell's 2.0 µm length, so efficient and predictable packaging is important to ensure that the genome remains accessible despite being compacted into a small space. Perhaps an even more significant role of DNA bending is in gene regulation. Nucleoid-associated proteins are responsible for genome packaging and DNA bending in prokaryotes, and have myriad varied roles in nucleoid structuring and gene expression, which will be discussed beginning in section 1.2.2, following a brief overview of protein structure and function.

1.2.1 Overview of proteins

Proteins are large biomolecules made up of one or more chains of amino acids (figure 2). There are twenty proteinogenic amino acids, distinguished by their substituent side chains. Encoded by DNA and produced in ribosomes, proteins are responsible for a titanic array of biological functions including cell structuring, intracellular transport, and the response of cells to stimuli. Among the most important roles of proteins is the catalysis of biochemical reactions by a category of proteins known as enzymes, which bind to substrates and facilitate their metabolism into useful products; enzymatic catalysis is required in order for the vast majority of metabolic reactions to occur at a rate sufficient to sustain life. Another key function is gene regulation, which involves the binding of specialised proteins to specific DNA sites, resulting in structural changes and crowding effects

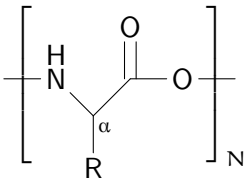


Figure 2: A polypeptide consists of a chain of amino acids with a repeating N−C−C−O backbone; an amino acid monomer is illustrated here. The label R indicates the site of attachment of the side chain to the α carbon; there are twenty possible side chains, which vary in size and charge and are not illustrated here. (In one amino acid, proline, the side chain is also attached to the nitrogen atom of the amine group; this attachment is not illustrated in this diagram.)

that either enhance or inhibit the expression of particular genes; an example of this was discussed in section 1.1 in the context of the *lac* operon.

Proteins consist of long, unbranched polypeptide chains—that is, chains of covalently linked amino acids. These fold, through noncovalent interactions,* into a number of secondary structural motifs including α helices (right-handed helices in which every backbone N−H group is joined by a hydrogen bond to the C=O group occurring three or four positions earlier in the sequence) and β sheets (sets of strands connected laterally by at least two or three hydrogen bonds to form a pleated sheet).

These local structures further organise into an overall tertiary structure, with the secondary structural motifs folded together by nonlocal interactions, most notably the tendency of hydrophobic amino acids to orient towards the core of the protein to avoid the surrounding solvent. Hydrogen bonds and other local interactions, however, continue to play a role.

Furthermore, many important proteins actually consist of multiple disconnected polypeptide chains joined together into an overall quaternary structure that functions as a single entity. In this context, each polypeptide chain is considered a "subunit" of the protein; proteins consisting of a single chain are termed monomers, while those with two subunits are dimers, those with three subunits trimers, and so forth. These subunits may be identical to one another, forming a homodimer (or homotrimer, homotetramer, *et cetera*), or at least one may be distinct, forming a heterodimer (or heterotrimer, *et cetera*).

1.2.2 Integration host factor (IHF)

Integration host factor (IHF) (illustrated in figure 3) is one of the most abundant proteins associated with the bacterial chromosome, present in Gram-negative bacteria such as *E. coli*. A 22 kDa nucleoid-associated protein (NAP), it binds to

* Covalent disulphide bridges between distal cysteine residues are also important to the secondary and tertiary structures of many proteins, but these occur relatively infrequently.

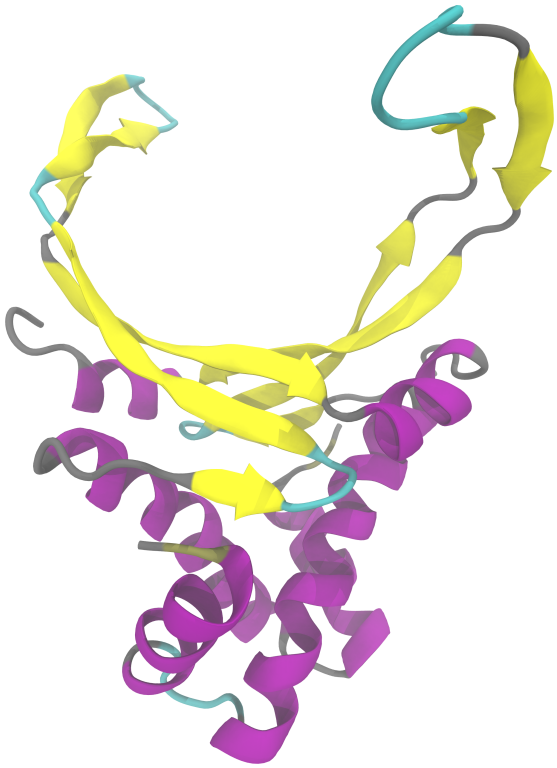


Figure 3: IHF consists of an α-helical core (purple helices) and a pair of β-ribbon arms (yellow arrows) that slot into the DNA major groove and have proline residues at their apices which intercalate between base pairs. In this image, the DNA, if present, would pass between the arms with the helix axis roughly perpendicular to the page.

DNA and induces a sharp bend of up to 160° [36]. While capable of nonspecific binding, IHF is observed to bind much more strongly to the consensus sequence WATCARNNNNTTR [37] (W is A or T; R is A or G; N is any nucleotide—A, C, G, or T).

The protein exists *in vivo* as a heterodimer, with two similar but distinct subunits each consisting of a primarily α-helical core attached to a β-ribbon loop with a proline at its apex. These arms extend into the minor groove of bound DNA with the prolines located 9 bp apart, each intercalated between two base pairs, where they disrupt the local structure of the DNA and form a flexible hinge that is thought to be responsible for the protein's strong DNA-bending ability [36]. The consensus sequence sits to the right (downstream) of this central region, with an AT-tract located to the left (upstream) in most *in vivo* IHF binding sites.

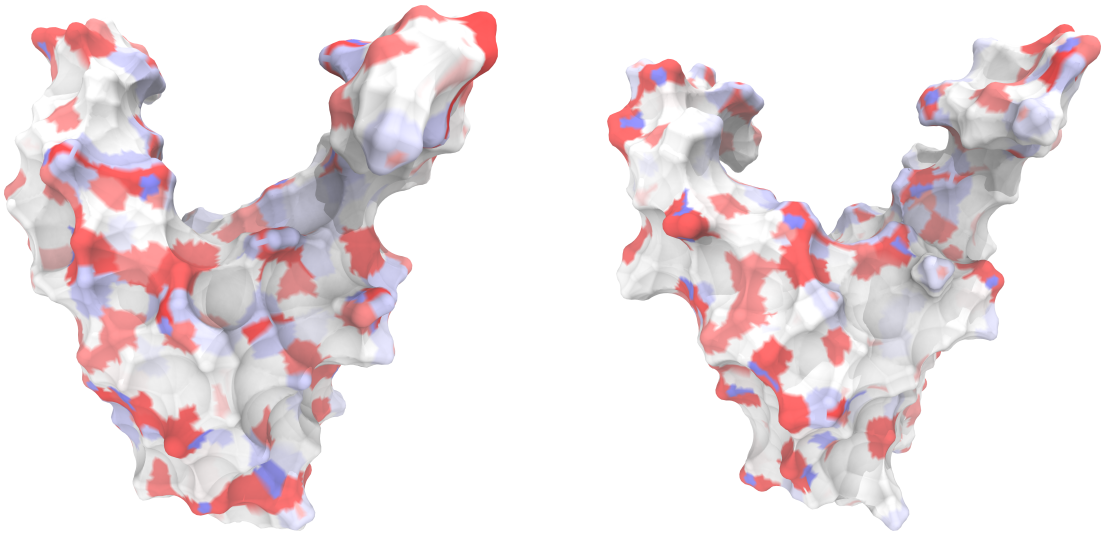
The primary function of IHF is to compact prokaryotic DNA, facilitating the assembly of higher-order nucleoprotein complexes, but it also plays myriad other roles in governing chromosomal architecture and dynamics. DNA bending by one or both of IHF and the structurally similar HU has been implicated in the regulation of around 120 genes in *E. coli* [38], as well as the initiation of chromosome replication [39] and the facilitation of recombination reactions [40]; these roles

make IHF important to the specificity of the CRISPR-Cas gene-editing system [41]. IHF-mediated supercoiling is associated with undertwisting and negative writhe, resulting in the formation of closed DNA loops [42]. These are an important gene regulatory mechanism; for example, binding of IHF to the promoter region of a *Pseudomonas* plasmid allows RNA polymerase to interact with a distally bound regulatory protein [42].

IHF has also been demonstrated to be of vital importance to the stability of certain biofilms, a type of microbial community that is present in approximately 80 % of chronic infections and can present significant challenges to treatment [43]. In a number of biofilms, such as those formed by *Pseudomonas æruginosa*, large quantities of DNA are pumped out of the cell to form an interwoven lattice of extracellular DNA (eDNA) with IHF located at the vertices [44]. The importance of IHF to biofilm stability is demonstrated by observations that an anti-IHF serum can reduce the size and mass of a *Burkholderia cenocepacia* biofilm by 44–56 % [45], and its synergy with traditional antibiotics makes it a promising target for next-generation treatments.

In the canonical structure of the IHF–DNA complex derived through X-ray crystallography (as archived in the Protein Data Bank (PDB) entries `1IHF` [46] and `5J0N` [47]), the protein induces a bend of around 160° by interacting with a 35 bp region of DNA. More recent research based on fluorescence lifetime measurements indicates the existence of at least one additional partially bound state with an unknown structure and a significantly smaller bend angle [48].

Previous work has shown that binding occurs through a two-step process, beginning with a fast, sequence-nonspecific step that occurs on a ~100 µs timescale, followed by a slower, site-specific step on a millisecond timescale [49]. It has been suggested that these two steps are binding of IHF to straight DNA followed by bending of the DNA as it wraps around the protein [50]. The second (bending) step is associated with an activation energy of approximately 14 kcal mol^{-1} (~24k_BT, where k_B is the Boltzmann constant and T is the thermodynamic temperature), which may be associated with proline intercalation [51], while the free energy associated with wrapping is found to be only around 3.6 kcal mol^{-1} (~6k_BT) [52]. Although these values are known, the underlying binding mechanism remains poorly understood. Little is known about the structure of the IHF–DNA complex at each stage of the binding process, and it is unclear why the free-energy associated with wrapping should be so close to the thermal energy scale.



(a) IHF **(b)** HU

Figure 4: IHF (a) and HU (b) have very similar surface charges, with a mostly neutral surface (white) studded with individual positively charged residues (blue) amidst patches of negative charge (red). The surface charge is here defined by the charge of the closest residue to each point on the MSMS surface [54]. This figure is not to scale.

1.2.3 HU

HU, the heat-unstable protein,* is superficially similar to IHF, with the same α-helix body and β-ribbon arms with intercalating prolines at their apices, but their amino acid sequences differ substantially; together, they are members of the DNABII family of proteins [36]. Like IHF, HU is dimeric, but, while IHF is usually an obligate heterodimer, with α and β subunits combining to form a single (αβ) IHF protein, HU can in most cases also exist as an αα or ββ homodimer (although the latter only in small quantities, typically close to the end of a cell's life) [53].† While IHF exhibits a strong affinity for specific binding to a particular DNA sequence, far outweighing its weak nonspecific binding, HU does not have any such preference, but instead binds almost exclusively to bent or damaged DNA [36].

While it is generally agreed that IHF binding is associated with a bend angle of around 160°, the range of observed values for HU varies considerably from 70° [55] to 140° [56]. The similar surface electrostatic profiles of the two proteins (figure 4) indicate broadly similar bending mechanisms despite these differences. An alternative binding mode has been observed for HU, in which the arms do not

* HU is alternatively taken to stand for "histone-like protein from *E. coli* strain U93".

† Note that the α and β subunits are two separate polypeptide chains that bind together non-covalently to form a single protein; these are not to be confused with α helices and β sheets, which are structural motifs within a single polypeptide chain. That they happen to be denoted by the same set of Greek letters is an unfortunate coincidence.

wrap around the DNA and the prolines do not intercalate; instead, the body of the protein lies parallel to the DNA, resulting in only a small bend [57]. While the proteins' similar electrostatic profiles suggest that a similar binding mode may exist for IHF, this has not yet been observed.

Most prokaryotic genomes contain at least one member of the DNABII family, and in the majority of cases this protein is more similar to *E.coli* HU than to *E.coli* IHF; those bacteria that do encode an IHF-like protein typically also encode an HU-like protein [36]. HU is thus an almost universal bacterial protein and plays a number of important roles in gene regulation. Among them is its importance to the *gal* operon, which switches on and off production of the enzymes necessary for metabolism of the sugar galactose in *E.coli* based on the balance of sugars in its surroundings. In this process, DNA looping is necessary in order to block the access of RNA polymerase to the promoter region, preventing transcription; this looping is facilitated by HU [58, 59].

Exploring the similarities and differences between the binding and bending behaviours of IHF and HU can provide deeper insight into their distinct regulatory behaviours, their roles inside and outside the cell, and the DNA-binding behaviour of nucleoid-associated proteins in general. In particular, one could hope to elucidate the many factors that determine the strength of binding to, and degree of bending of, DNA by NAPs.

1.2.4 Other nucleoid-associated proteins

1.2.4.1 H-NS

The heat-stable nucleoid-structuring protein (H-NS) is an abundant bacterial protein of around 15 kDa, the primary function of which is mediation and moderation of DNA topology. It is aided in this pursuit by a specialised oligomerisation domain that allows it to form large complexes with other copies of itself bound to distal DNA sites. This clustering, which results in large loops of DNA bridged by many proteins, is the most distinctive signature of H-NS binding. The oligomerisation domain, situated at the protein's N terminal, consists primarily of an elongated cluster of intertwined, antiparallel α helices; this is joined by a linker of around 30 amino acids to the smaller DNA-binding domain [60].

H-NS and similar proteins exhibit a preference for binding to curved DNA such as the promoter regions of genes, but have also been observed to bind to RNA [61]. By binding to the promoter region of a gene, H-NS represses transcription of that gene; the exact mechanism underlying this repression is not known for all affected promoters, but is thought to involve the inhibition of transcriptional elongation by the trapping of RNA polymerase in an H-NS-bridged DNA loop. Gene repression by H-NS can be relieved by local topological changes [62], suggesting

a supercoiling-dependent function; bound H-NS also maintains the superhelical density of DNA, even if strands are nicked and religated [60]. Many transcriptional activators, including Fis, have been shown to be antagonistic to H-NS.

StpA is a less common paralogue of H-NS, thought to have originated from a common ancestral protein through a genetic duplication event. The two proteins have very similar sequences, structures, and behaviour, even to the extent that they are able to oligomerise together, but studies on *E. coli* have demonstrated distinct behaviour and codependence [63], and StpA has a unique RNA-annealing ability [64].

The bridging of DNA by H-NS has been well studied and is quite thoroughly understood. This stands in contrast to IHF, which appears to bridge DNA through an unknown mechanism with unknown stoicheiometry, structure, and properties. H-NS thus represents a potential model for this behaviour, as well as being an interesting protein in its own right with important roles in gene expression and the division of supercoiled DNA into topological domains.

1.2.4.2 Fis

Fis is an 11 kDa homodimeric protein, which folds into four α helices and a pair of β hairpins [65] and binds specifically to the consensus sequence KNNYRNNWNNYRNNM [66],* although it is thought that its true preferences are both more specific and more complex than this [65]. Unlike in IHF and HU, the β hairpins are not involved in Fis's primary DNA-binding activity.

Like IHF and HU, however, Fis does bend DNA, with bend angles in the range 40–90° having been measured. This bending results in the formation of tightly bent DNA microloops, which enhance gene expression when located in certain regions of the promoter sequence. These microloops are thought to compensate for Fis-induced reduction of the superhelical density of DNA towards the relaxed state [67]. This provides fine-grained control over *in vivo* supercoiling; *E. coli* uses this to significantly and dynamically adjust the distribution of supercoiling in its genome in response to environmental factors and its own growth phase, although the effects in the otherwise similar *Salmonella enterica* are much smaller for reasons that are not yet well understood [68].

DNA bending makes comparison of Fis to IHF and HU natural, while the extent of its impact on genome topology makes it relevant to any discussion of DNA supercoiling, and the differing size of this effect between species provides an interesting example of how complex and significant the many factors affecting DNA topology can be. Comparison of the bending and bridging behaviours of multiple NAPs is likely to lead to deeper understanding of the mechanisms by

* K is G or T; Y is C or T; R is A or G; W is A or T; M is A or C; N is any nucleotide.

which they all function and the complex interplay between DNA supercoiling and protein binding.

Chapter 2

Methods

2.1 Molecular dynamics

Molecular dynamics (MD) is a simulation method for the modelling and analysis of the movements of atoms and molecules. In an MD simulation, a system consists of distinct components with defined positions that are evolved forwards in time in discrete steps by the application of a set of simple physical rules. These components are frequently atoms, but in order to simulate larger systems it is often necessary to reduce the number of components by considering instead the movement of larger, multiatomic segments, at the expense of some accuracy.

Each step of a simple MD simulation consists roughly of the following sub-steps, each of which is much more complex than the simple description provided here might indicate and warrants individual discussion. For accuracy, efficiency, and numerical stability, MD software frequently performs these steps in a different order, or adds additional steps, as discussed in the sections below, but this framework should be sufficient for a general understanding.

1. Calculate the force, and thus the acceleration, experienced by each component according to a defined force field.
2. Move the components according to their velocities, and adjust their velocities in accordance with the calculated acceleration.
3. Apply boundary conditions, a temperature-regulating thermostat, and a pressure-regulating barostat, if necessary, to ensure the simulation remains within physical bounds and accurately models the canonical ensemble.
4. Repeat until the desired number of time steps have been completed.

2.1.1 Initialising an MD simulation

A prerequisite for an MD simulation is something to simulate; before the forces can be calculated and applied, one must provide something to which to apply them.

This is less trivial than it may appear, and selecting the right initial structure can be crucial to the success of a simulation [69].

DNA has a predictable double-helical structure consisting of repeating units with known geometric attributes. It is thus trivial to generate an ideal piece of B-DNA with any nucleotide sequence using a utility such as the Nucleic Acid Builder (NAB) [70], which is included with the AMBER* software suite [71]. Building a protein, on the other hand, would be a decidedly non-trivial undertaking. While a simple polypeptide chain could be produced following a similar methodology, the vast majority of useful proteins are folded in intricate and important ways to produce complex three-dimensional structures that cannot be easily predicted from first principles. The structures of interesting proteins must, therefore, be obtained experimentally. The atomic structures of biological systems containing proteins are usually determined through X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or electron cryomicroscopy. These structures are deposited in the Protein Data Bank [72] by researchers worldwide and are accessible online; this resource is highly valuable as a starting point for MD simulations, although it is common to wish to simulate a system somewhat different from that which has been crystallised and imaged experimentally.

In such cases, it is necessary to manipulate the structure. In many cases, this means simply extracting the interesting part of a larger system in order to simulate it separately; in others, more involved manipulation is required. For example, consider a crystal structure in which a protein of interest is bound to a segment of DNA and disrupts its structure. In order to simulate the same protein bound to a longer piece of DNA with a different sequence, it is necessary to build the desired sequence (for example, with NAB), cut out a section containing the protein's binding site, and insert the protein and nearby DNA from the crystal structure, mutating the nucleobases as necessary. This is a complex, multi-step process that is very difficult to automate; some graphical tools such as PyMol [73] exist to facilitate this type of manipulation, or one can align the structures in such a way as to minimise the root-mean-square deviation (RMSD) between the desired parts of the structures and edit PDB files in a text editor or using a simple script. An element of trial and error is usually necessary, and it is important to carefully check that the resulting structure is sensible and evolves as expected when simulated.

It is very unlikely that atom positions and bond lengths following such a process will be optimal. If two atoms are too close together when the system is simulated, they are likely to experience a repulsive force strong enough to tear the system apart, a turn of events that does not result in a useful trajectory. Bond lengths and angles that are outside the range of physical values for which the force field

* Assisted Model Building with Energy Refinement

was calibrated can result in similar problems and, even in the absence of such manipulations, the insertion of solvent molecules can also cause clashes. It is thus prudent to find the structure that minimises a system's free energy prior to initiating molecular dynamics [69]; thankfully, this can be done automatically and is a standard feature of most MD software.

2.1.2 Force fields

After defining a sensible initial structure, it is necessary to also define the set of rules by which the atoms in a system will interact. This is done by means of a potential field constructed by considering the interactions between each pair of atoms. The set of parameters defining this potential is known as the "force field".

The simplest interatomic potential capable of accurately describing a real system is a Lennard-Jones potential [74–76], which has the form

$$V_{LJ} = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right], \tag{11}$$

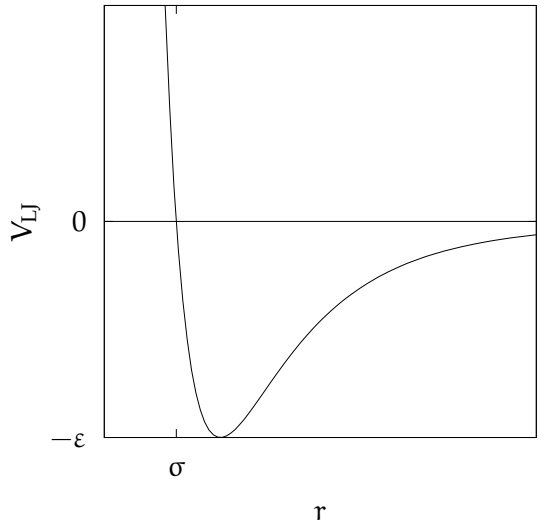
where ε is the depth of the potential well, r is the distance between the interacting particles, and σ is the value of r at which the potential is zero (figure 5a). The term proportional to r^{-6} represents attractive van der Waals forces due to the fluctuation of partial charges; its form arises from the observation that London dispersion forces fall off with r^{-6} [77]. The term proportional to r^{-12} represents the repulsive effect of the Pauli exclusion principle; the power of 12 was selected for computational convenience since it allows the Lennard-Jones potential to be computed as $V_{LJ} = 4\varepsilon(a^2 - a)$, where $a = (\sigma/r)^6$, so the expensive division and sixth power need only be computed once.

A potential of this form would be entirely sufficient for many simulations, such as a fluid of uncharged argon atoms [74, 75], and even accurately models swarm behaviour with small modifications [78], but systems of molecules are somewhat more complicated. Perhaps the most obvious modification one could make is to add a Coulomb potential [79, 80] of the form

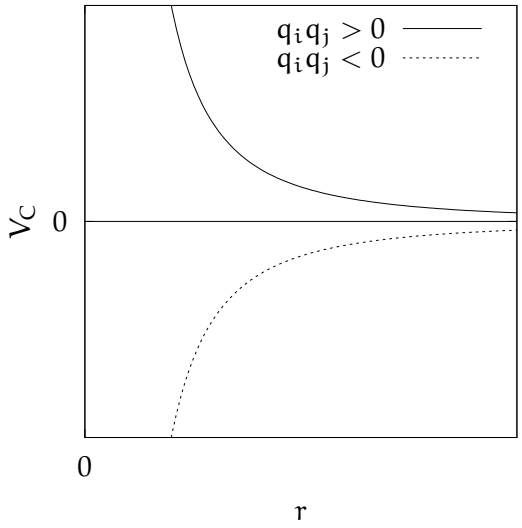
$$V_C = \frac{q_1 q_2}{4\pi\varepsilon_0 r} \tag{12}$$

(figure 5b) in order to model charged particles such as ions. Here, q_1 and q_2 are the charges of the atoms and r is the distance between them; ε_0 is the permittivity of free space. The charges involved need not be associated with ionisation of atoms i and j—the majority of atoms in a system differ in electronegativity from their neighbours and thus have some partial charge.

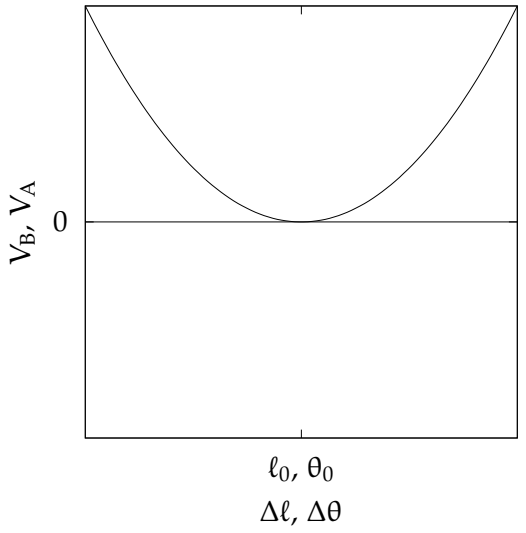
The inclusion of a Coulomb term brings us closer to a useful potential for the simulation of interesting systems, but still cannot model covalent bonds, which are of course an essential component of molecular modelling. A covalent bond between



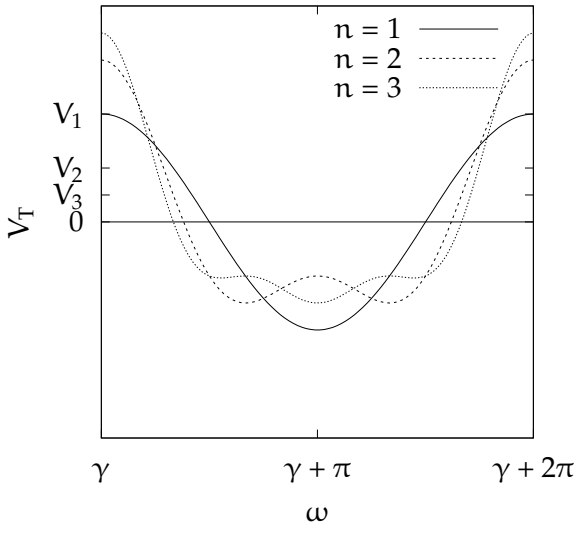
(a) Lennard-Jones potential



(b) Coulomb potential



(c) Hooke potential



(d) Torsion potential

Figure 5: Forms of potential terms in the AMBER potential

a particular pair of atoms will typically vibrate, the bond length ℓ oscillating about an equilibrium value ℓ_0; we can model this as a Hookean spring [81],

$$V_L = k_L(\ell - \ell_0)^2 , \tag{13}$$

with a spring constant k_L (figure 5c). This is good approximation for $\ell \approx \ell_0$ but becomes less appropriate at bond lengths far from equilibrium.

Due to the arrangement of electron orbitals, each pair of bonds involving the same atom will in general be separated by an equilibrium angle θ_0, but once again the true angle θ may fluctuate about this value; this can be modelled in a similar fashion using a second Hookean potential,

$$V_A = k_A(\theta - \theta_0)^2 , \tag{14}$$

with a different spring constant k_A.*

Finally, twisting about a bond requires significant torsion due to bond order (that is, while single bonds can generally rotate freely, double and triple bonds generally cannot), neighbouring bonds, or nearby lone electron pairs. To account for this, we can add an additional term to the potential representing the torsion of each bond. The torsion due to the relative rotation of bonds A−B and C−D about bond B=C is a non-trivial periodic function of the dihedral angle ω (the angle between A−B and C−D when projected onto the plane to which B=C is perpendicular) and is generally expressed as a Fourier series [82],

$$V_T = \sum_n V_n \cos(n\omega - \gamma) , \tag{15}$$

where V_n is the amplitude of the n-th term of the Fourier series (figure 5d). This is a periodic function (accounting for rotational symmetry) with its first maximum at the phase angle γ. The number of terms required varies, but three are usually sufficient. The parameters are chosen to account for various factors including the charge and size of the A and D groups. When combined with other terms as part of an MD force field, it is common to define each term of the Fourier series as $V_n[1 + \cos(n\omega - \gamma)]$ in order to ensure that this component is always positive, although this offset is superfluous when considering this term alone as it has no effect on the locations or depths of the minima.

* Due to the cyclic nature of angles, the comparison of θ and θ_0 should actually be performed *modulo* 2π, lest the difference between the angles ε and $2\pi - \varepsilon$ (for small ε) be overstated.

The forms of all four types of potential are shown in figure 5. Combining these terms allows us to construct a generalised Amber potential for N atoms [83],

$$\begin{aligned} V &= \sum_{\text{bonds}} V_L + \sum_{\text{angles}} V_A + \sum_{\text{torsions}} V_T + \sum_{\text{pairs}} (V_{LJ} + V_C) \\ &= \sum_{i=1}^{n_B} k_{Li}(\ell_i - \ell_{0i})^2 + \sum_{i=1}^{n_A} k_{Ai}(\theta_i - \theta_{0i})^2 + \sum_{i=1}^{n_T} \sum_{n} V_{in}[1 + \cos(n\omega_i - \gamma_i)] \\ &\quad + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right), \end{aligned} \tag{16}$$

where n_B, n_A, and n_T are the numbers of covalent bonds, bond angles, and bond torsions, respectively. Modern force fields make a number of refinements to this model to improve accuracy by adjusting weights or correcting for computational error, or to improve computational efficiency by reducing unnecessary calculations [84, 85]. It is also rarely desirable to loop over every possible pair of atoms in an explicitly solvated system of interesting size, since this would scale with $\mathcal{O}(N^2)$,* the vast majority of the N atoms are likely to belong to uninteresting water molecules, and the interactions between distant atoms are likely to be very weak; the Lennard-Jones, Coulomb, and van der Waals terms in such a case may be calculated only for values of r_{ij} smaller than a defined cutoff.

All that remains is to determine values for the various constants in equation 16. This is a large set of constants, with each taking on a distinct value for every possible pair of atoms. It is obvious, for example, that an oxygen atom will behave differently from a carbon atom, but a carbon atom bonded to four of its own kind will also be distinct from a carbon atom whose bonded neighbours include an oxygen. Attempting to parametrise the Amber force field for every possible configuration of atoms from across the periodic table would be a Heracleian endeavour. Thankfully, biology is composed of a limited number of repeating units—four nucleotides and the twenty amino acids for which they code—which in turn comprise primarily the four elements carbon, oxygen, nitrogen, and hydrogen; other elements, such as phosphorus and sulphur, occur only in very limited positions. It is thus possible to parametrise a force field for exclusively biochemical application by determining, for example, the constants governing the interaction between the alpha carbon of a given amino acid and its neighbouring nitrogen, and so on. Of course, the differences between many types of atom are likely to be very small or nonexistent; for example, most alpha carbons can probably be expected to behave similarly. In this way, the size of the set of atom types—and with it the number of constants one must define—can be reduced substantially.

* This is "big O" notation, which describes the limiting behaviour of a function as its argument tends to infinity. When describing the time complexity of an algorithm, this is given by the order of the fastest-growing term as a function of the size of the input.

Perhaps the most satisfying way to derive these constants is *ab initio*—through quantum-mechanical calculations. This is indeed possible [86], and a number of force fields exist that are parametrised in this way [87]. However, such methods are still prone to inaccuracy when compared to experimental results, so state-of-the-art force fields such as ff14SB [88] and BSC1 [89] apply empirically derived corrections to carefully tune the values of the constants so that simulations most accurately replicate experimental observations.

The AMBER potential is not the only type of force field available. Alternatives include, for example, the CHARMM (Chemistry at Harvard Macromolecular Mechanics) potential [90], which includes bond length and angle, dihedral, and nonbonded (van der Waals and Coulomb) terms like those in AMBER but adds to these an "improper" term of the form

$$V_\omega = k_\omega(\omega - \omega_0)^2 \tag{17}$$

accounting for out-of plane bending of sets of atoms that would be expected to lie in a plane (where $\omega - \omega_0$ is the angle by which the system deviates from the plane), and a Urey–Bradley term accounting for interactions between pairs of atoms that are both bonded to the same atom (atoms A and C in the chain A−B−C),

$$v_u = k_u(r - r_0)^2 . \tag{18}$$

There is considerable debate about the best force field, but the latest versions of the AMBER and CHARMM force fields, with appropriately chosen parameters, are both known to give similarly accurate results [91]. Other force fields include GROMOS [92], OPLS [93], MMFF [94], and CFF [95], although these are not yet as widely used for biomolecular simulations.* While no force field is clearly better than all its competitors, it is of course necessary to choose one, and force fields of the AMBER form (ff14SB and BSC1) were used in this work.

2.1.3 Solvent models

2.1.3.1 Explicit solvent

Biology as we know it does not occur *in vacuo*. The biological environment is overwhelmingly aquæous, and charged molecules like DNA can be unstable if not surrounded by counterions. In order for a simulation to accurately describe the physics of biomolecules, the solvent environment must be modelled.

This is unfortunate, since the behaviour of water molecules, especially those far from the biomolecules in question, is generally of little interest, and yet water fills the vast majority of most systems. For example, a 10 bp piece of B-DNA has

* The details of these force fields are out of the scope of this discussion, but are available to the interested reader in the references provided.

a length of around 33 Å and contains around 650 atoms (with the exact number depending on its sequence); a cube of side length 33 Å would contain around 3600 water atoms at standard temperature and pressure, thus accounting for around 85 % of the atoms in the system despite being of little to no interest. To make matters worse, while the number of atoms in a piece of DNA of length N scales with $\mathcal{O}(N)$, the number of water atoms required to solvate it scales with $\mathcal{O}(N^3)$, so any system of interesting size becomes vastly outnumbered very rapidly indeed.

This is, of course, an oversimplification. The size of the solvent box can be reduced by simple modifications such as positioning the DNA so that it spans the box diagonally, or bending it so that it fits in a smaller box, and interesting systems are also likely to occupy a larger volume than does a simple linear piece of DNA. Of more universal utility is the observation that the box need not be a perfect cube. In fact, the only requirement is that it tessellate so that periodic boundary conditions can be applied, although it is sensible to restrict the search space to convex polyhedra whose faces are regular since such a box may be most easily constructed around a system of arbitrary shape. Among the convex polyhedra whose faces are regular, only five are space-filling: the cube, the triangular and hexagonal prisms, the truncated octahedron, and the gyrobifastigium (a solid formed by joining two triangular prisms along corresponding square faces with a relative rotation of 90° [96]). Furthermore, it is important that the molecule be allowed to rotate freely without interacting strongly with itself across the box boundary. The most desirable shape that maintains this constraint is one of maximum compactness, so that its volume (and thus the amount of solvent it contains) is minimised relative to the longest vector that can span it in all orientations, minimising the quantity of unnecessary solvent molecules in the corners distant from the molecules of interest. The most compact shape in three dimensions is a sphere, but this has the serious disadvantage that it does not

Table 2: Sphericity and properties of space-filling convex polyhedra

Polyhedron	Sphericity, Ψ	Regular	Parallelohedron
Triangular prism	0.716	✓	
Gyrobifastigium	0.767	✓	
Cube	0.806	✓	✓
Hexagonal prism	0.816	✓	✓
Elongated dodecahedron	0.880		✓
Rhombic dodecahedron	0.905		✓
Truncated octahedron	0.910	✓	✓

tessellate; we must thus seek the space-filling polyhedron whose sphericity [97],

$$\Psi = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}, \tag{19}$$

where A is the polyhedron's surface area and V is its volume, is closest to unity. The values of the sphericity for each regular space-filling polyhedron are shown in table 2; with a volume of $8\sqrt{3}a^3$ and a surface area of $\left(6 + 12\sqrt{3}\right) a^2$ for side length a, the truncated octahedron has sphericity 0.910, making it the optimal shape for an explicitly solvated MD simulation. An additional beneficial property of the truncated octahedron is that it is a parallelohedron [98], which means its tesselation requires only translations, making its periodic boundary conditions somewhat easier to implement. The cube, hexagonal prism, rhombic dodecahedron, and elongated dodecahedron are also parallelohedra; as table 2 demonstrates, the truncated octahedron is also the most spherical shape in this category.

Despite this optimisation, however, the underlying issue remains: the solvent molecules are still numerous and contribute a great deal of computational expense. Thankfully, a number of simplifications are possible. These simplifications are based on the assumption that vibrations of the O−H covalent bonds within water molecules do not affect the system's overall behaviour; this assumption allows the bond length and angle terms to be discarded from the potentials experienced by water atoms, with the molecules instead behaving as rigid triangles with fixed side lengths and angles. The form of the potential for a water atom in such a model is thus

$$V_{H_2O} = V_{LJ} + V_C\,; \tag{20}$$

in practice, the Lennard-Jones term need only be applied to the oxygen atom, with the two hydrogens interacting only via the Coulomb force. This is the basis for the TIP3P (transferable intermolecular potential with 3 points) model [99], which is among the most widely used and has an interaction site corresponding to each of the three atoms; four-site, five-site, and even six-site models exist [100], each adding additional dummy atoms for increased accuracy, but the gains associated therewith are minimal for a typical simulation and generally do not justify the increased computation required.

While this simplification substantially reduces the number of calculations required, the presence of solvent continues to slow down simulations for another, more physical reason: solvent is viscous, particularly at the length scales typical of atomistic simulations.

2.1.3.2 Implicit solvent

An alternative approach is to represent the solvent as a continuous medium with particular dielectric properties. Such models are referred to as "implicit solvent"

models, and typically include the effect of ions in their estimation of the dielectric properties of the solvent, so that both water molecules and solvent ions become unnecessary. This is best achieved by considering the free energy of the solvent, ΔG_s, which can be divided into an electrosatic part, ΔG_{el}, and a non-electrostatic part, ΔG_{nonel}, so [101, 102]

$$\Delta G_s = \Delta G_{el} + \Delta G_{nonel} \,. \tag{21}$$

The nonelectrostatic part, which represents the free energy of solvating a solute molecule from which all charges have been removed, originates from solvent–solute van der Waals forces and a cost associated with disrupting the structure of the solvent around the solute; a simple approximation is to consider this value to be proportional (with an empirically derived constant of proportionality) to the total solvent-accessible surface area of the solute.

The electrostatic part represents the free energy of removing all the charges from the solute in a vacuum and adding them back in the presence of the solvent. This is given by [103]

$$\Delta G_{el} = \frac{1}{2} \sum_{i=1}^{N} q_i \psi(\mathbf{r}_i) \,, \tag{22}$$

where q_i is the charge on solute atom i (of N), which has position $\mathbf{r}_i$; ψ is the electric potential, the bulk distribution of which can in principle be described by the Poisson equation [104],

$$\nabla^2 \psi = -\frac{\rho_e}{\varepsilon \varepsilon_0} \,, \tag{23}$$

where ρ_e is the local electric charge density and ε is the dielectric constant of the solvent. In order to account for the free movement of ions in solution, this can be combined with the Boltzmann equation, which gives the local ion density c in terms of the bulk ion concentration c_0, as

$$c = c_0 \exp\left(\frac{-W}{k_B T}\right) , \tag{24}$$

where W is the work required to move an ion from an infinite distance. Since $W = \pm e\psi$, where e is the charge of an electron, with a sign depending on the sign of the ion's charge,

$$c_\pm = c_0 \exp\left(\frac{\mp e\psi}{k_B T}\right) . \tag{25}$$

The local electric charge density is thus

$$\begin{aligned} \rho_e &= e(c_+ - c_-) \\ &= c_0 e \left[\exp\left(\frac{-e\psi}{k_B T}\right) - \exp\left(\frac{e\psi}{k_B T}\right)\right] \\ &= -2c_0 e \sinh\left(\frac{e\psi}{k_B T}\right) . \end{aligned} \tag{26}$$

Substituting this into the Poisson equation (equation 23) results in the Poisson–Boltzmann equation [105],

$$\nabla^2\psi = \frac{2c_0 e}{\varepsilon\varepsilon_0}\sinh\left(\frac{e\psi}{k_B T}\right), \tag{27}$$

a nonlinear differential equation that cannot be solved efficiently. When the potential is small (strictly, $e|\psi| \ll k_B T$, although the results are generally valid for a broader range of potentials close to room temperature), the Poisson–Boltzmann equation can be linearised, in each dimension, to

$$\psi(x) = \psi_0 \exp\left(-\sqrt{\frac{2c_0 e^2}{\varepsilon\varepsilon_0 k_B T}}\,x\right); \tag{28}$$

this low-potential approximation is valid for normal MD simulations.

While a number of specialised Poisson–Boltzmann solvers exist [106], particularly for the linearised case, solving this equation remains inefficient. A further approximation results from modelling the solute as a set of spheres with an internal dielectric constant that differs from that of the surrounding solvent. This is the basis for the generalised Born model [102], which has the functional form

$$\Delta G_{el} = -\frac{1}{8\pi\varepsilon_0}\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon}\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{q_i q_j}{f_{GB}}, \tag{29}$$

where

$$f_{GB} = \sqrt{r_{ij}^2 + R_i R_j \exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)}; \tag{30}$$

ε is again the dielectric constant of the solvent, ε_{in} is the dielectric constant of the solute (usually set to 1), N is the number of solute atoms, r_{ij} is the distance between atoms i and j, and q_i and R_i are the charge and effective Born radius, respectively, of atom i. The effective Born radius of an atom is a measure of how deeply it is embedded within the solute, and represents the distance from the atom to the molecular surface; this is typically calculated using the Coulomb field approximation as

$$\begin{aligned} R_i^{-1} &= \rho_i^{-1} - I_i \\ &= \rho_i^{-1} - \frac{1}{4\pi}\iiint_\Omega \frac{1}{r^4}\,d^3\mathbf{r}, \end{aligned} \tag{31}$$

where ρ_i is the atom's intrinsic radius and I_i is its Coulomb integral, in which Ω is the volume outside atom i but inside the molecule ($r > \rho_i$). Numerically computing this integral is computationally expensive, so its value is frequently approximated by the pairwise descreening approximation method described by Hawkins, Cramer, and Truhlar (the GB-HCT model) [107]. Unfortunately, this method tends to underestimate the effective Born radii of buried atoms; Onufriev,

Bashford, and Case corrected for this in their GB-OBC model [108], which scales up the Born radii of buried atoms using an empirically derived set of parameters α, β, and γ, so

$$R_i^{-1} = \tilde{\rho}_i^{-1} - \rho_i^{-1} \tanh\left(\alpha\phi - \beta\phi^2 + \gamma\phi^3\right), \tag{32}$$

where $\tilde{\rho}_i$ is ρ_i minus some offset and $\phi = \tilde{\rho}_i I_i$. Both of these models use the van der Waals surface to define the solvent–solute boundary, since this is much simpler to compute than the true molecular surface, at the expense of some accuracy. The GB-neck model by Mongan *et aliis* partially corrects for this by additionally integrating over the "neck" regions formed by the molecular surface (but not included in the van der Waals surface) between pairs of nearby atoms [109]. The parameters used to define these neck regions have been derived separately for proteins and nucleic acids [110, 111].

The effects of electrostatic screening by monovalent ions in the solvent can be accounted for by incorporating into equation 29 the Debye–Hückel screening parameter κ [112], which is given by

$$\kappa = \sqrt{\frac{8\pi I}{\varepsilon k_B T}} \tag{33}$$

for a solution of ionic strength I, so that

$$\Delta G_{el} = -\frac{1}{8\pi\varepsilon_0} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{q_i q_j}{f_{GB}} \left(1 - \frac{e^{-\kappa f_{GB}}}{\varepsilon}\right). \tag{34}$$

The implicit generalised Born model provides the most efficient reasonable approximation to the linearised Poisson–Boltzmann equation available to current software, and is the standard choice of implicit solvation model for MD simulations. The Gibbs free energy is simple to integrate into a force field, provided the Born radii can be estimated accurately, and the parameters have been refined several times to give accurate results for most systems [111].

2.1.4 Integrators

Having defined the system to simulate and the rules by which it should evolve, there remains only the seemingly simple task of evolving it forward in time. This cannot, of course, be done continuously, as this would necessitate solving exactly the very complex system of differential equations governing the system's behaviour—there is no general analytical solution to the three-body problem, let alone for the motion of thousands of atoms in a complicated potential. Instead, the system must be evolved forward in discrete time steps of finite duration Δt.

The simplest explicit method for the numerical integration of ordinary differential equations is the Euler method. In this method, the position, $\boldsymbol{x}$, of a particle at time step $n + 1$ is given by

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{v}_n \Delta t\,. \tag{35}$$

The velocity, of course, also varies with time, as

$$\boldsymbol{v}_{n+1} = \boldsymbol{v}_n + \boldsymbol{a}_n \Delta t \,. \tag{36}$$

The acceleration, $\boldsymbol{a}$, can be calculated at each time step from the potential. The Euler method is simple to implement, but is inappropriate for most practical applications. For example, consider applying the Euler method to a simple harmonic oscillator obeying the differential equation

$$\frac{d^2x}{dt^2} = -\frac{k}{m}x \,, \tag{37}$$

which has the analytic solution

$$x(t) = x_0 \cos(\omega t) + \frac{v_0}{\omega} \sin(\omega t) \tag{38}$$

where

$$\omega = \sqrt{\frac{k}{m}} \,. \tag{39}$$

In order to solve this with Euler integration, the second-order differential equation can be replaced by the pair of first-order differential equations

$$\frac{dx}{dt} = v; \qquad \frac{dv}{dt} = -\frac{k}{m}x \tag{40}$$

and discretised so that

$$x_{n+1} = x_n + v_n \Delta t \tag{41a}$$

$$v_{n+1} = v_n - \frac{k}{m} x_n \Delta t \,. \tag{41b}$$

The total energy of this simple harmonic oscillator is given by

$$E = \frac{1}{2}kx^2 + \frac{1}{2}mv^2 \,, \tag{42}$$

so the energy at each time step under Euler integration is

$$\begin{aligned} E_{n+1} &= \frac{1}{2}kx_{n+1}^2 + \frac{1}{2}mv_{n+1}^2 \\ &= \frac{1}{2}k(x_n + v_n \Delta t)^2 + \frac{1}{2}m\left(v_n - \frac{k}{m}x_n \Delta t\right)^2 \\ &= E_n + \frac{1}{2}k\left(v_n^2 + \frac{k}{m}x_n^2\right)\Delta t^2 \,. \end{aligned} \tag{43}$$

Since $k > 0$ and $m > 0$, $E_{n+1} > E_n$ for all values of Δt. The total energy of a physical system should of course remain constant, so the Euler method is unstable for oscillating systems. This includes the Hookean potentials used to model covalent bonds in MD simulations.

Thankfully, this problem can be corrected by updating the position and velocity at staggered intervals, rather than simultaneously. That is,

$$\boldsymbol{v}_{n+\frac{1}{2}} = \boldsymbol{v}_{n-\frac{1}{2}} + \boldsymbol{a}_n \Delta t \tag{44a}$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{v}_{n+\frac{1}{2}} \Delta t \,. \tag{44b}$$

This is the leapfrog integration method, since the position and velocity calculations "leapfrog" over one another. Provided Δt is constant and $\Delta t \leqslant \frac{2}{\omega}$, the leapfrog method is stable for oscillatory motion and requires only the same number of calculations per step as the Euler method [113]. This is the integration method used by MD software including AMBER, albeit with modifications in order to apply the effects of restraints, temperature-conserving thermostats, and pressure-conserving barostats. The size of the time step Δt is limited by the frequency of the highest-frequency oscillation in the system.

Note also that the calculation of $\boldsymbol{v}_{\frac{1}{2}}$ is dependent on $\boldsymbol{v}_{-\frac{1}{2}}$, which may be poorly defined; estimating this value accurately can be crucial to the stability of a simulation. This is handled during the equilibration phase of an MD simulation, a short phase following minimisation during which the system is gradually heated from 0 K (at which every atom has velocity $\boldsymbol{0}$) to the desired temperature and the system is allowed to explore the conformation space until an equilibrium structure is obtained. This gradual heating allows atoms to take on appropriate velocities, and a Maxwell–Boltzmann distribution of velocities will normally arise naturally from the equations of motion during this stage. It is necessary to ensure that almost all the degrees of freedom in the system reach equilibrium at this stage if one wishes to extract from the following production stage any thermodynamic or structural measurements, as a failure to reach equilibrium is likely to result in incorrect values for many of the system's properties [69].

2.1.5 Constraints and restraints

The highest-frequency oscillation in a biological system is typically the vibration of covalent bonds involving hydrogen atoms, since these are particularly light. They are also, however, of little importance in most systems. The lengths of these bonds can therefore be constrained—held absolutely fixed—using the SHAKE algorithm to remove these oscillations [114]. Since the highest-frequency oscillation limits the size of the time step, SHAKE constraints allow a longer time step, thus allowing the simulation to progress faster.

Restraints are superficially similar to constraints, but are subtly different in important ways. Instead of holding a value absolutely fixed, as constraints like SHAKE do, a system of restraints instead adds an additional term to the potential in order to bias a certain coordinate in favour of a particular value while still allowing it to fluctuate about this value to an extent. This is more useful than an absolute constraint for most purposes other than removing an oscillation for computational expedience. The textbook use of restraints is for the refinement of molecular structures obtained through NMR spectroscopy, which provides an upper and lower bound on certain lengths, and they are thus referred to as NMR

restraints. These can act on distances, angles, and torsions, defined by two, three, and four atoms respectively. Distance restraints, for example, act on the distance, r, between a pair of atoms and are defined using four distances, $r_1 \leqslant r_2 \leqslant r_3 \leqslant r_4$. In the region $r_2 \leqslant r \leqslant r_3$, the potential term is zero; in each of the regions $r_1 \leqslant r \leqslant r_2$ and $r_3 \leqslant r \leqslant r_4$ it is a half-parabola corresponding to a Hooke potential with spring constants k_2 and k_3, respectively; for $r < r_1$ and $r > r_4$, the potential term becomes linear so that particularly large or small distances do not result in a potential strong enough to tear the system apart. In the case $r_1 = 0$, $r_2 = r_3$, $k_2 = k_3$, $r_4 \to \infty$, the potential is harmonic. Angle and torsion restraints work similarly [71].

NMR restraints can be used to further restrain the lengths of certain bonds, to prevent certain undesirable behaviours such as the diffusion of a ligand molecule out of a particular region of a protein, or to nudge the system towards a desired state, among other things. They provide a convenient mechanism by which to carefully modify the system's potential, so are of substantial utility for advanced sampling methods.

2.1.6 Boundary conditions

When allowed to evolve freely, all or part of the simulated system may find its way, through random Brownian diffusion, to the edge of the simulation box. This is particularly a concern for explicitly solvated systems, since these contain a solvent box of a fixed size; implicitly solvated systems can in principle translate freely as long as all their coordinates remain within the limits of floating-point arithmetic. The simple solution to this is to use periodic boundary conditions. Under such a scheme, a particle exiting the box on one side will reenter from the opposite side.* For a simple cubic box, periodic boundary conditions can be implemented by simply taking each coordinate *modulo* the side length of the box. For more complex shapes like the truncated octahedron described in section 2.1.3, the arithmetic becomes more involved but the principle remains the same.

This is important because biology rarely consists of isolated water droplets floating in a vacuum; a system of biomolecules exists in a crowded aqueous environment of effectively infinite size. In the absence of periodic boundary conditions, the solvent would simply diffuse into the surrounding vacuum, resulting in ever-decreasing density and leaving the molecules of interest in a near-vacuum. Periodic boundary conditions maintain the number and density of particles in the system. In order to avoid surface effects, it is further necessary that molecules can interact even across the boundaries, so that atoms very close to opposite ends of the box interact as though they were located very close together.

* Perhaps the most familiar example of periodic boundary conditions to most people is the game Pac-Man.

Of course, there are infinitely many periodic repeats of the system, so this would entail performing an infinite number of calculations at every step. This is easily rectified by adding a cutoff distance, so that no two atoms farther apart than a certain specified distance interact; the potential tends to zero for large enough distances, and the lower bound on the distance for which this approximation is valid is typically less than the size of the box.

2.1.7 Thermostats and barostats

2.1.7.1 Berendsen thermostat and barostat

The physical coordinates are not the only properties of the system that can deviate too far from their initial values. MD simulations are performed in a microcanonical (NVE) ensemble, with the number of particles N, volume V and total energy E taking on constant values. Experiments, however, are usually performed on a canonical (NVT) ensemble, in which the temperature, T, is fixed instead of the energy. These ensembles are distinct from the perspective of statistical mechanics, so it is desirable to maintain a constant temperature within a microcanonical MD simulation.

The equilibrium temperature, T_0, of the system is related to the time-average kinetic energy $\langle E_k \rangle$ such that

$$\langle E_k \rangle = \frac{3}{2} N k_B T_0 \,; \tag{45}$$

while the total energy remains fixed, the instantaneous kinetic energy E_k fluctuates as some is transformed into potential energy and back. The total instantaneous kinetic energy of the system can thus be used to determine, at each time step, an effective temperature T. By considering the system to be weakly coupled to a heat bath with a constant temperature T_0, fluctuations of the kinetic energy can be suppressed by forcing the effective temperature to decay exponentially towards T_0 with a time constant τ, by applying the adjustment

$$\frac{\Delta T}{\Delta t} = \frac{T_0 - T}{\tau} \,. \tag{46}$$

Of course, the temperature is not a parameter of the simulation, but it is trivial to obtain from this equation the factor, λ, by which the velocities of the particles in the system should be rescaled. The velocity of each particle is adjusted such that

$$\boldsymbol{v} \to \lambda \boldsymbol{v} \,, \tag{47}$$

where

$$\lambda = \left[1 + \frac{\Delta t}{\tau} \left(\frac{T_0}{T} - 1 \right) \right]^{\frac{1}{2}} . \tag{48}$$

This is the Berendsen thermostat, also called the weak-coupling method [115]. The suppression of kinetic energy fluctuations means this method cannot produce

trajectories consistent with the canonical ensemble, but the results converge on the canonical ensemble for systems with a sufficiently large number of particles and collisions therebetween; it is thus suitable for most explicitly solvated simulations, and widely used due to its efficiency.

The pressure, P, of the system must also remain constant; this is of particular importance for explicitly solvated systems with periodic boundary conditions, the properties of which should agree with those obtained in the NPT (isothermal-isobaric) ensemble. The pressure can be controlled using the Berendsen barostat [115], which is similar to the Berendsen thermostat described above. Under this system, the lengths in the system are scaled by a scale factor μ, so, for example, the position coordinate of each particle is adjusted such that

$$\mathbf{r} \to \mu\mathbf{r}\,. \tag{49}$$

The scale factor is calculated for a cubic system as

$$\mu = \left[1 + \frac{\Delta t}{\tau_P}(P - P_0)\right]^{\frac{1}{3}}, \tag{50}$$

where τ_P is the "rise time" of the barostat, a time constant, and P_0 is the target pressure. P is, here, the instantaneous pressure, approximated for box volume V as

$$P = \frac{1}{V}\left(Nk_BT + \frac{1}{3}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbf{F}_{ij}\cdot\mathbf{r}_{ij}\right), \tag{51}$$

where $\mathbf{F}_{ij}$ is the force particles i and j (of N) exert on each other and $\mathbf{r}_{ij}$ is the vector joining them.

2.1.7.2 Langevin dynamics

The Berendsen thermostat is less appropriate for simulations in implicit solvent, which benefit from an alternative approach such as that provided by Langevin dynamics [116]. Transfer of kinetic energy occurs almost exclusively through collisions with the solvent, which is of course not present; these collisions are not modelled by the generalised Born model, which considers only electrostatic screening. There are thus two degrees of freedom not accounted for in the system's dynamics. The first of these is friction due to the movement of solvent molecules around the system, which exerts a force

$$\mathbf{F}_F = -\gamma\boldsymbol{v} \tag{52}$$

on a particle with velocity $\boldsymbol{v}$, where γ is the solvent viscosity, usually defined in terms of the frequency of solvent–solute collisions. Some of these collisions will be of particularly high energy, resulting in a perturbation to the system; this provides an additional random force,

$$\mathbf{F}_R(t) = \sqrt{2\gamma k_B T}\mathbf{R}(t)\,, \tag{53}$$

where $\mathbf{R}(t)$ is a stationary Gaussian process satisfying

$$\langle \mathbf{R}(t) \rangle = \mathbf{0} \tag{54a}$$
$$\langle \mathbf{R}(t) \cdot \mathbf{R}(t') \rangle = \delta(t - t'), \tag{54b}$$

where $\delta(t - t')$ is a Dirac delta function, defined as

$$\delta(x) = \begin{cases} \infty & \text{if } x = 0, \\ 0 & \text{otherwise;} \end{cases} \tag{55a}$$
$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1\,; \tag{55b}$$

that is, $\mathbf{R}(t)$ is an uncorrelated stochastic process occurring in individual "kicks" at random intervals, and results in no net acceleration. Thus, the total force experienced by particle i at time t is given by

$$\mathbf{F}_i(t) = -\nabla V_i(t) - \gamma \boldsymbol{\nu}_i(t) + \sqrt{2\gamma k_B T}\mathbf{R}_i(t)\,, \tag{56}$$

where V_i is the potential experienced by the particle due to its interactions with the other solute particles according to the MD force field and modified by the generalised Born model. This stochastic differential equation accounts for solvent viscosity and, due to its explicit dependence on temperature, can function as a thermostat. Generally, the viscosity, γ, should be kept small (though of course non-zero) because a viscous solvent retards the simulation and because overdamped Langevin dynamics ceases to be inertial and instead becomes Brownian, allowing no net acceleration to take place.

2.1.8 Performance optimisation and hardware

While the main loop in an MD simulation is iteration over time steps, and is thus necessarily serial, most of the calculations performed during each step are independent and thus trivial to parallelise. For example, a summation over atoms, $\sum_{i=1}^{N}$, is common in many of the calculations described above; since the order over which the atoms are iterated within a single step should be of no consequence, it is trivial to simply allocate a set of atoms to each core of a parallel processing unit and add together the results once all cores have finished their computations. In this manner, simulations can be accelerated considerably, especially on modern supercomputer clusters consisting of very many nodes [117].

An even more useful observation is that MD calculations are vectorisable, with the same operation applied to a large number of elements [118]. This operation—for example, calculating a term of the potential—can be expressed as a single operation on an array, or vector, containing multiple elements, and applied to all simultaneously on specialised vector-processing hardware such as a graphics

processing unit (GPU). For example, a modern GPU could perform the computation

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 \\ y_1 + y_2 \\ z_1 + z_2 \end{pmatrix} \tag{57}$$

in a single operation, rather than the three that would be required by a traditional processor. While this hardware was intended, as the name suggests, for the processing of computer graphics, general-purpose programming on GPUs (GPGPU) is possible using platforms such as CUDA,* an application programming interface (API) provided by GPU manufacturer Nvidia [119].

Using this principle, it is possible to perform an entire MD simulation exclusively on a GPU, resulting in substantially increased simulation speed and requiring only a single central processing unit (CPU) core for overall control of the process and input/output. Since even basic consumer CPUs nowadays comprise four or more cores, it is possible to run multiple CUDA-accelerated simulations with a small number of inexpensive CPU chips, provided there is sufficient GPU capacity. In the AMBER suite, there exists a CUDA version of the pmemd (particle-mesh Ewald MD) program, which runs many times faster than does the parallel CPU version on a reasonable number of cores [120, 121].

However, GPGPU does have some shortcomings. The highly parallel nature of GPU computations makes it more difficult to provide detailed debugging information, and GPU support for double-precision floating-point arithmetic (that is, the storage of non-integer numbers in 64 bit of computer memory, for increased precision compared to single-precision 32 bit numbers) remains imperfect; older cards lacked this capability altogether, while modern hardware typically has reduced capacity for double-precision arithmetic, resulting in degraded performance if too many double-precision values are operated on simultaneously. These issues can be accounted for by careful algorithm design [122], but can still result in numerical instability in extreme cases; for example, the CUDA version of pmemd may be unable to correctly minimise structures containing close interatomic contacts that would be correctly resolved in a traditional multi-CPU computation. For this reason, minimisation on CPUs is strongly preferred.

2.1.9 Simulation parameters

In this work, all MD simulations were performed using versions 14 to 18 of the AMBER software suite [71, 123, 124].

Both explicit and implicit solvent models were used in this work, depending primarily on the size of the system. Larger constructs were implicitly solvated

* Compute Unified Device Architecture

using the implicit generalized Born model [125] at a salt concentration of 0.2 M with GBneck2 corrections [109], mbondi3 Born radii set [108] and no cutoff for a better reproduction of molecular surfaces, salt bridges and solvation forces [111, 126]. Langevin dynamics was employed for temperature regulation at 300 K with a collision frequency of 0.01 ps^{-1}, which reduces the effective solvent viscosity and thus accelerates the exploration of conformational space. Where present, prolines were restrained to remain intercalated by using NMR restraints to restrict the distances between key atoms in the proline side chain and the neighboring bases. Shorter constructs were explicitly solvated using a truncated octahedral TIP3P box [99] and neutralised with a 0.2 M-equivalent concentration of K and Cl ions [127]. In both cases, the protein and DNA were represented using the ff14SB [88] and BSC1 [89] force fields, respectively.

Further details regarding special parameters used for certain simulations are provided in the following chapters.

2.2 Analysis of simulation trajectories

2.2.1 The WrLINE molecular contour

Accurate estimation of the writhe of DNA requires a well defined helical axis. If DNA were a uniform rod, this would be trivial, but the particular structure of the DNA double helix complicates matters. There is a handedness to the molecular structure, with the two grooves—the vertical spaces between the strands—having distinctly different widths; the wider major groove of B-DNA is 22 Å wide, compared to the minor groove's 12 Å. This results in a consistent bend towards the major groove and a periodic deviation in the local helix axis. Failing to account for this periodicity results in overestimation of writhe. A number of definitions of the molecular contour exist, but the WrLINE contour is shown to provide the most accurate results for minicircles with reasonable amounts of supercoiling [128].

The WrLINE contour forms the helix axis at a set of points corresponding to each dinucleotide step, with local irregularities smoothed out over the surrounding helical turn. Consider, for example, the neighbouring base pairs A–B and C–D, which form the i-th base-pair step in the DNA sequence; the midpoint of this base-pair step is

$$\mathbf{r}_i = \frac{1}{4}(\mathbf{r}_{C1'A} + \mathbf{r}_{C1'B} + \mathbf{r}_{C1'C} + \mathbf{r}_{C1'D}), \tag{58}$$

where $\mathbf{r}_{C1'}$ is the position of the 1′ carbon atom of the nucleotide, as labelled in figure 1e on page 16. The local helix axis, z_i, is the vector joining the midpoints of

base pairs A–B and C–D,

$$\begin{aligned} z_i &= r_{AB} - r_{CD} \\ &= \frac{1}{2}(r_{C1'A} + r_{C1'B}) - \frac{1}{2}(r_{C1'C} + r_{C1'D}) ; \end{aligned} \tag{59}$$

this is perpendicular to the plane Z. The twist of the helix over this base-pair step, θ_i, can be determined by projecting into Z the vectors y_{AB} and y_{CD}, which join the two C1′ atoms in each base pair, and calculating the angle between them as

$$\theta_i = \cos^{-1}(y_{ABZ} \cdot y_{CDZ}) , \tag{60}$$

where y_{ABZ} and y_{CDZ} are the projected vectors. The periodicity in the local helix axis is averaged out over a full turn consisting of $2m$ base pairs, with m chosen so that

$$\Theta_m = \theta_i + \sum_{k=1}^{m} (\theta_{i+k} + \theta_{i-k}) ; \tag{61a}$$

$$\Theta_m > 2\pi > \Theta_{m-1} . \tag{61b}$$

This, of course, results in a value of Θ_m slightly greater than 2π, systematically biasing the averaged helix axis. This can be corrected by reweighting the two base-pair steps at the ends of the turn by multiplying them by a factor w chosen so that

$$\Theta_{m-1} + (\theta_{i+m} + \theta_{i-m})w = 2\pi ; \tag{62}$$

the solution to this is

$$w = \frac{2\pi - \Theta_{m-1}}{\theta_{i+m} - \theta_{i-m}} = \frac{2\pi - \Theta_{m-1}}{\Theta_m - \Theta_{m-1}} . \tag{63}$$

Finally, the helix axis at step i is calculated by averaging the positions of all the midpoints within the helical turn, with these weights applied. The full set of these points for all dinucleotide steps defines the molecular contour.

The writhe of this contour can be calculated by numerically approximating the Gauß writhing integral (equation 7, page 22); the reference implementation of WrLine [128]* does this. The WrLine contour is also used by software such as SerraLine [129],† which projects the contour onto a plane and calculates structural properties such as bend angles and overall compaction in a manner broadly comparable to results obtained through experimental techniques such as atomic force microscopy (AFM).

2.2.1.1 Modifications to the WrLINE molecular contour

The reference implementation of WrLine [128] is written in Python 2, which is no longer supported,‡ and works only for circular DNA; the averaging of the

* https://github.com/agnesnoy/WrLINE

† https://github.com/agnesnoy/SerraLINE

‡ Peterson B 2008 "Python 2.7 Release Schedule" PEP 373 (Python Software Foundation) URL https://www.python.org/dev/peps/pep-0373

midpoints over a helical turn results in the obtained molecular contour for a significant region at either end of a piece of linear DNA being "smeared" over the vector joining the two ends. This has a significant effect on the overall contour and impacts measurements of properties such as writhe and bending angles.

For this project, the WrLine software was rewritten (with permission) in Python 3, with more informative names for functions and variables, clearer code formatting, a pleasant user interface, and support for linear DNA. The mathematics used for circular DNA is unchanged from the original implementation, and all functions retain the same functionality and input and output formats. The most significant change is the addition of a "`linear`" flag, which indicates that the input structure corresponds to a piece of linear DNA. If this flag is set, the midpoints close to the ends are averaged over less than a full turn, with the summation in equation 61 performed only until the end of the DNA is reached.

This does result in some deviation from the ideal helix axis close to the ends, but these end effects are small and affect only a small region of around one helical turn at each end, with the errors decreasing sharply away from the ends so that they are noticeable only for around 5 bp at each end. This can be expected to have a negligible effect on the writhe for any sequence of sufficient size to have non-zero writhe, and in any event represents a significant improvement over the original implementation of WrLine. Further, since the Gauß writhing integral is a path integral over a closed contour, so that one effectively views the system from all possible angles, the value calculated for a piece of linear DNA cannot be considered a true measure of writhe, and there is no reason to expect the calculated writhes for each linear subsection of a contour to sum to the global writhe; regardless, the method should result in a reasonable linear analogue of writhe suitable for comparison of the structures of fragments of equal length.

The modified code is listed in appendix 1.

2.2.2 Hierarchical agglomerative clustering

It can be helpful to categorise the frames of an MD simulation into structural categories such as open and closed states or distinct binding modes. This can be done by comparison with a set of known reference structures, but the selection of these structures can bias the results and known structures are not always available. An alternative approach is to construct clusters of similar results by simply comparing MD frames to one another.

This can be achieved using the method of hierarchical agglomerative clustering. In this method, each frame (or, for the sake of reducing computational expense, a random subset of the frames, known as a "sieve") is initially assigned to its own cluster of size 1. At each iteration, the clusters are compared to one another by

some distance metric, and the closest two are merged to form a single, larger cluster. This process can be repeated until some criterion—usually a chosen number of clusters or inter-cluster distance threshold—is met.

The distance metric used for cluster comparison can, in principle, be any quantity that is well-defined for an individual frame (and, in the case of average linkage, as discussed below, can be averaged over a cluster). For example, one could use the difference between two distances or angles, or the RMSD between the positions of corresponding atoms, which is defined as

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |\mathbf{r}_{i1} - \mathbf{r}_{i2}|^2}\,, \tag{64}$$

where N is the number of atoms of interest and $\mathbf{r}_{i1}$ and $\mathbf{r}_{i2}$ are the position vectors of the i-th atom in the two frames to be compared.

Irrespective of the chosen distance metric, there are multiple ways to define the distance between clusters containing more than one element; this is termed "linkage". Most commonly, the distance between clusters is defined as one of the mean, minimum, or maximum distance between elements (termed average, single, and complete linkage, respectively).

The standard algorithm for hierarchical agglomerative clustering has time complexity $\mathcal{O}(n^3)$ for n initial clusters (and requires $\Omega(n^2)$ memory);* while the special cases of single and complete linkage can be solved more efficiently [130, 131], the process remains computationally expensive. Sieves are used to reduce the computational expense by reducing the size of n. Furthermore, none of these methods is guaranteed to find the optimum solution, which can be guaranteed only by performing an exhaustive search (which has particularly poor time complexity of $\mathcal{O}(2^n)$); thankfully, while the results of the standard algorithms may not be completely optimal, with appropriate parameters this process should be entirely sufficient for all practical purposes.

Clustering of the frames of an MD trajectory is possible using the Amber cpptraj package [132, 133]. In this work, average linkage of the RMSD was used, usually applied only to the system's backbone atoms, with other parameters (such as the desired number of clusters) chosen individually for each simulation.

2.2.3 Hydrogen bond determination

Hydrogen bonds are an important type of interaction, forming the strongest non-covalent bond available to most biomolecules. They represent one of the main ways in which proteins and DNA interact with other molecules and themselves, and play

* See the footnote on page 38 for more discussion of "big O" notation; $\mathcal{O}$ represents an asymptotic bound from above while Ω represents an asymptotic bound from below.

an important role in stabilising intramolecular binding. It is thus important to be able to identify these bonds.

Hydrogen bonds are formed by a donor atom—usually oxygen or nitrogen—that is covalently bonded to a hydrogen atom, and a nearby electronegative acceptor atom. From an MD trajectory, hydrogen bonds can be identified by simply locating potential acceptor–donor pairs separated by a distance less than a chosen cutoff; a distance cutoff of 3.5 Å was used in this work. A further restriction on hydrogen bonding results from the arrangement of the necessary lone electron pair relative to the donor−H covalent bond, so the donor–hydrogen–acceptor angle is required to fall within some range of values close to its maximum value of 180°; any angle above 120° was considered acceptable in this work.

Of most interest for protein–DNA binding is the extent to which each amino acid interacts with DNA. This is captured by the time-average number of intermolecular hydrogen bonds, a value that takes into account both the strength and the persistence of interactions. For example, a residue that forms only a single hydrogen bond with the DNA but does so in every frame of the trajectory obviously forms one bond on average, but so does a residue that forms two hydrogen bonds in 50 % of frames; this may seem counterintuitive, but more hydrogen bonds mean a stronger interaction, and these residues are therefore of more significance and should be weighted accordingly.

Hydrogen bonds can be identified by cpptraj [132]; intramolecular bonds can be removed using a simple script (appendix 2.1), and the time-average number of bonds formed by each residue can be calculated by summing the “frac” column of the cpptraj output over all the donor or acceptor atoms belonging to a residue (appendix 2.2).

2.2.4 Identifying denatured regions of DNA

DNA denaturation is the disruption of the local structure of the double helix, identified by the breaking of hydrogen bonds between complementary bases and the separation of the helical strands. A region of denatured DNA is frequently referred to as a “bubble”.

Denatured regions can usually be identified quite easily by visual inspection of the atomistic trajectory, since the associated structural disruption is so large, but this is of course a qualitative and unreliable process that does not scale effectively for large numbers of frames. A programmatic mechanism for the identification of denatured base pairs on a frame-by-frame basis is therefore desirable. Perhaps the most obvious approach would be to identify those base pairs whose hydrogen bonds are disrupted in each frame, and this is indeed possible; however, the hydrogen bonding information is most easily obtained (using cpptraj) as a time-

average, which provides no information on the time-evolution of denatured bubbles, or as a very large matrix capable of representing all possible donor–acceptor pairs at each frame, an enormous quantity of data with which it is difficult to work effectively.

A more tractable approach is to consider the base-pair and base-step parameters for each base pair at each frame, as provided by the Curves+ software [134]. This provides a large number of time series representing base-pair and base-step properties at each point along the double helix. It would be both impractical and unnecessary to consider them all, but three angles were found to be sufficient to capture the presence of denatured regions with few false positives: twist, the relative rotation of adjacent base pairs about the helical axis; roll, the relative rotation of adjacent base pairs about the axis parallel to a single base pair; and propeller twist, the rotation of each base relative to its complement on the opposite strand. These rotational parameters were found to be both less noisy in canonical DNA and more significantly disrupted by denaturation than the available translational parameters.

The mean value and standard deviation of each of these parameters over all base pairs and frames are trivial to calculate. Then, each value x_{it} can be converted into a number of standard deviations away from the mean, N_{xit}, at each position i and time t as

$$N_{xit} = \left| \frac{x_{it} - \langle \bar{x} \rangle}{\sigma_x} \right|, \tag{65}$$

where $\langle \bar{x} \rangle$ is the mean value of the property x over all values of i and t and σ_x is its standard deviation. Finally, these values can be summed over the three properties considered to obtain an arbitrary measure of "how denatured" each base pair is at each time step, which will be referred to herein as the "index of denaturation". While the value obtained has little physical meaning in its own right, plotting its value on an $i \times t$ heatmap results in a large bright streak corresponding to the position in space and time of each denatured bubble, while no such features appear otherwise. This substantially simplifies and enhances the reliability of the process of identifying and locating denatured regions of DNA in an MD trajectory. A Python script to perform this analysis on a set of Curves+ output files is provided in appendix 2.3.

Furthermore, if the twist, roll, and propeller twist can be reasonably expected to be normally distributed about their means in the absence of disruption, one can expect to find 68 %, 95 %, and 99.7 % of values within one, two, or three standard deviations, respectively; if these are far from the observed proportions, it is very likely that significant structural disruption has occurred. It would be simple to estimate this from the index of denaturation: With three contributing parameters,

we expect fewer than 0.3 % of pairs (i, t) to have $N_{it} > 9$,* for example, and this presents an opportunity to automate the process further without requiring visual inspection of heatmaps.

2.3 Free energy calculation

The Helmholtz free energy, A, of a system is equal to the amount of reversible work performed on or available to be performed by the system at a constant temperature, T, and volume, and is given by [135]

$$A = U - TS\,, \tag{66}$$

where U is the internal energy of the system and S is its entropy. Free energy is an important thermodynamic property of a system, since a negative change in the Helmholtz free energy between two states is a necessary condition for a transition between those states to occur spontaneously.†

For a system of a constant number of particles N held at a fixed volume V, such as a typical molecular dynamics simulation, the free energy A_{NVT} can be formulated in terms of the partition function Z_{NVT}, so that

$$A_{NVT} = -k_B T \ln(Z_{NVT})\,, \tag{67}$$

where k_B is the Boltzmann constant.

For any non-trivial system, Z_{NVT} is, of course, a multiple integral over $3N$ degrees of freedom,

$$Z_{NVT} = \int \cdots \int_{3N} \exp\left(\frac{-U(x_1, \ldots, x_{3N})}{k_B T}\right) dx_1 \ldots dx_{3N}\,, \tag{68}$$

rendering its calculation a difficult undertaking, not least since extracting this information from a simulation would necessitate sampling the entire conformation space, the size of which scales exponentially with N.

Thankfully, many interesting properties of a system can be reduced to a single reaction coordinate, such as a distance between two atoms or an angle defined by three. It is possible in these cases to extract the partition function over the reaction coordinate x by multiplying the integral in equation 68 through by the Dirac delta function $\delta(x - x_0)$; that is,

$$\begin{aligned} Z_{NVT}(x) &= \int \cdots \int_{3N} \exp\left(\frac{-U(x_0, \ldots, x_{3N})}{k_B T}\right) \delta(x - x_0)\, dx_1 \ldots dx_{3N} \\ &= \int \exp\left(\frac{-U(x)}{k_B T}\right) dx\,. \end{aligned} \tag{69}$$

---

* In fact, we would expect even fewer than 0.3 %, since the three parameters are not perfectly correlated—if they were, there would be little need to measure all three.

† This is a direct corollary of the generalised Clausius inequality, $dS \geqslant \delta Q / T_{surr}$ (where δQ is an absorbed infinitesimal amount of heat and T_{surr} is the temperature of the surroundings).

The dependence of Z_{NVT} on the other degrees of freedom can in this way be integrated out.

We can thus obtain the Helmholtz free energy as a function of our reaction coordinate:

$$A_{NVT}(x) = -k_B T \ln\left(Z_{NVT}(x)\right). \tag{70}$$

The free-energy surface across a single coordinate is correctly referred to as the potential of mean force (PMF).

Calculating the partition function still requires knowledge of the internal energy of the system at every point along the reaction coordinate. However, a state's free-energy is, of course, related to its probability, P, such that

$$P(x) \propto \exp\left(\frac{-A(x)}{k_B T}\right); \tag{71}$$

we can thus determine the PMF up to some constant of proportionality by calculating the relative probability of each value of the reaction coordinate.

In principle, it is possible to do this by simply running a normal MD simulation and extracting from it the distribution of values of x, but the presence of barriers in the PMF means that sufficiently sampling the conformation space would require a prohibitively (in practice, infinitely) long simulation.

It is thus usually necessary to add a biasing umbrella potential $U'(x)$ to the system in order to help it over the potential barriers. The use of this method is known as umbrella sampling. From an MD simulation with this biased potential, we can trivially extract the biased probability $P'(x)$, and thence recover the unbiased PMF,

$$A(x) = -k_B T \ln\left(P'(x)\right) - U'(x) + F, \tag{72}$$

where F is an undetermined constant.

In the simplest case, $U'(x)$ could be chosen to flatten a known barrier, requiring only a single simulation and allowing F to be discarded. Frequently, however, multiple barriers exist, and their positions are not known in advance.* In these cases, it is necessary to perform a series of simulations with different umbrella potentials and combine the results of all these simulations in order to estimate the PMF.

Consider, for example, restraining the reaction coordinate according to Hooke's law, by adding an additional harmonic umbrella potential of the form

$$U'(x) = (x_0 - x)^2 k. \tag{73}$$

The observed values of x over a sufficiently long time period, assuming that the system undergoes Brownian motion in this harmonic potential alone, are expected

* The astute reader may notice that this method of calculating the free-energy surface requires prior knowledge of the free-energy surface.

to have a known distribution about x_0, but the observed distribution will deviate from this due to the underlying unbiased potential. This fact allows the change in the PMF over the sampled range to be extracted as in equation 72.

By performing a series of such simulations with different values of x_0, with k sufficiently strong that even the most unfavourable parts of the free-energy landscape are sampled but sufficiently weak that the expected distributions are broad enough to overlap, it is possible to sample the entire range of the reaction coordinate. The value of the offset F varies for each simulation, and each simulation should be weighted differently according to the favourability of the free energy landscape over its range. The optimal weights and offsets can be determined using the weighted histogram analysis method (WHAM) [136]; this entails dividing the values of x into bins, as if one were constructing a histogram, and iteratively solving the coupled equations

$$P(x) = \frac{\sum_{i=1}^{N} n_i(x)}{\sum_{i=1}^{N} N_i \exp\left([F_i - U_i'(x)]/k_B T\right)} \qquad \text{and} \tag{74a}$$

$$F_i = -k_B T \ln\left(\sum_{\text{bins}} P(x) \exp\left(\frac{-U_i'(x)}{k_B T}\right)\right), \tag{74b}$$

in which N is the number of simulations, N_i is the number of frames in the i-th simulation, $n_i(x)$ is the number of those frames in which the value of the reaction coordinate falls within the bin associated with x, and F_i is the value of the offset F for the i-th simulation.

The PMF can in this way be computed over the entire sampled range of the reaction coordinate. It is in fact possible to use the weights derived by WHAM to calculate the thermodynamic average of any property of a system.

There exist a number of software implementations of WHAM; in this work, the implementation by Alan Grossfield [137] was used.

2.4 Experimental techniques

The simulations described in the forthcoming chapters were developed in parallel with complementary AFM experiments by a collaborator. This close collaboration was facilitated by the parallel development of analogous theoretical and experimental methodologies. This included carefully selecting the systems to be studied—choosing DNA short enough to be simulated efficiently but sufficiently long as to be easily measurable experimentally, and with a sequence that can be easily produced for experimental use—as well as guiding the choice of methodology. For example, AFM images are two-dimensional and provide only limited volume information, so structural properties such as distances and angles

can be measured accurately only as projected onto a plane; thus, simulation data is best compared to AFM measurements when similarly projected.

This type of close collaboration, guided by the strengths and limitations of both methods, pervaded this work and allowed both techniques to enhance one another. An understanding of atomic force microscopy will therefore similarly enhance understanding of the results to be discussed.

2.4.1 Atomic force microscopy

Atomic force microscopy is an imaging technique in which a physical probe moves over a surface and generates a heightmap by measuring the vertical displacement of the probe at each point [138]. This allows very high resolution imaging, often far exceeding the optical diffraction limit [139, 140].

An AFM instrument consists of a very sharp tip mounted at the end of a flexible cantilever. As the tip passes over the surface (whether or not it makes physical contact with the sample), it experiences a variable force of an approximately Lennard-Jones nature (recall equation 11 from page 35). This results in a slight bending of the cantilever, which can be measured using a sophisticated detector; for example, a slight deviation in the cantilever position will change the amount of laser light it reflects into a photodiode. This allows the surface topography to be precisely determined.

Among the downsides of AFM is that high-resolution imaging generally requires that the system to be imaged be immobilised on a surface, making it difficult to observe dynamic events and causing some interference in the results due to surface chemistry. The resolution of the image is limited by the flexibility of the cantilever, the precision to which its bending can be determined, and the sharpness of the tip; the behaviour of the tip in air is further affected by the thin water meniscus that forms around the sample due to ambient humidity. While each of these issues can be individually addressed—with high-speed AFM allowing for imaging of dynamic systems, in-liquid AFM removing the effects of the water meniscus, and increasingly elaborate setups reducing the errors associated with each component—these all involve some trade-off and can result in worse resolution, longer imaging times, or higher costs [140].

Chapter 3

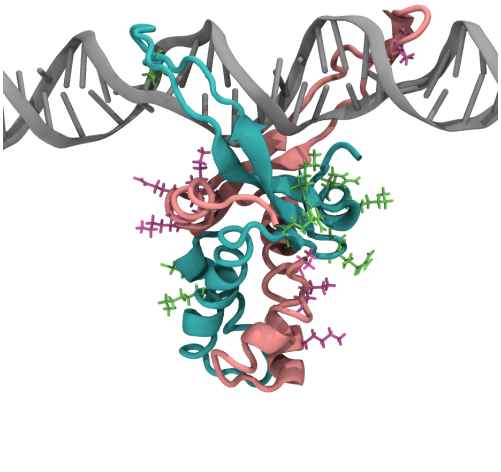
Multiplicity of topological states of IHF-bound DNA

3.1 Modelling DNA bending by IHF

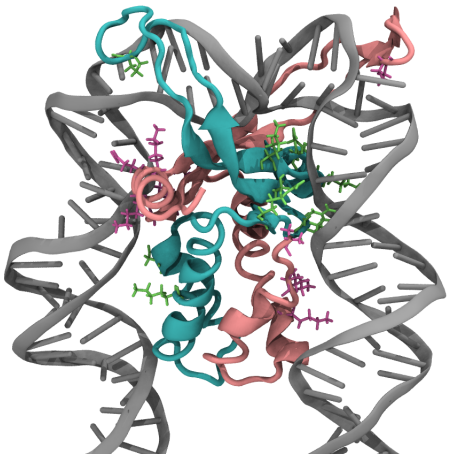
In order to describe the bending of DNA by IHF in MD simulations, an initial structure was created consisting of IHF bound to its H2 binding site in a length of unbent B-DNA, with the prolines intercalated as in the crystal structure described in PDB entry 5J0N [47]. The protein and an 11 bp segment of DNA were extracted from this crystal structure and embedded in a long 302 bp construct, which was implicitly solvated (figure 6a); and in a shorter 61 bp construct centred on the binding site, which was explicitly solvated. The simulation parameters are described in section 2.1.9.

No structural insight is yet available into the observed two-step binding mechanism; thus, in order to verify that this was an appropriate choice, the system was allowed to evolve. Recovery of the canonical structure would indicate that this choice was reasonable, as well as providing evidence in favour of the hypothesis that the two-step binding mechanism consists of intercalation followed by bending; in this model, the IHF arms would bind to DNA first, with the proline residues intercalating to induce flexible hinges prior to wrapping of the DNA around the protein body. Such a model would explain the high activation energy required for initial binding [51] and the smaller free energy of bending [52].

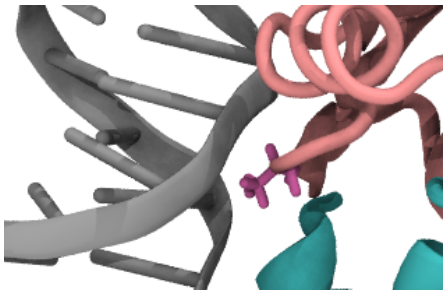
The canonical structure was indeed recovered in both implicit- and explicit-solvent simulations (figure 6b); the key amino acids Ser 47α and Arg 46β were observed to be inserted into the DNA minor groove, as previously reported (figures 6c–d), and all of the important hydrogen bonds present in the crystal structure described in PDB entry 1IHF [46] were reproduced (figure 6e). This demonstrates the validity of this simulation methodology, as well as suggesting that the above model of the two-step binding mechanism is plausible.



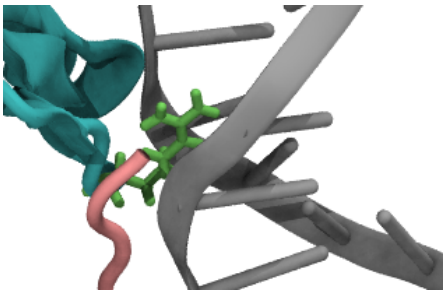
(a) Initial (intercalated) structure



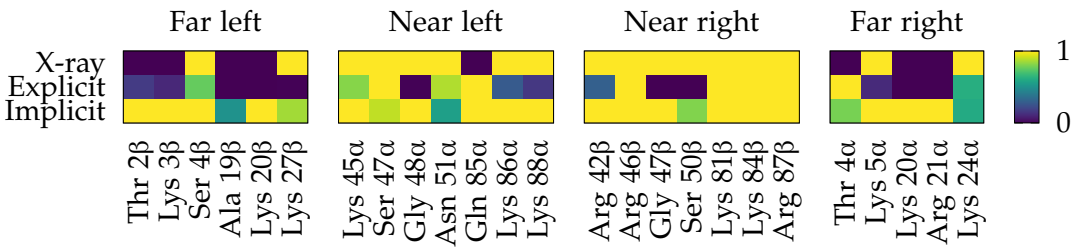
(b) Final (canonical) structure



(c) Ser 47α



(d) Arg 46β



(e) Contact map

Figure 6: Starting from an initial structure in which IHF is intercalated into straight DNA (a), canonical wrapping can be reproduced (b). Here, the α subunit is shown in pink and the β subunit in blue, with the intercalating prolines and amino acids that interact with the lateral DNA arms highlighted with an atomic representation; the DNA is shown in grey. In the obtained fully-wrapped structure, the key amino acids Ser 47α (c) and Arg 46β (d) are inserted into the DNA minor groove, as previously reported, and the interactions present in the crystal structure (labelled X-ray) are also present in both explicit and implicit solvent (e). This contact map shows the time-average number of intermolecular hydrogen bonds formed by a given amino acid, with the scale capped at 1; the X-ray structure can give only integer values, since it consists only of a single frame. Note that the DNA in the crystal structure is too short to capture some of the interactions in the far regions.

Hydrogen bond interactions were broadly divided into four regions based on the position of the involved DNA relative to the centre of the binding site and the involved protein subunit. On the left-hand side, which contains the A-tract, the α subunit is the closest to the centre of the binding site and thus interactions therewith are assigned to the "near left", while the β subunit interacts with DNA more distal to the binding site and thus its interactions are assigned to the "far left". This situation is reversed on the right-hand side, which contains the consensus sequence; here, the β subunit constitutes the "near right" while the α subunit constitutes the "far right".

3.2 Multimodality of the IHF–DNA complex

In addition to the canonical structure, additional states were observed over a series of MD simulations. This agrees with previous experimental observations of non-canonical binding states with bend angles less than 160° [48, 141], as well as with AFM observations by collaborators as part of this work (described in appendix 3).

The bend angles present in this AFM data are well fitted by three Gaussian distributions, suggesting the presence of three distinct populations (figure A1). This is also apparent from qualitative observation of MD trajectories and AFM images (figure A2). The MD trajectories were therefore divided into three populations using hierarchical agglomerative clustering. The bend angle was calculated for each frame of the MD simulations using a methodology designed to approximate that used for analysis of AFM images. To facilitate this, the WrLine molecular contour [128] of the DNA was projected onto the best-fit plane using SerraLine [129] and the bend angle was defined as the angle in this plane between two vectors each joining points 30 bp apart, with these vectors separated by a further 30 bp centred on the binding site.

Table 3: Mean values (± standard deviation) of the bend angle and radius of gyration, R_g, (both projected onto a plane) for each cluster of simulation frames (MD) compared to values measured from AFM images of two constructs with (1λ302) and without (0λ361) a specific IHF binding site.

		Associated	Half-wrapped	Fully wrapped
Bend angle / °	MD	66 ± 27	113 ± 30	157 ± 31
	1λ302	73 ± 7	107 ± 9	147 ± 30
	0λ361	70 ± 6	116 ± 5	–
R_g / nm	MD	24.5 ± 0.7	22.12 ± 0.02	20.78 ± 0.02

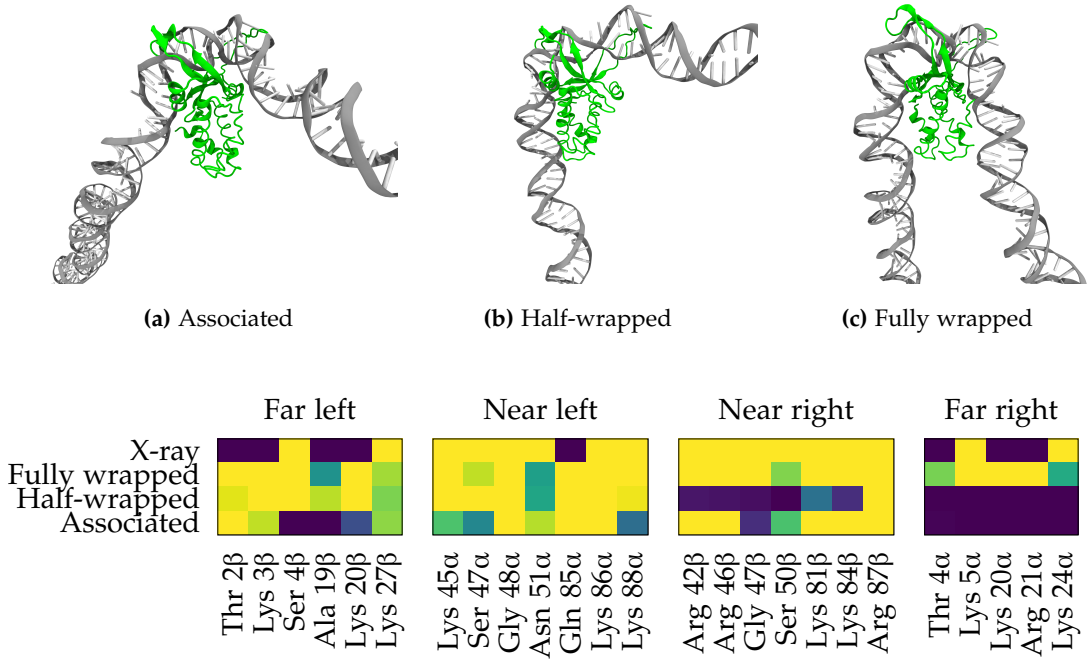
The resulting populations had mean bend angles of $66 \pm 27°$, $113 \pm 30°$, and $157 \pm 31°$ (mean $\pm$ standard deviation), with a larger bend angle associated with a reduction in the radius of gyration;* as table 3 demonstrates, these values are in good agreement with those obtained via AFM.

By visual inspection of representative structures (figure 7), these clusters were assigned names based on the extent to which the DNA interacts with IHF. In the state with the smallest (~66°) bending angle (figure 7a), the DNA and protein are in a relatively loose association, with the DNA bent only slightly and interacting only with the near subunit of the protein on each side; this cluster is thus termed the "associated" state. In the intermediate state (figure 7b), with a bending angle of ~113°, the DNA is fully bound to the protein on the left-hand side but does not interact with the protein at all on the right-hand side; this is the "half-wrapped" state. The largest bending angle (~157°) corresponds to the canonical binding mode (figure 7c), in which the DNA is bound to the maximum possible extent to both sides of the protein; this is thus the "fully wrapped" state. These states can further be characterised by mapping the hydrogen bonds present between the protein and the DNA (figure 7d); the fully wrapped state forms hydrogen bonds with both protein subunits on both sides of the binding site, while the half-wrapped state interacts similarly on the left but not at all on the right and the associated state forms contacts with only the parts of the protein on each side closest to the binding site.

It is notable that the right-hand side binds less frequently (or perhaps less strongly) than the left, since this is the side that contains the consensus sequence necessary for specific binding. Indeed, the consensus sequence does not interact with the protein at all in the associated or half-wrapped states, suggesting that these binding modes should occur without sequence specificity. This is confirmed by AFM experiments performed on a construct of similar length with no specific binding site (figure A1). In this case, the third peak—corresponding to the fully wrapped state—disappears from the angle distribution, while peaks remain present at angles that correspond to the associated and half-wrapped states, suggesting that these two states can occur nonspecifically. It is thus apparent that the presence of a consensus sequence is not necessary for the binding of IHF to DNA, but does increase the maximum possible degree of DNA bending by IHF.

These results strengthen the existing model of the two-step binding mechanism of IHF; the initial intercalation step occurs without sequence specificity (since there is no interaction between the protein and its consensus sequence at this stage, and this is also observed to occur when no consensus sequence is present), but full

* The radius of gyration, R_g, of an object is the root-mean-square distance of the object's components (in this case, atoms) from its axis of rotation or (more usefully for biological and polymer physics) its centre of mass, and provides a reliable measure of overall compaction.



(a) Associated **(b)** Half-wrapped **(c)** Fully wrapped

(d) Contact map

Figure 7: Representative structures of IHF binding modes. Simulations can be classified into three clusters with mean bending angles that correspond well with experimental data (table 3). In the associated state (a), with a bend angle of 66°, the DNA interacts only with the top subunit of the protein on each side of the binding site. In the half-wrapped state (b), the A-tract on the left-hand side fully binds to the protein while the consensus bases to the right are free and do not interact with the protein at all, resulting in a bend angle of 113°. The fully-wrapped state (c) is the previously observed canonical state, in which the DNA is bound to both subunits of the protein on both sides, resulting in a large bend of 157°. The DNA–protein interactions characteristic of each binding mode can be observed in a contact map (d) showing the time-average number of hydrogen bonds formed by each protein residue; the values here were calculated using all simulation frames belonging to each cluster.

bending occurs more slowly and only in the presence of a consensus sequence. The model is further refined by the observation of additional partially wrapped binding modes that can occur nonspecifically, which agree with previous results [48] but were not captured in previous investigations into the two-step binding mechanism.

Across four replica simulations, the system was observed to always pass first into either the half-wrapped (3 replicas) or the associated (1 replica) state. Of the replicas in the half-wrapped state, one remained in the half-wrapped state for at least 50 ns, until it was terminated; one transitioned to the fully wrapped state after ~10 ns; and one began to oscillate between the half-wrapped and fully wrapped states after ~25 ns, gradually becoming biased in favour of the fully wrapped state. The replica in the associated state transitioned to the fully wrapped state after ~11 ns. No transitions were observed between the associated and half-wrapped states. An explanation for this behaviour comes from consideration of the Debye screening length, λ, which represents how far the electrostatic effect of a charge carrier persists in a given solution (and is equal to $1/\kappa$, where κ is the Debye–Hückel screening parameter given in equation 33 on page 44). For the solution used here, with a salt concentration of 0.2 M, the Debye length $\lambda \approx 6.8$ Å, a very small fraction of the overall system. This precludes the DNA from interacting electrostatically with most of the protein, necessitating a "zipper" mechanism in which the near subunits interact with the DNA immediately next to the binding site, bending the DNA and allowing a new set of interactions; this process repeats until the DNA is fully bound to the protein, and explains the presence of transitional partially bound states.

3.3 Asymmetry of IHF binding

In order to better understand the asymmetry of the IHF–DNA complex, it may be helpful to consider its energetics. The construction of a free-energy landscape should provide some information about where the most energetically favourable states lie and the paths by which the system may pass between them.

To this end, a reaction coordinate was selected on each side of the complex and a set of umbrella-sampling simulations were performed to obtain potentials of mean force reflecting the positions of each DNA arm relative to the protein body. NMR restraints in Amber can only be applied to a distance defined by two atoms (or an angle defined by three or a torsion by four, but these are less useful in this case), so one must restrain a pair of individual atoms rather than, say, the centres of mass of two bodies. In this case, the reaction coordinates were chosen by selecting backbone atoms (since the positions of these should be less flexible than those of side-chain atoms) located in the far interaction regions that are close together in the minimised crystal structure. The reaction coordinates for the left and right sides, respectively,

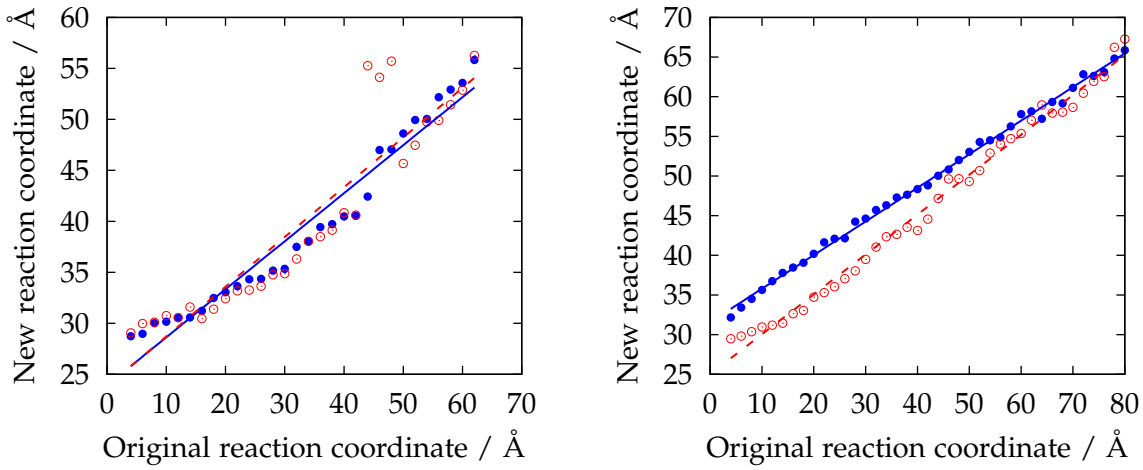
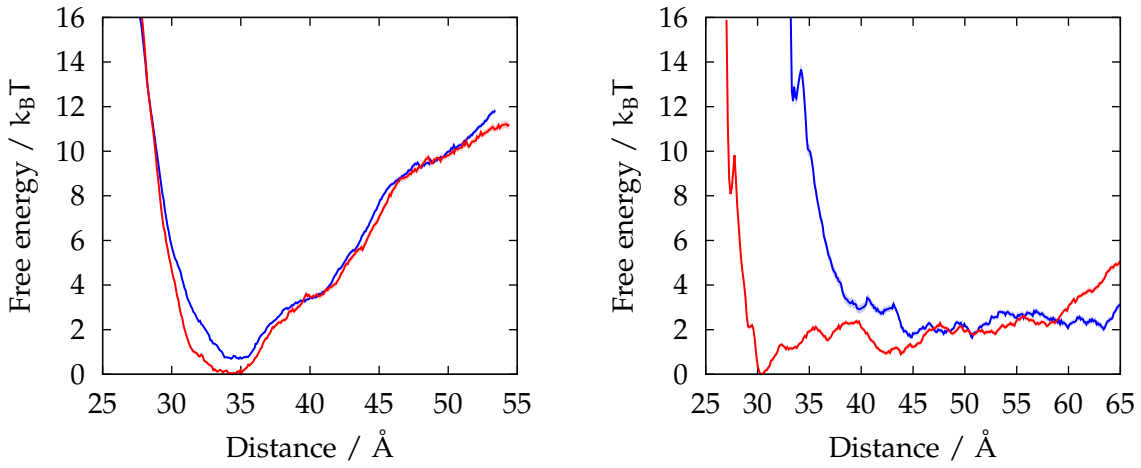


Figure 8: The original and new reaction coordinates were converted between by performing a series of linear fits with R^2 = 0.84–0.99, with the other arm free (hollow/dashed red) and restrained away from the protein (solid blue). The "original reaction coordinate" for the purpose of these fits is the position of the minimum of the umbrella potential for each window. This is sufficient for the purposes of this fit, but in some cases the actual value of the reaction coordinate deviates from this, as in the three outliers visible on the left; the transformation is performed on the free-energy landscapes calculated through WHAM, so these outliers are correctly transformed according to the actual value of their reaction coordinate.

were the distances from the Cα atoms of the amino acids Pro 18β and Ser 19α to the DNA backbone phosphorus atom that is closest in the crystal structure. Both reaction coordinates were reduced in 2 Å steps between 5 ns simulations, with a spring constant of 2 kcal mol^{-1} Å^{-2}, from their positions in the minimised initial structure until the PMF was observed to increase sharply, spanning ranges of 4–62 Å and 4–80 Å respectively, for a total simulation length of 150 ns for the left arm and 195 ns for the right. The final frame of each window was used as the initial structure for the next.

Two sets of umbrella-sampling simulations were performed for each arm. In the first, the reaction coordinate corresponding to the arm of interest was varied while the other arm was unrestrained and allowed to bind to the protein; in the second, the other arm was held away from the protein by a one-sided potential with spring constant 2 kcal mol^{-1} Å^{-2} if the value of its reaction coordinate fell below 40 Å, preventing it from fully binding to the protein body. Comparison of these two sets of results should demonstrate the independence, or otherwise, of the PMF from the position of the other arm.

In order to account for shifts in the free-energy landscape due to structural flexibility between simulations, a linear fit was performed in gnuplot to translate



(a) Left

(b) Right

Figure 9: Free-energy landscapes for binding of DNA arms. The left-hand side (a) presents a potential well with a depth of ~10k_BT, a minimum at 34 Å, and an additional plateau around 40 Å regardless of whether the right arm is free to bind (red) or held away from the protein (blue). The right-hand side (b) is associated with a much flatter potential, dominated by fluctuations on the order of k_BT, above ~45 Å. Below this point, the behaviour begins to depend on the position of the left arm; a small minimum is present around ~30 Å when the left arm is allowed to bind, but when the left arm is held away from the protein this region is blocked by a sharp increase in the free energy below ~40 Å. Shaded regions in these graphs represent the statistical error in the free energy, estimated using 50 iterations of Monte Carlo bootstrap analysis, but the errors are very small.

the reaction coordinates to the root-mean-square distance between the centres of mass of the protein and a 10 bp region of DNA on the appropriate side (figure 8). This measure should be relatively inflexible and consistent between simulations, resulting in well aligned free energies.

The obtained free-energy landscapes for the two DNA arms are shown in figure 9. The left arm (figure 9a) is observed to present a deep well potential with a depth of ~10k_BT, a minimum around ~34 Å (which corresponds to full binding) and an additional plateau around ~40 Å (which corresponds to partial binding as observed in the associated state) regardless of the restraints on the right arm, suggesting a strong preference for full binding and no dependence on the position of the other arm. The right arm (figure 9b), however, presents a flat potential dominated by fluctuations on the order of k_BT at distances greater than ~45 Å; below this distance, the behaviour has a strong dependence on the position of the left arm. When the left arm is allowed to bind, the potential continues to be mostly flat and presents small minima around ~43 Å and ~30 Å; when the left arm is prevented from binding, shorter distances are inaccessible due to a sharp increase in the right-arm PMF below ~40 Å. It is thus necessary that the left arm is permitted

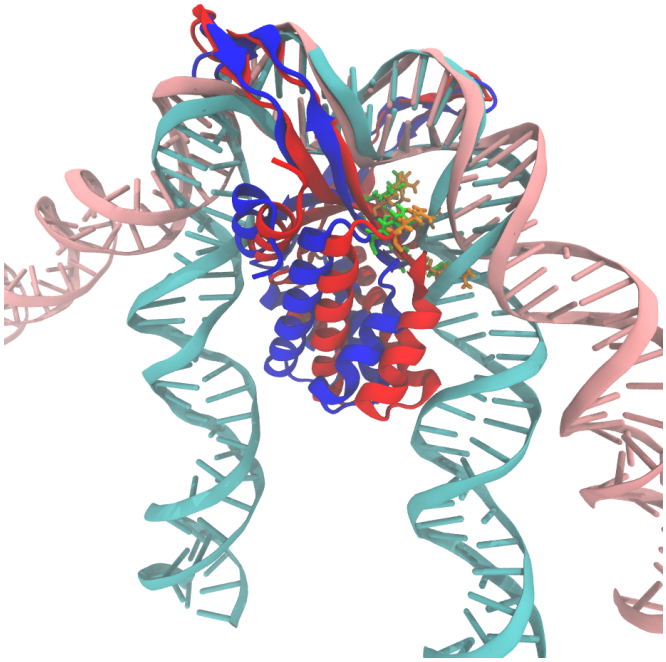
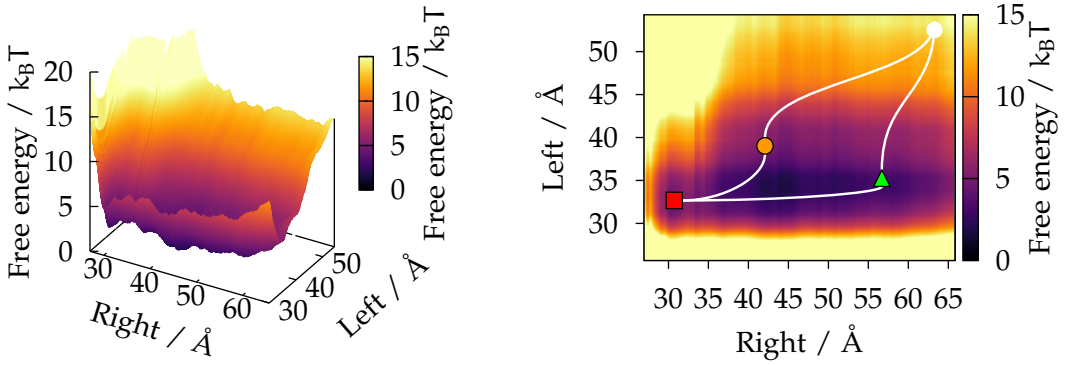


Figure 10: When the left arm is unbound (protein red and orange, DNA pink), the upper subunit of the protein protrudes on the right-hand side and blocks the right DNA arm from binding to the lower subunit without bending significantly. When the left arm binds (protein blue and green, DNA cyan), the protein body is pulled to the left, flattening the right-hand side and allowing the right arm to bind without bending.

to bind in order for the right arm to do so.

This explains the absence of any observed states in which the right arm is bound but the left arm is not. The physical mechanism for this behaviour can be elucidated by comparison of a representative structure of the IHF–DNA complex in which the left arm is bound to one in which it is not (figure 10). When the left arm is unbound, the upper subunit of the protein protrudes on the right-hand side, so the right arm cannot bind fully without bending significantly over an inflexible region; this bending results in a large free-energy barrier. When the left arm binds fully, the protein body is pulled leftwards, presenting a flatter surface to the right arm and allowing it to bind fully without bending. This phenomenon is not observed in reverse; there is no barrier preventing the left arm from binding, regardless of the position of the right arm.

A two-dimensional free-energy landscape can be constructed by considering the two reaction coordinates to be orthogonal axes, allowing for the estimation of the free energy of any possible binding mode. Unfortunately, the dependence of the right arm's PMF on the position of the left arm demonstrates that the reaction coordinates are not truly orthogonal. It is, however, still desirable to develop such a



(a) 3D **(b)** 2D

Figure 11: The overall shape of the free-energy landscape becomes apparent when it is plotted in three dimensions (a), but analysis of its main features may be simpler on a two-dimensional heatmap (b). The minima and plateaux in this landscape correspond to the observed binding modes: the fully wrapped (red square), half-wrapped (green triangle), and associated (yellow circle) states, with the paths through them shown as white lines from the initial structure (white circle, top right).

two-dimensional landscape, and it would be prohibitive to simulate every possible combination of arm positions.* In order to account for this, the two obtained PMFs for each arm can be taken to represent the extremes of the other arm's position and interpolated between; while this will result in some loss of accuracy far from the extremes, the differences between the two PMFs are small across most of the region of interest, and the main features of the landscape—the positions of minima, for example—should be reproduced despite inaccuracies in the exact values estimated for the free energy. Such an estimate is still valuable and represents the best that can be reasonably achieved with existing methods.

The result of this is shown in figure 11. When plotted as a three-dimensional landscape (figure 11a), the difference between the left and right arms becomes clear, a steep slope along the axis corresponding to the left arm contrasting sharply with the flatter, noise-dominated right arm. The main features of the landscape are perhaps best viewed, however, on a two-dimensional heatmap (figure 11b). This shows the restricted region corresponding to binding of the right arm without the left, as well as the multiple local minima and plateaux.

In order to both validate and further characterise this landscape, it is possible to map onto it the three populations previously observed—each of which represents

* To the 2 Å resolution used here, there are 29 possible left-arm and 39 possible right-arm positions, so sampling the entire conformation space would require $29 \times 39 = 1131$ simulations, far too many to reasonably perform.

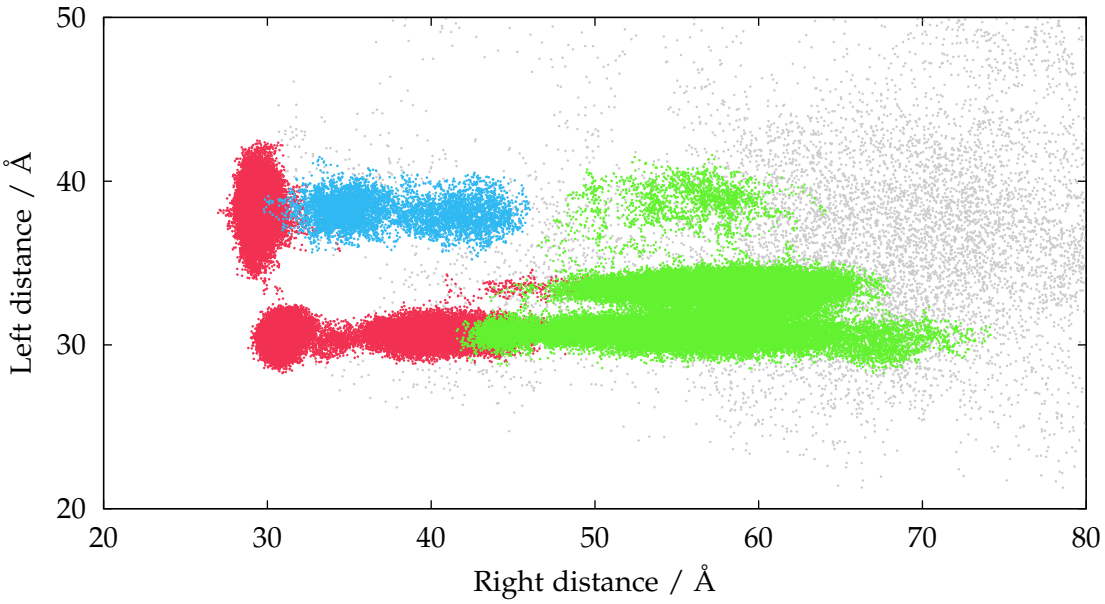


Figure 12: Positions of binding modes on the conformational landscape. The fully wrapped binding mode (red) occupies the lower-left corner of this landscape, extending across both the minima at right-arm positions of 30 Å and 40 Å; both the right arm and the left arm have some flexibility, but only if the other remains fully bound. The half-wrapped state (green) occupies the remainder of the left arm's minimum, while the associated state (blue) is confined to a relatively small region. Grey dots represent transition states observed in the first 5 ns of each simulation.

a distinct binding mode—by extracting the values of the reaction coordinates in each frame belonging to a cluster (figure 12). The clusters line up with minima or plateaux in the free-energy landscape, allowing the nature of these features to be identified and (recalling that the clustering and the free-energy calculations were performed on different systems, solvated and simulated differently) providing some validation for both sets of results. The two minima corresponding to states in which the left arm is fully bound and the right arm is either partially or fully bound are revealed to be part of a single broad minimum, being separated only by a small free-energy barrier of less than $2k_BT$ that can be reversibly traversed under the influence of thermal noise; the frames present in this region are all clustered together into the fully-wrapped state and have indistinguishable bend angles.

There is an overall free-energy reduction of ~$15k_BT$ between the initial intercalated state, in which the DNA is completely straight, and the canonical fully wrapped bending mode. The intermediate states are less favourable overall than the canonical state, but correspond to plateaux in the free-energy landscape and are therefore metastable. The steepness of the free-energy landscape of the left-hand side makes it very likely that this side will bind first, resulting in the half-

wrapped state; this is more favourable than the initial state by ~$13k_BT$ and exists on a broad plateau. Although the fully-wrapped state is, from here, more favourable (at least for this sequence), this results in a further free-energy reduction of only ~$2k_BT$ and requires the system to first reach the edge of its current plateau, so one would expect this state to be metastable. The associated state, meanwhile, exists in a more precarious balance; while it corresponds to a local minimum in the right arm's free-energy landscape, restricting its movement along this axis, it occupies only a narrow plateau in the overall landscape. Transition from the initial conditions to this state results in a free-energy reduction of ~$9k_BT$ and the path the system must traverse to reach it is less steep; from here, transition to the fully wrapped state would result in a further free-energy reduction of ~$6k_BT$. While also metastable, it would seem unusual for this state to remain occupied for an extended period.

By integrating over the regions of the free-energy landscape that correspond to each bending mode, it is possible to estimate the probability with which each state will occur. In this case, the free-energy landscape was divided into a grid of squares with side lengths of 1 Å, with the free energy taken to be homogeneous within each square, and the contribution of each square to a cluster's probability was weighted according to the proportion of the frames falling within that square that belonged to the cluster. The results were found to be the same to within 1 percentage point whether or not the weightings were normalised by the number of points in each cluster. Recalling equation 71, the probability $P(x, y)$ of the reaction coordinates taking on the values x and y is related to the free energy such that

$$P(x, y) \propto \exp\left(\frac{-A(x, y)}{k_B T}\right). \tag{75}$$

When considering the relative probabilities of a set of states, the constant of proportionality can be discarded by *post hoc* normalisation of the probabilities to sum to 100 %.

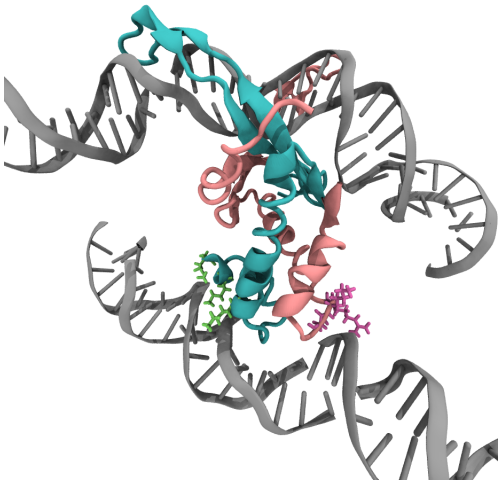
This method predicts that the associated state should occur with a relative probability of 18 %, the half-wrapped state 50 %, and the canonical fully wrapped state 32 %. Experimentally observed proportions can be determined by taking the area under the Gaussian curves fitted to the histogram of bend-angle data measured using AFM; the associated state is in fact observed in 32 % of samples, the half-wrapped state 27 %, and the fully wrapped state 41 %. While these two sets of figures do not agree perfectly, the experimental observations confirm the prediction that all three states should occur in significant proportions. The differences between the predicted and observed proportions may be an artefact of the different solvent and surface conditions, or the use of quite different techniques to estimate them (integrating over the free energy landscape, compared to integrating under a set of fitted probability density curves), but the largest effect is likely to come from inaccuracies in the free-energy landscape far from the extremes due to the

linear interpolation that was performed. The size of this effect can be estimated by considering the proportions predicted using each possible pair of the four simulations to obtain a set of different (less accurate) estimates of the state probabilities. This technique provides error bars of ±13 percentage points (pp) for the associated state, ±4 pp for the half-wrapped state, and ±10 pp for the fully wrapped state. Another source of error comes from the choice of force field, and it would be interesting to prepare the system using a range of force fields, such as CHARMM and GROMOS, and use the deviations in static energies between these and the AMBER force field used for the simulations to estimate the same for the free energy; this would, however, require substantial work and the differences should be small compared to the error introduced by the interpolation and accounted for above. A corresponding estimate of the confidence intervals for the experimental proportions could be obtained from the output of the algorithm used to fit the Gaussian distributions to the data, but was not provided by the collaborator in this work.

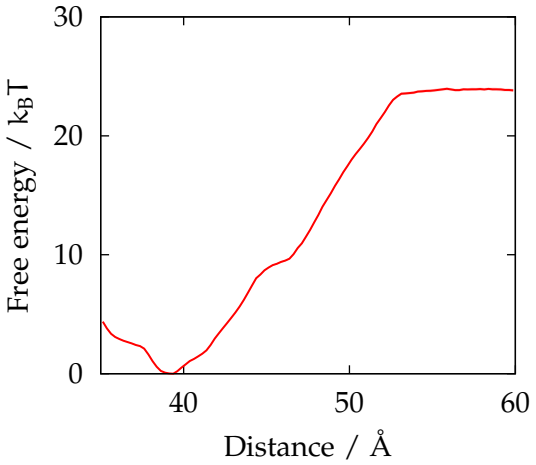
This free-energy landscape is likely to vary significantly with the DNA sequence. The fact that the fully wrapped state is not observed in AFM images of DNA lacking the IHF consensus sequence suggests that this state would become unfavourable if this sequence were to be mutated away. The other states are unlikely to be affected by such a change, because they feature no interactions between IHF and its consensus sequence, although a complex indirect readout mechanism is always possible. Mutating away the A-tract may have a wider-reaching effect—positioned on the left arm, this interacts with the protein to some extent in every binding mode and is known to be important to IHF binding, but it is difficult to say much about this based on these data. The effects of DNA sequence are likely to provide a fascinating third dimension to the free-energy landscape of IHF binding and bending.

3.4 DNA bridging by IHF

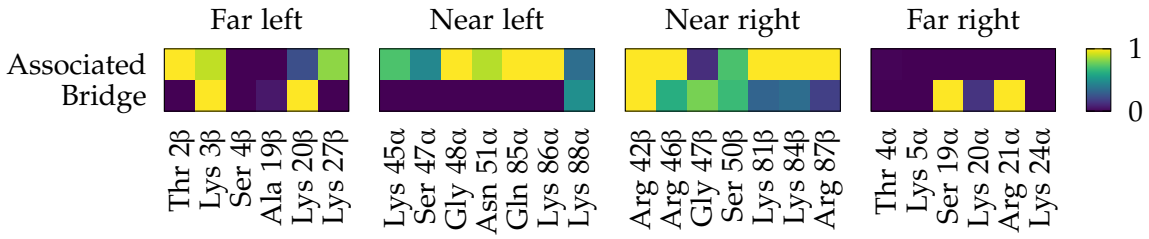
AFM imaging at high IHF concentrations revealed the formation of large DNA–protein aggregates (figure A3); this agrees with previous observations that indicate the presence of IHF at DNA crossing points [44], and suggests that a single IHF molecule is capable of being bound to more than one DNA strand simultaneously. In simulations of IHF bound to supercoiled minicircles, of the type that will be discussed in chapter 4, IHF was observed to form spontaneous contacts with DNA distal to its binding site. The observed bridges result from nonspecific interactions between positively charged or polar amino acids and the negatively charged DNA backbone; these contacts were seen to be remarkably delocalised, involving various parts of the protein including the arms, and remarkably stable, with no such



(a) Structure of DNA–IHF–DNA bridge



(b) Free-energy landscape



(c) Contact map

Figure 13: A DNA–IHF–DNA bridge is formed by interactions between DNA and IHF far from the canonical binding site (a). Weighted histogram analysis of umbrella-sampling simulations reveals that this structure is very energetically favourable compared to a structure where the two DNA strands are far apart, with a free energy reduction of ~24k_BT on bridging (b). The bridge is formed by residues from the "far" regions, with the strand bound to the canonical site mostly unbent and interacting even less than in the associated state (c).

bridge observed to spontaneously break. When a bridge forms, the main strand (containing the classic binding site, into which the prolines are intercalated) remains mostly unwrapped—one could hypothesise that this is due to electrostatic repulsion between the two DNA strands.

In order to investigate this phenomenon, which has significance in the study of biofilms, two 61 bp sections of DNA were extracted from a bridged minicircle structure, one segment centred on the traditional IHF binding site and the other on the region forming additional contacts with the protein (figure 13a). This system was then explicitly solvated, and umbrella-sampling simulations were performed to gradually pull the two pieces of DNA apart. The reaction coordinate was selected by choosing the backbone atoms closest to the centres of mass of the protein and the

second DNA molecule (A55 OP2–Phe81α Cα), and was increased in 1 Å increments over a series of 5 ns windows from its initial value until the PMF was seen to plateau.

The resulting PMF reveals that DNA bridging by IHF is very energetically favourable, with a free-energy change of ~24k_BT between the bridged and unbridged states (figure 13b). This is much larger than the free energy of wrapping, suggesting that, given a choice between wrapping one DNA strand and bridging two, bridging is more probable. It appears this is a choice IHF is doomed to face frequently; in simulations of the bridged system, the DNA strand bound to the canonical site is not observed to fully wrap, generally interacting even less than in the associated state (figure 13c). It is probable that this is due to electrostatic repulsion between the DNA strands, which would make full wrapping highly unfavourable. The effects of bridge formation on the free energy of wrapping were not quantified in this work, but would be a valuable topic for further study.

3.5 A complete model of IHF binding, bending, and bridging

The results presented in this chapter can be combined to produce a full model of IHF binding, bending, and bridging, as presented in figure 14. This model proposes that the two-step binding mechanism begins with intercalation into mostly unbent DNA. From here, bridge formation is favourable if possible but requires a nearby piece of DNA. Otherwise, the system passes to either the half-wrapped or the associated state, with a slight preference for the half-wrapped state. Both of these states appear to be either stable or metastable. If the consensus sequence is present, full wrapping like that previously observed through X-ray crystallography is possible from either of these states and further reduces the system's free energy. While the half-wrapped and associated states are here shown to be nonspecific, they may not be the only conformations in which IHF can bind nonspecifically to DNA. The similar electrostatic profiles of IHF and HU suggest that a probable candidate for an additional nonspecific mode is one similar to that observed for HU by Hammel *et aliis* [57], in which the protein aligns parallel to the DNA without intercalating and induces only a minor bend; such states would require a distinct simulation methodology outside the scope of this work, but represent an interesting target for further investigation. There is little to suggest, however, that additional specific binding modes exist beyond those described in this work.

This work suggests that the reduction in free energy associated with wrapping of DNA around IHF is around 15k_BT, much larger than the ~6k_BT previously measured [52]. This may reflect differences between IHF binding sites, with this work conducted on the H2 site while most previous work has focused on the H′ site; it is known that the strength of the H2 site is greater than or equal to that of

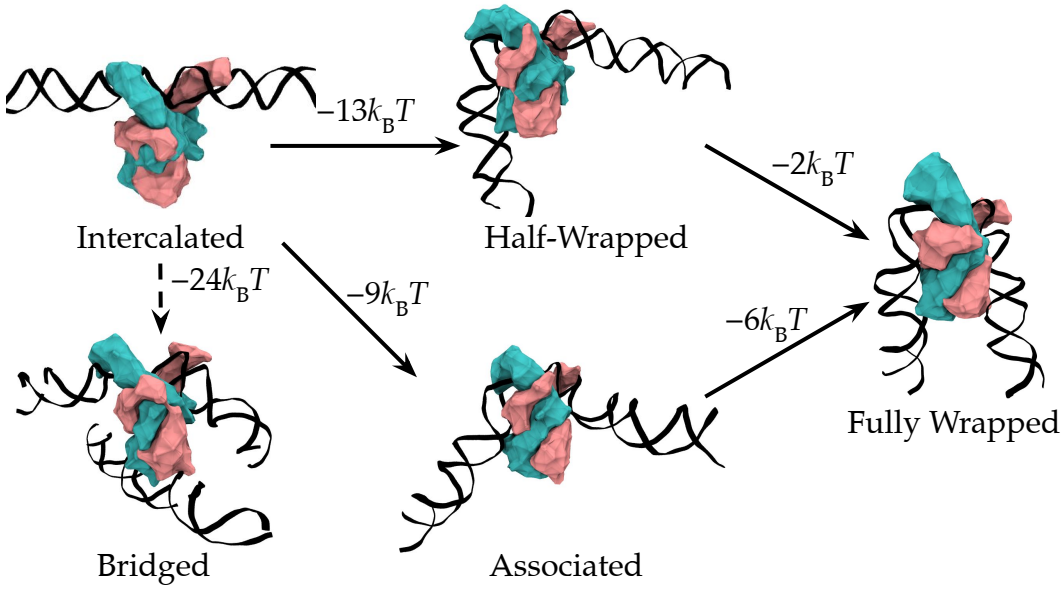


Figure 14: A complete model of IHF binding, bending, and bridging. IHF first binds to DNA, the prolines intercalating to form flexible hinges. If bridge formation is possible, it is preferred; otherwise, the protein will begin to bend DNA. Either the half-wrapped or the associated state may occur next, although the half-wrapped state is slightly favoured. In either case, once the A-tract has fully bound (and only once the A-tract has fully bound), the consensus sequence may bind if one exists. In this way, the canonical bending is reproduced. Arrows represent possible transitions between states, annotated with an approximation of the corresponding change in free energy; a transition from the associated to the bridged state requires that a second piece of DNA be present and appropriately located relative to the protein.

the H′ site. While a difference of this magnitude is not impossible, it would be surprising. Some of the difference may result from the experimental techniques used to measure the free energy of wrapping; it seems quite possible, for example, that the cited work in fact measured the free-energy change associated with a transition to the half-wrapped or associated state, rther than complete wrapping. Furthermore, since DNA *ex silico* never occupies the almost perfectly straight structure used in this work to define the initial (intercalated but unbent) state, these simulations will somewhat overestimate the reduction in the free energy.

A fascinating possibility opened up by these results is that the IHF–DNA complex may act as a mechanical switch, capable of occupying multiple distinct states with populations moderated by the position of the DNA to its left; in this way, IHF's bending or regulatory behaviour could be modulated—or even switched on or off—by tension or structural influences upstream of its binding site.

3.6 Multiple binding sites

Simulations were also performed of a 343 bp construct containing three IHF binding sites, which was also imaged using AFM. Unusual behaviour in the presence of another IHF bound nearby would indicate cooperativity between IHF molecules.

No such cooperativity was observed. The behaviour of each binding site was consistent with the description provided in the previous section, and the local bending induced by IHF at one binding site resulted in no effects at the other sites. This does not rule out cooperative behaviour between binding sites separated by a shorter distance, but does suggest that the influence of IHF on the structure of linear DNA is very localised, and that distal structural changes are not among IHF's regulatory mechanisms.

AFM images demonstrated the formation of large IHF–DNA aggregates in the presence of multiple binding sites, which did not form for constructs containing only a single binding site. This aggregation can be explained by simple stoicheiometry: a single bound IHF can form a bridge with only one other DNA molecule, resulting in a closed system consisting of a simple bridge between the two pieces of DNA; in a system with more than one binding site, each IHF can, if oriented differently, bind a different piece of DNA, allowing the aggregate to grow exponentially. Unfortunately, these structures are very large, and thus are not well suited to atomistic simulation.

Chapter 4

Interactions between IHF and supercoiled DNA

4.1 Introduction

As discussed in chapter 1, there is significant interest in DNA topology as a regulator of gene expression and protein binding. DNA supercoiling is known to have a significant effect on a number of important regulatory processes, and to itself be the product of a large number of complex interactions. It is thus of interest to explore the interplay between DNA topology and IHF binding.

The relationship between the two structural influences is likely to be complex and bidirectional: By altering the structure and flexibility of DNA, supercoiling is likely to have some effect on the binding modes exhibited by IHF; meanwhile, the bending of DNA by IHF can be expected to affect the structure of, and distribution of topology within, the DNA sequence.

The exact nature of these effects is, of course, unknown—otherwise there would be little need to do this work. One can, however, begin with some reasonable hypotheses. For example, supercoiled DNA is known to form plectonemes, writhed structures with tightly bent regions of DNA at their apices. The formation of these structures is dependent on the local flexibility of the DNA in each region, a property which is significantly enhanced by IHF, so one might expect to observe the formation of plectonemes with IHF at their apices. This would manifest as a consistent alignment of plectoneme features between replicas, at least in the presence of a single bound IHF. Meanwhile, the inherent bending induced by plectonemes, especially if it is concentrated around the IHF binding site, may shift the previously observed free energy landscape in favour of more tightly bound states, and changes in twist may affect DNA recognition and binding by IHF, further complicating the picture. Finally, intercalation of a small ligand into DNA is associated with a significant reduction of local twist, so the two intercalating

proline residues on IHF's arms should be expected to have some influence on the local topology, an effect which—due to the global conservation of Lk—should be compensated for elsewhere, perhaps resulting in a nonuniform distribution of twist and writhe along the DNA contour.

While unlikely if IHF is indeed located at the apex, it is also interesting to note that plectonemes bring distal DNA sites into close proximity, and have previously been observed to allow interactions between regions of DNA and distally bound ligands [24, 25]; should IHF find itself located close to a DNA crossing point upon the formation of a plectoneme, this may present interesting opportunities for the formation of a bridge such as that described in the previous chapter.

4.2 Modelling DNA supercoiling

The study of DNA topology requires the simulation of pieces of DNA with $Lk \neq Lk_0$. In order to maintain a constant linking number in an MD simulation, the ends must not be allowed to rotate relative to one another, as this would allow twist to pass along the double helix to the end of the strand and thereby be relieved. Maintaining this restraint in a piece of linear DNA without also placing the strand under tension is impractical and has no experimental analogue. While it might be valuable to study DNA supercoiling under tension, for example for comparison with experiments using magneto-optical tweezers or to extract certain physical properties of the polymer, the additional restraints limit the direct biological relevance of such results and make comparisons to other single-molecule imaging techniques difficult.

For these reasons, it is desirable to simulate DNA minicircles. To these can be applied an arbitrary amount of supercoiling,* which will remain fixed throughout the simulation. Covalently closed circular DNA has direct biological relevance, with examples of such structures including plasmids and the bacterial chromosome, and there is growing interest in minicircle DNA as a hallmark of cancer in humans [142, 143]. Minicircles have previously been used for studies using a range of experimental and simulation techniques [33] and can be obtained for experimental use through a complex and interesting technique involving λ recombination [144], a convenient side effect of which is the presence of IHF binding sites, and minicircles of sufficient size can be imaged through techniques such as AFM.

A tool like NAB can be used to generate a piece of linear DNA, which can then be deformed to lie along the circumference of a perfect circle. A hydrogen atom

* Of course, reasonable physical limits apply; the structure will be torn apart by extreme forces if one attempts to apply amounts of supercoiling far beyond the normal physical range, but there is little reason to do so.

can then be removed from each end of each strand and covalent bonds formed to complete a pair of intertwined circles before the topology file governing the simulation is created. The atom labels used by the ff14SB force field must be changed slightly in order to allow these covalent bonds. This is by now a standard procedure that has been used in a number of publications [18, 129]. Proteins and other ligands can then be attached as to linear DNA, although additional care may be required to avoid clashes between atom positions, unusual bond parameters, and in some cases changes to the linking number when a segment of straight DNA is inserted in place of an arc of equal length.* Nevertheless, careful construction and minimisation make the production of such systems possible. DNA *in vivo* is not, of course, a perfect circle, but this initial structure is likely to be sufficient as a first approximation and can be expected to evolve towards the correct structure over the course of minimisation and equilibration.

Such a process was followed to produce DNA minicircles with a length of 336 bp and the sequence given in appendix 4.4, which contains a single IHF binding site. A range of topologies were considered, with linking numbers between 29 and 34. Assuming B-DNA with 10.5 bp per turn, one would expect that $Lk_0 = 32$. MD simulations have traditionally systematically underestimated twist [145], but this is no longer true since the BSC1 force field solved this problem; in fact, the use of implicit solvent has been shown to result in an overestimation of twist [146]. To account for these effects, the actual value of Lk_0 for a system under a chosen force field can be determined by performing simulations of bare minicircles with a variety of topologies, calculating the mean writhe of each, and performing a linear fit to determine the point at which $\langle Wr \rangle = 0$. For this system, $Lk_0 = 31.08$. The most negatively supercoiled minicircle considered here therefore has superhelical density $\sigma = -0.067$, very close to the $\sigma \approx -0.06$ at which the prokaryotic genome is typically maintained [147].

Three replica simulations were performed for each minicircle, each with a length of 30 ns, with the last 10 ns used for most analyses in order to allow the systems time to equilibrate fully. Due to the size of the systems, implicit solvent was necessary; the generalised Born method was used with the parameters described in section 2.1.9 on page 51. This unfortunately makes a study of the energetics, as in chapter 3, impractical, since a reliable estimate of the free energy requires a realistic solvent model accounting for its entropy and viscosity, and more effectively modelling electrostatic screening.

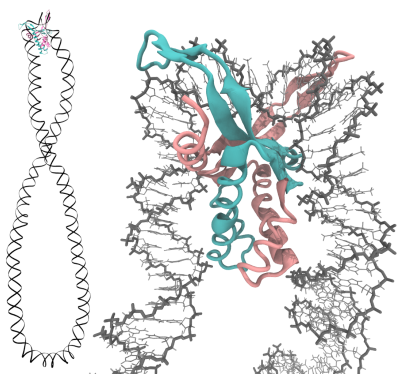
* This follows from basic geometry, in that the length of an arc differs from that of its corresponding chord, yet in this application it is generally desirable to replace an arc with a chord of equal length.

4.3 Effect of supercoiling on local IHF–DNA interactions

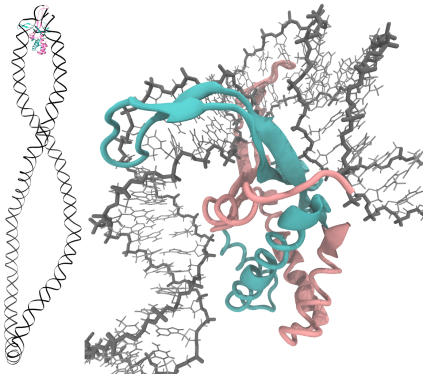
In order to study the binding modes of IHF to supercoiled DNA circles, hierarchical agglomerative clustering was once again performed to classify the frames present in the last 10 ns of each simulation, using the RMSD of backbone atoms belonging to the protein and the 61 bp region of DNA surrounding the binding site as the distance metric.

However, the inherent topological differences between minicircles with different linking numbers present difficulties for automated clustering algorithms, since the significant local structural differences introduced by changes in twist result in an inherently larger RMSD when comparing them, even if the molecular contours are otherwise identical. It is thus difficult to correctly calibrate the stopping condition for the clustering algorithm: One obtains either a large number of clusters with very similar bend angles and contact maps, or a smaller number of clusters in which states that appear different upon visual inspection or present distinct contact maps are wrongly clustered together. For this reason, a hybrid approch was used in this analysis. In this method, the simulation frames are first classified into a large set of candidate binding states through hierarchical agglomerative clustering; the number of clusters, n, is then reduced by one at a time. For some value of n, the formation of $n - 1$ clusters using the automated algorithm will result in two clusters with different bend angles or contact patterns being combined together; the clustering is halted at n clusters, and the remaining clusters are manually categorised according to their bend angles and contact maps. This occurred here at $n = 9$, and these nine states were classified by hand into five distinct binding modes; these states are illustrated in figure 15, and their contact maps in figure 16. (One cluster, consisting of around 2 % of the frames with Lk = 32 and having no well defined bend angle, was discarded.)

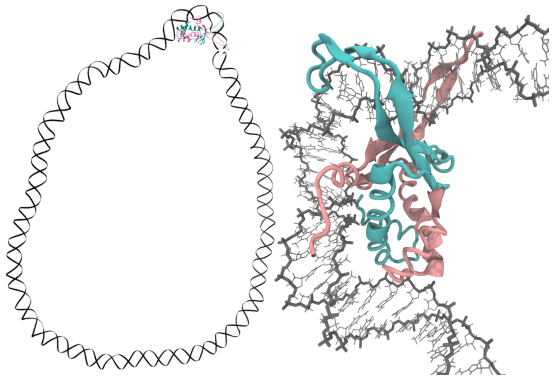
The fully wrapped state was once again recovered, with a mean bend angle of $136 \pm 18°$, in reasonable agreement with the results described for linear DNA in chapter 3. (The procedure used to calculate these angles differs slightly from that used in chapter 3, and will be discussed in detail in section 4.4.) However, the half-wrapped and associated states do not make an appearance, probably due to the inherent curvature of circular DNA—over the length of the IHF-interacting region, a perfectly circular piece of DNA of this size would have a curvature of around 64°, and one might expect this to bias the system in favour of more tightly bent conformations. Instead, a state emerges in which DNA binds fully on the left but only to the upper subunit on the right, which is herein termed the “three-quarters state” for want of better nomenclature. This contact pattern is associated with a bend angle of $140 \pm 12°$, surprisingly slightly sharper than that of the fully wrapped state, although well within one standard deviation.



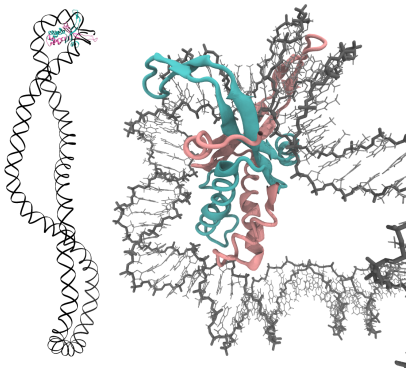
(a) Fully wrapped ($136 \pm 18°$)



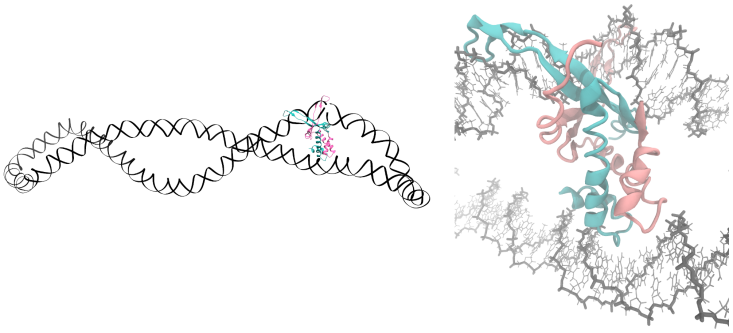
(b) Three quarters ($140 \pm 12°$)



(c) Half + bottom ($112 \pm 12°$)



(d) Three quarters + bottom ($117 \pm 13°$)



(e) Bridge ($66 \pm 14°$)

Figure 15: The IHF binding modes observed in simulations of minicircles can be divided into five principal states. These are the fully wrapped state (a) with a bend angle around 136°; a new state in which the left-hand side is fully bound while the right-hand side is partially bound (b), with a bend angle of 140°; two different states featuring binding of DNA on the left of the binding site to the side of the protein opposite the canonical site, one in which the right arm is unbound (c, 112°) and one in which it is partially bound (d, 117°); and a bridged state (e), the spontaneous formation of which supports the conclusion in chapter 3 that such a state is favourable. Once again, the bridged state features only minimal bending of around 66°.

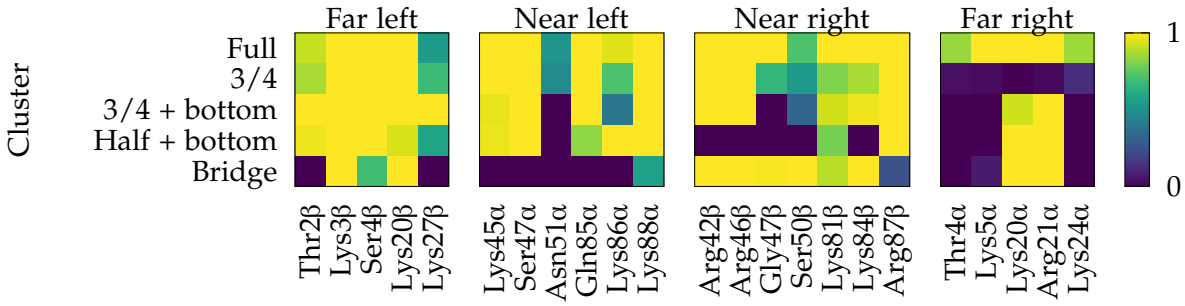


Figure 16: The binding modes can once again be classified according to the time-average number of hydrogen bonds formed by key residues (here capped at 1 for clarity).

The emergence of the new three-quarters state can be explained by considering the bending energy of the DNA away from the IHF binding site. That is, DNA away from the binding site is relatively stiff at these length scales and can sustain only a certain amount of curvature; pinching in of the minicircle at the IHF binding site must therefore be balanced by bulging out of the rest of the minicircle, which is associated with some elastic bending energy. This has the effect of making the fully wrapped state slightly less favourable and increasing the probability of states with more obtuse departure angles. One would expect this effect to be less pronounced for larger minicircles or those with a defect directly opposite the IHF binding site.

Two clusters feature binding of the left DNA arm to the bottom of the protein, involving a similar set of amino acids to the DNA bridges described previously. In one of these states, the right arm is entirely unbound, as in the half-wrapped state, corresponding to a mean bend angle of $112 \pm 12°$; in the other, the right arm binds to the near subunit as in the three-quarters state, resulting in a mean bend angle of $117 \pm 13°$. The left arm is fully bound in both cases.

Finally, one simulation of the most positively supercoiled minicircle was observed to spontaneously form a bridge as described in chapter 3. This is an exciting and serendipitous result, as it represents the first known observation of spontaneous IHF bridging in an MD simulation and lends significant weight to the conclusion previously discussed that DNA bridging by IHF is energetically favourable. This state featured bending over the binding site of just $66 \pm 14°$.

The populations of these states vary with the superhelical density of the DNA as in table 4. While torsionally relaxed minicircles bind to IHF in a variety of states, positively supercoiled DNA occupies exclusively the fully wrapped state unless a bridge is formed; the same holds for small amounts of negative supercoiling ($\sigma \approx -0.04$), but this pattern breaks down at a superhelical density close to the native level ($\sigma \approx -0.07$), returning to increased variability.

A possible explanation for this is that defects form in the DNA at these levels of torsional stress, as shown in figure 17; an example of a particularly large defect is

Table 4: Populations of binding states by linking number, Lk, and superhelical density, σ. ¾ refers to the three-quarters contact pattern, in which the full left side and right upper subunit of the protein bind to DNA, which is associated with a slightly sharper bend than the fully wrapped state. "Full" and "Half" refer to the fully and half-wrapped states, respectively; "+ B" indicates that the left DNA arm also binds to the "bottom" of the protein; and "Bridge" indicates the bridging of two distal DNA sites by IHF. Numbers in superscript identify the replicas in which each state appeared (arbitrarily numbered 1, 2, and 3 for each topoisomer).

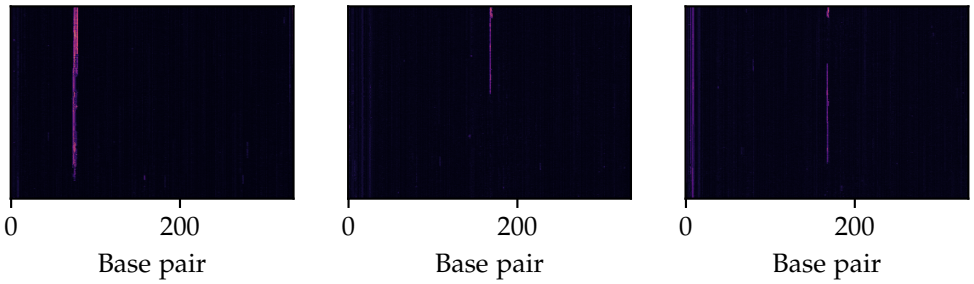
		Proportion of time in state / %				
Lk	σ	¾	Full	¾ + B	Half + B	Bridge
29	−0.067	33^{1}	33^{3}	33^{2}	–	–
30	−0.035	–	$100^{1,2,3}$	–	–	–
31	0.000	$45^{1,3}$	21^{3}	–	33^{2}	–
32	+0.030	–	$98^{1,2,3}$	–	–	–
33	+0.062	–	$100^{1,2,3}$	–	–	–
34	+0.094	–	$67^{2,3}$	–	–	33^{1}

shown in figure 18. These defects are not present—or are at least much smaller—in other topoisomers. This increased flexibility, concentrated in a region other than the IHF binding site, provides an additional hinge proportionally limiting IHF's dominance.

4.4 Bending of supercoiled DNA by IHF

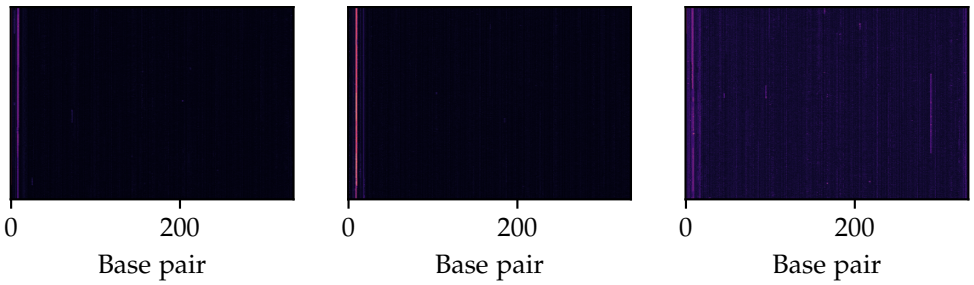
In order to further quantify the relationship between DNA supercoiling and the binding mode of IHF, it is instructive to consider the bend angle. In chapter 3, this was defined as the angle between two 30 bp vectors, separated by a further 30 bp centred on the binding site, when the whole structure was projected onto its best-fit plane. This methodology was selected as the closest analogue to the angle measurements performed on AFM images, which are necessarily two-dimensional and suffer from resolution limits.

The size of the minicircles studied here makes it difficult to accurately measure a bend angle from an AFM image, so there is no longer any reason to use an analogous methodology. Since biology does not occur on a plane, and projecting a three-dimensional structure into two dimensions inherently results in some distortion of angles, it is perhaps more instructive to measure the angles without first performing this projection. Furthermore, the substantial variability of the length of the IHF-interacting region of DNA requires that a different pair of vectors be defined for each structure: The states featuring bottom binding include



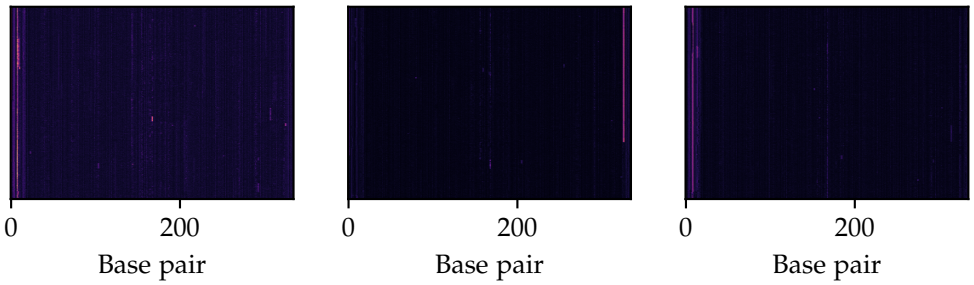
(i) Replica 1 **(ii)** Replica 2 **(iii)** Replica 3

(a) Lk = 29



(i) Replica 1 **(ii)** Replica 2 **(iii)** Replica 3

(b) Lk = 31



(i) Replica 1 **(ii)** Replica 2 **(iii)** Replica 3

(c) Lk = 33

Figure 17: These heatmaps represent the "index of denaturation", as described in section 2.2.4 on page 56, with the colour scale normalised individually for each panel. The x axis represents position along the helix, with the IHF binding site located at the far left (base pairs 1 to 11), and the y axis time, progressing upwards. Bright streaks indicate denatured regions of DNA; these are common in the presence of significant negative supercoiling (a), but not observed (other than at the IHF binding site) in torsionally relaxed DNA (b) or in positively supercoiled DNA of similar absolute superhelical density (c). One simulation in particular (a(i)) features a very large denatured bubble that continues to grow throughout the simulation, illustrated in figure 18.

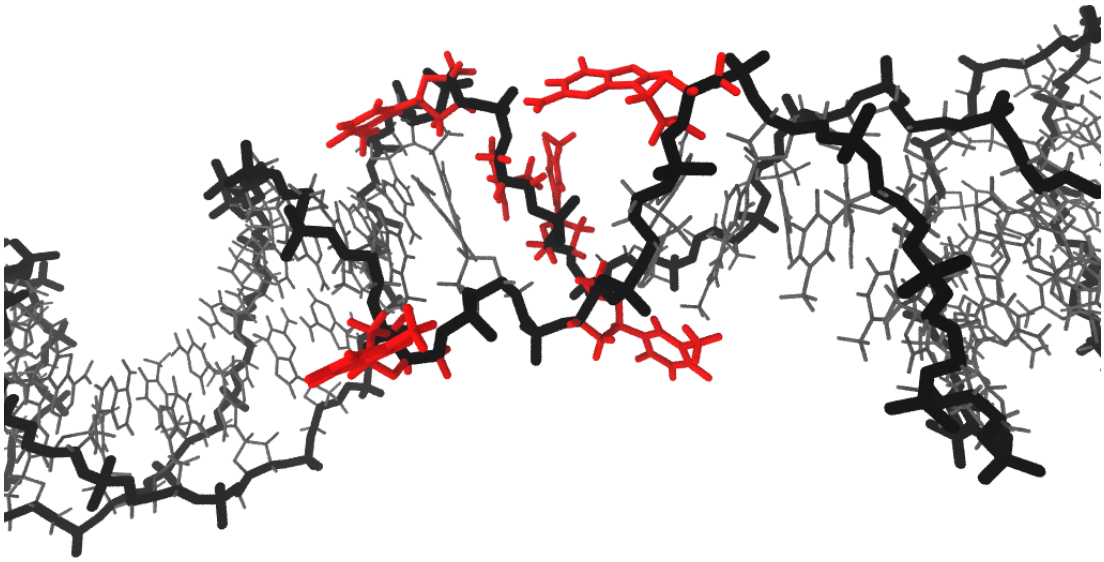


Figure 18: This figure illustrates a snapshot of the denaturation bubble formed in replica 1 with Lk = 29, visible in the index of denaturation in figure 17, with particularly disrupted nucleotides indicated in red.

substantial bending within the vectors previously used, while in other states the same piece of DNA is relatively far from the protein. When measuring bend angles in DNA minicircles, unlike in the linear case, it is desirable to use the shortest possible vectors. To illustrate this, consider that the 30 bp vectors, which corresponded to mostly straight regions of DNA in the linear case, each account for around 9 % of the minicircle's circumference, and thus represent a chord bypassing a significant segment of the circle's curvature. This results in a very unreliable measure of protein-induced bending.

For all of these reasons, the bend angles were in this case calculated by locating the last base pair on each side of the binding site that interacts with the protein and defining a vector joining the molecular contour corresponding to each of these base pairs to that 10 bp farther from the centre of the binding site (with this length chosen to represent approximately one full helical turn). The angle between these vectors was then calculated—without first projecting the structures onto a plane—using SerraLine. This results in more consistent bend angles for each state, with a smoother distribution, than the previous method, suggesting that it does indeed produce more reliable measurements for this system.

The mean bend angles (and associated standard deviations) of each binding mode were given, using this metric, in the previous section.

Based on the results described in the previous section, one might expect the bend angle about the IHF binding site to be positively correlated with the degree of supercoiling, since supercoiling enhances more tightly bound states, and that the distribution of bend angles will be broader for topoisomers that exhibit greater variability. However, this is not obvious from the clustering alone, since the angle

distributions for the various clusters are broad and overlapping—for example, the distribution of angles measured for the fully wrapped state fully encompasses both of the three-quarter states. Direct measurement of the bend angle distribution for each simulated degree of supercoiling will more clearly illuminate the relationship.

In fact, while significant differences do exist between the topoisomers, as illustrated in figure 19, there is no obvious correlation between the linking number and the position of the bend angle distribution. There is, however, a relationship between the frequency with which each bend angle is observed and the clusters populated by the frames of the simulations belonging to each topoisomer. Smaller bend angles, in the range 100–120°, are more common in topoisomers that occupy the states featuring bottom binding, angles between 120° and 150° in those that occupy the fully wrapped state, and angles above 150° in those that occupy the three-quarter state. However, this is only a rough and unreliable rule of thumb and distinct peaks are difficult to distinguish due to the breadths of the underlying distributions and the small differences between their means.

Regardless, relatively clean and sensible distributions are obtained, demonstrating the accuracy with which the bend angles can be calculated and further illustrating that DNA bending by IHF has a strong but complex relationship with DNA topology.

Supercoiling in either direction appears to be associated with an overall reduction in the mean, and certainly in the standard deviation, of the Gaussian distribution that best fits the main peak of each distribution (table 5), a potentially interesting and counterintuitive result indicating that IHF is able to bend DNA less efficiently but more consistently when it is supercoiled.

Table 5: The means and standard deviations of the Gaussian distributions that best fit the principal peak of the bend angle distribution for each value of Lk.

Lk	σ	Bend angle / °
29	−0.067	126 ± 20
30	−0.035	143 ± 26
31	0.000	139 ± 41
32	+0.030	140 ± 20
33	+0.062	135 ± 14
34	+0.094	130 ± 12

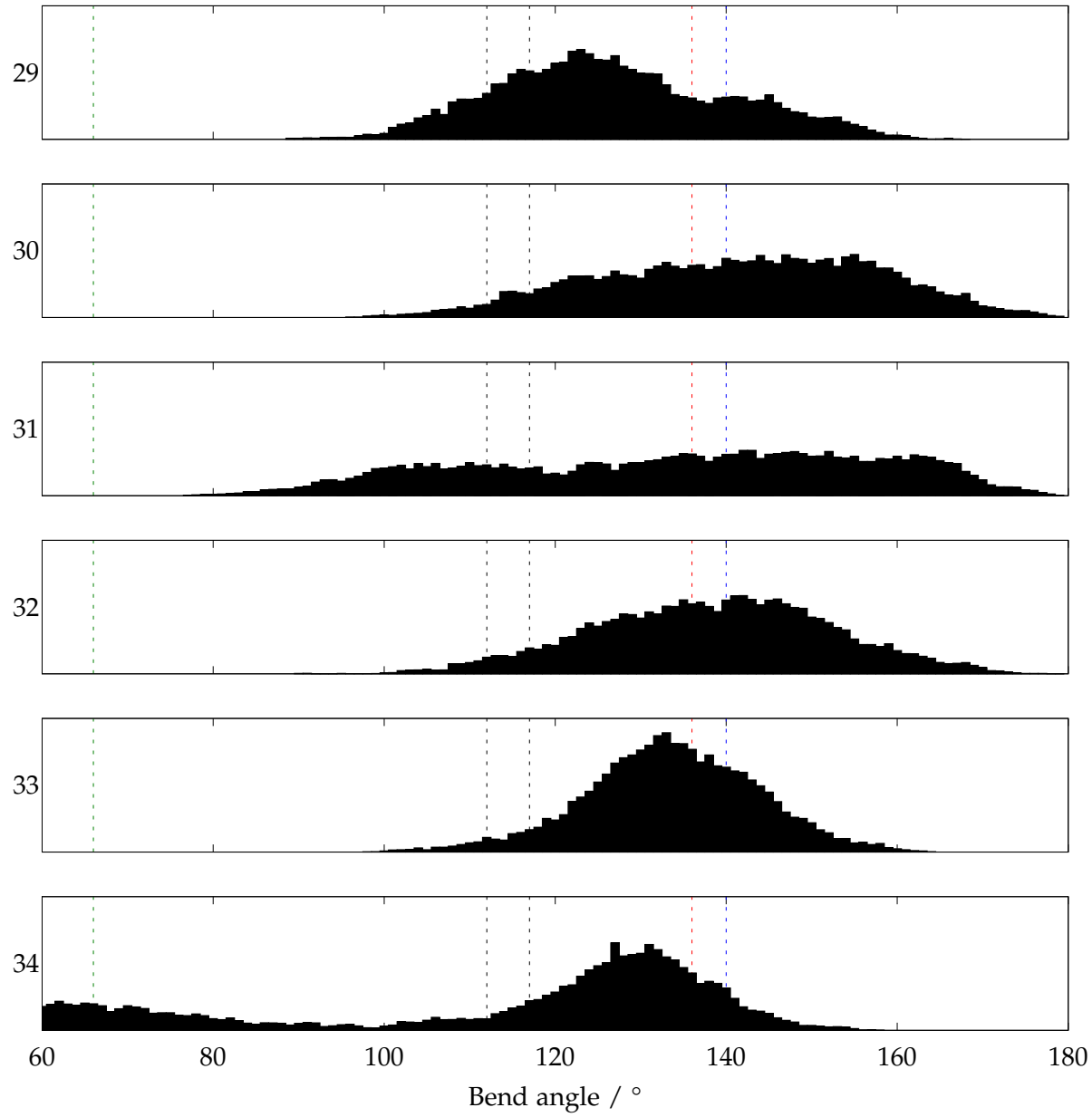


Figure 19: The distribution of bend angles about the IHF binding site for each value of Lk is broad and individual peaks are difficult to distinguish. However, it does illustrate the differing variabilities of the bend angles for different topoisomers. The bridged state (green, leftmost dotted line) is clearly visible as a broad but distinct peak for the one replica in which it was observed. The states with Lk = 29 and Lk = 31 feature significant populations in the two states featuring bottom binding (black), which have similar bend angles. The remaining measurements are clustered around the fully wrapped (red) and three-quarter (blue, rightmost dotted line) states, with the widths and locations of the peaks providing some clues about the underlying behaviour.
(Readers with reduced colour vision or a monochrome version of this document may find it helpful to note that the dotted lines correspond to the bend angles given in figure 15).

4.5 Influence of IHF on global plectoneme structure

The significant bending imposed upon DNA by IHF can be expected to influence the formation of plectonemes. In particular, one might imagine that increasing the favourability of DNA bending might shift the twist–writhe balance in favour of increased writhe.

Indeed, while there is no difference between the mean twist or writhe with and without IHF for the torsionally relaxed (Lk = 31) case, a small but significant reduction of twist is observed upon binding of IHF to minicircles with moderate degrees of supercoiling ($|\Delta Lk| = 1$), as shown in figure 20. This is associated with a corresponding change in writhe. This is interesting, as we observe in the absence of IHF a distinct asymmetry between positive and negative supercoiling—negative supercoiling is associated with more writhed structures. IHF appears to correct this asymmetry by shifting the writhe of both topoisomers in the positive direction. This effect is less strong for more strongly supercoiled DNA.

However, this does not rule out a role for IHF in plectoneme positioning, perhaps an even more important biological function. By visual inspection, it can be observed that IHF is always positioned at an apex of the plectoneme. This is sensible, as the apices are generally the most strongly bent part of a plectoneme, and the locations in which defects such as kinks and denaturation bubbles are most likely to form; relieving the free energy of plectoneme formation and of IHF binding in the same place allows the system to kill two birds with one stone. Whether this remains true for significantly larger DNA circles such as plasmids, which may be able to accommodate multiple sites of extreme bending, remains unknown.

To more robustly quantify this qualitative result, one can measure the distance of each point along the WrLine molecular contour from the centroid of the contour. In a plectoneme which becomes more rodlike with increasing writhe, the apices are the maxima of this distance, while the minima indicate the points at which distal DNA sites most closely approach one another.

In figure 21, torsionally relaxed mincircles are shown to be mostly circular in the absence of IHF, with IHF inducing some pinching-in of the structure centred on the binding site. Supercoiled minicircles, on the other hand, feature distinct maxima and minima corresponding to apices and points of closest approach. In the absence of IHF, these features can form in many positions and no alignment is observed between independent replica simulations. The addition of IHF causes these features to align across all replicas, with IHF located at an apex; the points at which distal sites most closely approach one another are also consistently located across replicas, a non-obvious result that indicates a possible role for IHF in controlling biologically

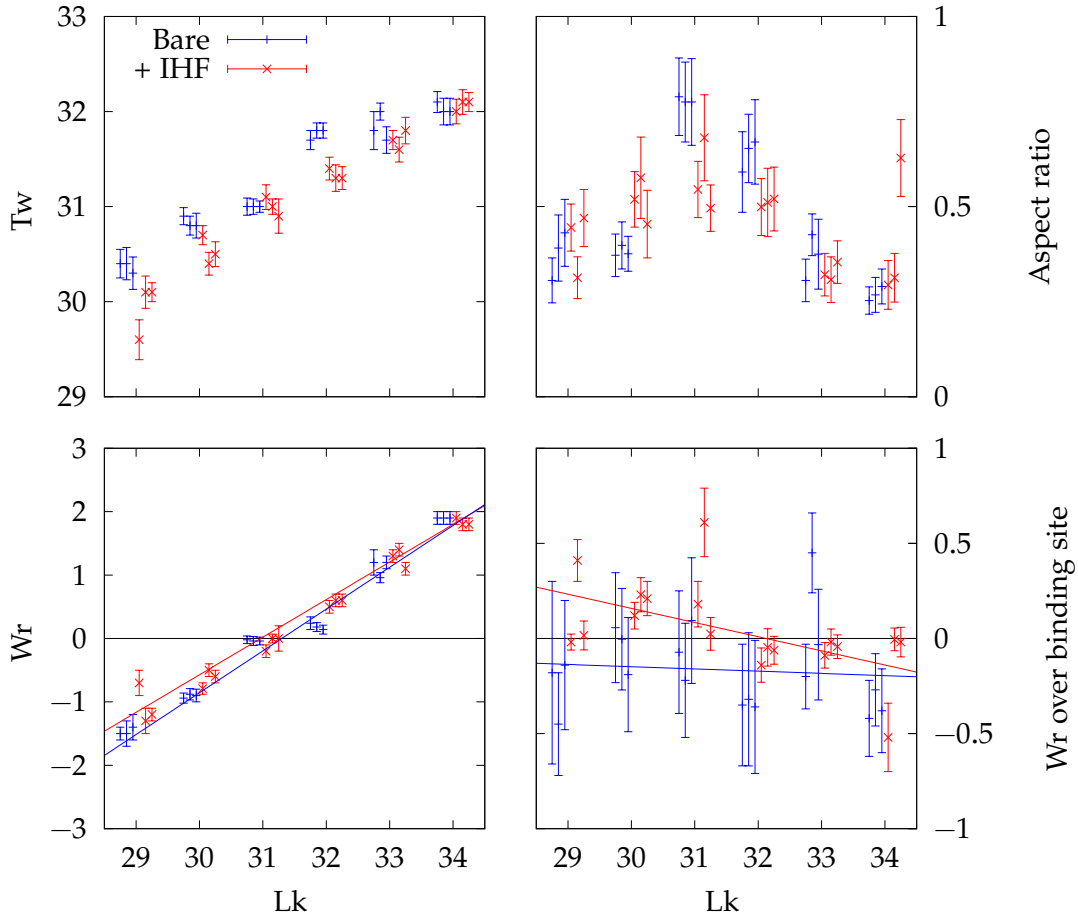


Figure 20: In the absence of IHF (blue), there is a distinct asymmetry between the +1 (Lk = 32) and −1 (Lk = 30) topoisomers in terms of both twist (top left) and writhe (bottom left). The effect is here most strongly identified by observing the asymmetry in the values of Wr. When IHF is bound (red), this asymmetry is corrected, with both topoisomers shifted towards slightly more positive (or less negative) values of Wr. This effect disappears for larger values of |ΔLk|.
Supercoiling causes minicircles to compact along the minor axis (defined as described in the main text), becoming more rodlike and resulting in a reduction in aspect ratio (top right); while bound IHF slightly enhances compaction of a torsionally relaxed minicircle, this effect disappears in the presence of even small amounts of supercoiling, and the aspect ratio remains relatively constant despite changes to the linking number.
Interestingly, however, IHF does significantly reduce the variability of the writhe around its binding site (bottom right, defined here as the familiar 61 bp region) and appears to cause a local reversal of the global supercoiling.
Horizontal jitter is applied to all points in this figure, with replicas displayed in an order corresponding to their assigned numbers as in table 4, and the + IHF data is displaced to the right of the bare data. The outlying values observed in replica 1 for Lk = 29 are associated with the formation of a large denatured bubble (figure 17), and for replica 1 with Lk = 34 with the formation of a bridge. The variablility in local writhe observed for bare DNA (e.g. for Lk = 33) is due to variable positioning of the plectonemes with respect to the binding site (see figure 21).

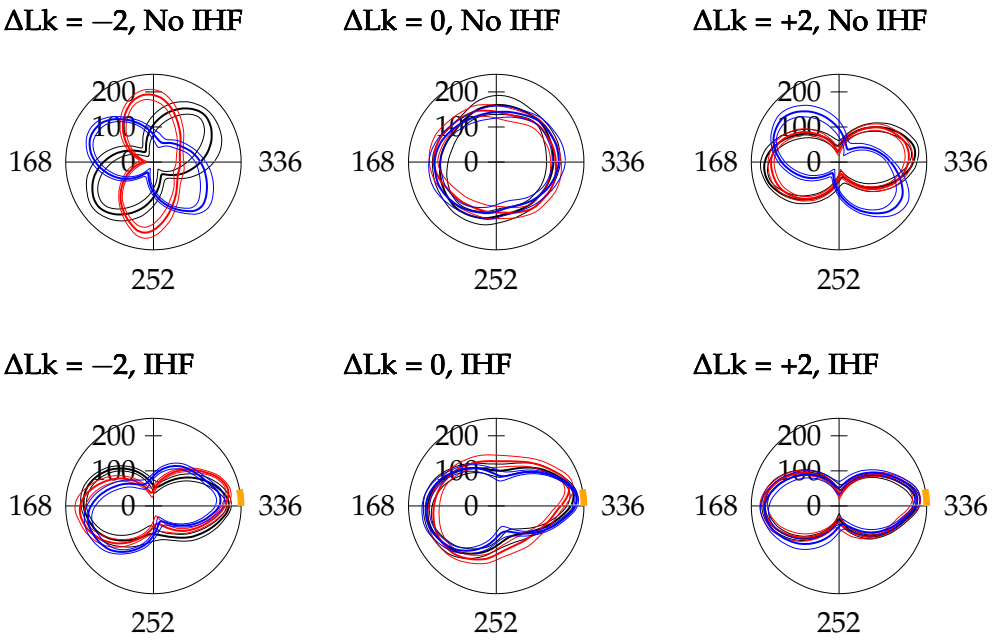


Figure 21: These plots show the distance of each point along the WrLine molecular contour from the centroid of the contour; the contour runs anticlockwise from the right of each plot, while the distance from the centroid in ångströms is represented by the distance of the coloured lines from the centre of the plot. Thick lines show the mean value over a replica, and thin lines represent a range of one standard deviation each side of the mean; the location of the IHF binding site is indicated by the thick orange portion of the outer ring. In the absence of IHF (top row), supercoiled minicircles are observed to form plectonemes in many positions (top left, top right); when IHF is present, the apices and crossing points of the plectoneme are consistent between replicas (bottom left, bottom right), with IHF always located at an apex regardless of the direction of supercoiling. Corresponding results for torsionally relaxed minicircles are shown in the centre column.

important interactions between DNA sites and distally bound regulatory proteins—perhaps even the bridging of DNA by another copy of itself.

IHF binding is also associated with a substantial reduction of the variability of the local writhe over its binding site (figure 20), which will of course also reduce the variability of the writhe available to other parts of the minicircle. This is consistent with the alignment of plectonemes, and seems a natural consequence of the additional electrostatic restraints placed on local DNA by bound IHF. More interestingly, the local writhe over the binding site appears to be negatively correlated with the global linking number of the mincircle when IHF is bound; in contrast, the variability of bare DNA is too large for any trend to be discerned. This is a highly counterintuitive observation that raises fascinating questions about the nature and extent of IHF's role in genome structuring. While it is possible that this observation is simply the result of limited sampling, the trend is sufficiently strong and apparent (with a gradient of -0.08 and a coefficient of determination $R^2 = 0.32$) to justify further investigation by an interested party.

DNA bending by IHF can also be expected to induce compaction. As the amount of supercoiling within a minicircle increases, it becomes more rodlike. That is, the minor axis (calculated here as the length of the short sides of the smallest rectangle that wholly contains the molecular contour after it has been projected onto the best-fit plane) is reduced by supercoiling, whatever the sign of ΔLk; in order to maintain a constant contour length, this is usually accompanied by an increase in the major axis (the length of the long sides of the same rectangle).* Perhaps a slightly more useful quantity is the aspect ratio—that is, the quantity m/M, where m is the minor axis and M is the major axis. Since changes in these quantities tend to be coupled, a seemingly small change in the axis lengths can result in a larger change in the aspect ratio, making clearer the underlying trend. A perfect circle (or square) would have an aspect ratio of 1, while an ideal rod with zero width would have an aspect ratio of 0.

IHF does aid in compaction of the minor axis of superhelically relaxed DNA, but this effect appears to be reduced by supercoiling; the difference in the aspect ratio between bare DNA minicircles and those bound to IHF is smaller for $|\sigma| \approx 0.03$ than for $\sigma = 0$, and disappears altogether for $|\sigma| > 0.05$ (figure 20). Thus, contrary to the traditional description of IHF as a source of nucleoid compaction, these results demonstrate that its effect on the overall compaction of DNA is small, particularly in the presence of physiological levels of DNA supercoiling. The true role of IHF is not to compact the genome, but instead to determine the axis along which it is to be compacted by other factors, which may include supercoiling and other DNA-

* This is not strictly necessary, as the contour length is not necessarily conserved when projected from three dimensions into two, but holds as a general rule.

binding proteins.

4.6 Structure of minicircles with multiple bound proteins

Of course, additional bound proteins complicate this simple model. For example, we have already seen that IHF has a strong preference for a position at the apex of a plectoneme; this preference cannot be simultaneously satisfied for multiple nearby copies of IHF bound to the same small minicircle. How this conflict is resolved is an important question if one wishes to truly understand the complex and interconnected structural regulatory mechanisms underlying gene expression. It is worth noting, additionally, that this is not simply an academic exercise; the genome *in vivo* is crowded by myriad proteins, all providing competing or cooperative structural influences, but alas we must start small, tuning into the symphony played by this molecular orchestra one note at a time.

To investigate these complex effects, minicircles were simulated that featured two IHF binding sites (of types H1 and H2), with the sequence listed in appendix 4.5. These minicircles, also of length 336 bp, are identical to those used for complementary AFM imaging, and possess two IHF binding sites as a consequence of the λ recombination technique by which they were produced [144]. The H1 site has a different sequence and a lower affinity than the H2 site.

The addition of another protein of course increases the variability of the system by adding additional degrees of freedom. Sampling the entire conformation space of these systems would be challenging, and the scope of this study is thus simply to expose some of the ways in which multi-IHF systems differ from those containing only a single IHF. Simulations were focused on the most relaxed ($\sigma = 0.000$, Lk = 31) and the most negatively supercoiled ($\sigma = -0.067$, Lk = 29) minicircles; again, three replicas were performed for each topoisomer in implicit solvent, each of 30 ns, with most analysis performed on the last 10 ns.

The structural effects of the additional protein once again appear to be dependent on DNA supercoiling. In the absence of supercoiling, the minicircle structural features are consistent and aligned with the locations of the binding sites, with the H1 site resulting in more significant bending and thus positioning itself farther from the minicircle's centroid than the H2 site, although both represent local maxima of the contour's distance from its centroid (figure 22). Despite the additional degrees of freedom contributed by the presence of an additional IHF, we see that the structure of such a minicircle is remarkably consistent and all features are well aligned. In the presence of supercoiling, one might expect to observe similar alignment of the main plectoneme features, much like in the single-IHF case. This is not what is observed. Instead, the supercoiled minicircle with two bound IHF proteins exhibits a surprising degree of variability; the observed plectoneme

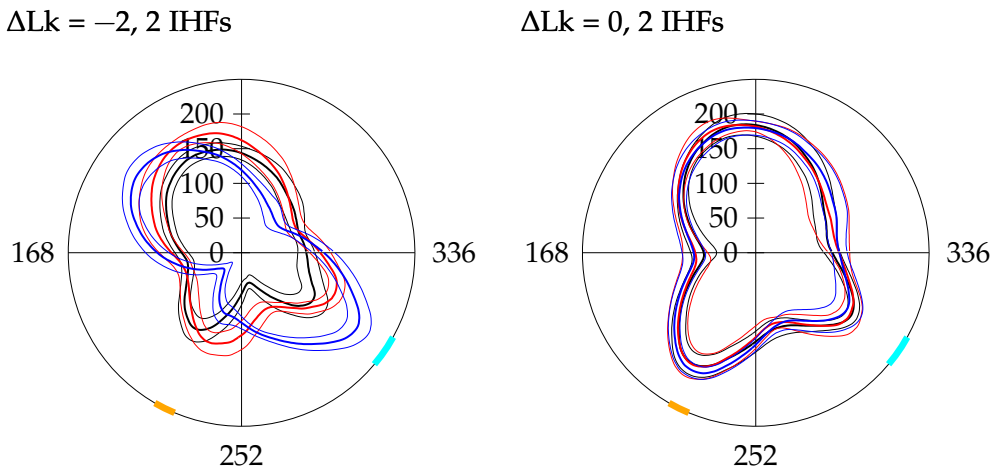
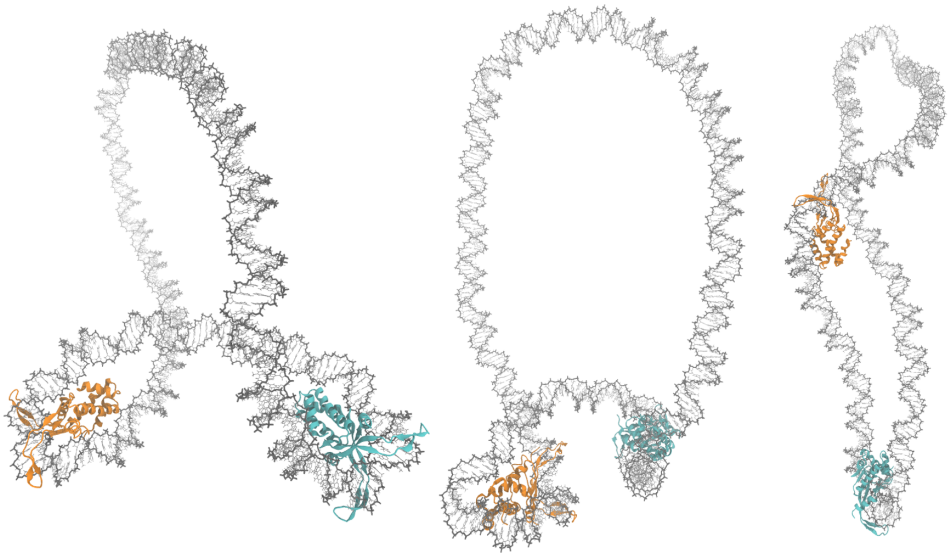


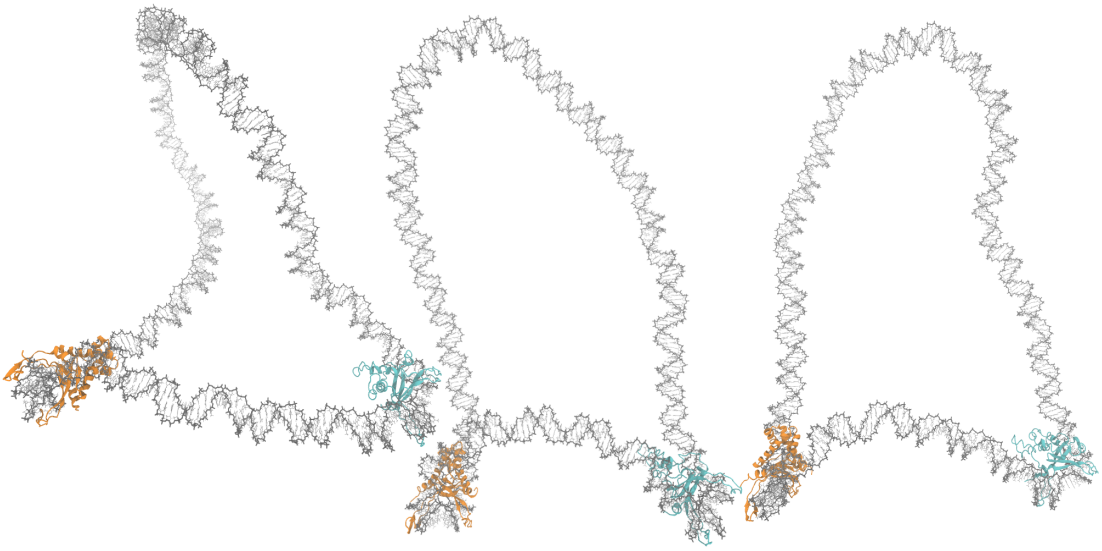
Figure 22: The global structure of minicircles bound to two copies of IHF is somewhat more complex than the single-IHF case shown in figure 21. Torsionally relaxed minicircles show remarkable consistency between replicas, with the H1 site (shown in orange, beginning around base pair 221) resulting in more bending and positioning itself farther from the centroid while the H2 site (shown in cyan, beginning around base pair 299) forms a secondary structural feature. In the presence of supercoiling, this order breaks down and no clear alignment is observed.

structures are complex and the features do not align. In fact, the IHF H1 binding site—which in the relaxed minicircle was consistently observed at the apex of the structure—is in one simulation instead the location at which the DNA most closely approaches itself. This is particularly interesting as it provides an opportunity for bridge formation, although in this case the opportunity was not taken. Visual inspection of the minicircle structures (figure 23) reveals similar results: Supercoiled minicircles with two binding sites display greater variety than otherwise identical torsionally relaxed minicircles.

This structural variety can be explained to an extent by the binding modes exhibited by the proteins in each case. In all three replica simulations of the relaxed minicircle, the H1 site binds in the fully wrapped state and the H2 site in the three-quarters state; that this results in a consistent minicircle structure is unsurprising but emphasises IHF's remarkable DNA-structuring abilities. In the presence of supercoiling, however, there is no such consistency; the H1 site binds variously in the fully wrapped state, the half-wrapped state (which here makes its only appearance in this chapter), and the half-wrapped state with bottom binding, while the H2 site binds either fully or in the three-quarters state with bottom binding. This appears to suggest that the relative strengths of the binding sites may be inverted in the presence of supercoiling, with the H1 site consistently binding more tightly than the H2 site when torsionally relaxed but showing greater variety when



(a) Lk = 29



(b) Lk = 31

Figure 23: The structures of supercoiled minicircles with two copies of IHF bound (a) are significantly more variable than those of otherwise similar minicircles with no supercoiling (b). The IHF bound to the H1 site is shown in orange and to the H2 site in cyan, with the binding sites aligned as closely as possible to their positions in figure 22.

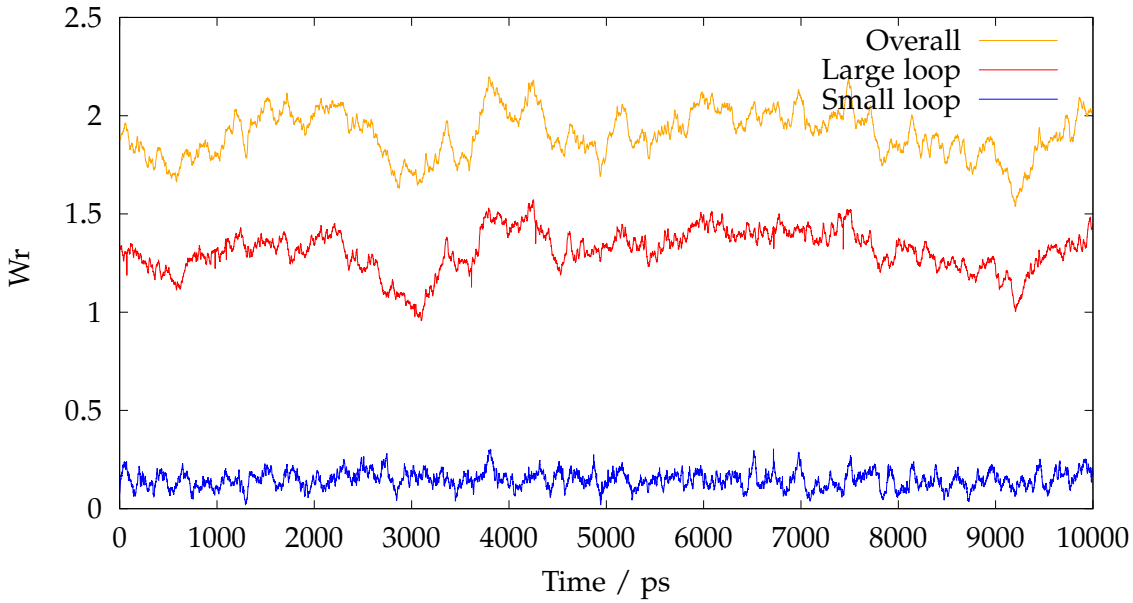


Figure 24: The overall writhe of the minicircle (orange, top line) is partitioned between the two loops closed by IHF, with most—as expected—lying within the larger of the two loops (red, middle line). While the writhe of the larger loop varies somewhat, representing the exchange of torsion between writhe and twist, the smaller loop (blue, bottom line) retains a constant writhe. There is no evidence of writhe passing between the loops, consistent with the formation of two distinct topological domains.

supercoiled.

The extent of the observed variety of supercoiled structures in the presence of multiple proteins makes it difficult to draw firm conclusions from these simulations, but constitutes an interesting result worthy of further investigation.

4.7 Bridging of supercoiled minicircles by IHF

The spontaneous formation of DNA bridges mediated by IHF was observed in simulations of supercoiled minicircles. That this occurred by chance despite the relatively limited number of simulations performed lends credence to the conclusion in section 3.4 that DNA–IHF–DNA bridges are a highly favourable, and thus strongly preferred, state.

Among these was a simulation with Lk = 34 (σ = +0.094) in which a bridge very similar to that induced through umbrella sampling, involving the same amino acids on the "bottom" of the protein, formed spontaneously and remained in place for the duration of the simulation.

This divides the minicircle into two closed loops, which one can expect to constitute distinct topological domains. The larger of these loops has a length of

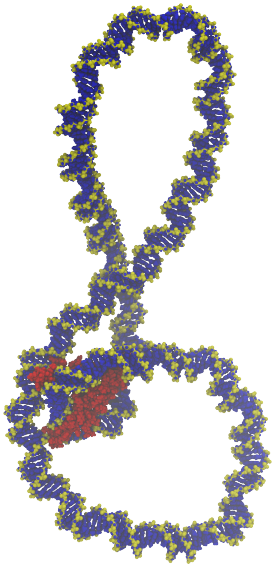


Figure 25: An unusual bridge involving the "arms" of the protein was observed in a minicircle with Lk = 29 during a preliminary simulation performed using a different set of parameters. It is presented here only as an interesting observation and an indication that IHF bridges may exhibit greater flexibility than otherwise demonstrated.

around 255 bp, while the shorter is approximately 81 bp. The writhe within each of these loops can be calculated in the normal manner, treating them as closed loops, and is displayed in figure 24. There is no evidence of writhe passing between the two loops, with the shorter loop maintining a consistent writhe of around 0.15 while the longer loop exhibits greater variability around a mean value of approximately 1.3. Perhaps interestingly, the writhe is not partitioned in perfect accordance with the lengths of the loops: While the larger loop accounts for 76 % of the minicircle's contour length, it holds 90 % of the combined writhe. That this asymmetry was not corrected by the diffusion of some writhe into the smaller loop provides some further evidence of the formation of topological domains, but may be in part an artefact of the differing lengths of the loops' molecular contours; note that the two writhes do not sum perfectly to the writhe of the whole minicircle for this very reason.

It therefore appears likely that IHF bridges do indeed divide topological domains, which are an important factor in topological regulation and genome structuring with a significant influence on plectoneme position and long-range intrachromosomal communication. This provides another mechanism by which IHF can control plectoneme position and adds another string to its nucleoid-structuring bow. Further studies designed specifically to investigate this effect would be valuable.

A quite distinct bridge (figure 25) was observed in a preliminary simulation of a minicircle with Lk = 29 ($\sigma = -0.067$), this time involving one of the protein's

arms, which formed additional contacts with a distal DNA site while remaining intercalated and bound to the minor groove of the canonical binding site in the usual fashion. This simulation was performed using a different set of conditions from the other simulations discussed herein, but is presented here as a further indication that bridging is possible and as a hint that the process may be even more flexible than previously suggested. Alternative bridging arrangements such as this would be a valuable target of further study.

It is also worth noting that the observed bridges occurred only at relatively high levels of supercoiling (in the most positively and negatively supercoiled structures, in fact). This makes sense, since distal DNA sites approach one another more closely in supercoiled DNA and such approaches are a prerequisite for bridge formation. This hints at a relationship between supercoiling and the favourability of IHF-mediated bridges, and that the interplay between IHF bridging and DNA topology is bidirectional.

Chapter 5

Discussion

This work constitutes some of the first simulations of DNA–protein binding dynamics at the atomistic level, reliant as it is on relatively recent advances in force field design and hardware availability. In particular, the DNA constructs simulated in the preceding chapters are relatively large for atomistic simulation, even in implicit solvent. Furthermore, principally due to historic difficulties in combining force fields for proteins and nucleic acids [91], a large number of DNA–protein interactions remain unsimulated. In fact, the atomistic simulation as described herein of a NAP bound to supercoiled DNA appears to be entirely novel. By demonstrating the feasibility of large-scale atomistic simulations of DNA–nucleoprotein interactions in both implicit and explicit solvent, it is hoped that this work will stimulate further application of these methods. While this study does not represent the first application of the weighted histogram analysis method to DNA–protein binding [148], the recovery of a multidimensional free-energy landscape is a significant step towards understanding the nature of a complex interaction such as those described in the preceding chapters. The combination of distinct experimental and simulation methodologies validates the results thereby obtained and provides a simultaneously broader and more detailed analysis of the system's behaviour than could any one technique alone; this should be taken as a framework for future research into the physics of life.

While the existence of multiple binding modes for IHF was already known [48], this work provides the first look at their structures, and does so in atomistic detail, accompanied by an estimate of the relative free energies of the states—a significant advance towards understanding of IHF's many important biological roles. Much of this work is likely to be transferable to other NAPs, of which IHF can be reasonably expected to be representative. The similarities between IHF and HU are immediately apparent, so it would be reasonable to assume that there are similarities in their binding behaviour. Both IHF and HU are vital for the regulation of a number of genetic processes across the bacterial domain, and

the multimodal mechanical switching described for the first time in chapter 3 sheds new light on the factors by which this behaviour may be controlled. That this behaviour is observed even for the H2 binding site, which is thought to be the strongest, suggests that similar alternative binding modes are likely to be present, perhaps in even higher proportions, when the protein is bound to weaker binding sites; the behaviour for sites as weak as those favoured by HU (which have dissociation constant K_d = 200–2500 nM [149] compared to K_d = 2–20 nM [150] for specific binding of IHF) remains unexplored. These conclusions may be transferable even to other DNA-bending NAPs, such as Fis, a very important protein involved in a number of complex regulatory pathways [65], since the binding of most NAPs is thought to be driven by similar electrostatics.

The bridging of two DNA molecules, and of DNA minicircles, by IHF is reminiscent of the behaviour of a number of other NAPs, including H-NS. The bridging of DNA by IHF specifically is of particular importance to the study of biofilms, in which IHF frequently stabilises the extracellular DNA scaffold. This work provides a deeper understanding of the bridging behaviour of IHF, and demonstrates that the strength of bridging interactions far outweighs IHF's more familiar DNA bending. While this is not of immediate utility to antibiotic development, these results do enhance our understanding of the mechanisms by which microbial communities protect themselves, and understanding the enemy is a necessary step towards vanquishing them. Unfortunately, large DNA–IHF aggregates, such as those observed under AFM in the presence of multiple binding sites, are too large to be simulated atomistically; the development or application of a coarse-grained model to explain the behaviour of these structures in bulk would be of some interest.

The complex interplay between supercoiling and DNA bending by nucleoid proteins remains an unmapped frontier in the physics of life, but the data herein shines a candle upon its edges, illustrating and roughly quantifying the effect of DNA supercoiling on IHF binding. The role of IHF in positioning plectonemes is explored and to some extent explained, and intriguing data is presented suggesting that IHF binding may be associated with a local reversal of the global supercoiling, which must be accompanied by an equal and opposite enhancement of the change in the superhelical density elsewhere.

At the very least, this work has demonstrated the validity of this simulation methodology, resulting in a procedure that can be applied with relative ease to similar DNA–protein systems in order to extract detailed binding information and establish the energetics driving the behaviour of biological systems. The combination of MD simulations and single-molecule experiments is shown to be both possible and fruitful, suggesting a productive framework for future studies.

5.1 Future work

This work focused almost exclusively on binding of IHF to its H2 binding site. It would be trivial but supremely interesting to repeat this work on sequences containing a different binding site, or none at all. Simply mutating away the DNA consensus sequence would be expected, based on these observations, to result in a different free-energy landscape prohibiting the fully wrapped state; producing and quantifying this would further validate this work. A more involved study considering a number of different sequences could even result in an understanding of exactly which features of the consensus sequence are of the greatest importance, and even suggest paths along which the IHF binding sites may have evolved, as has been done recently for the ParB–*parS* and Noc–*NBS* complexes [151]. Sequences lacking an upstream A-tract could be considered in a similar manner.

Replicating this methodology for HU would also be interesting, allowing the similarities and differences between the binding behaviours of IHF and HU to be elucidated. Recovery of free-energy landscapes for HU bound to damaged and undamaged DNA, alongside corresponding landscapes for nonspecific binding of IHF, may explain the extreme differences between the proteins' binding strengths and establish whether HU exhibits the same multimodality and mechanical switching observed for IHF.

The range of superhelical densities herein considered is relatively narrow, and sampling of the conformation space for each topoisomer is limited. Additional simulations covering a broader range of values of Lk would help to refine these results. Unfortunately, the size of DNA mincircles makes their conformations inaccessible at present to explicit-solvent simulations of the type required for methods such as WHAM, but the recovery of free energy landscapes for the DNA–IHF system in different topologies would represent a significant step forward into the frontier at which this work dare only peek.

Further afield, collaborations between single-molecule experiments and advanced atomistic simulations as described in this work could in principle be applied to any interacting system sufficiently small that it can be simulated in explicit solvent. Other nucleoid-associated proteins, such as H-NS and Fis, are good candidates for this, but there is little reason to expect that these methods should not work for any system of the interested reader's choosing.

Appendix 1

Modified WrLINE code

The code in this appendix is based on the original WrLine implementation by Sutthibutpong, Harris, and Noy as described in their 2015 paper [128]. This is availabe under version 3.0 of the GNU General Public License* in the agnesnoy/WrLINE GitHub repository.†

This version is made available under the same terms and can be found in the georgewatson/reWrLINE GitHub repository.‡ Additional input files used by `test.py` are available there, but are required only to run the test suite.

Appendix 1.1 WrLINE.py

```
#! /usr/bin/env python3

"""
         __         __    _     ___ _   _ _____
 _ __ __\ \       / / __| |   |_ _| \ | | ____|
| '__/ _ \ \ /\ / / '__| |    | ||  \| |  _|
| | |  __/\ V  V /| |  | |___ | || |\  | |___
|_|  \___| \_/\_/ |_|  |_____|___|_| \_|_____|

reWrLINE: A reimplementation of WrLINE

(c) 2019 George D. Watson, University of York
https://georgewatson.me

Based on WrLINE
by Thana Sutthibutpong, Sarah Harris, and Agnes Noy.
Please cite
Sutthibutpong T, Harris S A and Noy A 2015 J. Chem. Theory Comput. 11 2768-75
https://doi.org/10.1021/acs.jctc.5b00035
"""
```

---

* https://www.gnu.org/licenses/gpl-3.0.en.html

† https://github.com/agnesnoy/WrLINE

‡ https://github.com/georgewatson/reWrLINE

```
import sys
import os
import writhe
import caxislib

print(__doc__)
print("---\n")

name = sys.argv[1]
top = sys.argv[2]
traj = sys.argv[3]
num_bp = int(sys.argv[4])
num_steps = int(sys.argv[5])
try:
    linear = sys.argv[6] not in ('0', 'False')
except IndexError:
    linear = False

# Strip trajectory to get C1' coordinates
os.system(f'mkdir -p {name}')
os.system('\n'.join(['cpptraj <<EOF',
                     f'parm {top}',
                     f'trajin {traj}',
                     "strip !(@C1') outprefix C1",
                     f'trajout {name}/C.mdcrd',
                     'EOF']))

print(f"Processing {name}")
print(f"Treating system as {'linear' if linear else 'circular'}")

print("Reading files & initialising arrays")
strand_a, strand_b, midpoints = caxislib.read(name, num_bp, num_steps,
                                              linear=linear)

print("Calculating first-order helical axis")
helix_axis = caxislib.helix_axis(num_bp, num_steps, midpoints, strand_a,
                                 linear=linear)

print("Calculating twist")
twist = caxislib.full_twist(name, num_bp, num_steps, strand_a, strand_b,
                            helix_axis, linear=linear)

print("Calculating helical axis")
caxis = caxislib.caxis(name, num_bp, num_steps, midpoints, twist,
                       linear=linear)

print("Calculating register angles")
sinreg = caxislib.sinreg(name, num_bp, num_steps, midpoints, caxis)

print("Writing output .xyz and .3col files")
caxislib.make_files(name, num_bp, num_steps, midpoints, caxis)

print("Calculating writhe")
wr = writhe.main(name, num_bp, num_steps, linear)

print(f"Job {name} done!")
```

Appendix 1.2 writhe.py

```
import numpy as np


def read_3col(filename, num_bp, num_steps):
    """
    Reads a 3col coordinate file and splits it by timestep
    """
    coords = np.loadtxt(filename)
    x = []
    t = 0
    for i in range(num_steps):
        x.append(coords[t*num_bp:(t+1)*num_bp, :])
        t += 1
    return np.array(x)


def writhe(coords, t, length, axis=2, linear=False):
    """
    Calculates write for a single timestep
    """
    shape = np.shape(coords[t])
    y = np.zeros((shape[0]+(0 if linear else 1), shape[1]))
    # Read one bp-step
    y[:shape[0], :] = coords[t]
    if not linear:
        # Add the head at the bottom
        y[shape[0], :] = coords[t, 0]
    result = 0
    for j in range(length):
        for k in range(j):
            tangents_j = y[j+1] - y[j]
            tangents_k = y[k+1] - y[k]
            # Vector joining j and k
            vector = y[j] - y[k]
            # Discretised Gauss integral
            # Add each individual contribution from a pair of tangent vectors
            result += (np.dot(vector, np.cross(tangents_j, tangents_k)) /
                       (np.linalg.norm(vector)**3 * 2 * np.pi))
    return result


def main(name, num_bp, num_steps, linear=False, write=True):
    # Read file
    coords = read_3col(name + '/C1.3col', num_bp, num_steps)
    length = len(coords[0])
    # Calculate writhe for num_steps timesteps
    wr = []
    for t in range(num_steps):
        print(f"\r\tStep {t}...", end=" ")
        wr.append([t+1, writhe(coords, t, length, linear)])
    wr = np.array(wr)
    if write:
        np.savetxt(name+'/writhe.ser', wr, fmt='%5d %9.4f')
    print("Done!")
```

```
    return wr
```

Appendix 1.3 caxislib.py

```
import numpy as np


# Maths


def cross(a, b):
    """
    Cross product of two arrays of 3-vectors
    """
    assert(np.shape(a)[0] == 3)
    assert(np.shape(a) == np.shape(b))
    return np.array([a[1]*b[2] - a[2]*b[1],
                     a[2]*b[0] - a[0]*b[2],
                     a[0]*b[1] - a[1]*b[0]])


def dot(a, b):
    """
    Dot product of two arrays of vectors
    """
    assert(np.shape(a) == np.shape(b))
    return sum(a[i] * b[i] for i in range(len(a)))


def norm(a):
    """
    Norms of an array of vectors
    """
    return np.sqrt(sum(x**2 for x in a))


def rotate_to_x(vector):
    """
    Returns the rotation matrix that rotates a vector to the x axis
    """
    normalised = vector / np.linalg.norm(vector)

    # C = -arctan(y / x)
    c = -np.arctan2(normalised[1], normalised[0])
    return np.array([[np.cos(c), -np.sin(c), 0.0],
                     [np.sin(c), np.cos(c),  0.0],
                     [0.0,       0.0,        1.0]])


def rotate_to_z(vector):
    """
    Returns the rotation matrix that rotates a vector to the z axis
    as follows:
      Define rotation matrix about (x, y) axis & Euler angles as
        Rxy(A, B) = Ry(B) `dot` Rx(A)
      Solve
        (0, 0, 1) = Rxy `dot` (x, y, z)
    """
```

```
    normalised = vector / np.linalg.norm(vector)

    # A = arctan(y / z)
    a = np.arctan2(normalised[1], normalised[2])
    # B = arctan(-x / |y, z|)
    # Explicitly NOT arctan2
    b = np.arctan(-normalised[0] /
                  np.sqrt(normalised[1]**2 + normalised[2]**2))

    rotate_x = np.array([[1.0,        0.0,       0.0],
                         [0.0,        np.cos(a), -np.sin(a)],
                         [0.0,        np.sin(a), np.cos(a)]])
    rotate_y = np.array([[np.cos(b),  0.0,       np.sin(b)],
                         [0.0,        1.0,       0.0],
                         [-np.sin(b), 0.0,       np.cos(b)]])

    return np.dot(rotate_y, rotate_x)


def twist(a1, a2, b1, b2, z):
    """
    Returns the twist
        a1 & a2 are the points defining the first vector
        b1 & b2 are the points defining the second vector
        z is the vector defining the z axis
    """
    vector_a = a2 - a1
    vector_b = b2 - b1

    unit_a = vector_a / np.linalg.norm(vector_a)
    unit_b = vector_b / np.linalg.norm(vector_b)
    unit_z = z / np.linalg.norm(z)

    # Generate rotation matrix rotating basis to the z axis
    rotation_matrix = rotate_to_z(unit_z)
    # Perform rotation
    unit_a = np.dot(rotation_matrix, unit_a)
    unit_b = np.dot(rotation_matrix, unit_b)

    # Now rotate such that a lies along the x axis
    rotated_b = np.dot(rotate_to_x(unit_a), unit_b)
    return np.arctan2(rotated_b[1], rotated_b[0]) * 180.0 / np.pi


# IO


def read(name, num_bp, num_steps, linear=False):
    """
    Reads a .mdcrd file & create returns a 3D array of atomic coordinates
    """
    with open(name + '/C.mdcrd', 'r') as file:
        file.readline()
        data = []
        for line in file:
            line_length = len(line)
            num_fields = line_length // 8
```

```
            for i in range(num_fields):
                data.append(float(line[i*8:(i+1)*8]))

        data_array = np.reshape(np.array(data), (len(data)//3, 3))

        x = np.reshape(data_array[:, 0], (num_steps, num_bp*2))
        y = np.reshape(data_array[:, 1], (num_steps, num_bp*2))
        z = np.reshape(data_array[:, 2], (num_steps, num_bp*2))

        strand_a = np.array([x[:, 0:num_bp], y[:, 0:num_bp], z[:, 0:num_bp]])
        strand_b = np.array([x[:, (2*num_bp - 1):(num_bp - 1):-1],
                             y[:, (2*num_bp - 1):(num_bp - 1):-1],
                             z[:, (2*num_bp - 1):(num_bp - 1):-1]])

        # Coordinate representation of a base-pair step
        midpoints = np.zeros(np.shape(strand_a))
        for i in range(num_bp):
            if linear and i+1 >= np.shape(strand_a)[2]:
                midpoints[:, :, i] = 0.5 * (strand_a[:, :, i] +
                                            strand_b[:, :, i])
            else:
                midpoints[:, :, i] = (0.25 *
                                      (strand_a[:, :, i] +
                                       strand_a[:, :, ((i + 1) % num_bp)] +
                                       strand_b[:, :, i] +
                                       strand_b[:, :, ((i + 1) % num_bp)]))

        return strand_a, strand_b, midpoints


def make_files(name, num_bp, num_steps, midpoints, caxis):
    """
    Makes xyz files of average C1' single helix
        midpoints is the array of midpoints of the C1' atoms of neighbouring
            base pairs
        caxis is the fully processed helical CAXIS
    """
    with open(name + '/C.xyz', 'w') as c_xyz, \
            open(name + '/C1.xyz', 'w') as c1_xyz, \
            open(name + '/C.3col', 'w') as c_3col, \
            open(name + '/C1.3col', 'w') as c1_3col:
        for i in range(num_steps):
            c_xyz.write(f"{num_bp}\n\n")
            c1_xyz.write(f"{num_bp}\n\n")
            print(f"\r\tStep {i}...", end=" ")
            for j in range(num_bp):
                c_xyz.write(" ".join(["H",
                                      f"{midpoints[0][i][j]:8.3f}",
                                      f"{midpoints[1][i][j]:8.3f}",
                                      f"{midpoints[2][i][j]:8.3f}",
                                      "\n"]))
                c1_xyz.write(" ".join(["H",
                                       f"{caxis[0][i][j]:8.3f}",
                                       f"{caxis[1][i][j]:8.3f}",
                                       f"{caxis[2][i][j]:8.3f}",
                                       "\n"]))
                c_3col.write(" ".join([f"{midpoints[0][i][j]:8.3f}",
```

```
                                        f"{midpoints[1][i][j]:8.3f}",
                                        f"{midpoints[2][i][j]:8.3f}",
                                        "\n"]))
                c1_3col.write(" ".join([f"{caxis[0][i][j]:8.3f}",
                                        f"{caxis[1][i][j]:8.3f}",
                                        f"{caxis[2][i][j]:8.3f}",
                                        "\n"]))
        print("Done!")


def sinreg(name, num_bp, num_steps, midpoints, caxis, write=True):
    """
    Calculates sine of register angles
    """
    result = np.zeros((num_steps, num_bp + 1))
    for j in range(num_bp):
        print(f"\r\tBase pair {j}...", end=" ")
        m = list(map(int, np.linspace(j-1, j+1, num=3) % num_bp))
        # Vectors bent on a plane
        v0 = caxis[:, :, m[1]] - caxis[:, :, m[0]]
        v1 = caxis[:, :, m[2]] - caxis[:, :, m[1]]
        plane_vector = cross(v1, v0)
        minor_groove = midpoints[:, :, m[0]] - caxis[:, :, m[0]]
        # sinreg[:, j+1] = |M <cross> C| / (|M| |C|)
        # where M = minor_groove & C = plane_vector
        result[:, j+1] = (norm(cross(minor_groove, plane_vector)) /
                          (norm(minor_groove) * norm(plane_vector)))
        # To obtain the sign of the result,
        # where positive is a minor groove pointing into the circle,
        # take the dot product of the minor groove with the unit normal vector
        for i in range(num_steps):
            if dot(v1 - v0, minor_groove)[i] < 0:
                result[i, j+1] = -result[i, j+1]
    result[:, 0] = np.linspace(0.01, 0.01*num_steps, num=num_steps)
    if write:
        np.savetxt(name + '/sinreg.ser', result, fmt='%8.3f')
    print("Done!")
    return result


def helix_axis(num_bp, num_steps, midpoints, strand_a, linear=False):
    """
    Calculates the first-order helical axis
    without taking the weight.
    Used for twist calculation.
    """
    result = np.zeros(np.shape(strand_a))
    for j in range(num_bp):
        print(f"\r\tBase pair {j}...", end=" ")
        # Summation of coordinates
        summation = np.zeros((3, num_steps))
        for t in range(num_steps):
            # Sum single helix position
            summation[:, t] += midpoints[:, t, j]
            k = 0
            while k < 5:
                k += 1
```

```
                if linear:
                    try:
                        summation[:, t] += (midpoints[:, t, j-k] +
                                            midpoints[:, t, j+k])
                    except IndexError:
                        k -= 1
                        break
                else:
                    summation[:, t] += (midpoints[:, t, (j-k) % num_bp] +
                                        midpoints[:, t, (j+k) % num_bp])
            # Average helix (almost full turn)
            result[:, t, j] = summation[:, t] / (2*k + 1)
    print("Done!")
    return result


def full_twist(name, num_bp, num_steps, strand_a, strand_b, haxis,
               linear=False, write=True):
    """
    Calculates twist
    """
    result = np.zeros((num_steps, num_bp))
    for j in range(num_bp):
        print(f"\r\tBase pair {j}...", end=" ")
        for t in range(num_steps):
            # Linear special cases
            if linear:
                # Does haxis[:, :, j+1] exist?
                if np.shape(haxis)[2] > j + 1:
                    # If haxis[:, :, j-1] doesn't exist,
                    # approximate using half the range
                    if j > 0:
                        z = haxis[:, t, j+1] - haxis[:, t, j-1]
                    else:
                        z = 2 * (haxis[:, t, j+1] - haxis[:, t, j])
                    result[t, j] = twist(strand_a[:, t, j],
                                         strand_b[:, t, j],
                                         strand_a[:, t, (j+1)],
                                         strand_b[:, t, (j+1)],
                                         z)
                # If not, just return zero
                # This may not be the best way to handle this
                else:
                    result[t, j] = 0
            else:
                z = haxis[:, t, (j+1) % num_bp] - haxis[:, t, (j-1) % num_bp]
                result[t, j] = twist(strand_a[:, t, j],
                                     strand_b[:, t, j],
                                     strand_a[:, t, (j+1) % num_bp],
                                     strand_b[:, t, (j+1) % num_bp],
                                     z)
    if write:
        np.savetxt(name + '/tw.ser', result, fmt='%8.3f')
    print("Done!")
    return result
```

```
def caxis(name, num_bp, num_steps, midpoints, tw, linear=False):
    """
    Calculates the central helical axis
    by performing the running average of each bp with its 2*k neighbours
    & including the weight of the excess base pair
    """
    result = np.zeros(np.shape(midpoints))
    for j in range(num_bp):
        print(f"\r\tBase pair {j}...", end=" ")
        total_twist = np.zeros(num_steps)
        summation = np.zeros((3, num_steps))
        for t in range(num_steps):
            total_twist[t] += tw[t, j]
            summation[:, t] += midpoints[:, t, j]
            k = 0
            # Find the point where two more flanking steps would make twist
            # exceed 360 degrees
            while total_twist[t] < 360.0:
                k += 1
                prev = total_twist[t]
                # In linear DNA, only go as far as the ends
                # This might not be the best approach
                if linear and (j-k < 0 or j+k >= num_bp):
                    break
                else:
                    total_twist[t] += (tw[t, (j-k) % num_bp] +
                                       tw[t, (j+k) % num_bp])
                    # Sum single helix position
                    summation[:, t] += (midpoints[:, t, (j-k) % num_bp] +
                                        midpoints[:, t, (j+k) % num_bp])
            # If the linear condition was met,
            # the twist must be less than 360 degrees,
            # so there's no need to remove the last two flanking steps
            if linear and (j-k < 0 or j+k >= num_bp):
                weight = 0
            else:
                # Add the flanks with weight < 1
                weight = (360.0 - prev) / (total_twist[t] - prev)
                summation[:, t] -= ((1 - weight) *
                                    (midpoints[:, t, (j-k) % num_bp] +
                                     midpoints[:, t, (j+k) % num_bp]))
            weight_3d = np.array([weight, weight, weight])
            result[:, t, j] = summation[:, t] / (2*(k + weight_3d) - 1)
    print("Done!")
    return result
```

Appendix 1.4 test.py

```
#! /usr/bin/env python3

# pylint: disable=anomalous-backslash-in-string

"""
         __         __    _     ___ _   _ _____
 _ __ __\ \       / / __| |   |_ _| \ | | ____|
| '__/ _ \ \ /\ / / '__| |    | ||  \| |  _|
| | |  __/\ V  V /| |  | |___ | || |\  | |___
|_|  \___| \_/\_/ |_|  |_____|___|_| \_|_____|

reWrLINE: A reimplementation of WrLINE

(c) 2019 George D. Watson, University of York
https://georgewatson.me

Based on WrLINE
by Thana Sutthibutpong, Sarah Harris, and Agnes Noy.
Please cite
Sutthibutpong T, Harris S A and Noy A 2015 J. Chem. Theory Comput. 11 2768-75
https://doi.org/10.1021/acs.jctc.5b00035

This is the test suite.
It is recommended that you run this before using this implementation.
Tests ensure consistency with WrLINE, not correctness.
"""  # noqa

import filecmp
import sys
import numpy as np
import caxislib
import writhe

try:
    from termcolor import colored
except ImportError:
    def colored(s, _):
        return s

print(__doc__)
print("---\n")

tolerance = 0.1

name = 'test'
num_bp = 336
num_steps = 8

a = np.array([[330.0, 330.0],
              [50.0,  40.0],
              [0.0,   10.0]])
b = np.array([[330.0, 335.0],
              [45.0,  40.0],
              [20.0,  15.0]])
```

```
a1 = a[:, 0]
a2 = a[:, 1]
b1 = b[:, 0]
b2 = b[:, 1]
z = np.array([5.0, 0.0, -2.0])


print(f"Processing {name}")
print("Reading files & initialising arrays")
strand_a, strand_b, midpoints = caxislib.read(name, num_bp, num_steps)

print("Calculating first-order helical axis")
helix_axis = caxislib.helix_axis(num_bp, num_steps, midpoints, strand_a)

print("Calculating twist")
twist = caxislib.full_twist(name, num_bp, num_steps, strand_a, strand_b,
                            helix_axis)

print("Calculating helical axis")
caxis = caxislib.caxis(name, num_bp, num_steps, midpoints, twist)

print("Calculating register angles")
sinreg = caxislib.sinreg(name, num_bp, num_steps, midpoints, caxis)

print("Writing output .xyz and .3col files")
caxislib.make_files(name, num_bp, num_steps, midpoints, caxis)

print("Reading coordinates")
read_coords = writhe.read_3col(name + '/C1.3col', num_bp, num_steps)

print("Calculating writhe")
wr = writhe.writhe(read_coords, 2, len(read_coords[0]))
full_writhe = writhe.main(name, num_bp, num_steps)

# Linear

print("Reading files & initialising arrays as if linear")
linear_strand_a, linear_strand_b, linear_midpoints = caxislib.read(name,
                                                                   num_bp,
                                                                   num_steps)


print("Calculating linear first-order helical axis")
linear_axis = caxislib.helix_axis(num_bp, num_steps, linear_midpoints,
                                  linear_strand_a, linear=True)

print("Calculating linear twist")
linear_twist = caxislib.full_twist(name, num_bp, num_steps, linear_strand_a,
                                   linear_strand_b, linear_axis, linear=True,
                                   write=False)

print("Calculating linear helical axis")
linear_caxis = caxislib.caxis(name, num_bp, num_steps, linear_midpoints,
                              linear_twist, linear=True)

tests = {
    "cross\t": [sum(sum(caxislib.cross(a, b))), -8850],
```

```
    "dot\t": [sum(caxislib.dot(a, b)), 223450],
    "norm 1\t": [sum(caxislib.norm(a)), 666.33216833189],
    "norm 2\t": [sum(caxislib.norm(b)), 671.36690822942],
    "rotate_to_x": [np.linalg.norm(caxislib.rotate_to_x(a1)), 1.7320508075689],
    "rotate_to_z": [np.linalg.norm(caxislib.rotate_to_z(a1)), 1.7320508075689],
    "twist 1\t": [caxislib.twist(a1, a2, b1, b2, z), 71.997556736987],
    "twist 2\t": [caxislib.twist(a1, b1, a2, b2, z), -15.06995574963],
    "twist 90deg": [caxislib.twist(np.array([0, 0, 0]),
                                   np.array([0, 1, 0]),
                                   np.array([0, 0, 0]),
                                   np.array([1, 0, 0]),
                                   np.array([0, 0, 1])), -90],
    "read strand_a": [sum(sum(sum(strand_a))), 1057248.34],
    "read strand_b": [sum(sum(sum(strand_b))), 1057277.65],
    "read midpoints": [sum(sum(sum(midpoints))), 1057262.995],
    "helix_axis": [sum(sum(sum(helix_axis))), 1057262.995],
    "helix_axis lin": [sum(sum(linear_axis[:, :, 150])),
                       sum(sum(helix_axis[:, :, 150]))],
    "full_twist": [sum(sum(twist)), 83325.202562320],
    "full_twist lin": [sum(linear_twist[:, 150]), sum(twist[:, 150])],
    "caxis\t": [sum(sum(sum(caxis))), 1057251.5419271],
    "caxis lin": [sum(sum(linear_caxis[:, :, 150])),
                  sum(sum(caxis[:, :, 150]))],
    "sinreg\t": [sum(sum(sinreg)), 46.523100971271],
    "read_3col": [sum(sum(sum(read_coords))), sum(sum(sum(caxis)))],
    "writhe.writhe": [wr, -1.013515045594],
    "writhe.main": [sum(sum(full_writhe)), 28.1093914578840],
    "C.3col\t": [filecmp.cmp(f'{name}/C.3col',
                             f'{name}/C.3col.original'), True],
    "C.xyz\t": [filecmp.cmp(f'{name}/C.xyz',
                            f'{name}/C.xyz.original'), True],
    "C1.3col\t": [filecmp.cmp(f'{name}/C1.3col',
                              f'{name}/C1.3col.original'), True],
    "C1.xyz\t": [filecmp.cmp(f'{name}/C1.xyz',
                             f'{name}/C1.xyz.original'), True],
    "tw.ser\t": [filecmp.cmp(f'{name}/tw.ser',
                             f'{name}/tw.ser.original'), True],
    "sinreg.ser": [filecmp.cmp(f'{name}/sinreg.ser',
                               f'{name}/sinreg.ser.original'), True],
    "writhe.ser": [filecmp.cmp(f'{name}/writhe.ser',
                               f'{name}/writhe.ser.original'), True],
}

pass_text = colored('[PASS]', 'green')
fail_text = colored('[FAIL]', 'red')

failures = []

print()
for test, assertion in tests.items():
    print(test, end="\t")
    try:
        if abs(assertion[1] - assertion[0]) < tolerance:
            print(pass_text)
        else:
            failures.append(test)
            print(fail_text)
```

```
            print(f"\tExpected\t{assertion[1]}")
            print(f"\tGot\t\t{assertion[0]}")
    except Exception as e:
        failures.append(test)
        print(fail_text)
        print("\tThe following exception was raised:")
        print(f"\t{e}")

print()
if failures:
    print(f"Total failures:\t{len(failures)}")
    for test in failures:
        print(f"\t{test}")
    sys.exit(len(failures))
else:
    print("All tests passed!")
```

Appendix 2

Analysis scripts

Appendix 2.1 Remove intramolecular hydrogen bonds

```
#!/bin/sh

grep '\bD' "$1" | grep -v '^D.* D.* D.*' > "${1}_cleaned"
```

Appendix 2.2 Calculate time-average number of hydrogen bonds

```
#!/usr/bin/env python3

"""
COMBINE H-BONDS
George Watson, Department of Physics, University of York, YO10 5DD
May 2018
https://www-users.york.ac.uk/~gw639

Usage:
    combine_hbonds2.py <filename>
"""

import sys
import operator
import re

if len(sys.argv) > 1:
    FILENAME = sys.argv[1]
else:
    FILENAME = input("Filename: ")

INPUT_FILE = open(FILENAME, 'r')
OUTPUT_FILE = open(FILENAME + "_total", 'w')

DATA = []
for line in INPUT_FILE.readlines():
    if not line.startswith('#'):
        cols = line.split()
        trimmed = [cols[0], cols[1], cols[4]]
        trimmed[0] = trimmed[0].split('@')[0]
```

```
        trimmed[1] = trimmed[1].split('@')[0]
        DATA.append(trimmed)

RESIDUES = {}
for bond in DATA:
    donor = bond[0]
    acceptor = bond[1]
    resname = ""

    donor_match = re.match(r'([a-ce-z][a-z]+)(_)([0-9]+)', donor, re.I)
    if donor_match:
        donor_items = donor_match.groups()
        donor_id = int(donor_items[2])
        resname = donor_items[0] + str(donor_id)

    acceptor_match = re.match(r'([a-ce-z][a-z]+)(_)([0-9]+)', acceptor, re.I)
    if acceptor_match:
        acceptor_items = acceptor_match.groups()
        acceptor_id = int(acceptor_items[2])
        resname = acceptor_items[0] + str(acceptor_id)

    if resname in RESIDUES:
        RESIDUES[resname] = RESIDUES[resname] + float(bond[2])
    else:
        RESIDUES[resname] = float(bond[2])

SORTED_RESIDUES = sorted(RESIDUES.items(),
                         key=operator.itemgetter(1),
                         reverse=True)

for key, value in SORTED_RESIDUES:
    OUTPUT_FILE.write(key + "\t" + str(value) + "\n")
```

Appendix 2.3 Find denatured regions of DNA

```
#!/usr/bin/env python3

"""
Script to find denatured regions, kinks, and flipped bases in DNA

Arguments (* = required):
    *   -f  --file  The file (Canal output) to process, without e.g. _twist.ser
        -n  --num   Number of base pairs (default: 336)
        -s  --start First column containing useful data, 0-indexed (default: 1)
        -l  --line  First line containing useful data, 0-indexed (default: 0)

(c) 2017 George Watson, University of York
"""

import csv
import argparse
import numpy as np
import matplotlib.pyplot as plt


def check_array(array, upper, lower):
    """
    Find denatured regions in a specified array

    Arguments (* = required):
        *   array   The array to process
        *   upper   The upper threshold (typically mean + N*stdev)
        *   lower   The lower threshold (typically mean - N*stdev)

    Returns an array in the same format as the input where
        0 => bp was not denatured at this time step
        >0 => number of standard deviations above/below limit
    """
    out_array = np.zeros(shape=array.shape)
    for bp in range(0, array.shape[1]):
        # Estimate the standard deviation
        # Start with the assumption that this bp is not denatured
        denatured = False
        # Loop over all time steps
        for t in range(1, array.shape[0]-1):
            # Is the base pair denatured at the next time step?
            if (array[t+1][bp] > upper) or (array[t+1][bp] < lower):
                # Is it also denatured at at this time step?
                # If so, in which direction?
                if (array[t][bp] > upper or array[t][bp] < lower):
                    # Mark this step as denatured in the array
                    out_array[t][bp] = 1

                    # Did we already know this bp was denatured?
                    if not denatured:
                        # We now know this bp is denatured
                        denatured = True
            # If the base pair isn't denatured now
            else:
```

```
                # Was it denatured at the previous step?
                if denatured:
                    # Is it also not denatured at the next step?
                    if (array[t+1][bp] < upper) and (array[t+1][bp] > lower):
                        # Mark the base pair as normal again
                        denatured = False

    # Return the list
    return out_array


def all_diffs(array, mean, stdev):
    """
    Find the number of standard deviations away from the mean for every entry

    Arguments (* = required):
        *   array   The array to process
        *   mean    The mean of the quantity
        *   stdev   The standard deviation of the quantity

    Returns an array in the same format as the input where
        0 => bp was not denatured at this time step
        >0 => number of standard deviations above/below limit
    """
    out_array = np.zeros(shape=array.shape)

    for bp in range(0, array.shape[1]):
        # Loop over all time steps
        for t in range(1, array.shape[0]-1):
            out_array[t][bp] = abs(array[t][bp] - mean) / stdev

    # Return the list
    return out_array


# Set up CLI argument parser
PARSER = argparse.ArgumentParser(
    description="Script to find denatured DNA regions in Canal output")

# REQUIRED ARGUMENTS
# -f    --file  Filename to process
PARSER.add_argument(
    "-f", "--file", help="Filename to process", required=True)

# OPTIONAL ARGUMENTS
# -n    --num   Number of entries per line
PARSER.add_argument(
    "-n", "--num", help="Number of entries per line", required=False,
    default=336)
# -s    --start First useful column, 0-indexed
PARSER.add_argument(
    "-s", "--start", help="First column containing values; zero-indexed",
    required=False, default=1)
# -l    --line  First useful line, 0-indexed
PARSER.add_argument(
    "-l", "--line", help="First line containing values; zero-indexed",
    required=False, default=0)
```

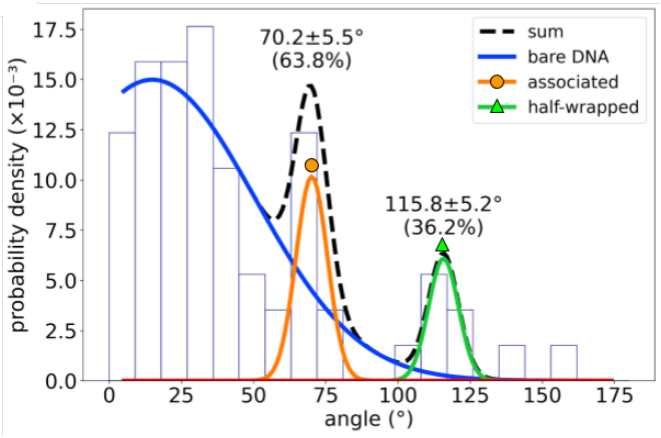
```
# -t    --threshold     Num. of sd. required to consider a bp denatured
PARSER.add_argument(
    "-t", "--threshold", help="Num. sd. to consider bp denatured",
    required=False, default=3)

# Parse arguments
ARGUMENT = PARSER.parse_args()
# Turn arguments into easier-to-use variables
FILENAME = ARGUMENT.file
N = int(ARGUMENT.num)
START = int(ARGUMENT.start)
LINE = int(ARGUMENT.line)
THRESHOLD = int(ARGUMENT.threshold)

# Read files into 2D arrays
TWISTS = np.loadtxt(
    FILENAME+"_twist.ser", usecols=tuple(range(START, START+N-1)),
    skiprows=LINE)
ROLLS = np.loadtxt(
    FILENAME+"_roll.ser", usecols=tuple(range(START, START+N-1)),
    skiprows=LINE)
# The last column of propel is ignored here to make the arrays the same size
PROPELS = np.loadtxt(
    FILENAME+"_propel.ser", usecols=tuple(range(START, START+N-1)),
    skiprows=LINE)

# Calculate means of arrays
MEAN_TWIST = np.mean(TWISTS)
MEAN_ROLL = np.mean(ROLLS)
MEAN_PROPEL = np.mean(PROPELS)

# Calculate standard deviations of arrays
SD_TWIST = np.std(TWISTS)
SD_ROLL = np.std(ROLLS)
SD_PROPEL = np.std(PROPELS)

# Find regions of unusual twist
print("Processing twist...")
TWIST_OUT = check_array(
    TWISTS, MEAN_TWIST + THRESHOLD*SD_TWIST,
    MEAN_TWIST - THRESHOLD*SD_TWIST)
TWIST_FULL = all_diffs(TWISTS, MEAN_TWIST, SD_TWIST)
print("Done")
print("Processing roll...")
ROLL_OUT = check_array(
    ROLLS, MEAN_ROLL + THRESHOLD*SD_ROLL, MEAN_ROLL - THRESHOLD*SD_ROLL)
ROLL_FULL = all_diffs(ROLLS, MEAN_ROLL, SD_ROLL)
print("Done")
print("Processing propeller twist...")
PROPEL_OUT = check_array(
    PROPELS, MEAN_PROPEL + THRESHOLD*SD_PROPEL,
    MEAN_PROPEL - THRESHOLD*SD_PROPEL)
PROPEL_FULL = all_diffs(PROPELS, MEAN_PROPEL, SD_PROPEL)
print("Done")

# Combine results, weighted by how many quantities appear to be denatured
DENATURED_ARRAY = TWIST_OUT + ROLL_OUT + PROPEL_OUT
```

```
FULL_ARRAY = (TWIST_FULL + ROLL_FULL + PROPEL_FULL) / 3

# Make a histogram
HIST_BARS = []
for i in range(0, DENATURED_ARRAY.shape[1]):
    HIST_BARS.append([i, np.count_nonzero(DENATURED_ARRAY[:, i])])

# Write histogram
with open("denatured_histogram.dat", "w") as outfile:
    WRITER = csv.writer(outfile, delimiter=" ")
    for row in HIST_BARS:
        WRITER.writerow(row)

# Count denatured bp-steps
DENATURED_BP_STEPS = np.count_nonzero(DENATURED_ARRAY)
print("\nMean num. denatured bp:\t" +
      str(DENATURED_BP_STEPS / DENATURED_ARRAY.shape[0]))

# Plot results
# plt.figure(1, figsize=(9, 6))
plt.rc('pgf', rcfonts=True)
plt.figure(1, figsize=(1.9, 1.3))
plt.imshow(
    FULL_ARRAY, cmap='magma', interpolation='none', aspect='auto',
    origin='lower')
plt.xlabel("Base pair")
plt.yticks([])
plt.savefig(f"{FILENAME}_denatured_heatmap.pgf", bbox_inches='tight')
```

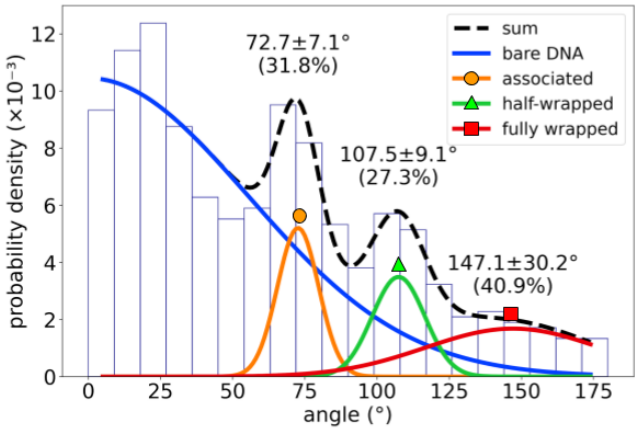
Appendix 3

Experimental data

All results in this chapter were obtained by Samuel Yoshua.



(a) 0λ361



(b) 1λ302

Figure A1: The distribution of bend angles measured using AFM is well fitted by two Gaussian distributions in the absence of an IHF binding site once the distribution for bare DNA is removed (a), and by three Gaussian distributions when a binding site is present (b).

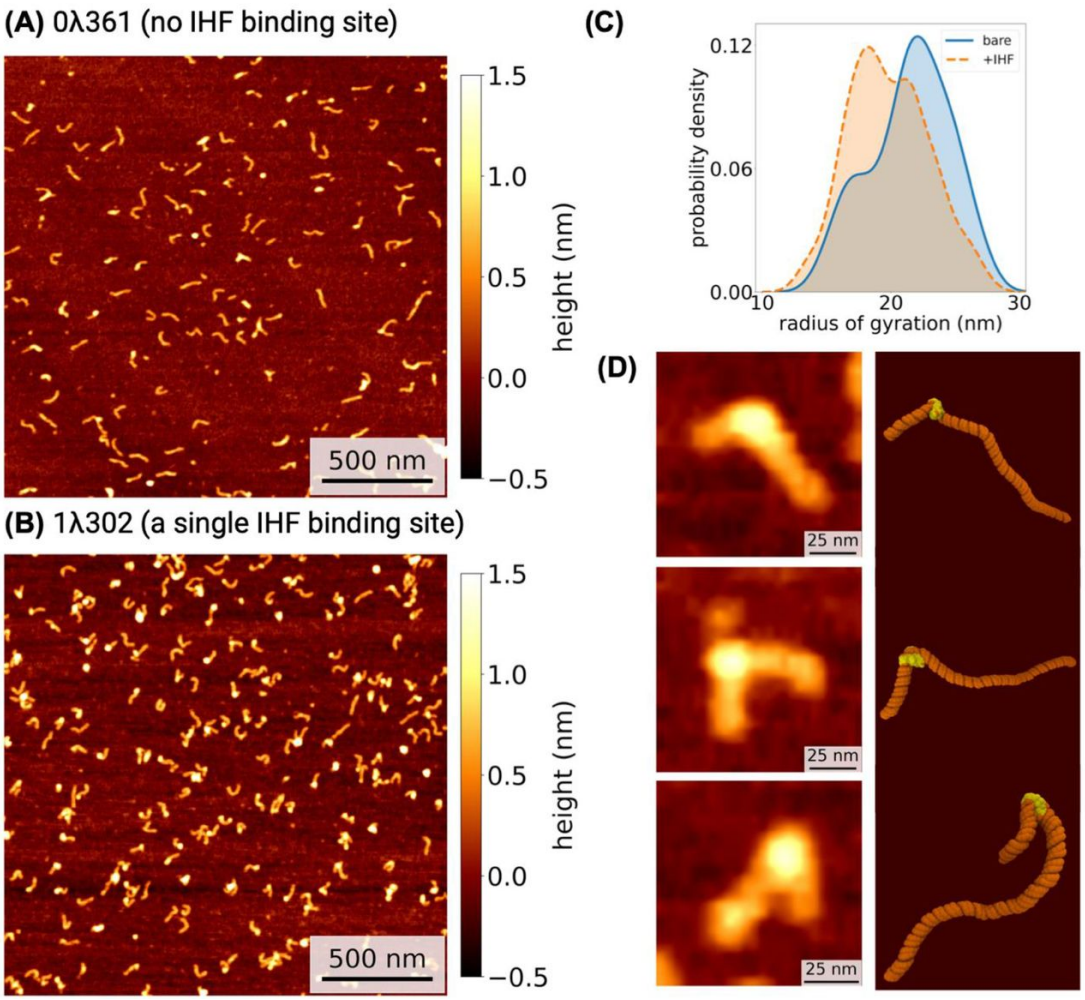


Figure A2: IHF is observed to bind minimally in the absence of a specific binding site (A), but much more binding is observed when a binding site is present (B). A corresponding reduction in the radius of gyration is observed (C). The three states observed in the AFM data (D, left) correspond to the three states obtained through MD (D, right).

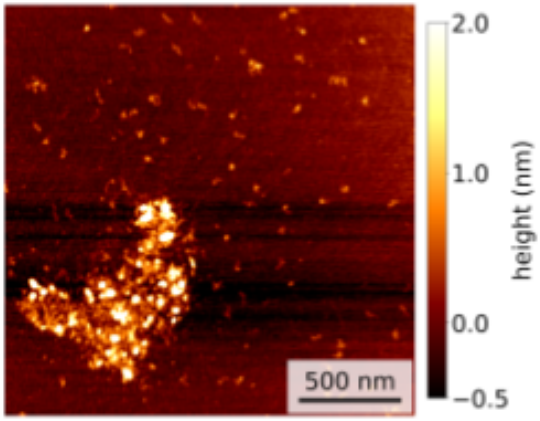


Figure A3: Large aggregates were observed in AFM images of a DNA construct with three IHF binding sites, which can be explained by the bridging behaviour described in the main text.

Appendix 4

DNA sequences

Appendix 4.1 1λ302

This sequence is 302 bp long and was extracted from the *xis* gene of bacteriophage λ. The IHF consensus sequence occurs once in this sequence (indicated in bold and underlined). This sequence was used for most simulations and AFM imaging of linear DNA.

CAAGACACCGGATCTGCACATTGATAACGCCCAATCTTTTTGCTCAGACTC**TAA**CTCA**TTGATA**CTCATT
TATAAACTCCTTGCAATGTATGTCGTTTCAGCTAAACGGTATCAGCAATGTTTATGTAAAGAAACAGTAA
GATAATACTCAACCCGATGTTTGAGTACGGTCATCATCTGACACTACAGACTCTGGCATCGCTGTGAAGA
CGACGCGAAATTCAGCATTTTCACAAGCGTTATCTTTTACAAAACCGATCTCACTCTCCTTTGATGCGAA
TGCCAGCGTCAGACATCATATG

Appendix 4.2 1λ61

This sequence is 61 bp long and was extracted from the 1λ302 sequence (appendix 4.1). The IHF consensus sequence occurs once in this sequence (indicated in bold and underlined). This sequence was used for umbrella sampling simulations.

TAGAAAAACGAGTCTGAG**ATT**GAGT**AACTAT**GAGTAAATATTTGAGGAACGTTACATACAC

Appendix 4.3 3λ343

This sequence is 343 bp long and was extracted from the *xis* gene of bacteriophage λ. The IHF consensus sequence occurs three times in this sequence (indicated in bold and underlined). This sequence was used for AFM imaging of large DNA–IHF clusters, and was also simulated.

CTTTGTGCTTCTCTGGAGTGCGACAGGTTTGATGACAAAAAATTAGCGCAAGAAGACAAAAATCACCTTG
CGCTAATGCTCTGTTACAGGTCACTAATACCATC**TAA**GTAG**TTGATT**CATAGTGACTGCATATGTTGTGT
TTTACAGTATTATGTAGTCTGTTTTTTATGCAAAATCTAATT**TAA**TATA**TTGATA**TTTATATCATTTTAC
GTTTCTCGTTCAGCTTTTTTATACTAAGTTGGCATTATAAAAAAGCATTGCT**TATCAA**TTTG**TTG**CAACG
AACAGGTCACTATCAGTCAAAATAAAATCATTATTTGATTTCAATTTTGTCCCACTCCCTGCC

Appendix 4.4 336 bp minicircle (1 binding site)

This construct is a 336 bp piece of covalently closed circular DNA. The IHF consensus sequence occurs once in this construct (indicated in bold and underlined). This sequence was used for most simulations of DNA minicircles.

TTGATATTTATATCATTTTACGTTTCTCGTTCAGCTTTTTTATACTAACTTGAGCGATATCACCGCAAGG
GATAAATATCTAACACCGTGCGTGTTGACTATTTTACCTCTGGCGGTGATACTTCCCGGAAAACGCGGTG
GAATATTTCGTTTCCTACTACGACTACTATCAGCCGGAAGCCTATGTACCGAGTTCCGACACTTTCATTG
AGAAAGATGCCTCAGCTATCACCGCCAGTGGTATTTATGTCAACACCGCCAGAGATAATTTATCACCGCA
GATGGTTTTTACAGTATTATGTAGTCTGTTTTTTATGCAAAATCTAATT**TAA**TATA

Appendix 4.5 336 bp minicircle (2 binding sites)

This construct is a 336 bp piece of covalently closed circular DNA. The IHF consensus sequence occurs twice in this construct (indicated in bold and underlined). This sequence was used for simulations of DNA minicircles with two bound copies of IHF, and for AFM imaging of DNA minicircles.

TTTATACTAACTTGAGCGAAACGGGAAGGGTTTTCACCGATATCACCGAAACGCGCGAGGCAGCTGTATG
GCATGAAAGAGTTCTTCCCGGAAAACGCGGTGGAATATTTCGTTTCCTACTACGACTACTATCAGCCGGA
AGCCTATGTACCGAGTTCCGACACTTTCATTGAGAAAGATGCCTCAGCTCTGTTACAGGTCACTAATACC
ATC**TAA**GTAG**TTGATT**CATAGTGACTGCATATGTTGTGTTTTACAGTATTATGTAGTCTGTTTTTTATGC
AAAATCTAATT**TAA**TATA**TTGATA**TTTATATCATTTTACGTTTCTCGTTCAGCTTT

Abbreviations

A	adenine	16
AFM	atomic force microscopy	53
AMBER	Assisted Model Building with Energy Refinement	34
API	application programming interface	51
bp	base pair	17
C	cytosine	16
ccDNA	covalently closed circular DNA	20
CHARMM	Chemistry at Harvard Macromolecular Mechanics	39
CPU	central processing unit	51
CUDA	Compute Unified Device Architecture	51
DNA	deoxyribonucleic acid	16
E. coli	*Escherichia coli*	18
eDNA	extracellular DNA	27
et al.	*et alii* or *et aliæ* (Latin: "and others")	133
G	guanine	16
GB-HCT	Generalised Born (Hawkins, Cramer, Truhlar)	43
GB-neck	Generalised Born with neck corrections	44
GB-OBC	Generalised Born (Onufriev, Bashford, Case)	44
GPGPU	general-purpose programming on GPUs	51
GPU	graphics processing unit	51
H-NS	heat-stable nucleoid-structuring protein	29
HU	heat-unstable protein, *or* histone-like protein from *E.coli* strain U93	28
IHF	integration host factor	25
Lk	linking number	20

MD	molecular dynamics	33
NAB	Nucleic Acid Builder	34
NAP	nucleoid-associated protein	25
NMR	nuclear magnetic resonance	34
PDB	Protein Data Bank	27
pmemd	particle-mesh Ewald MD	51
PMF	potential of mean force	59
pp	percentage points	75
RMSD	root-mean-square deviation	34
RNA	ribonucleic acid	16
T	thymine	16
TIP3P	transferable intermolecular potential with 3 points	41
Tw	twist	21
U	uracil	16
WHAM	weighted histogram analysis method	60
Wr	writhe	21

Journal title abbreviations used in references (starting on page 133) are in accordance with the International Organization for Standardization's recommendations in ISO 4:1997.*

* ISO 4:1997 "Rules for the abbreviation of title words and titles of publications" (Geneva: International Organization for Standardization) URL `https://www.iso.org/standard/3569.html`

References

[1] Crick F H C 1958 "On protein synthesis" *Symp. Soc. Exp. Biol.* vol 12 (Society for Experimental Biology) pp 138–63

[2] Yakovchuk P, Protozanova E and Frank-Kamenetskii M D 2006 "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix" *Nucleic Acids Res.* **34** 564–574 doi:10.1093/nar/gkj454

[3] Svoboda P and Di Cara A 2006 "Hairpin RNA: A secondary structure of primary importance" *Cell. Mol. Life Sci.* **63** 901–908 doi:10.1007/s00018-005-5558-5

[4] Watson J D and Crick F H C 1953 "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid" *Nature* **171** 737–738 doi:10.1038/171737a0

[5] Pray L A 2008 "DNA replication and causes of mutation" *Nat. Educ.* **1** 214

[6] Clancy S 2008 "DNA transcription" *Nat. Educ.* **1** 41

[7] Clancy S and Brown W 2008 "Translation: DNA to mRNA to protein" *Nat. Educ.* **1** 101

[8] Lander E S *et al.* 2001 "Initial sequencing and analysis of the human genome" *Nature* **409** 860–921 doi:10.1038/35057062

[9] Hong E L *et al.* 2016 "Principles of metadata organization at the ENCODE data coordination center" *Database* **2016** doi:10.1093/database/baw001

[10] Pray L A 2008 "What is a gene? Colinearity and transcription units" *Nat. Educ.* **1** 97

[11] Jacob F and Monod J 1961 "Genetic regulatory mechanisms in the synthesis of proteins" *J. Mol. Biol.* **3** 318–356 doi:10.1016/S0022-2836(61)80072-7

[12] Normanno D, Vanzi F and Pavone F S 2008 "Single-molecule manipulation reveals supercoiling-dependent modulation of *lac* repressor-mediated DNA looping" *Nucleic Acids Res.* **36** 2505–2513 doi:10.1093/nar/gkn071

[13] Rudd K E and Menzel R 1987 "*his* operons of *Escherichia coli* and *Salmonella typhimurium* are regulated by DNA supercoiling" *Proc. Natl. Acad. Sci. USA* **84** 517–521 doi:10.1073/pnas.84.2.517

[14] Fulcrand G, Dages S, Zhi X, Chapagain P, Gerstman B S, Dunlap D and Leng F 2016 "DNA supercoiling, a critical signal regulating the basal expression of the *lac* operon in *Escherichia coli*" *Sci. Rep.* **6** 19243

[15] Dickerson R E 1989 "Definitions and nomenclature of nucleic acid structure components" *Nucleic Acids Res.* **17** 1797–1803 doi:10.1093/nar/17.5.1797

[16] Dickerson R E "DNA structure from A to Z" *Methods Enzymol.* **211** 67–111 doi:10.1016/0076-6879(92)11007-6

[17] Wing R, Drew H, Takano T, Broka C, Tanaka S, Itakura K and Dickerson R E 1980 "Crystal structure analysis of a complete turn of B-DNA" *Nature* **287** 755–758 doi:10.1038/287755a0

[18] Noy A, Sutthibutpong T and Harris S A 2016 "Protein/DNA interactions in complex DNA topologies: Expect the unexpected" *Biophys. Rev.* **8** 145–155 doi:10.1007/s12551-016-0241-7

[19] Burns H and Minchin S 1994 "Thermal energy requirement for strand separation during transcription initiation: The effect of supercoiling and extended protein DNA contacts" *Nucleic Acids Res.* **22** 3840–3845 doi:10.1093/nar/22.19.3840

[20] Champoux J J 2001 "DNA topoisomerases: Structure, function, and mechanism" *Annu. Rev. Biochem.* **70** 369–413 doi:10.1146/annurev.biochem.70.1.369

[21] Madigan M T and Orent A 1999 "Thermophilic and halophilic extremophiles" *Curr. Opin. Microbiol.* **2** 265–269 doi:10.1016/S1369-5274(99)80046-0

[22] Baranello L, Levens D, Gupta A and Kouzine F 2012 "The importance of being supercoiled: How DNA mechanics regulate dynamic processes" *Biochim. Biophys. Acta* Gene Regul. Mech. **1819** 632–638 doi:10.1016/j.bbagrm.2011.12.007

[23] Norregaard K, Andersson M, Sneppen K, Nielsen P E, Brown S and Oddershede L B 2013 "DNA supercoiling enhances cooperativity and efficiency of an epigenetic switch" *Proc. Natl. Acad. Sci. USA* **110** 17386–17391 doi:10.1073/pnas.1215907110

[24] D'Annessa I, Coletta A, Sutthibutpong T, Mitchell J, Chillemi G, Harris S and Desideri A 2014 "Simulations of DNA topoisomerase 1B bound to supercoiled DNA reveal changes in the flexibility pattern of the enzyme and a secondary protein–DNA binding site" *Nucleic Acids Res.* **42** 9304–9312 doi:10.1093/nar/gku654

[25] Noy A, Maxwell A and Harris S A 2017 "Interference between triplex and protein binding to distal sites on supercoiled DNA" *Biophys. J.* **112** 523–531 doi:10.1016/j.bpj.2016.12.034

[26] Kim S H, Ganji M, Kim E, van der Torre J, Abbondanzieri E and Dekker C 2018 "DNA sequence encodes the position of DNA supercoils" *eLife* **7** e36557 doi:10.7554/eLife.36557

[27] Catanese D J, Fogg J M, Schrock D E, Gilbert B E and Zechiedrich L 2012 "Supercoiled minivector DNA resists shear forces associated with gene therapy delivery" *Gene Ther.* **19** 94–100 doi:10.1038/gt.2011.77

[28] Gauß C F 1874 *Werke* vol 6 (Göttingen: Königliche Gesellschaft der Wissenschaften) ISBN 978-1-139-05827-8 (German: *Works*)

[29] Swigon D 2009 "The mathematics of DNA structure, mechanics, and dynamics" *Mathematics of DNA Structure, Function and Interactions* ed Benham C J, Harvey S, Olson W, Sumners D L and Swigon D (London: Springer) chap 14, pp 293–320 ISBN 978-1-4419-0669-4

[30] Călugăreanu G 1961 "Sur les classes d'isotopie des noeuds tridimensionnels et leurs invariants" *Czech. Math. J.* **11** 588–625 (French: "On the isotopy classes of three-dimensional nodes and their invariants")

[31] White J H 1969 "Self-linking and the Gauss integral in higher dimensions" *Am. J. Math.* **91** 693–728 doi:10.2307/2373348

[32] Fuller F B 1971 "The writhing number of a space curve" *Proc. Natl. Acad. Sci. USA* **68** 815–819 doi:10.1073/pnas.68.4.815

[33] Irobalieva R N *et al.* 2015 "Structural diversity of supercoiled DNA" *Nat. Commun.* **6** 1–11 doi:10.1038/ncomms9440

[34] Leng F, Chen B and Dunlap D D 2011 "Dividing a supercoiled DNA molecule into two independent topological domains" *Proc. Natl. Acad. Sci. USA* **108** 19973–19978 doi:10.1073/pnas.1109854108

[35] Blattner F R *et al.* 1997 "The complete genome sequence of *Escherichia coli* K-12" *Science* **277** 1453–1462 doi:10.1126/science.277.5331.1453

[36] Swinger K K and Rice P A 2004 "IHF and HU: Flexible architects of bent DNA" *Curr. Opin. Struct. Biol.* **14** 28–35 doi:10.1016/j.sbi.2003.12.003

[37] Yang C C and Nash H A 1989 "The interaction of *E. coli* IHF protein with its specific binding sites" *Cell* **57** 869–880 doi:10.1016/0092-8674(89)90801-5

[38] Arfin S M, Long A D, Ito E T, Tolleri L, Riehle M M, Paegle E S and Hatfield G W 2000 "Global gene expression profiling in *Escherichia coli* K12: The effects of integration host factor" *J. Biol. Chem.* **275** 29672–29684 doi:10.1074/jbc.M002247200

[39] Hwang D S and Kornberg A 1992 "Opening of the replication origin of *Escherichia coli* by DnaA protein with protein HU or IHF" *J. Biol. Chem.* **267** 23083–23086

[40] Kobryn K, Lavoie B D and Chaconas G 1999 "Supercoiling-dependent site-specific binding of HU to naked Mu DNA" *J. Mol. Biol.* **289** 777–784 doi:10.1006/jmbi.1999.2805

[41] Nuñez J K, Bai L, Harrington L B, Hinder T L and Doudna J A 2016 "CRISPR immunological memory requires a host factor for specificity" *Mol. Cell* **62** 824–833 doi:10.1016/j.molcel.2016.04.027

[42] Seong G H, Kobatake E, Miura K, Nakazawa A and Aizawa M 2002 "Direct atomic force microscopy visualization of integration host factor-induced DNA bending structure of the promoter regulatory region on the *Pseudomonas* TOL plasmid" *Biochem. Biophys. Res. Commun.* **291** 361–366 doi:10.1006/bbrc.2002.6443

[43] Jamal M, Ahmad W, Andleeb S, Jalil F, Imran M, Nawaz M A, Hussain T, Ali M, Rafiq M and Kamil M A 2018 "Bacterial biofilm and associated infections" *J. Chin. Med. Assoc.* **81** 7–11 doi:10.1016/j.jcma.2017.07.012

[44] Gustave J E, Jurcisek J A, McCoy K S, Goodman S D and Bakaletz L O 2013 "Targeting bacterial integration host factor to disrupt biofilms associated with cystic fibrosis" *J. Cyst. Fibros.* **12** 384–389 doi:10.1016/j.jcf.2012.10.011

[45] Novotny L A, Amer A O, Brockson M E, Goodman S D and Bakaletz L O 2013 "Structural stability of *Burkholderia cenocepacia* biofilms is reliant on eDNA structure and presence of a bacterial nucleic acid binding protein" *PLoS One* **8** e67629 doi:10.1371/journal.pone.0067629

[46] Rice P A, Yang S W, Mizuuchi K and Nash H A 1996 "Crystal structure of an IHF–DNA complex: A protein-induced DNA U-turn" *Cell* **87** 1295–1306 doi:10.1016/S0092-8674(00)81824-3

[47] Laxmikanthan G, Xu C, Brilot A F, Warren D, Steele L, Seah N, Tong W, Grigorieff N, Landy A and Van Duyne G D 2016 "Structure of a Holliday junction complex reveals mechanisms governing a highly regulated DNA transaction" *eLife* **5** e14313 doi:10.7554/eLife.14313

[48] Connolly M, Arra A, Zvoda V, Steinbach P J, Rice P A and Ansari A 2018 "Static kinks or flexible hinges: Multiple conformations of bent DNA bound to integration host factor revealed by fluorescence lifetime measurements" *J. Phys. Chem.* B **122** 11519–11534 doi:10.1021/acs.jpcb.8b07405

[49] Velmurugu Y, Vivas P, Connolly M, Kuznetsov S V, Rice P A and Ansari A 2018 "Two-step interrogation then recognition of DNA binding site by integration host factor: An architectural DNA-bending protein" *Nucleic Acids Res.* **46** 1741–1755 doi:10.1093/nar/gkx1215

[50] Khrapunov S, Brenowitz M, Rice P A and Catalano C E 2006 "Binding then bending: A mechanism for wrapping DNA" *Proc. Natl. Acad. Sci. USA* **103** 19217–19218 doi:10.1073/pnas.0609223103

[51] Kuznetsov S V, Sugimura S, Vivas P, Crothers D M and Ansari A 2006 "Direct observation of DNA bending/unbending kinetics in complex with DNA-bending protein IHF" *Proc. Natl. Acad. Sci. USA* **103** 18515–18520 doi:10.1073/pnas.0608394103

[52] Dixit S, Singh-Zocchi M, Hanne J and Zocchi G 2005 "Mechanics of binding of a single integration-host-factor protein to DNA" *Phys. Rev. Lett.* **94** 118101 doi:10.1103/PhysRevLett.94.118101

[53] Claret L and Rouviere-Yaniv J 1997 "Variation in HU composition during growth of *Escherichia coli*: The heterodimer is required for long term survival" *J. Mol. Biol.* **273** 93–104 doi:10.1006/jmbi.1997.1310

[54] Sanner M F, Olson A J and Spehner J C 1996 "Reduced surface: An efficient way to compute molecular surfaces" *Biopolymers* **38** 305–320 doi:10.1002/(SICI)1097-0282(199603)38:3<305::AID-BIP4>3.0.CO;2-Y

[55] Wojtuszewski K and Mukerji I 2003 "HU binding to bent DNA: A fluorescence resonance energy transfer and anisotropy study" *Biochemistry* **42** 3096–3104 doi:10.1021/bi0264014

[56] Swinger K K, Lemberg K M, Zhang Y and Rice P A 2003 "Flexible DNA bending in HU–DNA cocrystal structures" *EMBO J.* **22** 3749–3760 doi:10.1093/emboj/cdg351

[57] Hammel M, Amlanjyoti D, Reyes F E, Chen J H, Parpana R, Tang H Y H, Larabell C A, Tainer J A and Adhya S 2016 "HU multimerization shift controls nucleoid compaction" *Sci. Adv.* **2** e1600650 doi:10.1126/sciadv.1600650

[58] Aki T, Choy H E and Adhya S 1996 "Histone-like protein HU as a specific transcriptional regulator: Co-factor role in repression of *gal* transcription by GAL repressor" *Genes Cells* **1** 179–188 doi:10.1046/j.1365-2443.1996.d01-236.x

[59] Lia G, Bensimon D, Croquette V, Allemand J F, Dunlap D, Lewis D E, Adhya S and Finzi L 2003 "Supercoiling and denaturation in Gal repressor/heat unstable nucleoid protein (HU)-mediated DNA looping" *Proc. Natl. Acad. Sci. USA* **100** 11373–11377 doi:10.1073/pnas.2034851100

[60] Dorman C J 2004 "H-NS: A universal regulator for a dynamic genome" *Nat. Rev. Microbiol.* **2** 391–400 doi:10.1038/nrmicro883

[61] Deighan P, Beloin C and Dorman C J 2003 "Three-way interactions among the Sfh, StpA and H-NS nucleoid-structuring proteins of *Shigella flexneri* 2a strain 2457t" *Mol. Microbiol.* **48** 1401–1416 doi:10.1046/j.1365-2958.2003.03515.x

[62] Schnetz K and Wang J C 1996 "Silencing of the *Escherichia coli bgl* promoter: Effects of template supercoiling and cell extracts on promoter activity *in vitro*" *Nucleic Acids Res.* **24** 2422–2428 doi:10.1093/nar/24.12.2422

[63] Müller C M, Dobrindt U, Nagy G, Emödy L, Uhlin B E and Hacker J 2006 "Role of histone-like proteins H-NS and StpA in expression of virulence determinants of uropathogenic *Escherichia coli*" *J. Bacteriol.* **188** 5428–5438 doi:10.1128/JB.01956-05

[64] Cusick M E and Belfort M 1998 "Domain structure and RNA annealing activity of the *Escherichia coli* regulatory protein StpA" *Mol. Microbiol.* **28** 847–857 doi:10.1046/j.1365-2958.1998.00848.x

[65] Finkel S E and Johnson R C 1992 "The Fis protein: It's not just for DNA inversion anymore" *Mol. Microbiol.* **6** 3257–3265 doi:10.1111/j.1365-2958.1992.tb02193.x

[66] Kostrewa D, Granzin J, Koch C, Choe H W, Raghunathan S, Wolf W, Labahn J, Kahmann R and Saenger W 1991 "Three-dimensional structure of the *E. coli* DNA-binding protein FIS" *Nature* **349** 178–180 doi:10.1038/349178a0

[67] Travers A, Schneider R and Muskhelishvili G 2001 "DNA supercoiling and transcription in *Escherichia coli*: The FIS connection" *Biochimie* **83** 213–217 doi:10.1016/S0300-9084(00)01217-7

[68] Cameron A D, Stoebel D M and Dorman C J 2011 "DNA supercoiling is differentially regulated by environmental factors and FIS in *Escherichia coli* and *Salmonella enterica*" *Mol. Microbiol.* **80** 85–101 doi:10.1111/j.1365-2958.2011.07560.x

[69] Roe D R and Brooks B R 2020 "A protocol for preparing explicitly solvated systems for stable molecular dynamics simulations" *J. Chem. Phys.* **153** 054123 doi:10.1063/5.0013849

[70] Macke T J and Case D A 1998 "Modeling unusual nucleic acid structures" *Molecular Modeling of Nucleic Acids* (*ACS Symp. Series* vol 682) ed Leontis N B (San Francisco: ACS Publications) chap 24, pp 379–393 ISBN 978-0-8412-3541-0

[71] Case D A *et al.* 2014–2018 "AMBER" v14–18 URL http://ambermd.org/

[72] wwPDB consortium 2019 "Protein Data Bank: The single global archive for 3D macromolecular structure data" *Nucleic Acids Res.* **47** D520–D528 doi:10.1093/nar/gky949

[73] Schrödinger LLC 2015 "The PyMol molecular graphics system" v1.8 URL https://pymol.org

[74] Jones J E 1924 "On the determination of molecular fields. I. From the variation of the viscosity of a gas with temperature" *Proc. R. Soc. Lond.* A **106** 441–462 doi:10.1098/rspa.1924.0081

[75] Jones J E 1924 "On the determination of molecular fields. II. From the equation of state of a gas" *Proc. R. Soc. Lond.* A **106** 463–477 doi:10.1098/rspa.1924.0082

[76] Lennard-Jones J E 1931 "Cohesion" *Proc. Phys. Soc.* **43** 461–482 doi:10.1088/0959-5309/43/5/301

[77] Eisenschitz R and London F 1930 "Über das Verhältnis der van der Waalsschen Kräfte zu den homöopolaren Bindungskräften" *Z. Phys.* **60** 491–527 doi:10.1007/BF01341258 (German: "About the relationship between the van der Waals forces and the homœopolar binding forces")

[78] Son J H, Ahn H S and Cha J 2017 "Lennard-Jones potential field-based swarm systems for aggregation and obstacle avoidance" *Int. Conf. Control Autom. Sys.* (IEEE) pp 1068–1072 doi:10.23919/ICCAS.2017.8204374

[79] de Coulomb C A 1785 "Premier mémoire sur l'electricité et le magnétisme" *Histoire de l'Académie Royale des Sciences* (Paris: Académie Royale des Sciences) pp 569–577 (French: "First dissertation on electricity and magnetism")

[80] de Coulomb C A 1785 "Second mémoire sur l'electricité et le magnétisme" *Histoire de l'Académie Royale des Sciences* (Paris: Académie Royale des Sciences) pp 578–611 (French: "Second dissertation on electricity and magnetism")

[81] Hooke R 1678 *Lectures de Potentia Restitutiva, or of Spring: Explaining the Power of Springing Bodies* (London: Royal Society)

[82] Weiner S J, Kollman P A, Case D A, Singh U C, Ghio C, Alagona G, Profeta S and Weiner P 1984 "A new force field for molecular mechanical simulation of nucleic acids and proteins" *J. Am. Chem. Soc.* **106** 765–784 doi:10.1021/ja00315a051

[83] Cornell W D, Cieplak P, Bayly C I, Gould I R, Merz K M, Ferguson D M, Spellmeyer D C, Fox T, Caldwell J W and Kollman P A 1995 "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules" *J. Am. Chem. Soc.* **177** 5179–5197 doi:10.1021/ja00124a002

[84] Ponder J W and Case D A 2003 "Force fields for protein simulations" *Protein Simulations* (*Adv. Protein Chem.* vol 66) ed Daggett V (Amsterdam: Elsevier) chap 2, pp 27–85 ISBN 978-0-12-034266-2

[85] Cheatham III T E and Case D A 2013 "Twenty-five years of nucleic acid simulations" *Biopolymers* **99** 969–977 doi:10.1002/bip.22331

[86] Liu Y P, Kim K, Berne B J, Friesner R A and Rick S W 1998 "Constructing *ab initio* force fields for molecular dynamics simulations" *J. Chem. Phys.* **108** 4739–4755 doi:10.1063/1.475886

[87] Khoury G A, Thompson J P, Smadbeck J, Kieslich C A and Floudas Ch A 2013 "Forcefield_PTM: *Ab initio* charge and AMBER forcefield parameters for frequently occurring post-translational modifications" *J. Chem. Theory Comput.* **9** 5653–5674 doi:10.1021/ct400556v

[88] Maier J A, Martinez C, Kasavajhala K, Wickstrom L, Hauser K E and Simmerling C 2015 "ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB" *J. Chem. Theory Comput.* **11** 3696–3713 doi:10.1021/acs.jctc.5b00255

[89] Ivani I *et al.* 2016 "Parmbsc1: A refined force field for DNA simulations" *Nat. Methods* **13** 55 doi:10.1038/nmeth.3658

[90] Brooks B R *et al.* 2006 "CHARMM: The biomolecular simulation program" *J. Comput. Chem.* **30** 1545–1615 doi:10.1002/jcc.21287

[91] MacKerell Jr A D and Nilsson L 2008 "Molecular dynamics simulations of nucleic acid–protein complexes" *Curr. Opin. Struct. Biol.* **18** 194–199 doi:10.1016/j.sbi.2007.12.012

[92] Soares T A, Hünenberger P H, Kastenholz M A, Kräutler V, Lenz T, Lins R D, Oostenbrink C and van Gunsteren W F 2005 "An improved nucleic acid parameter set for the GROMOS force field" *J. Comput. Chem.* **26** 725–737 doi:10.1002/jcc.20193

[93] Kaminski G A, Friesner R A, Tirado-Rives J and Jorgensen W L 2001 "Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides" *J. Phys. Chem.* B **105** 6474–6487 doi:10.1021/jp003919d

[94] Halgren T A 1996 "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94" *J. Comput. Chem.* **17** 490–519 doi:10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P

[95] Ewig C S *et al.* 2001 "Derivation of class II force fields. VIII. Derivation of a general quantum mechanical force field for organic compounds" *J. Comput. Chem.* **22** 1782–1800 doi:10.1002/jcc.1131

[96] Johnson N W 1966 "Convex polyhedra with regular faces" *Can. J. Math.* **18** 169–200 doi:10.4153/CJM-1966-021-8

[97] Wadell H 1935 "Volume, shape, and roundness of quartz particles" *J. Geol.* **43** 250–280 doi:10.1086/624298

[98] Alexandrov A D 2005 "Parallelohedra" *Convex Polyhedra* (Berlin: Springer-Verlag) chap 8.1, pp 349–359 ISBN 978-3-540-23158-5

[99] Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 "Comparison of simple potential functions for simulating liquid water" *J. Chem. Phys.* **79** 926–935 doi:10.1063/1.445869

[100] Nada H and van der Eerden J P J M 2003 "An intermolecular potential model for the simulation of ice and water near the melting point: A six-site model of H_2O" *J. Chem. Phys.* **118** 7401–7413 doi:10.1063/1.1562610

[101] Still W C, Tempczyk A, Hawley R C and Hendrickson T 1990 "Semianalytical treatment of solvation for molecular mechanics and dynamics" *J. Am. Chem. Soc.* **112** 6127–6129 doi:10.1021/ja00172a038

[102] Onufriev A V and Case D A 2019 "Generalized Born implicit solvent models for biomolecules" *Annu. Rev. Biophys.* **48** 275–296 doi:10.1146/annurev-biophys-052118-115325

[103] Lin Y L, Aleksandrov A, Simonson T and Roux B 2014 "An overview of electrostatic free energy computations for solutions and proteins" *J. Chem. Theory Comput.* **10** 2690–2709 doi:10.1021/ct500195p

[104] Poisson S D 1823 "Mémoire sur la théorie du magnétisme en mouvement" *Mémoires de l'Académie Royale des Sciences* (Paris: Académie Royale des Sciences) pp 441–570 (French: "Thesis on the theory of magnetism in motion")

[105] Fogolari F, Brigo A and Molinari H 2002 "The Poisson–Boltzmann equation for biomolecular electrostatics: A tool for structural biology" *J. Mol. Recognit.* **15** 377–392

[106] Höfinger S 2005 "Solving the Poisson–Boltzmann equation with the specialized computer chip MD-GRAPE-2" *J. Comput. Chem.* **26** 1148–1154 doi:10.1002/jcc.20250

[107] Hawkins G D, Cramer C J and Truhlar D G 1995 "Pairwise solute descreening of solute charges from a dielectric medium" *Chem. Phys. Lett.* **246** 122–129 doi:10.1016/0009-2614(95)01082-K

[108] Onufriev A, Bashford D and Case D A 2004 "Exploring protein native states and large-scale conformational changes with a modified generalized Born model" *Proteins* **55** 383–394 doi:10.1002/prot.20033

[109] Mongan J, Simmerling C, McCammon J A, Case D A and Onufriev A 2007 "Generalized Born model with a simple, robust molecular volume correction" *J. Chem. Theory Comput.* **3** 156–169 doi:10.1021/ct600085e

[110] Nguyen H, Roe D R and Simmerling C 2013 "Improved generalized Born solvent model parameters for protein simulations" *J. Chem. Theory Comput.* **9** 2020–2034 doi:10.1021/ct3010485

[111] Nguyen H, Pérez A, Bermeo S and Simmerling C 2015 "Refinement of generalized Born implicit solvation parameters for nucleic acids and their complexes with proteins" *J. Chem. Theory Comput.* **11** 3714–3728 doi:10.1021/acs.jctc.5b00271

[112] Debye P and Hückel E 1923 "Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen" *Phys. Z.* **24** 185–206 (German: "The theory of electrolytes. I. Lowering of freezing point and related phenomena")

[113] Birdsall C K and Langdon A B 2005 "Introduction to the numerical methods used" *Plasma Physics via Computer Simulation* (Oxford: Taylor & Francis) chap 4, pp 55–80 ISBN 978-0-7503-1025-3

[114] Ryckaert J P, Ciccotti G and Berendsen H J C 1977 "Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes" *J. Chem. Phys.* **23** 327–341 doi:10.1016/0021-9991(77)90098-5

[115] Berendsen H J C, van Postma J P M, van Gunsteren W F, DiNola A R H J and Haak J R 1984 "Molecular dynamics with coupling to an external bath" *J. Chem. Phys.* **81** 3684–3690 doi:10.1063/1.448118

[116] Langevin P 1908 "Sur la théorie du mouvement brownien" *C. R. Acad. Sci. Paris* **146** 530–533 (French "On the theory of Brownian motion")

[117] Plimpton S 1995 "Fast parallel algorithms for short-range molecular dynamics" *J. Comput. Phys.* **117** 1–19 doi:10.1006/jcph.1995.1039

[118] Teleman O and Jönsson B 1986 "Vectorizing a general purpose molecular dynamics simulation program" *J. Comput. Chem.* **7** 58–66 doi:10.1002/jcc.540070108

[119] Nickolls J, Buck I, Garland M and Skadron K 2008 "Scalable parallel programming with CUDA" *Queue* **6** 40–53 doi:10.1145/1365490.1365500

[120] Götz A W, Williamson M J, Xu D, Poole D, Le Grand S and Walker R C 2012 "Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born" *J. Chem. Theory Comput.* **8** 1542–1555 doi:10.1021/ct200909j

[121] Salomon-Ferrer R, Götz A W, Poole D, Le Grand S and Walker R C 2013 "Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald" *J. Chem. Theory Comput.* **9** 3878–3888 doi:10.1021/ct200909j

[122] Le Grand S, Götz A W and Walker R C 2013 "SPFP: Speed without compromise—a mixed precision model for GPU accelerated molecular dynamics simulations" *Comput. Phys. Commun.* **184** 374–380 doi:10.1016/j.cpc.2012.09.022

[123] Case D A, Cheatham III T E, Darden T, Gohlke H, Luo R, Merz Jr K M, Onufriev A, Simmerling C, Wang B and Woods R J 2005 "The Amber biomolecular simulation programs" *J. Comput. Chem.* **26** 1668–1688 doi:10.1002/jcc.20290

[124] Salomon-Ferrer R, Case D A and Walker R C 2013 "An overview of the Amber biomolecular simulation package" *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3** 198–210 doi:10.1002/wcms.1121

[125] Srinivasan J, Trevathan M W, Beroza P and Case D A 1999 "Application of a pairwise generalized Born model to proteins and nucleic acids: Inclusion of salt effects" *Theor. Chem. Acc.* **101** 426–434 doi:10.1007/s002140050460

[126] Perez A, MacCallum J L, Brini E, Simmerling C and Dill K A 2015 "Grid-based backbone correction to the ff12SB protein force field for implicit-solvent simulations" *J. Chem. Theory Comput.* **11** 4770–4779 doi:10.1021/acs.jctc.5b00662

[127] Dang L X 1995 "Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-Crown-6 ether: a molecular dynamics study" *J. Am. Chem. Soc.* **117** 6954–6960 doi:10.1021/ja00131a018

[128] Sutthibutpong T, Harris S A and Noy A 2015 "Comparison of molecular contours for measuring writhe in atomistic supercoiled DNA" *J. Chem. Theory Comput.* **11** 2768–2775 doi:10.1021/acs.jctc.5b00035

[129] Pyne A L B *et al.* 2020 "Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and recognition" *bioRxiv* 863423 doi:10.1101/863423 (Preprint)

[130] Sibson R 1973 "SLINK: An optimally efficient algorithm for the single-link cluster method" *Comput. J.* **16** 30–34 doi:10.1093/comjnl/16.1.30

[131] Defays D 1977 "An efficient algorithm for a complete link method" *Comput. J.* **20** 364–366 doi:10.1093/comjnl/20.4.364

[132] Roe D R and Cheatham III T E 2013 "PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data" *J. Chem. Theory Comput.* **9** 3084–3095 doi:10.1021/ct400341p

[133] Roe D R and Cheatham III T E 2018 "Parallelization of CPPTRAJ enables large scale analysis of molecular dynamics trajectory data" *J. Comput. Chem* **39** 2110–2117 doi:10.1002/jcc.25382

[134] Lavery R, Moakher M, Maddocks J H, Petkeviciute D and Zakrzewska K 2009 "Conformational analysis of nucleic acids revisited: Curves+" *Nucleic Acids Res.* **37** 5917–5929 doi:10.1093/nar/gkp608

[135] von Helmholtz H 1891 "On the thermodynamics of chemical processes" *Physical Memoirs Selected and Translated from Foreign Sources* (London: Taylor & Francis) pp 43–62 (Originally published in German, 1882)

[136] Kumar S, Rosenberg J M, Bouzida D, Swendsen R H and Kollman P A 1992 "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method" *J. Comput. Chem.* **13** 1011–1021 doi:10.1002/jcc.540130812

[137] Grossfield A 2017 "WHAM: The weighted histogram analysis method" v2.0.9.1 URL http://membrane.urmc.rochester.edu/?page_id=126

[138] Binnig G, Quate C F and Gerber Ch 1986 "Atomic force microscope" *Phys. Rev. Lett.* **56** 930–933 doi:10.1103/PhysRevLett.56.930

[139] Jalili N and Laxminarayana K 2004 "A review of atomic force microscopy imaging systems: Application to molecular metrology and biological sciences" *Mechatronics* **14** 907–945 doi:10.1016/j.mechatronics.2004.04.005

[140] Liu S and Wang Y 2010 "Application of AFM in microbiology: A review" *Scanning* **32** 61–73 doi:10.1002/sca.20173

[141] Le S, Chen H, Cong P, Lin J, Dröge P and Yan J 2013 "Mechanosensing of DNA bending in a single specific protein-DNA complex" *Sci. Rep.* **3** 3508 doi:10.1038/srep03508

[142] Møller H D *et al.* 2018 "Circular DNA elements of chromosomal origin are common in healthy human somatic tissue" *Nat. Commun.* **9** 1–12 doi:10.1038/s41467-018-03369-8

[143] Wu S *et al.* 2019 "Circular ecDNA promotes accessible chromatin and high oncogene expression" *Nature* **575** 699–703 doi:10.1038/s41586-019-1763-5

[144] Fogg J M, Kolmakova N, Rees I, Magonov S, Hansma H, Perona J J and Zechiedrich E L 2006 "Exploring writhe in supercoiled minicircle DNA" *J. Phys. Condens. Matter* **18** S145 doi:10.1088/0953-8984/18/14/S01

[145] Zgarbová M, Jurečka P, Lankaš F, Cheatham III T E, Šponer J and Otyepka M 2017 "Influence of BII backbone substates on DNA twist: A unified view and comparison of simulation and experiment for all 136 distinct tetranucleotide sequences" *J. Chem. Inf. Model.* **57** 275–287 doi:10.1021/acs.jcim.6b00621

[146] Dans P D, Faustino I, Battistini F, Zakrzewska K, Lavery R and Orozco M 2014 "Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA" *Nucleic Acids Res.* **42** 11304–11320 doi:10.1093/nar/gku809

[147] Collin F, Karkare S and Maxwell A 2011 "Exploiting bacterial DNA gyrase as a drug target: Current state and perspectives" *Appl. Microbiol. Biotechnol.* **92** 479–497 doi:10.1007/s00253-011-3557-z

[148] Chu X, Liu F, Maxwell B A, Wang Y, Suo Z, Wang H, Han W and Wang J 2014 "Dynamic conformational change regulates the protein–DNA recognition: An investigation on binding of a Y-family polymerase to its target DNA" *PLoS Comput. Biol.* **10** e1003804 doi:10.1371/journal.pcbi.1003804

[149] Pinson V, Takahashi M and Rouviere-Yaniv J 1999 "Differential binding of the *Escherichia coli* HU, homodimeric forms and heterodimeric form to linear, gapped and cruciform DNA" *J. Mol. Biol.* **287** 485–497 doi:10.1006/jmbi.1999.2631

[150] Yang S W and Nash H A 1995 "Comparison of protein binding to DNA *in vivo* and *in vitro*: Defining an effective intracellular target" *EMBO J.* **14** 6292–6300 doi:10.1002/j.1460-2075.1995.tb00319.x

[151] Jalal A S B, Tran N T, Stevenson C E, Chan E W, Lo R, Tan X, Noy A, Lawson D M and Le T B K 2020 "Diversification of DNA-binding specificity by permissive and specificity-switching mutations in the ParB/Noc protein family" *Cell Rep.* **32** 107928 doi:10.1016/j.celrep.2020.107928