IN SILICO EXAMINATION OF THE STRUCTURE OF CLOSED NAKED DNA AND PROTEIN/DNA COMPLEXES

by

LAUREN ADRIAN BRITTON

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Chemical and Biochemical Engineering

written under the direction of

Dr. Wilma K. Olson

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2012

ABSTRACT OF THE DISSERTATION

IN SILICO EXAMINATION OF THE STRUCTURE OF CLOSED NAKED DNA AND PROTEIN/DNA COMPLEXES

by Lauren Adrian Britton

Dissertation Director: Wilma K. Olson

The focus of this thesis is to study the overall global shape of DNA and its dependence on various intrinsic parameters both when naked and bound to proteins. We developed a definition of the twist of DNA base-pair steps, Tw^{SC}, that can be related to the overall folding of the DNA molecule and is a new contribution to the field of DNA topology. We developed a software package, titled 3DNATwSC, to calculate Tw^{SC}. We established, by studying seven simplified ideal structures, the value of Tw^{SC} as an effective gauge for the topological landscape of DNA due to its sensitivity to changes in chirality.

We constructed a web-based user-friendly database, named TwiDDL, to show the impact of protein binding on Tw^{SC}. TwiDDL makes it easy to look for unusual values of Tw^{SC} in biologically relevant structures. We identified a number of highly unusual over- and under-twisted molecules by comparing Tw^{SC} to the twist of ideal B-DNA. We developed 2D and 3D-plots to highlight patterns and unusual deformations in Tw^{SC} in a

variety of structures. We examined the twisting of DNA in representative A, B, and Z-DNA structures as well as in a collection of 45 nucleosomes. We also studied the effects of shearing on Tw^{SC}. When shearing was removed from both a nucleosome and an HU-DNA structure, the value of Tw^{SC} approaches that of Tw^{SP}, the rigid-body step parameter twist.

We created a user-friendly application, called 3DNAdesigner, that allows a user to construct minimum-energy configurations of open linear and spatially confined DNA molecules. We presented the software design, described its features, and gave three detailed examples. The software was used to create minimum-energy configurations of DNA mini-circles and DNA/protein complexes. The changes in Lk and energy introduced by binding protein to a mini-circle might be of interest to a lab performing *in vitro* studies. We also found numerous minimum-energy configurations of the Lac operon bound to the Lac repressor in the absence or presence the HU protein.

Acknowledgments

Since I have been part of the Rutgers University family so very long, both as an undergraduate and then, after working for a few years, as a graduate student, my path has crossed many wonderful people. I would like to have a chance to thank so many who helped guide me in one way or another.

First and foremost is my husband, Jim Britton, who gave me both the moral support and the technical support I needed, including access to state of the art computers I would have otherwise never been able to use.

I would like to thank Prof. Wilma Olson, my thesis advisor, for taking me in at a pivotal point in my studies and opening up a new field for me. You have been astoundingly diligent about being deeply involved in all of my work. I know anything that leaves your lab will always hold perfection. You were always there for me and offering kindness and advice through both professional and personal growth experiences. I also want to thank the one we called my "not so silent advisor" Prof. Irwin Tobias for having me be a part of his wonderful theories and amazing mind for mathematics. You both have truly been my heroes. A lot of my research was extended from work done by Prof. David Swigon, his graduate advisor Prof. Bernard Coleman, and Dr. Yoav Biton (Coleman's graduate student). Thanks to the three of you for the time spent on

discussions about our work and for your patience when I was just joining the lab.

My journey to the Olson lab was not an easy one, and I would like to thank Prof. Helen Buettner and Martin Yarmush for pointing me in Wilma's direction. Prof. Buettner also helped me immensely in preparing for my master's thesis defense. I am so lucky to have had you by my side.

A significant part of my time at Rugers graduate school was spent teaching. Whether engineering or chemistry, I was never happier than when I was able to help students with their education. I spent much more time than "necessary" in order to help my students grasp concepts. Thank you Profs. Yee Chiew, Marianthi Ierapetritou, Donald Siegel, and Peter Couchman for having me TA and lecture. I also was able to teach statistics at the Passaic County Community College to fill in for my dad, Art Weintraub, when he had bypass surgery. I am happy that he is recovered and teaching again all on his own.

Thanks to all my group members over the years and especially Drs. Andrew Colasanti, Ariella Sasson, Fei Xu, Yun Li, Yangyang Shen, Tony Felts, and Yurong Xin who have always been there for me to bounce ideas off of and to chat about just anything.

Thanks to my friends Jack Chen, Nobel Shelby, and Mike Burke for their support of me when I needed it. Mike and Nobel, it was great to have you there for encouragement. Jack, oh Jack. You have been the best friend Jim and I could ever have

asked for. You helped us pack up and move and were always the first one there and the last one to leave. You are truly a brother to us and a kindred spirit.

Thank you to all of my professors who put in many hours of work into teaching me and my peers the course material.

Thanks to the engineering assistant Dean Don Brown and the graduate schools very own Barbara Sirman for all of their help and lending me their ears.

Thank you to the Department of Chemical and Biochemical Engineering for having me as a part of you. Special thanks to my friend and teacher Professor Alkis Constantinides. Thanks also to Prof. Fernando Muzzio for helping find my first job after undergraduate with Catalytica and for introducing me to one of my best mentors Dr. Daniel Löffler. Thank you also to my committee members, Profs. Yee Chiew, Henrik Pedersen, Irwin Tobias, and Wilma Olson for reading my huge thesis and listening to my defense.

Dedication

I dedicate this thesis to the one person who always gives me love, joy, and unconditional support, my best friend and husband, Jim Britton. I don't think a more magnificent person has ever walked the earth. Smart, loving, caring, patient beyond belief, talented in many genres, gorgeous inside and out, giving, nurturing, a chef worthy of many Michelin stars, a terrific horseman, a farrier who goes beyond, dog walker, cat adorer, thesis helper, parallel parker, and a great many other positive things. We have been through a lot in our almost 20 years together and no matter what, we are the perfect team. You are my other half and my light. Thank you sweetheart, I love and adore you.

I also want to dedicate this thesis to my family who have in their own ways given lots of love and support. To Fran Britton, my wonderful mother-in-law, you are missed and I am truly blessed to have had you in my life for as long as you were. I will always remember your sweet smiles, warm hugs, and zest for life. I was lucky enough to have the support of my dad and stepmother (Art and Cheryl), mom and stepfather (Marilyn and Larry), sisters (Wendy and Amy), brother (Josh), cousin/sister (Monique), grandmother (Rosalie), and father-in-law (Joe). Wendy you are my best friend next to Jim. What an amazing woman you have become, the world could use more like you, thanks for everything. A special thanks to my grandma and dad for always asking me if I

was done yet ;-) and a really special thanks for all of those who didn't ask ;-P

And last but not least, I want to thank my menagerie for keeping me sane while also adding to the insanity. Peek-a-boo, Socrates, Romeo, Bagels, Shadow, and Wembley.

Table of Contents

List of Tables

List of Figures

Chapter 1: Background/Significance

1.1 What is DNA?

DNA is a polymer which contains codes about the most basic level of life's building blocks. These DNA codes are translated, by a series of actions made by protein machines in the cell, into useful entities for a living organism to be able to exist and function. At this point in time not much is known about the information encoded in the molecule. Only 1-5 % of the DNA inside each cell is expressed as protein [1]. The idea that creating proteins is the only productive phenomena that DNA carries out in a cell is beginning to change. Many researchers are attempting to decipher the 95% of the DNA function that has been, until now, completely hidden from us. Even though much of DNA's function is hidden from us today, many researchers aim to unlock its secrets through a better understanding of its structure. Of primary importance is understanding how DNA, at lengths on the order of a meter, is accommodated within a single cell. The compaction required to achieve this is accomplished through a series of bends and folds. The DNA is packed into an assembly that ensures an organized network that is necessary for proteins to be able to work with the encoded information in the long molecule. To make use of one of the codes in DNA, namely the genetic code, DNA needs to be translated into RNA with the help of many proteins, including a key enzyme called RNA polymerase [4]. Messenger RNA, one type of RNA product, in turn, is used as a code to

form proteins. When a eukaryotic cell undergoes cell division it will make a copy of the DNA using various proteins, including an enzyme called DNA polymerase, so that each cell will have its own copy. All of these processes require that the RNA or DNA created by these processes form an organized network; however, this does not always happen correctly and other enzymes are then needed to untie or untwist portions of the newly created sequence. Sometimes errors or misincorporation of the wrong bases can be repaired, while other times the whole process needs to start again and throw away the work that brought the DNA into the current improper state [5 ,6].

It is important to understand protein/DNA interactions and what happens to the topology of protein-bound DNA because of their interdependence in processes like those mentioned above. In order to gain that understanding, the structure and the chemistry of these molecules must be further analyzed. The picture of DNA that most people can recognize is the double-helical structure containing a series of steps that twist like a spiral staircase. These "steps" consist of two nucleotides from separate strands held together by interactions known as base pairs. On the level of base-paired nucleotides, DNA can be understood in terms of an assembly of atoms. When viewed at a higher polymeric level, DNA can be understood in terms of its global properties, such as the topology.

DNA is an acronym for deoxyribonucleic acid, a polymer that consists of a repeating sugar-phosphate backbone and laterally attached heterocyclic base side groups. The sugar-phosphate backbone to which the nucleic acid bases are attached is soluble in

water. The phosphates on this backbone carry negative charges. The bases are highly hydrophobic, i.e., repelled by water, and tend to stack one above the other along the chain. The backbone and attached bases define a single strand, and each base has a complementary nucleic acid base on the adjacent strand that it prefers to bind to. The bases are of two types, the purines, adenine (A) and guanine (G), and the pyrimidines, thymine (T) and cytosine (C). The most common arrangement of base pairs is called a Watson-Crick base-pair configuration, which takes advantage of the isosteric nature of four common base pairs (A-T, T-A, G-C, C-G) and allows all four to fit in the same double-helical framework. When these two strands have bonded, the polymer as a whole takes on a double-helical form [1]. Each nucleotide has only one preferred partner and the strands run opposite to each other, thusly are anti-parallel.

1.2 Basic Underlying Reason for the Shape of DNA

The famous double-helical twist of the DNA is due to a combination and balancing of the forces acting upon it. Hydrogen bonding is the force that holds the two bases from opposite strands together. These bonds which solder the two strands are situated between the complementary bases and constitute the common Watson-Crick base-pair, as seen in Figure 1.2. Because the interior of the cell consists mostly of water and the nucleic acid bases are hydrophobic, Calladine [7] believes that the twisting is induced by the hydrophobicity, i.e., tendency to be repelled by water, that the bases

demonstrate towards the cell media. The bases try to pack tightly, as stacked arrays, to avoid being exposed to the water in the cell. The optimum packing structure for the shapes of the bases is a helical twisted array resembling the steps of a circular staircase shown in Figure 1.1 [1].

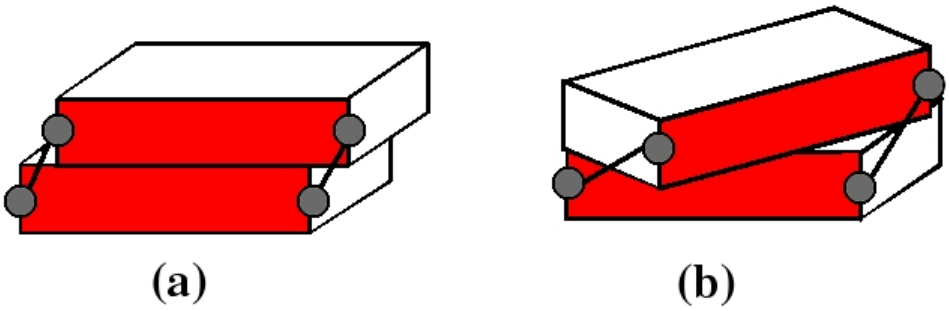


Figure 1.1: Schematic of one base-pair step of generic DNA showing (a) optimum stacking without twist and (b) optimal stacking using a helical structure. The paired bases are represented by the slabs. The red highlighted area indicates the minor groove edge of the base pair. The circles are points where the backbone meets the base and the lines that connects the circles show that the distance between the base-pairs can vary from one configuration to another.

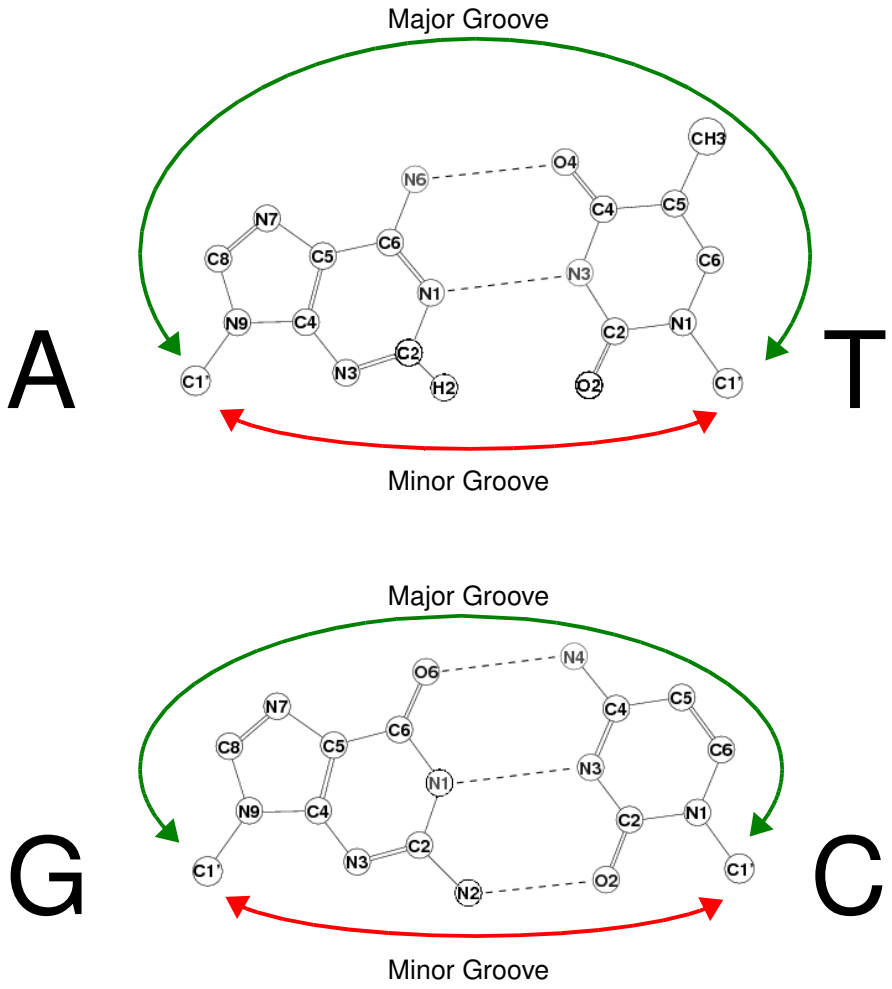


Figure 1.2: Atomic renditions of Watson-Crick A-T and G-C base pairs, where the dashed lines represent the hydrogen bonds that link the bases together. The minor groove is highlighted by the red arrows, similar to the red highlighted minor groove in Figure 1.1, and the major groove is highlighted by the green arrows.

A third influence on the packing of the structure, other than hydrogen bonding and stacking, comes from the electrostatics of the macromolecule. DNA is a polyanion since the phosphates on the sugar-phosphate backbone are negatively charged. The arrangement of phosphates creates a repulsive force between base-pair steps. The amount of tight packing that can take place is limited due to both the base and phosphate

repulsions. The greater the salinity, the more the charges on the phosphates are neutralized and the more DNA becomes flexible. The lesser the salinity, the greater the repulsion between the negatively charged phosphates and, the stiffer DNA becomes [2, 3]. The bases themselves are polar, which means that the charge is not uniformly dispersed and can be influenced by surrounding charges. When the cell media has a salinity change, the shape of DNA is affected at both the level of double-helical structure (A-, B-, C-type helices) and chain deformability.

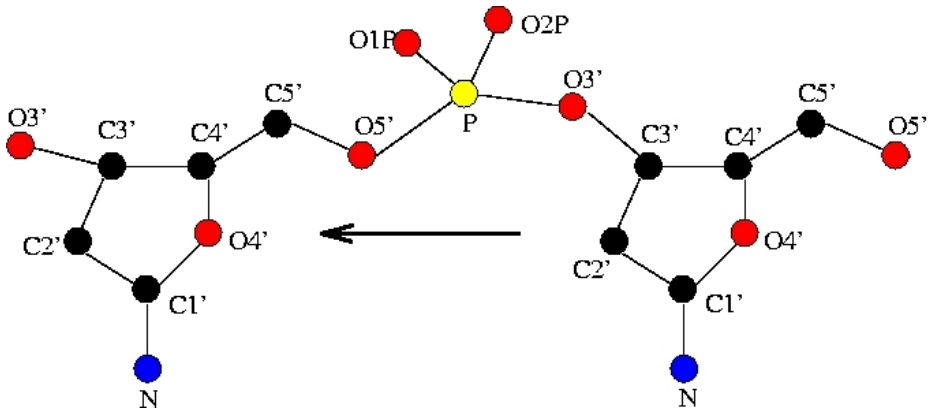


Figure 1.3: Sugar-phosphate backbone with all atoms color-coded and labeled with the arrow indicating the 5´ → 3´ direction of the chain. Red denotes oxygen, yellow denotes phosphorus, black denotes carbon, and blue denotes nitrogen. The 5´ → 3´ notation refers to the ends of the DNA.

Now that the form of DNA has been declared a double helix in the above paragraphs, it is important to note that the complexities do not stop there. The helix has a directionality. This means that, depending on the way successive nucleotides rotate and translate with respect to the global axis, the double helix can be left- or right-handed. Also, each strand of the DNA is read in a specific order. The sugar-phosphate backbone

connects all the bases together and is read from the 5´ to the 3´ end of the chain. Successive nucleotides are attached through the 3´-oxygen of one nucleotide and the 5´-oxygen of the next. This is shown in Figure 1.3.

There are many forms of double-helical DNA, which are designated by letters ranging from A, B, C, all the way to the Z form. One major difference between the forms is the helical twist. Helical twist is the number of bases that can fit in one 360° revolution of the sugar-phosphate backbone. The A form was the first DNA structure that was discovered [8], and like the second B-type structure , is right-handed. The A form contains 11 base pairs per helical turn and the B form, the typical form seen under normal cell conditions, contains 10 base pairs per turn [9]. The C form is also right-handed and has nine base pairs per turn [10], whereas the Z form is left-handed and has 12 base pairs per turn. Other than the handedness of DNA there exists variant forms such as single-stranded structure (e.g., broken or melted double helices, triple-stranded helices, four-stranded helices, and four-stranded Holiday junctions).

As illustrated in this section, the shape of DNA is determined by many internal and external factors. Each of these factors plays a unique role in affecting the forces that twist and turn DNA into the various forms that have been observed in high-resolution (NMR, X-ray) structures, and deduced from various other measurements, such as gels.

1.3 Supercoiling

The length of the double-helical DNA that resides in each human cell is around 2 meters and holds roughly three billion base pairs. The diameter of the helix is $2\times10^{-9}\text{m}$ [1]. DNA is able to fit into a cell by a series of "folds" called supercoils. Since DNA is already a double helix, any other coils which form on top of the helix are therefore referred to as supercoils [11, 12, 13]. Supercoiling is a natural occurrence when two ends of the same piece of a DNA segment are forced to an orientation or spacing other than that in its relaxed state. An easy way to picture this is by looking at a telephone cord and thinking of it as a helix. Then when the telephone cord becomes twisted, it forms a supercoil. Following this train of thought, it becomes easier to visualize the DNA inside the nucleus of a cell. DNA wraps around proteins, called histones, to form small superhelices in eukaryotes, or living systems which contain a cell nucleus. These protein/DNA complexes, known as nucleosomes, are strung together with others in a chain. The chain of nucleosomes is folded into a tightly packed structure in order to form a chromosome.

Covalently closed sequences of DNA, are of primary interest in this research study due to their unique supercoiling properties [14]. Closed space curves have spatial characteristics which can be described in a unique way [15, 16, 17]. A very interesting attribute of closed DNA is that the overall shape is coupled to the twisting of the double

helix. When a double helix is closed there is a finite number of times that one strand wraps itself around its mate. This is referred to as the linking number. The linking number, *Lk*, is related to two other quantities by the White equation [18, 19]:

$$Lk = Tw + Wr \qquad \text{Eq. 1.3.1}$$

Here *Tw* is the twisting of the helix, an additive property of the twist of successive base pairs as you travel along the space curves, and *Wr* is the writhe, a measure of the overall folding of the closed piece of DNA. When a circular section of DNA, known as a plasmid, lays in one plane $Wr = 0$, as is also the case when all displacements out of a plane are symmetric. When sections of a plasmid come out of the plane the writhe may increase or decrease depending on the directionality of the displacement. When DNA gets over or under twisted the shape will compensate to counteract the change in twist by exhibiting supercoiling. These definitions are purely mathematical. Figure 1.4

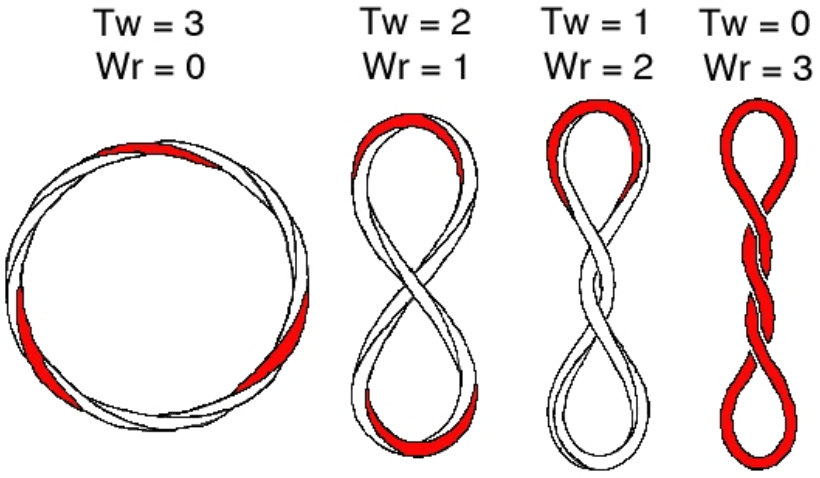


Figure 1.4: Illustration of Twist versus Writhe for Lk = +3. Here DNA is represented by a four-sided structure with the "minor-groove" edge highlighted in red with the remaining three sides shown in white [1].

demonstrates the usefulness of Eq. 1.3.1. The only way to change the linking number of a molecule is to create a new configuration of the molecule by physically opening the loop, changing the number of times one strand wraps around its partner, and re-closing the loop. Two or more configurations of a molecule with the exact same chemical makeup but different linking numbers are called topoisomers. Topoisomers of a known sequence can be created and made to take on different shapes, with varying degrees of supercoiling, by the addition of outside forces, such as salt or drugs.

In a closed DNA molecule, the linking number needs to be an integer because both ends of a strand must bind to each other to close the molecule, and because the nucleotides at the ends of the strands can only bond in a single orientation. This means that, for relaxed DNA, the ends of the strand can only bind to each other when the backbone of the last base pair lines up to form a continuous pathway with the backbone of the first base pair. Eq. 1.3.1 shows how the total twisting of the DNA steps combined with the writhe of the closed molecule gives the linking number. When the total twist of the DNA molecule, *Tw*, is calculated using the step parameter twist the linking number is not necessarily an integer. In the work done by David Swigon and the subsequent program developed based on his work, the linking number calculations were performed incorrectly by rounding the calculated sum of the writhing number and twist to the nearest integer. This was a large red flag that prompted us to search for a better way to calculate the twist of the DNA base-pair steps.

1.3.1 DNA Simplifications

This thesis uses the Kirchhoff rod model to measure the cost of deforming DNA from its rest state and to calibrate the energies of different topoisomers [20]. One model used in this study treats DNA as an inextensible, isotropic, deformable rod that is small in diameter compared to its contour length. However, actual DNA is not a continuous rod; it is made up of discrete steps and broken helical fragments. These discrete steps are made up of base pairs. We will discuss in Section 1.4 how we view a base pair as a rectangular slab. A base-pair step is composed of two base pairs which are stacked on top of each other. This stacking is described by six degrees of freedom which specify the orientation and displacement of one base pair to its neighbor.

There is a new method derived by Dr. Irwin Tobias to describe the supercoiling topology of a discrete base pair sequence [21]. One difference between the Kirchoff and discrete representations of DNA is that the former model uses an idealized rod and an ideal uniform sequence whereas the latter model is sequence-dependent with capabilities that reflect the natural spatial arrangements and motions of DNA base-pair steps. Knowing the sequence and being able to model that sequence can bring a greater understanding of the how and why DNA can take on different shapes. That is why the discrete model of DNA is a large improvement on just being able to model DNA as an ideal rod. The more accurate sequence-dependent model is based on the properties of real DNA in in vitro experiments.

1.4 Model Simplifications

Studying a large DNA system makes it necessary to simplify the way that the molecule is modeled. One of the biggest simplifications that is made in this work is the representation of each nucleotide base as a slab. The bases adenine (A), guanine (G), cytosine (C), and thymine (T) are relatively planar, aromatic structures that form flat configurations consistent with a slab shape. When a base from each strand on the double helix finds and forms a Watson-Crick pair with its mate on the other side, the resulting shape of the base pair can be approximated by a rectangular slab. Because the nucleotide is a very rigid chemical unit within the double-helical structure owing to the stacking and hydrogen bonding, this is a valid reason to model the bases in this way.

At present there is not enough information available to know all of the chemical interactions that take place in a long DNA sequence. Everything we know about long DNA sequences comes from only a few thousand X-ray crystal structures. It is also impossible to know all of the base and chain interactions between helices. We have limited information about the packing of helices in high-resolution structures. X-ray crystal studies cannot detect every atom/bond because of limitations on the resolution and size of structures that can be studied. The largest piece of DNA studied to-date at high resolution using X-ray crystallography is a 147 base-pair sequence wrapped around the nucleosome core particle depicted in Figure 1.5 [22]. Even though there are models of

DNA based on all-atom simulations, such as molecular dynamics [24, 25, 26, 27, 28], it is only now becoming feasible to study a large molecule with such methodologies. State-of-the-art molecular simulations are typically used to treat a few turns of a double helix (20-30 base pairs) in water and surrounding ionic media [29]. To do otherwise takes very long, needs large amounts of computer time, and requires more computer power than what is available to most users.

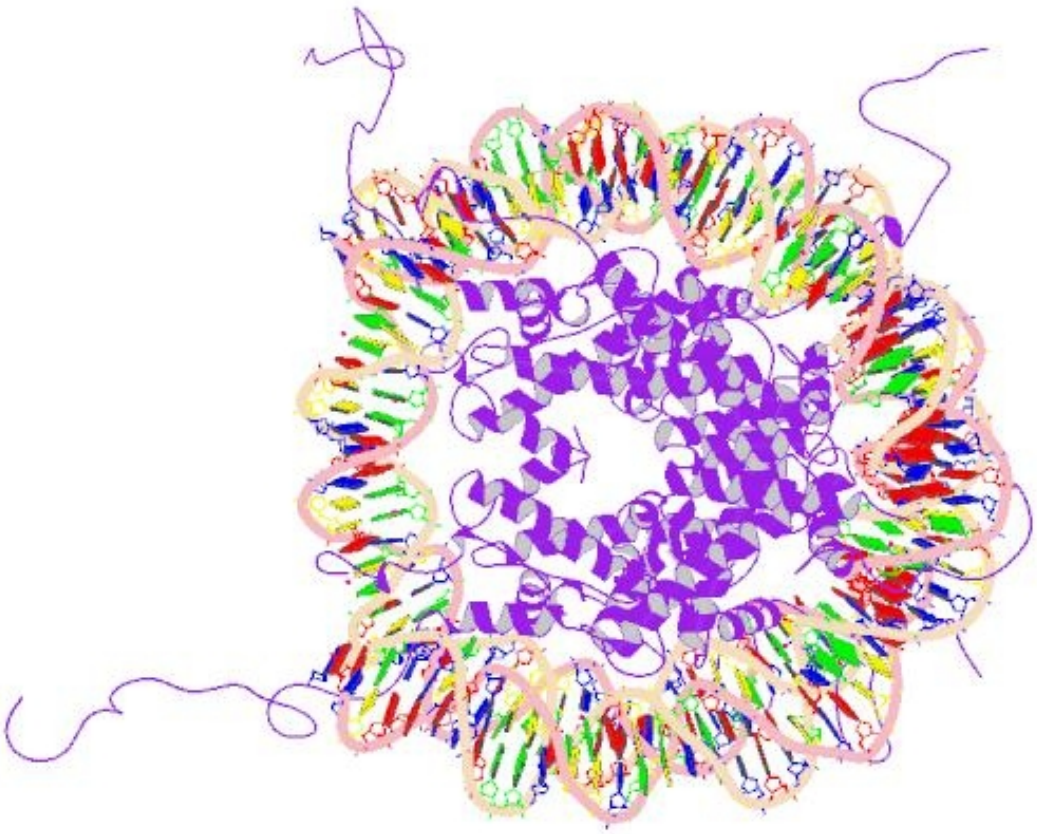


Figure 1.5: DNA double helix wrapped around a nucleosome core particle represented by PDB ID 1kx5 [22]. The histone core particle is represented by the ribbon like curve in purple, the double helical DNA backbone is a thicker curve in pink, and the base pairs consist of adenine in red, thymine in blue, guanine in green, and cytosine in yellow. This image is from the PDB [23].

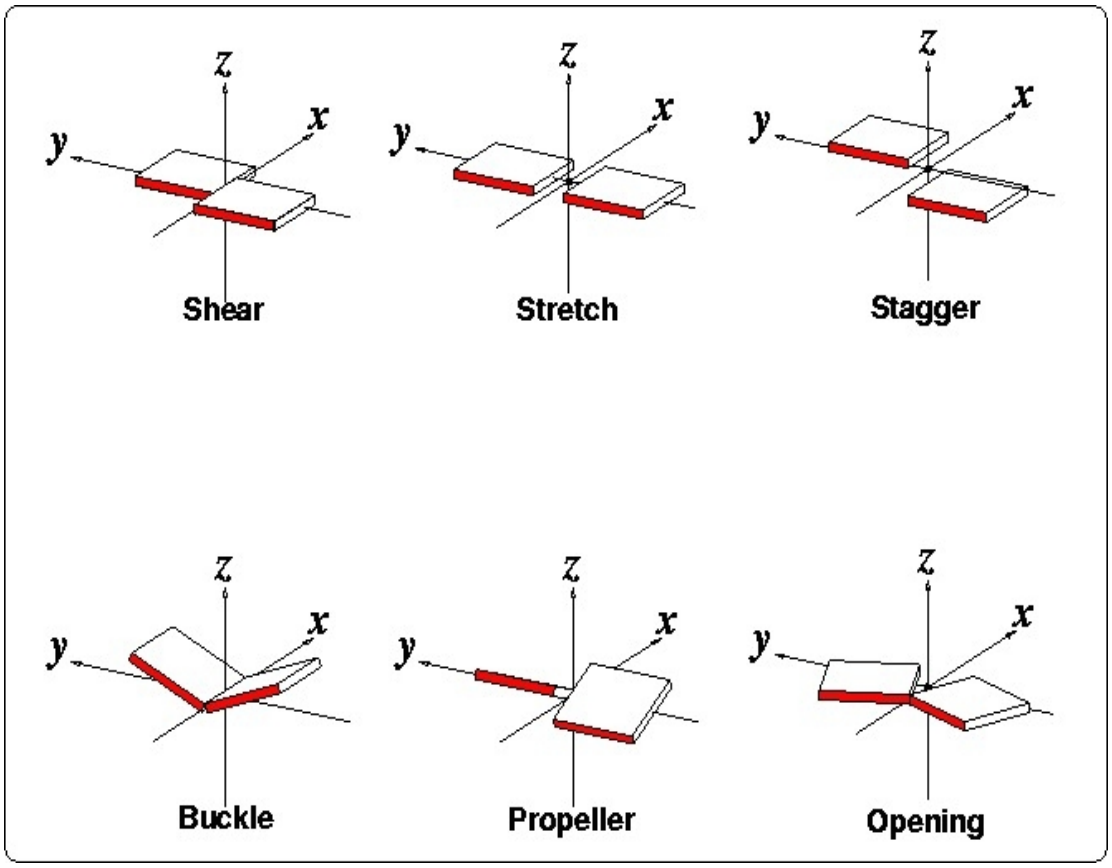


Figure 1.6: Base-pair parameters used to describe the translational and rotational movements between bases on adjoining strands. The red highlighted area denotes the minor groove edge of the base pair. The triple axes origin is placed locally at the center of the base pair.

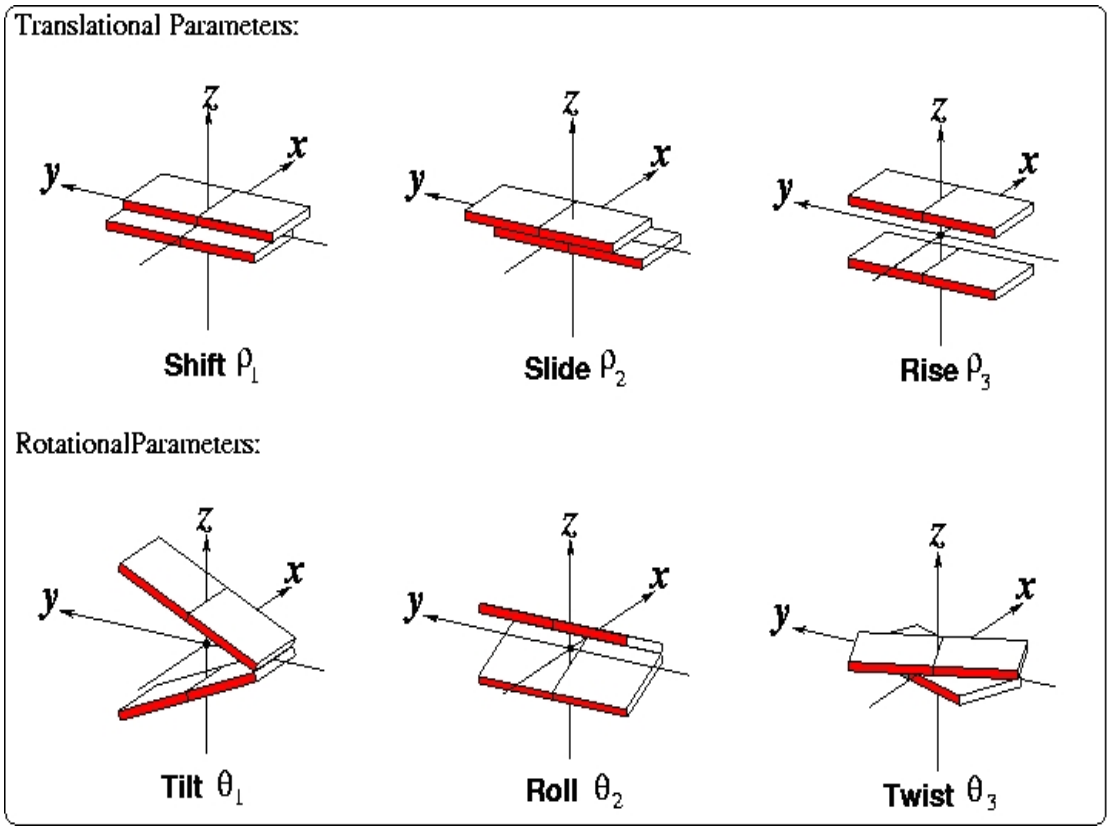


Figure 1.7: Step parameters used to describe the motions between two consecutive base-pair steps. The red highlighted area denotes the minor groove edge of the base pairs. The triple axes origin is placed locally at the center of the base-pair step.

In order to address these limitations, our research assumes that the base pair can be modeled as a rectangular slab. In actuality the rectangle is an over simplification of what really happens. Complementary bases do not lay completely flat within the same plane. Instead, between the two strands of the double helix there exists buckling, opening, propellering, shearing, stretching, and staggering, as shown in Figure 1.6. However, the model used here will not deal with the parameters that govern how the bases interact with their mates, but rather with the interactions between adjacent neighbors, i.e., base pair steps [30,31].

1.4.1 Intrinsic Properties of Base-Pair Steps

A base-pair step consists of two base-pair slabs (in actuality four bases) and the intervening sugar-phosphate backbones. The parameters that govern how each base pair interacts with its nearest neighbor consist of six degrees of freedom, corresponding to deformations from the rest state, a state taken from averaging over a very large data set of structures of the same step type (e.g., AG) to find an average of the step parameters, and a set of sequence-dependent elastic force constants. Three translational parameters specify the direction in which the base pairs move with respect to the axes of the base-pair step; Three rotational parameters specify the angles of rotation about these axes. The diagram shown in Figure 1.7 illustrates positive representation of each movement. As previously discussed, DNA is made up of four bases. The base pairs can form a base-pair (or dimer) step in 16 different combinations (AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG,

GA, GT, GC, GG). Since each strand is unidirectional AT does not mean the same thing as TA, but AA and TT do represent the same dimeric fragment. Each of the 10 unique dimer steps has unique intrinsic parameters associated with it as compared with any other dimer step [32]. For example, TT and AA are not unique. The base-pair step parameters relating the side groups for the two strands of the step are identical except for shift and tilt, which differ in sign. TT and AA are complementary sequences, as are GG and CC, GA and TC, AG and CT, CA and TG, and AC and GT. The remaining four base-pair steps AT, TA, GC, and CG are called self-complementary or autocomplementary base-pair steps in that the sequences of the two strands are identical. Coupling may exist between step parameters, such as twist and roll. That is, a variation in one parameter prompts a particular change in another. Translational parameters are coupled to angular parameters and/or to other translational parameters in some steps. Rise is the most restricted of all of the six parameters due to the van der Waals interactions between adjacent hydrophobic base pairs.

The intrinsic rest-state values for the six parameters and the associated force constants, also called elastic moduli, have been determined through the analysis of the X-ray structures of many oligomers [32]. Gō and Gō [33] originally provided a way to calculate the force constant matrix $\boldsymbol{F}$ from an all-atom energy function and then to derive its inverse, the covariance matrix, $\boldsymbol{F}^{-1}$ which can be used to deduce the elastic moduli.

The sequence-dependent elasticity moduli calculated in this thesis are based on that method. The energy of a base-pair step is taken to be zero in the rest state and assigned values which are proportional to the deformations from the rest state; therefore, we can only measure a relative energy value. Changes in a particular translational or rotational parameter may have a higher or lower cost to the energy depending on where in the sequence the step is located (i.e., in the context of its neighboring base pairs), how many parameters are affected, and to what degree each parameter is changed. Sometimes a lower energy change may occur if two or more parameters are coupled with each other rather than move independently of one another [32].

The rest states and elastic moduli used in this research are taken from the analysis of the step parameters in a large collection of B-DNA and protein-DNA structures by the Olson group [32, 34]. The step parameters are obtained using the 3DNA suite of programs, a software package developed by Lu in the Olson group to "analyze, reconstruct, and visualize" DNA [35]. This is the software that is used to collect base-pair-step parameters between adjacent Watson-Crick base pairs in known X-ray structures. There is work in progress in the Olson lab that looks at the properties of dimers in difference sequence contexts, including tetramers. Now there are sufficient data to extract acceptable ensemble averages.

Orthogonal base-pair step triads are used to determine the base-pair-step parameters. If the base pair is viewed from the top down, the positive y-axis of the triad,

referred to in this study as the long axis, travels from the complementary strand to the leading strand or sequence. The short axis is perpendicular to the long axis and pointed toward the major-groove edge of a base pair. The normal is perpendicular to the average or middle plane of the base-pair step. In practice, the long axis is found from the cross product of the normal and the short axis. The three axes, or triads, describe the orientation of the base-pair step or dimer coordinate frame and aid in establishing the step parameters. The triads are determined from the atomic coordinates of the DNA in a Protein Data Bank (PDB) file with 3DNA. Triads are placed on every base using a least-squares process that fits an ideal, planar base structure to the observed base. The triads are added with the ideal base structures and then used to find the coordinate frame of the base-pair steps and the base-pair step parameters. The values of the step parameters do not depend on the direction that the dimer step is viewed, i.e., the leading strand or complementary strand. The only caveat to this rule about the directional independence of step parameters involves the x-axis step parameters, shift and tilt, which are of the same magnitude in either direction but change in sign (i.e., + or -) when measured in terms of the two strands [36, 37].

1.5 DNA and Protein Interactions

Now that it has been established that naked DNA has a sequence-dependent shape it needs to be noted that proteins and the ionic surroundings can affect that shape. Since

DNA is a polyanion it attracts and is easily deformed by the presence of positively charged counterions. Such ions, therefore, have an effect on the overall shape and the energy of the molecule. Ions can neutralize a phosphate group on the backbone and thereby may interfere with the phosphate-phosphate repulsion that pushes one nucleotide from the next or may allow parts of the molecule to get close to one another.

When proteins or other ligands bind to DNA, the double helical structure may deform it into a conformation different from that which it assumes when uncomplexed [38, 39]. There are three main types of proteins that interact with DNA: enzymes, transcription factors, and structural proteins. The cutting and joining of DNA strands by enzymes, such as topoisomerases, is one application of how proteins can act on DNA [40, 41, 42, 43]. DNA can be copied into RNA or an exact copy of the original DNA. Polymerases are enzymes which are able to bind to sites on the desired gene and perform these copying processes. Transcription factors control the copying of genes by sending out recognition signals to the polymerase enzymes [44, 45]. Such factors work for or against transcription because the transcription factors can either block or promote the action of polymerases [44]. Structural proteins are proteins that generally bind with DNA to aid in the packaging, and thus the organization, of DNA within a cell. A prime example of this is the nucleosome mentioned previously. Intercalating drugs can also affect the shape of the DNA molecule. When inserted between base pairs, these drugs lengthen the distance between the base pairs and change the rigid-body parameters of

affected steps.

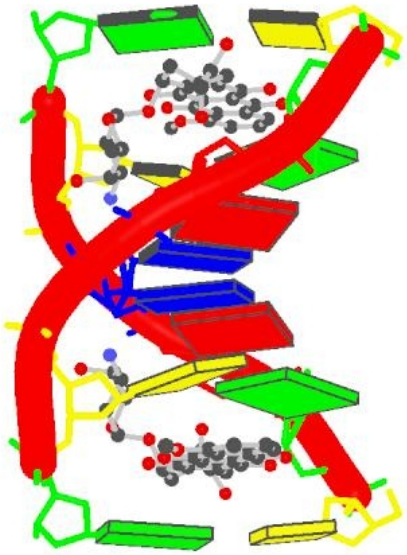


Figure 1.8: The association of the intercalating and groove-binding drug daunomycin with DNA creates additional spacing between base pairs at two sites [48]. The daunomycin drug is represented by the ball and stick molecule models near the top and bottom of the figure. The double-helical DNA backbone is the thick curve in red, and the base pairs consist of adenine in red, thymine in blue, guanine in green, and cytosine in yellow. This image is from the PDB [49].

1.6 Previous Model

The program used to build models of DNA in this thesis was originally designed by Dr. David Swigon when a member of the Olson group. It has been used to model ideal systems of naked and protein-bound DNA [46, 47]. New analysis of a multitude of X-ray crystal structures has provided intrinsic parameters and elasticity moduli which can be used in modeling more realistic DNA [32]. The Swigon computation finds the step parameters of a configuration of minimum elastic energy using a Newton-Raphson scheme starting from an initial state (set of step parameters) with boundary constraints placed upon its ends. The ability to study real DNA opens up a new world of possibilities, the first of which is working with biologist counterparts, such as Lynn Zechiedrich at Baylor College of Medicine, and physics collaborator, such as Sarah Harris at the University of Leeds.

1.7 Objectives

With a defined set of assumptions for DNA and a rudimentary description of how DNA base pairs are modeled, our study aims to provide an advancement in our understanding of the structure of large DNA fragments that can be generated within the limitations and accuracy of these models. This research utilizes state-of-the-art calculations for determining DNA shape with the intention to improve on the model as more technology becomes available. Prior to this research biologists could not reliably

determine shapes of the DNA sequences that they are interested in. The work in this thesis helps by using data our group has collected over the years on DNA base-pair step parameter values and examining their impact on the overall structure of any DNA helical segment.

Now that we have been able to describe DNA as a physical system that can be simplified, a main goal of this research is to establish a modeling technique that yields useful and accurate output. Biologists are the ones who should benefit most from this study. The reason for this is the simple fact that they are the ones looking at DNA in vivo and in vitro. Our work gives insight into the structure and the relative energies of those shapes and topoisomers more reliably than running a gel experiment. In fact, our work may even give more information about what is happening in the gels a biologists can study today. Our research benefits do not stop there. We can help biologists find specific sequences tailored to suit their interests by having them supply some simple preliminary requirements which are then used to generate three-dimensional structures that can answer their questions.

Biologists today cannot easily find out how binding a protein/drug/ligand to their DNA molecule affects the overall shape of that DNA. Currently, they could go to the Protein Data Bank or the Nucleic Acid Database to determine the step parameters and sequence of the DNA that is bound to the protein (not always reliable) and to visualize the molecular complex. However, as noted above, the largest DNA structure resolved from

X-ray crystallography to date is the 147 base pairs long DNA that is wrapped around the core of eight histone proteins, and forms the complex known as a nucleosome. These structural databases do not show how the rest of the DNA that is unbound is affected by the formation of the nucleosome. There exists a new database, which is part of this thesis, called TwiDDL, that will help scientists understand the effects on the twist of DNA that occurs from the protein binding. TwiDDL, located at http://twiddl.rutgers.edu, has a large number of structures that are taken from the Nucleic Acid Database and the Protein Data Bank and fit certain requirements. The structures in TwiDDL are fully searchable and supplemented with visualizations, including color-coded tables, comparison charts, and three-dimensional models.

Digging deep into what has been done in the field of DNA topology and what could be done to make the task of predicting the outcome and/or interpreting results from biological wet lab experiments, we have come up with the following specific aims that comprise the main foci of this thesis study.

1.7.1 Specific Aim 1: To construct a model of the twist of DNA base-pair steps as it relates to the overall DNA molecule and the effects that bound proteins have on DNA twist.

The purpose of this specific aim was to improve the accuracy of the calculations

of the twist in order to provide a better understanding of the structure of supercoilied DNA. The linking number *Lk*, defined by Eq. 1.3.1, helps us understand the topology of a closed molecule by counting how many times one strand of the DNA double helix wraps around the other. As shown in this thesis, in order for *Lk* to be an integer, the twist variable *Tw* in Eq. 1.3.1 must utilize the step parameter twist, Tw^{SP}. This means that for each base-pair step the Tw^{SP} for that step is added to each and every Tw^{SP} along the entire molecule. This number is divided by 360º to determine the number of turns the base-pair steps have undergone around the molecule and that is the number used for *Tw*. This widely used method to calculate the linking number had not yielded acceptable results since the values were non-integers when only an integer value would be valid. When looking into how the twist of base-pair steps was defined, it became obvious that there was a problem using the step parameter twist with the writhe.

A new contribution to the field of DNA topology came out of this specific aim. We have defined a new twist of supercoiling which is not only appropriate to use in determining the linking number but a twist that gives insight into the chirality of DNA as you travel around the molecule. This twist is useful in both closed DNA molecular structures as well as in open molecules.

1.7.2 Specific Aim 2: To create a web accessible interface to showcase the impact of the twist of supercoiling on DNA/protein complexes taken from NMR and X-ray crystal structures.

The new methodology developed in Specific Aim 1 yielded a new understanding about the twisting of DNA structures, but as a mathematical expression only it is hard to see its value. In the second specific aim of this research we set out to develop a widely available set of data based on existing DNA structures that would show the meaning of the twist of supercoiling and benefits of the new measurement. There are plenty of resolved DNA/ligand structures that are available to anyone who is interested to browse through the Protein Data Bank and the Nucleic Acid Database. TwiDDL, the outcome of Specific Aim 2, takes those structures and calculates the twist of supercoiling, the six step parameters, and differences between the two twists. TwiDDL also provides a three-dimensional visualization tool to get a closer look at the twists and the differences between the two twists. The database includes descriptions of the structures, and some graphical tools and files to give further information to look over. All of these data have been stored in a database and provided a user friendly web interface that quickly and easily allows end users, like biologists and other researchers, to find a structure of interest, to access our analysis of the structures, and to see visualizations that show how the measurement differs from previously reported descriptions of DNA twists.

1.7.3 Specific Aim 3: To design/engineer a user-friendly graphical user interface for biologists to use to create minimum-energy DNA/configurations of open linear and spatially constrained supercoiled DNA molecules

The third specific aim of this research was to extend the modeling software first established by Dr. David Swigon to the point where it could be used by biologists without the need to understand or change the underlying code. Since we already had a program that was designed to model sequence-specific DNA with and without bound entities it seemed a natural progression to enhance that program for use by non-programmers to achieve their own goals. The application that was developed can run on various operating systems, allows an end user to create their own molecule in a variety of ways, and supplies them with an understanding of the energy and topology of their own system and the physical appearance of their molecule.

1.8 Organization of Thesis

This thesis has the following layout:

Chapter 2 focuses on the twist of supercoiling and the contributions this parameter will bring to understanding DNA topology. After the twist of supercoiling is defined and discussed, Chapter 3 introduces TwiDDL, a database focused on the twist of supercoiling in known DNA structures. The database has many unique features which allow all types

of users regardless of computer savvy to access the twist of supercoiling and many other values that characterize many of the published structures stored in the PDD/NDB. The development of TwiDDL has made it much easier to look at various structures and to see how the twist of supercoiling provides insight into the topology of the DNA molecules. Specific examples of how to use TwiDDL in the examination of a structure are given in Chapter 4. Chapter 5 presents 3DNAdesigner, a computer program that allows biologists to model naked and protein-bound DNA molecules of their own design and to have clear output of the topologies and relative energies of these generated structures. Examples of DNA structures generate with 3DNAdesigner are discussed in depth in Chapter 6. Chapter 7 sums up the work accomplished from this thesis study and gives recommendations for future work to be done to continue and develop what has already been accomplished during my tenure with the Olson group.

1.9 References

[1] C.R. Calladine, H.R. Drew, B.F. Luisi, and A.A. Travers. (2004). *Understanding DNA: The Molecule and How It Works, 3rd Edition*. Elsevier Academic Press, San Diego, CA.

[2] J.R. Wenner, M.C. Williams, I. Rouzina, and V.A. Bloomfield. (2002). Salt Dependence of the Elasticity and Overstretching Transition of Single DNA Molecules. *Biophys J.*, **82(6)**, 3160–3169.

[3] G.S. Manning. (2006). The Persistence Length of DNA Is Reached from the Persistence Length of Its Null Isomer through an Internal Electrostatic Stretching Force. *Biophys J.*, **91(10)**, 3607–3616.

[4] H.M. Lim, D.E.A. Lewis, H.J. Lee, M. Liu, and S. Adhya. (2003). Effect of Varying the Supercoiling of DNA on Transcription and Its Regulation. *Biochem.*, **42(36)**, 10718-10725.

[5] A.B. Robertson, A. Klungland, T. Rognes, and I. Leiros. (2009). DNA Repair in Mammalian Cells Base Excision Repair: The Long and Short of It. *Cell. Mol. Life Sci.*, **66(6)**, 981-993.

[6] J. Baute, and A. Depicker. (2008). Base Excision Repair and Its Role in Maintaining Genome Stability. *Crit. Revs. Biochem. Mol. Biol.*, **43(4)**, 239-276.

[7] M.A. El Hassan, and C.R. Calladine. (1997). Conformational Characteristics of DNA: Empirical Classifications and a Hypothesis for the Conformational Behaviour of Dinucleotide Steps. *Phil. Trans. R. Soc. Lond. A*, **355(1722)**, 43-100.

[8] R.E. Franklin, and R.G. Gosling. (1953). Molecular Configuration in Sodium Thymonucleate. *Nature,* **171(4356),** 740–741.

[9] M.H.F. Wilkins, A.R. Stokes, and H.R. Wilson. (1953). Molecular Structure of Deoxypentose Nucleic Acids. *Nature*, **171(4356)**, 738–740.

[10] D.A. Marvin, M. Spencer, M.H.F. Wilkins, and L.D. Hamilton. (1958). A New Configuration of Deoxyribonucleic Acid. *Nature*, **182(4632)**, 387-388.

[11] W.R. Bauer, R.A. Lund, and J.H. White. (1993). Twist and Writhe of a DNA Loop Containing Intrinsic Bends. *Proc. Natl. Acad. Sci. USA*, **90(3)**, 833-837.

[12] M.D. Frank-Kamenetskii. (1993). *Unraveling DNA: Circular DNA*. VCH Publishing, New York, NY.

[13] M.A. El Hassan, and C.R. Calladine. (1996). Structural Mechanics of Bent DNA. *Endeavour*, **20(2)**, 61-67.

[14] B.D. Coleman, D. Swigon, and I. Tobias. (2000). Elastic Stability of DNA Configurations. II. Supercoiled Plasmids with Self-Contact. *Phys. Rev. E.*, **61(1)**, 759-770.

[15] F.B. Fuller. (1971). The Writhing Number of a Space Curve. *Proc. Natl. Acad. Sci. USA*, **68(4)**, 815-819.

[16] J. Langowski, W.K. Olson, S.C. Pedersen, I. Tobias, and T.P. Westcott. (1996). DNA Supercoiling, Localized Bending and Thermal Fluctuations. *Trends Biochem. Sci.*, **21(2)**, 50.

[17] A. Vologodskii. (1992). *Topology and Physics of Circular DNA*. CRC Press, Boca Raton, FL.

[18] J.H. White. (1969). Self-Linking and the Gauss Integral in Higher Dimensions. *American J. Math.*, **91(3)**, 693-728.

[19] J.H. White. (1989). *Mathematical Methods for DNA Sequences: An Introduction to the Geometry and Topology of DNA Structure*. CRC Press, Boca Raton, FL, 225-253.

[20] B.D. Coleman, and D. Swigon. (2004). Theory of Self-Contact in Kirchhoff Rods with Applications to Supercoiling of Knotted and Unknotted DNA Plasmids. *Phil. Trans. R. Soc. Lond. A*, **362(1820)**, 1281-1299.

[21] L.A. Britton, I. Tobias, and W.K. Olson. (2009). Two Perspectives on the Twist of DNA. *J. Chem. Phys.,* **131(24)**, 245101.

[22] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 A Resolution. *J. Mol. Biol.*, **319(5)**, 1097-1113.

[23] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. (2012). RCSB PDB - Images for 1KX5. Available at *http://www.rcsb.org/pdb/explore/images.do?structureId=1KX5*.

[24] T.C. Bishop. (2009). VDNA: The Virtual DNA Plug-In for VMD. *Bioinformatics*, **25(23)**, 3187-3188.

[25] R. Lavery, K. Zakrzewska, D. Beveridge, T.C. Bishop, D. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J.H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, N. Spackova, and J. Sponar. (2010). A Systematic

Molecular Dynamics Study of Nearest-Neighbor Effects on Base Pair and Base Pair Step Conformations and Fluctuations in B-DNA. *Nucleic Acids Res.*, **38(1),** 299-313.

[26] S.A. Harris, C.A. Laughton, and T.B. Liverpool. (2008). Mapping the Phase Diagram of the Writhe of DNA Nanocircles Using Atomistic Molecular Dynamics Simulations. *Nucleic Acids Res.*, **36(1)**, 21-29.

[27] J. Curuksu, M. Zacharias, R. Lavery, and K. Zakrzewska. (2009). Local and Global Effects of Strong DNA Bending Induced During Molecular Dynamics Simulations. *Nucleic Acids Res.*, **37(11)**, 3766-3773.

[28] J. Z. Ruscio, and Alexey Onufriev. (2006). A Computational Study of Nucleosomal DNA Flexibility. *Biophys. J.*, **91(11)**, 4121-4132.

[29] K.M. Kosikov, A.A. Gorin, X. Lu, W.K. Olson, and G.S. Manning. (2002). Bending of DNA by Asymmetric Charge Neutralization: All-Atom Energy Simulations. *J. Am. Chem. Soc.*, **124(17)**, 4838-4847.

[30] B.D. Coleman, and D. Swigon. (2002). Theory of Self-Contact in DNA Molecules Modeled as Elastic Rods. *Nuovi Progressi Nella Fisica Mathematica Dall'Eredita Di Dario Graffi*, **177**, 281-295.

[31] D. Swigon, B.D. Coleman, and I. Tobias. (1998). The Elastic Rod Model for DNA and Its Application to the Tertiary Structure of DNA Minicircles in Mononucleosomes. *Biophys J.*, **74(5)**, 2515-2530.

[32] W.K. Olson, A.A. Gorin, X. Lu, L.M. Hock, and V.B. Zhurkin. (1998). DNA Sequence-Dependent Deformability Deduced from Protein–DNA Crystal Complexes. *Proc. Natl. Acad. Sci. USA*, **95(19)**, 11163–11168.

[33] M. Gō, and N. Gō. (1976). Fluctuations of an α-Helix. *Biopolymers*, **15(6)**, 1119-1127.

[34] S. Balasubramanian, F. Xu, and W.K. Olson. (2009). DNA Sequence-Directed Organization of Chromatin: Structure-Based Computational Analysis of Nucleosome-Binding Sequences. *Biophys J.,* **96(6),** 2245-2260.

[35] X. Lu and W.K. Olson. (2003). 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.,* **31(17)**, 5108-5121.

[36] X. Lu, M.A. El Hassan, and C.A. Hunter. (1997). Structure and Conformation of Helical Nucleic Acids: Analysis Program (SCHNAaP). *J. Mol. Biol.,* **273(3)**, 668-680.

[37] M.A. El Hassan, and C.R. Calladine. (1995). The Assessment of the Geometry of Dinucleotide steps in Double-Helical DNA; a New Local Calculation Scheme. *J. Mol. Biol*., **251(5)**, 648-664.

[38] W. Ge, B. Schneider, and W.K. Olson. (2005). Knowledge-Based Elastic Potentials for Docking Drugs or Proteins with Nucleic Acids. *Biophys J.*, **88(2)**, 1166-1190.

[39] J.D. Kahn, and D.M. Crothers. (1998). Measurement of the DNA Bend Angle Induced by the Catabolite Activator Protein Using Monte Carlo Simulation of Cyclization Kinetics. *J. Mol. Biol*., **276(1)**, 287-309.

[40] T. Gruger, J.L. Nitiss, A. Maxwell, E.L. Zechiedrich, P. Heisig, S. Seeber, Y. Pommier, and D. Strumberg. (2004). A Mutation in Escherichia coli DNA Gyrase Conferring Quinolone Resistance Results in Sensitivity to Drugs Targeting Eukaryotic Topoisomerase II. *Antimicrob. Agents Chemother.*, **48(12)**, 4495-4504.

[41] E.L. Zechiedrich, A.B. Khodursky, S. Bachellier, R. Schneider, D. Chen, D.M.J. Lilley, and N.R. Cozzarelli. (2000). Roles of Topoisomerases in Maintaining Steady-state DNA Supercoiling in Escherichia coli*. *J. Biol. Chem.*, **275(11)**, 8103-8113.

[42] G.R. Buck, and E.L. Zechiedrich. (2004). DNA Disentangling by Type-2 Topoisomerases. *J. Mol. Biol.*, **340(5)**, 933-939.

[43] C.R. Lopez, S. Yang, R.W. Deibler, S.A. Ray, J.M. Pennington, R.J. DiGate, P.J. Hastings, S.M. Rosenberg, and E.L. Zechiedrich. (2005). A Role for Topoisomerase III in a Recombination Pathway Alternative to RuvABC. *Mol. Microbiol.*, **58(1)**,80-101.

[44] C. Lawson, D. Swigon, K.S. Murakami, S.A. Darst, H. Berman, and R.H. Ebright. (2004). Catabolite Activator Protein: DNA Binding and Transcription Activation. *Cur. Opin. Struct. Biol.*, **14(1)**, 10-20.

[45] T. Harmer, M. Wu, and R. Schleif. (2001). The Role of Rigidity in DNA Looping-Unlooping by AraC. *Proc. Natl. Acad. Sci. USA*, **98(2)**, 427-431.

[46] W.K. Olson, D. Swigon, and B. Coleman. (2004). Implications of the Dependence of the Elastic Properties of DNA on Nucleotide Sequence. *Phil. Trans. R. Soc. Lond. A.*, **362(1820)**, 1403-1422.

[47] B.D. Coleman, W.K. Olson, and D. Swigon. (2003). Theory of Sequence-Dependent DNA Elasticity. *J. Chem. Phys.*, **118(15)**, 7127-7140.

[48] K. Shi, B. Pan, and M. Sundaralingam. (2003). Structure of a B-form DNA/RNA Chimera (dC)(rG)d(ATCG) Complexed with Daunomycin at 1.5 A Resolution. *Acta Crystallogr., Sect.D*, **59(8)**, 1377-1383.

[49] K. Shi, B. Pan, and M. Sundaralingam. (2012). RCSB PDB - Images for 1JO2. Available at *http://www.rcsb.org/pdb/explore/images.do?structureId=1JO2*.

Chapter 2: A New Twist on DNA

2.1 An Introduction to DNA Twist

This chapter presents a new way to calculate the twist of DNA and that of a complex DNA molecule. The new calculations focus on an in-depth look at the twist in supercoiled DNA, as introduced in Section 1.3, and introduce a new parameter called the twist of supercoiling. Comparisons drawn between the traditional step-parameter twist, Tw^{SP}, and the twist of supercoiling, Tw^{SC}, emphasize the significance of the new twist of supercoiling. Further, we show how Tw^{SC} is relevant to the interpretation of DNA topology, and present a few simple models to highlight the usefulness of the new parameter.

The twist of DNA is commonly calculated using the Tw^{SP} amongst the community of structural biologists. For a molecule with a closed axial curve, the writhing number or writhe, a measure of chiral distortion can be easily determined. The linking number, a topological invariant, can also be calculated for a closed double-stranded structure and will yield an integer. An easy way to find the linking number is to add the writhe to the total twist, which is obtained by summing up the twists along the chain. Up to now the Tw^{SP} was used for this calculation instead of a twist value consistent with the topological properties of supercoiled DNA.

The writhe of a closed piece of DNA can be calculated using more than one

technique. We currently use a calculation of writhe that was derived by Swigon and Tobias for a smooth continuous curve [1]. However, that model does not accurately account for the discrete nature of real base-pair steps. We have derived a new formula for the writhe that starts from a discrete sequence-dependent system. This chapter shows the derivation and validation of this discrete approach. The resultant formula is incorporated into a program to scrutinize new configurations of supercoiled DNA which then can be characterized in terms of the discrete writhe, as opposed to the writhe approximated by the formula based on a continuous curve. The combined efforts of generating a new discrete formula, integrating it into a program, and utilizing that program to analyze supercoiled DNA are pioneering our ability to successfully describe the topology of a discrete sequence of DNA base pairs, and this approach is the first of its kind.

Also of interest, is the difference in the values of Tw^{SC} and Tw^{SP} for DNA in various environments. Thus, when the DNA is relaxed, and not interacting with any proteins, we find, and understand why, the two twists for each step are close in value. On the other hand, when the DNA is interacting with a protein, such as the histones in chromatin, there are certain steps for which one observes significant differences. By considering various model structures, we find instances in which we are able to relate observed differences in the twists at a particular step to the specific nature and chirality of the structural distortion being imposed by the protein on the DNA in the region of that particular step.

Many scientists mistakenly use Tw^{SP} when dealing with the topology of DNA, rather than Tw^{SC}, which was introduced many decades ago to treat the topology of DNA described by ideal space curves. We derived a mathematical expression for Tw^{SC} which is applicable to DNA treated as a succession of discrete base pairs, not as two smooth curves winding about each other as was done in the original theory. Given the confusion among scientists about twist, it was a natural extension of our work to study the difference in the properties of these two twists. We then discovered that for DNA steps in the region of close protein-DNA interactions, often these two twists are markedly different in value. As a very relevant example, when dealing with DNA supercoiling and packaging, we illustrate the difference between the two twists by examining a key site responsible for the curvature and superhelical rise in nucleosomal DNA, shown in Chapter 4.

The derivation of Tw^{SC} for a DNA step involves subtleties, and arriving at a deep understanding of the difference in the effect of a given distortion of the structure of DNA on the values of the two twists is a worthwhile goal. It should also be noted that in the paper, which is presented in Section 2.2, we gently point out that if you are adding together the twist and writhe to get a linking number, you should not be using Tw^{SP}. The calculations we set forth in this paper should help guide scientists along the correct path for calculating and interpreting the topology of supercoiled DNA.

It should be noted that the values of both Tw^{SP} and Tw^{SC} depend on the method

used to locate the origin of an ideal base-pair plane. We employ the widely used method recommended by the IUPAC – IUBMB Joint Commission on Biochemical Nomenclature [2], which places the origins in such a way to give rise to a Tw^{SC} sensitive to structural distortions such as the buckling, or the out-of-plane "pinching", between the two bases of a base pair. Buckling can, in regards to the method we employ, move the origin of the base pair, but has no effect on the origin if other methods are used to determine the origin. Had we placed the origin following the methods developed by El Hassan and Caladine [3], by Dickerson [4], or by Bansal [5], for example, the calculated twist would have been insensitive to that particular structural change.

At the conference "From DNA–Inspired Physics to Physics–Inspired Biology" at the International Centre for Theoretical Physics (ICTP) in Trieste a participant questioned why when the twist was added to the writhe of certain closed DNA molecules, the sum was not exactly an integer, as expected. This is precisely one of the problems we address in this thesis, and one of the outcomes was the *Journal of Chemical Physics* manuscript "Two Perspectives on the Twist of DNA" as shown below in Section 2.2 [7]. A copy of the manuscript, and calculations were performed for the scientist in question. When Tw^{SC} and the writhing number, as we have defined them, were added together we showed that, indeed, the structures simulated by the scientist had integral linking numbers.

2.2 Two Perspectives on the Twist of DNA [7]

2.2.1 Abstract

Because of the double-helical structure of DNA, in which two strands of complementary nucleotides intertwine around each other, a covalently closed DNA molecule with no interruptions in either strand can be viewed as two interlocked single-stranded rings. Two closed space curves have long been known by mathematicians to exhibit a property called the linking number, a topologically invariant integer, expressible as the sum of two other quantities, the twist of one of the curves about the other, and the writhing number, or writhe, a measure of the chiral distortion from planarity of one of the two closed curves. We here derive expressions for the twist of supercoiled DNA and the writhe of a closed molecule consistent with the modern view of DNA as a sequence of base-pair steps. Structural biologists commonly characterize the spatial disposition of each step in terms of six rigid-body parameters, one of which, coincidentally, is also called the twist. Of interest is the difference in the mathematical properties between this step-parameter twist and the twist of supercoiling associated with a given base-pair step. For example, it turns out that the latter twist, unlike the former, is sensitive to certain translational shearing distortions of the molecule that are chiral in nature. Thus, by comparing the values for the two twists for each step of a high-resolution structure of a protein-DNA complex, we may be able to determine how the binding of various proteins

contributes to chiral structural changes of the DNA.

2.2.2 Introduction

The structure of a DNA molecule is often described as a succession of base pairs, each represented as a rectangular plane [6]. A knowledge of the relative locations of origins positioned within these planes and the relative orientations of the short and long axes of the rectangles allows one to determine for each pair of adjacent base pairs in the molecule – a so-called base-pair step – the numerical values of six rigid-body parameters, three translational: shift, slide, and rise, and three angular: tilt, roll, and twist [8-13].

Some forty years ago mathematicians defined a "twist" which, shortly after its introduction, was applied to DNA and used, like the step-parameter twist mentioned above, to characterize the secondary structure of the molecule [14-16]. This twist was defined as the value of a certain integral involving two continuous space curves. In the application to DNA, the structure of which at that time was often depicted in terms of space curves [17], one of the curves was taken to be the axis of the double helix, and the other one of the strands winding about this axis. One then went on to compute the twist of the DNA, a unitless scalar representing the number of times the strand wound about the helical axis.

Here we are concerned with differences in the properties of these two twists, the step-parameter twist, and the twist of the preceding paragraph, which, because of its connection with the global shape of the helical axis of a closed DNA molecule, a

plasmid, for example, we shall refer to as the twist of supercoiling. In the next section we begin by reviewing the definition of the twist of supercoiling, and its well-known connection with the writhing number and the linking number [18,19]. Then, after characterizing two space curves consistent with today's picture of DNA as a succession of discrete rectangular planes, we go on to describe a method for the computation of the twist of supercoiling for a single base-pair step. We also point out how easy it is to compute the writhe for the case of a closed molecule with an axial curve envisioned as a succession of line segments connecting the origins.

Of particular interest is the difference in properties between the step-parameter twist and the twist of supercoiling. We note that in a relaxed, undeformed configuration of a DNA molecule, the two twists are expected to be close in value for all base-pair steps. However, we find that although translations of the base pairs leave the step-parameter twist unchanged, that is not generally the case for the twist of supercoiling. It, instead, is sensitive to translational distortions that are chiral in nature. To illustrate the point, we compare the values of the two types of twist for base-pair steps in a model DNA structure having a both a bend and a shear. We find in this case a base-pair step for which the value of the step-parameter twist and the twist of supercoiling are significantly different.

2.2.3 The Twist of Supercoiling

As we pointed out above, in much of the early theoretical work describing the

equilibrium configurations of DNA, the atomistic details of the molecule were ignored, and instead, the structure of the molecule was, in fact, described in terms of two space curves. The tertiary structure of the molecule was represented, as shown in Figure 2.2.1, by the shape of a smooth space curve *C* mirroring the shape of the axis of the double helix.

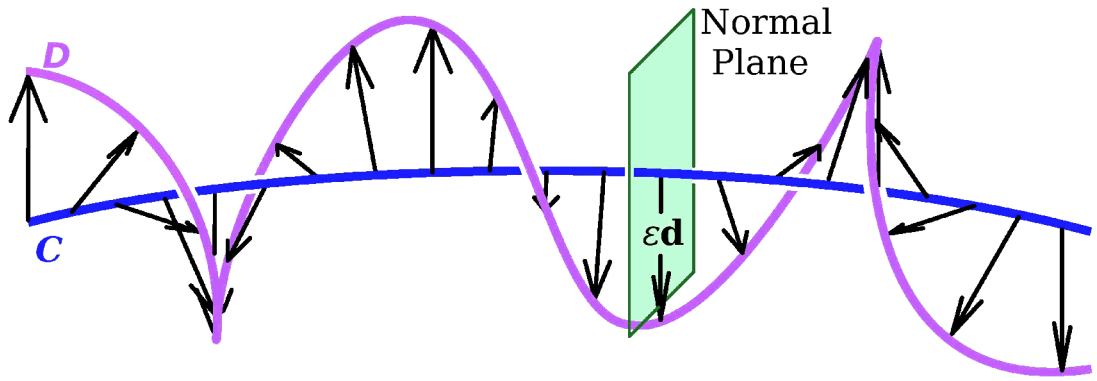


Figure 2.2.1: Schematic representation of DNA. The double-helical axis is given by curve C and one of the helical strands by curve D. For purposes of calculation of the twist of D about C, D, D is to be thought of as being traced out by the head of a vector $\varepsilon \mathbf{d}(s_C)$ *everywhere normal to the tangent vector* $\mathbf{t}_C(s_C)$.

A second curve D, identified with one of the helical strands, was that traced out by the head of a vector $\varepsilon\,\mathbf{d}(s_C)$, where ε is a constant, and $\varepsilon\,\mathbf{d}(s_C)$ is a unit vector normal to the tangent $\mathbf{t}_C(s_C)$ to C at a position having an arc length s_C along C.

The $\mathbf{t}_C(s_C)$, $\mathbf{d}(s_C)$ pairs associated with two nearby points along C having arc lengths s_C and $s_C + ds_C$ are, in general, somewhat rotated with respect to each other. The relative rotational orientation of the two pairs before and after a small change in arc length ds_C is such that there exists a single rotation of the initial pair (at s_C) by a small angle about some axis that leads to $\mathbf{t}_C(s_C + ds_C)$ and $\mathbf{d}(s_C + ds_C)$. The vector $d\,\boldsymbol{\Omega}$, the differential of the Darboux vector $\boldsymbol{\Omega}$, points in the direction of this axis of rotation and has a magnitude equal to the small angle of rotation. The change $d\,\mathbf{t}_C(s_C)$ and $d\,\mathbf{d}(s_C)$ that the vectors $\mathrm{t}_C(s_C)$ and $\mathbf{d}(s_C)$ undergo during this change in arc length ds_C is simply the cross product of $d\,\boldsymbol{\Omega}$ with the vector itself. That is, if $\mathbf{v}_C(s_C)$ stands for either $\mathbf{t}_C(s_C)$ or for $\mathbf{d}(s_C)$,

$$d\,\mathbf{v}_C(s_C) = d\,\boldsymbol{\Omega} \times \mathbf{v}_C(s_C) \qquad \textit{Eq. 2.2.1}$$

.

For the case of the tangent, Eq. 2.2.1 leads to the equation

$$d\,\boldsymbol{\Omega} = \mathbf{t}_C(s_C) \times d\,\mathbf{t}_C(s_C) + (d\,\boldsymbol{\Omega} \cdot \mathbf{t}_C(s_C))\,\mathbf{t}_C(s_C) \qquad \textit{Eq. 2.2.2}$$

.

One of the Frenet-Serret equations, the three equations relating the tangent to the principal normal $\mathbf{n}_C(s_C)$ and the binormal $\mathbf{b}_C(s_C)(=\mathbf{t}_C(s_C) \times \mathbf{n}_C(s_C))$, allows us to write

$d\mathbf{t}_C(s_C) = \mathbf{n}_C(s_C)\,\kappa_C(s_C)ds_C$ where $\kappa_C(s_C)$ is the curvature of C. The first term on the right-hand-side of Eq. 2.2.2 then becomes

$$\mathbf{t}_C(s_C)\times d\,\mathbf{t}_C(s_C)=\mathbf{b}_C(s_C)\kappa(s_C)d\,s_C \qquad \textit{Eq. 2.2.3}$$

.

The second term in that equation, the one containing the component of $d\,\boldsymbol{\Omega}$ along $\mathbf{t}_C(s_C)$, is proportional to the twist density. Replacing $\mathbf{t}_C(s_C)$ by $\mathbf{d}(s_C)$ in Eq. 2.2.2 and then taking the projection of the resulting expression for $d\,\boldsymbol{\Omega}$ along the tangent shows that

$$d\,\Omega\cdot\mathbf{t}_C(s_C)=(\mathbf{d}(s_C)\times d\,\mathbf{d}(s_C))\cdot\mathbf{t}_C(s_C) \qquad \textit{Eq. 2.2.4}$$

.

The twist $T(D,C)$, in units of number of turns, of D about a length l of C is

$$T(D,C)=\left(\frac{1}{2\pi}\right)\int_{s_{C_2}}^{s_{C_1}} d\boldsymbol{\Omega}\cdot\mathbf{t}_C(s_C) \qquad \textit{Eq. 2.2.5}$$

,

where $s_{c_2}-s_{c_1}=l$.

If the curves C, given by $\mathbf{r}_C(s_C)$, and D, given by $\mathbf{r}_D(s_D)$, are closed, the conventional twist is simply related to two other integrals [18,19], so-called Gauss integrals, the linking number $L(D,C)$

$$L(D,C)=\left(\frac{1}{4\pi}\right)\oint\!\!\oint \frac{\mathbf{t}_D(s_D)\times\mathbf{t}_C(s_C)\cdot\left(\mathbf{r}_C(s_C)-\mathbf{r}_C(s_C)\right)}{\left|\mathbf{r}_D(s_D)-\mathbf{r}_C(s_C)\right|^3}d\,s_D\,d\,s_C \qquad \textit{Eq. 2.2.6}$$

and the writhing number $W(C)$, or writhe for short,

$$W(C)=\left(\frac{1}{4\pi}\right)\oint\!\!\oint \frac{\mathbf{t}_C(s_C)\times\mathbf{t}_C(s'_C)\cdot\left(\mathbf{r}_C(s_C)-\mathbf{r}_C(s'_C)\right)}{\left|\left(\mathbf{r}_C(s_C)-\mathbf{r}_C(s'_C)\right)\right|^3}d\,s_C\,d\,s'_C \qquad \textit{Eq. 2.2.7}$$

.

The linking number is an integer, a topological invariant, equal to the number of

times that the curve *D* passes through a surface bounded by *C* [20]. (In computing this sum each pass-through is assigned a value of either +1 or –1 according to a convention consistent with the form of the Gauss integral.) This integer remains unchanged for all distortions in shape of the curves *C* and *D* as long as the curves do not intersect each other during the distortions. The writhe, a property of closed curve *C* alone, is a measure of the chiral distortion of the curve from planarity. Fuller pointed out that its value is also what one would get by averaging, over all orientations of a plane *P*, the sum of the signed self-crossings occurring in the planar curves resulting from the perpendicular projection of *C* on *P* [16].

The connection between the twist, the writhe, and the linking number mentioned above is given by the well-known equation [15] (For another derivation, see Section 2.28 (Appendix A).):

$$L(D,C)=W(C)+T(D,C) \qquad \textit{Eq. 2.2.8}$$

Some additional mathematical properties of the twist, writhe, and linking number integrals are discussed in a more recent paper [21].

2.2.4 The Twist of Supercoiling for the Multistep DNA Molecule

As we have mentioned, details of the structure of DNA are now more realistically represented, not in terms of smooth space curves, but as a sequence of base pairs. Two adjacent base pairs enclose a base-pair step. Various mathematical procedures have been

formulated for going from the atomic coordinates (determined from high-resolution structural measurements and simulations) of two associated bases to a base-pair plane such as that shown in Figure 2.2.2. Each plane contains an origin **o** from which a triad of mutually orthogonal unit vectors emanate, a short axis **s**, a long axis **l**, and a normal **n**(= $\mathbf{s} \times \mathbf{l}$). The method employed here for determining the position and orientation of these planes is the one agreed upon by DNA structural and computational biologists in 1999 at the Tsukuba Workshop on Nucleic Acid Structure and Interactions and subsequently reviewed and approved by the IUBMB Commission on Biochemical Nomenclature [22].

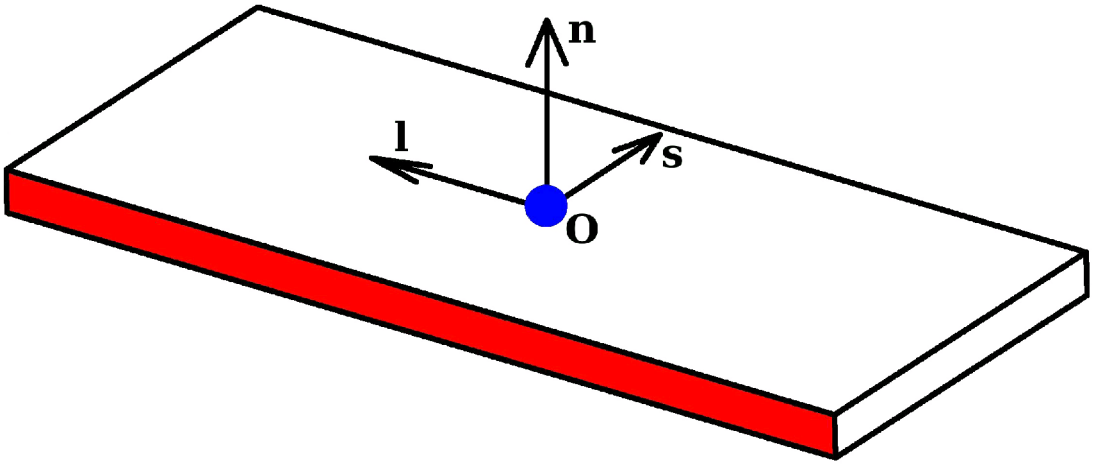


Figure 2.2.2: The vectors associated with a base-pair plane: an origin ***o****, and a mutually orthogonal triad of unit vectors, the short axis* ***s****, the long axis* ***l****, and the normal* ***n****.*

We begin the application of the concept of the twist of supercoiling, described above, to a succession of base-pair planes by imagining that there are simply line segments connecting the origins of adjacent base pairs. How then can a smooth curve *C* be chosen that gives such a picture of the DNA, a picture that seems to show a curve with a discontinuous change in its tangent at each base pair? That is, for each base pair, the i^{th} for example, the incoming line segment has a unit tangent we call $\mathbf{t}_{(i-1)}$. The tangent of the outgoing segment is $\mathbf{t}_{(i)}$. Both of these vectors are defined in terms of the origins $\mathbf{o}_{i-1}$, $\mathbf{o}_i$, and $\mathbf{o}_{i+1}$ of the i^{th} base pair and the two base pairs adjoining it

$$\mathbf{t}_{(i-1)}=\frac{\mathbf{o}_i-\mathbf{o}_{i-1}}{|\mathbf{o}_i-\mathbf{o}_{i-1}|}$$

$$\mathbf{t}_{(i)}=\frac{\mathbf{o}_{i+1}-\mathbf{o}_i}{|\mathbf{o}_{i+1}-\mathbf{o}_i|} \qquad \textit{Eq. 2.2.9}$$

(Note: Subscripts enclosed in parentheses label base-pair steps and those not enclosed in parentheses label individual base pairs.)

We can envision a limiting process, shown in Figure 2.2.3, in which we start with a curve along which the tangent changes smoothly from $\mathbf{t}_{(i-1)}$ to $\mathbf{t}_{(i)}$ in the vicinity of base pair *i* as one moves along a circular arc lying in a plane spanned by these two vectors, i.e., lying in the plane having as its normal $\mathbf{b}_i$ given by

$$\mathbf{b}_i=\frac{\mathbf{t}_{(i-1)}\times\mathbf{t}_{(i)}}{|\mathbf{t}_{(i-1)}\times\mathbf{t}_{(i)}|} \qquad \textit{Eq. 2.2.10}$$

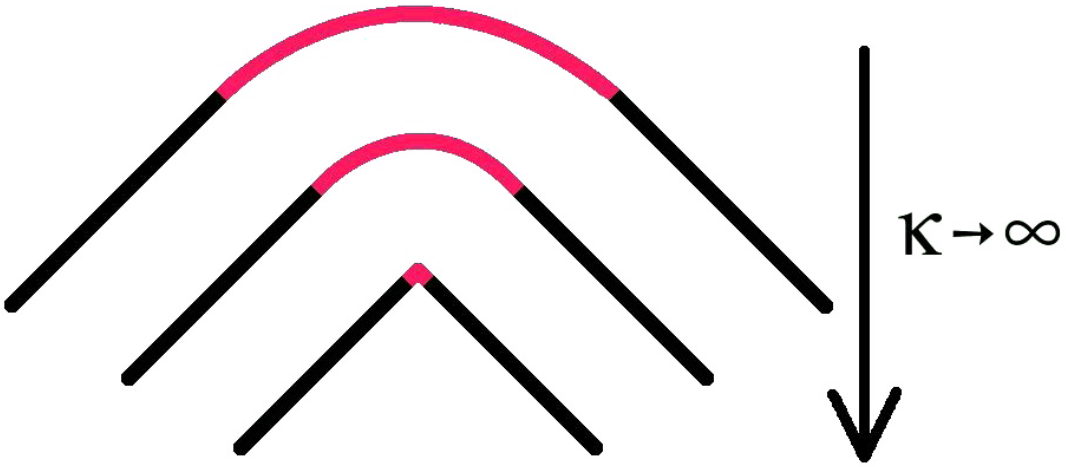


Figure 2.2.3: The passage from a smooth curve with circular segments of curvature κ to the entirely linear segments that connect the origins of the DNA base pairs extracted from a high-resolution structure.

Then the arc of the circle is allowed to decrease to zero as its curvature increases to infinity, with the tangent vectors $\mathbf{t}_{(i-1)}$ and $\mathbf{t}_{(i)}$ remaining unchanged. This process is repeated at each base pair. Thus, base pair i+1 has an incoming tangent $\mathbf{t}_{(i)}$ and an outgoing tangent $\mathbf{t}_{(i+1)}$ determined as in Eqs. 2.2.9 from the origins of base pairs i+1 and i+2. When the tangent becomes $\mathbf{t}_{(i)}$ the curve abruptly becomes a line segment of length $|\,\mathbf{o}_{i+1} - \mathbf{o}_i\,|$ directed along $\mathbf{t}_{(i)}$. At base pair i+1 the curvature abruptly becomes infinitely large again but with the tangent still changing smoothly from $\mathbf{t}_{(i)}$ to $\mathbf{t}_{(i+1)}$, along a circular arc lying in a plane with

$$\mathbf{b}_{i+1} = \frac{\mathbf{t}_{(i)} \times \mathbf{t}_{(i+1)}}{|\mathbf{t}_{(i)} \times \mathbf{t}_{(i+1)}|} \qquad \textit{Eq. 2.2.11} \quad ,$$

as its normal.

We now can define in greater detail the nature of a single step, the i^{th}. At base pair i curve C for the step begins at that point in the circular arc where the tangent, call it $\tilde{\mathbf{t}}_i$, is midway between $\mathbf{t}_{(i-1)}$ and $\mathbf{t}_{(i)}$ i.e.,

$$\tilde{\mathbf{t}}_i = \frac{\mathbf{t}_{(i-1)} + \mathbf{t}_{(i)}}{|\mathbf{t}_{(i-1)} + \mathbf{t}_{(i)}|} \qquad \textit{Eq. 2.2.12} \quad ,$$

and ends at the corresponding point at base pair i+1, i.e., where the tangent is

$$\tilde{\mathbf{t}}_i = \frac{\mathbf{t}_{(i)} + \mathbf{t}_{(i+1)}}{|\mathbf{t}_{(i)} + \mathbf{t}_{(i+1)}|} \qquad \textit{Eq. 2.2.13} \quad .$$

The unit vector $\mathbf{d}_i^{\mathbf{l}}$ at the beginning of the step we first take to be in the direction

of the projection of the long axis $\mathbf{l}_i$ on the plane containing $\mathbf{o}_i$ perpendicular to $\tilde{\mathbf{t}}_i$. The unit vector $\mathbf{d}^{\mathbf{l}}_{\boldsymbol{i+1}}$ at the end of the step points along the projection of $\mathbf{l}_{i+1}$ on the plane containing $\mathbf{o}_{i+1}$ perpendicular to $\tilde{\mathbf{t}}_{i+1}$:

$$\mathbf{d}_i^l = \frac{(\mathbf{l}_i \cdot \mathbf{b}_i)\mathbf{b}_i + (\mathbf{l}_i \cdot \tilde{\mathbf{t}}_i \times \mathbf{b}_i)(\tilde{\mathbf{t}}_i \times \mathbf{b}_i)}{|(\mathbf{l}_i \cdot \mathbf{b}_i)\mathbf{b}_i + (\mathbf{l}_i \cdot \tilde{\mathbf{t}}_i \times \mathbf{b}_i)(\tilde{\mathbf{t}}_i \times \mathbf{b}_i)|}$$

and *Eq. 2.2.14*

$$\mathbf{d}_{i+1}^l = \frac{(\mathbf{l}_{i+1} \cdot \mathbf{b}_{i+1})\mathbf{b}_{i+1} + (\mathbf{l}_{i+1} \cdot \tilde{\mathbf{t}}_{i+1} \times \mathbf{b}_{i+1})(\tilde{\mathbf{t}}_{i+1} \times \mathbf{b}_{i+1})}{|(\mathbf{l}_{i+1} \cdot \mathbf{b}_{i+1})\mathbf{b}_{i+1} + (\mathbf{l}_{i+1} \cdot \tilde{\mathbf{t}}_{i+1} \times \mathbf{b}_{i+1})(\tilde{\mathbf{t}}_{i+1} \times \mathbf{b}_{i+1})|} .$$

Let us call $\alpha^{\mathbf{l}}_i$ the angle (in radians) that $\mathbf{d}^{\mathbf{l}}_{\boldsymbol{i}}$ makes with $\mathbf{b}_i$, and $\alpha^{\mathbf{l}}_{i+1}$ the angle that $\mathbf{d}^{\mathbf{l}}_{\boldsymbol{i+1}}$ makes with $\mathbf{b}_{i+1}$, or, more precisely,

$$\cos\alpha_i^l = \mathbf{b}_i \cdot \mathbf{d}_i^l, \qquad \sin\alpha_i^l = \tilde{\mathbf{t}}_i \cdot \mathbf{b}_i \times \mathbf{d}_i^l$$

and

$$\cos\alpha_{i+1}^l = \mathbf{b}_{i+1} \cdot \mathbf{d}_{i+1}^l, \qquad \sin\alpha_{i+1}^l = \tilde{\mathbf{t}}_{i+1} \cdot \mathbf{b}_{i+1} \times \mathbf{d}_{i+1}^l \qquad \textit{Eq. 2.2.15} .$$

These vectors and angles are pictured in Figure 2.2.4

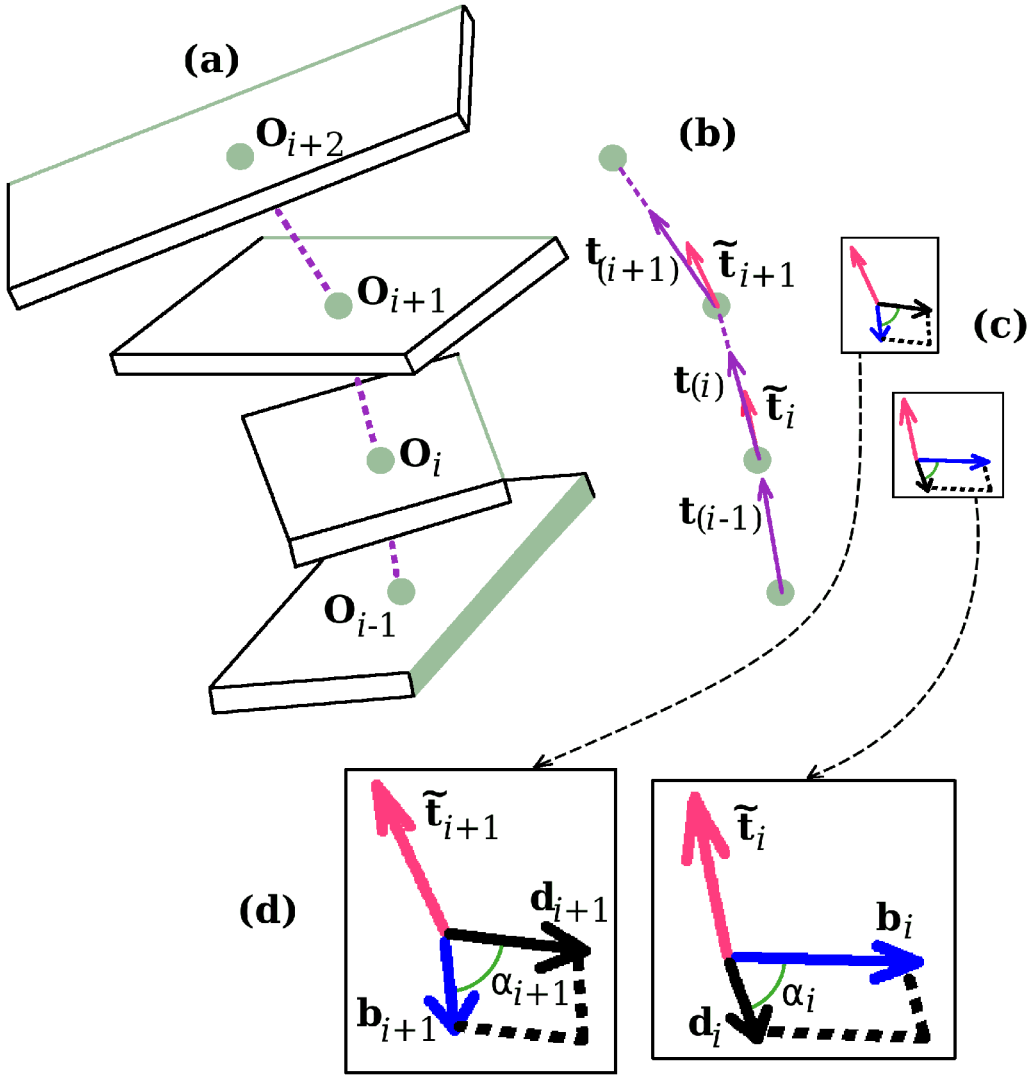


*Figure 2.2.4: The vectors involved in the calculation of the twist of supercoiling of the DNA base-pair step bounded by the i^{th} and the $(i+1)^{st}$ planes. Shown in (a), the four base-pair plane origins needed for the determination of the three vectors depicted in (b), $\boldsymbol{t}_{(i-1)}$, $\boldsymbol{t}_{(i)}$, and $\boldsymbol{t}_{(i+1)}$ which, in turn, are needed for specifying $\tilde{\mathbf{t}}_i$ and $\tilde{\mathbf{t}}_{i+1}$. These last two are normal to the two planes seen in (c). Each of these planes contains a **d** vector and a **b** vector. Two of the angles needed for the twist calculation as given by Eq. 2.2.16 are denoted in (d), an enlargement of (c).*

To determine the twist of supercoiling of the i^{th} base-pair step we carry out five rotations during which $\mathbf{d}_i^{\mathbf{l}}$ is converted to $\mathbf{d}_{i+1}^{\mathbf{l}}$. All of the vectors involved during this process can be thought of as emanating from a single point. For each of the rotations of the $\mathbf{d}^{\mathbf{l}}$ vector, the rotational axis, $d\,\Omega$, also passing through this point, will either (*a*) lie along a tangent which is not changing its direction, or (*b*) the axis will coincide with one of the two constant **b** vectors. For *a*-type rotations, according to Eq. 2.2.5, the twist of supercoiling is simply the angle of rotation while for *b*–type rotations with the axis perpendicular to the tangent, the twist is zero. One: We start with the plane containing $\mathbf{d}_i^{\mathbf{l}}$ and $\mathbf{b}_i$. The normal to this plane is $\tilde{\mathbf{t}}_i$. Now rotate $\mathbf{d}_i^{\mathbf{l}}$ about $\tilde{\mathbf{t}}_i$ until the angle between $\mathbf{d}^{\mathbf{l}}$ and $\mathbf{b}_i$ changes from $\alpha_i^{\mathbf{l}}$ to $\bar{\alpha}_{(i)}^{\mathbf{l}}$ where $\bar{\alpha}_{(i)}^{\mathbf{l}}=(\alpha_i^{\mathbf{l}}+\alpha_{i+1}^{\mathbf{l}})/2$. Two: Rotate $\mathbf{d}^{\mathbf{l}}$ about $\mathbf{b}_i$ until the normal to the plane containing these vectors changes from $\tilde{\mathbf{t}}_i$ to $\mathbf{t}_{(i)}$. At this point the plane contains $\mathbf{b}_{i+1}$ as well as $\mathbf{b}_i$. Three: Rotate $\mathbf{d}^{\mathbf{l}}$ about $\mathbf{t}_{(i)}$ until $\mathbf{d}^{\mathbf{l}}$ makes an angle of $\bar{\alpha}_{(i)}^{\mathbf{l}}$ with $\mathbf{b}_{i+1}$. Four: Rotate $\mathbf{d}^{\mathbf{l}}$ about $\mathbf{b}_{i+1}$ until the normal changes from $\mathbf{t}_{(i)}$ to $\tilde{\mathbf{t}}_{i+1}$. Five: Rotate $\mathbf{d}^{\mathbf{l}}$ about $\tilde{\mathbf{t}}_{i+1}$ until the angle between $\mathbf{d}^{\mathbf{l}}$ and $\mathbf{b}_{i+1}$ changes from $\bar{\alpha}_{(i)}^{\mathbf{l}}$ to $\bar{\alpha}_{i+1}^{\mathbf{l}}$. The $\mathbf{d}^{\mathbf{l}}$ vector has now become $\mathbf{d}_{i+1}^{\mathbf{l}}$. There is a nonzero twist associated with the *a*-type rotations one, three, and five, since, for each of these steps, the vector $d\,\Omega$ is directed along the tangent. Rotations two and four, on the other hand, since the axis of rotation is perpendicular to the tangent, are *b*–type rotations with zero twist. In the case of rotation one and five the twist angle is $\Delta\alpha_{(i)}^{\mathbf{l}}/2$ where $\Delta\alpha_{(i)}^{\mathbf{l}}=\alpha_{i+1}^{\mathbf{l}}-\alpha_i^{\mathbf{l}}$. For rotation

three the angle of rotation, which we call $\beta_{(i)}$, has as its cosine and sine: $\beta_{(i)}=\mathbf{b}_i\cdot\mathbf{b}_{i+1}$ and $\beta_{(i)}=\mathbf{t}_{(i)}\cdot\mathbf{b}_i\times\mathbf{b}_{i+1}$. The twist of supercoiling of the base-pair step associated with these rotations $T^{\mathbf{l}}_{(i)}$ (in units of number of turns) is thus

$$T^{\mathbf{l}}_{(i)}=\frac{\Delta\alpha^{\mathbf{l}}_{(i)}+\beta_{(i)}}{2\pi} \qquad \textit{Eq. 2.2.16}$$

The beginning and ending **d** vectors in the rotations leading to the twist given by Eq. 2.2.16 are those defined in Eqs. 2.2.14 and 2.2.15, the unit vectors in the direction of the projections of the long axes $\mathbf{l}_i$ and $\mathbf{l}_{i+1}$ onto the planes perpendicular to $\tilde{\mathbf{t}}_i$ and $\tilde{\mathbf{t}}_{i+1}$, respectively.

For an unnicked closed DNA molecule with n_B base pairs, Eq. 2.2.8 tells us that

$$\sum_{i=1}^{n_B} T^{\mathbf{l}}_{(i)}=L(D,C)-W(C) \qquad \textit{Eq. 2.2.17}$$

If one now were to calculate the $T_{(i)}$'s for the same five rotations but using as the **d**-vectors those derived from the short axes $\mathbf{s}_i$ and $\mathbf{s}_{i+1}$, we would find that for a single step the two twists, call them, $T^{\mathbf{l}}_{(i)}$ and $T^{\mathbf{s}}_{(i)}$, would be somewhat different. The sum $\sum_{i=1} T^{\mathbf{s}}_{(i)}$, however, would have exactly the same value as that obtained before, that given by the right-hand-side of Eq. 2.2.17. We define the twist of supercoiling of the step, $T_{(i)}$, as the average of , $T^{\mathbf{l}}_{(i)}$ and $T^{\mathbf{s}}_{(i)}$. Clearly, Eq. 2.2.17 is also satisfied for $T_{(i)}$'s so defined. The average of $\alpha^{\mathbf{l}}_{(i)}$ and $\alpha^{\mathbf{s}}_{(i)}$ will be denoted as α_i.

In the calculation of the twist of supercoiling for the case of a DNA molecule with an open helical axis, for the initial step we take $\tilde{\mathbf{t}}_1$ to be in the direction of $\mathbf{t}_{(1)}$, and for the final step we take $\tilde{\mathbf{t}}_{n_B}$ in the direction of $\tilde{\mathbf{t}}_{(n_B-1)}$.

The step-parameter twist of the base pair step can be derived in a similar way. In this case the **d** vectors are the long and short axes themselves. Carrying out the rotations starting with $\mathbf{l}_i$ and ending with $\mathbf{l}_{i+1}$ gives the same value for the twist as starting with $\mathbf{s}_i$ and ending with $\mathbf{s}_{i+1}$. There is thus no need to do the averaging indicated in the definition of the twist of supercoiling. Figure 2.2.5 shows the vectors and angles that play a role in the determination of the step-parameter twist.

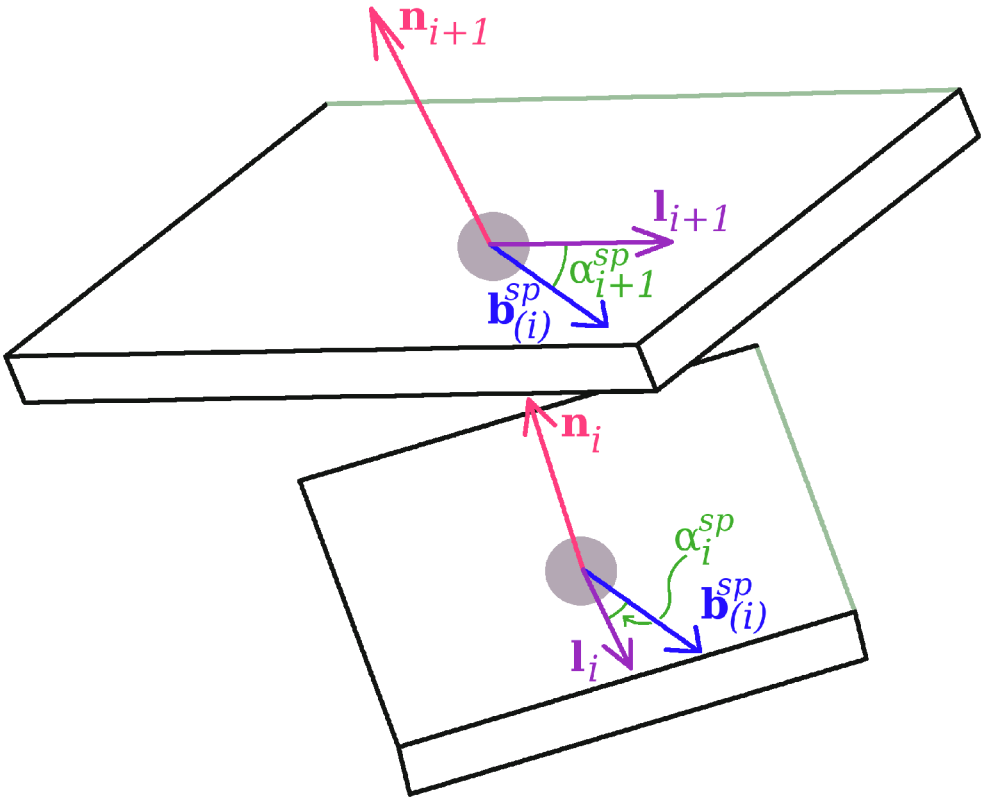


Figure 2.2.5: The vectors needed for the calculation of the step-parameter twist of the same step shown in the previous figure. Here knowledge of the direction of the two normals allows the determination of the single vector $\mathbf{b}^{sp}_{(i)}$*, which lies in each of the two base-pair planes. Also indicated are the two angles needed for the use of Eq. 2.2.18 for the twist calculation.*

The analog of $\tilde{\mathbf{t}}_i$ and $\tilde{\mathbf{t}}_{i+1}$ are the normals $\mathbf{n}_i$ and $\mathbf{n}_{i+1}$, and $\mathbf{b}_i$ and $\mathbf{b}_{i+1}$ become the single vector, $\mathbf{b}_{(i)}^{sp}=\mathbf{n}_i\times\mathbf{n}_{i+1}/|\mathbf{n}_i\times\mathbf{n}_{i+1}|$. The step-parameter twist $T_{(i)}^{\mathbf{sp}}$ of the step (in units of number of turns) thus, is, simply,

$$T_{(i)}^{sp}=\frac{\Delta\alpha_{(i)}^{sp}}{2\pi} \qquad \textit{Eq. 2.2.18}\ ,$$

where $\Delta\alpha_{(i)}^{sp}$ is the difference between the angle α_{i+1}^{sp} that $\mathbf{l}_{i+1}$ (or $\mathbf{s}_{i+1}$) makes with $\mathbf{b}_{(i)}^{sp}$ and the angle α_i^{sp} that $\mathbf{l}_i$ (or $\mathbf{s}_i$) makes with $\mathbf{b}_{(i)}^{sp}$.

2.2.5 The Write of the Closed Multistep DNA Molecule

The shape of the closed axial curve of the DNA molecules we are dealing with, consistent with the smooth space curves we have described, is a succession of line segments connecting the origins of the base pairs. A method for computing the writhe of such a segmented closed curve was elegantly put forth many years ago by Levitt [23]. Here we get Levitt's result using a different approach. We first note that for this type of curve, the Gauss integral (Eq. 2.2.8) for the writhe for a molecule with n_B base pairs, and therefore, n_B base pair steps, takes the form of a sum of contributions $w_{(i)(j)}$ of all pairs of base-pair steps,

$$W(C)=\sum_{i,j} w_{(i,j)} \qquad \textit{Eq. 2.2.19}$$

with

$$w_{(i,j)}=\left(\frac{1}{2\pi}\right)\iint \frac{\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}\cdot\mathbf{r}_{(i,j)}(s_{(i)},s_{(j)})}{|\mathbf{r}_{(i,j)}(s_{(i)},s_{(j)})|^3}\,ds_{(i)}ds_{(j)} \qquad \textit{Eq. 2.2.20}\ ,$$

where

$$\begin{aligned}\mathbf{r}_{(i,j)}(s_{(i)},s_{(j)}) &= \mathbf{r}_{(i)}(s_{(i)}) - \mathbf{r}_{(j)}(s_{(j)}) \\ &= \mathbf{X}_{(i,j)} + \mathbf{t}_{(i)}s_{(i)} - \mathbf{t}_{(j)}s_{(j)}\end{aligned} \qquad \textit{Eq. 2.2.21}$$

is a vector which points from a point on the j^{th} line segment to a point on the i^{th}. The constant vector $\mathbf{X}_{(i,j)}$ is perpendicular to each of the tangents. Its magnitude is thus the distance of closest approach of the segments. Terms involving a base-pair step with itself, and terms involving pairs of adjacent steps are zero.

Given Eq. 2.2.21, Eq. 2.2.20 can be cast into the form

$$w_{(i,j)} = \left(\frac{1}{2\pi}\right)\int_{s_{(i)}(1)}^{s_{(i)}(2)}\int_{s_{(j)}(1)}^{s_{(j)}(2)} \frac{\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}\cdot\mathbf{X}_{(i,j)}}{\left|\mathbf{X}_{(i,j)} + \mathbf{t}_{(i)}\mathbf{s}_{(i)} - \mathbf{t}_{(j)}\mathbf{s}_{(j)}\right|^3} ds_{(i)}ds_{(j)} \qquad \textit{Eq. 2.2.22}\ ,$$

where $s_{(i)}(1)$, $s_{(i)}(2)$, $s_{(j)}(1)$, and $s_{(j)}(2)$ denote the arc lengths of the endpoints of the segments.

The three vectors $\mathbf{t}_{(i)}$, $\mathbf{t}_{(j)}$, and $\mathbf{r}_{(i,,j)}(s_{(i)}, s_{(j)})$ determine a dihedral angle $\mu_{(i,j)}(s_{(i)}, s_{(j)})$ we define as follows in terms of its sine and cosine

$$\begin{aligned}\sin\mu_{(i,j)} &= -\frac{\mathbf{r}_{(i,j)}\cdot(\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)})\times(\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)})}{|\mathbf{r}_{(i,j)}||\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)}||\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)}|} \\ &= -\frac{|\mathbf{r}_{(i,j)}|(\mathbf{r}_{(i,j)}\cdot\mathbf{t}_{(i)}\times\mathbf{t}_{(j)})}{|\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)}||\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)}|}, \\ \cos\mu_{(i,j)} &= \frac{(\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)})\cdot(\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)})}{|\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)}||\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)}|}\end{aligned} \qquad \textit{Eq. 2.2.23}\ .$$

It follows from its definition that $\mu_{(i,j)}(s_{(i)}, s_{(j)})$ is the angle between the vectors normal to two planes, the one spanned by $\mathbf{r}_{(i,,j)}(s_{(i)}, s_{(j)})$ and $\mathbf{t}_{(i)}$, and the one spanned by $\mathbf{r}_{(i,,j)}(s_{(i)}, s_{(j)})$

and $\mathbf{t}_{(j)}$.

In Section 2.2.9 (Appendix B), it is shown that the integrand in the integral appearing in Eq. 2.2.22 is equal to $-\partial^2\mu_{(i,j)} / \partial s_{(i)}\partial s_{(j)}$ so that the contribution to the writhe of base-pair step i and base-pair step j, $w_{(i,j)}$, is simply related to the four dihedral angles associated with the endpoints of the steps, i.e.,

$$w_{(i,j)}=\left(\frac{1}{2\pi}\right)\Big(-\mu_{(i,j)}(s_{(i)}(2),s_{(j)}(2))+\mu_{(i,j)}(s_{(i)}(2),s_{(j)}(1)) \\ -\mu_{(i,j)}(s_{(i)}(1),s_{(j)}(1))+\mu_{(i,j)}(s_{(i)}(1),s_{(j)}(2))\Big)$$

Eq. 2.2.24 .

We note that $w_{(i,j)}$ has the same sign as $\mathbf{t}_{(i)} \times \mathbf{t}_{(j)} \cdot \mathbf{X}_{(i,j)}$.

2.2.6 Comparison of the Supercoiling and Step-Parameter Twists

Like $T^{sp}_{(i)}$ of Eq. 2.2.18, the twist of supercoiling of a step in the form as expressed in Eq. 2.2.16 is independent of the direction of propagation along the DNA molecule. But whereas $T^{sp}_{(i)}$ does not depend at all on the properties of the two steps adjacent to the i^{th} base-pair steps, $T_{(i)}$ depends on the previous step (i–1), and on the following step (i+1). In particular, $\tilde{\mathbf{t}}_i$ depends on the origin $\mathbf{o}_{i-1}$ and $\tilde{\mathbf{t}}_{i+1}$ depends on $\mathbf{o}_{i+1}$. For Eq. 2.2.17 to be satisfied for unnicked closed DNA molecules, the twist of supercoiling must depend on the way the tangent vectors are changing. In the case of the step-parameter twist, on the other hand, the sum $\sum_{i=1} T^{sp}_{(i)}$ has no particular significance.

Because in relaxed DNA, the normals of the base pairs and the associated $\tilde{\mathbf{t}}$

vectors are approximately collinear, for each step we expect the two twists to be close in value. When the base pairs are forced to undergo translations of a chiral nature, however, unlike the step-parameter twist, which is unaffected by pure translations, we expect the twist of supercoiling to change.

To demonstrate how differences in values of the step-parameter twist and the twist of supercoiling can arise we consider the simple model depicted in Figure 2.2.6

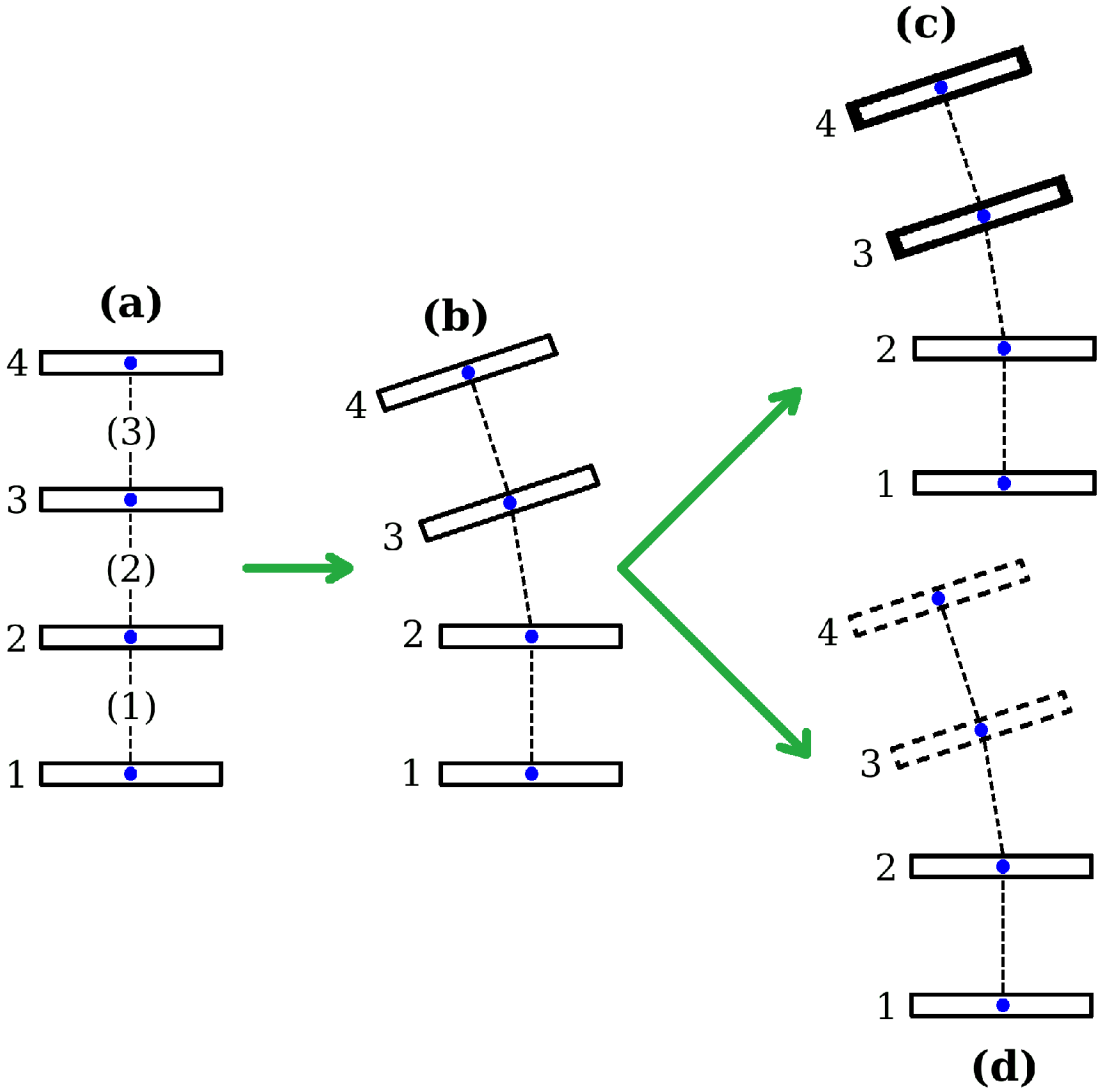


Figure 2.2.6: Construction of a model DNA structure characterized by a chiral deformation. Image labeled (a) shows four equally spaced and parallel base-pair planes having their origins lying on a line. The sequence of base-pair planes in (b) depicts the structure after the bend described in the text is introduced. The four origins are still coplanar, and the viewing direction is chosen to be normal to this plane. A translation of base pairs 3 and 4 as a single unit along the viewing direction, depending on the direction of the motion, results either in (c), a structure with a right-handed jog, or (d), one with a left-handed jog.

We start with a DNA molecule with four base pairs having origins that are equally spaced and collinear, Figure 2.2.6(a). In addition, the base-pair planes are oriented so that their normals coincide with the **t**-vectors. Here, for each of the three steps, the twist of supercoiling equals the step-parameter twist. To change the structure to that labeled (b), base pairs 3 and 4 are first rotated as a unit by 9° about an axis lying in base pair 2 and passing through the origin of that base pair, and then rotated as a unit by 9° about an axis parallel to the first axis but lying in base pair 3 and passing through the origin of that base pair. After this total bend of 18° is introduced, the origins of the base pairs are no longer collinear, but they remain coplanar. (In the figure the common plane of the origins is taken to be the plane of the page.) A calculation shows that for the new structure neither the step-parameter twist nor the twist of supercoiling for each of the three steps has changed in value so that the equality seen in (a) of the two types of twist for each step carries over to (b). Finally, base pairs 3 and 4 are translated in a direction perpendicular to the plane on which they had been lying in (b) either up to give (c), or by the same amount down to give (d). This shearing motion has no effect on the step-parameter twists, but the connection between the twist of supercoiling and the writhing number in closed molecules given by Eq. 2.2.17 demands that the chirality that now characterizes the positions of the origins in the structures (c) and (d) lead to a change in the value of the twist of supercoiling. We find that if the size of the translational shear is taken to be half

as large as the spacing between the base pairs in (a), that the twist of supercoiling of the middle step, (2), of the structure with the right-handed jog, (c), is 4.3° greater in value than the step-parameter twist for that step, and 4.3° lower in value for the middle step of the structure with the left-handed jog, (d). Shearing of this magnitude is typical of values seen in high resolution structures. For the two steps, (1) and (3), adjacent to the middle step in both (c) and (d), the two types of twist are now very close in value but not exactly equal.

We have pointed out that there is not one, but various procedures that have been proposed for going from the measured atomic coordinates of two associated bases to the ideal base-pair plane containing an origin **o** and the vectors **s** and **l**. The approach employed here is the one agreed upon by the international structural and computational biology community [22]. The question arises – how might the results reported here change had we used a different method. For one thing, the values of the step-parameter twists would not have changed by any significant amount. These twists are defined in terms of the vectors **s**, **l**, and **n** only, and it turns out that **s**, **l**, and **n** are little dependent on which method is chosen [12]. The twists of supercoiling, however, also depend on the position of the origins. In the absence of distortions such as buckling, which change the angle between the normals of the bases when viewed in projection along a plane perpendicular to the short, pseudo-dyad axis of the base pair, the various methods used to

describe the spatial arrangements of DNA bases and base pairs introduce a coordinate frame with an origin located in the midst of the hydrogen – bonding of the base pairs. In this case, the values of the twist of supercoiling, like those of the step-parameter twist, would be method-independent.

The origins introduced in three analysis schemes — NUPARM from Bansal and associates [24,25], CEHS from El Hassan and Calladine [10], and FREEHELIX from Dickerson [11] — lie midway between the positions of two carbon atoms, one on each base (C6 on pyrimidine and C8 on purine) near the DNA backbone. For these three methods, buckling has the effect of moving the origin outside of the hydrogen-bonded region of the base pair. By contrast, the origins defined here would remain near the center of the purine-pyrimidine complex.

In Figure 2.2.7 an example is given of a protein-induced structural distortion that would not have changed the twist of supercoiling for a DNA step had the NUPARM, CEHS, or FREEHELIX schemes been employed (origin shown as a red dot), but do produce such a change with the method we use here (origin shown as a blue dot). The structure labeled (a) in Figure 2.2.7 is identical to structure (d) in Figure 2.2.6, the one with the left-handed jog. During the passage to Figure 2.2.7(b), base-pair 3 is subjected to a buckling-type distortion. Before the distortion the two origins are superimposed, but after the distortion, they are separated. For the placement of the two origins shown, we find that our method gives a twist of supercoiling for the middle step in the buckled

structure Figure 2.2.7(b) greater than what it was in structure Figure 2.2.7(a). For the other method, the twist of supercoiling is insensitive to this particular structural change, one commonly found when a protein inserts an amino acid side-group between two base pairs [26].

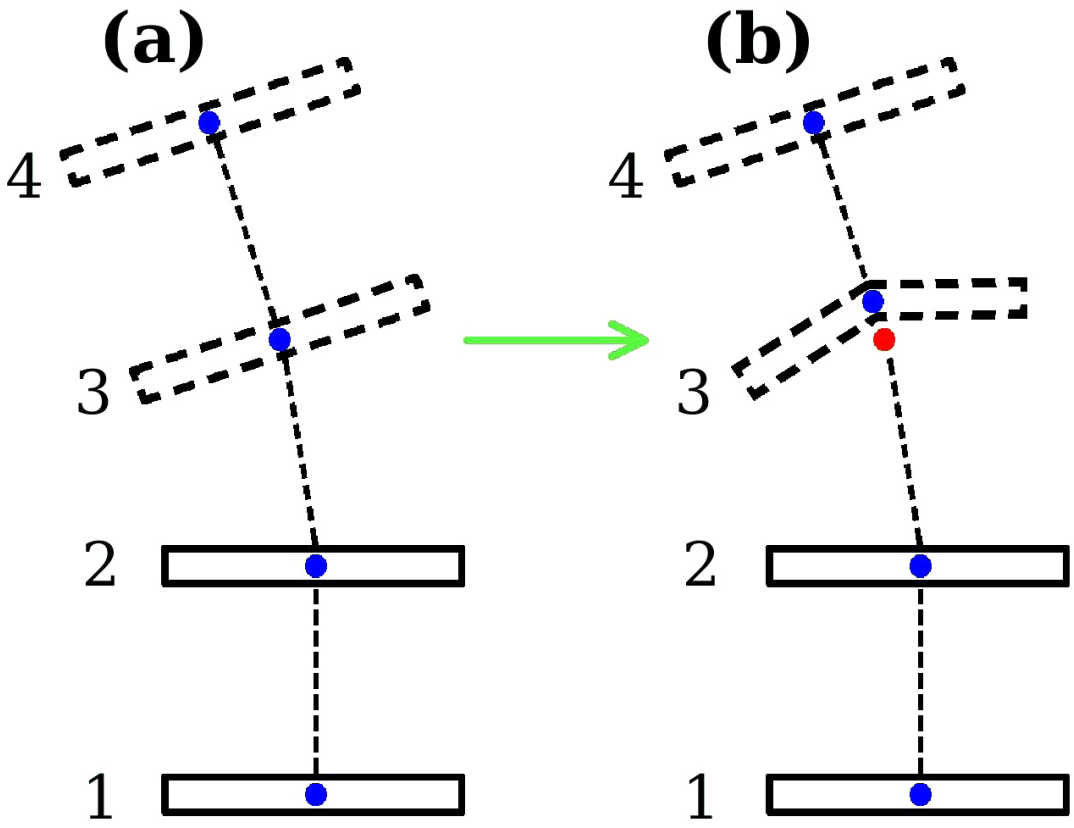


Figure 2.2.7: Structural deformation of DNA leading to a twist of supercoiling change dependent on the method used to determine the position of the origin, the method used here, or that in which the origin lies on a line connecting two carbon atoms on the bases. In structure (a), the same as structure (d) in Figure 2.2.6, both methods lead to an origin for all four base pairs located in the same position. However when base pair 3 is buckled as shown to form structure (b), the origin determined by our method (blue dot) moves, but that of the other method (red dot) does not. One then observes a method-dependent twist of supercoiling for the middle step.

2.2.7 Summary

We have defined a twist of supercoiling consistent with the high-resolution atomic structure of DNA starting from the original definition of the twist of one smooth space curve about another. The twist of supercoiling is to be distinguished from one of six step parameters, also called the twist, now in common usage to describe DNA structure. The twist of supercoiling, unlike the step-parameter twist, is connected to the topological invariant, the linking number, and the writhing number of closed DNA molecules. Given this association, the twist of supercoiling, as we show, must be sensitive to chiral structural distortions. We give an example of a chiral distortion that has no effect on the step-parameter twist. In fact, in future work, we plan to compare the values of these two twists for DNA steps in a selection of protein-DNA complexes to gain insight into details of the distortion that the proteins may be inducing.

2.2.8 Appendix A

We begin the derivation of Eq. 2.2.8 by expressing the connection between curve *D* and curve *C*, by writing for $\mathbf{r}_D$,

$$\mathbf{r}_D(s_D)=\mathbf{r}_C(s'_C)+\varepsilon\,\mathbf{d}(s'_C) \qquad \textit{Eq. 2.2.25}$$,

where ε is a constant and $\mathbf{d}(s'_C)$ is a unit vector perpendicular to the tangent $\mathbf{t}_C(s'_C)$ to curve *C*, i.e., $\mathbf{d}(s'_C) \cdot \mathbf{t}_C(s'_C) = 0$. Eq. 2.2.25 implies that the unit tangent $\mathbf{t}_D(s_D)$ to curve *D* is of the form

$$\mathbf{t}_D(s_D)=k\left(\mathbf{t}_C(s'_C)+\varepsilon\frac{d\,\mathbf{d}(s'_C)}{ds'_C}\right) \qquad \textit{Eq. 2.2.26}$$

where k is the reciprocal of the magnitude of the vector contained in the parentheses. These two expressions are substituted into the integrand appearing in the Gauss integral form for the linking number, Eq. 2.2.6. We then allow to approach zero. Beyond a certain point in this limiting process curve C and curve D no longer can intersect each other, and thereafter the linking number remains unchanged [16]. The factor k approaches one as ε approaches zero. We also find that, in this limit, the terms in the integrand that are linear in ε are zero. Thus one can write for the linking number

$$L(D,C)=W(C)+\lim_{\varepsilon\to 0}\left(\frac{\varepsilon^2}{4\pi}\right)\oint\!\!\oint\frac{\mathbf{d}(s_C)\times\dfrac{d\,\mathbf{d}(s_C)}{ds_C}\cdot\mathbf{t}_C(s'_C)}{\left(|\mathbf{r}_C(s_C)-\mathbf{r}_C(s'_C)|^2+\varepsilon^2\right)^{3/2}}ds'_C\,ds_C \qquad \textit{Eq. 2.2.27}$$

where $W(C)$ is the writhing number as given in Eq. 2.2.7. Furthermore, because

$$\lim_{\varepsilon\to 0}\frac{\varepsilon^2}{\left(|\mathbf{r}_C(s_C)-\mathbf{r}_C(s'_C)|^2+\varepsilon^2\right)^{3/2}}=\lim_{\varepsilon\to 0}\frac{\varepsilon^2}{\left((s_C-s'_C)^2+\varepsilon^2\right)^{3/2}}$$
$$=2\delta(s_C-s'_C) \qquad \textit{Eq. 2.2.28}\ ,$$

where $\delta(s_C - s'_C)$ is a Dirac delta function, the integration over s'_C in the second term of Eq. 2.2.27 can be readily carried out:

$$\lim_{\varepsilon\to 0}\left(\frac{\varepsilon^2}{4\pi}\right)\oiint \frac{\mathbf{d}(s_C)\times\frac{d\mathbf{d}(s_C)}{ds_C}\cdot\mathbf{t}_C(s'_C)}{\left(|\mathbf{r}_C(s_C)-\mathbf{r}_C(s'_C)|^2+\varepsilon^2\right)^{3/2}}ds'_C\,ds_C$$
$$=\left(\frac{1}{2\pi}\right)\oiint \delta(s_C-s'_C)\mathbf{d}(s_C)\times\frac{d\mathbf{d}(s_C)}{ds_C}\cdot\mathbf{t}_C(s'_C)\,ds'_C\,ds_C$$
$$=\left(\frac{1}{2\pi}\right)\oint \mathbf{d}(s_C)\times\frac{d\mathbf{d}(s_C)}{ds_C}\cdot\mathbf{t}_C(s_C)ds_C \qquad \textit{Eq. 2.2.29.}$$

This completes the proof of Eq. 2.2.8.

2.2.9 Appendix B

We begin by deriving an expression for $-\partial\mu_{(i,j)}/\partial s_{(j)}$. Since, for example,

$$\frac{\partial\cot\mu}{\partial s_{(j)}}=\frac{d\cot\mu}{d\mu}\frac{\partial\mu}{\partial s_{(j)}} \qquad \textit{Eq. 2.2.30}\ ,$$

we see that

$$-\frac{\partial\mu}{\partial s_{(j)}}=\sin^2\mu\frac{\partial\cot\mu}{\partial s_{(j)}} \qquad \textit{Eq. 2.2.31}\ .$$

Given the definition of the dihedral angle (Eq. 2.2.23), and the fact that $\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}\cdot\mathbf{r}_{(i,j)}$ is independent of $s_{(i)}$ and $s_{(j)}$, this last equation can be rewritten as

$$-\frac{\partial\mu}{\partial s_{(j)}}=-\frac{|\mathbf{r}_{(i,j)}|^2\left(\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}\cdot\mathbf{r}_{(i,j)}\right)}{|\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)}|^2|\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)}|^2}$$
$$\times\frac{\partial}{\partial s_{(j)}}\left(\frac{(\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)})\cdot(\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)})}{|\mathbf{r}_{(i,j)}|}\right) \qquad \textit{Eq. 2.2.32}\ .$$

After noting that the explicit dependence of $\mathbf{r}_{(i,j)}$ on $s_{(i)}$ and $s_{(j)}$ as given in Eq. 2.2.21 leads to the fact that

$$|\mathbf{r}_{(i,j)}|^2=|\mathbf{X}_{(i,j)}|^2+s_{(i)}^2+s_{(j)}^2-2(\mathbf{t}_{(i)}\cdot\mathbf{t}_{(j)})s_{(i)}s_{(j)}$$

$$\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}\cdot\mathbf{r}_{(i,j)}=\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}\cdot\mathbf{X}_{(i,j)}$$

$$|\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)}|^2=|\mathbf{X}_{(i,j)}|^2+|\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}|^2 s_{(j)}^2$$

$$|\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)}|^2=|\mathbf{X}_{(i,j)}|^2+|\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}|^2 s_{(i)}^2$$

$$(\mathbf{r}_{(i,j)}\times\mathbf{t}_{(j)})\cdot(\mathbf{t}_{(i)}\times\mathbf{r}_{(i,j)})=-\left((\mathbf{t}_{(i)}\cdot\mathbf{t}_{(j)})|\mathbf{X}_{(i,j)}|^2+|\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}|^2 s_{(i)}s_{(j)}\right) \qquad \textit{Eq. 2.2.33}$$

We find, after performing the indicated differentiation with respect to $s_{(j)}$ in Eq. 2.2.32 and simplifying the resulting expression, that

$$-\frac{\partial\mu}{\partial s_{(j)}}=\frac{\left(\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}\cdot\mathbf{X}_{(i,j)}\right)(s_{(i)}-(\mathbf{t}_{(i)}\cdot\mathbf{t}_{(j)})s_{(j)})}{(|\mathbf{X}_{(i,j)}|^2+|\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}|^2 s_{(j)}^2)|\mathbf{r}_{(i,j)}|} \qquad \textit{Eq. 2.2.34}$$

Differentiating Eq. 2.2.34 again, this time with respect to $s_{(i)}$, yields

$$-\frac{\partial^2\mu_{(i,j)}}{\partial s_{(i)}\partial s_{(j)}}=\frac{\mathbf{t}_{(i)}\times\mathbf{t}_{(j)}\cdot\mathbf{X}_{(i,j)}}{|\mathbf{r}_{(i,j)}|^3} \qquad \textit{Eq. 2.2.35}$$

2.3 Calculate the Twist of a Known Structure, the User-Friendly Way: 3DNATwSC

Once we defined Tw^{SC}, it became clear that there was a need for a user friendly program, which we called 3DNATwSC, that would allow others to calculate the twist that is defined in Section 2.2.

In order to compute Tw^{SC} one needs base-pair-slab data, such as the origins, short axis, long axis, and normal axis, and all the variables that follow from calculations based

upon the base-pair-slab data. This assumes, however, that the user has already either (a) converted the atomic coordinates of a Protein Data Bank (PDB) [27] file into base pair slabs and interpreted those slabs to display usable data including the long, short, and normal axes and the origin of the base pair; or (b) converted base-pair-step parameters that are found in either an existing structure, or one that the user is interested in creating into the base-pair-slab data described above.

The ease of calculation of Tw^{SC} brought about by 3DNATwSC opened up the possibility of looking for links between this twist and all of the six step parameters. These relationships could be evaluated not only as a one-to-one linear relationship, but also in terms of any coupling between Tw^{SC} and two of the step parameters. Unfortunately, we have not found a correlation between Tw^{SC} to any of the step parameters in any combination. We did hope to see a correlation between them because we hoped it would have been useful to see how, for example, the shift in a molecule might affect the twist. The inability to identify a direct correlation may stem from the fact that the data analyzed were confined to single steps, and further analysis along the molecule may be needed to discover a connection.

2.4 Seven Closed Loop Structures and Their Effect on the Twist of Supercoiling

2.4.1 Model Background and Assumptions

Before finding Tw^{SC} in a real DNA structure stored in the Nucleic Acid Database (NDB) [28], we studied a simplified rendition of a closed molecule. However, at the same time, we needed a structure greater than four base pairs, already examined in Section 2.2. Our group had been looking at models of an 80-base-pair closed system made up of ideal base pairs.

In this case, ideal means that the base-pair-step parameters have values of ideal B-DNA, assumed to have 10.5 base-pairs per turn with 3.4Å of vertical displacement, or rise, between the base pairs. The assumed twist of 10.5 base pairs per turn comes from experimental studies in solution [29,30]. The choice of a 3.4Å rise is based on the spacing seen in early fiber diffraction studies and later borne out in the base-pair steps of crystal structures[13]. When we think of ideal DNA, we mean unbound, naked DNA segments. Ideal DNA should not have any external forces acting on it. Therefore, when we speak of ideal DNA, we also know that such a creature cannot really exist in vivo, and if it did, this would be an anomaly. That is because, in a cell, DNA will most likely be subject to changes in salinity, interactions from other DNA, protein interactions, drug or other ligand interactions, or any other errata that exists inside a cell or in an in vitro situation.

I would like to point out that there are only two non-zero values for the step parameters for ideal B-DNA. The first is rise which, as seen in Figure 2.4.1.1, is the perpendicular displacement between the planes of a base pair step. The rise is located along the normal of the base-pair step in B-DNA. The second non-zero step parameter, which is defined with respect to the normal of the base pair step, is Tw^{SP}.

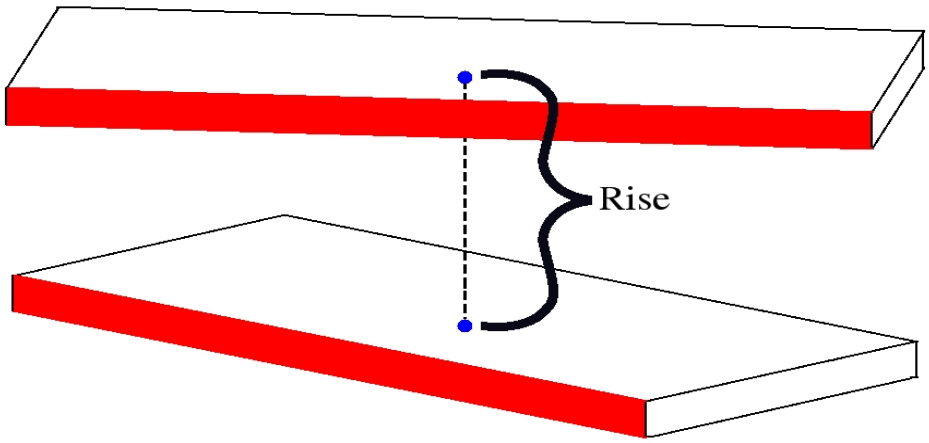


Figure 2.4.1.1: Two base pairs (one base-pair step) of ideal B-DNA. The dashed line between the bases is the step parameter Rise. The minor-groove edge, where the sugar-phosphate backbones of the strand are closest to one another, is highlighted in red. The blue spheres are the origins of the base pairs.

As pointed out in Section 2.2 Tw^{SP} and Tw^{SC} are the same for idealized B-DNA. This is simply due to the fact that both the normal of the base pair and the tangent going from one base pair to the other base pair in the step are the same. This, however, assumes that the base pairs sandwiching the base pair step of interest are also that of ideal B-DNA, seen in Figure 2.4.1.2, or that there are no base pairs at all surrounding the base pair step of interest.

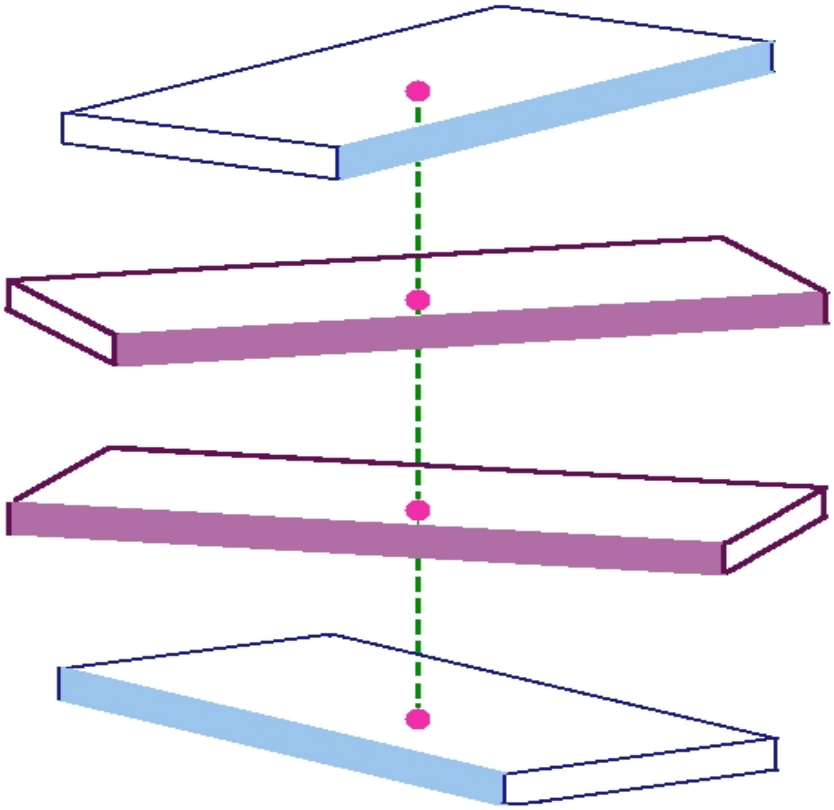


Figure 2.4.1.2: Four base pairs of ideal B-DNA. The space between the two purple base pairs is the base-pair step of interest. The two blue base pairs flanking the step of interest are required to calculate Tw^{SC}. The pink spheres are the origins and the dashed green line is the axial curve. All highlighted sides denote the minor-groove edge.

B-DNA has 10.5 base pairs per helical turn, but here we will simplify the calculations by assuming 10. A helical turn spans 360 degrees. For example, along a strand of ideal B-DNA the first 360º helical turn results in the first base pair of the next helical turn being positioned in the exact same orientation as the first base pair of the first helical turn. See Figure 2.4.1.3 for a helical turn of ideal B-DNA. If we divide 360 degrees by 10 base pairs, we have 36 degrees / base pair. Thus any value of Tw^{SP} greater than 36° signifies over-twisting of a step and any Tw^{SP} less than 36° is considered under-twisting of a base pair step. The amount of over- or under-twisting is valuable information for a biologist to know because it can help predict whether or not a particular segment can undergo changes such as specific protein binding.

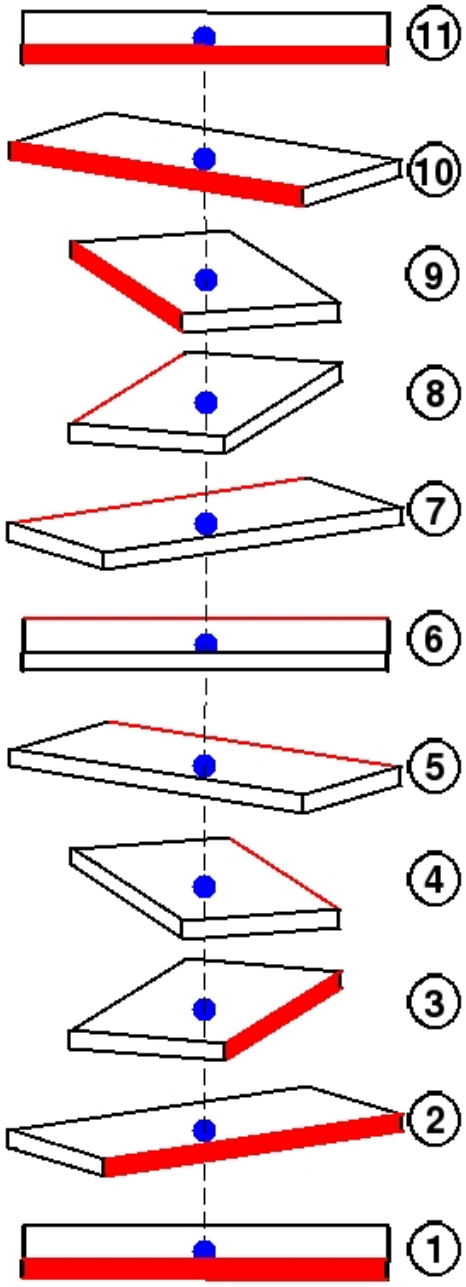


Figure 2.4.1.3: Base-pair slab representation of an ideal B-DNA helical turn plus an extra base pair. The extra base pair, number 11, is to show how the next helical turn starts just like the previous one. Notice halfway through the helical rotation, at base pair 6, that we are now looking at the opposite side of the base pair (the major-groove edge) compared to base pair 1 (where we see the minor-groove edge).

2.4.2 Closed Loop Model B-DNA

Each of the models described below is that of a closed structure. The step parameters for ideal B-DNA are as shown in Table 2.4.2.1.

Translational Parameters	Angstroms (Å)	Rotational Parameters	Degrees (°)
Shift	0	Tilt	0
Slide	0	Roll	0
Rise	3.4	Twist	36

Table 2.4.2.1: The step parameters for ideal B-DNA.

Ideal DNA is a linear molecule, which cannot form a closed structure without some sort of deformation from its rest state. Thus, we need to add other non-zero values to the bending step parameters (tilt and roll) in order to form a closed structure. In order to keep these models as simple as possible, we will mostly change only one or two variables at a time, depending on the model, to achieve this simplicity. The change to create a closed structure, in general, will only occur at two base pair steps for every helical turn. So, every fifth base pair step will have non-ideal B-DNA bending step parameters. Therefore, these are the steps studied for their effect on the difference between Tw^{SC} and Tw^{SP}, and they will be flanked by ideal B-DNA. It is imperative that the steps of interest are surrounded by ideal B-DNA because this will let it be consistent when we compare other steps of the same type. If ideal B-DNA is somehow affecting these steps of interest, the impact can be disregarded because it would have the same effect on each and every

step of interest.

The data collected for each structure located in the next few sections were obtained using the procedure described in Figure 2.4.2.1. It is necessary to mention, again, that we do not expect Tw^{SC} to differ much, if at all, from Tw^{SP} unless sections of the molecule show non-zero chirality. Chirality is the amount of nonplanarity a structure exhibits and gives it a left or right handedness. This stems from the expression for the linking number, Eq. 2.2.8, which clearly states that the linking number is a sum of Tw^{SC} and the writhe. For a closed molecule this number would have to be an integer. If the linking number has not changed, e.g. the DNA molecule has not been nicked and reattached in a different state, then the writhe must change in an equal and opposite way from Tw^{SC}. The writhe is a measure of how far from planarity a closed structure is. If the segment we are looking at is not chiral, it contributes zero change in writhe. Therefore, Tw^{SC} will also not change. To put this simply, a segment which is not chiral means that Tw^{SC} will equal Tw^{SP}.

1. Calculate the necessary step parameters that will create a circular structure for an 80 base-pair molecule, see Table 2.4.2.2 for the values used.
2. Use the 3DNA software package [13] to obtain a PDB formatted file based on results of Step (1).
3. Run 3DNATwSC to calculate Tw^{SC}, the linking number, and writhe of the molecule.
4. Plot 3-dimensional graphical depictions from 3DNATwSC for full rotational visualization of the molecule. There are two perspectives shown for the same molecule. The first is a top down view and the second is a side view. The third plot displays the difference, in degrees, between Tw^{SC} and Tw^{SP} along the contour of the molecule with color-coded 3-dimensional rendering.
5. Prepare tables to help visualize how changes in the step parameters influence Tw^{SC} and provide the ability to contrast it with Tw^{SP}.

Figure 2.4.2.1: The procedure used to generate the models, including all data and figures, discussed in Section 2.4.2.

2.4.2.1 Model Kink 1: Circle formed by kinking every five base pairs

We have looked at seven structures involving kinks, referred to as Model Kink 1, Model Kink 2, and so on to Model Kink 5, as well as two additional variations on Model Kink 2 called Model Bubble 1 and Model Bubble 2. The first model that is shown is the simplest model in our series of seven modeled mini-circles because the only step parameter that is changing is the roll, which changes at every fifth base pair. The first non-zero number is –22.5° and the second is +22.5°, see Table 2.4.2.2. This is because half way through a helical turn of DNA, you are now looking at the backside of the structure as seen in Figure 2.4.1.3. Therefore, the roll needs to change signs. If the signs did not change we would have a snakelike s-curve instead of a mini-circle. We do not expect any real difference between Tw^{SC} and Tw^{SP} because we have not introduced any non-zero chirality into the equation. Remember, only chiral structures will bring a change between the two twists. Since roll is the only variable changing and it is a rotational parameter without a translational variable in the mix, we can safely say that Model Kink 1 has a zero chirality and zero writhe. Figures 2.4.2.1.1 - 2.4.2.1.3 show different representations of the same closed molecule. The first shows a top down view of the base-pair steps along a minicircle that closely resembles a nearly perfect circle. The second, a side shot, shows how all the origins lie along the same line. This is a visual indicator that the molecule has zero, or close to zero, writhe. The third is a very interesting depiction of how the difference between Tw^{SC} and Tw^{SP} changes at each base-

pair step. In this instance, we see that both twists are very close to each other with their colors depicting difference values very close to zero.

Figures 2.4.2.1.3 and 2.4.2.1.4 show how all six step parameters and Tw^{SC} change along the sequence. The results show that the total Tw^{SC} and Tw^{SP} are, in fact, equal. However, in order to close the molecule, a synthetic "last step" was added which joins the first base pair to the 80th base pair. Since this step does not conform to the "just add the rotational parameter roll" it does add a small amount of translation. This leads to a minute amount of writhe. The writhe for this "closed" structure equals 0.00031, and the total turns of Tw^{SC} is 7.99969. The sum gives a perfect B-DNA linking number of 8 for the 80 base-pair model. For comparison, for the "open" structure before the closing step was added, the total turns for both Tw^{SC} and Tw^{SP} were the same, as expected, and came to 7.90022 turns.

Step	Shift	Slide	Rise	Tilt	Roll	Twist
	(Å)	(Å)	(Å)	(°)	(°)	(°)
1	0	0	3.4	0	0.0	36
2	0	0	3.4	0	0.0	36
3	0	0	3.4	0	0.0	36
4	0	0	3.4	0	0.0	36
5	0	0	3.4	0	–22.5	36
6	0	0	3.4	0	0.0	36
7	0	0	3.4	0	0.0	36
8	0	0	3.4	0	0.0	36
9	0	0	3.4	0	0.0	36
10	0	0	3.4	0	22.5	36

Table 2.4.2.2: Step parameters for the first 10 steps of the 80 base-pair mini-circle Model Kink 1. This repeats eight times to complete the 80 base-pair mini-circle and is the basis for the other models in this chapter except for Model Kink 5. All values except for roll are those used for ideal B-DNA. Roll is 0° for ideal B-DNA.

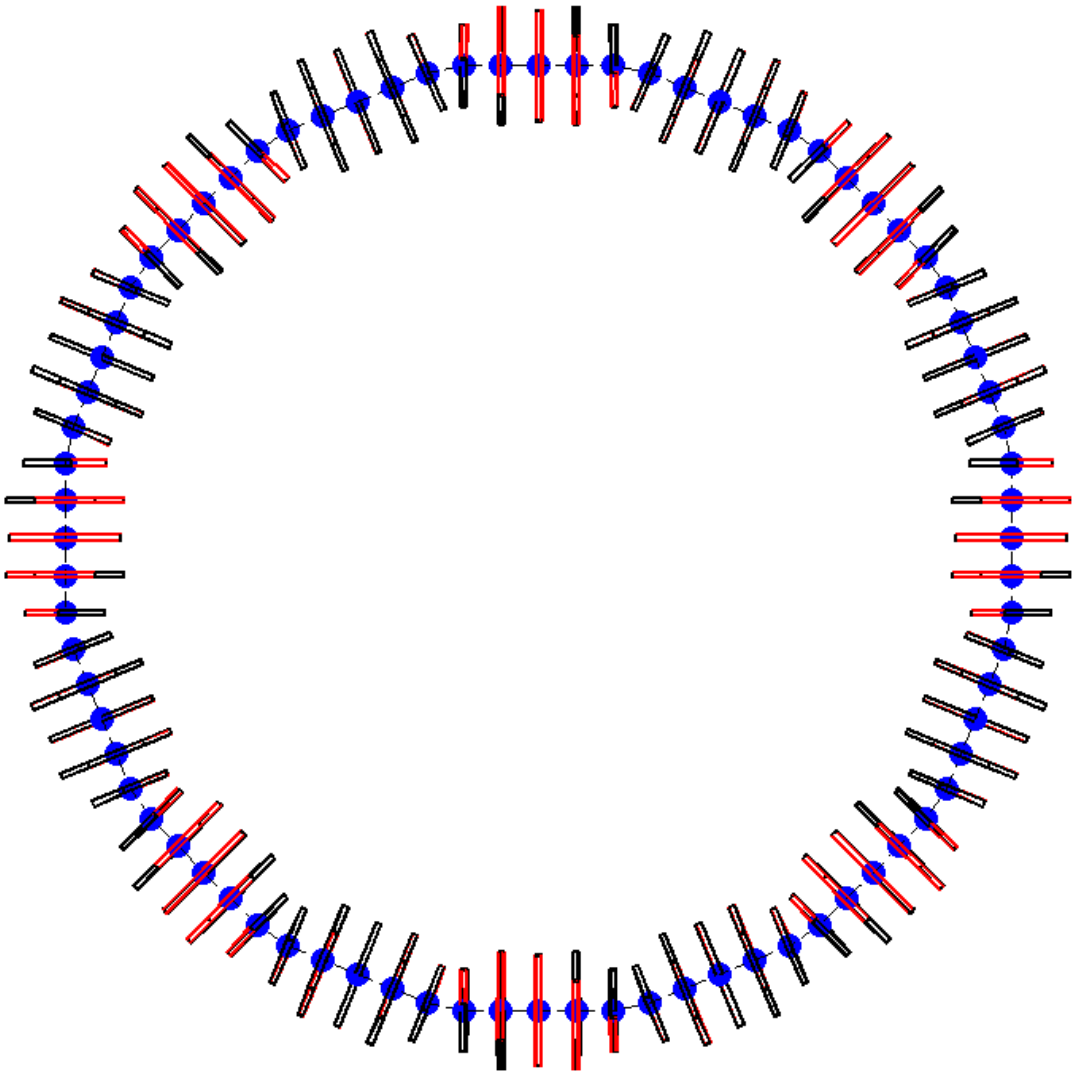


Figure 2.4.2.1.1: Top-down view of the 80 base-pair mini-circle, Model Kink 1. The opening is near the 9 o'clock position. The minor-groove edge is highlighted in red, and the origins of the base pairs are the blue spheres. The central axis of the molecule is a dashed line.

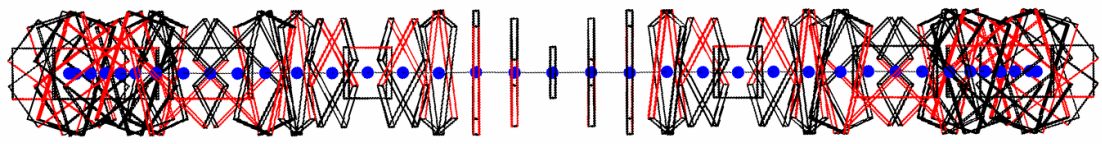


Figure 2.4.2.1.2: Side view of Model Kink 1. Notice how all of the origins lie along the same line.

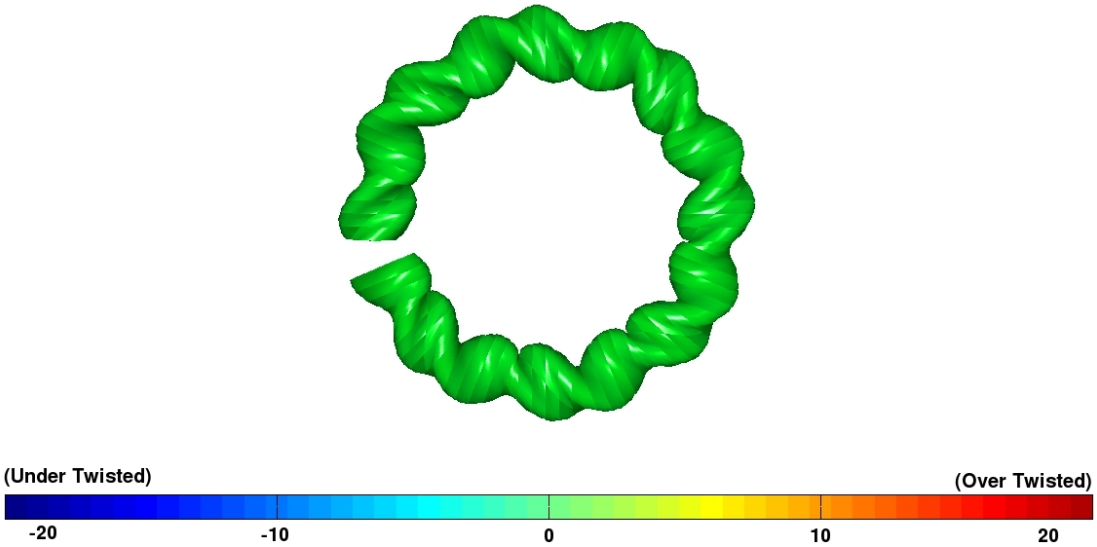


Figure 2.4.2.1.3: Top down view of Model Kink 1, color coded to show how the difference between Tw^{SC} and Tw^{SP} changes along the molecule.

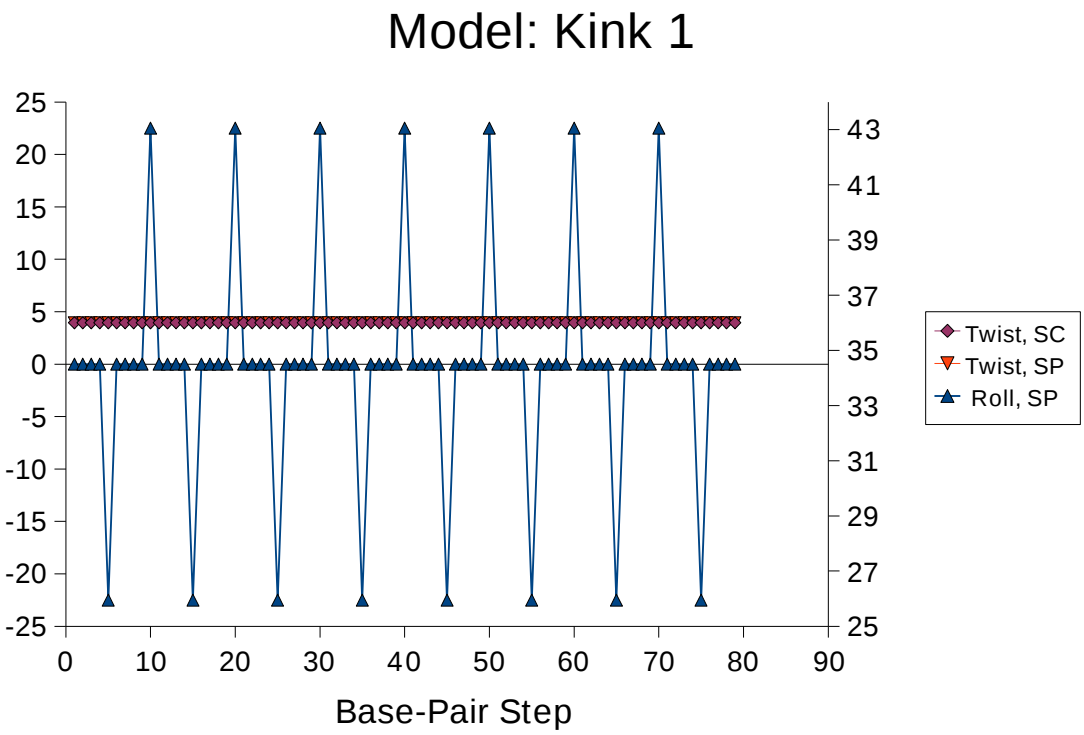


Figure 2.4.2.1.4: Graph for Model Kink 1 that shows the non-zero rotational step parameters, twist and roll, for the 80 base pair mini-circle and Tw^{SC} *for each base pair. The left axis on the graph denotes the roll, in degrees, and the right axis denotes the twist, in degrees. The graph shows the non-zero roll steps alternating between –25° and +25°, and the overlap of* Tw^{SP} *and* Tw^{SC} *at 36° for all base-pair steps.*

2.4.2.2 Model Kink 2: Left-handed superhelix formed by kinking and sliding every 5 base pairs

Model Kink 1 was one of the simplest forms that could be made to create a mini-circle using very few base pairs to create an almost circular pathway. The model in this section, Model Kink 2, adds one level of complexity to the first model. The roll has the same values as the previous one but the pathway also includes the step parameter slide into mix. Slide is the motion of one base translating along the long axis relative to its partner in the base-pair step. Just like the previous model, the slide will change signs due to its location within a double-helical turn. Every five base-pair steps the edge of the base pair that was facing outward is currently facing inward. The slide alternates between –1.75Å and +1.75Å at these steps.

At first glance the top down view of the 3-dimensional structure in Figure 2.4.2.2.1 suggests we have again generated a circular molecule. Examination of the side view in Figure 2.4.2.2.2, however, shows that the circle is not planar, but rather the DNA pathway is superhelical. This is because the slide creates a chiral structure. Figure 2.4.2.2.4 shows the primary variables at each step of the molecule. At the steps where there is ideal B-DNA Tw^{SC} and Tw^{SP} are equal, or at least very close to equal. However, Model Kink 2 has created a distinct change between the two twists. The difference between them is positive and thusly indicates a right-handed chiral structure. Also note that the superhelical pathway is left-handed.

Figure 2.4.2.2.3 depicts the changes between Tw^{SC} and Tw^{SP} in Model Kink 2 through the color-coded graphical representation. There is an eye catching color difference in ΔTwist every five base pairs compared to that in the B-DNA step when ΔTwist is zero. Comparisons of the overall twist values for Model Kink 2, expressed as the total turns for the open (no closing step added) structure, given 7.90000 turns for Tw^{SC} and 8.12779 turns for Tw^{SP}. A "synthetic" closing step from base-pair 1 to base-pair 80 was added such that this molecule could close tightly, and that step had Tw^{SP} of 0.10000 turns and Tw^{SC} of 0.002795 turns. This closing step showed that the molecule had a writhe of –0.172895 and Tw^{SC} of 8.172895 turns, making the linking number 8. In contrast the closing step yields Tw^{SP} of 8.00000 turns and this, when added to the writhe of the structure, definitely does not lend itself to an integer value for the linking number. Since the linking number is the same for Model Kink 1, we can say that these are the same molecules with just differences in their overall shape. This can happen in vitro/vivo with outside forces, e.g., salinity changing the step parameters. Note that the sum of the writhe and Tw^{SP} is not an integer.

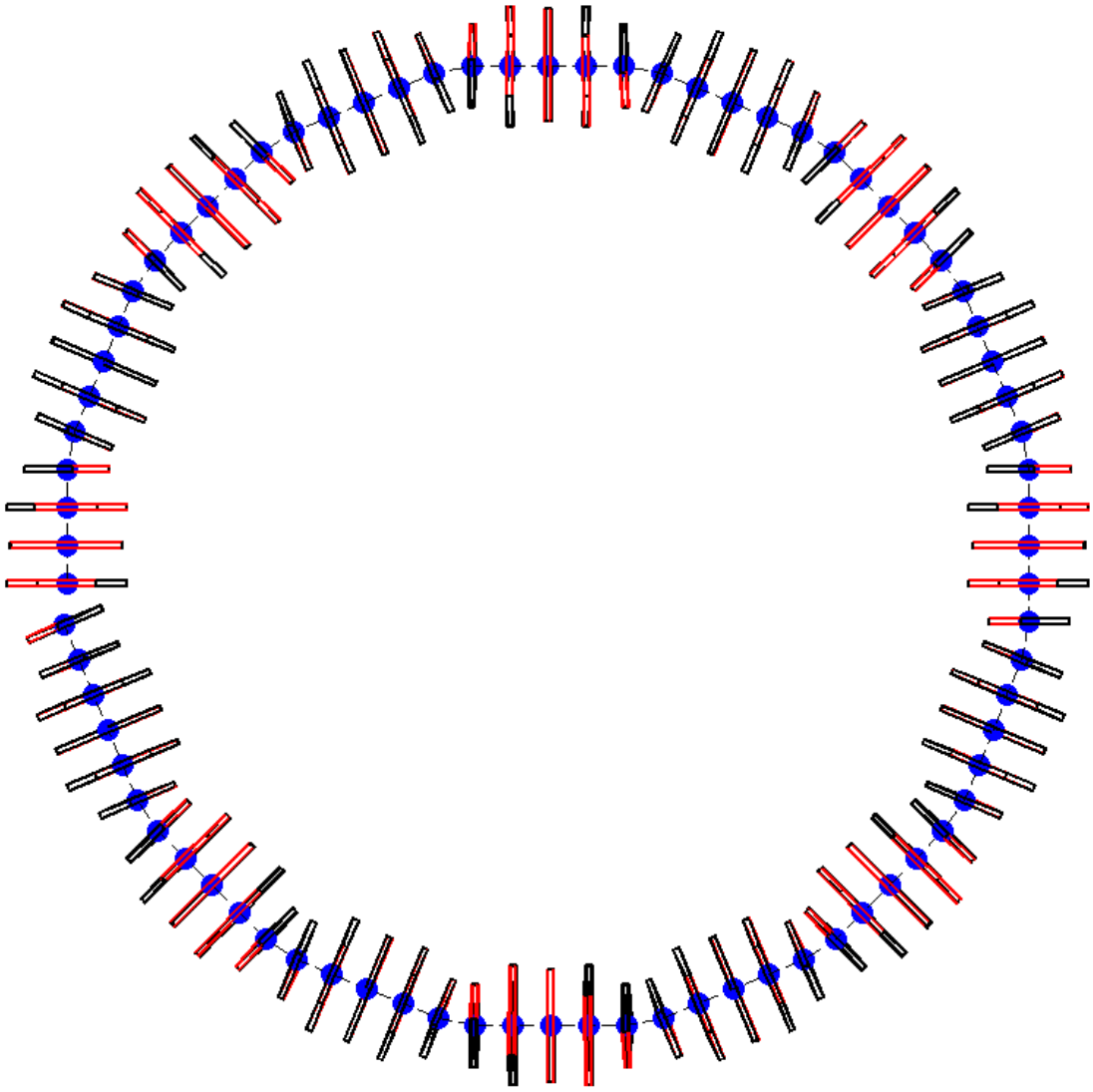


Figure 2.4.2.2.1: Top-down view of Model Kink 2. The minor-groove edge is highlighted in red, and the origins of the base pairs are the blue spheres. The central axis of the molecule is a dashed line. Notice that the DNA looks very circular and planar from this angle. Also notice that there is no closing step at the opening near the 9 o'clock position.

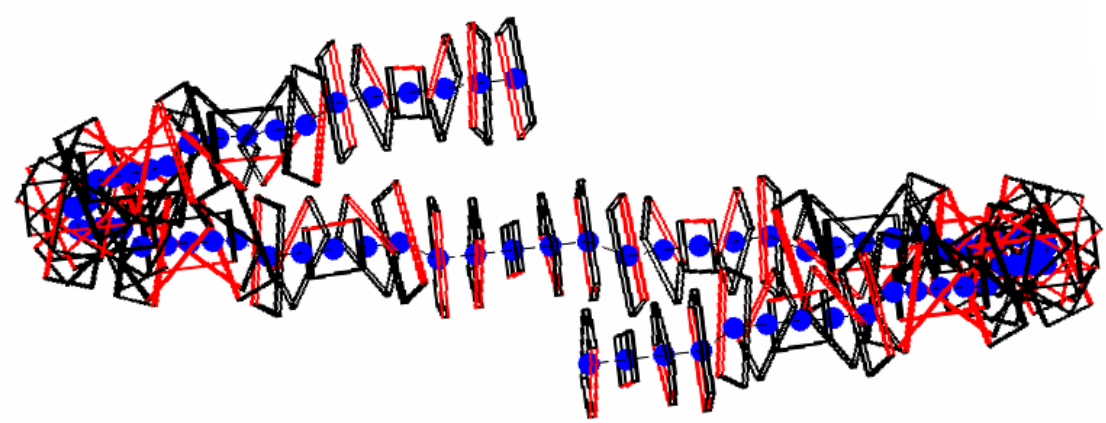


Figure 2.4.2.2.2: Side view of Model Kink 2. The structure depicted here is left-handed. To calculate the writhe a straight line is drawn to connect the first and last base pairs.

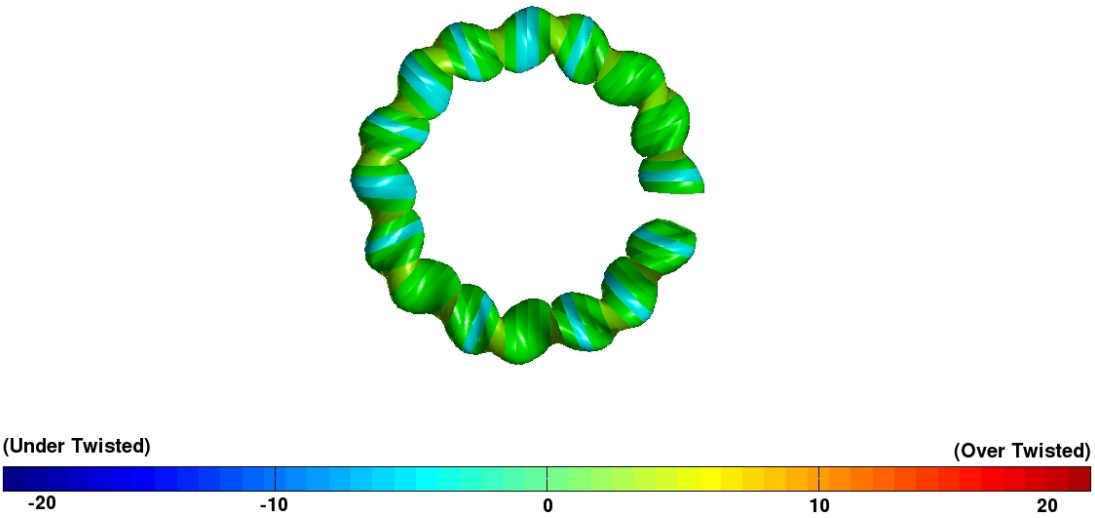


Figure 2.4.2.2.3: Difference in Tw^{SC} and Tw^{SP} color coded around Model Kink 2. Notice the cyan lines every five base pairs. This is where the chirality is introduced in the model.

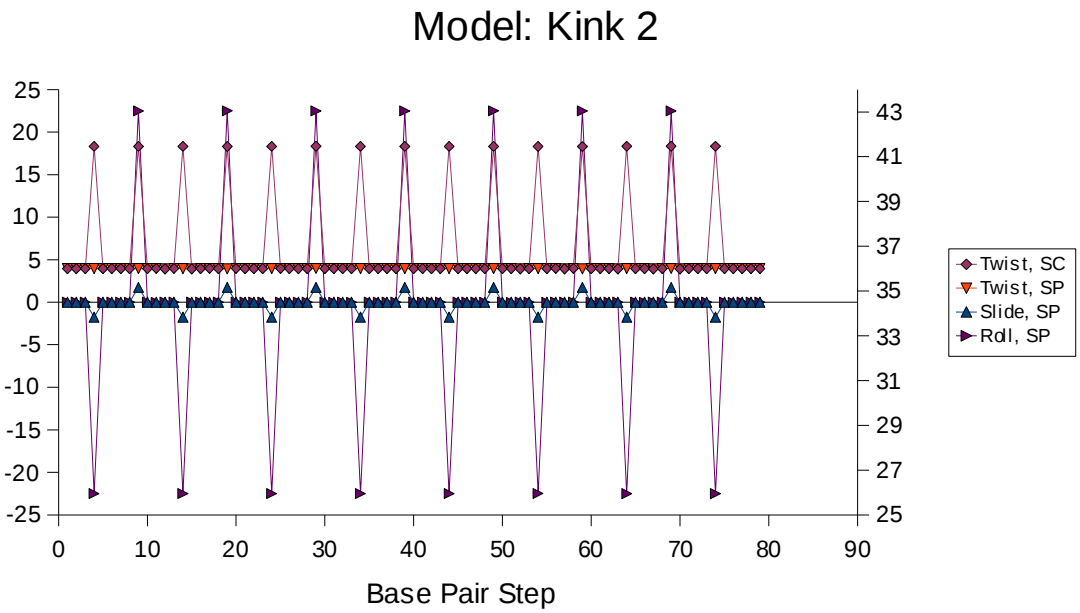


Figure 2.4.2.2.4: Graph of the variable step parameters, slide and roll, used to generate Model Kink 2 and the variation in Tw^{SC} and Tw^{SP}. The left axis on the graph denotes the slide, in Å and roll, in degrees. The right axis denotes the twist, in degrees. Notice the difference in Tw^{SC} and Tw^{SP} at the sites of non-zero slide.

2.4.2.3 Model Kink 3: Circle formed by kinking with twisting every 5 base pairs

Model Kink 3 is similar to Model Kink 2 in that it also has two variables changing at the same base pair location, but the variables are Roll and Twist. The Roll and the step-parameter Twist change at the same place, with Roll either positive or negative 22.5° while the Twist alternates between 41° and 31°. As mentioned in Chapter 1, Roll and Twist are both rotational step parameters, and therefore no translation is taking place at these base pairs other than the necessary Rise. Without the combination of translation and rotation, writhe will remain zero, and as stated above, without writhe there will be no chirality. In this situation, Tw^{SC} and Tw^{SP} are expected to be the same for both the total molecule and for each individual step.

Figure 2.4.2.3.4 shows the variation of the step parameters Roll and Twist and Tw^{SC} at each base-pair step. It is very easy to see from the right axis that both twists are identical along the entire strand. The left axis, also set in degrees, only tracks the step-parameter Roll. The total turns for Tw^{SC} is 7.91381 and for Tw^{SP} is 7.91389 for the 80 base-pair molecule. When a final closing step is placed between the first and the last base pair, the writhe is 0, Tw^{SC} is 8 and the linking number becomes 8, a perfect integer. These numbers mathematically help us have a feel that the molecule should be a mini-circle that closely resemble a circle, but to get an even better feel, we can look at Figures 2.4.2.3.1 and 2.4.2.3.2 to see how very planar and circular Model Kink 3 really is. As is clear from the color-coded representation of Model Kink 3, in Figure 2.4.2.3.3, all values

of the differences in the two twists are close to zero and there are no distinct patterns along the molecule contour, indicating that ΔTwist is not a factor when studying Model Kink 3.

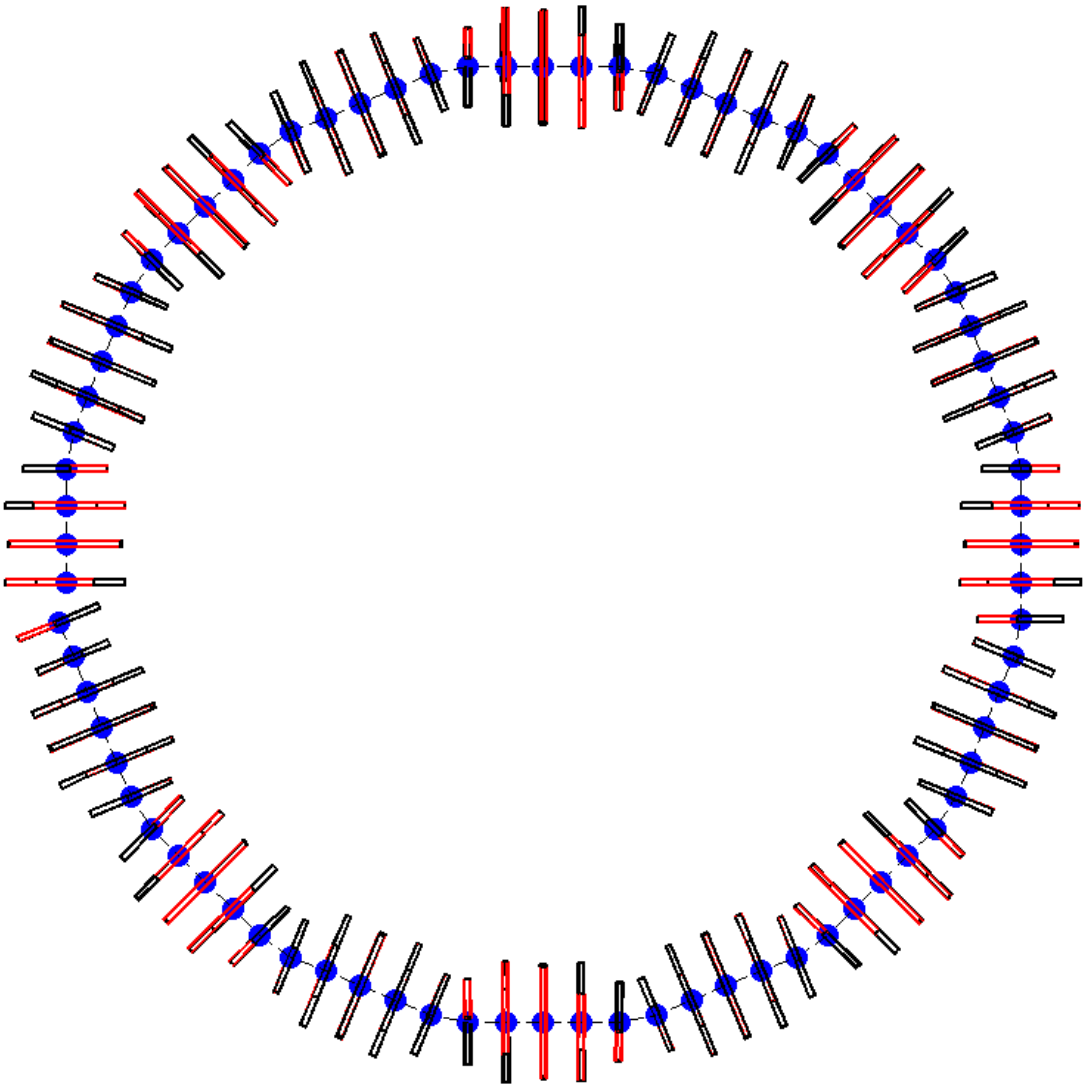


Figure 2.4.2.3.1: Top-down view of the base-pair slabs and central axis of Model Kink 3. Look how close to a perfect circle this is. The minor-groove edge is highlighted in red, and the origins of the base pairs are the blue spheres. The central axis of the molecule is a dashed line.

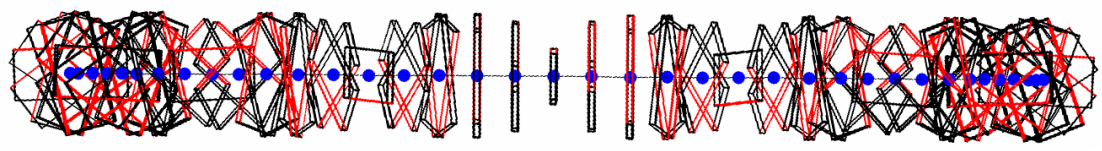


Figure 2.4.2.3.2: Base-pair slabs, origins, and central axis of Model Kink 3. Notice how all the origins lie along the same line. This is a clue that the structure is planar.

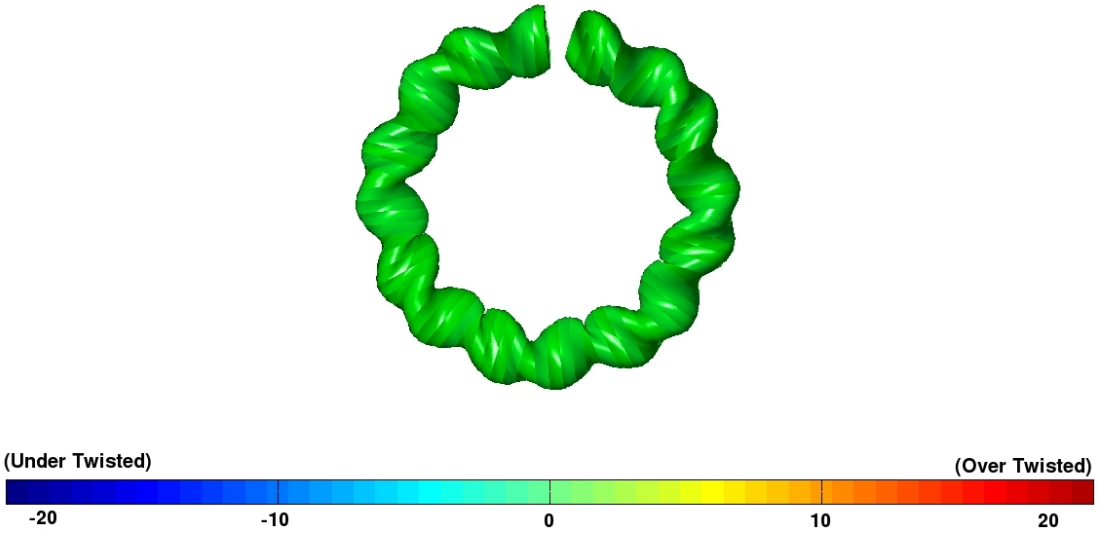


Figure 2.4.2.3.3: Three-dimensional rendering of Model Kink 3 color coded to show ΔTwist = $Tw^{SC} - Tw^{SP}$.

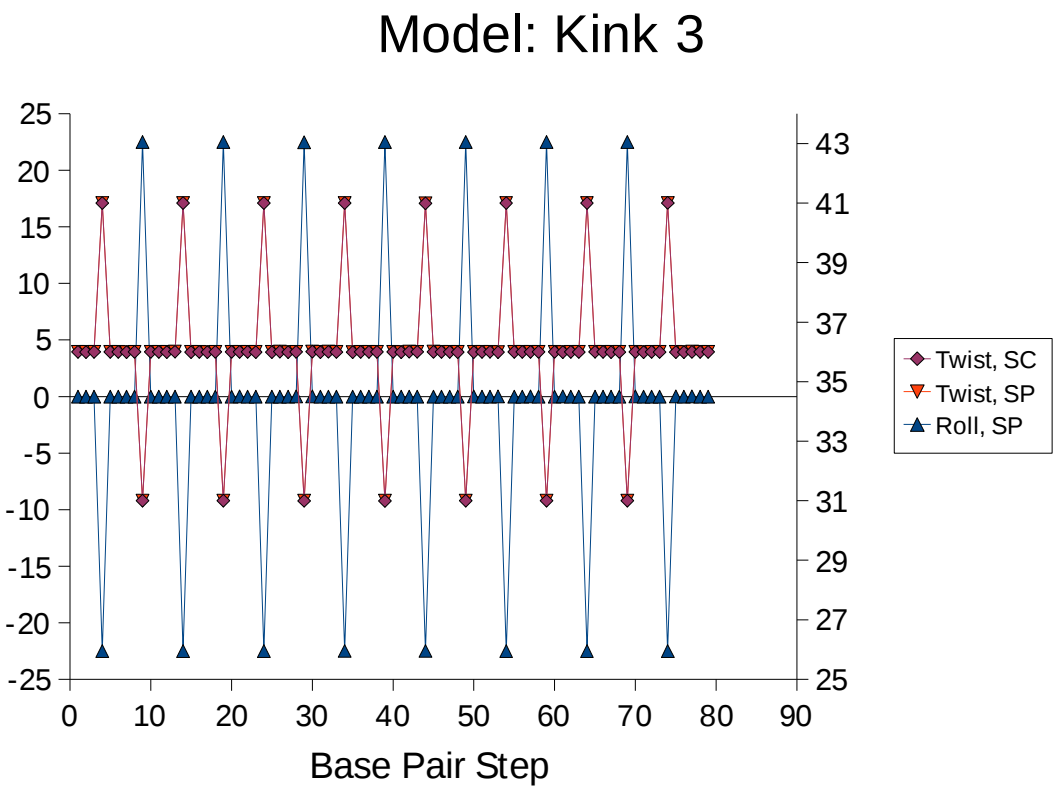


Figure 2.4.2.3.4: Graph of how Tw^{SC} and Tw^{SP} and roll change at each of the 80 base-pair steps of Model Kink 3. Roll, the values of which are shown on the left axis in degrees, alternates between positive and negative 22.5° while Tw^{SP}, the values of which are shown on the right axis in degrees, goes from 41° to 31°.

2.4.2.4 Model Kink 4: Left-handed superhelix formed by kinking and sliding with twisting every 5 base pairs

Continuing the theme of adding complexity, Model Kink 4 has three variables. The step parameters Twist and Roll vary as in Model Kink 3 in combination with changes in the translational step parameter Slide. The values for Slide are slightly different than those used in Model Kink 2. Here, the Slide alternates between 2.5Å and –1Å when (twist, roll) are respectively (41°, –22.5°) and (31°, +22.5°). To get a better feel for Model Kink 4, examination of Figure 2.4.2.4.1 shows a top-down view of a nearly circular pathway. However, examination of Model Kink 4 from the side in Figure 2.4.2.4.2 shows that this molecule is not planar and has a non-zero chirality and, therefore, writhe. Instead, the two ends of the DNA diverge from one another.

Figure 2.4.2.4.4 shows the variation in the step parameters Twist, Roll, and Slide along with Tw^{SC}. It is very easy to see how Tw^{SC} and Tw^{SP} differ. Remember, these molecules are first created by assigning specified step parameters to each of the 79 base-pair steps (80 base pairs). The program 3DNA takes these step parameters and pulls out the information needed to calculate the origins, long and short axes, and the base-pair normals [13]. With this in mind it is doubly interesting to see how the twists do not show the same overall pattern (e.g., over and under twisting). Instead, despite Tw^{SP} defining the long and short axes positions of the molecule, Tw^{SC} shows a different dependence on the chiral changes brought about by Roll and Slide.

The color-coded graphical representation of the value of ΔTwist = (Tw^{SC} – Tw^{SP}) in Model Kink 4, shown in Figure 2.4.2.4.3, reveals a pattern with two different ΔTwists within each helical turn. The values of Slide alternate between –1 and +2.5 at steps 5 and 9 of each helical turn. Moreover, the magnitude of Slide at theses two steps is either 1 or 2.5, and the larger magnitude of Slide leads to a larger magnitude in ΔTw. Looking at the final results for the molecule as a whole, Model Kink 4 shows two different results for the total turn in the open molecule. For Tw^{SP} the turns are 7.68614 and for Tw^{SC} the turns are 7.91383. When a final synthetic closing step is added, Tw^{SP} becomes 8.18650 turns and Tw^{SC} becomes 7.82739 turns. When Tw^{SC} calculated by the addition of a closing step is combined with the writhe of 0.17261 it brings the linking number to a perfect 8.

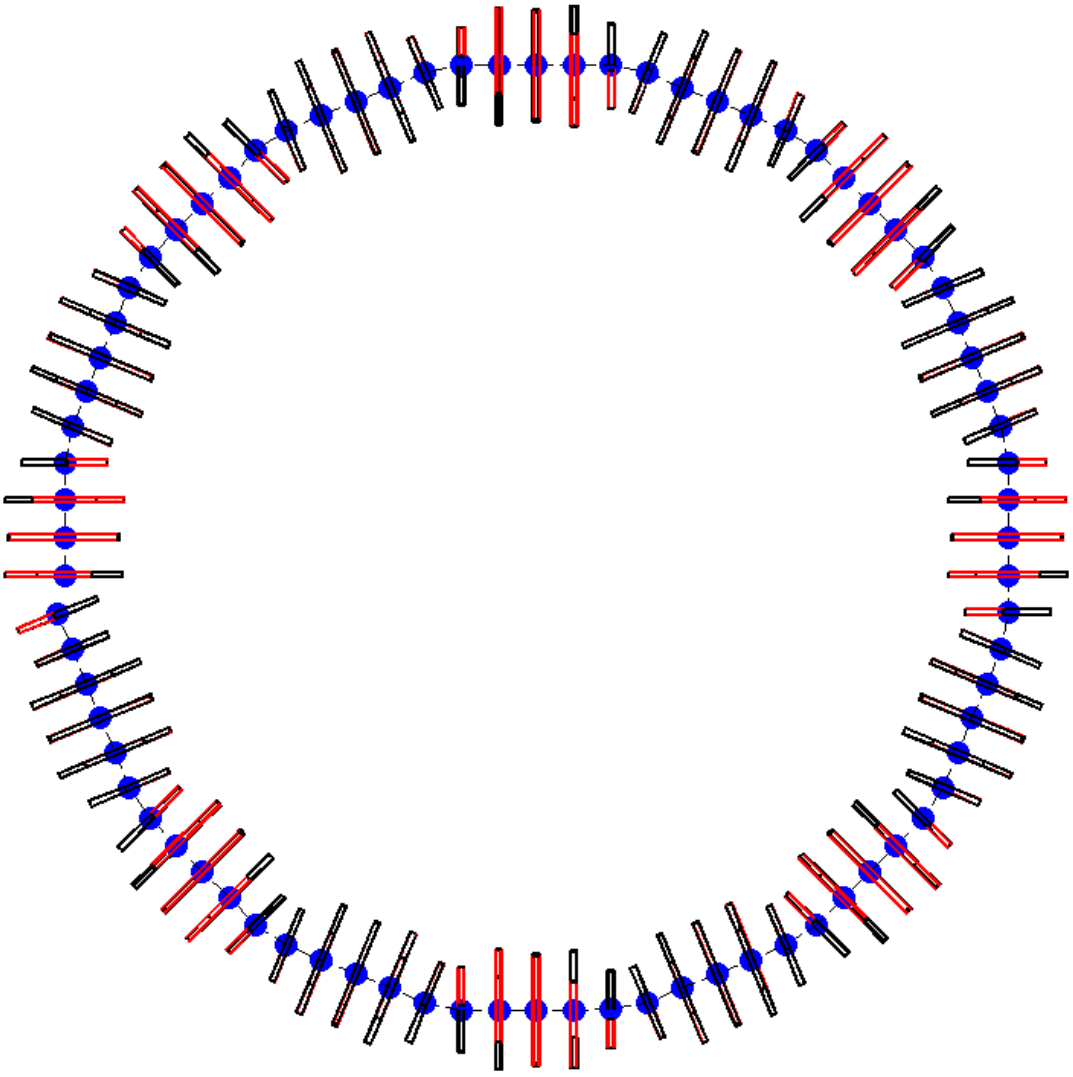


Figure 2.4.2.4.1: Top-down view of Model Kink 4 showing all base pairs as slabs. The minor-groove edge is highlighted in red, and the origins of the base pairs are the blue spheres. The central axis of the molecule is a dashed line.

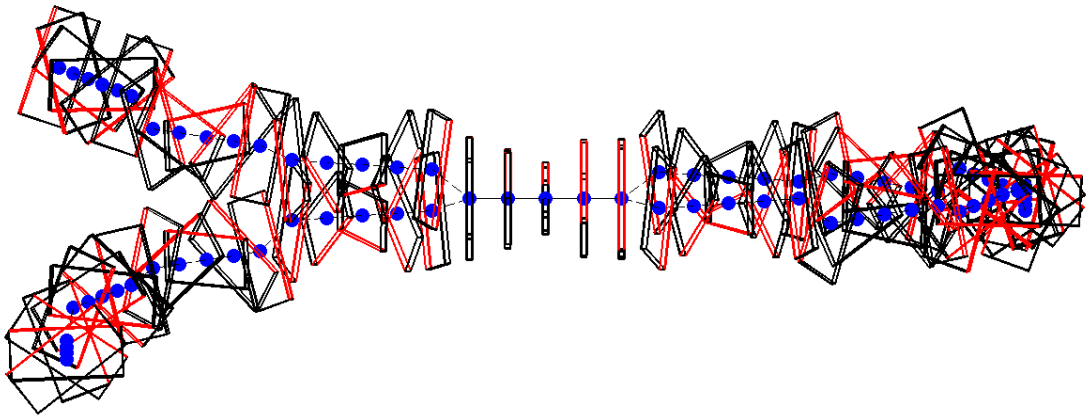


Figure 2.4.2.4.2: Side view of Model Kink 4 using slabs to represent the base pairs, blue spheres for the origins, and the dashed line for the axial curve of the molecule.

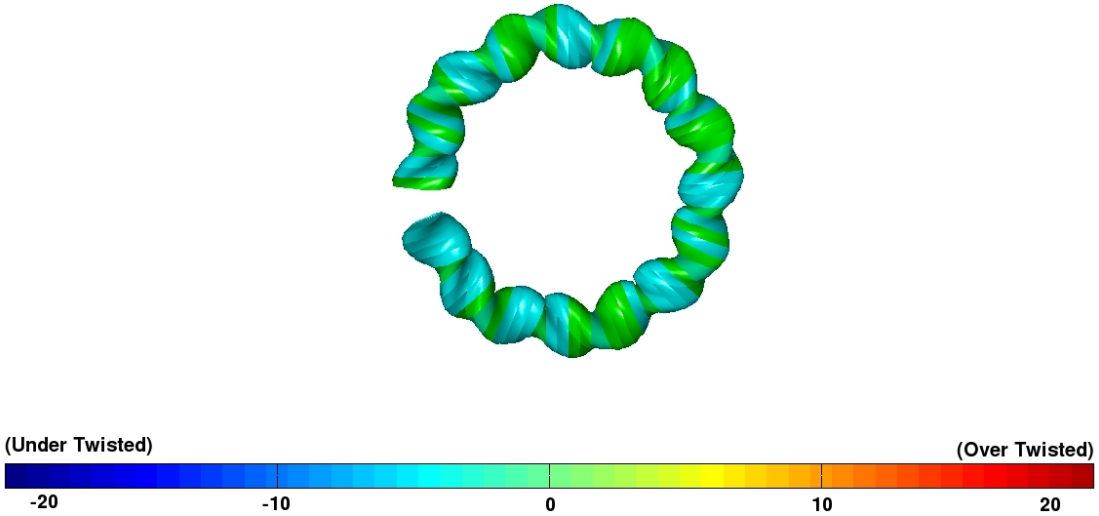


Figure 2.4.2.4.3: Graphical representation for Model Kink 4 with each base-pair step color coded to depict the value of ΔTwist = $Tw^{SC} - Tw^{SP}$. Notice two distinct patterns, the first one every helical turn and the second one halfway through the turn.

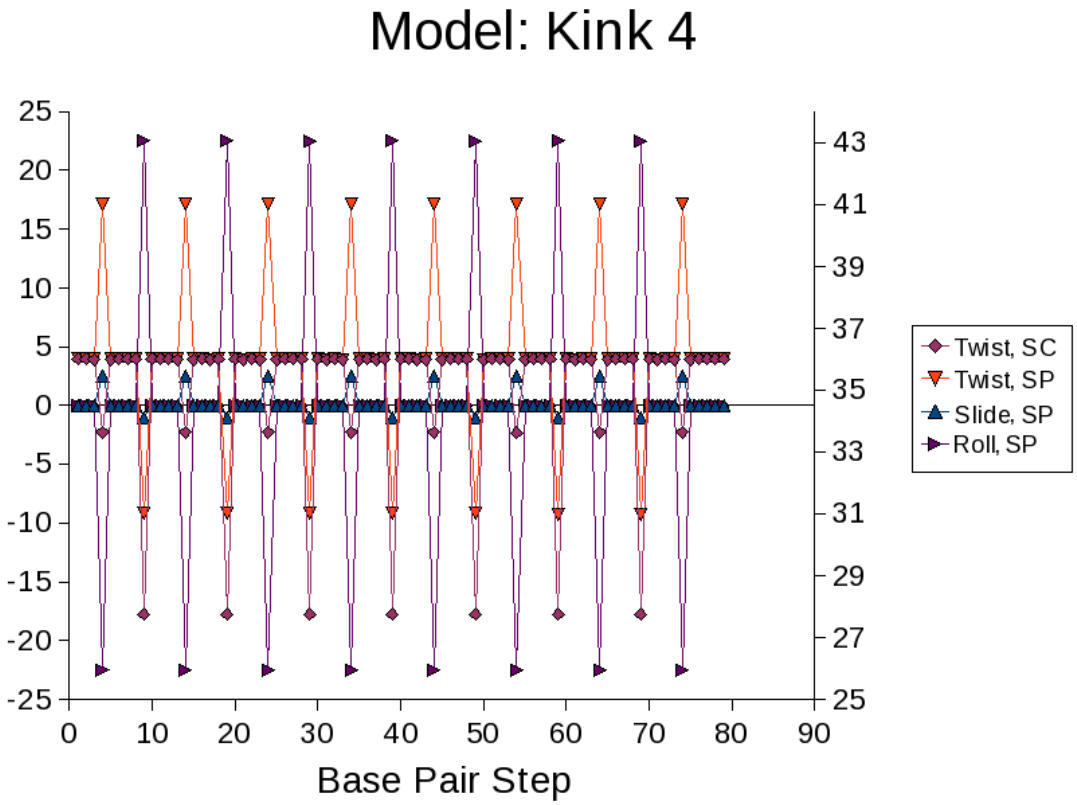


Figure 2.4.2.4.4: Graph of the two rotational step parameters Twist and Roll and the translational step parameter Slide along with the calculated value for Tw^{SC}. Notice that Tw^{SC} is either 36° or under twisted at every step. Compare this to Tw^{SP}, which alternates between a more over twisted or under twisted value.

2.4.2.5 Model Kink 5: Mini-circle superhelix formed by smooth (roll and tilt) at each base-pair step

Model Kink 5 can be considered the most complex mini-circle out of the entire Kink series. Examinations of Figures 2.4.2.5.1 and 2.4.2.5.2 shows how close the structure is to a perfect planar mini-circle. To construct this mini-circle we have introduced a constant bend angle at each base-pair step by sinusoidal variation in the Roll and Tilt step parameters. The translational step parameters, Shift and Slide are set to zero, and Rise assumes a constant value of 3.4Å throughout. The value of Tw^{SP} is kept at a constant value of 36° throughout the model. Table 2.4.2.5.1 lists the step parameters for the first 10 steps of the helix. The pattern is repeated seven more times throughout the molecule. Model Kink 5 is definitely a structure that we would expect Tw^{SC} and Tw^{SP} to be the same at each base-pair step. Looking at Figure 2.4.2.5.4 it is easy to see that this is indeed the case. This figure also shows very nice fluid waves of the Tilt and Roll values along the molecule. The reason that the two twists are expected to be the same is because chirality is not present at any point along the molecule. To reiterate the point again, there must be translation, not including Rise, and rotation, not including Tw^{SP}, coexisting at the same base-pair step to introduce chirality. Figure 2.4.2.5.3 shows that all the ΔTw values are very close to zero as expected.

When a final closing step is added, because of the design of the 80 base-pair mini-

circle, the total turns for Tw^{SC} is 8 with a writhe of 0 thus giving a linking number of 8. This implies that the final closing step is in perfect alignment with the rest of the sequence in regards to how all six step parameters and the Tw^{SC} are distributed throughout the molecule.

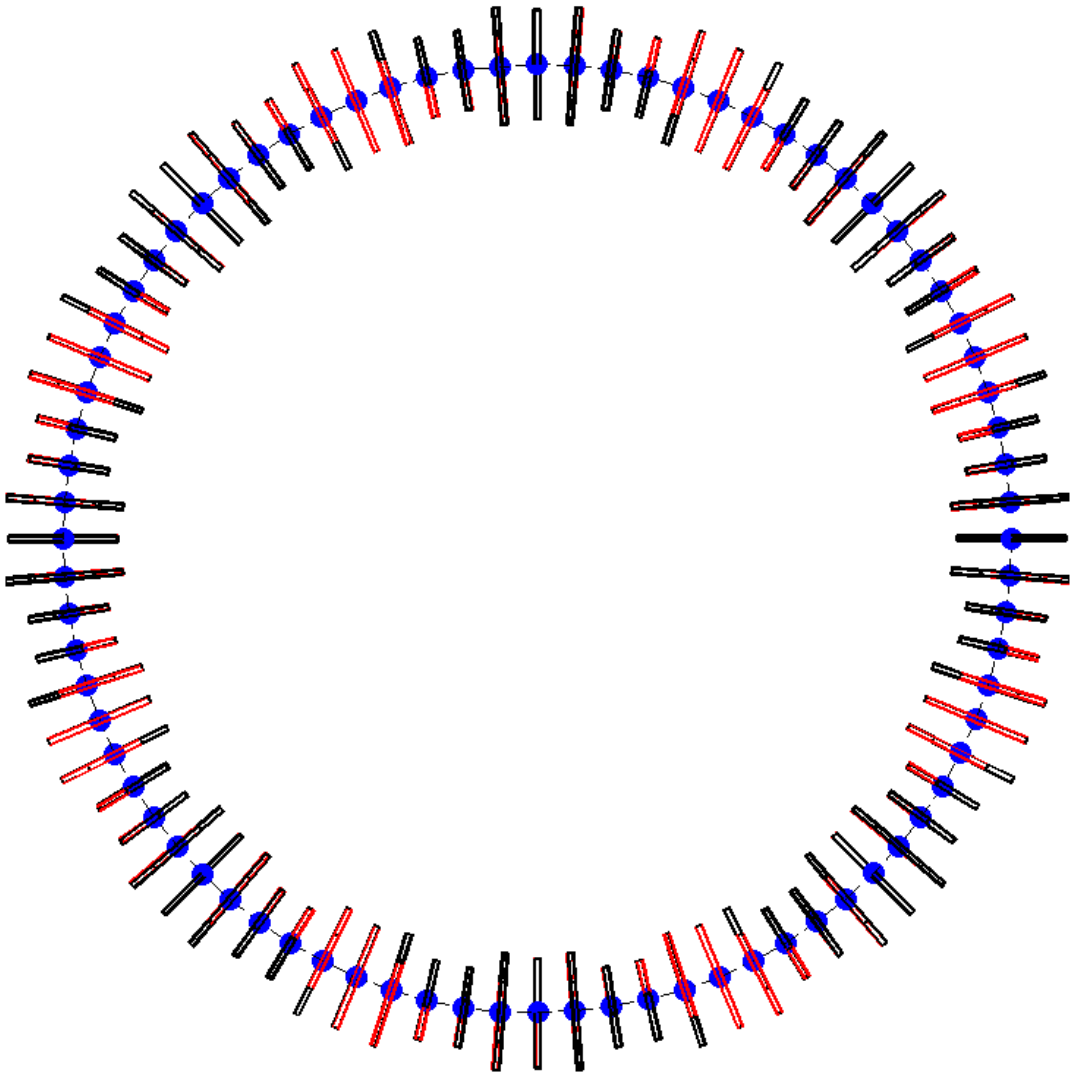


Figure 2.4.2.5.1: Top-down view of the base-pair slabs of Model Kink 5. The minor-groove edge is highlighted in red, and the origins of the base pairs are the blue spheres. The central axis of the molecule is a dashed line. The first and last base pair are located on the right-hand side close to the 3-o'clock position.

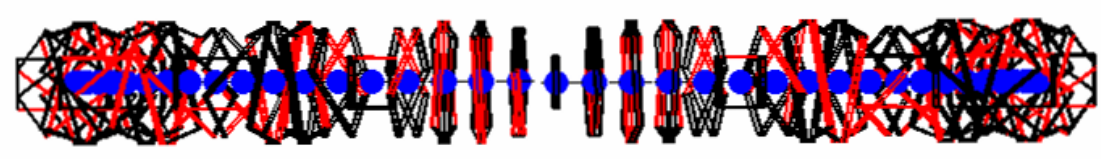


Figure 2.4.2.5.2: Side view of Model Kink 5. Notice how all of the blue spherical origins lie in a perfect line.

Step	Shift	Slide	Rise	Tilt	Roll	Twist
	Å	Å	Å	°	°	°
1	0	0	3.4	2.65	–3.63	36
2	0	0	3.4	–0.01	–4.51	36
3	0	0	3.4	–2.65	–3.64	36
4	0	0	3.4	–4.28	–1.39	36
5	0	0	3.4	–4.28	1.39	36
6	0	0	3.4	–2.66	3.63	36
7	0	0	3.4	0.01	4.50	36
8	0	0	3.4	2.65	3.64	36
9	0	0	3.4	4.28	1.40	36
10	0	0	3.4	4.28	–1.40	36

Table 2.4.2.5.1: Values for one helical turn of Model Kink 5. This is a pattern that is repeated 7 times in total for an 80 base-pair molecule.

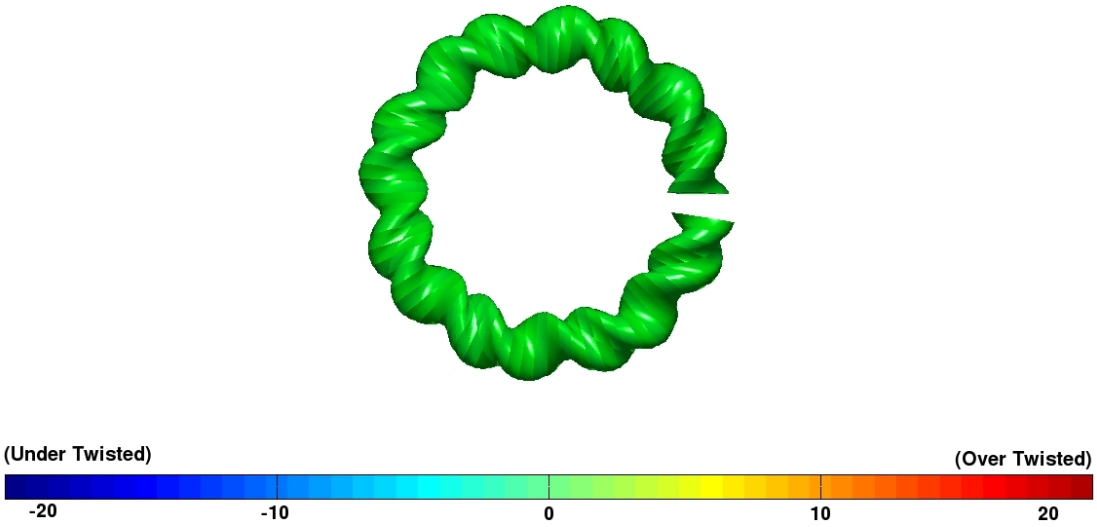


Figure 2.4.2.5.3: Color-coded graphical representation of ΔTwist = Tw^{SC} – Tw^{SP} of each base-pair step for Model Kink 5.

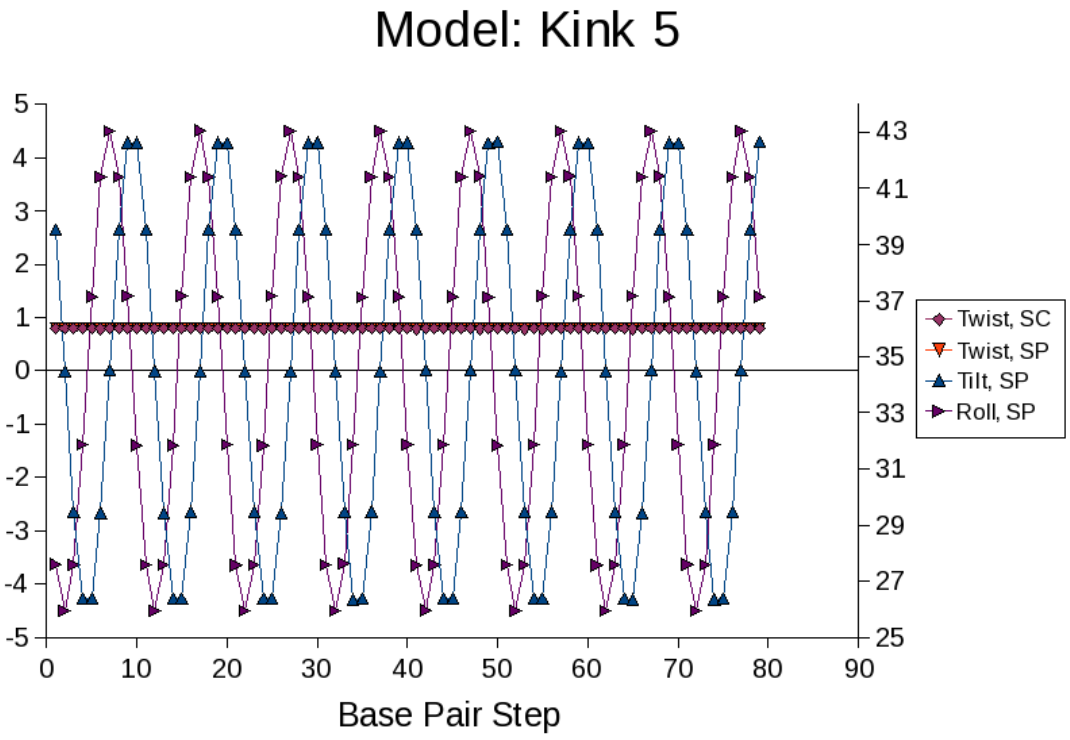


Figure 2.4.2.5.4: Graph of the two rotational step parameters Tilt and Roll that are changing at each base-pair step. Notice that Tw^{SC} and Tw^{SP} do not have any deviation from 36°. The ordinate on the left is for Tilt and Roll expressed in degrees, and that on the right is for both twists also in units of degrees.

2.4.2.6 Model Bubble 1: Left-handed superhelix formed by kinking and sliding with twisting every 5 base pairs

The next two sections introduce a bubbled portion into a mini-circle. The mini-circles are based on the ones from the previous five sections. More specifically, these next two sections have the same step parameters as the mini-circle in Model Kink 2 with the added bonus of a non-perfect bubble at or close to the start of the sequence. Model Bubble 1, the step parameters of which are listed in Table 2.4.2.6.1, has the bubbled portion starting at base pair 1. These values were used to generate the origins and axes required to calculate Tw^{SC}. Figures 2.4.2.6.1 and 2.4.2.6.2 show the top-down and side views of Model Bubble 1. These figures make it easy to see how the bubble deforms Model Kink 2 into Model Bubble 1. The extra bubble makes certain that this molecule can no longer be thought of as circular. Figure 2.4.2.6.2 shows that the origins do not lie in a single plane. This non-planarity implies that the writhe and chirality will be non-zero. The color coded representation of $\Delta Twist = (Tw^{SC} - Tw^{SP})$ in Figure 2.4.2.6.3 reveals a very distinct and clearly visible region with large $\Delta Twist$ values located in the bubbled portion. The dark blue color implies large under-twisting.

Figure 2.4.2.6.4 shows the overall variation of the two twists and the variable step parameter in Model Bubble 1. Figure 2.4.2.6.5 is a zoomed-in area containing only the bubble. The value of $\Delta Twist$ for each of the nine steps is easily tracked and at step five where the step parameters Roll and Slide are at their highest level, meaning maximum

chirality is introduced, the magnitude of ΔTwist is also the largest. The values of Tw^{SC} and Tw^{SP} are not the same for most of the steps in the bubble and the differences in the two values of twist are mostly negative.

The total turns for Model Bubble 1 is 7.90006 for Tw^{SC} and 7.66542 for Tw^{SP}. When the artificial closing step is added the total turns for Tw^{SC} is 7.74351 and the writhe is 0.25649 making the linking number 8. We can compare this writhe of 0.25649 to that of –0.17290 for Model Kink 2. Although most of Model Bubble 1 is taken exactly from Model Kink 2, the bubble leads to a negative writhe, with the highest absolute value of writhe generated thus far. This means that the writhe contributed from the bubble not only counter balanced the negative writhe from Model Kink 2, but was sufficiently positive to create a large writhe, large for structures with a nearly circular overall shape. This is why it is important for us to have a way to describe the topology of DNA in a useful way to see when areas can be over or under-twisted as this may affect how it is "seen" by other proteins/ligands/drugs that require specific conditions to bind.

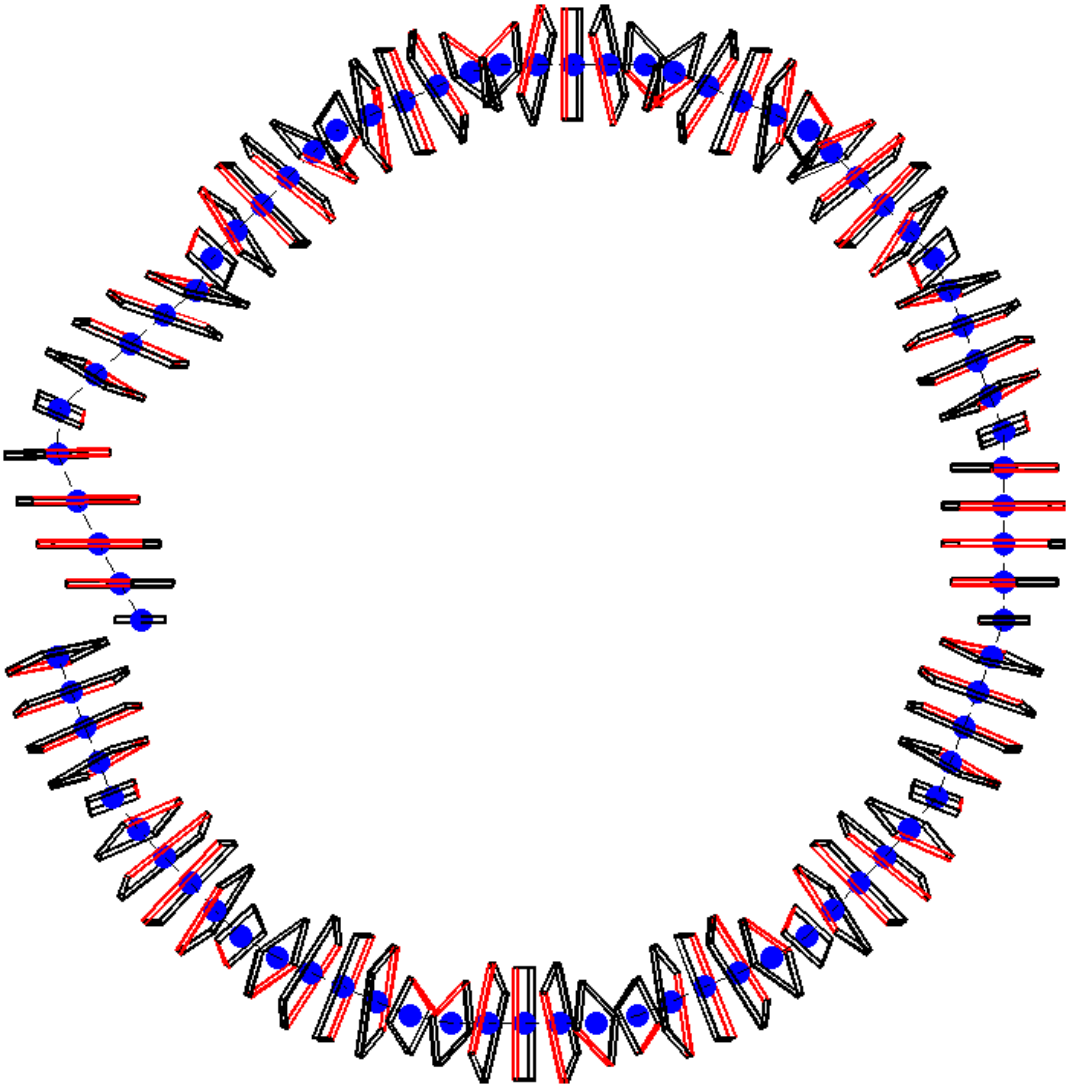


Figure 2.4.2.6.1: Top-down view of Model Bubble 1. The minor-groove edge is highlighted in red, and the origins of the base pairs are the blue spheres. The central axis of the molecule is a dashed line. The bubble is on the left-hand side in the middle of the molecule.

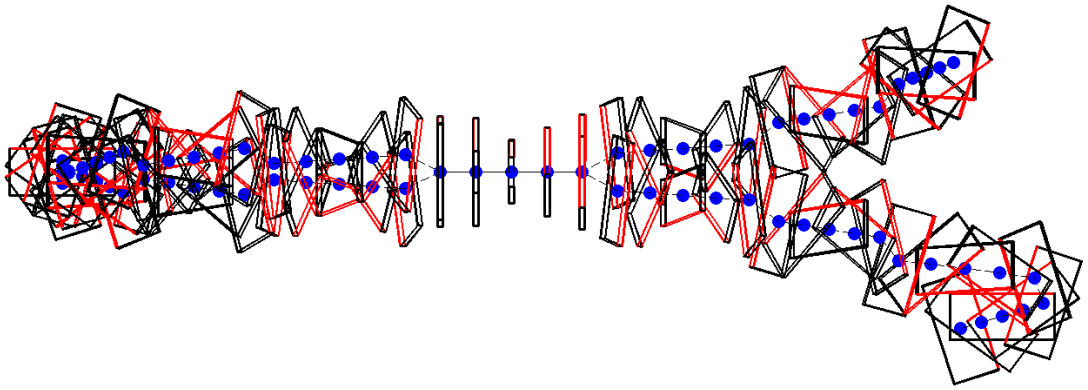


Figure 2.4.2.6.2: Side view of Model Bubble 1. Notice the bubble on the lower right-hand side of the picture.

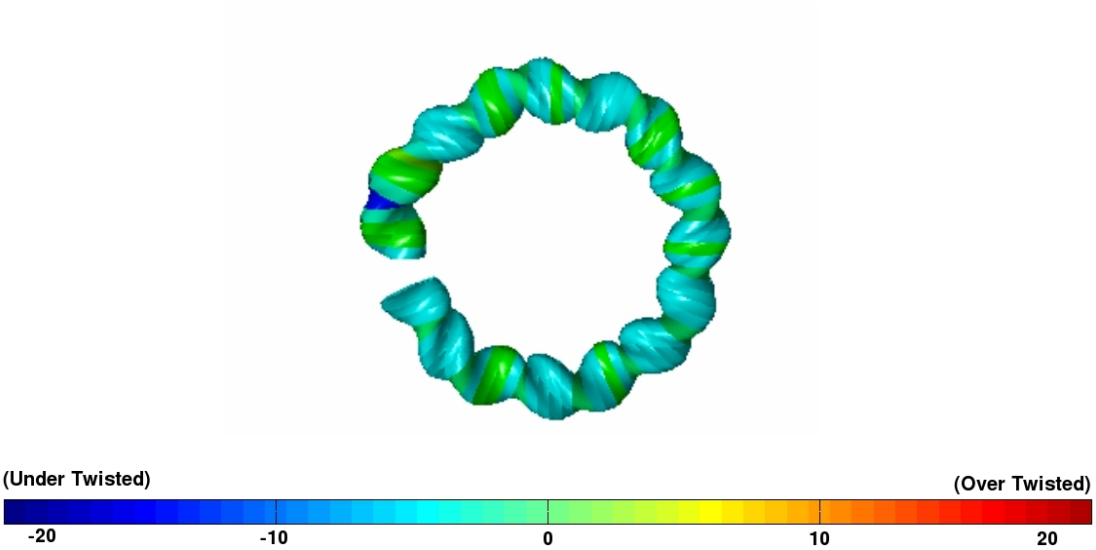


Figure 2.4.2.6.3: Color-coded graphical representation of ΔTwist = Tw^{SC} – Tw^{SP} for Model Bubble 1. The bubble is contained in the first nine base-pair steps at the start of the sequence.

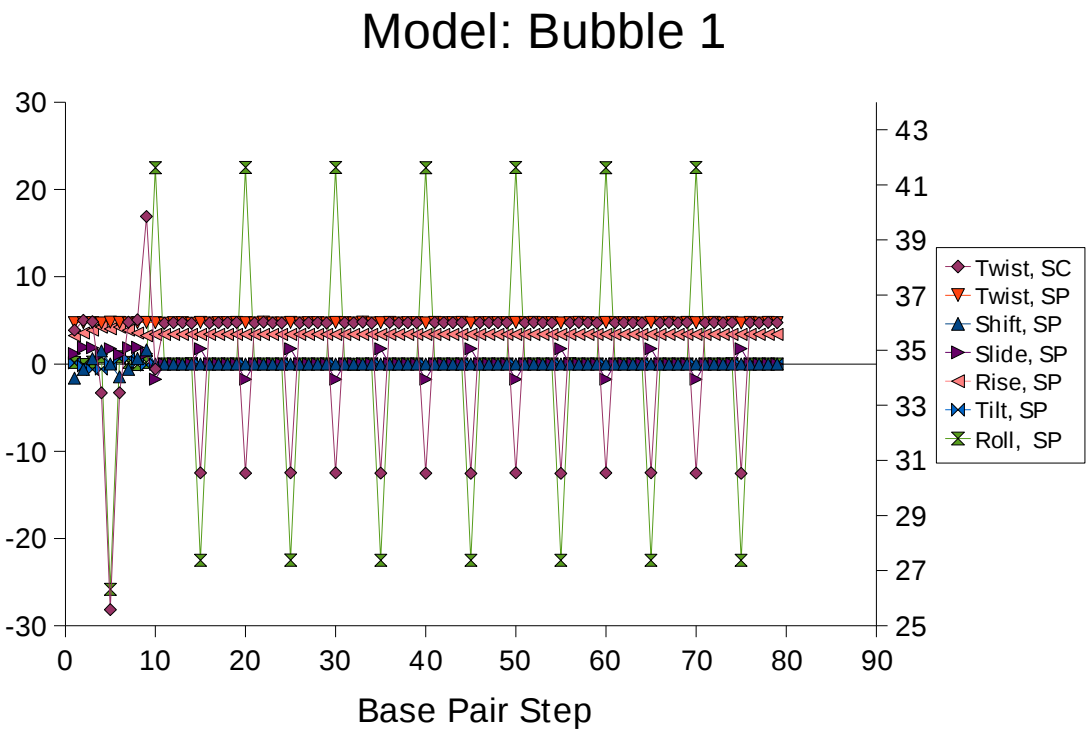


Figure 2.4.2.6.4: Graphical representation of Model Bubble 1. The most interesting steps are the first nine steps in the sequence. The rest of the molecule is the same as Model Kink 2.

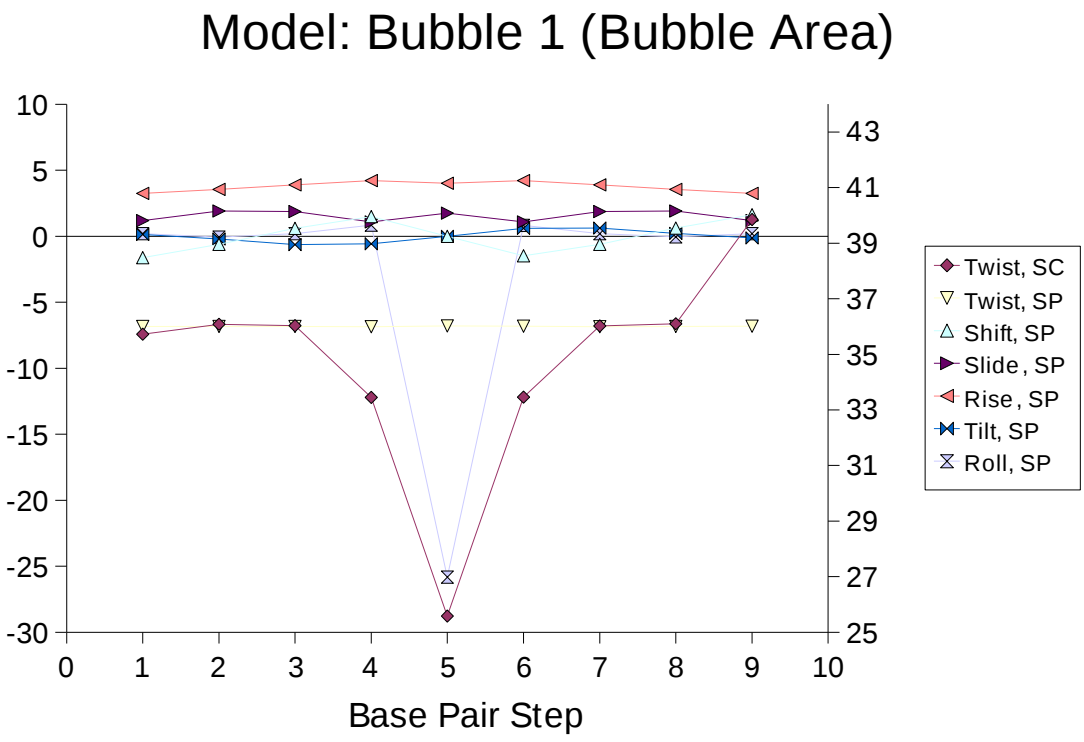


Figure 2.4.2.6.5: Close-up view of what is taking place at the bubbled portion of Model Bubble 1. Notice the large changes in the two types of twist plotted with respect to the ordinate on the right. The ordinate on the left is used to plot the rest of the step parameters — Shift, Slide, Rise, Tilt, and Roll.

Step	Shift Å	Slide Å	Rise Å	Tilt °	Roll °	Twist °
1	–1.62	1.18	3.24	0.13	0.19	36
2	–0.62	1.91	3.55	–0.21	–0.07	36
3	0.60	1.86	3.88	–0.63	0.20	36
4	1.48	1.08	4.22	–0.58	0.83	36
5	0.00	1.75	4.02	–0.02	–25.82	36
6	–1.48	1.08	4.22	0.60	0.82	36
7	–0.61	1.86	3.88	0.62	0.19	36
8	0.62	1.91	3.55	0.22	–0.06	36
9	1.62	1.18	3.24	–0.14	0.19	36

Table 2.4.2.6.1: The specific values of the base-pair step parameters describing the bubbled portion of Model Bubble 1.

2.4.2.7 Model Bubble 2: Left-handed superhelix formed by kinking, sliding, and twisting every 5 base pairs

Model Bubble 2 is a structure similar to Model Bubble 1. This bubble starts six base-pair steps from the first base pair. The remainder of the molecule that is not part of the bubble is still the same as Model Kink 2. The values for the bubble are very close to those of Model Bubble 1 but are slightly closer to ideal B-DNA. This means we expect the chirality of the overall structure will be slightly less than that of Model Bubble 1. The top-down view of the bubble, in Figure 2.4.2.7.1, resembles a closed structure. Remember, the bubble starts half a helical turn from the initial step. To see the start of the sequence, look at roughly the 9 o'clock position of Figure 2.4.2.7.1 to see where the dashed line is missing between base pairs. Figure 2.4.2.7.2 shows the side view of Model Bubble 2 with the bubble on the lower left-hand side of the image.

Comparisons of the two bubble models show that the middle of Model Bubble 2 is less over-twisted than the middle of Model Bubble 1 (note the lesser variations in color in Figure 2.4.2.7.3 when compared to Figure 2.4.2.6.3). The values for Tw^{SC} and Tw^{SP} are mapped out for the whole molecule in Figure 2.4.2.7.4 and the variation in the parameters of the bubble is seen in more detail in Figure 2.4.2.7.5. Like all of the previous models, these graphs display only the variables (step parameters) that are changing. In this case, due to the bubbled section, the image includes all six step parameters and Tw^{SC}.

The values for the step parameters chosen for the bubbled section are listed in

Table 2.4.2.7.1. The overall molecule results for the topology descriptions are similar to those for Model Bubble 1. For the open molecule, where no closing step is added, the number of turns for Tw^{SC} is 7.68743 and the number of turns for Tw^{SP} is 7.89994. These are not exactly the same and thus we can tell from these values that the writhe will not be zero. In fact, when a closing step is added, the writhe is 0.15763. This is less than the writhe of Model Bubble 1, as expected, since the bubble is not as dramatic in its chirality. The total turns for Tw^{SC} is 7.84237 for the closed molecule leaving a perfect 8 for the linking number.

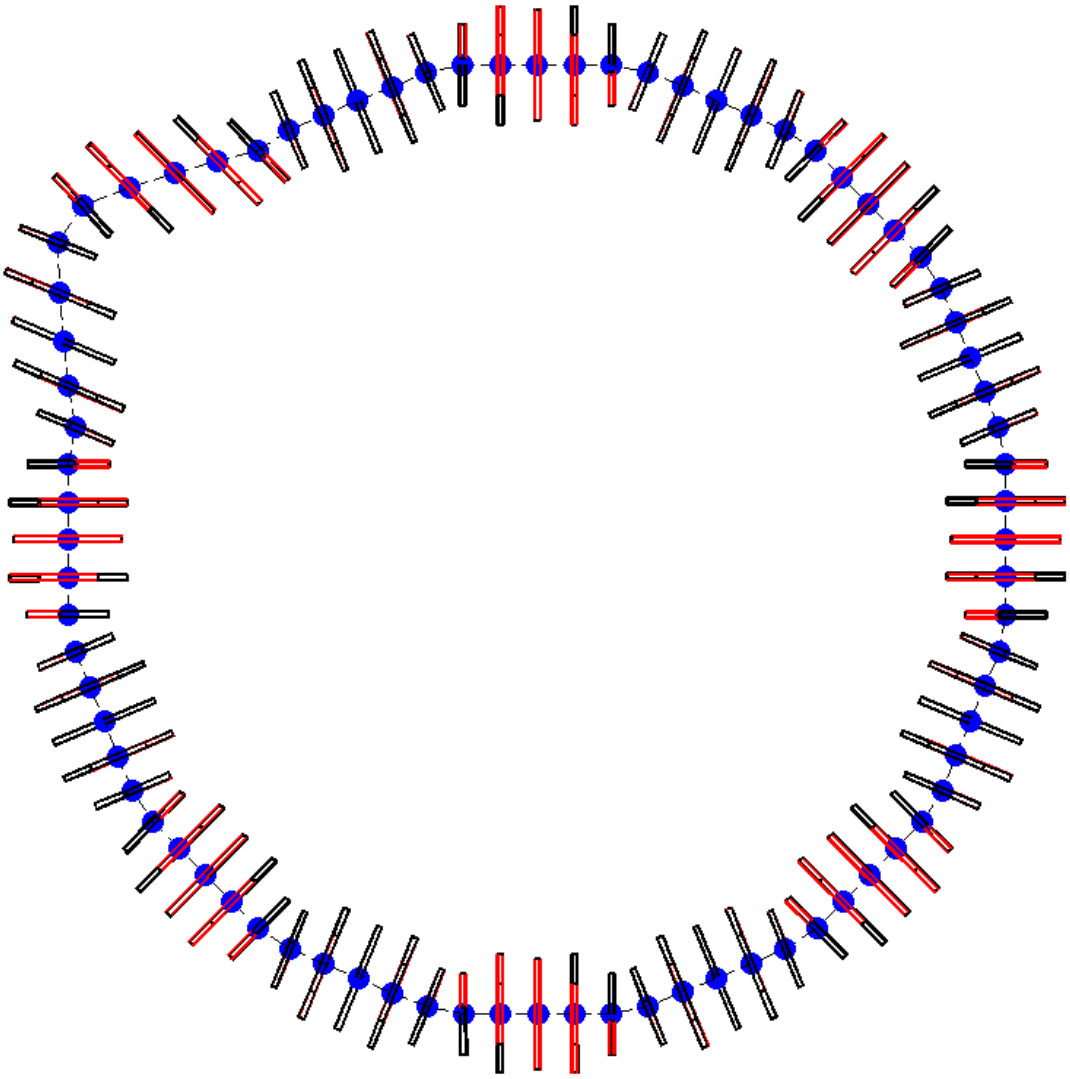


Figure 2.4.2.7.1: Top-down view of Model Bubble 2. The origins of the base pairs are represented by blue spheres, the base pairs are shown as rectangular slabs, and the axial curve is a dashed black line.

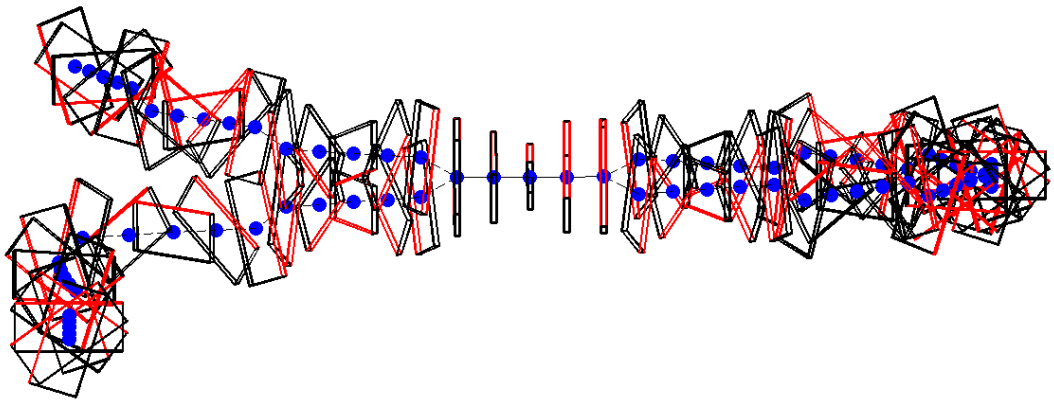


Figure 2.4.2.7.2: Side view of Model Bubble 2. The origins of the base pairs are small blue spheres, the base pairs are represented by rectangular slaves, and the black dotted line is the axial curve of the molecule.

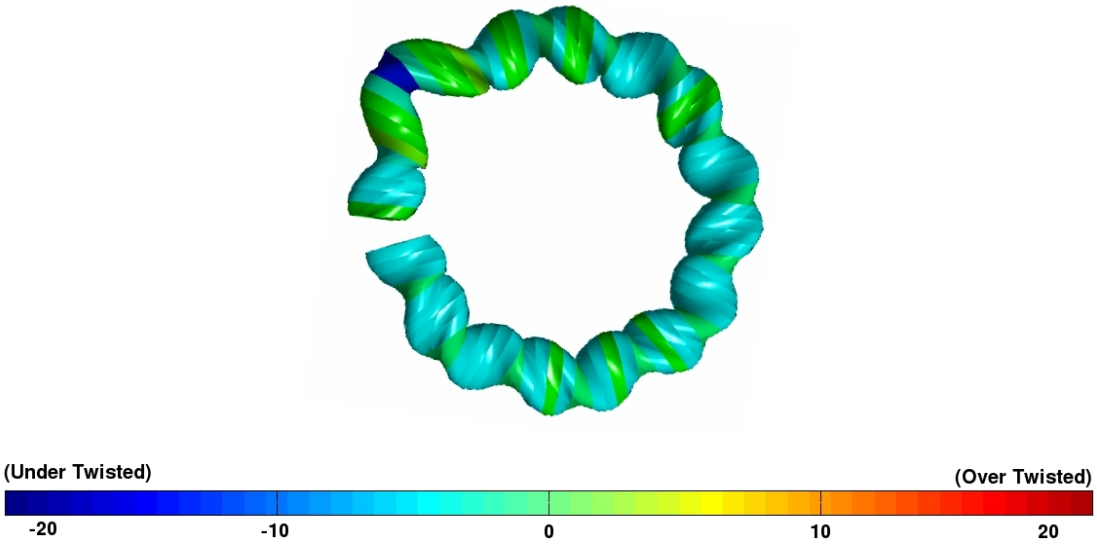


Figure 2.4.2.7.3: Color coded representation of the base-pairs step showing the degree of ΔTwist = $Tw^{SC} - Tw^{SP}$ of Model Bubble 2. Notice the large value of ΔTwist at the lower left-hand side of the figure where the peak of the bubble is located.

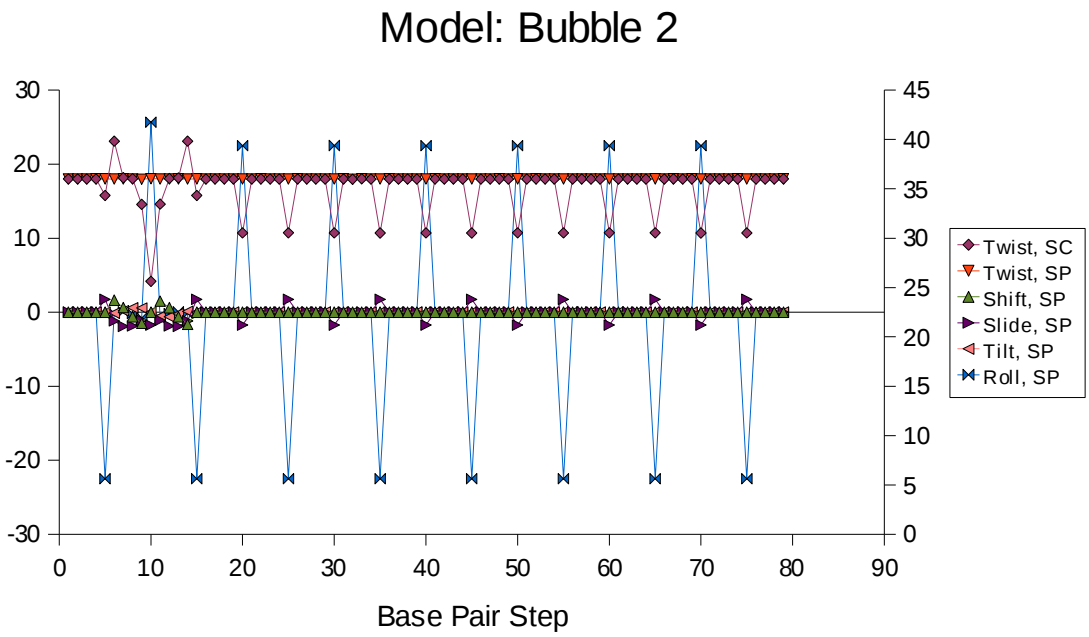


Figure 2.4.2.7.4: Graphical representation of Model Bubble 2. The most interesting steps are six through fourteen in the sequence. The rest of the molecule is the same as Model Kink 2.

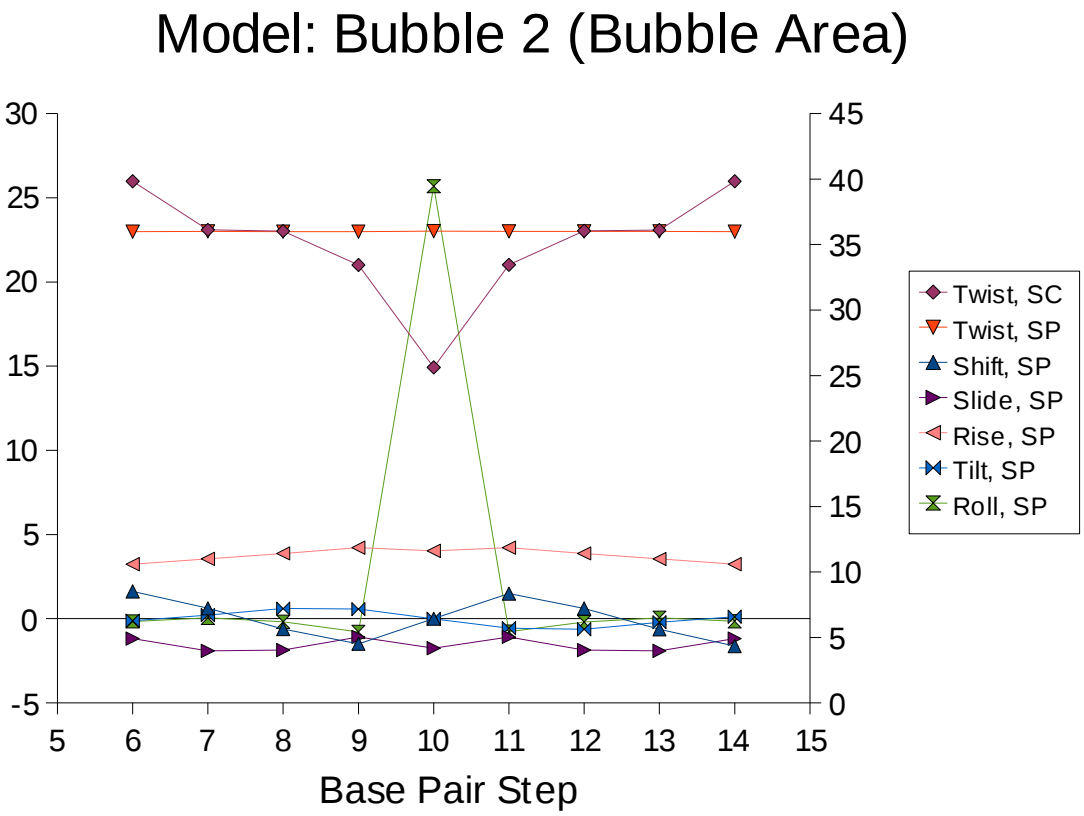


Figure 2.4.2.7.5: Close-up view of what is taking place at the bubbled portion of Model Bubble 2. Notice the large changes in the two types of twist plotted with respect to the ordinate on the right. The ordinate on the left is used to plot the rest of the step parameters — Shift, Slide, Rise, Tilt, and Roll.

Step	Shift	Slide	Rise	Tilt	Roll	Twist
	Å	Å	Å	°	°	°
6	1.62	–1.18	3.24	–0.12	–0.16	36
7	0.62	–1.91	3.56	0.21	0.05	36
8	–0.61	–1.86	3.88	0.60	–0.18	36
9	–1.49	–1.08	4.22	0.57	–0.78	36
10	0.00	–1.75	4.03	–0.01	25.69	36
11	1.49	–1.08	4.22	–0.56	–0.77	36
12	0.61	–1.86	3.88	–0.62	–0.20	36
13	–0.62	–1.91	3.55	–0.21	0.06	36
14	–1.62	–1.18	3.24	0.12	–0.15	36

Table 2.4.2.7.1: The specific values of the base-pair step parameters describing the bubbled portion of Model Bubble 2.

2.4.2.8 Summary of the 80 base-pair models

The previous sections were designed to give insight into how the variation of base-pair orientation and displacement contributes to the topology of a DNA segment or a closed molecule. The change in values of the step parameters can result in differences between the two types of twist. These differences are more noticeable when there are large chiral segments present. Tw^{SC} is very sensitive to changes in chirality and thus is an efficient way to gage the topological landscape of DNA. Table 2.4.2.8.1 summarizes the step parameter variation in all of the models discussed in Sections 2.4.2.1 - 2.4.2.7.

	Model						
	Kink 1	**Kink 2**	**Kink 3**	**Kink 4**	**Kink 5**	**Bubble 1**	**Bubble 2**
Shift, Å	0	0	0	0	0	0	0
Slide, Å	0	+/–1.75	0	2.5/–1	0	+/– 1.75	+/– 1.75
Rise, Å	3.4	3.4	3.4	3.4	3.4	3.4	3.4
Tilt, °	0	0	0	0	***	0	0
Roll, °	+/– 22.5	+/– 22.5	+/– 22.5	+/– 22.5	***	+/– 22.5	+/– 22.5
Twist, °	36	36	41/31	41/31	36	36	36
Total Turns (SP)	7.90022	8.12781	7.91389	7.68614	7.90003	7.66542	7.68743
Total Turns (SC)	7.90022	7.90000	7.91381	7.91383	7.90003	7.90006	7.89994
Total Turns (SP)**	8.00000	8.00000	8.01389	8.01389	8.00000	7.89992	7.89997
Total Turns (SC)**	7.99969	8.17290	8	7.82735	8	7.74351	7.84237
Writhe	0.00031	–0.17290	0	0.17265	0	0.25649	0.15763
Linking Number	8	8	8	8	8	8	8

Table 2.4.2.8.1: The rows outlined in cyan list the step parameter values for the varied base-pair steps. All other steps are: Shift = 0Å, Slide = 0Å, Rise = 3.4Å, Tilt = 0°, Roll = 0°, and Twist = 36°. The rows in the pink outlined section list the values of the two types of twists for an open molecule; the values are divided by 360° to give the total number of turns in the molecule. The section outlined in green adds a closing step between the last and first base pair to enable the linking number calculation. The Roll and Tilt values for Model Kink 5 are given in Table 2.4.2.5.1. The step parameter values listed in this table are for the steps that are not bubbled. The step parameters for the bubbled sections in Model Bubble 1 and 2 are located in Table 2.4.2.6.1 and Table 2.4.2.7.1.

*** Indicates a closing step was added.*

2.5 References

[1] D. Swigon, B.D. Coleman, and I. Tobias. (1998). The Elastic Rod Model for DNA and Its Application to the Tertiary Structure of DNA Minicircles in Mononucleosomes. *Biophys J.,* **74(5)**, 2515-2530.

[2] W.K. Olson, M. Bansal, S.K. Burley, R.E. Dickerson, M. Gerstein, S.C. Harvey, U. Heinemann, X.J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.S. Tung, E. Westhof, C. Wolberger, and H.M. Berman. (2001). A Standard Reference Frame for the Description of Nucleic Acid Base-Pair Geometry. *J. Mol. Biol*., **313(1)**, 229-237.

[3] M.A. El Hassan, and C.R. Calladine. (1995) The Assessment of the Geometry of Dinucleotide Steps in Double-Helical DNA: A New Local Calculation Scheme with an Appendix., *J. Mol. Biol.* **251(5)**, 648-664.

[4] K. Yanagi, G.C. Privé, and R.E. Dickerson. (1991). Analysis of Local Helix Geometry in Three B-DNA Decamers and Eight Dodecamers. *J. Mol. Biol.*, **217(1)**, 201-214.

[5] D. Bhattacharyya, and M. Bansal. (1989). A Self-Consistent Formulation for Analysis and Generation of Non-Uniform DNA Structures. *J. Biomol. Struct. Dynam.,* **6(4)**, 635-653.

[6] R.E. Dickerson, M. Bansal, C.R. Calladine, S. Diekmann, W.N. Hunter, O. Kennard, E. von Kitzing, R. Lavery, H.C.M. Nelson, W.K. Olson, W. Saenger, Z. Shakked, H. Sklenar, D.M. Soumpasis, C.S. Tung, A.H.J. Wang, and V.B. Zhurkin. (1989). Definitions and Nomenclature of Nucleic Acid Structure Parameters. *J. Mol. Biol.,* **205(4)**, 787-791.

[7] L.A. Britton, I. Tobias, and W.K. Olson. (2009). Two Perspectives on the Twist of DNA. *J. Chem. Phys.,* **131(24)**, 245101.

[8] V.B. Zhurkin, Y.P. Lysov, and V.I. Ivanov. (1979). Anisotropic Flexibility of DNA and the Nucleosomal Structure. *Nucleic Acids Res.*, **6(3)**, 1081-1096.

[9] A. Bolshoy, P. McNamara, R.E. Harrington, and E.N. Trifonov. (1991). Curved DNA without A-A: Experimental Estimation of All 16 DNA Wedge Angles.

Proc. Natl. Acad. Sci. USA, **88(5)**, 2312-2316.

[10] M.A. El Hassan, and C.R. Calladine. (1995). The Assessment of the Geometry of Dinucleotide Steps in Double-Helical DNA; a New Local Calculation Scheme. *J. Mol. Biol*., **251(5)**, 648-664.

[11] R.E. Dickerson. (1998). DNA Bending: The Prevalence of Kinkiness and the Virtues of Normality. *Nucleic Acids Res.*, **26(8)**, 1906-1926.

[12] X.J. Lu, and W.K. Olson. (1999). Resolving the Discrepancies Among Nucleic Acid Conformational Analyses. *J. Mol. Biol*., **285(4)**, 1563-1575.

[13] X.J. Lu, and W.K. Olson. (2003). 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res*., **31(17)**, 5108-5121.

[14] G. Călugăreanu. (1961). Sur les classes d'isotopie des noeuds tridimensionnels et leurs invariants. *Czech. Math. J*., **11(4)**, 588-625.

[15] J.H. White. (1969). Self-Linking and the Gauss Integral in Higher Dimensions. *American J. Math.*, **91(3)**, 693-728.

[16] F.B. Fuller. (1971). The Writhing Number of a Space Curve. *Proc. Natl. Acad. Sci. USA*, **68(4)**, 815-819.

[17] J.D. Watson, and F.H.C. Crick. (1953). Genetical Implications of the Structure of Deoxyribonucleic Acid. *Nature,* **171(4361)**, 964-967.

[18] F.B. Fuller. (1978). Decomposition of the Linking Number of a Closed Ribbon: A Problem from Molecular Biology. *Proc. Natl. Acad. Sci. USA,* **75(8)**, 3557-3561.

[19] J.H. White. (1989). *Mathematical Methods for DNA Sequences: An Introduction to the Geometry and Topology of DNA Structure*. CRC Press, Boca Raton, FL.

[20] R. Courant. (1959). *Differential and Integral Calculus, Volume II*. Blackie & Son Limited, London.

[21] M.R. Dennis, and J.H. Hannay. (2005). Geometry of Călugăreanu's Theorem. *Proc. R. Soc. A,* **461(2062)**, 3245-3254.

[22] W.K. Olson, M. Bansal, S.K. Burley, R.E. Dickerson, M. Gerstein, S.C. Harvey, U. Heinemann, X.J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.S. Tung, E. Westhof, C. Wolberger, and H.M. Berman. (2001). A Standard Reference Frame for the Description of Nucleic Acid Base-pair Geometry. *J. Mol. Biol.,* **313(1)**, 229-237.

[23] M. Levitt. (1983). Protein Folding by Restrained Energy Minimization and Molecular Dynamics. *J. Mol. Biol*., **170(3)**, 723-764.

[24] D. Bhattacharyya and M. Bansal. (1989). A Self-Consistent Formulation for Analysis and Generation of Non-Uniform DNA Structures. *J. Biomol. Struct. Dynam*., **6(4),** 635-653.

[25] M. Bansal, D. Bhattacharyya, and B. Ravi. (1995). NUPARM and NUCGEN: Software for Analysis and Generation of Sequence Dependent Nucleic Acid Structures. *Comput. Appl. Biosci*., **11(3)**, 281-287.

[26] M.H. Werner, A.M. Gronenborn, and G.M. Clore. (1996). Intercalation, DNA Kinking, and the Control of Transcription. *Science,* **271(5250)**, 778-784.

[27] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. (2000). The Protein Data Bank. *Nucleic Acids Res.,* **28(1)**, 235-242.

[28] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.H. Hsieh, A.R. Srinivasan, and B. Schneider. (1992). The Nucleic Acid Database. A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys J.,* **63(3),** 751-759.

[29] L.J. Peck, and J.C. Wang. (1981). Sequence Dependence of the Helical Repeat of DNA in Solution. *Nature*, **292(5821)**, 375–378.

[30] D. Rhodes, and A. Klug. (1981). Sequence-dependent Helical Periodicity of DNA. Nature, 292(5821), 378-380.

Chapter 3: TwiDDL (Twist of DNA Data Log)

3.1 Introduction to TwiDDL

The Twist of DNA Data Log, TwiDDL, is a database that focuses on the twisting of DNA and RNA double-helical structures. The objective in creating TwiDDL was to provide a database that presents a new conformational parameter called the twist of supercoiling, Tw^{SC} [1], and multiple factors which can influence that twist, including, but not limited to the six rigid-body base-pair-step parameters that describe the spatial disposition of neighboring base pairs [2]. TwiDDL has the capability to find structural examples where Tw^{SC} and other selected quantities lie in specified ranges and to search a wide range of criteria across every stored structure.

Each entry in TwiDDL includes easily manipulated, color-coded 3-D models that map the variation in Tw^{SC} on the double-helical structure. Data are reported in terms of the difference in twist, ΔTw, with respect to the twisting of canonical B DNA (10.5 base pairs per helical turn),

$$\Delta Tw^{B-DNA} = Tw^{SC} - Tw^{B-DNA} \qquad \textit{Eq 3.1}$$

,

and relative to the step parameter twist,

$$\Delta Tw^{SP} = Tw^{SC} - Tw^{SP} \qquad \text{Eq 3.2}$$

The models help the user to visualize and gain a better understanding of how the twisting is distributed throughout a DNA or RNA molecule and how the presence of proteins, drugs, and other forces affects the twist.

Structures that meet specific criteria can be found with two search methods. The first is a simple search with a few search options, and the second is an advanced search with many search features. Please note, there is also a quick search option in the upper right-hand corner of every TwiDDL page in which the user can enter a TwID (TwiDDL structural identifier), a PDB ID (Protein Data Bank structural identifier) [3], or an NDB ID (Nucleic Acid Database structural identifier) [4].

The simple search page gives the user the option of searching through a minimal number of fields to find structural examples with the desired features. While the advanced search page is similar to the simple search, it extends the search abilities by giving the user the option of searching most of the fields that are in the database and allowing a greater refinement in the structures of interest. The user can search any, all, or various combinations of the fields listed below (leaving all the fields empty will return the entire database). Clicking on any header in the search field creates a pop-up definition window, which explains what each field name means. The database is updated weekly, adding all new structures from the Protein Data Bank and Nucleic Acid Database

that meet TwiDDL's requirements for use.

3.1.1 What data are stored

The TwiDDL database contains a large variety of data about the physical construction of double-helical DNA and RNA. The data are organized in such a way that the information can be quickly and easily searched for characteristics that are related to the twist. In this section we will discuss the various data points and their relationship to the twist.

To make the search capabilities within TwiDDL very flexible, the database has been designed to evaluate data at both a base-pair step and a base-pair level for finer granularity searches. The database also contains a variety of details about the structure as a whole. When considering the data points discussed in this section, it is important to understand that some structures may have a considerable number of data points pertaining to them while others may have very few. This is due to the size of the structure and not a reflection on the importance of the data the structure provides in understanding the twist of supercoiling.

For each base-pair step in a molecule the database stores the dimeric (e.g., AC/GT) and tetrameric (e.g., GACT/AGTC) contexts of the step, the six base-pair-step parameters (Shift, Slide, Rise, Tilt, Roll, Twist also referred to as Tw^{SP}), the twist of

supercoiling (Tw^{SC}), the net kinking per base-pair step, and the net shearing per base-pair step. The individual base-pair-step parameter twist, or Tw^{SP} , one of the six rigid-body parameters commonly used to describe the spatial arrangements of neighboring base pairs in high-resolution DNA and RNA structures [2], is a primary focus for this database in its comparison to the twist of supercoiling or Tw^{SC}. The sum of the Tw^{SP} over all base-pair steps of a closed double helix plus the writhing number, a mathematical description of the global folding of the double-helical axis, does not necessarily give an integer. The twist of supercoiling, Tw^{SC}, a local conformational parameter, newly described for real DNA structures [1] and only available through TwiDDL, is consistent with the definitions introduced over forty years ago by mathematicians to characterize the intertwining of two closed space curves [5,6,7]. That is, the sum of the Tw^{SC} over a closed double helical structure plus the writhing number, a mathematical description of the global folding of the helical axis, gives a topologically invariant integer called the linking number. We have found that Tw^{SC} can give insight into the molecule as a whole or segments of interest, including the sensitivity of this parameter to molecular distortions associated with bound proteins and melted states. The database can be searched for a range of values or magnitudes for the difference in Tw^{SC} when compared to Tw^{SP}. Typical values of Tw^{SC} span the range 1° - 100°.

Additionally, at the base-pair step level the database stores, for easy search and

comparison, the calculated difference between the twist of supercoiling, Tw^{SC}, and that of either relaxed, ideal B-DNA (with 10.5 base pairs per full 360° helical turn), or the step-parameter twist (Eqs 3.1 and 3.2). Similarly, the database stores kinking, which is a measure of the net bending of a DNA or RNA base-pair step, calculated from the angles of the step parameters Tilt and Roll,

$$Kinking = \sqrt{\left(Tilt^2 + Roll^2\right)} \qquad Eq\ 3.3$$

The value of Tw^{SC} is correlated to the kinking and shearing of the base-pair step of interest. Shearing is a measure of the net lateral displacement of a DNA or RNA base-pair step measured in terms of the step parameters Shift and Slide,

$$Shearing = \sqrt{\left(Shift^2 + Slide^2\right)} \qquad Eq\ 3.4$$

The database can be searched for a range of values for either kinking or shearing, where typical kinking values span the range 0° to 175° and observed shearing values span the range 0Å to 9Å.

For each base pair in a molecule the database stores the chemical identities and residue numbers (Base ID I, Base ID II) of the paired bases in the selected structure, the identities of the complementary strands (Chain ID I, Chain ID II), the chemical identities of the paired bases (Residue ID I, Residue ID II), the numbers of the two bases (Residue Num I, Residue Num II), the six rigid-body base-pair parameters that describe the spatial

arrangement of paired bases (Shear, Stretch, Stagger, Buckle, Propeller, Opening), the coordinates of the origin of the base pair (OriginX, OriginY, OriginZ), and the coordinates for the three vectors of the base-pair slab. The latter parameters include the directions of (i) the short axis (ShortAxisX, ShortAxisY, ShortAxisZ) – the vector pointing across the base pair toward the major groove, (ii) the long axis (LongAxisX, LongAxisY, LongAxisZ) – the vector pointing toward the strand that carries the base sequence, and (iii) the normal (NormalX, NormalY, NormalZ) – the vector defined perpendicular to the mean base-pair plane. Buckle, one of the six rigid-body parameters commonly used to describe the spatial arrangements of hydrogen-bonded Watson-Crick base pairs in high-resolution DNA and RNA structures [2], is of particular interest because of its potential correlation with the twist. The values of Buckle stored in the database correspond to the angle between the normals of the bases when viewed in projection along a plane perpendicular to the short, pseudo-dyad axis of the base pair [8]. The database can be searched for a range of values or magnitudes for Buckle, where typical values span the range –90° to 90°.

For the structure as a whole, the database contains fields for the TwID (TwiDDL's unique structural identifier), the PDB ID (Protein Data Bank structural identifier), the NDB ID (Nucleic Acid Database structural identifier), the experimental method used to characterize the atomic structure, the sequence, the resolution, the space group, the title,

the primary citation, the classification, the number of base pairs, the Total Tw^{SC}, the Total $Turns^{SC}$, the Total Tw^{SP}, the comments field, and the data workup directory. The totals contain all the values for twist summed up for the whole DNA segment. The data stored in TwiDDL is based on structures taken either from the Protein Data Bank (PDB) or the Nucleic Acid Database (NDB). TwiDDL stores both database identifiers, PDB ID and NDB ID, when available for the same structure, in order to allow ability to look up familiar structures and obtain data describing the twisting properties of the double-helical DNA or RNA in those structures. It is important to note that not all structures in the PDB or NDB pass the standards needed to be included in TwiDDL, so there will be structures that cannot be found because of those requirements.

One distinguishing characteristic within the database is how it handles the TwID, TwiDDL's unique structural identifier, versus the PDB's use of the PDB ID. Both TwiDDL and the PDB store data about the experimental method, either NMR (nuclear magnetic resonance) spectroscopy or X-Ray crystallography, utilized to characterize the structure, but within TwiDDL the TwID actually corresponds to a single structure, making it possible to differentiate each of the structures that make up the ensemble of models derived from NMR studies and sometimes constitute a single crystal structure. That is, even though all the models included in an NMR-based structural entry have the same PDB ID, each NMR model has its own TwID. TwiDDL therefore can have multiple

entries for the same PDB ID. TwiDDL does this because each NMR model may yield different values significant for understanding the fluctuations in twist in the structure. X-Ray crystallographic based models, however, typically will have a single TwID for each PDB ID because there is only a single model within that PDB entry.

In TwiDDL the resolution entry associated with a structure refers to the resolution of the X-ray crystallographic data. X-ray crystal structures are determined at different levels of accuracy, or resolution, based on the interplanar spacing of reflected X-rays. The lower the resolution, the more accurate (or better resolved) the data. The sequence stored in TwiDDL refers to the nucleotides that constitute the leading strand of the DNA or RNA in the specified structure. The sequence is limited to a combination of the five common bases (A, C, G, T, U) and is determined, along with the number of base pairs, by 3DNA [8]. Thus, chemically modified bases are ignored. The space group, the notation describing the symmetry of the crystal and specifying the combination of symmetry operations that generate the crystal lattice from the unit cell, is also stored in the database. The resolution and space group, along with the crystal title, the journal name and primary citation, and the structure classification are all taken from the Protein Data Bank.

The Total Tw^{SC} is the sum of the Tw^{SC} values for all base-pair steps in the structure,

$$Total\,Tw^{SC} = \sum_{i=1}^{nbp} Tw_{(i)}^{SC} \qquad Eq\ 3.5$$,

where nbp is the number of base pairs in the structure. Similarly, the Total Tw^{SP} is the sum of the Tw^{SP} values for all base-pair steps in the structure,

$$Total\,Tw^{SP} = \sum_{i=1}^{nbp} Tw_{(i)}^{SP} \qquad Eq\ 3.6$$.

The Total $Turns^{SC}$, the number of helical turns in the selected structure, is derived only from the Total Tw^{SC},

$$Total\,Turns^{SC} = \frac{Total\,Tw^{SC}}{360^{o}} \qquad Eq\ 3.7$$.

These calculated values help demonstrate the significance of the differences between the classical step parameter twist, Tw^{SP}, and the new twist of supercoiling, Tw^{SC}, in the context of the overall structure.

In order to support TwiDDL as a fully automated system that updates its content automatically the database was designed to include a comment field. This comment field is hidden from the end users because it stores information about the capability of the structure to conform to TwiDDL's requirements for generating valid data about the twist of the structure. The comment field is essentially used as a flag to indicate whether the data should or should not be displayed about a structure. If there are no comments

present and the comments field is null, then it is assumed that the data are valid and can be displayed to end users. However, if there are comments about the structure, then the data are essentially flagged for review by the administrator, and cannot be viewed through the web interface. The automation in TwiDDL currently catches three known cases that are flagged in the comments field, and supplies information about how the structure does not comply with TwiDDL's requirements. These cases include structures, such as (i) the first NMR model of PDB entry 1fzs (the structure of DNA with pyrene paired at abasic sites [9]), in which Chain ID I or II has multiple Chain letters (meaning that one of the strands of the double-helical structure is nicked, i.e., made up of two different chains); (ii) the fourth and seventh NMR models of PDB entry 3php (the structure of the 3´-hairpin of the TYMV pseudoknot [10]), in which the Residue Number I or II deviates by more than one between steps (meaning that the intervening base pairs loop/flip out or form some other type of structure); or (iii) the sixth NMR model of PDB entry 1wwf (the structure of a nucleocapsid protein of Moloney murine leukemia virus bound to RNA [11]), which is made up of only a single base-pair (meaning that the fragment is insufficient for the twist of supercoiling calculations, which require at least three successive base-pair steps in the structure). Beyond the three automated cases, the administrator can review the data manually and choose to insert a comment about why the structure should be removed from the list. This gives the administrator full control over the data displayed, and is why any comment is automatically picked up and used to remove the structure from the list of

viewable data. After the administrator reviews the data of such a structure and determines the data to be valid, the structure can easily be listed in the TwiDDL database by manually removing the comment. Many important structures may not be found in TwiDDL due to the three criteria discussed.

In the following sections we will discuss the sources for all of the data in the database, as well as the calculations based on that data. As it will be shown, there are a variety of sources for the data both from external sources, as well as from applications that are used in TwiDDL's new model for the twist of supercoiling. To address this, the database stores an entry for each structure that points to a directory on the server that contains all of the sources of data, including the new data that were obtained using the twist of supercoiling model. This database entry for these derived data provides a clear connection between the data stored in the database, and the raw sources that were used to supply this information. It also allows the raw data to be viewed if desired.

3.1.2 Raw Data

3.1.2.1 Data Sources

The data outlined in Section 3.1.1 are obtained through several sources. The PDB and NDB were already noted as sources of data for the structures stored in TwiDDL. The remainder of the data in TwiDDL are derived by manipulation of the data supplied via the

PDB through either the 3DNA software package [8] and new programs that calculate the twist of supercoiling. The NDB is used primarily as a source for correlating the TwID, the PDB ID, and the NDB ID such that a structure has a single entry. The atomic coordinates from the PDB are the basis for all of the calculations used to generate the other data within the database. In particular, the PDB is the primary source of information about the structure, including its crystal title, journal name and primary citation, classification, resolution, and space group, as well as the coordinate data used to generate the twist of supercoiling models that this database focuses on. All of this can easily be accomplished because PDB documents follow a standardized format that can be parsed to extract the subsets of relevant information [12].

As previously discussed, TwiDDL generates a single entry for each X-Ray crystallographic structure, as well as multiple entries for NMR structures, in which case an entry is made for each NMR model. For each of these entries, a copy of the PDB file is stored in a directory signifying each model, where each model is either stored in a directory name based on the PDB ID, such as 1aay, which is the identifier for the Zif268 zinc finger-DNA complex crystal structure [13], or in the case of multiple NMR structures the directory name contains the PDB ID with an additional enumeration for each model using a dash and a number, such as 4kbd-1, 4kbd-2, and 4kbd-3, which are three NMR models of a DNA structure with a mutated kappa B site [14].

Within the directories for each model, several files contain the raw data for entry into the database. The PDB data are the first file stored in raw form, i.e., the original PDB document format. To address structures with multiple models, this file is stored in one of two ways, either as provided by the PDB or in the per NMR model format generated by TwiDDL. The files in the per NMR model format are nearly identical to that in the original (first) PDB file, except for the removal of all NMR models other than the single model denoted by the TwID. The naming scheme for the PDB file is identical to that of the directory in which it is stored except for the addition of the three-letter ".pdb" filename extension. That is, the filename is either based on the PDB ID, such as 1aay.pdb in the case of the crystal structure for the Zif268 zinc finger-DNA complex, or 4kbd-1.pdb, 4kbd-2, and 4kbd-3.pdb in the case of the NMR structure of a DNA structure with a mutated kappa B site.

In the directory of every structure the PDB file is copied to a file called out.pdb, which is used as input for the 3DNA software package. The files generated by 3DNA include: (i) auxiliary.par, which contains the base, base-pair, middle, and helical reference frame coordinates utilized by the twist of supercoiling calculations; (ii) bestpairs.pdb, which contains the atomic coordinates of the adjacent bases in the DNA strand and identifies the complimentary bases in the adjoining strand; (iii) bp_helical.par, which contains the six rigid body base-pair and helical parameters of the structure; (iv)

bp_order.dat, which contains base-pair data prior to re-ordering; (v) bp_step.par, which contains the base-pair and base-pair step parameters for the final structure; (vi) cf_7methods.par, which contains the base-pair and base-pair step parameters in formats used by seven popular programs for analyzing nucleic acids (CEHS, CompDNA, Curves, FreeHelix, NGEOM, NUPARM, and RNA); (vii) col_chains.scr, which contains the coloring settings for each strand in the structure; (viii) col_helices.scr, which contains the coloring settings for each helix in the structure; (ix) hel_regions.pdb, which contains the atomic coordinates of the DNA atoms separated by base number and helical strand; (x) hstacking.pdb, which contains the atomic coordinates of each dinucleotide step relative to its middle helical frame; (xi) inputcoord.txt, which contains the coordinates for the origins and the vectors for the short and long axis of each base-pair; (xii) origins.csv, which contains the coordinates for the origins of each base-pair; (xiii) out.inp, which contains the data for matching bases on individual strands into their base-pairs; (xiv) out.out, which contains the main x3DNA output including detailed information about bases, base-pairs, base-pair steps, helices, strands, and more; (xv) ref_frames.dat, which contains the origins, short, long and normal axis for each base-pair; and (xvi) stacking.pdb, which contains the atomic coordinates of each dinucleotide step relative to its middle step frame [15]. TwiDDL currently uses only four of the 3DNA generated files: origins.csv, out.out, bp_step.par, and inputcoord.txt files.

The origins.csv file is used to calculate the number of base pairs within a structure. The out.out file provides, for each base-pair step, the dimeric context, tetrameric context, rigid body parameters of each base pair (shear, stretch, stagger, buckle, propeller, opening), the base identifiers (Base ID I & II), the chain identifiers (Chain ID I & II), the residue identifiers (Residue ID I & II), and the residue numbers (Residue Num I & II). The out.out file is used to generate the sequence as it is stored by TwiDDL, and is also the source for the automated evaluation of whether the structure conforms to TwiDDL's requirements and the storage of this information in the comments. The bp_step.par and inputcoord.txt files are used as input for the twist of supercoiling calculations performed by a stand-alone Matlab application created specifically for TwiDDL. Both files are copied into the same directory with the stand-alone twist of supercoiling application. In addition to these two files, the twist of supercoiling calculation also uses the number of base pairs (given as a command line option) to generate the twistwrithe.csv file.

The directories, containing all of the raw data, store a few other items utilized by TwiDDL as critical pieces of the web interface that is specific for each structure. Beyond the database, one of the key features of the TwiDDL web site is the visualizations of the derived data. The files that support these visualizations are stored for each structure in two formats. A png format is used for a series of three images; graph-scatter.png, graph-

points.png, and graph-lines.png. These three png files respectively contain the dynamically generated line, point, and scatter graphs, which are based on the values of the Tw^{SP} and Tw^{SC} along the double helical structure. A txt format is used for a text file, called thetarhotwsc.txt,that contains the data required to generate the real-time 3D model displayed on the detail page for the structure. The thetarhotwsc.txt file is read by the TwiDDLPlots Java Servlet which utilizes Matlab code to render the 3D image. This image shows the difference ΔTw between the twist of supercoiling Tw^{SC} and, depending on the view selected, either the step parameter twist Tw^{SP} or the ideal twist of B-DNA $Tw^{B\text{-}DNA}$ with a value of 34.28º based on 10.5 base-pairs per helical turn.

3.1.2.2 Calculations

The data and files discussed in the previous sections are a result of various calculations. Some are derived from the twist of supercoiling calculation developed in this thesis, some are performed to highlight the significance of the twist of supercoiling, and some are previously established calculations. The latter are used as references to compare and contrast the twist of supercoling from conventional measures of DNA twist. This section focuses on the new calculations performed to obtain the twist of supercoiling or to display the data on the web interface.

Each structure entered into the TwiDDL database has a twistwrithe.csv file generated with the twist of supercoiling application. This file contains all of the critical

data about the structure generated by the twist of supercoiling model. This file is then parsed by TwiDDL into its various sub-components and the individual data points are stored in the database for quick and easy searching. The fields in the database that come from twistwrithe.csv include the base-pair index, Tw^{SC}, ΔTw^{SP}, the six rigid-body parameters at each base-pair step (Shift, Slide, Rise, Tilt, Roll, Twist or Tw^{SP}), the coordinates of the origin and the three vectors (short axis, long axis, normal) of the base-pair slab, the sum of Tw^{SC} over the DNA sequence, and the number of helical $Turns^{SC}$. These data supply the basic structural information that is required for the analysis, comparison, and understanding of the twist of supercoiling that TwiDDL aims to provide.

Within the TwiDDL database there are several other stored values which relate to and are based on data extracted from the twistwrithe.csv file. These data points are calculated at the time of insertion into the database, and are not in any of the raw data files. In other words, the following database entries are found only within TwiDDL because their values are calculated as a byproduct of the analyis of the raw data files. The value of ΔTw^{SC}, the difference between the twist of supercoiling and the step parameter twist, is calculted for each base-pair step in the structure. The former quantity is based on the coordinates of four base pairs (three base-pair steps) and the latter on the coordinates of only two base pairs (one base-pair step),

$$\Delta Tw^{SC} = Tw^{SC} - Tw^{SP} \qquad \textit{Eq 3.8}$$,

where the step parameter twist, Tw^{SP}, and the twist of supercoiling, Tw^{SC}, are both data points that exist in the twistwrithe.csv. The difference $\Delta Tw^{B\text{-}DNA}$ between the twist of supercoiling and the twist of canonical B DNA, $Tw^{B\text{-}DNA}$, is also calculated and inserted in the database. The canonical B DNA is assumed in this thesis to contain 10.5 base pairs per helical turn, corresponding to a twist angle $Tw^{B\text{-}DNA}$ of ~34.3º, i.e., the quotient of the number of degrees in a single helical turn and the number of base pairs in that helical turn (360º/10.5). The value of $\Delta Tw^{B\text{-}DNA}$ is thus given by:

$$\begin{aligned} \Delta Tw^{B-DNA} &= Tw^{SC} - Tw^{B-DNA} \\ &= Tw^{SC} - \left(\frac{360^o}{10.5}\right) \end{aligned} \qquad \textit{Eqn 3.9}$$

The kinking and shearing of the base-pair step are calculated with Eq 3.3 and Eq 3.4 using four of the six step parameters in the twistwrithe.csv, which are tilt, roll, shift, and slide.

In addition to the preceding descriptors of structure at the base-pair-step level, TwiDDL includes other data for comparison and analysis of the structure as a whole. The latter calculations focus on the total twist of the overall structure, Total Tw^{SP}, which is calculated as the sum of all of the base-pair step parameter twist values (Eq 3.6). The value of Total Tw^{SP} shows how the overall step parameter twist compares to the overall twist of supercoiling of the structure. The direct comparison can be seen in the ΔTw value for the overall structure:

$$\Delta Tw = Total\,Tw^{SC} - Total\,Tw^{SP} \qquad Eq\ 3.10$$

The value of Total Tw^{SC}, described in Eq 3.5, is calculated in the twist of supercoiling calculations. The differences described in Eq's 3.9 and 3.10 help to truly demonstrate the relative difference in the overall twist measured by the twist of supercoiling and the step parameter twist.

In addition to the calculations performed as part of the twist of supercoiling calculations or as part of processing that data for insertion into the TwiDDL database, the database itself includes internal calculations. While a database is invaluable for storing information, it is also beneficial in the analysis of that data. Section 3.1.1 presents the various data points stored in the database, which are valuable in showing the nuances that the twist of supercoiling shows about the structures. In the design of TwiDDL a primary focus was to show how the twist of supercoiling differs from the step parameter twist. An important part of demonstrating that difference relies on the ability to analyze the trends in the data that arise in the evaluation of the twist of specific structures. In order to obtain this perspective on the data the minimum, maximum, and average values are calculated for the Tw^{SC}, ΔTw^{SC}, $\Delta Tw^{B\text{-}DNA}$, Tw^{SP}, $Shift^{SP}$, $Slide^{SP}$, $Rise^{SP}$, $Tilt^{SP}$, and $Roll^{SP}$ in the selected structures. These values are grouped by the respective dimer or tetramer step contexts, and can be calculated for a single structure or multiple structures found by a search. In the case of the average values, the number of matching dimers or tetramers

used to calculate the average are counted and displayed in order to help highlight the significance of the average. For example, an average based on a handful of matching dimers may or may not be reflective of typical values for that dimer step context. An average based on thousands of matching dimers would hold greater statistical significance.

3.2 Database Design

All of the data discussed in previous sections of this chapter could be accessed in a variety of ways. The first and most basic approach would be to run the applications to calculate the data for each structure. This approach is inefficient, cumbersome, and time consuming, especially if the information is not reliably stored. The user would have to run and rerun the calculations each time he or she wanted to see the information about a structure. The second and slightly more effective way of accessing this data would be to store the output of the applications into a file, such as the twistwrithe.csv discussed previously, and then collect those flat files in a hierarchy of directories. The problem with this methodology is that there is no simple way to go through the data, especially since there are thousands and thousands of structures that could be of potential interest. This problem quickly identifies some basic goals, the first being the ability to retrieve the data provided in the applications quickly and efficiently, and the second being able to sort and search through data for a large number of structures.

Throughout my research I have noticed that many others develop and maintain databases to solve these challenges [16]. In fact, much research on nucleic acid structure is based on information in the Nucleic Acid Database (NDB). In keeping with that tradition, and in order to solve the challenges of efficient data search and retrieval, I developed TwiDDL using MySQL as the database of choice. MySQL is a free open source package that is highly reliable and used not only by researchers, but also by large companies [17]. To use and maintain this database one needs to know how to use its command line language, which is based on the ANSI standard SQL (Structured Query Language) [18]. The manuals online provided by MySQL are very detailed and helpful, but one needs to know what to look for and to have a starting idea of how to make one's way through this documentation [19]. I invested considerable time during my thesis work developing a solid understanding of the SQL in the MySQL database in order to develop the backend TwiDDL schema.

A database is just that, a place in which to store data. It is only as good as the information entered into it in terms of both accuracy and completeness. In its original form, the TwiDDL database only held the PDB ID, the title of the crystal data, a pointer to the twist of supercoiling workup file, the resolution of the crystal structure, a pointer to a file from the 3DNA workup, and some twist of supercoiling values taken from the workup file. As my thesis progressed the requirements for data stored in the database

radically changed, leading to the storage of all of the data discussed in the previous sections of this chapter. In this section we will focus on how that data are stored in a database, how the schema are designed based on that data, and how the schema are used to drive features in the web interface discussed below. This is the foundation by which TwiDDL derives its features and functionality, utlimately enabling an ease of use that hides the details of the underlying data and its structure. Even though the schema are hidden to the end users, it was critically important to the development of the database to understand database design fully and then develop a plan baseed on this design to make access to the data simple and useful.

I invested considerable effort on enhancing the database to enable greater flexibility in the way the information could be accessed. Proper design is critical to provide the storage and query abilities the end user will require from a database like this. I spent several weeks learning about the different data types that MySQL supports and looking into proper database structure. Once I had a good idea of what data types were available, I drafted several potential relational database options for my data. Since my programs generate a variety of data that would be beneficial to search through, it was very important to choose the proper data type for each of the items to be stored and the proper organization of the data types into relevant groupings.

3.2.1 Database Structure

The final design for the database has expanded the amount of data that could be contained compared to the earlier versions developed during the course of my research. As one can see from Figure 3.2.1.1, the TwiDDL database is constructed from two major tables. The table names are not as important as the data contained within them, but proper name selection makes the database intuitive for writing SQL Queries. The two tables that make up the TwiDDL database and store all the data discussed in the previous sections are the Summary table and the Details table. The names are somewhat descriptive of the contents of each table.

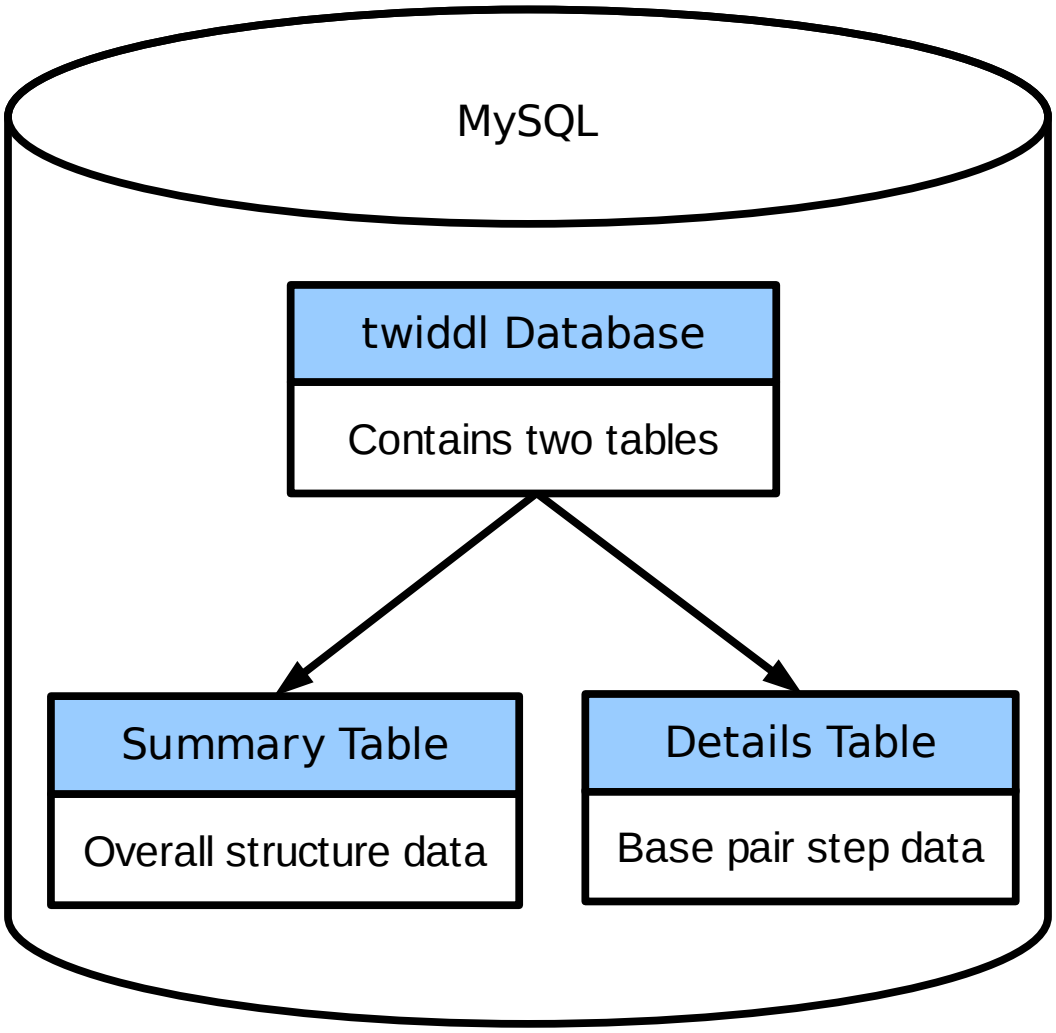


Figure 3.2.1.1: The TwiDDL database is built upon MySQL and designed to contain two tables of information about the structures. The Summary table contains the data pertaining to the overall structure, and the Details table contains the fine grain data about each base-pair step in the structure.

Summary Table

	Field	Type	Null	Key	Default	Extra
1	TWID	int(8) unsigned zerofill	NO	PRI	NULL	auto_increment
2	PDBID	varchar(16)	NO	MUL		
3	NDBID	varchar(16)	NO	MUL		
4	CrystalTitle	text	YES		NULL	
5	PrimaryCitation	text	YES		NULL	
6	DataWorkUp	text	YES		NULL	
7	Resolution	decimal(4,2)	YES		NULL	
8	SpaceGroup	text	YES		NULL	
9	TotTwSC	decimal(20,15)	YES		NULL	
10	TotTurnsSC	decimal(18,15)	YES		NULL	
11	NBP	int(10)	YES		NULL	
12	Sequence	text	YES		NULL	
13	DeltaTw	decimal(20,15)	YES		NULL	
14	TotTwSP	decimal(12,8)	YES		NULL	
15	ExpMethod	enum('XRayCrystal','NMR')	YES		XRayCrystal	
16	Title	text	YES		NULL	
17	Classification	text	YES		NULL	
18	Comments	text	YES		NULL	

Figure 3.2.1.2: The Summary table consists of 18 columns of data that summarize the structure as a whole. Various data types are used to store the information appropriately.

The data in the Summary table are contained in eighteen columns that summarize the results from running a particular crystal structure from the PDB through the twist of supercoiling program. As seen from Figure 3.2.1.2 the Summary table contains details of the crystal structure from the various sources discussed in Section 3.1.2.1, where all of these details pertain to the crystal structure as a whole giving a high level summary of the structure. The Summary table contains data generated by TwiDDL, as indexed by the TwID (item 1), and extracted from the raw data work up directory (item 6), which was calculated by executing the twist of supercoiling program. The extracted data in the Summary table include the total twist of supercoiling (item 9), the total turns calculated for that structure (item 10), the number of base pairs (item 11), the sequence of the crystal

structure (item 12), the difference between the twist of supercoiling and the step parameter twist (item 13), the total step parameter twist (item 14), and any comments about the compliance of the structure with the requirements of the twist of supercoiling calculations (item 18). The Summary table also contains information supplied from the PDB file such as the PDB ID (item 2), the crystal title (item 4), the primary citation (item 5), the resolution (item 7), the space group (item 8), the experimental method (item 15), the journal name (item 16), and the classification (item 17). As discussed above the PDB ID may be modified to reflect the respective model within an NMR structure. The last data source is the NDB which only supplies the NDB ID (item 3).

Details Table

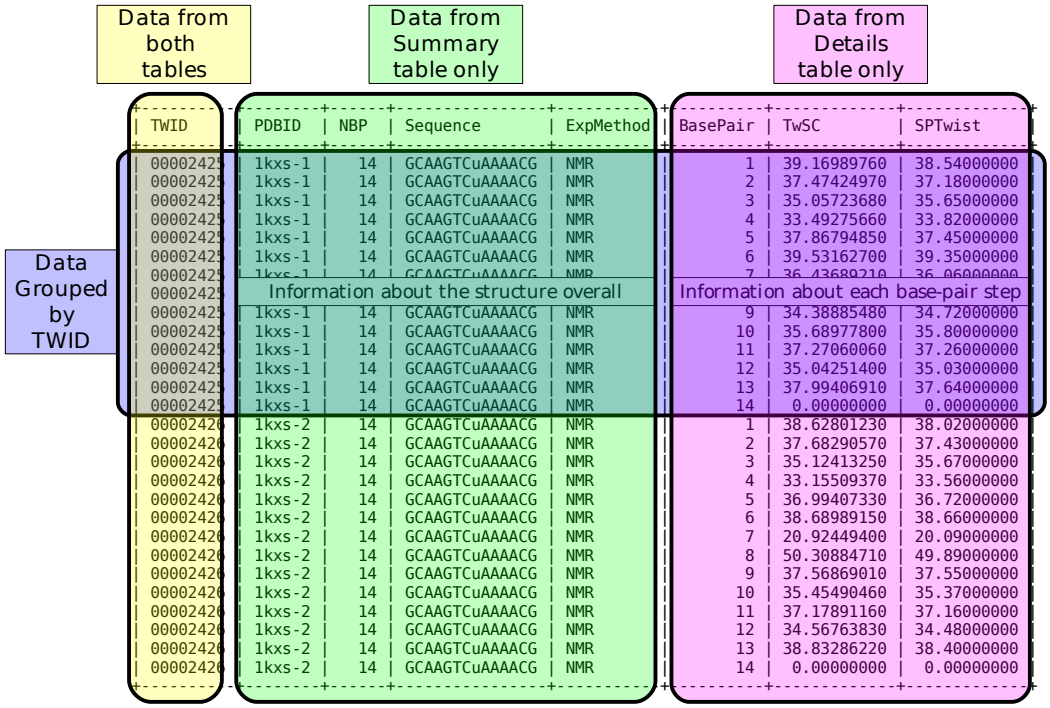
	Field	Type	Null	Key	Default	Extra
1	TWID	int(8) unsigned zerofill	NO	MUL	NULL	
2	BasePair	int(6)	YES	MUL	NULL	
3	TwSC	decimal(12,8)	YES		NULL	
4	DeltaTwSC	decimal(12,8)	YES		NULL	
5	SPTwist	decimal(12,8)	YES		NULL	
6	SPShift	decimal(12,8)	YES		NULL	
7	SPSlide	decimal(12,8)	YES		NULL	
8	SPRise	decimal(12,8)	YES		NULL	
9	SPTilt	decimal(12,8)	YES		NULL	
10	SPRoll	decimal(12,8)	YES		NULL	
11	BaseID	varchar(8)	YES		NULL	
12	BaseID'	varchar(8)	YES		NULL	
13	ChainIDI	varchar(2)	YES		NULL	
14	ChainIDII	varchar(2)	YES		NULL	
15	ResidueIDI	varchar(2)	YES		NULL	
16	ResidueIDII	varchar(2)	YES		NULL	
17	ResidueNumI	int(7)	YES		NULL	
18	ResidueNumII	int(7)	YES		NULL	
19	Step	varchar(5)	YES	MUL	NULL	
20	OriginX	decimal(12,8)	YES		NULL	
21	OriginY	decimal(12,8)	YES		NULL	
22	OriginZ	decimal(12,8)	YES		NULL	
23	ShortAxisX	decimal(12,8)	YES		NULL	
24	ShortAxisY	decimal(12,8)	YES		NULL	
25	ShortAxisZ	decimal(12,8)	YES		NULL	
26	LongAxisX	decimal(12,8)	YES		NULL	
27	LongAxisY	decimal(12,8)	YES		NULL	
28	LongAxisZ	decimal(12,8)	YES		NULL	
29	NormalX	decimal(12,8)	YES		NULL	
30	NormalY	decimal(12,8)	YES		NULL	
31	NormalZ	decimal(12,8)	YES		NULL	
32	Shear	decimal(12,8)	YES		NULL	
33	Stretch	decimal(12,8)	YES		NULL	
34	Stagger	decimal(12,8)	YES		NULL	
35	Buckle	decimal(12,8)	YES		NULL	
36	Propeller	decimal(12,8)	YES		NULL	
37	Opening	decimal(12,8)	YES		NULL	
38	Tetramer	varchar(9)	YES		NULL	
39	Kinking	decimal(12,8)	YES		NULL	
40	Shearing	decimal(12,8)	YES		NULL	
41	DeltaTwBDNA	decimal(12,8)	YES		NULL	

Figure 3.2.1.3: The Details table consists of 41 columns of data that provide insight into the physical structure at the base-pair and base-pair step level. Like the Summary table, various data types are used in the Details table to store the information appropriately.

Fitting all of the data from the twist of supercoiling calculations into a single table would not be a manageable solution, so in addition to the Summary table there is the Details table that contains the remainder of the data in a coordinated fashion. Figure 3.2.1.3 describes the organization of the Details table. The data stored in the Details table

contain information from the twist of supercoiling calculations, for each base-pair in a given structure, as well as information about the structure at a base-pair-step level. This is in contrast to the Summary table that stores information that is applicable to the structure as a whole. The Details table contains the TwID (item 1), the base-pair index (item 2), Tw^{SC}(item 3), ΔTw^{SP}(item 4), $\Delta Tw^{B\text{-}DNA}$ (item 41), the six base-pair-step parameters (Shift, Slide, Rise, Tilt, Roll, Twist or Tw^{SP}) (items 6-10 and 5, respectively), the coordinates of the origin of the base-pair slab (items 20-22), the three vectors (short axis, long axis, normal) of the base-pair slab (items 23-25, 26-28, and 29-31, respectively), the dimeric context of the base-pair step (item 19), the tetrameric context of the base-pair step (item 38), the six rigid-body base-pair parameters, describing the spatial arrangement of paired bases (Shear, Stretch, Stagger, Buckle, Propeller, Opening) (items 32-37), the identifiers of the paired bases (items 11,12), the identifiers of the chain to which the paired bases are attached (items 13,14), the names of the individual bases attached to either strand I or II (items 15,16), the number of the residue which contain the paired bases (items 17,18), the kinking and the shearing of the base-pair step (items 39,40). The incorporation of base-pair and base-pair-step data into a single table improves performance, as discussed in Section 3.2.2. The combination is accomplished by assuming there will always be a NULL value for the last base-pair step, since there is usually one more base-pair than base-pair step in a structure.

These two tables cross reference one another through a common identifier, the TwID. In previous implementations of the database the PDB ID was used as this identifier, but a change was required in order to extend the functionality of the database to PDB files that contained multiple models of the same structure. This change to a TwID instead of the PDB ID makes it possible to distinguish between multiple PDB models easily and examine the variation in the twist of supercoiling in NMR structures. The diagram in Figure 3.2.1.4 shows how the TwID can be used to link the data from the two tables, and concomitantly keep individual models separate despite having the same PDB. This common link allows for greater flexibility in the storage and search of various data points generated by the models, and in turn this link increases the value of the model by making the results easily accessible and manageable.



TWID	PDBID	NBP	Sequence	ExpMethod	BasePair	TwSC	SPTwist
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	1	39.16989760	38.54000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	2	37.47424970	37.18000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	3	35.05723680	35.65000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	4	33.49275660	33.82000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	5	37.86794850	37.45000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	6	39.53162700	39.35000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	7	36.43689210	36.06000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	8	36.43689210	36.06000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	9	34.38885480	34.72000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	10	35.68977800	35.80000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	11	37.27060060	37.26000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	12	35.04251400	35.03000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	13	37.99406910	37.64000000
00002425	1kxs-1	14	GCAAGTCuAAAACG	NMR	14	0.00000000	0.00000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	1	38.62801230	38.02000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	2	37.68290570	37.43000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	3	35.12413250	35.67000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	4	33.15509370	33.56000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	5	36.99407330	36.72000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	6	38.68989150	38.66000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	7	20.92449400	20.09000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	8	50.30884710	49.89000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	9	37.56869010	37.55000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	10	35.45490460	35.37000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	11	37.17891160	37.16000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	12	34.56763830	34.48000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	13	38.83286220	38.40000000
00002426	1kxs-2	14	GCAAGTCuAAAACG	NMR	14	0.00000000	0.00000000

Figure 3.2.1.4: This sample set of data illustrates the design of the TwiDDL database and the resulting flexible retrieval of data from both the Summary and Details tables via the TWID column from each. The TwID appears in the first column and is assigned a unique numerical identifier for each structure (00002425 and 00002426). The TwID is indicative of the order that the structure was entered in the database (00002425 for 1kxs-1 was entered before 00002426 for 1kxs-2 [20]). Grouping based on values in the TWID column allows a search of the database to return information about both the overall structure and each base-pair step.

Within Figures 3.1.1.2 and 3.1.1.3, there are descriptions of the various data types used to store these data in the database. After analyzing the data required for the database, I narrowed the possible data types to five categories; INT, DECIMAL, VARCHAR, TEXT, and ENUM. The INT data type is an integer that can be either positive or negative, and is utilized by TwiDDL for a few values, including the TwID. In the case of the TWID column, the INT data type is used in combination with the INT(8)

UNSIGNED ZEROFILL features, which force the TwID to be a positive integer with eight digits and with zeros filling the leading empty spaces, i.e., 00012345 instead of 12345. The DECIMAL data type handles floating point numbers where the total number of digits or precision, M, and the number of digits after the decimal, D, can be set by defining the type as DECIMAL(M,D). The DECIMAL data type is used in many of the columns in the Summary and Details tables. The VARCHAR(M) data type handles literal data, with M equal to the total number of characters in a variable length string of characters. This data type is useful for small strings of a known length, such as the dimer or teramer step contexts. Some items in the database require longer and less regular strings, which can be handled with the TEXT data type. The final data type, ENUM or enumeration, holds a string of characters and limits the values in the string to the values set with a syntax such as ENUM('value1','value2',...). In TwiDDL, the ENUM data type is only used for the ExpMethod entry, which limits the values to either XRayCrystal or NMR, the two methods that the database was designed to handle.

3.2.2 Performance Optimizations

The database design went through several iterations. Initially the tables were too simple and the requirements coming from the web interface demanded that more data be stored. The growth in the data stored at first did not seem to have an effect due to the small number of structures. As the number of structures began to grow significantly the

redesign of the database became necessary due to the impact on performance. The original design of the database had four tables, with the data divided on the basis of origin, either the out.out file or the twist of supercoiling twistwrithe.csv file, and the type of information, i.e., descriptors of the base-pair or the base-pair-step geometry. The Summary table was much like it is today, but problems arose in the design of the other tables when performing complex queries or calculating the average, minimum, and maximum values resulting from a search.

In the original design some queries and calculations would take on the order of minutes to complete. Although the time lag might be acceptable for the technical user who understands SQL to retrieve information, an average user like a biologist visiting a website expects a much quicker response on the order of single digit seconds or less. At first the cause for the delays was not clear. In order to debug the problem the delay needed to be quantified. Luckily, the MySQL database makes this easy by supplying the amount of time it takes to retrieve a result as part of the queries output. Section 3.4 will discuss the web interface more closely. The web interface, which first showed this performance issue, was analyzed and found to add no significant delay in data retrieval over running the queries natively within the MySQL command line database environment. This indicated that the database itself was the cause for the slow response.

After various experiments monitoring the time that queries took to complete the

database schema was changed, and a few optimizations were made to address the performance issues. First, the original database schema of three tables were combined into one single table, the Details table, which contained all of the data. Essentially making a new schema with just two tables; a Summary table, as before, and a new Details table with the three original tables in one. This was beneficial because the three tables had roughly 400,000 entries each, totaling roughly 1,200,000 rows of data to analyze for each TwID in a search. The act of analyzing and combining these tables took considerably longer than necessary. The combined Details table makes a significant improvement in the database design when it came to efficient access to the data during complex queries. A second major improvement in the performance was done in the queries themselves. It was discovered that the use of nested queries yielded another significant improvement in performance over the original queries. Nested queries use one query inside another query in order to generate results. In TwiDDL a nested query is used first to select the list of TwID's from the Summary table and then to gather data from the Details table based on the TwID's found in the first query. The original queries would simply list all of the desired TwID's and directly query the Details table. This was efficient with a small list of TwIDs, but as the number of TwID's in the database grew this type of query became very slow and the need for nested queries emerged. The third and final performance improvement was in an increased use of the KEYS feature in the MySQL database for columns that were often used for queries, beyond the PRIMARY

KEY and UNIQUE KEY feature already used on the TWID column in the Summary table. The columns that improved the queries when declared as a key included the PDBID and NDBID columns in the Summary table, and TWID, BasePair, and Step columns in the Details table. The improvements from these changes were significant. For example a query of average, minimum, and maximum step parameter values shortened from over 6 minutes to 2.13 seconds.

3.3 Data Handling

As discussed in Section 3.1.2, the data stored in TwiDDL came from a variety of sources. In this section we will focus on the handling of the raw data files stored in the directory of a given structure, which includes post processing or parsing of the PDB file, the out.out file from 3DNA, and the twistwrithe.csv file. This section does not address the data handling done by 3DNA and the twist of supercoiling applications or how these programs process the files used as input.

3.3.1 Parsing

The term parse as related to computing is defined as “to analyze (a string of characters) in order to associate groups of characters with the syntactic units of the underlying grammar”[22]. In this section we will discuss how the major sources of data in TwiDDL are obtained by parsing the text stored within the raw data files. In the

implementation of TwiDDL, the parsing is accomplished with a set of subroutines written in Perl. These subroutines are broken into three routines to match the three files that need to be processed. The subroutines consist of (i) the parse_pdb() routine, which processes the PDB files, (ii) the parse_outout() routine, which processes the 3DNA out.out file, and (iii) the parse_twistwrithe() routine, which processes the twistwrithe.csv file.

The first step in the parsing process determines which PDB file(s) to download. This is done dynamically through the administration interface of the web page or through a script designed to be run on a schedule to keep the database updated, as detailed in Section 3.5. The PDB file is then downloaded and stored in a directory of the same name. The parse_pdb() routine is provided two inputs, the PDB ID that needs to be parsed, and a reference to a Perl data structure that stores the information that will be inserted into the database. The referenced data structure can be thought of as a temporary database entry that all of the TwiDDL routines can commonly access and store information, which is eventually used to populate the database. The PDB ID is more straightforward because it locates the PDB file downloaded into the directory of the same name.

The execution of the parse_pdb() routine first creates a header file to store the common information contained in the PDB file that pertains to all models within the PDB file. This header file is only used for an NMR structure. The parse_pdb() routine

then opens the PDB file and processes the data line by line based on the standardized PDB file format [12]. This processing is the heart of the parsing needed to extract useful information from the PDB file. This is accomplished by using a combination of Perl regular expression syntax and built in Perl functions such as substr(), which extracts a subset of the string, or section of the current line of text, based on position and length. These two techniques very effectively allow the parse_pdb() routine to comply to the PDB file format standard, and to capture the specific information TwiDDL uses from the PDB.

The parse_pdb() routine stores the classification, title, space group, resolution, crystal title, and primary citation in the referenced data structure that was given as input when the routine was called. It also stores all the PDB file information in the header file until it hits the line that starts the MODEL section of the PDB file. Before it hits the MODEL section, a check of whether the EXPDTA value in the PDB file is defined as an NMR is made. If it is an NMR structure, then the header file is used to create a new directory and PDB file named PDB ID-#, like 4kbd-1, 4kbd-2, or 4kbd-3, as previously discussed. If it is not an NMR structure, then only the single PDB ID directory is required. The routine, once complete, returns an array or list of file names for all of the PDB files that pertain to the PDB ID input at the call of the parse_pdb() routine. This allows any calling code to work from the list of processed or parsed PDB file(s) to make

further use of them within TwiDDL. In particular, the individual PDB files listed in the array are then run one at a time through the 3DNA and the twist of supercoiling applications to generate the raw data files discussed in Section 3.1.2.

Once the raw data files have been generated, the next two routines parse these files. The parse_twistwrithe() routine processes the twistwrithe.csv file. Like the parse_pdb() routine, the parse_twistwrithe() routine is called with same two inputs, the PDB ID and the referenced data structure. The PDB ID is similarly used to locate the directory that contains the twistwrithe.csv file, but it is important to note that the PDB ID here can be the modified PDB ID-# for an individual NMR model. Once the directory and file are located, the routine runs through the text line by line processing the data contained within it. During the parsing of the twistwrithe.csv file, there is a significant amount of data captured as well as calculated (see Section 3.1.2). This parsing is accomplished with a series of flags which identify which section of the file is process. The different sections have differing syntax and thus require slightly different parsing codes. Each section has a key word to identify it, for setting the flag. The key words include TwistCAve, TwistOAve, Origins, Axis, and Normal, the corresponding respectively to twist values based on the structure with an added closing step, corresponding values obtained for the open structure, the coordinates of the origins of the base-pair slabs, the unit vectors along the X, Y, and Z axis of the slabs, and the normal

vectors of the slabs. Within the parse_twistwrithe() routine the primary built-in Perl function splits a line into separate entries based on a delimiter. Since the twistwrithe.csv file is a csv, or comma separated values based file, the use of a comma in the split routine was a natural fit in processing the data. This made processing the data simply a matter of knowing the key word that identifies the current section and the relevant data of that section on a given line. As each section was parsed, the relevant data were stored in the referenced data structure. In addition to collecting and storing the data and calculations based off of the twistwrithe.csv, the parse_twistwrithe() routine also extracts the NDB ID.

Next the parse_outout() routine is called, like the two prior parsing routines, with the PDB ID and the referenced data structure as inputs. The routine again uses the PDB ID to locate the out.out file, and to construct the PDB ID-# for the respective NMR models. Like the previous parsing routines, parse_outout() runs through the out.out file line by line and processes the contents for storage into the referenced data structure. The techniques used within this routine are a combination of those used in the previous routines, i.e., syntax specific to sections within the out.out file and to the delimited and positional storage information for the TwiDDL database. The sections within the out.out file similarly have key words – such as RMSD, step, or Local base-pair parameters – that could be used to indicate the section. Additional logic was applied to eliminate redundant

processing of the same information when some key words repeated. This was accomplished by keeping a counter that tracked the repeat of a particular section and using the counter to determine when to skip a repeated section.

Processing the individual sections within the out.out file had some added complexity beyond the variation of how the data were stored. The desire to infer information based on the contents of the out.out file added to the challenges. There were also some data points stored in the out.out file that did not have a simple built-in Perl function to solve. Those data, which required a more customized parsing code included (i) the sequence, (ii) the dimeric and tetrameric step contexts, (iii) the chemical identities and residue numbers (Base ID I, Base ID II) of the paired bases in the selected structure, (iv) the identities of the complementary strands (Chain ID I, Chain ID II), (v) the chemical identities of the paired bases (Residue ID I, Residue ID II), (vi) the numbers of the two bases (Residue Num I, Residue Num II). Once this information is successfully parsed from the out.out file, it is then analyzed for anomalies as will be discussed in Section 3.3.2.

The data from the three parse routines are inserted into the database using the load_db() routine. This routine takes the referenced data structure for the database values, a list of PDB ID's and a flag to indicate whether to overwrite any existing entries for the PDB ID. In summary, the load_db() routine runs all of the subroutines required to

input a new structure into the database by downloading the PDB file, generating the out.out and twistwrithe.csv files, parsing the files through the three parse routines, and then inserting the data into the database. This load_db() function is the primary routine required to enter a structure into the database by simply supplying the PDB ID of the selected structure. Once this routine has been run, the data are ready for use through the web interface unless the structure is flagged as not being in compliance with certain expectations or requirements of TwiDDL.

3.3.2 Checks for anomalies

At the tail end of the parse_outout() routine there is code specifically designed to check for anomalies and conditions in the data that are not handled by TwiDDL. These new entries are flagged in the Comments column of the Summary table in order to prevent them from being displayed to users without a manual review of the structure. The first check analyzes the identities of the complementary strands (Chain ID I, Chain ID II) to ensure they have only a single one-letter designator, since that is all TwiDDL can handle. The second check is the numbering of the paired bases (Residue Num I, Residue Num II) to ensure that the bases are sequential, i.e., the residue number must vary at increments of +/–1 along complimentary strands. The third and final check ensures that the structure consists of more than a single base pair. The text stored in the Comments column reflects how a flagged structure deviates from these expectations. An entry in the

Comments column can consist of any of the three following phrases or any combination of the three: (i) “ChainID I - contains multiple chains => X”, where X provides more details about the chains in the structure; (ii) “Residue Number I Deviates”; and (iii) "Only a single base pair”. The comments are also provided for ChainID II and Residue Number II. This information can then be reviewed and confirmed through examination of the respective entries in the database. The structure will be shown through the TwiDDL web interface once the comments have been manually removed.

3.4 Web Interface

There has been significant investment in the design of the web interface to TwiDDL in order to streamline the usability and features of the software for the biologists and other researchers that we intend to use it. A sound web site design can go a long way to improving acceptance and interest in the twist of supercoiling data stored in TwiDDL. As such I spent considerable time drafting ideas for what the key components of this site should be, and learning what new web standards existed that could enable them.

First, to address the design aspect of the improved website I identified four major subsections in order to categorize the information. This design is captured in the menu for the web interface which can be seen in Figure 3.4.1. The first section is the Home menu option, which contains a brief Introduction with information that a casual reader can use to get started. The Introduction page is also the default page that a user sees

when connecting to the website at http://twiddl.rutgers.edu.

The second section on the website is the Background menu option, which links to more in-depth documentation and other related materials. The link to the Information page presents the theory behind the method used for calculating the twist of supercoiling. This link to the Related Articles page provides a list of references relevant to the TwiDDL web site. The link to the Olson Group offers more general information for nucleic acid structures. The link to the Contact page contains information about myself.

The third section on the site, shown as the Software menu item, allows the user to download the 3DNAdesigner software package via the 3DNAdesigner Download page. If the request to download the software comes from a computer that is not an .edu or an .gov domain, the 3DNAdesigner Download page presents the end user with a short form to request access to the software. The form collects the user's name, email address, and organization. This information is then sent in email to me, so that I can respond to the request. If the request comes from an .edu or an .gov, then the user is brought to a download page, which allows quick and easy access to the 3DNAdesigner software packages created for both Linux and Windows Operating Systems.

The fourth section on the website is the TwiDDL menu option, which provides background information on and access to TwiDDL. The link to the About page presents the purpose and background behind the data stored in TwiDDL. The link to the Search

page is the primary web interface access to the database. The link to the Definitions page is used for two purposes. The page contains a list of defined terms, and is the source for the content of the TwiDDL context help that pops up at the bottom of the screen when a user places a mouse over a TwiDDL term.

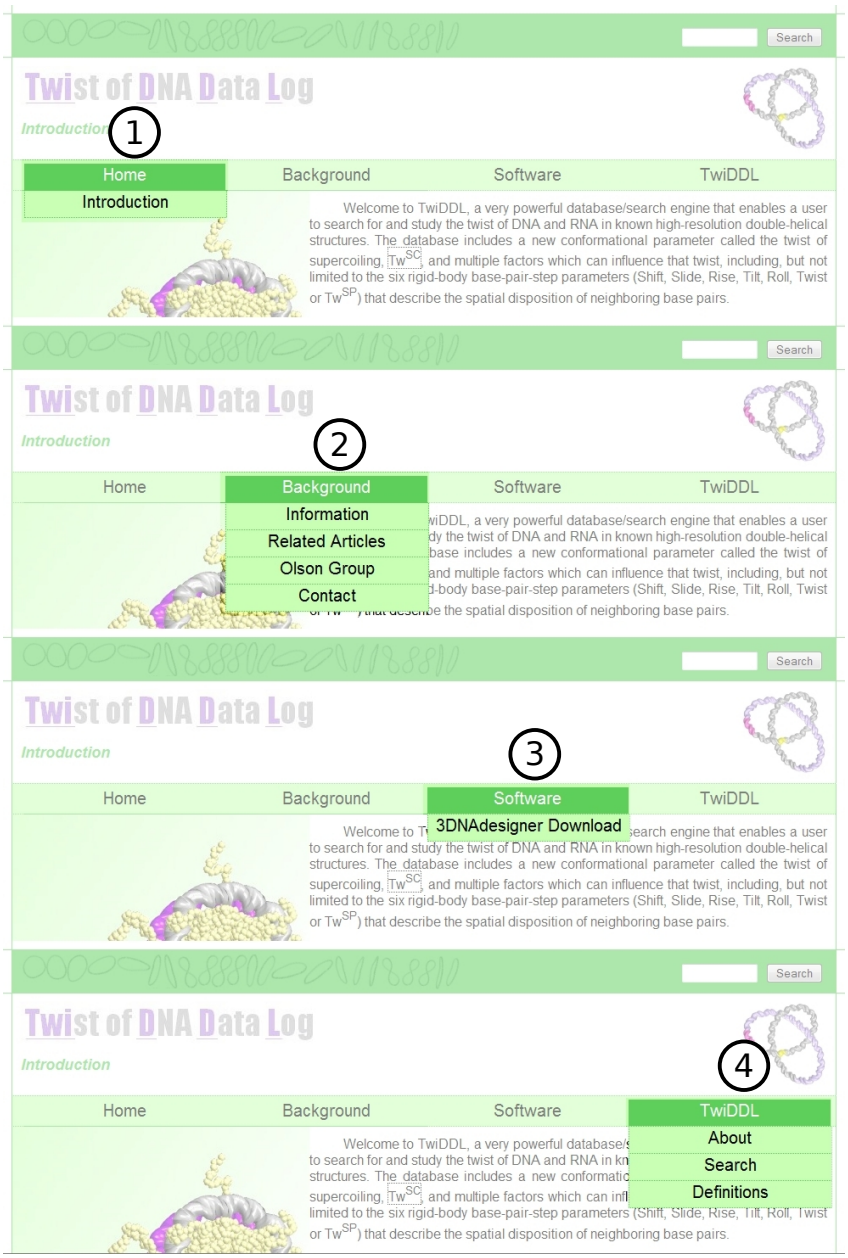


Figure 3.4.1: The TwiDDL web interface is broken up into four major sections. (1) The Home menu contains the introduction to the site. (2) The Background menu provides deeper information on the goals and methodology in TwiDDL. (3) The Software menu provides access to stand-alone programs. (4) The TwiDDL menu provides access to the database.

In addition to considering how the overall site layout would come together, I invested time into considering the critical components that all pages would need to have.

I then devised a design that I think best addresses the overlapping needs of each section of the site, to provide a common and easy to use page design for every section. Pulling all of this together was accomplished with the use of flow charts to help minimize the potential for creating a site too deep or complex for an end user to navigate.

The current design of each page consists of a single column layout that is broken into five horizontal sections. Starting from the top down, I will detail the intention for each of the sections as follows. First is the quick search bar, which enables the user to search the database quickly using the TwID, PDB ID, or NDB ID of a structure. Second is the title bar, which displays the title for the web site, as well as the title of the page that the user is currently accessing. The third section is the menu bar. The menu bar provides navigation to the four major topical subsections documented above: Home, Background, Software, and TwiDDL. The fourth section is where the bulk of the information is displayed, and could be considered the body of the web pages. The second and fourth sections change for each page to provide the user proper contextual information for the selected link. The fifth section is simply a footer for any disclaimers, copyright information, simple links, or other related information yet to be placed. This section is intended to be kept consistent across all pages, unless there are specific copyright differences or disclaimers that need to be made for particular materials on the site. All of these details can be seen in Figure 3.4.2, which is a screen capture of the current design

and highlights the five major sections described here as well as the four menu subsections discussed above.

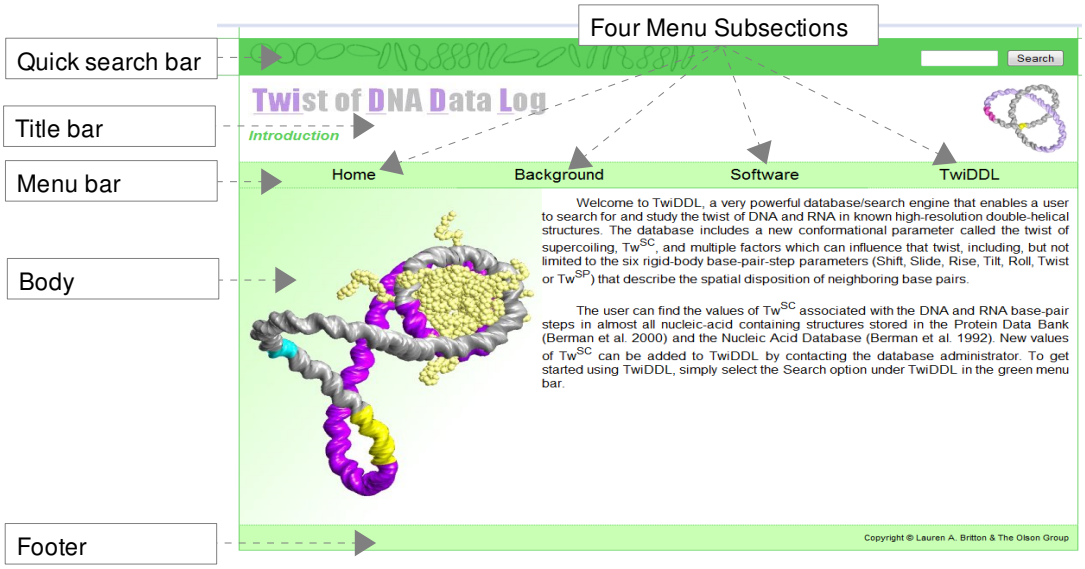


Figure 3.4.2: Diagram of the web page layout showing the four menu subsections, as well as the five default regions used in the page layout throughout the site.

The design of the website was accomplished through the combination of several web technologies. Throughout the course of my thesis I spent considerable time developing and learning these new technologies in order to produce the current website. The site is currently a combination of HTML (Hyper Text Markup Language), CSS (Cascading Style Sheets), PHP (Php: Hypertext Preprocessor), and JavaScript. Each of these technologies allowed for flexibility that traditional static HTML-only based pages could not allow.

For example, the menu bar uses a combination of HTML, CSS and PHP to make it easy to use and easy to maintain. The text of the menu is stored in single traditional HTML file called menu.html, which uses very basic tags that normally generate a hierarchical list of items. I use CSS to format those tags so that they operate like drop-

down lists when the mouse moves over them. This how the page would work if the menu.html were used on its own. In order to simplify things, all of the other web pages in the TwiDDL web site use PHP to embed the single HTML file for the menu dynamically rather than have the exact same HTML for the list of menu items stored separately in each page. The benefit is that I only have to maintain one menu.html file and all of my pages will update automatically based on that one HTML file to have a consistent and identical menu system.

Additionally, CSS has enabled me to simplify the HTML for each of the pages greatly, so that I can create templates in HTML and use CSS to format them all in a consistent manner. The web site design extends the CSS functionality by using PHP to generate the CSS file dynamically via the labcss.php file, which enables even more flexibility than CSS alone because one can use PHP to generate the CSS text dynamically. Using PHP like this enables changes to a variable or set of variables at the beginning of the labcss.php file to generate the CSS dynamically in all places the variables are used. This is helpful in changing things like the border colors of all tables on every page throughout the entire site with one quick edit. What all of this does is allow the greatest flexibility to change the site quickly, without a lot of work, as well as provide a very professional look and feel that is easily customizable.

3.4.1 Relationships with the Database

One of the primary functions of the website was to provide a user interface to access the information contained within the TwiDDL database. In addition to providing access to users who would like to view the data, the website was also designed to provide a user interface for the insertion, removal, and maintenance of entries in the database. In order for the website to accomplish these goals several different layers within the website and the database were designed to interact with each other. These layers are primarily based on a set of technologies referred to as LAMP, where LAMP typically stands for the combination of the Linux operating system, the Apache HTTP server, the MySQL database, and either Perl or PHP. TwiDDL uses LAMP along with a few other software packages such as 3DNA, Matlab, and Apache Tomcat to provide a feature-rich experience when interacting with the database.

The user interface of the site is primarily built up of either web pages created in PHP or dynamically generated by a Perl script. For example, when the user goes to the TwiDDL menu option, and selects the Search item, he or she is brought to a landing page that presents links for both a simple and an advanced search as described in Section 3.4.4. All of theses steps are driven by PHP to generate the HTML that make up the menu options and corresponding pages to which it links. However, once the user clicks either the simple or advanced search link, the content is then driven by a Perl script, which is

the primary interface into the MySQL database and which dynamically generates the web interface based on the user input. In this section we will look closely at how the Perl script interacts with the database and the rest of TwiDDL.

The Perl code that makes up the scripts driving the user interaction with the database and user experience with the website are broken into two major pieces. The first is twdl.cgi, which is the primary code used for driving the use of the various functions of the web interface. Primarily twdl.cgi is used to set up the variable information input through the web-based interface in a way for it to be utilized effectively by the various functions written for TwiDDL in Perl. The twdl.cgi code next completes some checks and may correct some of the inputs provided by users. For example, when inputting ranges to search the ΔTw, Tw^{SC}, Tw^{SP}, buckle, kinking, or shearing data, twdl.cgi ensures the low value is actually lower then the high value input. Otherwise it swaps them assuming that is what the user really meant to input. Similarly, it ensures that searches using a magnitude are input using a positive number by providing the appropriate error for the user to correct the input. Once this is set up, twdl.cgi evaluates which function it should perform based on the information input from the web interface. The default is for twdl.cgi to display the simple search page. There are several other functions that twdl.cgi supports in order to provide a robust interface to the TwiDDL database, and those functions are detailed in Table 3.4.1.1.

Function	Description	Subroutines
admin (hidden)	Handles administration inputs for TwiDDL. Hidden from the web interface.	admin
graph	Displays and generates 2D graphs of a structure.	graph
html_search (default)	Displays the Simple or Advanced Search Page based on user selection. Used by default if no function is specified to twdl.cgi.	html_search
html_search_results	Displays search results based on user inputs.	html_seach_results
html_twid_details	Shows the TwID details page containing all of the data for that structure	html_twid_details
quicksearch	Processes user input to the quick search bar.	html_seach_results
reload (hidden)	Causes all existing database entries to be rerun and processed into the database. Hidden from the web interface.	reload
run_makedbtwist (hidden)	Runs the basic twist of supercoiling calculations only. Hidden from the web interface.	run_makedbtwist
show_details (hidden)	Shows an HTML version of the raw data in the MySQL tables. Hidden from the web interface.	show_details
step_stats	Displays statistics about structures' base-pair steps.	html_step_stats
toolbar	Processes user input to the toolbar on the results page.	html_step_stats, html_seach_results, html_error

Table 3.4.1.1: This table contains a summary of the various twdl.cgi functions, their respective descriptions, and the subroutines they call from twdl_funcs.pl.

In order to keep the code for twdl.cgi simple and straightforward there is a second Perl file called twdl_funcs.pl that contains all of the subroutines used by twdl.cgi to accomplish each of the functions described in Table 3.4.1.1. The subroutines in twdl_funcs.pl can be broken up into several major functional categories, as described in Tables 3.4.1.2-4. There are the HTML-focused subroutines (Table 3.4.1.3), which generate the HTML used to display the various sections of the website. There are the SQL-focused subroutines (Table 3.4.1.2), which are used to interact with MySQL to access the TwiDDL database. There are subroutines (Table 3.4.1.4) focused on running the twist of supercoiling calculations, and other supporting subroutines that run 3DNA and download PDB files. There are data-focused subroutines (Table 3.4.1.4) used for parsing the data as discussed in Section 3.3.1, as well as routines for interaction between the raw data and the database, such as insertion and removal from the database. There are also subroutines (Table 3.4.1.2) focused on dynamically graphing the data from the database, or even administrating the database. Many of these routines call each other, and benefit from the modular design that their dependence on each other provides. This modular design allows changes that improve one function to benefit all of the functions that rely on it, and keeps a common interface for flexible and easy access to the data stored within TwiDDL.

Subroutine	Description	Category
reload	Reloads the database with existing entries. Useful for various incarnations of the twist of supercoiling calculations	Administration
areyousure_db	Confirms deletion of database entries.	Administration
admin	Handles administration inputs for TwiDDL. Hidden from the web interface.	Administration
show_details	Displays the raw database table in HTML.	Administration
remove	Removes specified structures from TwiDDL.	Administration
graph	Displays and generates 2D graphs of a structure.	Graphing
makegraph	Creates the 2D graph images.	Graphing
sql_textsearch	Generates the SQL for user inputs to text search fields in the advanced search page, such as the Structure Primary Citation(s) and the Structure Title(s).	SQL
sql_summary	Generates the SQL for creating the search results page, as well as statements to get the list of either TWIDs, PDBIDs, or NDBIDs returned by the search.	SQL
sql_subtables	Generates the SQL for creating the various tables displayed in the search results and TwID details pages.	SQL
sql_step_stats	Generates the SQL for creating the step statistics tables.	SQL
sql_add_andor	Appends logical AND or OR into SQL statements dynamically.	SQL

Table 3.4.1.2: Subroutines related to the Administration, Graphing, and SQL functions in the twdl_funcs.pl script, along with their descriptions and category.

Subroutine	Description	Category
html_context_highlight	Displays a highlight within the sequence based on the step or partial sequence searched for.	HTML
html_contexthelp	Displays the HTML for the pop up context help.	HTML
html_error	Displays a simple error page.	HTML
html_footer	Displays the standard footer section.	HTML
html_header	Displays the standard header section	HTML
html_search	Displays either the Simple or Advanced Search Page depending on user inputs.	HTML
html_search_results	Displays the results from a search.	HTML
html_searchcriteria	Displays the search criteria section.	HTML
html_step_stats	Displays the step statistics based on user inputs.	HTML
html_subtable	Displays the show/hide tables within search results, and the TwID details pages. Such as the buckle, kinking, shearing, step details, twist comparison, step parameter, and base-pair details tables.	HTML
html_toolbar	Displays the toolbar at the top of the search results table.	HTML
html_twid_details	Displays the TwID details page.	HTML
string_to_width	Converts a long single line, into multiple lines of a specified character width. For example, displaying long sequences as 25 characters per line.	HTML

Table 3.4.1.3: Subroutines related to the HTML generating functions in the twdl_funcs.pl script, along with their descriptions and category.

Subroutine	Description	Category
parse_twistwrithe	Parses the twistwrithe.csv file, from the twist of supercoiling calculations, for input into TwiDDL.	Parsing
parse_pdb	Parses the PDB file for input into TwiDDL.	Parsing
parse_outout	Parses the 3DNA out.out file for input into TwiDDL	Parsing
make_newtwistwrith	Parses and performs modifications to the output from the twist of supercoiling calculations.	Parsing
load_db	Runs all of the subroutines in the Categories running and Parsing in order to load the TwiDDL database with new structures.	Running & Parsing
run_twistwrithe	Runs the twist of supercoiling calculations for the specified structure.	Running
run_makedbtwist	Performs the automated steps to complete the twist of supercoiling calculations based on a new PDBID.	Running
ndb2pdb	Converts input NDBIDs to PDBIDs, and input PDBIDs to NDBIDs.	Running
get_pdb	Downloads the original PDB file and stores in the TwiDDL directory structure.	Running
get_nbp	Returns the number of base pairs.	Running
check_status	Converts subroutine return values to meaningful phrases	Running
run_x3dna	Runs 3DNA against the PDB file input.	Running

Table 3.4.1.4: Subroutines related to the Running and Parsing functions in the twdl_funcs.pl script, along with their descriptions and category.

Besides the Perl subroutines, that generate most of the web interface and allows access to the database, there are several other layers that make up TwiDDL. Many of them have been discussed above, but here they will be discussed as a whole in order to understand how TwiDDL as an application operates among the various layers. These layers are made up of several parts as depicted in Figure 3.4.1.1. The first layer is made up of the Linux operating system that provides the fundamental environment that everything runs on top of, and supplies standard interfaces for executing code and file storage or manipulation. The second layer is the underlying applications, such as 3DNA and the twist of supercoiling calculations run through Matlab, that provide the raw data. The third layer contains both the raw data stored in directories for the respective files of each structure, as well as the MySQL database that stores selected information from those raw data files. The fourth layer consists of a combination of PHP and CSS that provides the basic web interface for the more static components of TwiDDL. The fourth layer also contains the Perl code that drives many aspects of the web interface including parsing of the raw data for insertion into the database, as well as all aspects of the user's interaction with the database. The fifth layer is the web application layer made up of the web server and Java servlet applications which provide the services required to make TwiDDL widely available to our target end users, such as biologists and other researchers.

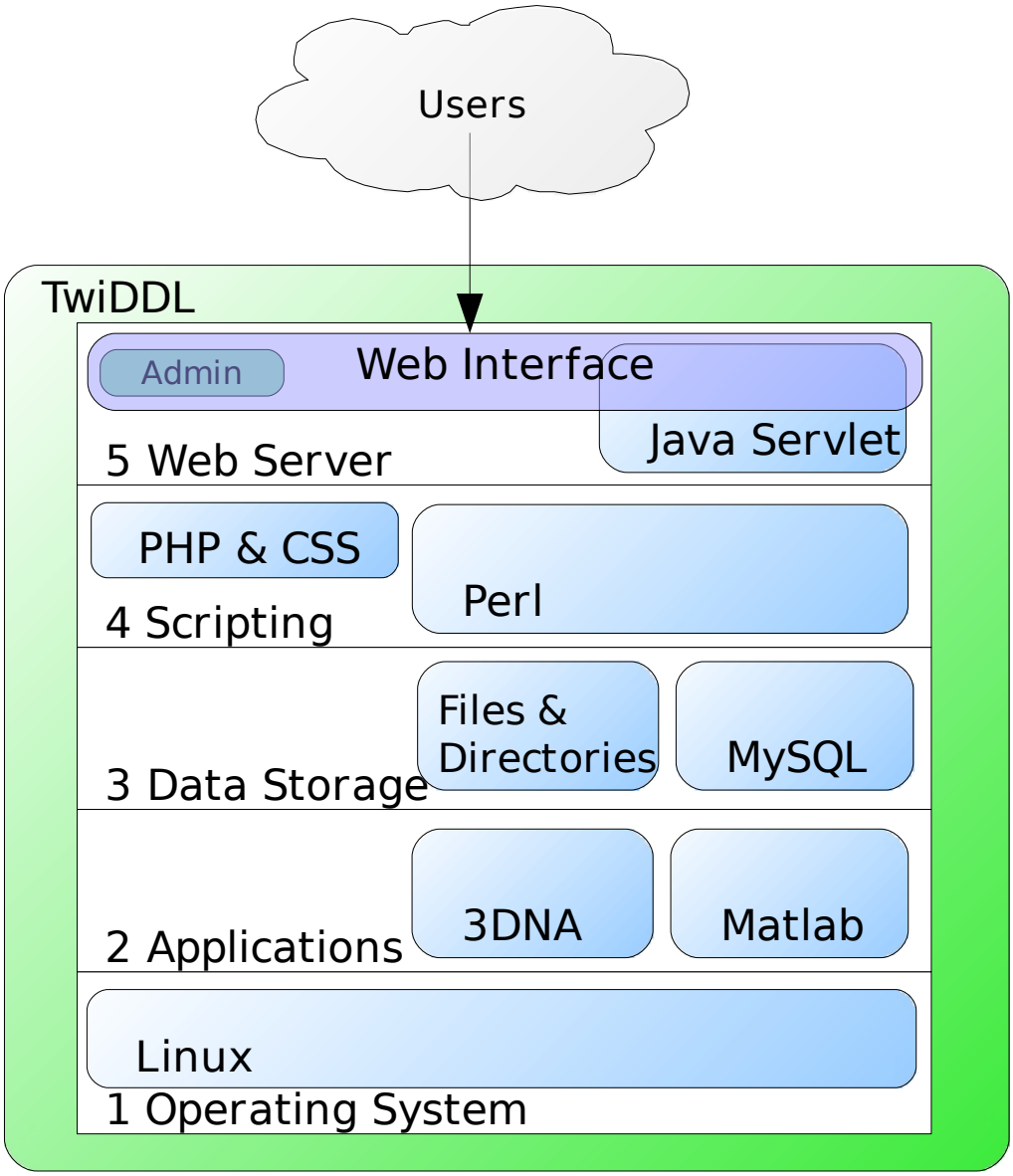


Figure 3.4.1.1: The five layers of software used by TwiDDL to drive the user interaction over the web interface, including the underlying database, raw files, and applications for generating the twist of supercoiling data.

The main code for TwiDDL was written in Perl, which allows for easily extending use of these functions beyond the web interface. In Figure 3.4.1.2 one can see that many of the capabilities available through the web interface can concomitantly be achieved through command line scripts written to take advantage of the features that exist in the fourth layer. This is a very powerful tool, since this allows for features that otherwise are not traditionally able to be accomplished in a web environment. Comparison of Figures 3.4.1.1 and 3.4.1.2 shows that the PHP block is missing in the fourth layer of the latter figure. PHP is not utilized in the command line environment because the intent of PHP is to provide web-focused interfaces. Examination of Figure 3.4.1.2 also shows a command-line interface in the fifth layer that replaces the web server and Java servlet-based web interface in Figure 3.4.1.1. This command-line interface is created with command-line Perl scripts that make use of the subroutines discussed in Tables 3.4.1.2-4.

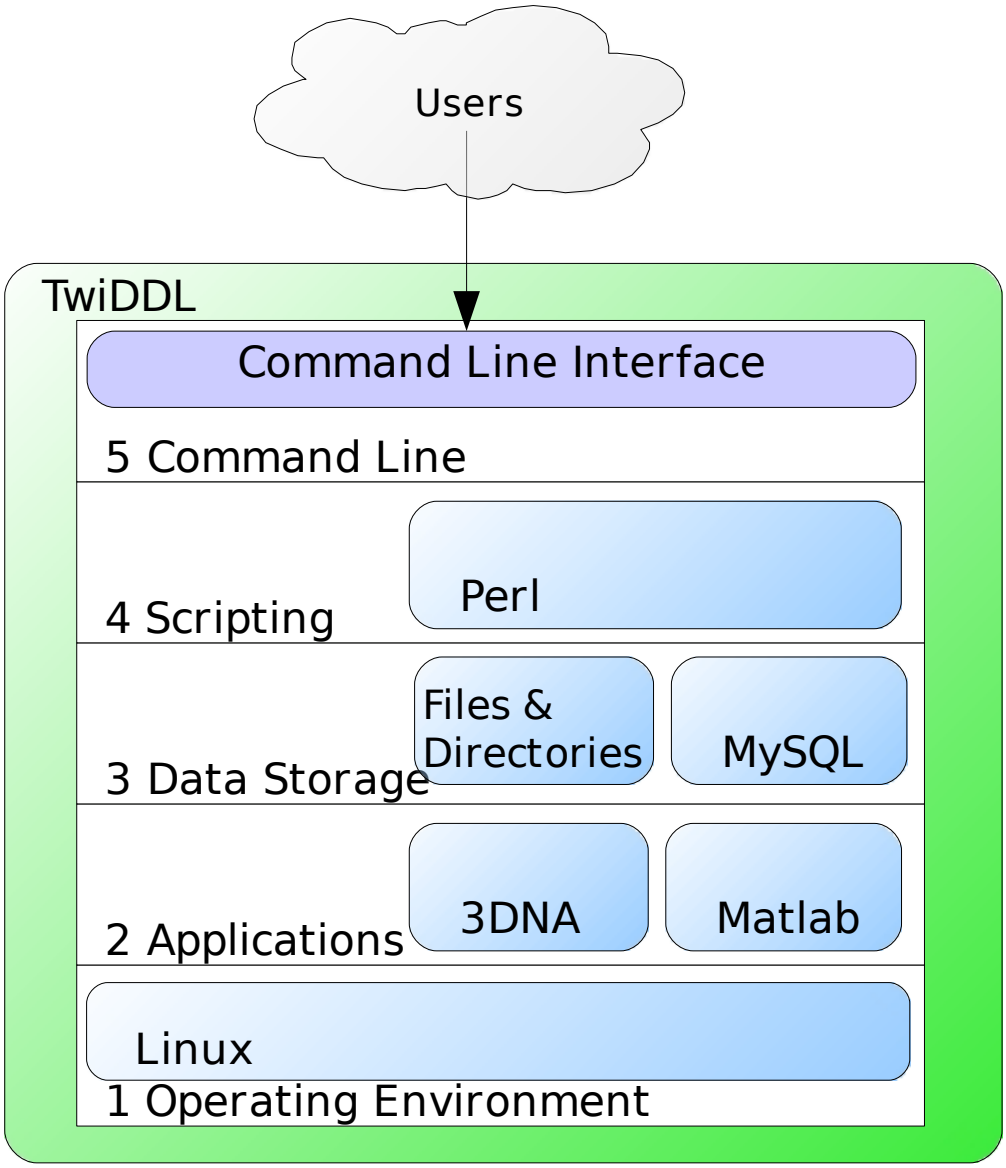


Figure 3.4.1.2: The five layers of software used by TwiDDL to drive the user interaction through a command line interface.

3.4.2 Use of CGI

As discussed so far, the web interface for TwiDDL uses HTML generated either by PHP or Perl to display the majority of the contents. In HTML a user may enter information for interacting with web applications through forms. This is accomplished in HTML by using the <FORM> tag, which most commonly takes the action and method attributes to determine how the user's information is processed. Within TwiDDL, the forms all use /cgi-bin/twdl.cgi as the value for the action attribute, and GET for the method attribute, with the exception of the download section which uses the POST method for reasons explained later. What this means is that, by default, a form filled out on TwiDDL will always use the twdl.cgi file described in Section 3.4.1, and almost always will display the list of inputs provided to twdl.cgi in the location field of most web browsers. The CGI interface to TwiDDL in this case is the twdl.cgi Perl script, which is easily utilized by web browsers to access features and functions in the database.

To understand the use of CGI in TwiDDL better, the example of the simple search page can show how twdl.cgi easily processes user inputs and provides dynamic content through the web interface. Figure 3.4.2.1(a) shows the URL used to access the simple search page for TwiDDL. Figure 3.4.2.1(b) dissects the same URL into its major components: the server URL; the script path as specified by the action attribute of the form tag; and the query string input by the user. In this case, the user is simply following

the link to the simple search page (1a). For TwiDDL, the server URL and script path will always be constant. The query string however will change with the user input and twdl.cgi responses based on that input. Figure 3.4.2.1(c) breaks down the query string into its component search words. Each of these search words consist of a variable and a value assigned to that variable as shown in Figure 3.4.2.1(d). The variables and their values are critical in driving the functionality required from TwiDDL. For twdl.cgi, the searchmode variable will handle two values, either simple or advanced. In this example the searchmode variable has been set to simple, but if the user were to click the advanced search page link then the searchmode variable is set to advanced as shown in Figure 3.4.2.1(e) & (f).

Figure 3.4.2.1: Presents the decomposition of the URL for the Simple Search Page, and shows how to turn it into the URL for the Advanced Search Page through CGI inputs. (a) Shows the URL used to access the simple search page for TwiDDL. (b) Dissects the same URL into its major components: the server URL; the script path; and the query string. (c) Breaks down the query string into its component search words. (d) Shows the respective variable and value from each search word in (c). (e) Setting the searchmode variable to advanced, would create the advanced search page URL in (f).

In Figure 3.4.2.1 the func variable is synonymous with the functions that were described in Table 3.4.1.1. In other words, to run a different function from the twdl.cgi Perl code, one simply specifies func=<function> to run it. For example, in Figure 3.4.2.2(a) the database could be quickly searched for the value of 1kx in any of the PDB, NDB, or TwiDDL identifiers based on two variables being set. The instruction func=quicksearch, with the func set to quicksearch tells twdl.cgi to run the quicksearch function. The value of the supplied quicksearch variable in Figure 3.4.2.2(b), 1kx, calls the html_search_results function from twdl_funcs.pl, and returns a web page of structures pertaining to 1kx in any of the PDB, NDB, or TwiDDL identifiers. This is exactly how the quick search box found at the top of every TwiDDL page functions.

```
(a)         func   =   quicksearch
     quicksearch   =   1kx

(b)   http://twiddl.rutgers.edu/cgi-bin/twdl.cgi?quicksearch=1kx&func=quicksearch
```

Figure 3.4.2.2: Presents the use of CGI to call twdl.cgi functions for performing a quick search in TwiDDL to access data about the structures containing the 1kx in their PDB, NDB, or TwiDDL identifiers. (a) The variable and value combinations required for the CGI. (b) The URL for the quick search based on the combinations in (a).

The two examples presented here show how TwiDDL uses CGI to accomplish various goals through a single script. In the first example we see how changing the input value can change the behavior of the function called by twdl.cgi. Figure 3.4.2.1(a) shows

how the TwiDDL CGI displays the simple search page, and by simply changing the input values, displays the advanced search page. Figure 3.4.2.2 shows how one can change the function to produce an entirely different set of functionality. Rather than displaying one of the search pages, we change the func variable from html_search to quicksearch resulting in the database being queried for the value input by the quicksearch variable and the search results page being displayed. These are both very simple examples of how TwiDDL utilizes CGI to respond dynamically to user inputs and requests.

In Figure 3.4.2.3 the example is a bit more complicated. This example will actually yield the same result as the quicksearch instruction in Figure 3.4.2.2, but instead this is done using the advanced search page. The quicksearch only shows a single variable because the twdl.cgi script hides the fact that three variables are used to check the PDB, NDB, and TwiDDL identifiers for the selected value. In the advanced search page, all of the possible variables are displayed as part of the query string, as shown in Figure 3.4.2.3, because the user could provide input for any of them. In our example we simplify the search results by only using the “Enter PDBID(s)” portion of the advanced search page, which also makes for a simple query. The various search capabilities will be discussed in Section 3.4.4, but here we have presented the fundamental underlying CGI that drives it.

```
   (a)   http://twiddl.rutgers.edu/cgi-bin/twdl.cgi?
dtwrm=range&lowdtw=&hidtw=&twscrm=range&lowtwsc=&hitwsc=&sptwistrm=range&lowsptwist=&
hisptwist=&bucklerm=range&lowbuckle=&hibuckle=&lowkinking=&hikinking=&lowshearing=&hi
shearing=&pdbids=1kx&ndbids=1kx&twids=1kx&notpdbids=&notndbids=&seq=&resolution=&titl
e=&citation=&limit=25&offset=0&func=html_search_results


   (b)    dtwrm   =   range         (A default value that is ignored in this search)
         lowdtw   =
          hidtw   =
         twscrm   =   range         (A default value that is ignored in this search)
        lowtwsc   =
         hitwsc   =
       sptwistrm   =   range         (A default value that is ignored in this search)
      lowsptwist   =
       hisptwist   =
        bucklerm   =   range         (A default value that is ignored in this search)
       lowbuckle   =
        hibuckle   =
      lowkinking   =
       hikinking   =
     lowshearing   =
      hishearing   =
          pdbids   =   1kx
          ndbids   =   1kx
           twids   =   1kx
       notpdbids   =
       notndbids   =
             seq   =
      resolution   =
           title   =
        citation   =
           limit   =   25              (Sets the number of results per page)
          offset   =   0               (Sets the page offset from page 1 to display)
            func   =   html_search_results (Function used for the search)
```

Figure 3.4.2.3: Displays the URL and CGI variables entered by an Advanced Search Page search that reproduces the same resulting data as the quick search in Figure 3.4.2.2. (a) The URL to the advanced search performed. (b) The list of variables and values passed to the CGI for completing the advanced search in (a).

3.4.3 Search features

The areas of the web interface and database that drove the largest number of changes within the design were those with a direct impact on how the user would access the data. The majority of these changes are handled through the features implemented as a search and display of result based on the criteria that a user enters. The search went through several changes and a few major additions to give it the flexibility and accuracy it needed. The complexity of these changes frequently required a complete redesign of the search web page, as well as the code used to search the database in order to establish the design as it is today. In this section we will focus on the design and features of the search portion of the web interface. In the following section, we will focus on how the data retrieved by the search are displayed.

The search interface for TwiDDL was broken into two major features because not all of the users would require the use of all of the more complex search options. To make this distinction clear to the users the two interfaces to the search functions were named the Simple Search Page and the Advanced Search Page. The basic search features that we wanted most users to have access to were kept in the Simple Search Page, whereas the broader set of search features were put in the Advanced Search Page.

The are several primary fields that can be searched within the database using the Simple Search Page as seen in Figure 3.4.3.1, and all structures within the database can

be retrieved using these fields. There are a few basic fields directly related to the twist of a structure that can be searched. These are primarily focused on the values of the difference in total twist ΔTw, where ΔTw can be defined as either $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$ or $\Delta Tw^{SP} = Tw^{SC} - Tw^{SP}$. Depending on the user input, all structures within the database will be searched for the respective range or magnitude of either $\Delta Tw^{B\text{-}DNA}$ or ΔTw^{SP} and the structures containing a ΔTw within the specified values will be returned in the results page. In addition to the ΔTw, the user can use the Simple Search Page to query four other types of information within the database: the PDB ID; the NDB ID; the Sequence; and the Step. Once the user has entered the search criteria, he or she can additionally choose the number of results to be displayed on each page , consisting of options for 25, 50, or 100 resulting structures.

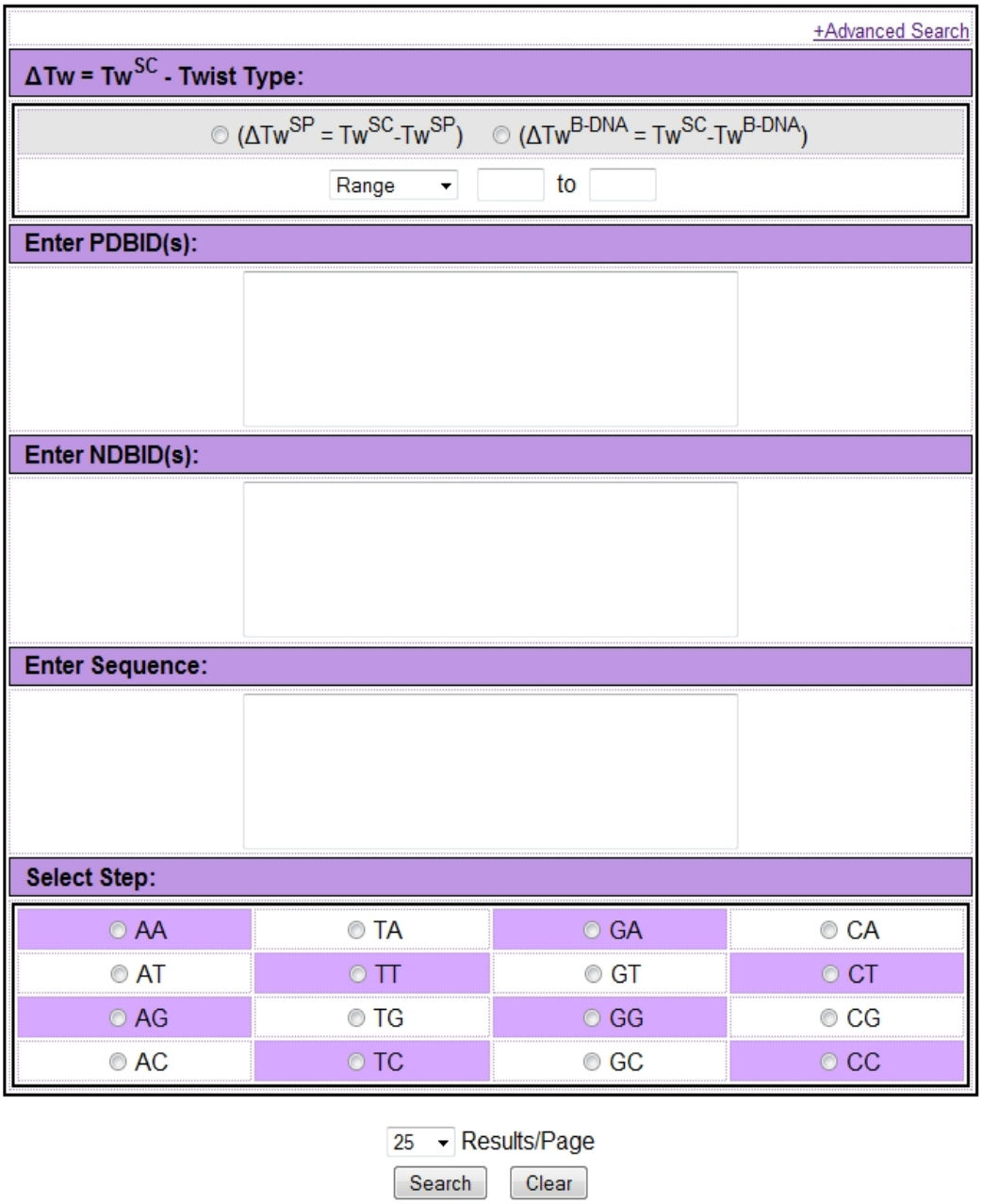


Figure 3.4.3.1: A screen capture of the Simple Search Page without user input.

When searching for the PDB ID or NDB ID of structures within the database the search may contain a single item or list of PDB IDs or NDB IDs. The list of PDB IDs or NDB IDs can be separated by any combination of spaces, or commas between entries in the list. Additionally, the user may search for a whole or partial PDB or NDB identifiers. The search will return all structures that match the criteria entered. For example, a search for 1kx will return a results page containing 1kx3, 1kx4, 1kx5, 1kxs-1, and 1kxs-2 [20, 21]. Similarly, a search using multiple PDBIDs or NDBIDs will return all of the structures that match all of the respective PDBIDs or NDBIDs, which can be helpful in getting a very specific list of structures. For example, a search could be made for the PDBIDs 1kx3, 1kx4, and 1kx5 to retrieve these exact PDBIDs only, and, in turn, fewer items than the broader 1kx search had retrieved in the previous example.

The user can additionally search for structures that contain all or part of a given base sequence. This option will return any structure with the specified entry and highlight the sequence of interest in green on the results page, making it clear where the sequence appears in the selected structures. The search is limited to the five common bases (A, C, G, T, U) and must only be a sequence containing those letters with no spaces. In addition to searching for long sequences, the user can search for a particular base-pair step. This step will be highlighted in purple in the results page. The search is made by clicking on one of the 16 common DNA and RNA base-pair steps, i.e., all

sequential combinations of successive A·T (or A·U) and G·C base pairs, as shown in the "Select Step" section in Figure 3.4.3.1.

An important and powerful feature of a TwiDDL search is the ability to combine the search fields in order to limit the results returned to the appropriate subset of structures that conform to all of the criteria. This feature exists for both the Simple Search Page and the Advanced Search Page, so the example here will focus on a very simple example that could be done using either search page. In the examples above it was shown how a partial PDB ID like 1kx would return a results page including the structures 1kx3, 1kx4, 1kx5, 1kxs-1, and 1kxs-2. It was also shown how the list of PDBIDs could be used to limit the structures returned by the search, but that is only useful with previous knowledge of the structures with given PDBIDs. The combining of search criteria handles the cases where the list of structures in the results page, often much longer than this example, need to be refined by something known about the structure, like the sequence. The combined search allows one to enter both the partial PDB ID of 1kx as well as the sequence of ATCTC. What this will achieve is identities of all structures that contain both 1kx in the PDB ID and ATCTC within the sequence. In this example those criteria return 1kx4 because it is the only structure from the original list that meets both criteria. This concept may be used to combine all of the fields offered for search through both the Simple and Advanced Search Pages of TwiDDL.

As can be seen from Figure 3.4.3.2 and Figure 3.4.3.3 the Advanced Search Page provides the capabilities required for greater refinement of the search compared to that available through the Simple Search Page. The search capabilities of the Simple Search Page are a subset of those available through the Advanced Search Page, so the Advanced Search Page is able to search for everything the Simple Search Page can and more. In addition to the search fields found in the Simple Search Page, the Advanced Search Page can also search the range and magnitudes of the total twist of supercoiling Tw^{SC}, the sum of the step parameter twist Tw^{SP}, and the buckle of paired bases in a structure. It also contains the ability to search a range of values for the kinking and the shearing of individual base-pair steps, to specify the experimental method used to characterize the atomic structure, to select the X-ray crystal resolution, and, for crystal structures, to pick the space group from a table. The search also allows the user to specify the structure classification from a list based on groupings supplied in the PDB file. In addition to searching for PDBIDs and NDBIDs, the Advanced Search Page allows for the exclusion of a single item or list of PDBIDs or NDBIDs so that certain structures are not included in the results page. The structure title and structure primary citation search options allow the user to search for any word included in title and literature citation associated of a Protein Data Bank file. The syntax used for these search boxes is much like others that are widely available. For example, desired phrases should be placed between quotation marks, a minus sign will exclude a word/phrase, and different search items need to be

separated by spaces.

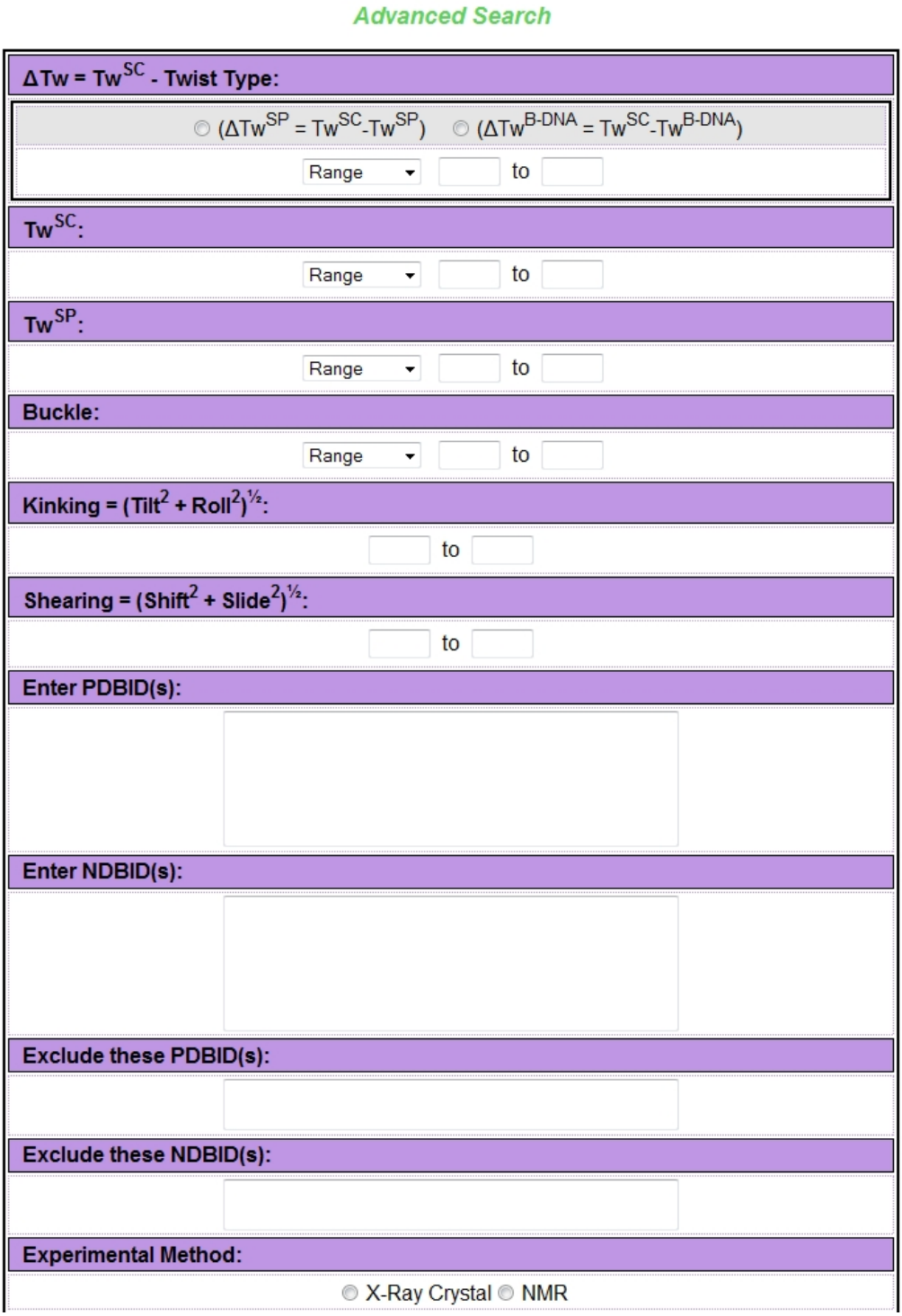


Figure 3.4.3.2: Screen capture of the top half of the Advanced Search Page without user input.

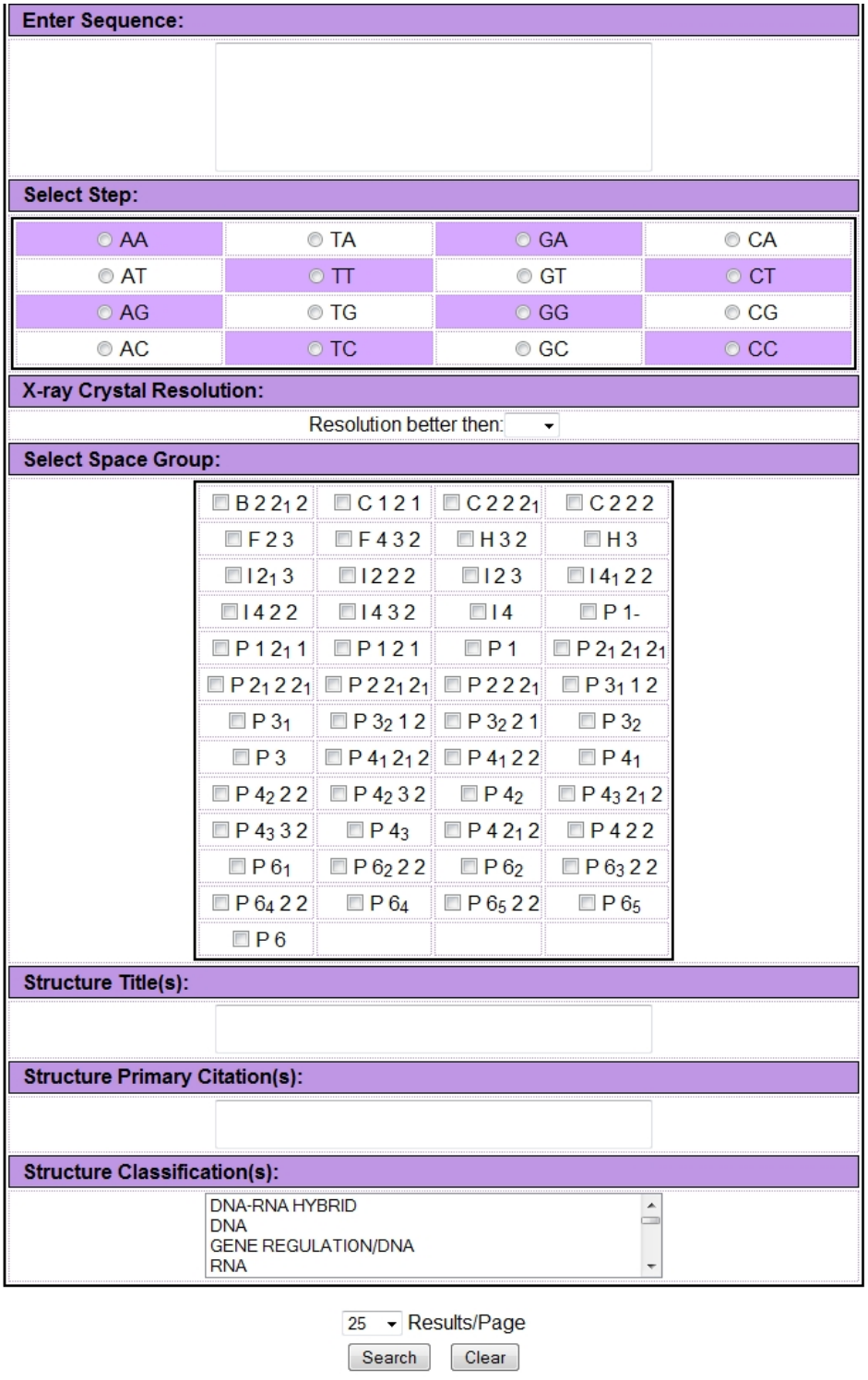
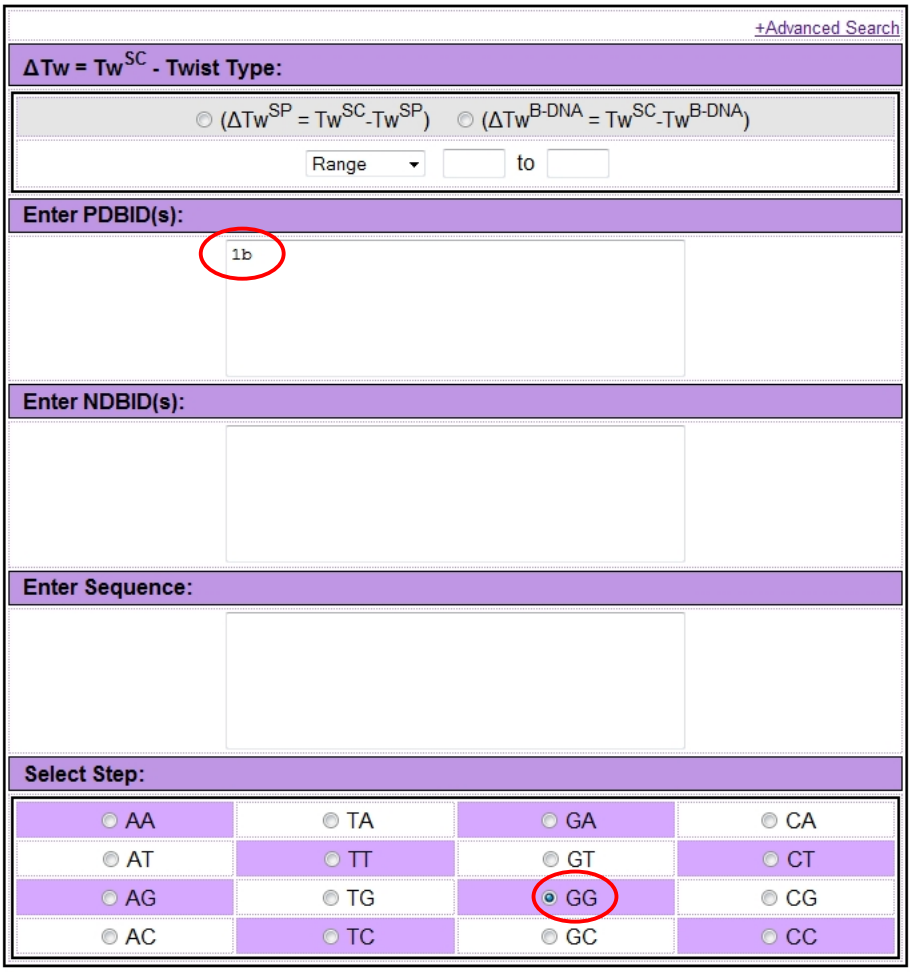


Figure 3.4.3.3: Screen capture of the bottom half of the Advanced Search Page without user input.

Both Simple and Advanced Search Pages use dynamically built-up SQL statements, generated by the user input, to retrieve their results. Each field in the search pages is gathered via the Perl script and turned into an SQL SELECT statement, that is submitted to the MySQL database when the user clicks search. For example, the Simple Search Page in Figure 3.4.3.4(a) produced the SQL statement in Figure 3.4.3.4(b). The created SELECT statement can be broken into three pieces in order to understand what it is doing. First it applies SELECT to Table.Column strings to identify which data we want from a table in the database, and it selects these data from all of the possible columns. In this example, the Table.Column names are renamed using the AS syntax. For example, Summary.TWID AS 'TwID' will return the column name as TwID. Next the SQL statement uses FROM to identify the table from which to get the columns, in this case the Summary table. The third section shows how SQL uses the WHERE clause to set the criteria determining which data to get from the columns specified in the SELECT statement and the table specified in the FROM statement. The criteria contained within the WHERE clause can be specific searches like Summary.PDBID LIKE '%1b%' and Summary.Sequence LIKE '%gg%', or empty as in "(Summary.Sequence LIKE '%%')" which would return all sequences in the database if it were used instead of the other criteria in this search. In SQL the % is a wild card that matches any alpha numeric characters, and this is what allows the PDB ID and sequence match to consist of more than 1b and gg, respectively. The WHERE clause in this case will only return results that

contain both PDBIDs containing 1b and sequences containing gg because each of the criteria in the WHERE clause is linked together by an AND, which ensures all of the criteria are met in each result returned. In searches for multiple specific PDBIDs an OR could be used to return more then one match for the same column of data. For example, a WHERE clause containing the criteria Summary.PDBID LIKE '1kx3' OR Summary.PDBID LIKE '1kx4' would return both 1kx3 and 1kx4, but an AND clause would return no data because there is no single entry in the database that contains two PDBIDs.



(a)

(b)
```
SELECT Summary.TWID AS 'TwID', Summary.NBP AS 'NBP',
Summary.Sequence, Summary.Title, Summary.PDBID,
Summary.NDBID, Summary.ExpMethod AS 'Experiment Method'

FROM Summary

WHERE ((Summary.PDBID LIKE '%1b%') ) AND
(Summary.Sequence LIKE '%gg%' )AND (Summary.Comments
LIKE 'NULL') GROUP BY Summary.TWID ORDER BY 1
LIMIT 25 OFFSET 0
```

Figure 3.4.3.4: An example of both the user input to the Simple Search Page (a) and the SQL that is generated to perform the query of the TwiDDL MySQL database (b).

In addition to the Simple and Advanced Search Pages there exists an extremely simplified quick search option. The quick search is available in every page on the TwiDDL website, and can be found at the top right hand side of the quick search bar described in Section 3.4. The quick search uses the same underlying code as the Simple and Advance Search Pages to generate the search that it performs. The primary function of the quick search feature is to allow the user to enter either a whole or partial TwID, PDB ID, or NDB ID and to retrieve the data about that structure in the results page.

3.4.4 Display of data

One of the primary goals for TwiDDL was to create a quick and effective tool for accessing, visualizing, and comparing large amounts of data based on the calculations for the twist of supercoiling. Early in the evaluation, of what was needed to demonstrate the value of this new method for calculating the twist, it became obvious that a database-driven interface to this information would be the best solution to meet these goals. In this section we will discuss the design decisions and implementation of TwiDDL and the variety of ways in which the data can be displayed.

As discussed in Section 3.4.3, the quick search and search pages are the primary ways in which a user may access the data within TwiDDL. In order to retrieve data from the database a user would enter search criteria in one of the search features, and in return the search would produce a results page. On the results page the user is presented with

the first high-level summary view of what structures stored in TwiDDL met the selected criteria. The results page can be described at a very high level as having three major sections related to the search results themselves. The results page utilizes the same overall web page design as the rest of the TwiDDL site, where the three search result sections are contained within the body of the common page layout. The basic overall layout can be seen in Figure 3.4.4.1.

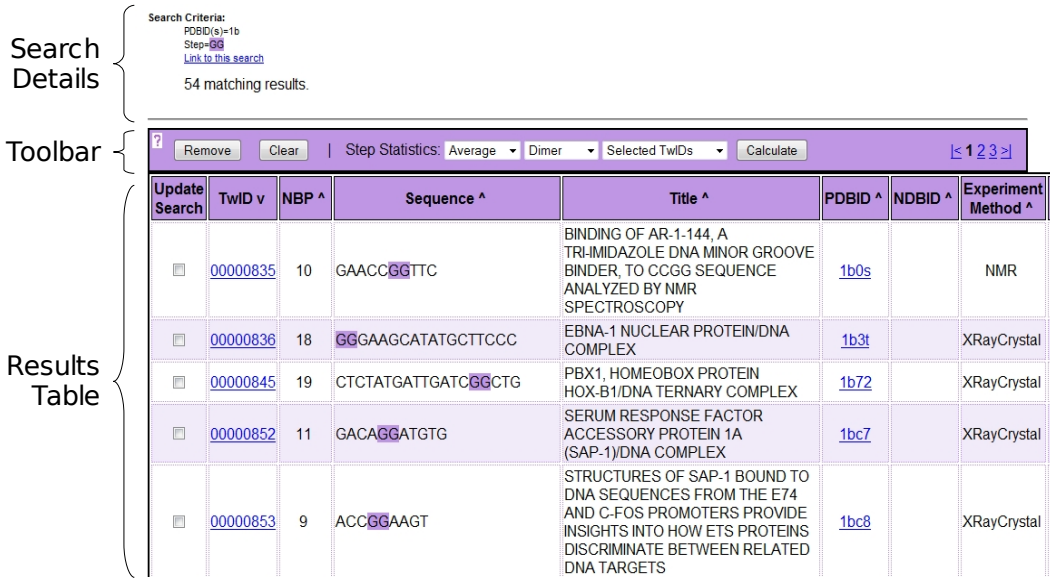


Figure 3.4.4.1: The search results page returned by any search input by a user to TwiDDL with the three major subsections highlighted, the search details, the toolbar, and the results table.

The first of the three major sections that make up the search results page contains details about the search itself, including the search criteria, a link to the current search, and a report of how many structures were returned as matching results to the search criteria that was entered. In the upper left portion of the body of the results page a list of

all criteria entered into the Simple or Advanced Search Page is displayed to show what produced the results being displayed. Below that list of search criteria there is a link to this search which is intended for easy copying and pasting into email to share with others, or to bookmark in the web browser in order to get back to these exact same results. Just below the link to this search the total number of entries in TwiDDL that matched the search criteria will be displayed here as "54 matching results.".

The other two major sections in the body of the search results page are the toolbar and the results table. They are considered separate sections because the form and function of each differs, even though they have a tightly coupled interaction with each other. Within the toolbar, there are a few basic functions for interacting with the information supplied in the results table. The toolbar functions consist of removal of structures from the results table, clearing any input to the web form, generating simple statistics, and paging through the search results. The results table consists of a matrix of rows and columns where the rows pertain to each structure returned due to matching the search criteria input by the user, with the top row containing a descriptive title for each of the columns, and the columns showing a short summary of details about each structure.

When using the toolbar, there are a few different ways to examine the list of structures in the results table. The primary link between the toolbar's functions and the results table entries can be found in the first column of the results table, called the Update

Search column. This column simply contains a check box for each row or structure in the table. Clicking or checking the small boxes on the left of each row allows the user to select certain TwIDs for further manipulation by the functions of the toolbar. The layout of the toolbar and the Update Search column can be seen in the example shown in Figure 3.4.4.2.

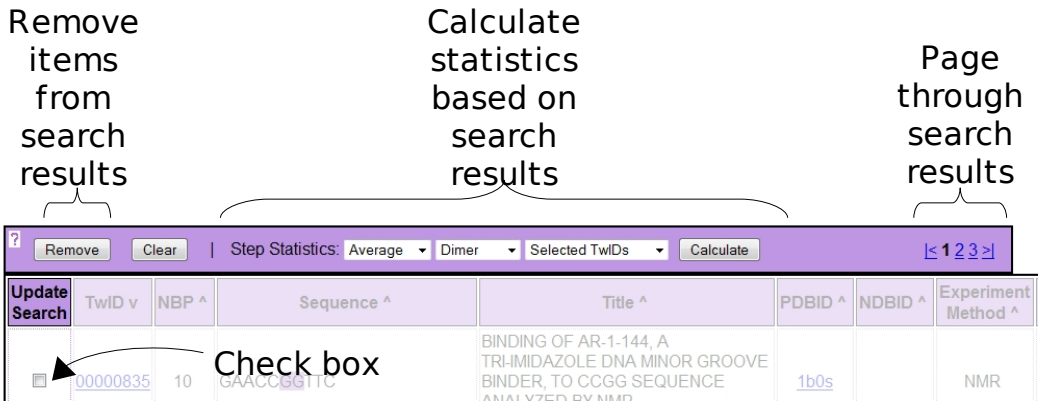


Figure 3.4.4.2: The toolbar from the TwiDDL results page with its three major sections highlighted, as well as the check boxes used to manipulate the search results via the toolbar.

The first function allows the user to remove the selected TwIDs from the current search, by selecting the structures to remove via the check boxes for these structures and then by clicking the Remove button on the toolbar. The selection can be added back to the set of structures by clicking on the Not TwID(s)= links found in the Search Criteria section at the top of the page. The remove function of the toolbar is very powerful because it allows users to refine a search to the exact list structures that they would like to see the details of and make comparisons between.

One of the ways the users can evaluate the structures returned by the search criteria is the ability to obtain simple statistical information through the Step Statistics section of the toolbar. The toolbar presents the ability to select the type of desired statistics about a structure based on three drop down lists. The first drop down list allows the user to choose whether to calculate the average, minimum, maximum, or all three respective values for the Tw^{SC} and the other local conformational parameters for each of the base-pair steps in the desired structures. The toolbar also provides the user a drop down list to select the sequence context, either dimeric or tetrameric, in which the statistics returned will be tabulated. The final option that the user may select regarding the statistics returned is the choice of structures for which statistics will be calculated. The drop-down list offers two options, either the Selected TwIDs or the All TwIDs Matched. The Selected TwiDs option will only calculate statistics for the structures with the check boxes selected in the Update Search column of the results table. The All TwIDs Matched option will calculate statistics for all of the structures returned by the search. Once the choices are made, the user may initiate the statistical calculations by clicking on the Calculate button in the toolbar. This action opens a new window for presenting the desired statistics about the chosen structures separate from the results page.

Additionally, the toolbar gives the user some basic navigation functions. The Clear button allows the user to deselect all of the rows that have been checked in the

Update Search column. The Clear button also resets the three Step Statistic drop down lists to their default values. The final feature of the toolbar is the page navigation on the far right side of the toolbar. This simply allows the user to scroll through the pages of the search results with the ability to jump to the beginning or the end of the results pages, as well as navigate directly to the five nearest pages of results.

The results table is designed to provide the user a typical set of columns pertaining to the structures that matched the input search criteria. When the search is performed using the Simple Search Page the results table is made up of nine columns, unless ΔTw^{SP} or $\Delta Tw^{B\text{-}DNA}$ is included in the search. In this case a tenth column is displayed for the ΔTw. However, when the search is input using the Advanced Search the set of columns displayed will change dynamically with the addition of several new columns relevant to the search that was made.

From the Simple Search the table typically displays the nine columns shown in Figure 3.4.4.3(a). The first column is the Update Search column whose rows contain the check boxes which can be used to modify the search or calculate the Step Statistics from the toolbar. The second column contains the TwID, TwiDDL's unique structural identifier, whose rows contain the TwID for each structure. The TwID in each row is also a link that brings the user to a new page that contains more details about the twist of supercoiling and other related parameters characterizing the specific structure. The third

column is NBP or the number of base pairs in the structure. The fourth and fifth columns are respectively the Sequence, and Title of the structure. The sixth column displays the PDB ID and links to the PDB entry for the structure, in the same way that the TwID column links to more details about the twist of supercoiling. The seventh column displays the NDB ID, when there is an NDB entry, and it links to the NDB entry for the structure. The eighth column lists the Experimental Method, either NMR (nuclear magnetic resonance spectroscopy) or XRayCrystal (X-Ray Crystallography), used to determine the structure. In the case where the ΔTw^{SP} or $\Delta Tw^{B\text{-}DNA}$ is included in the search, an additional column is inserted in the results table in the fourth position as can be seen in Figures 3.4.4.3(b) and (c). The title of this column describes which ΔTw was examined, and the column will display the first value found within the structure that matched the range or magnitude of the criterion that the user entered. In order to display a table containing the values for ΔTw^{SP} and $\Delta Tw^{B\text{-}DNA}$ at each base-pair step, the user may click the TwID link and then view the Twist Comparisons table by clicking Show to see the full table of values.

(a)

1 2 3 4 5 6 7 8 9

Update Search	TwID v	NBP ^	Sequence ^	Title ^	PDBID ^	NDBID ^	Experiment Method ^	Step Details
☐	00000835	10	GAACCGGTTC	BINDING OF AR-1-144, A TRI-IMIDAZOLE DNA MINOR GROOVE BINDER, TO CCGG SEQUENCE ANALYZED BY NMR SPECTROSCOPY	1b0s		NMR	Show
☐	00000836	18	GGGAAGCATATGCTTCCC	EBNA-1 NUCLEAR PROTEIN/DNA COMPLEX	1b3t		XRayCrystal	Show
☐	00000845	19	CTCTATGATTGATCGGCTG	PBX1, HOMEOBOX PROTEIN HOX-B1/DNA TERNARY COMPLEX	1b72		XRayCrystal	Show
☐	00000852	11	GACAGGATGTG	SERUM RESPONSE FACTOR ACCESSORY PROTEIN 1A (SAP-1)/DNA COMPLEX	1bc7		XRayCrystal	Show
☐	00000853	9	ACCGGAAGT	STRUCTURES OF SAP-1 BOUND TO DNA SEQUENCES FROM THE E74 AND C-FOS PROMOTERS PROVIDE INSIGHTS INTO HOW ETS PROTEINS DISCRIMINATE BETWEEN RELATED DNA TARGETS	1bc8		XRayCrystal	Show
☐	00000864	6	AAGGAA	INTRAMOLECULAR TRIPLEX, NMR,	1bce-1		NMR	Show

(b)

Update Search	TwID v	NBP ^	ΔTw^{SP} ^	Sequence ^	Title ^	PDBID ^	NDBID ^	Experiment Method ^	Step Details
☐	00000003	7	1.22	CCTTTTC	SOLUTION STRUCTURE OF THE COVALENT DUOCARMYCIN A-DNA DUPLEX COMPLEX	107d-1		NMR	Show
☐	00000004	7	1.79	CCTTTTC	SOLUTION STRUCTURE OF THE COVALENT	107d-2		NMR	Show

(c)

Update Search	TwID v	NBP ^	$\Delta Tw^{B\text{-}DNA}$ ^	Sequence ^	Title ^	PDBID ^	NDBID ^	Experiment Method ^	Step Details
☐	00000002	12	2.64	CGCAAATTTGCG	SEQUENCE-DEPENDENT DRUG BINDING TO THE MINOR GROOVE OF DNA: THE CRYSTAL STRUCTURE OF THE DNA DODECAMER D(CGCAAATTTGCG) 2 COMPLEXED WITH PROPAMIDINE	102d		XRayCrystal	Show
☐	00000004	7	1.24	CCTTTTC	SOLUTION STRUCTURE OF THE COVALENT	107d-2		NMR	Show

Figure 3.4.4.3: Three examples of the results table generated by the Simple Search Page.

The ninth column contains the Step Details table for each structure, and the word Show to link to this information, as seen in Figure 3.4.4.4(a). When the Show link is clicked it presents a table of details pertaining to each base-pair step in the structure, as

can be seen in Figure 3.4.4.4(b). After Show is clicked to display the table, the top of each Step Details table contains another link called Hide, which will remove the table from the user's view and display only the Show link in its place again. When displayed, the Step Details table consists of ten columns including the Base Pair Step, the six rigid-body base-pair step parameters (Shift, Slide, Rise, Tilt, Roll, Twist or Tw^{SP}), the twist of supercoiling Tw^{SC}, the net kinking per base-pair step, and the net shearing per base-pair step.

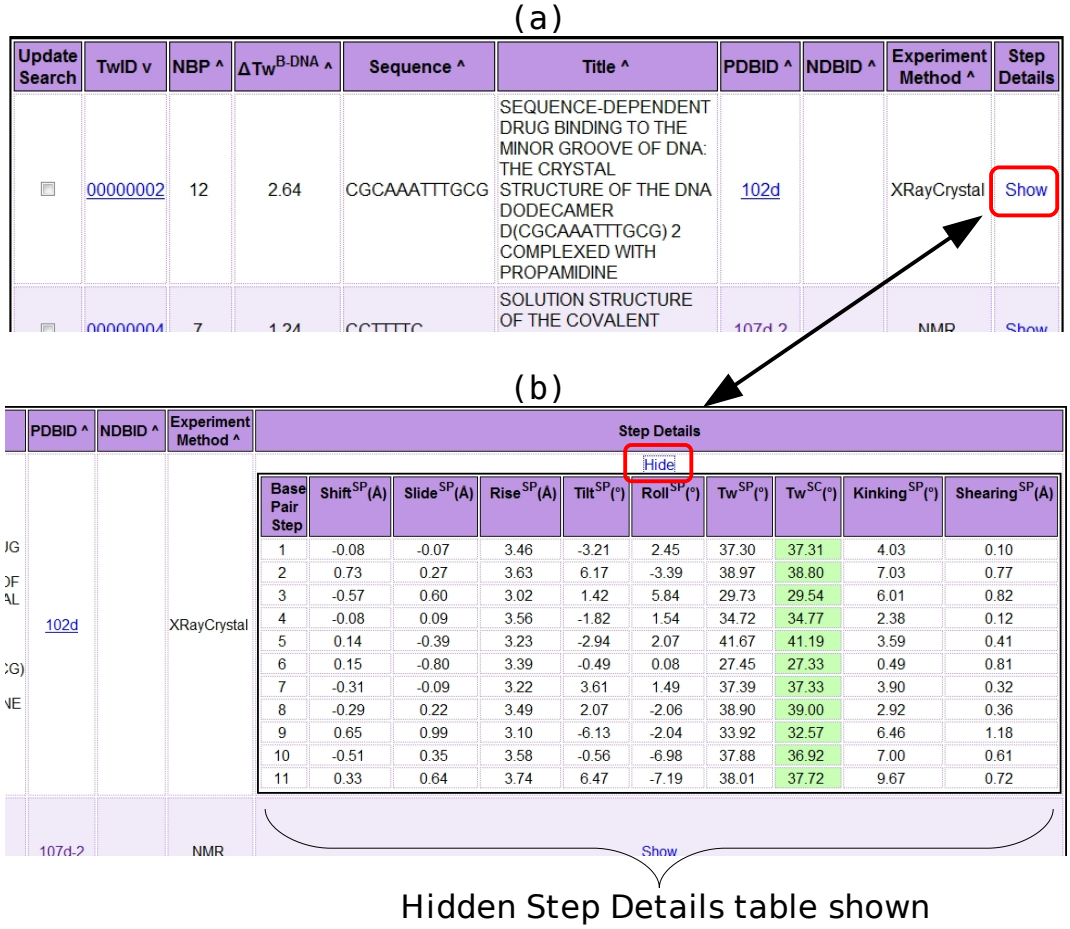


Figure 3.4.4.4: An example of the Show/Hide table feature in TwiDDL used to display a table of details about the step parameters, the twist of supercoiling, as well as the kinking and shearing, for each step in the structure.

The results table displayed by the Advanced Search Page can contain various columns that the results table from a Simple Search will not. In the results table generated by an Advanced Search these additional columns may include the Tw^{SP}, the Tw^{SC}, the Space Group, the Primary Citation, the Resolution, and the Classification of the structure. All of these columns simply display the structural information that pertains to

the respective column title, as can be seen in Figure 3.4.4.5(a). However, there are a few other columns, which like the Step Details column, contain a Show link and produce a table of values when Show is clicked. The quantities in these tables include the Buckle, Kinking, and Shearing in the structure. These tables can be shown or hidden using the respective Show or Hide links within the columns, as can be seen in Figure 3.4.4.5(a) and (b).

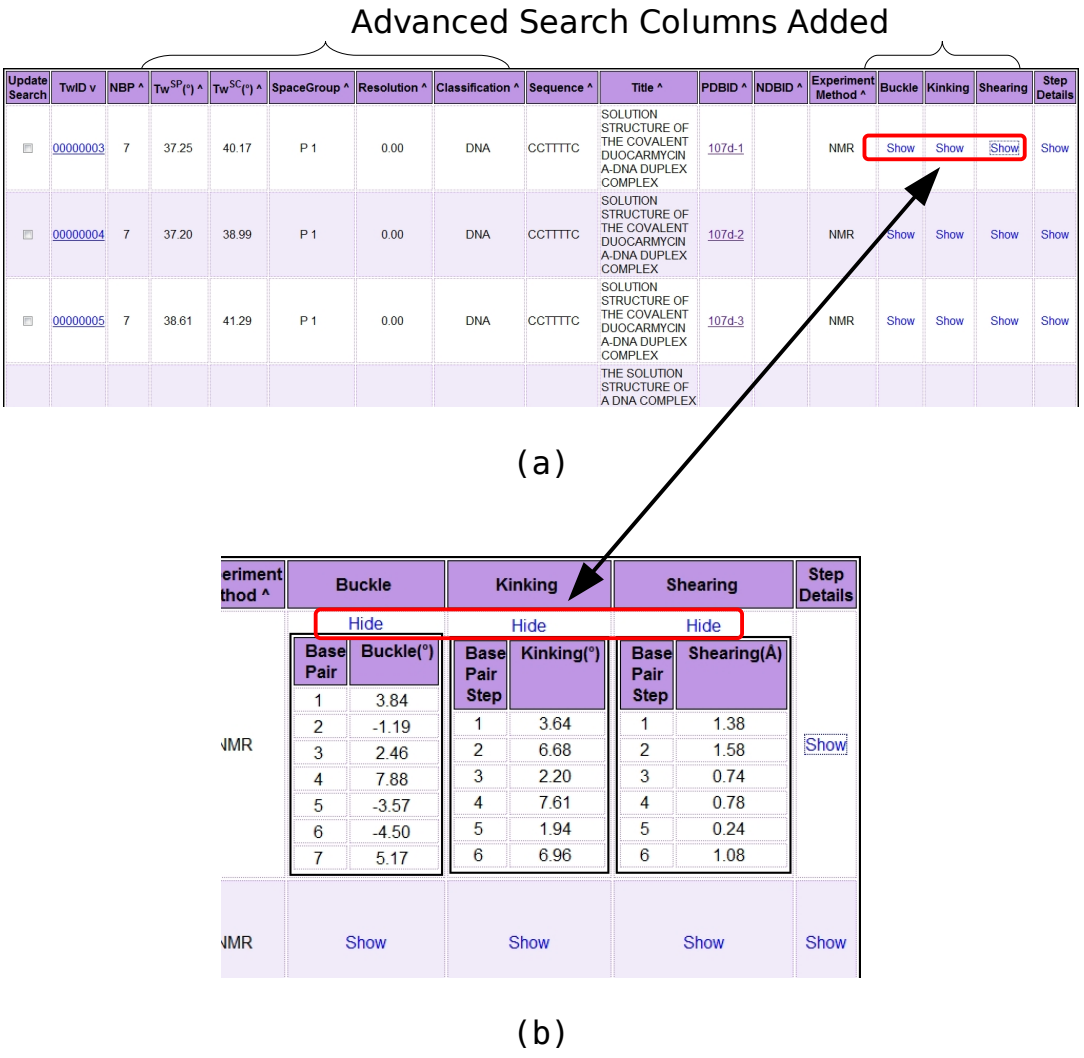


Figure 3.4.4.5: An example of the results table generated by the Advanced Search Page.

The results table discussed so far only presents a high-level overview about the selected structures in order to help the user decide what information they intend to review. Typically the results table is intended for the user to review enough details about the selected structures in order to determine whether they merit the more information analysis available about each structure and the twist of supercoiling values. The way that

TwiDDL allows a user to access more details about a specific structure is through the link presented in the TwID column. This link brings the user to the TwID details page as seen in Figure 3.4.4.6.

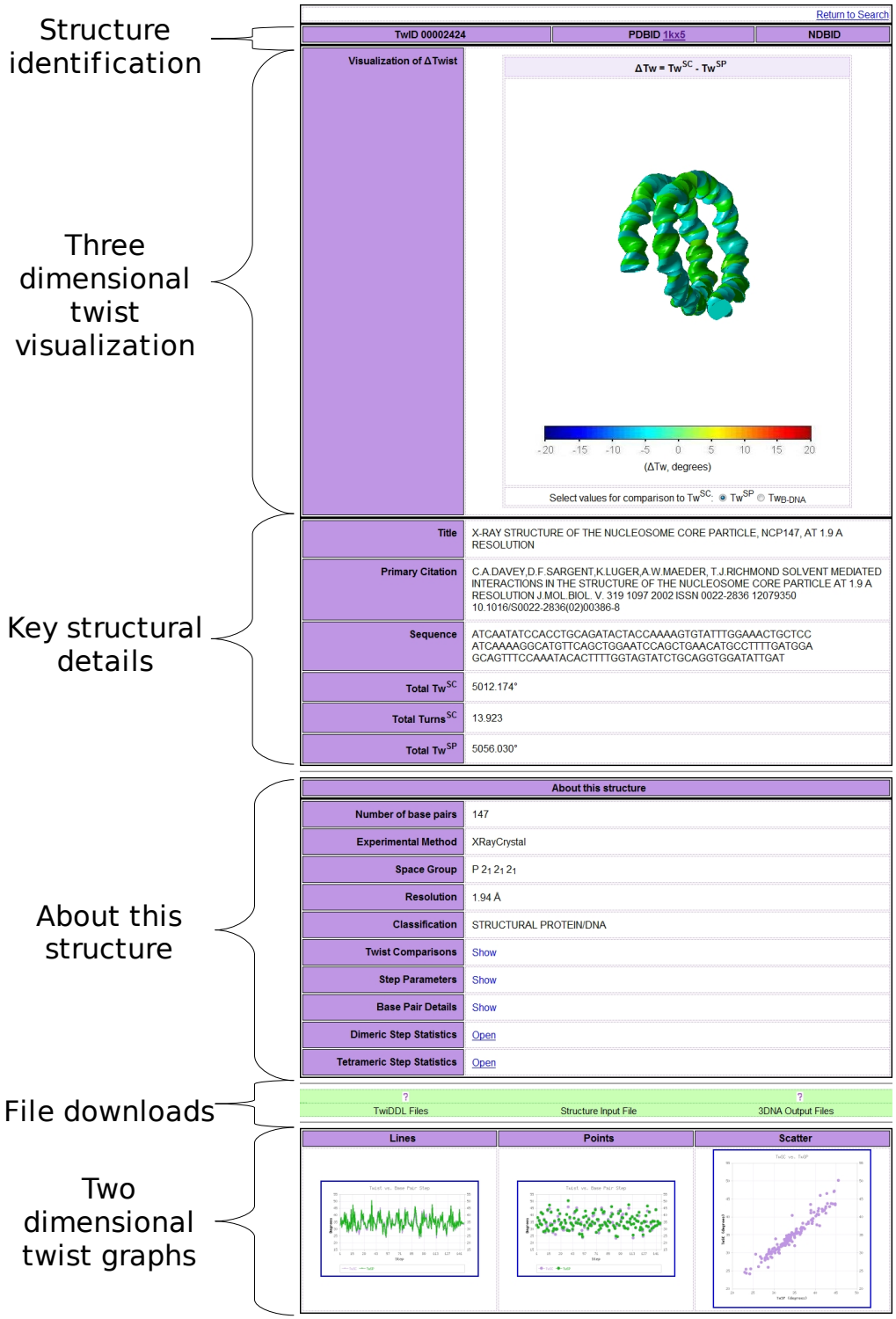
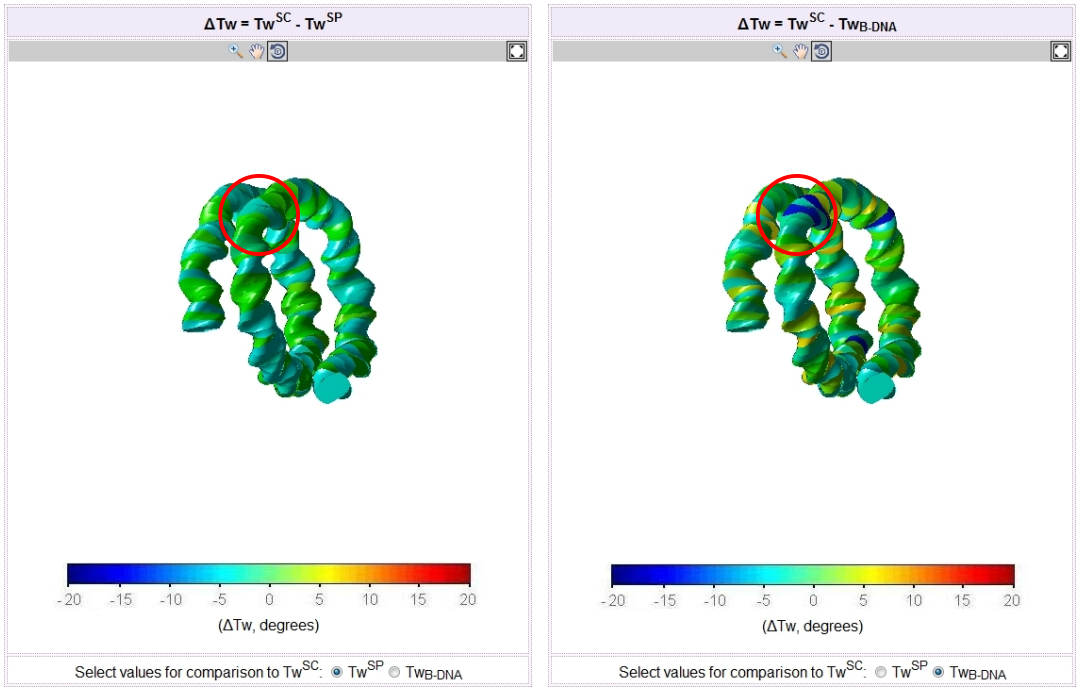


Figure 3.4.4.6: A screen capture of the TwID details page and its six sections highlighted.

The TwID details page is broken into six major sections as highlighted in Figure 3.4.4.6. The first section, used for structure identification, is made up of three fields. Each of the fields contain an identifier for the structure from the three major databases: TwiDDL, PDB, and NDB. In some cases, the NDB does not contain the respective structure, and therefore the NDB ID is not present.

The second section on the TwID details page contains a three-dimensional visualization of the ΔTw, that can dynamically be manipulated to view the structure from various angles. The visualization tool in this section graphically shows the variation in the difference, ΔTw, in the twist of supercoiling (Tw^{SC}) with respect to that of B-DNA ($Tw^{B\text{-}DNA}$) or relative to the step-parameter twist (Tw^{SP}), along the selected structure. The visualization displays color-coded images with the values mapped on to the base-pair steps using either the ΔTw^{SP} as seen in Figure 3.4.4.7(a) or the $\Delta Tw^{B\text{-}DNA}$ as seen in Figure 3.4.4.7(b). Each of these views is obtained by choosing the desired option with the radio button below the legend that displays the range of values corresponding to each color. This tool allows the user to zoom in or out, position the image, and rotate the model by using the tools at the top of the model box (a magnifying glass, a hand, and a circular arrow, respectively).



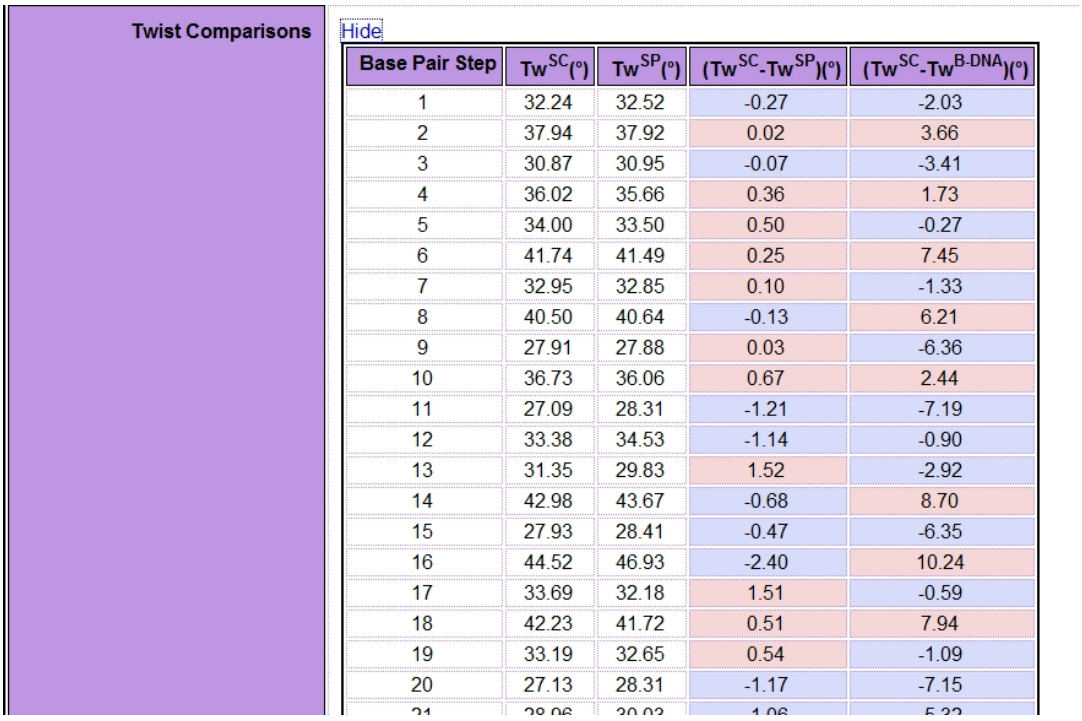
(a) (b)

Figure 3.4.4.7: Color coding ΔTw differs dependent on comparison versus (a) Tw^{SP} or (b) $Tw_{B\text{-}DNA}$. The red circled area in each model shows an example of the differences.

The third section in the TwID details page includes key structural details such as the Title, the Primary Citation, the Sequence, the Total Tw^{SC}, the Total $Turns^{SC}$, and the Total Tw^{SP}. This section gives the user a high-level overview of the structure, and some aggregate metrics that are indicative of how the twist of supercoiling calculations differ from the traditional step parameter twist, or Tw^{SP}. The fourth section is titled About this structure and contains more details about the structure as a whole, as well as some links that allow the user to view more details about the structure. The information in this section include the number of base pairs, the Experimental Method, the Space Group, the

Resolution, and the Classification, all of which are self explanatory.

The fourth section also contains three links to show more details through tables of Twist Comparisons, Step Parameters, and Base Pair Details. The Twist Comparisons row contains a Show link that displays a table, as shown in Figure 3.4.4.8, containing the Tw^{SC}, Tw^{SP}, ΔTw^{SP}, and $\Delta Tw^{B\text{-}DNA}$, and the base-pair step numbers at which these values occur in the selected structure. The list can be ordered by the entries in any column by clicking on the column header. Click once for ascending order and twice for descending order. The values of ΔTw^{SP} and $\Delta Tw^{B\text{-}DNA}$ are color-coded in blue and red to draw attention to negative or positive values, respectively.


Twist Comparisons

Hide

Base Pair Step	Tw^{SC}(°)	Tw^{SP}(°)	(Tw^{SC}-Tw^{SP})(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°)
1	32.24	32.52	-0.27	-2.03
2	37.94	37.92	0.02	3.66
3	30.87	30.95	-0.07	-3.41
4	36.02	35.66	0.36	1.73
5	34.00	33.50	0.50	-0.27
6	41.74	41.49	0.25	7.45
7	32.95	32.85	0.10	-1.33
8	40.50	40.64	-0.13	6.21
9	27.91	27.88	0.03	-6.36
10	36.73	36.06	0.67	2.44
11	27.09	28.31	-1.21	-7.19
12	33.38	34.53	-1.14	-0.90
13	31.35	29.83	1.52	-2.92
14	42.98	43.67	-0.68	8.70
15	27.93	28.41	-0.47	-6.35
16	44.52	46.93	-2.40	10.24
17	33.69	32.18	1.51	-0.59
18	42.23	41.72	0.51	7.94
19	33.19	32.65	0.54	-1.09
20	27.13	28.31	-1.17	-7.15
21	28.06	30.03	-1.96	-5.32

Figure 3.4.4.8: An example of the Twist Comparisons table expanded, showing both red cells for over twisted and blue cells for under twisted values relative to the Tw^{SC}.

The next row in the fourth section of the TwID details page is the Step Parameters table, as seen in figure Figure 3.4.4.9, which can be shown or hidden through the Show or Hide links. The table displays the six rigid-body base-pair-step parameters (here labeled Shift^{SP}, Slide^{SP}, Rise^{SP}, Tilt^{SP}, Roll^{SP}, Tw^{SP}), two quantities based on those step parameters (Kinking^{SP}, Shearing^{SP}), and the base-pair-step number at which these values occur. Like other tables in TwiDDL the list can be ordered by the entries in any column by clicking on the column header where one click is for ascending order and twice is for descending order.

Step Parameters

Hide

Base Pair Step	Shift^{SP}(Å)	Slide^{SP}(Å)	Rise^{SP}(Å)	Tilt^{SP}(°)	Roll^{SP}(°)	Tw^{SP}(°)	Kinking^{SP}(°)	Shearing^{SP}(Å)
1	-0.87	-0.61	3.07	2.35	0.78	32.52	2.47	1.06
2	0.78	0.02	3.46	-0.63	5.20	37.92	5.23	0.78
3	-0.17	0.13	3.35	1.73	6.59	30.95	6.81	0.21
4	-0.54	0.26	3.27	-5.56	2.17	35.66	5.96	0.59
5	0.59	-0.54	3.25	2.93	-2.17	33.50	3.64	0.79
6	0.06	-0.18	3.33	-1.97	-5.69	41.49	6.02	0.18
7	0.24	-0.80	3.35	0.52	-3.20	32.85	3.24	0.83
8	0.20	-0.31	3.24	2.48	-1.18	40.64	2.74	0.36
9	-0.19	-0.06	3.24	0.69	7.50	27.88	7.53	0.19
10	-0.48	0.31	3.25	-4.27	5.64	36.06	7.07	0.57
11	0.52	-0.78	3.38	0.42	4.46	28.31	4.47	0.93
12	0.38	-0.59	3.37	1.26	8.87	34.53	8.95	0.70
13	-0.63	0.27	3.47	0.93	9.81	29.83	9.85	0.68
14	0.22	1.41	3.21	-0.26	-6.25	43.67	6.25	1.42
15	-0.19	0.29	2.99	4.28	4.75	28.41	6.39	0.34
16	0.07	2.47	3.20	-2.79	-11.14	46.93	11.48	2.47
17	0.75	-0.13	3.52	0.76	-2.18	32.18	2.30	0.76
18	0.66	-0.35	3.53	2.35	-4.20	41.72	4.81	0.74
19	0.26	-0.44	3.15	3.87	-2.87	32.65	4.81	0.51
20	-0.17	-0.86	3.41	-2.53	10.91	28.31	11.19	0.87
21	1.06	0.49	3.02	3.49	12.68	30.03	13.15	1.16

Figure 3.4.4.9: An example of the Step Parameters table expanded.

Next in the fourth section is the Base Pair Details row that contains a table, as seen in Figure 3.4.4.10, that displays the chemical identities and residue numbers (Base

ID I, Base ID II) of the paired bases in the selected structure, the identities of the complementary strands (Chain ID I, Chain ID II), the names of the bases (Residue ID I, Residue ID II), the numbers of the bases (Residue Num I, Residue Num II), and the six rigid-body base-pair parameters (Shear, Stretch, Stagger, Buckle, Propeller, Opening) that describe the spatial arrangement of the pair bases. The list can be ordered by the entries in any column by clicking on the column header, where a single click for ascending order and a second for descending order. In this table a header that ends with a I holds information about the base on Strand I and a header that ends with a II holds information about the base on Strand II.

Base Pair Details

Hide

Base Pair	Base ID I	Base ID II	Chain ID I	Chain ID II	Residue ID I	Residue ID II	Residue Num I	Residue Num II	Shear (Å)	Stretch (Å)	Stagger (Å)	Buckle (°)	Propeller (°)	Opening (°)
1	DA_-73	DT_73	I	J	A	T	-73	73	-0.32	-0.58	-0.34	-3.80	4.48	-0.73
2	DT_-72	DA_72	I	J	T	A	-72	72	-0.54	-0.14	-0.64	6.21	-10.27	-2.31
3	DC_-71	DG_71	I	J	C	G	-71	71	-0.14	-0.03	-0.36	2.77	-4.61	1.05
4	DA_-70	DT_70	I	J	A	T	-70	70	0.07	-0.11	-0.33	7.19	-14.35	5.75
5	DA_-69	DT_69	I	J	A	T	-69	69	0.00	-0.10	0.19	3.61	-10.75	1.67
6	DT_-68	DA_68	I	J	T	A	-68	68	0.00	-0.40	0.10	9.23	-6.74	2.16
7	DA_-67	DT_67	I	J	A	T	-67	67	0.34	-0.05	0.48	5.13	-15.52	0.19
8	DT_-66	DA_66	I	J	T	A	-66	66	0.20	-0.08	0.30	2.47	-10.16	0.34
9	DC_-65	DG_65	I	J	C	G	-65	65	0.36	-0.10	0.04	-1.17	-8.49	3.25
10	DC_-64	DG_64	I	J	C	G	-64	64	0.25	-0.24	-0.16	1.26	-0.86	-4.52
11	DA_-63	DT_63	I	J	A	T	-63	63	0.57	0.00	0.18	4.21	-7.33	-1.95
12	DC_-62	DG_62	I	J	C	G	-62	62	-0.13	0.02	0.02	1.86	-6.28	4.94
13	DC_-61	DG_61	I	J	C	G	-61	61	0.28	-0.06	0.36	-5.88	-11.40	5.61
14	DT_-60	DA_60	I	J	T	A	-60	60	-0.40	-0.13	0.35	-13.25	-15.50	2.23
15	DG_-59	DC_59	I	J	G	C	-59	59	-0.76	-0.09	0.18	-9.02	-3.96	1.38
16	DC_-58	DG_58	I	J	C	G	-58	58	0.27	-0.08	-0.12	5.84	-5.87	-0.94
17	DA_-57	DT_57	I	J	A	T	-57	57	0.01	0.27	0.50	14.01	-15.00	-3.06
18	DG_-56	DC_56	I	J	G	C	-56	56	-0.13	0.13	0.28	8.37	-18.28	-3.15
19	DA_-55	DT_55	I	J	A	T	-55	55	0.40	-0.02	0.06	-4.76	-12.70	2.23
20	DT_-54	DA_54	I	J	T	A	-54	54	0.74	-0.29	-0.33	-1.08	-7.50	5.08
21	DA_-53	DT_53	I	J	A	T	-53	53	0.18	0.36	0.19	8.18	7.49	6.85

Figure 3.4.4.10: An example of the Base Pair Details table expanded.

The next two rows in the fourth section of the TwID details page consist of the Dimeric and Tetrameric Step Statistics links. Clicking on the Open link brings the user to a new page with statistical information — the average, minimum, and maximum values of

Tw^{SC}, ΔTw^{SP}, $\Delta Tw^{B\text{-}DNA}$, and the six rigid-body base-pair-step parameters (Tw^{SP}, $Shift^{SP}$, $Slide^{SP}$, $Rise^{SP}$, $Tilt^{SP}$, $Roll^{SP}$) — describing the base-pair steps within the selected structure. The Open link in the Dimeric Step Statistics row opens a new page as show in Figure 3.4.4.11, where the values are presented in a dimeric context for the 16 possible combinations of consecutive Watson-Crick base pairs. That is, all steps with a given dimeric sequence are grouped together regardless of the identities of the base pairs that flank them. The number of steps that contribute to the averages is listed under Number of Matching Steps.

Similarly, in Figure 3.4.4.12 the Tetrameric Step Statistics page is opened from the link in that row, and the values are presented in a tetrameric context for the 256 possible combinations of consecutive base pairs. That is, all steps with a common dimeric sequence are further grouped in terms of the identities of the base pairs that flank the step. The number of steps that contribute to the averages is listed under Number of Matching Steps. It is important to note that the first and last steps of the structure in a tetrameric context will not have a flanking 5´ or 3´ base pair. These terminal base-pair steps are easily identified from the blank spaces located on either side of the trimer.

Step	Avg Tw^{SC}(°)	Avg ΔTw^{SC}(°)	Avg $\Delta Tw^{B\text{-}DNA}$(°)	Avg Tw^{SP}(°)	Avg $Shift^{SP}$(Å)	Avg $Slide^{SP}$(Å)	Avg $Rise^{SP}$(Å)	Avg $Tilt^{SP}$(°)	Avg $Roll^{SP}$(°)	Number of Matching Steps
AA/TT	34.12	0.16	-0.16	33.95	-0.02	0.06	3.31	-0.32	4.32	13
AC/GT	29.44	-0.78	-4.84	30.22	0.04	-0.48	3.22	-0.43	7.49	7
AG/CT	32.76	-0.22	-1.52	32.98	0.20	0.31	3.38	-1.08	0.69	9
AT/AT	31.07	-0.22	-3.21	31.29	0.00	-0.62	3.12	0.82	0.53	15
CA/TG	37.96	-0.30	3.68	38.27	0.00	1.04	3.39	0.39	1.88	14
CC/GG	32.29	-0.88	-1.99	33.17	-0.18	0.41	3.36	0.94	3.83	7
CT/AG	34.10	-0.40	-0.18	34.50	-0.12	0.36	3.37	-0.85	0.73	9
GA/TC	36.81	-0.14	2.52	36.95	-0.49	0.14	3.28	-2.77	2.01	8
GC/GC	37.48	-0.09	3.19	37.57	-0.05	0.74	3.28	1.07	-3.16	8
GG/CC	32.16	0.00	-2.11	32.17	0.24	0.30	3.47	0.21	5.77	7
GT/AC	29.19	-1.08	-5.08	30.28	-0.16	-0.46	3.06	-0.66	6.21	7
TA/TA	34.64	-1.17	0.35	35.81	0.12	-0.41	3.39	-0.94	5.66	8
TC/GA	36.84	0.14	2.56	36.70	0.45	0.06	3.31	1.22	2.26	8
TG/CA	37.46	-0.59	3.17	38.06	0.00	0.93	3.36	-1.22	0.93	14
TT/AA	34.47	0.18	0.18	34.29	-0.02	-0.03	3.31	0.67	3.12	12

Step	Max Tw^{SC}(°)	Max ΔTw^{SC}(°)	Max $\Delta Tw^{B\text{-}DNA}$(°)	Max Tw^{SP}(°)	Max $Shift^{SP}$(Å)	Max $Slide^{SP}$(Å)	Max $Rise^{SP}$(Å)	Max $Tilt^{SP}$(°)	Max $Roll^{SP}$(°)
AA/TT	40.90	2.77	6.61	40.56	0.78	1.05	3.92	5.06	23.80
AC/GT	37.20	0.27	2.92	36.93	0.52	-0.13	3.41	4.62	13.86
AG/CT	39.00	1.51	4.71	40.99	1.56	1.56	3.81	6.43	12.67
AT/AT	34.92	0.54	0.64	35.08	0.89	-0.40	3.35	4.07	7.00
CA/TG	44.88	2.66	10.59	47.15	1.48	2.47	3.90	7.22	12.45
CC/GG	41.25	0.45	6.97	45.87	0.48	1.94	3.66	6.64	8.87
CT/AG	40.85	3.12	6.57	42.21	1.28	1.66	3.57	4.06	17.04
GA/TC	43.50	0.75	9.22	42.75	0.66	1.34	3.53	2.35	7.33
GC/GC	42.55	1.27	8.26	42.30	1.50	1.30	3.78	6.11	7.57
GG/CC	42.57	0.67	8.28	42.75	1.25	1.08	4.01	9.78	14.08
GT/AC	32.31	-0.08	-1.97	32.40	1.34	-0.15	3.38	8.34	10.67
TA/TA	43.87	0.27	9.59	43.60	0.52	0.19	3.67	3.26	17.84
TC/GA	44.88	1.63	10.59	43.25	1.33	0.97	3.80	6.12	8.83
TG/CA	45.62	3.63	11.34	50.04	1.07	2.58	3.81	6.45	21.63
TT/AA	40.85	1.09	6.57	40.48	0.52	0.46	3.66	3.22	11.25

Step	Min Tw^{SC}(°)	Min ΔTw^{SC}(°)	Min $\Delta Tw^{B\text{-}DNA}$(°)	Min Tw^{SP}(°)	Min $Shift^{SP}$(Å)	Min $Slide^{SP}$(Å)	Min $Rise^{SP}$(Å)	Min $Tilt^{SP}$(°)	Min $Roll^{SP}$(°)
AA/TT	23.45	-1.07	-10.83	24.24	-1.00	-0.31	2.83	-9.83	-6.93
AC/GT	23.34	-1.99	-10.94	25.34	-1.06	-0.78	3.02	-3.84	1.05
AG/CT	28.29	-1.98	-5.98	28.19	-1.34	-0.48	2.96	-6.70	-12.38
AT/AT	24.26	-2.06	-10.02	24.04	-0.88	-1.06	2.88	-3.13	-3.53
CA/TG	30.87	-4.48	-3.41	30.95	-1.27	0.08	3.00	-6.13	-15.60
CC/GG	27.91	-4.61	-6.36	27.88	-1.20	-0.59	3.17	-5.77	-8.32
CT/AG	30.48	-5.87	-3.79	29.83	-1.15	-0.48	3.21	-4.67	-9.28
GA/TC	32.58	-1.14	-1.70	33.63	-1.43	-0.55	3.04	-6.65	-4.39
GC/GC	27.93	-1.45	-6.35	28.41	-1.30	0.29	2.76	-3.00	-12.52
GG/CC	26.99	-1.12	-7.29	27.25	-1.32	-0.39	3.18	-3.52	-3.21
GT/AC	24.44	-1.93	-9.84	25.44	-0.78	-0.86	2.75	-5.46	2.45
TA/TA	25.71	-3.85	-8.57	28.31	-0.17	-1.05	3.02	-8.34	-7.49
TC/GA	32.02	-0.54	-2.25	32.02	-0.34	-0.60	3.08	-4.02	-3.18
TG/CA	23.01	-4.41	-11.27	24.50	-1.67	-0.53	3.18	-13.42	-18.43
TT/AA	29.57	-0.43	-4.71	28.61	-0.36	-0.54	3.02	-3.20	-5.65

Figure 3.4.4.11: An example of the Dimeric step statistics generated by TwiDDL. In this case it is for a single structure, but using TwiDDL's toolbar on the search results page the user can generate statistics about multiple structures.

Tetramer	Avg Tw^{SC}(°)	Avg ΔTw^{SC}(°)	Avg $\Delta Tw^{B\text{-}DNA}$(°)	Avg Tw^{SP}(°)	Avg $Shift^{SP}$(Å)	Avg $Slide^{SP}$(Å)	Avg $Rise^{SP}$(Å)	Avg $Tilt^{SP}$(°)	Avg $Roll^{SP}$(°)	Number of Matching Tetramers
ATC/GAT	32.24	-0.27	-2.03	32.52	-0.87	-0.61	3.07	2.35	0.78	1
AAAA/TTTT	32.17	-0.22	-2.10	32.40	-0.40	0.08	3.21	-2.38	-4.24	2
AAAC/GTTT	33.85	0.61	-0.43	33.24	0.33	0.19	3.01	1.39	8.81	1
AAAG/CTTT	35.55	0.05	1.27	35.50	0.01	-0.16	3.32	1.51	5.22	2
AAAT/ATTT	36.75	-0.26	2.46	37.02	-0.24	-0.29	3.92	4.30	0.78	1
AACA/TGTT	31.73	0.15	-2.55	31.58	-0.08	-0.40	3.41	-3.35	6.68	1
AACT/AGTT	28.67	-0.95	-5.61	29.63	0.24	-0.51	3.27	4.62	8.73	1
AAGG/CCTT	30.89	-1.34	-3.39	32.24	1.56	0.62	2.96	2.21	-12.38	1
AAGT/ACTT	33.98	-0.44	-0.29	34.43	-0.31	-0.14	3.80	-0.18	7.54	1
AATA/TATT	33.46	0.18	-0.82	33.28	0.38	-0.47	3.10	-0.10	-2.57	2
AATC/GATT	26.79	-2.06	-7.49	28.86	-0.13	-1.06	3.16	3.16	7.00	1
ACAC/GTGT	33.45	0.07	-0.83	33.38	-0.05	0.08	3.18	2.15	7.27	1
ACAT/ATGT	36.59	1.14	2.31	35.45	-0.47	0.57	3.42	-0.37	9.09	1
ACCA/TGGT	29.48	-0.19	-4.80	29.68	0.10	0.18	3.42	6.64	7.83	1
ACCT/AGGT	33.38	-1.14	-0.90	34.53	0.38	-0.59	3.37	1.26	8.87	1
ACTA/TAGT	34.42	-1.09	0.13	35.52	0.19	-0.48	3.57	-2.49	-2.25	1

Tetramer	Max Tw^{SC}(°)	Max ΔTw^{SC}(°)	Max $\Delta Tw^{B\text{-}DNA}$(°)	Max Tw^{SP}(°)	Max $Shift^{SP}$(Å)	Max $Slide^{SP}$(Å)	Max $Rise^{SP}$(Å)	Max $Tilt^{SP}$(°)	Max $Roll^{SP}$(°)
ATC/GAT	32.24	-0.27	-2.03	32.52	-0.87	-0.61	3.07	2.35	0.78
AAAA/TTTT	40.90	0.34	6.61	40.56	0.20	0.48	3.59	5.06	-1.56
AAAC/GTTT	33.85	0.61	-0.43	33.24	0.33	0.19	3.01	1.39	8.81
AAAG/CTTT	38.37	0.29	4.09	38.08	0.36	-0.10	3.40	3.05	7.81
AAAT/ATTT	36.75	-0.26	2.46	37.02	-0.24	-0.29	3.92	4.30	0.78
AACA/TGTT	31.73	0.15	-2.55	31.58	-0.08	-0.40	3.41	-3.35	6.68
AACT/AGTT	28.67	-0.95	-5.61	29.63	0.24	-0.51	3.27	4.62	8.73
AAGG/CCTT	30.89	-1.34	-3.39	32.24	1.56	0.62	2.96	2.21	-12.38
AAGT/ACTT	33.98	-0.44	-0.29	34.43	-0.31	-0.14	3.80	-0.18	7.54
AATA/TATT	34.00	0.50	-0.27	33.50	0.59	-0.40	3.25	2.93	-2.17
AATC/GATT	26.79	-2.06	-7.49	28.86	-0.13	-1.06	3.16	3.16	7.00
ACAC/GTGT	33.45	0.07	-0.83	33.38	-0.05	0.08	3.18	2.15	7.27
ACAT/ATGT	36.59	1.14	2.31	35.45	-0.47	0.57	3.42	-0.37	9.09
ACCA/TGGT	29.48	-0.19	-4.80	29.68	0.10	0.18	3.42	6.64	7.83
ACCT/AGGT	33.38	-1.14	-0.90	34.53	0.38	-0.59	3.37	1.26	8.87
ACTA/TAGT	34.42	-1.09	0.13	35.52	0.19	-0.48	3.57	-2.49	-2.25

Tetramer	Min Tw^{SC}(°)	Min ΔTw^{SC}(°)	Min $\Delta Tw^{B\text{-}DNA}$(°)	Min Tw^{SP}(°)	Min $Shift^{SP}$(Å)	Min $Slide^{SP}$(Å)	Min $Rise^{SP}$(Å)	Min $Tilt^{SP}$(°)	Min $Roll^{SP}$(°)
ATC/GAT	32.24	-0.27	-2.03	32.52	-0.87	-0.61	3.07	2.35	0.78
AAAA/TTTT	23.45	-0.78	-10.83	24.24	-1.00	-0.31	2.83	-9.83	-6.93
AAAC/GTTT	33.85	0.61	-0.43	33.24	0.33	0.19	3.01	1.39	8.81
AAAG/CTTT	32.73	-0.18	-1.55	32.92	-0.34	-0.23	3.25	-0.02	2.64
AAAT/ATTT	36.75	-0.26	2.46	37.02	-0.24	-0.29	3.92	4.30	0.78
AACA/TGTT	31.73	0.15	-2.55	31.58	-0.08	-0.40	3.41	-3.35	6.68
AACT/AGTT	28.67	-0.95	-5.61	29.63	0.24	-0.51	3.27	4.62	8.73
AAGG/CCTT	30.89	-1.34	-3.39	32.24	1.56	0.62	2.96	2.21	-12.38
AAGT/ACTT	33.98	-0.44	-0.29	34.43	-0.31	-0.14	3.80	-0.18	7.54
AATA/TATT	32.92	-0.13	-1.36	33.06	0.17	-0.54	2.96	-3.13	-2.98
AATC/GATT	26.79	-2.06	-7.49	28.86	-0.13	-1.06	3.16	3.16	7.00
ACAC/GTGT	33.45	0.07	-0.83	33.38	-0.05	0.08	3.18	2.15	7.27
ACAT/ATGT	36.59	1.14	2.31	35.45	-0.47	0.57	3.42	-0.37	9.09
ACCA/TGGT	29.48	-0.19	-4.80	29.68	0.10	0.18	3.42	6.64	7.83
ACCT/AGGT	33.38	-1.14	-0.90	34.53	0.38	-0.59	3.37	1.26	8.87

Figure 3.4.4.12: An example of the Tetrameric step statistics generated by TwiDDL. In this case it is for a single structure, but using TwiDDL's toolbar on the search results page the user can generate statistics about multiple structures.

The fifth section of the TwID details page contains the file downloads provided to the user. The files are organized as part of a drop down list broken into three categories where one is for the TwiDDL Files, one for the Structure Input File, and one for the 3DNA Output Files as denoted respectively in Figure 3.4.4.13 by (1), (2), and (3). The drop-down list allows the user to download files with all the raw data used to populate TwiDDL for characterizing the selected structure.

The TwiDDL Files drop-down list links to the twistwrith.csv file of the structure, which contains the values of Tw^{SC} and several closely related conformational variables. This file can be opened and edited with most spreadsheet programs. The file includes values of Tw^{SC} for both the Open linear structure stored in the Protein Data Bank and a hypothetical Closed structure, which includes an added step that connects the first and last base pairs. The latter model is included so that the user can see what the linking and writhing numbers (Lk and Wr) would be if the structure were closed. The characterizations of the two structural forms include data for each base-pair step as well as the sums of the various twists and the number of helical turns at the top of the twist columns. The data labeled TwistOAve are the Tw^{SC} values listed on the website. The data labeled TwistCAve are the corresponding values for the hypothetical Closed structure. The Structure Input File drop down contains a link to the input PDB file used to calculate Tw^{SC} and the related conformational parameters stored in TwiDDL. The drop down for

3DNA Output Files contains links to the complete set of output files generated upon 3DNA [8] analysis of the selected structure.

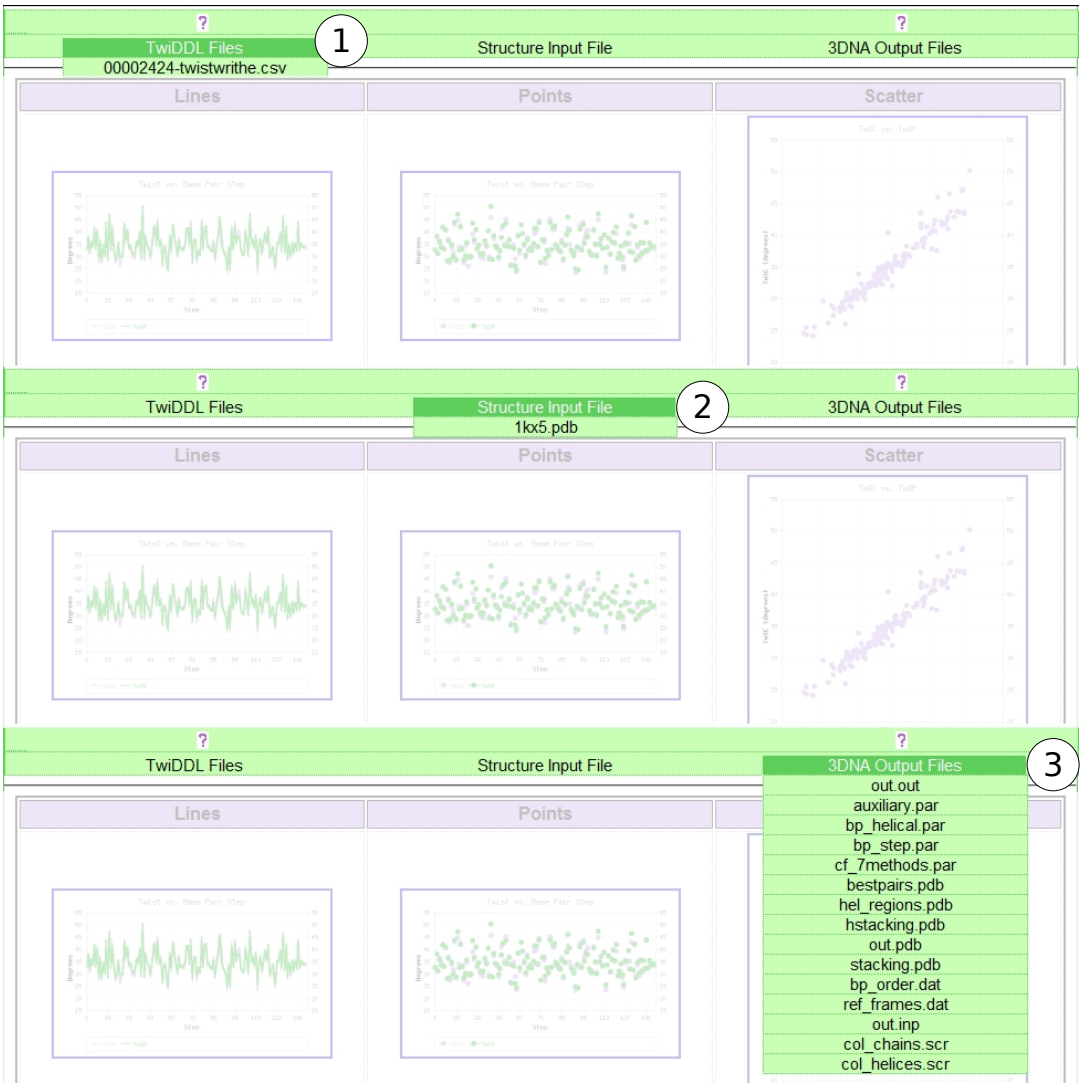
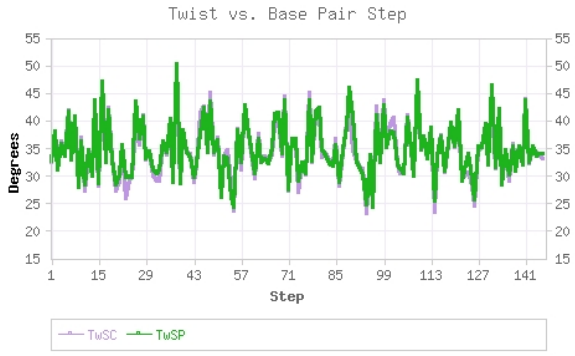


Figure 3.4.4.13: A screen capture of the file downloads menus.

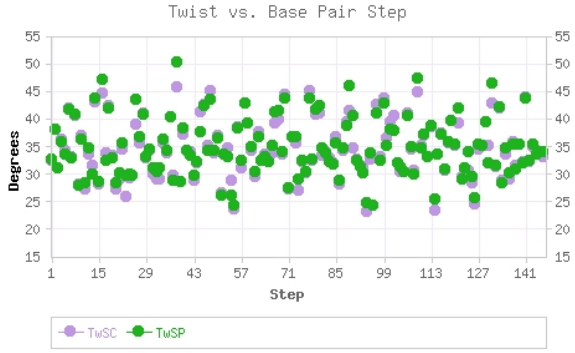
The sixth section of the TwID details page consists of three simple two-dimensional graphs that depict the variation of Tw^{SC} and Tw^{SP} along the chain and the sequential differences between the two values. The first is a line plot of the Twist vs. Base-pair Step, as seen in Figure 3.4.4.14(a). The color-coded line graph shows how the

two twists, Tw^{SC} in purple and Tw^{SP} in green, vary along the length of the molecule. The second is similar to the first where the primary difference is that it is instead a point plot of the Twist vs. Base-pair Step, as seen in Figure 3.4.4.14(b). To be consistent with the line plot the color-coded points show the Tw^{SC} in purple and Tw^{SP} in green, and again represent how the twists vary along the length of the molecule. The third plot, seen in Figure 3.4.4.14(c), is a scatter plot of Tw^{SC} versus Tw^{SP}. The scatter plot clarifies the differences in the two twist values at selected steps in the structure. If the twists are identical, the points will lie along the diagonal of the plot. The deviations from this line pinpoint the steps where Tw^{SC} and Tw^{SP} differ. If the molecule adopts a relaxed B-DNA structure, all the points will be located at a single point where $Tw^{SC} = Tw^{SP} = 34.3°$.

(a)



(b)



(c)

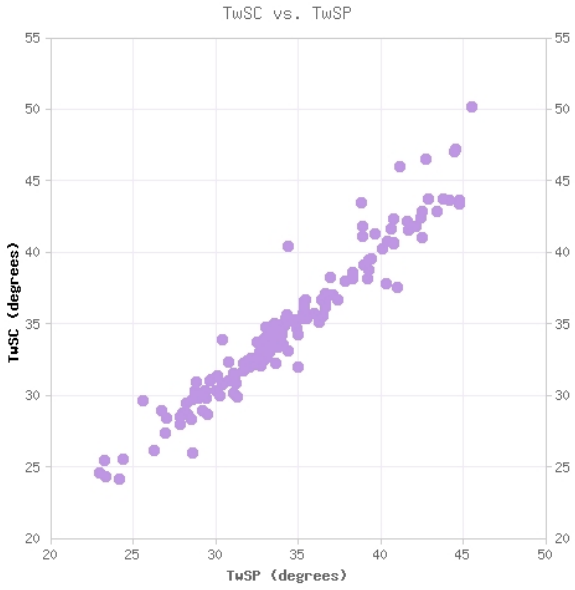


Figure 3.4.4.14: An example of the three two dimensional graphs generated for the TwID details page when clicking the images at the bottom of the page.

3.4.5 Java Servlets

As discussed in the previous section, one of the more powerful of the visualization methods used in TwiDDL is a color-coded 3D model that displays the difference in the twist of supercoiling from the step parameter twist ΔTw^{SP} or that of B-DNA $\Delta Tw^{B\text{-}DNA}$. This 3D model is generated with several different pieces of software, but the primary method for serving it through the web interface is the use of a Java Servlet implementation called Apache Tomcat. Apache Tomcat provides an extension to the typical services provided by web servers, such as the Apache HTTP Server, that allow it to directly serve Java applications known as servlets over the web. This is similar to how the Apache HTTP Server supplies HTML based web pages over the web, but instead it is a much richer software environment designed for distributing more complex Java based applications over the web.

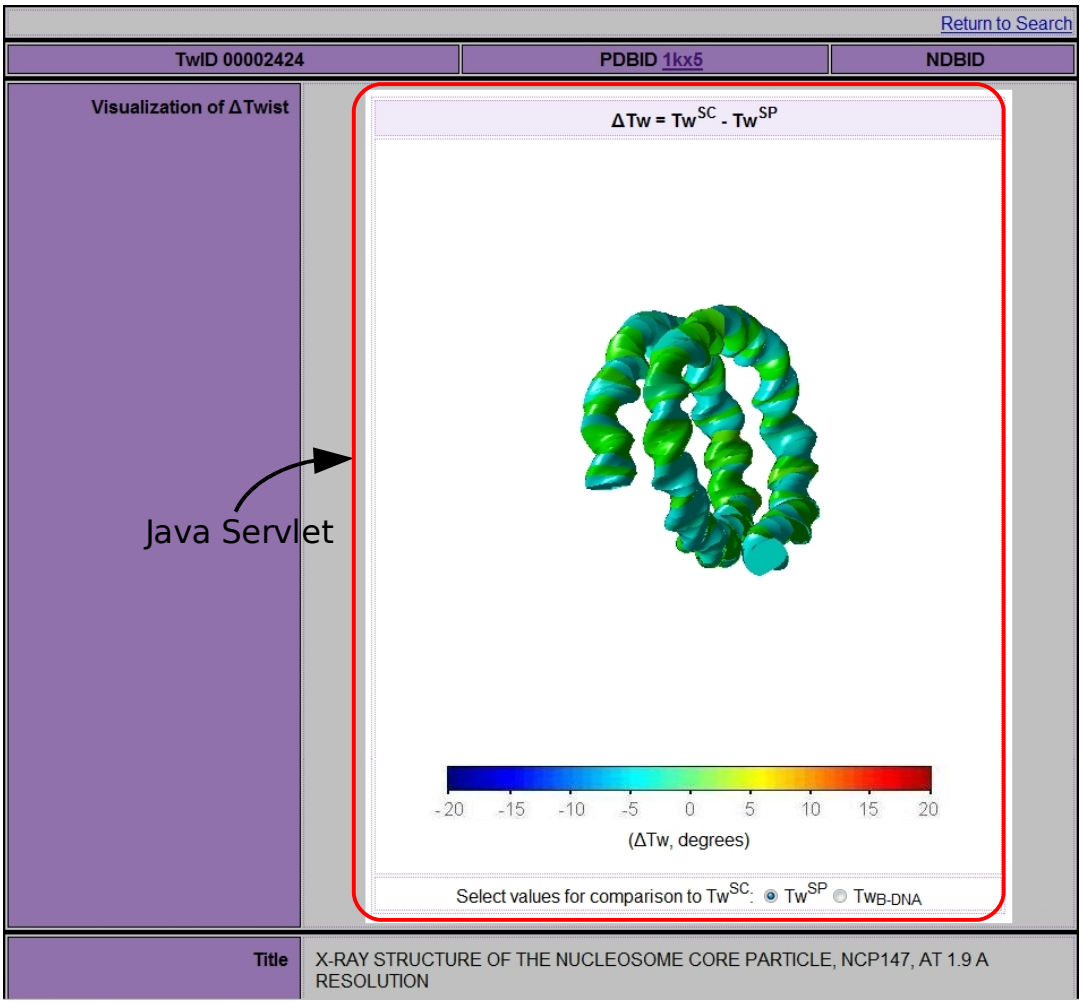


Figure 3.4.5.1: An example of how the Java Servlet is used to display a 3D visualization within the TwID details page.

Within TwiDDL the use of Java servlets is limited to the 3D visualization on the detailed TwID page as seen in Figure 3.4.5.1. In order to display the images, AJAX, or Asynchronous JavaScript and XML, is used to create a frame and dynamically load the Java servlet called TwiDDLPlots. The code for TwiDDLPlots is partly written in Java which calls the primary 3D modeling functions that were written in Matlab. This combination is integrated as a single Java Servlet and displayed to the end user through

the Apache Tomcat server. To display the data for the selected structure properly, two inputs must be supplied to the underlying Matlab plotting application. The first input is a pointer to the directory that contains the data used to generate the model, which is the same as the directories created to store all of the raw data for each of the structures within TwiDDL's database. As previously discussed, this directory contains a thetarhotwsc.txt file, which is the primary source of the data used by the 3D plotting servlet. The second input determines whether the plot will display the color-coded model based on the ΔTw^{SP} or $\Delta Tw^{B\text{-}DNA}$ values.

3.5 Database Maintenance

The TwiDDL database was designed with the intention for the database to grow as new structures become available. In order to grow like this the design required a few considerations that made it easy to grow and be maintained. The first thing that had to be done was the full automation of the twist of supercoiling calculations. This was accomplished by building modular Perl code that executed all of the steps required to generate the raw data discussed in Section 3.1.2. The next thing was to build scripts for parsing the raw data and inserting it into the database, as discussed in Sections 3.3 and 3.4. These few basic requirements set the foundation for long term automated maintenance of the TwiDDL database. These features also enabled a very simplistic administrative interface to be developed, as shown in Figure 3.5.1, that allows structures

to be added to or removed from the database by inputting either their PDB ID or NDB ID. The interface also allows for existing PDB ID or NDB ID structures in TwiDDL to be overwritten, which is valuable when bugs in TwiDDL are found and the data for structure in the database must be replaced. This also ensures that once a structure is in the database, that it keeps the same TwID when the data are updated, rather than a new entry with a new TwID being created. By default the TwiDDL Administration interface will not overwrite existing entries in the database.

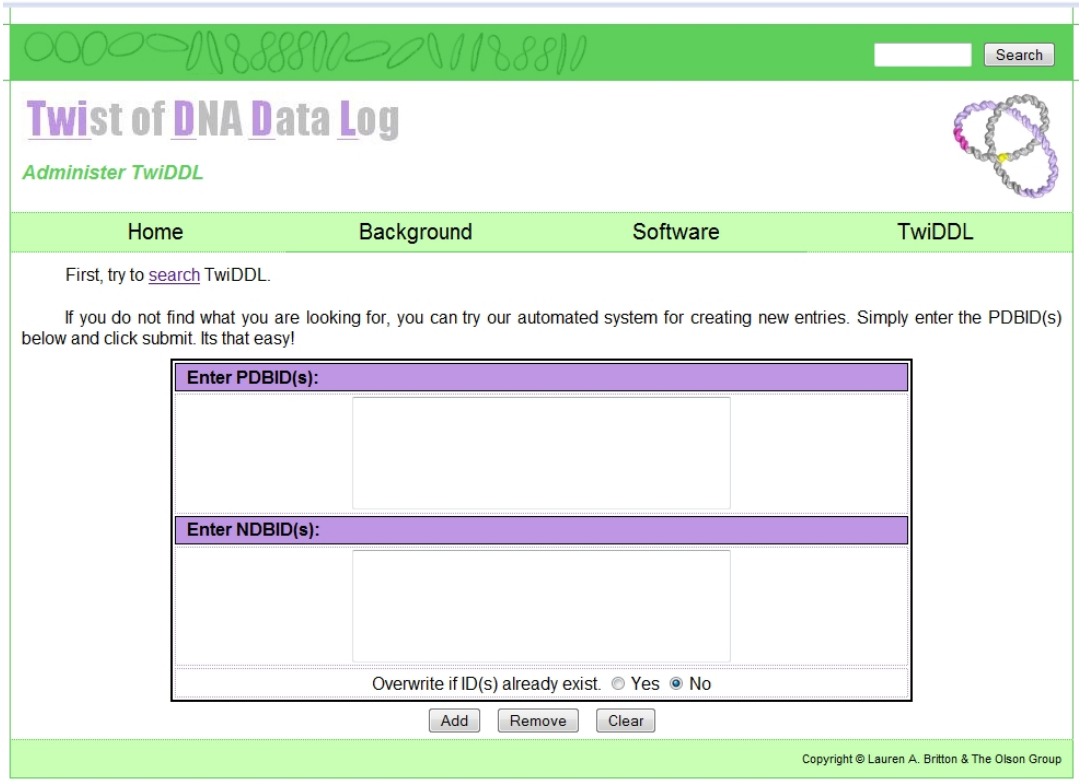


Figure 3.5.1: A screen capture of the administration interface for TwiDDL. Used to enter, overwrite, and remove structures from the database.

3.5.1 Automated data retrieval and update

As discussed in Section 3.4.1 the same code that was developed for use through the web interface can be used by command line scripts through Perl. One of the goals for TwiDDL was to update the database automatically as new structures are added to the PDB. In order to achieve this goal a Perl script was written that utilizes the same features described in the TwiDDL Administration interface, but operated at the Linux command line. This script, called updatedb-cron.pl, was intended for use by the Linux cron feature. The cron feature in Linux is intended for running commands on a schedule. The use of cron to run the updatedb-cron.pl script allows for a fully automated update of the data that are stored in TwiDDL.

In order to track the regular update of the database, each time the updatedb-cron.pl script is run to add data into TwiDDL the output from the run is stored in a directory. Each directory is named after the date and time that the script was run. The script is currently set to run once a week by cron, and the recurring update of the database was started on February 7, 2010 and continues today. The output from the script contains a brief status of how far it was able to proceed through the automated steps used for Getting, Parsing, and Running the PDB file through TwiDDL. At each of these three phases the script will either display the reason why it failed to move on to the next phase, or indicate completion of all three steps by providing a link that says succeeded and

points to the entry in TwiDDL for the newly entered structure. The logs from each of the automated updates to the database are stored at http://twiddl.rutgers.edu/cron/.

3.5.2 Data review & release

As discussed in Section 3.3.2 every structure stored in the database is automatically checked for compliance with some basic requirements that the twist of supercoiling has for the data being used to make calculations based on that structure. When the structure is found to have issues, a note about the issue is inserted into the Comments column of the Summary table, and any searches of the database will only return results that contain an empty or NULL Comments field. This can be seen in the SQL statement shown in FIGURE SQL1b as part of the WHERE clause which says "AND (Summary.Comments LIKE 'NULL')".

This combination of features in the database allows for the raw data and even the database entries to exist within TwiDDL but remain hidden from the end users until someone with enough understanding of TwiDDL can review the entry. This is a powerful combination because not only do we automatically update the database with new structures, but we ensure that only approved structures are posted for the biologists and other researches to access through TwiDDL. The other advantage is that all of the data for the twist of supercoiling have been calculated, so releasing the structure for public consumption can be as simple as a quick review of the data. In order to allow the

structure to show up in searches through TwiDDL the Comments field in the Summary table simply needs to be cleared or set to NULL. By design all of the SQL queries in TwiDDL will then automatically pick up the structure and include it in the search results with the other entries in the database with a NULL Comments field.

3.5.3 Debug

Due to the complexity of TwiDDL and all of the features that it has enabled, it had to be designed for being debugged in a flexible way. In Section 3.4.1, the various functions and subroutines that make up TwiDDL were discussed. In order to debug each of these functions or subroutines a flag has been used to turn on or off some verbose messages in the code that enable greater visibility into what the Perl source code is doing. For example, the SQL statement used in FIGURE SQL1b was obtained by setting the DEBUG=1 flag in the html_search_results subroutine. By setting that flag, the HTML generated by the search page would contain details that were otherwise not needed by the typical user, like hidden fields that contain the detailed SQL used to search the MySQL database and retrieve the data being displayed. This technique of setting a DEBUG flag in the Perl subroutines can be used for every subroutine, and even the main twdl.cgi calling script in order to debug the routines one at a time or even multiple routines at once. This is helpful to a developer trying to enhance or debug what the code is doing, but too much information like this, that is not valuable to the end user, can confuse the

value of the real data being displayed. By default all of the flags are set to DEBUG=0 which prevents this verbose debug information from being displayed until it is needed. In order to turn on the flags the source code needs to be manually edited and set to DEBUG=1.

3.6 References

[1] L.A. Britton, I. Tobias, and W.K. Olson. (2009). Two Perspectives on the Twist of DNA. *J. Chem. Phys.,* **131(24)**, 245101.

[2] R.E. Dickerson, M. Bansal, C.R. Calladine, S. Diekmann, W.N. Hunter, O. Kennard, E. von Kitzing, R. Lavery, H.C.M. Nelson, W.K. Olson, W. Saenger, Z. Shakked, H. Sklenar, D.M. Soumpasis, C.S. Tung, A.H.J. Wang, and V.B. Zhurkin. (1989). Definitions and Nomenclature of Nucleic Acid Structure Parameters. *J. Mol. Biol.,* **205(4)**, 787-791.

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. (2000). The Protein Data Bank. *Nucleic Acids Res.,* **28(1)**, 235-242.

[4] H.M. Berman , W.K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.H. Hsieh, A.R. Srinivasan, and B. Schneider. (1992). The Nucleic Acid Database. A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys J.,* **63(3),** 751-759.

[5] G. Călugăreanu. (1961). Sur les classes d'isotopie des noeuds tridimensionnels et leurs invariants. *Czech. Math. J.,* **11(4)**, 588-625.

[6] J.H. White. (1969). Self-Linking and the Gauss Integral in Higher Dimensions. *American J. Math.*, **91(3)**, 693-728.

[7] F.B. Fuller. (1971). The Writhing Number of a Space Curve. *Proc. Natl. Acad. Sci. USA,* **68(4)**, 815-819.

[8] X. Lu, and W. K. Olson. (2003). 3 DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.,* **31(17)**, 5108-5121.

[9] S. Smirnov, T.J. Matray, E.T. Kool, and C. de los Santos. (2002) Integrity of Duplex Structures without Hydrogen Bonding: DNA with Pyrene Paired at Abasic Sites. *Nucleic Acids Res.,* **30(24)**, 5561-5569.

[10] M.H. Kolk, M. van der Graaf, C.T. Fransen, S.S. Wijmenga, C.W. Pleij, H.A. Heus, and C.W. Hilbers. (1998) Structure of the 3'-Hairpin of the TYMV Pseudoknot: Preformation in RNA Folding. *EMBO J.*, **17(24)**, 7498-7504.

[11] A. Dey, D. York, A. Smalls-Mantey, and M.F. Summers. (2005) Composition and Sequence-dependent Binding of RNA to the Nucleocapsid Protein of Moloney Murine Leukemia Virus. *Biochem.,* **44(10)**, 3735-3744.

[12] H.M. Berman, et al. (2010). Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description Version 3.2. Available at *http://www.wwpdb.org/documentation/format32/v3.2.html*.

[13] M. Elrod-Erickson, M.A. Rould, L. Nekludova, and C.O. Pabo. (1996) Zif268 Protein-DNA Complex Refined at 1.6 A: A Model System for Understanding Zinc Finger-DNA Interactions. *Structure,* **4(10)**, 1171-1180.

[14] C. Tisne, B. Hartmann, and M. Delepierre. (1999) NF-kappa B Binding Mechanism: A Nuclear Magnetic Resonance and Modeling Study of a GGG --> CTC Mutation. *Biochem.,* **38(13)**, 3883-3894.

[15] X. Lu, and W.K. Olson. (2008). x3dna | Examples. Available at *http://3dna.rutgers.edu/x3dna/examples*.

[16] W.K. Olson. (2007). Olson Group: Databases. Available at *http://dnaserver.rutgers.edu/database.php*.

[17] Oracle Corporation and/or its affiliates. (2010) MySQL :: MySQL Customers. Available at *http://www.mysql.com/customers/*.

[18] Oracle Corporation and/or its affiliates. (2010) MySQL :: MySQL 5.1 Reference Manual :: 1.8 MySQL Standards Compliance. Available at *http://dev.mysql.com/doc/refman/5.1/en/compatibility.html*.

[19] Oracle Corporation and/or its affiliates. (2010) MySQL :: MySQL Documentation: MySQL Reference Manuals. Available at *http://dev.mysql.com/doc/*.

[20] Y. Boulard, G.V. Fazakerley, and L.C. Sowers. (2002) The Solution Structure of an Oligonucleotide Duplex Containing a 2'-Deoxyadenosine-3-(2-Hydroxyethyl)- 2'-Deoxyuridine Base Pair Determined by NMR and Molecular Dynamics Studies. *Nucleic Acids Res.*, **30(6)**, 1371-1378.

[21] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 A Resolution. *J. Mol. Biol.,* **319(5)**, 1097-1113.

[22] Dictionary.com. Dictionary.com Unabridged. Random House, Inc. (2011) parse. Available at *http://dictionary.reference.com/browse/parse*.

Chapter 4: TwiDDL in Use

4.1 Benefits of TwiDDL

DNA assumes a variety of different shapes during its biological processing. Knowing the topology of a segment of DNA gives us a better understanding of the structural changes that the segment can undergo during that biological processing. Proteins often alter the shape of the DNA segment to which they are bound, sometimes dramatically from the canonical double-helical structure. The ease of untwisting successive base pairs contributes to the "melting" that accompanies the biological processing of DNA. The process of strand separation during transcription and replication involves repetitive breakage of base pairs and formation of temporary single strands of DNA. This separation is accomplished by using proteins/enzymes binding to DNA. During the aforementioned biological processes, the double-helical DNA changes drastically under the influences of proximate proteins, drugs, and other factors in the environment [1].

Sharp kinks, induced by binding proteins and coupled to changes in shearing at the local level, are captured by changes in the twist of supercoiling, Tw^{SC}. Overall Tw^{SC} is coupled to global folding of DNA. The folding, in turn, is measured by the writhing number. Local deformations in structure that affect global folding can contribute to

action-at-a-distance characteristics of DNA.

Until this work structural biologists have characterized the twist of DNA base-pair steps solely in terms of the rigid-body step parameter, called the twist. This quantity, referred to here as Tw^{SP}, is very useful in the analysis and exact rebuilding of DNA structures. While Tw^{SP} is appropriate to use for this purpose, and for the purpose of estimating the energy costs of DNA deformation, it is not appropriate to use for finding the linking number of DNA, based on the well known equation relating the linking number Lk to the total twist and writhe. For this equation to be valid, one must use the Tw^{SC} described by Britton et al. [22]. While the original reason for deriving Tw^{SC} was for use in the determination of the linking number, we have found that the quantity yields new insight into the local twisting of DNA steps, as well as the net twisting of the molecule as a whole. The Tw^{SC} is an indirect measure, through writhe, of the global folding of the double-helical axis. The amount of twist and writhe in a section of DNA reflects how a protein can bind, a drug can intercalate, or the double helix retains the canonical B-DNA form.

Relating the global topology of protein-decorated DNA to the content and location of bound proteins requires a new quantity. The propensity of DNA to undergo chiral distortions determines where certain proteins can bind, and to what extent they work on DNA. Since the writhe measures the chiral distortion and knowing that the linking

number cannot change for a closed piece of DNA without being nicked, it follows that the Tw^{SC} must also be sensitive to chiral distortions. Chirality occurs when bending and shearing happen concurrently, and the Tw^{SC} captures this type of chirality. In this chapter we examine what effect various proteins bound to DNA have on both the chiral distortion and the amount of over/under twisting in comparison to protein-free B-DNA. The Tw^{SC} measures the degree to which segments of DNA are over- or undertwisted compared to relaxed B-DNA.

This chapter focuses on the twist of supercoiling in representative protein-bound DNA structures, including how Tw^{SC} captures the chiral distortions of base pairs in these complexes. This information is potentially useful to biologists interested in how much a protein will change the topology of the DNA. The TwiDDL database, described in Chapter 3, provides an easy way for biologists to collect the topological information of thousands of known structures.

4.2 Searching for over- and undertwisted protein/DNA complexes

These next three sections present examples of three proteins, two that undertwist DNA and one that overtwists DNA. These examples are taken from three high resolution X-ray crystals stored in the Protein Data Bank. The structures stand out in a search of TwiDDL as among the most topologically interesting ones. In this case, TwiDDL's advanced search page was used to find structures with a magnitude of $\Delta Tw^{B\text{-}DNA} = Tw^{SC} -$

$Tw^{B\text{-}DNA}$ between 10 and 100 degrees. A subset of the results returned from that search can be seen in Table 4.2.1.

Many of the returned structures from the TwiDDL search have been discounted for use in this study because they do not comply with our purposes. The structures that did not make the cut, and therefore are not included in TwiDDL at all, are omitted due to having:

1. RNA in them. These are omitted because we do not have the tools to define the topology of RNA in a consistent manner currently. This will probably change in the future.
2. Mismatched bases. Unused arrangements could lead to extraordinary values of twist.
3. Melted DNA. We do not define the topology of structures containing melted, i.e., single-stranded, DNA currently in a way that would allow us to define the twist. We do plan to incorporate this in future.
4. Very short DNA sequences. Such structures, i.e., a single base-pair step, are too short to have any real significance. Tw^{SC} relies on four base pairs. One cannot compute meaningful Tw^{SC} values for two or three base pairs.

Update Search	TwID v	NBP ^	$\Delta Tw^{B\text{-}DNA}$ ^	Sequence ^	Title ^	PDBID ^	NDBID ^	Experiment Method ^	Step Details
☐	00000001	10	-14.18	CCGGCGCCGG	CRYSTAL STRUCTURE OF THE HIGHLY DISTORTED CHIMERIC DECAMER R(C)D(CGGCGCCG)R(G)-SPERMINE COMPLEX-SPERMINE BINDING TO PHOSPHATE ONLY AND MINOR GROOVE TERTIARY BASE-PAIRING	100d		XRayCrystal	Show
☐	00000007	8	-10.55	CGCTAGCG	THE SOLUTION STRUCTURE OF A DNA COMPLEX WITH THE FLUORESCENT BIS INTERCALATOR TOTO DETERMINED BY NMR SPECTROSCOPY	108d-1		NMR	Show
☐	00000008	8	-17.73	CGCTAGCG	THE SOLUTION STRUCTURE OF A DNA COMPLEX WITH THE FLUORESCENT BIS INTERCALATOR TOTO DETERMINED BY NMR SPECTROSCOPY	108d-2		NMR	Show
☐	00000009	8	-14.78	CGCTAGCG	THE SOLUTION STRUCTURE OF A DNA COMPLEX WITH THE FLUORESCENT BIS INTERCALATOR TOTO DETERMINED BY NMR SPECTROSCOPY	108d-3		NMR	Show
☐	00000010	8	-12.56	CGCTAGCG	THE SOLUTION STRUCTURE OF A DNA COMPLEX WITH THE FLUORESCENT BIS INTERCALATOR TOTO DETERMINED BY NMR	108d-4		NMR	Show

Table 4.2.1: Sample results returned from TwiDDL's advanced search page consisting of the first five entries of over 3500 structures found in TwiDDL with a magnitude of $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$ between 10 and 100 degrees. The full table is too large to reasonably display, and therefore only a small sample of five is shown here.

4.2.1 Undertwisted Example 1 - Human TFIIA/TBP/DNA Complex

The 17 base-pair piece of DNA bound to the TATA-box binding protein and the transcription initiation factor IIA [2], PDB entry 1NVP, is highly undertwisted compared to B-DNA. The TATA-box binding protein is known to bend DNA very severely. The large deformation facilitates the positioning/opening of DNA for transcription factors to bind and recognize the sequence. The whole complex, an X-ray crystal structure resolved at 2.10Å resolution and published by Bleichenbacher et al [2], is depicted in Figure 4.2.1.1. The rectangular slabs are the nucleic acid bases and the purple ribbons, lines, and arrows are the proteins. In Figure 4.2.1.1 the TATA-box binding protein is located on the upper right hand side, and the alpha, beta, and gamma chains of the transcription initiation factor IIA are located off of the lower left-hand side. The DNA double helix has tube-like backbones and rectangular bases in either green (G), yellow (C), blue (T), or red (A).

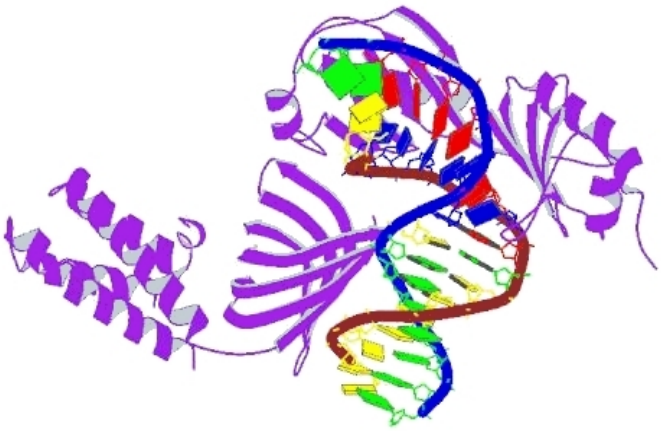


Figure 4.2.1.1: Representation of the DNA bound to the TATA-box binding protein and the alpha, beta, and gamma chains of transcription initiation factor II [2], PDB ID 1NVP. This figure is made with 3DNA [4] and taken from the PDB. The TATA-box binding protein and the alpha, beta, and gamma chains of transcription initiation factor II are represented by the ribbon like curves, lines, and arrows in purple. The double helical DNA backbones are the thicker curves in red and blue. The base pairs consist of adenine in red, thymine in blue, guanine in green, and cytosine in yellow. This image is from the PDB [3].

Figure 4.2.1.2 shows a color-coded tube representation of the DNA base pairs in the same complex taken from TwiDDL The color coding shows how the twist changes along the sequence. The twist is expressed in terms of the deviation, $\Delta Tw^{B\text{-}DNA}$, of Tw^{SC} with respect to the twist of supercoiling of B-DNA, $Tw^{B\text{-}DNA}$. The blue/green colors make it clear that the DNA is definitely undertwisted compared to relaxed B-DNA. $\Delta Tw^{B\text{-}DNA}$ for the overall molecule is –76.54° and the total number of turns based on Tw^{SC} is 1.314. This means that on average for each step Tw^{SC} would be 29.49° while $Tw^{B\text{-}DNA}$ value would be 34.28° (=360°/10.5), a difference of –4.78°. As is clear from the image, the total twist of the molecule does not tell the whole story of what is going on in smaller sections of the structure.

Tables 4.2.1.1 and 4.2.1.2 show the values of Tw^{SC} and the difference in twist, $\Delta Tw^{B\text{-}DNA}$, of Tw^{SC} compared to relaxed B-DNA. Table 4.2.1.2 is sorted by $\Delta Tw^{B\text{-}DNA}$, the difference in twist compared to relaxed B-DNA. It is interesting to note how the largest values of $\Delta Tw^{B\text{-}DNA}$ occur at steps with large values of kinking and shearing (compare Tables 4.2.1.1 and 4.2.1.3). Note that the regions of greatest undertwisting have appreciably different values of Tw^{SP} and Tw^{SC} (see steps 9 through 14 in Figure 4.2.1.3, which is generated from the data in Tables 4.2.1.1 and 4.2.1.3). Use of the incorrect parameter (Tw^{SP}) would lead to errors in the effect of protein binding on overall DNA twist.

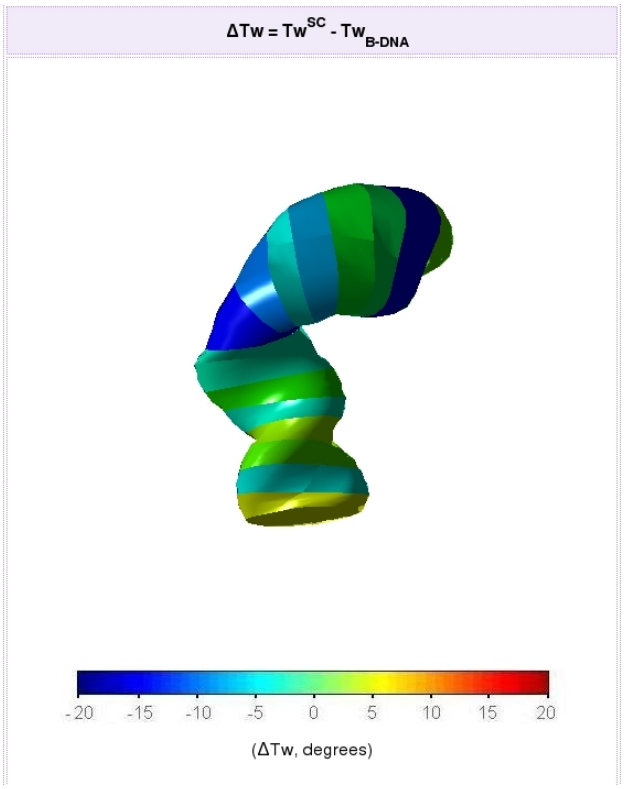


Figure 4.2.1.2: Representation of the human TFIIA/TBP/DNA complex taken from TwiDDL showing the deviation $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$ *with color coded base-pair steps.*

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°)
1	43.30	9.02
2	33.64	-0.64
3	37.65	3.36
4	40.52	6.24
5	32.34	-1.94
6	35.27	0.99
7	31.60	-2.67
8	32.10	-2.18
9	12.03	-22.25
10	19.30	-14.98
11	33.62	-0.66
12	16.97	-17.31
13	25.62	-8.65
14	27.75	-6.52
15	12.37	-21.91
16	37.84	3.56

Table 4.2.1.1: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the human TFIIA/TBP/DNA complex taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps. The data are sorted by the number of the base-pair step.

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°) ▼
9	12.03	-22.25
15	12.37	-21.91
12	16.97	-17.31
10	19.30	-14.98
13	25.62	-8.65
14	27.75	-6.52
7	31.60	-2.67
8	32.10	-2.18
5	32.34	-1.94
11	33.62	-0.66
2	33.64	-0.64
6	35.27	0.99
3	37.65	3.36
16	37.84	3.56
4	40.52	6.24
1	43.30	9.02

Table 4.2.1.2: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the human TFIIA/TBP/DNA complex taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps. The data are sorted numerically by the value of $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$.

Base Pair Step	Shift^{SP}(Å)	Slide^{SP}(Å)	Rise^{SP}(Å)	Tilt^{SP}(°)	Roll^{SP}(°)	Tw^{SP}(°)	Kinking^{SP}(°)	Shearing^{SP}(Å)
1	0.56	1.45	3.39	2.93	-5.99	47.86	6.66	1.55
2	-0.68	1.68	3.15	-1.68	6.66	35.36	6.86	1.81
3	0.25	1.03	3.42	-0.54	-1.02	37.72	1.15	1.05
4	0.28	0.26	3.44	-1.49	2.29	40.53	2.73	0.38
5	0.36	-0.05	3.14	-1.76	2.05	32.27	2.70	0.36
6	-0.01	-0.75	3.68	3.83	4.78	36.05	6.12	0.75
7	0.17	-0.77	3.40	-4.42	1.88	33.98	4.80	0.78
8	1.12	-0.23	2.96	7.55	3.77	32.96	8.43	1.14
9	0.16	-0.55	4.88	-0.20	51.69	16.05	51.69	0.57
10	-1.35	-0.29	3.36	0.02	23.03	14.98	23.03	1.38
11	-0.05	2.27	2.89	1.10	8.46	22.49	8.53	2.27
12	0.44	1.59	3.13	-0.68	21.89	9.50	21.90	1.64
13	-0.27	1.37	3.52	-0.18	22.83	18.82	22.83	1.39
14	0.39	0.74	3.26	-0.74	17.35	23.61	17.36	0.83
15	-1.51	0.59	4.52	-1.39	45.94	9.27	45.96	1.62
16	-1.01	0.30	3.40	-2.42	7.41	37.57	7.79	1.05

Table 4.2.1.3: Base-pair step parameters and the degree of kinking and shearing of each step in the human TFIIA/TBP/DNA complex taken from TwiDDL.

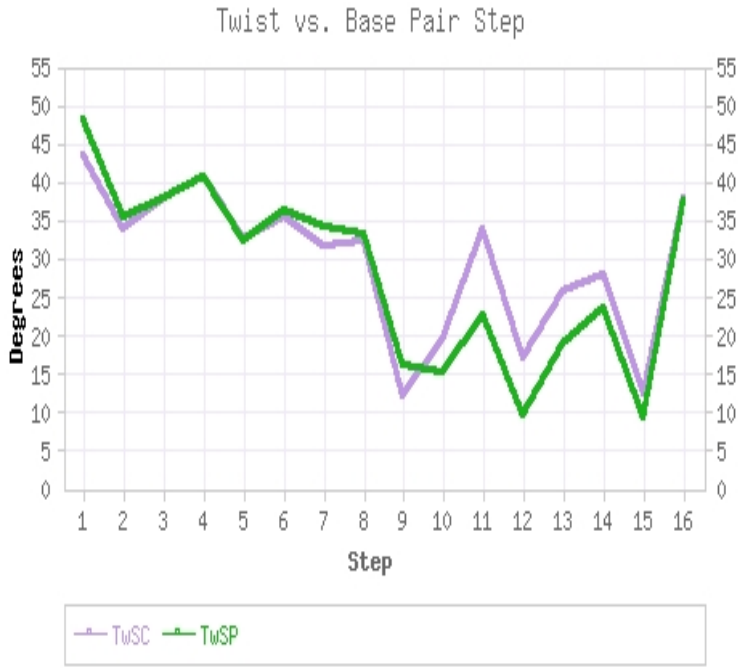


Figure 4.2.1.3: Sequential variations of Tw^{SC} (purple) and Tw^{SP} (green) in the human TFIIA/TBP/DNA complex graph taken from TwiDDL.

4.2.2 Undertwisted Example 2 - Crystal Structure of Lambda Repressor/DNA

Another structure with under twisted DNA is the complex of the bacteriophage lamda repressor and the 19 base-pair sequence depicted in Figure 4.2.2.1 PDB ID 3BDN [5]. The bacteriophage is a virus (DNA) that infects bacteria and uses a repressor switch (protein) to control gene expression.

As a whole the DNA has a Tw^{SC} average of around 34°. The DNA has both under- and overtwisted base-pair steps along the sequence, but the net undertwisting is slightly greater than the net overtwisting. In fact, close examination of Tables 4.2.2.1 and 4.2.2.2 shows that the under- and overtwisted base-pair steps occur right next to one another.

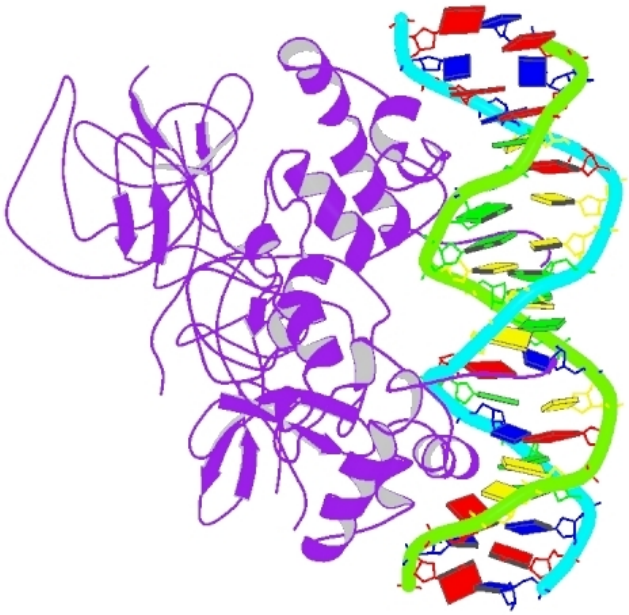


Figure 4.2.2.1: Representation of the DNA bound to the lambda repressor [5], PDB ID 3BDN. This figure is made with 3DNA [4] and taken from the PDB. The lambda repressor protein is represented by the ribbon like curves, lines, and arrows in purple. The double helical DNA backbone are the thicker curves in cyan and green. The base pairs consist of adenine in red, thymine in blue, guanine in green, and cytosine in yellow. This image is from the PDB [6].

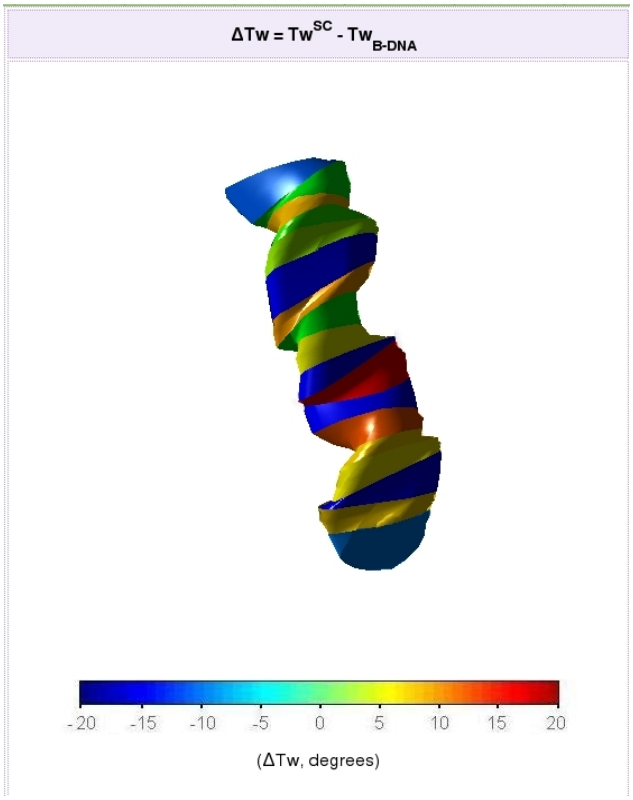


Figure 4.2.2.2: Representation of the crystal structure of the lambda repressor taken from TwiDDL showing the deviation $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$ *with color coded base-pair steps.*

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°)
1	19.05	-15.22
2	45.86	11.57
3	9.94	-24.34
4	44.49	10.20
5	45.91	11.63
6	50.78	16.50
7	11.03	-23.25
8	65.46	31.18
9	6.96	-27.31
10	43.98	9.69
11	24.94	-9.33
12	48.07	13.79
13	9.04	-25.23
14	40.57	6.28
15	35.88	1.59
16	46.77	12.48
17	25.12	-9.15
18	20.42	-13.85

Table 4.2.2.1: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the lambda repressor taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps. The data are sorted by the number of the base-pair step.

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°) ▼
9	6.96	-27.31
13	9.04	-25.23
3	9.94	-24.34
7	11.03	-23.25
1	19.05	-15.22
18	20.42	-13.85
11	24.94	-9.33
17	25.12	-9.15
15	35.88	1.59
14	40.57	6.28
10	43.98	9.69
4	44.49	10.20
2	45.86	11.57
5	45.91	11.63
16	46.77	12.48
12	48.07	13.79
6	50.78	16.50
8	65.46	31.18

Table 4.2.2.2: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the lambda repressor taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps. The data are sorted numerically by the value of $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$.

Base Pair Step	Shift^{SP}(Å)	Slide^{SP}(Å)	Rise^{SP}(Å)	Tilt^{SP}(°)	Roll^{SP}(°)	Tw^{SP}(°)	Kinking^{SP}(°)	Shearing^{SP}(Å)
1	-1.82	-1.91	3.65	23.07	-32.37	17.14	39.74	2.63
2	-0.72	0.31	3.74	7.21	1.98	47.42	7.47	0.78
3	-0.79	-1.96	2.81	-16.03	7.37	11.63	17.64	2.11
4	-0.81	-0.50	3.12	7.96	1.20	44.67	8.04	0.95
5	-0.47	1.13	2.09	2.20	-3.93	41.81	4.50	1.22
6	2.34	-0.83	3.95	3.36	-8.25	41.70	8.90	2.48
7	-2.03	-1.97	4.67	-6.93	23.75	6.62	24.74	2.82
8	-0.02	1.97	3.68	13.01	-4.44	66.01	13.74	1.97
9	0.46	1.86	1.92	-12.55	-3.79	17.99	13.10	1.91
10	-1.57	0.14	3.42	-6.43	7.20	48.06	9.65	1.57
11	-1.35	-0.51	3.36	-8.76	-12.02	26.65	14.87	1.44
12	2.25	-1.26	3.96	4.73	-6.88	42.11	8.34	2.57
13	-2.38	-2.10	3.99	5.95	13.37	9.28	14.63	3.17
14	0.61	-0.24	3.03	-0.64	7.01	42.14	7.03	0.65
15	0.05	-0.04	2.74	-4.34	-2.01	36.15	4.78	0.06
16	1.60	-1.02	3.77	0.42	-19.06	45.59	19.06	1.89
17	-0.22	0.14	2.99	9.56	27.07	27.16	28.70	0.26
18	0.42	-1.41	3.79	-12.34	-24.25	18.11	27.20	1.47

Table 4.2.2.3: Base-pair step parameters and the degree of kinking and shearing of each step. Based on the lambda repressor taken from TwiDDL.

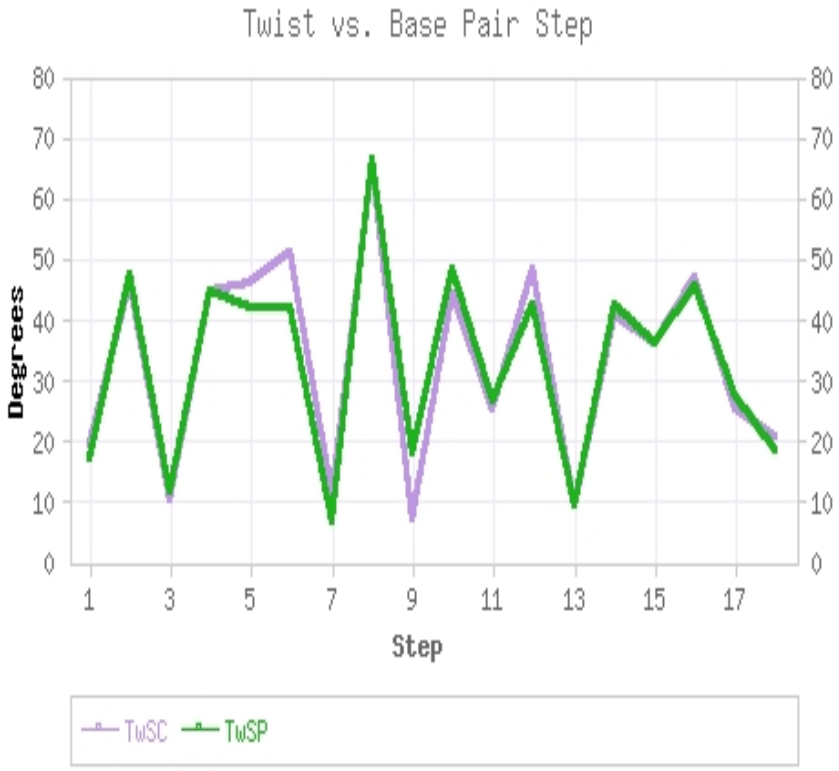


Figure 4.2.2.3: Sequential variations of Tw^{SC} (purple) and Tw^{SP} (green) in the lambda repressor graph taken from TwiDDL. The graph shows how in this structure Tw^{SC} has limited difference from Tw^{SP}.

4.2.3 Overtwisted Example - Crystal structure of Smad3-MH1/DNA

The structure chosen to show overtwisted DNA is the complex of the human Smad3-MH1 protein and the 15 base-pair sequence depicted in Figure 4.2.3.1 PDB ID 1OZJ [7]. Smad proteins bind to specific DNA sequences and regulate the expression of ligand-response genes.

As a whole the DNA has a Tw^{SC} average of around 37°. The DNA has both under- and overtwisted base-pair steps along the sequence, but the net overtwisting is slightly greater than the net undertwisting. As seen in Tables 4.2.3.1 and 4.2.3.2, the most significantly over twisted base-pair steps occur at steps 1, 7, and 14, which are clearly emphasized by the color coding of yellow, orange, and red, respectively, in Figure 4.2.3.2.

Figure 4.2.3.1: Representation of the DNA bound to the Smad3-MH1 protein [7], PDB ID 1OZJ. This figure is made with 3DNA [4] and taken from the PDB. The Smad3-MH1 protein are represented by the ribbon like curves, lines, and arrows in purple. The double helical DNA backbone are the thicker curves in cyan and green. The base pairs consist of adenine in red, thymine in blue, guanine in green, and cytosine in yellow. This image is from the PDB [8].

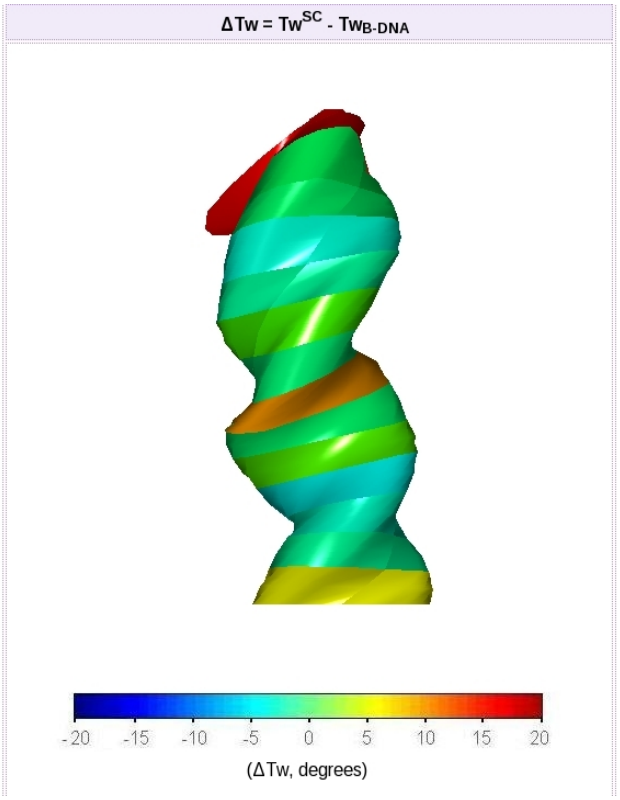


Figure 4.2.3.2: Representation of the human Smad3-MH1/DNA complex taken from TwiDDL showing the deviation $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$ *with color coded base-pair steps.*

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°)
1	44.02	9.73
2	30.31	-3.97
3	32.10	-2.18
4	33.79	-0.48
5	37.61	3.32
6	28.52	-5.76
7	48.20	13.92
8	29.03	-5.25
9	36.86	2.57
10	30.42	-3.85
11	32.63	-1.65
12	28.42	-5.85
13	27.74	-6.53
14	75.35	41.06

Table 4.2.3.1: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the human Smad3-MH1/DNA complex taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps. The data are sorted by the number of the base-pair step.

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°) ▾
13	27.74	-6.53
12	28.42	-5.85
6	28.52	-5.76
8	29.03	-5.25
2	30.31	-3.97
10	30.42	-3.85
3	32.10	-2.18
11	32.63	-1.65
4	33.79	-0.48
9	36.86	2.57
5	37.61	3.32
1	44.02	9.73
7	48.20	13.92
14	75.35	41.06

Table 4.2.3.2: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the human Smad3-MH1/DNA complex taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps. The data are sorted numerically by the value of $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$.

Base Pair Step	Shift^{SP}(Å)	Slide^{SP}(Å)	Rise^{SP}(Å)	Tilt^{SP}(°)	Roll^{SP}(°)	Tw^{SP}(°)	Kinking^{SP}(°)	Shearing^{SP}(Å)
1	-0.08	0.21	3.32	1.34	2.46	43.66	2.80	0.22
2	0.71	0.42	3.14	0.08	3.27	29.37	3.27	0.82
3	-0.30	-0.46	3.25	-1.72	7.22	32.58	7.42	0.54
4	-0.96	-0.74	3.55	-0.27	3.44	34.91	3.45	1.21
5	-0.14	-0.27	3.39	-1.54	2.80	37.91	3.19	0.30
6	-0.26	0.32	3.29	3.82	7.25	28.24	8.19	0.41
7	-0.07	2.73	3.02	0.05	-9.13	51.27	9.13	2.73
8	0.32	0.32	3.16	-3.15	6.46	28.64	7.18	0.45
9	0.03	-0.09	3.58	-0.79	5.97	37.11	6.02	0.09
10	0.80	-0.92	3.27	-0.38	3.39	31.88	3.41	1.21
11	0.21	-0.06	3.33	3.22	9.89	33.32	10.40	0.21
12	0.61	-0.96	3.17	-0.64	3.38	28.76	3.44	1.13
13	-0.52	-0.49	3.46	3.78	11.47	29.87	12.07	0.71
14	1.20	-1.86	3.16	9.24	9.74	77.63	13.42	2.21

Table 4.2.3.3: Base-pair step parameters and the degree of kinking and shearing of each step. Based on the human Smad3-MH1/DNA complex taken from TwiDDL.

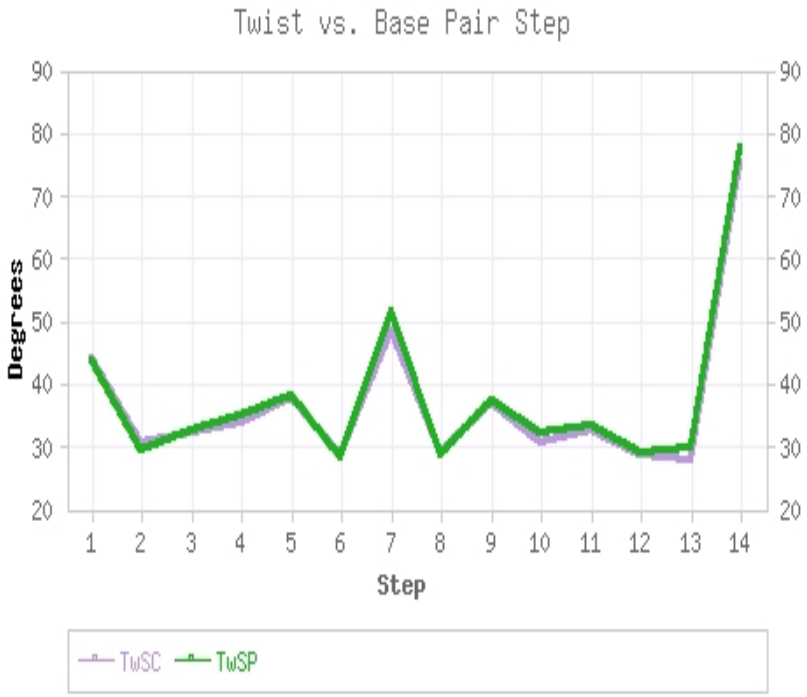


Figure 4.2.3.3: Sequential variations of Tw^{SC} (purple) and Tw^{SP} (green) in the human Smad3-MH1/DNA complex graph taken from TwiDDL The graph shows how in this structure Tw^{SC} has limited difference from Tw^{SP}.

4.3 Models for A, B and Z-DNA

The TwiDDL repository includes information about disparate types of DNA. The arrangements of successive base pairs in different DNA types occur in various DNA/protein interactions [9]. Here using TwiDDL we focus on the difference in Tw^{SC} in various DNA types compared to relaxed B-DNA.

We discuss three commonly known types of DNA: A-DNA, B-DNA, and Z-DNA. The step parameters for A-DNA contribute to the compression of the duplex as well as to the displacement and inclination of base pairs with respect to the helical axis. Here we discuss the TwiDDL data for a representative A-DNA model, the crystal structure of the octamer d(G-G-T-A-T-A-C-C) (PDB ID 1VJ4) [10], which is referred to as A-DNA in Table 4.3.1. The step parameters for a representative B-DNA helix produce a more extended helix with base pairs roughly centered on and perpendicular to the helical axis. We take the crystal structure of the synthetic DNA dodecamer d(CpGpCpGpApApTpTpCpGpCpG) (PDB ID 1BNA) [11] as a representative B-DNA model, and refer to it as B-DNA in Table 4.3.1. However, the representative Z-DNA helix entails two sets of parameters, which contribute to its zig-zag pathway. Z-DNA is usually made up of alternating C and G bases and depending on the order of bases (CpG vs. GpC steps) the step parameters are quite different. The values that describe the two Z-DNA steps are also very different from those of A- or B-DNA. We take the crystal

structure of the DNA fragment d(CpGpCpGpCpG) (PDB ID 2DCG) [12] as a representative Z-DNA model, and refer to the two alternating stepts as CpG and GpC for Z-DNA in Table 4.3.1. Table 4.3.8 lists the pertinent twist data and details about the representative molecules, such as the sequence, for each of the structures discussed in this section.

Step Parameter	Averages for A-DNA	Averages for B-DNA	Averages for Z-DNA (CpG)	Averages for Z-DNA (GpC)
Tilt	0.1°	–0.1°	0.74°	0.31°
Roll	8.0°	0.6°	–1.89°	–3.64°
Twist	31.1°	36.0°	–6.57°	–51.7°
Shift	0.00Å	–0.02Å	0.08Å	0.02Å
Slide	–1.53Å	0.23Å	5.36Å	–0.93Å
Rise	3.32Å	3.32Å	3.71Å	3.4Å

Table 4.3.1: Average values of base-pair step parameters in high resolution A-, B-, and Z-DNA crystal structures [14,15,16].

We first look at the A-DNA. Figure 4.3.1 shows a top-down view of a single helical turn of an ideal A-DNA structure based on the X-ray diffraction of DNA fibers under low water content. The figure shows lines representing the chemical bonds, and balls for the carbon, nitrogen, and phosphorus atoms. The ideal A-DNA consists of 11 base-pairs per helical turn with an ideal helical twist of 32.73° (=360°/11). The representative A-DNA is a synthetic self complementary octamer with the sequence GGTATACC (PDB ID 1VJ4). This is an X-ray crystal structure of a little more than half of a turn of A-type DNA, eight base pairs, reported by Shakked et al. [10]. This structure is prepared in TwiDDL and is taken here as representative of A-type DNA. Looking at the 3-D graph, shown in Figure 4.3.2 and provided at the top of the page in TwiDDL's individual TwID page, we see that all the base-pair steps are very under twisted compared to relaxed B-DNA. This can also be seen in the data shown in Table 4.3.2. These differences are clear from the dark blue steps in Figure 4.3.2. The six step parameters for the structure are listed in Table 4.3.3. The average Tw^{SC} per step, at 26.78° in Table 4.3.8, is a bit lower than the ideal A-DNA helical twist of 32.73°, which is different from the twist of supercoiling for A-DNA. The pathway of the base-pair steps used to compute Tw^{SC} does not coincide with the helical axis. This tells us that A-DNA is more chiral in nature and further away from B-DNA.

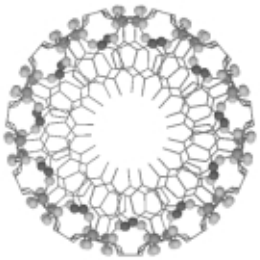


Figure 4.3.1: Top-down view of a single helical turn of an ideal A-DNA fiber with 11 base pairs per helical turn, as illustrated by Andrew Colasanti in the Handbook of Molecular Biophysics: Methods and Applications [9].

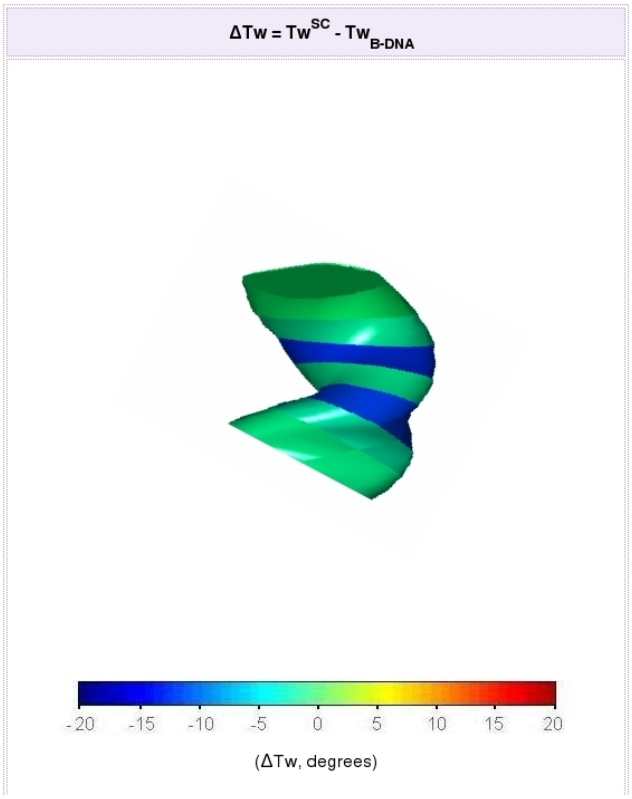


Figure 4.3.2: Representation of the octamer d(G-G-T-A-T-A-C-C) taken from TwiDDL showing the deviation $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$ with color coded base-pair steps.

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°)
1	29.96	-4.31
2	31.49	-2.79
3	21.54	-12.73
4	30.46	-3.81
5	22.29	-11.99
6	30.77	-3.51
7	28.51	-5.77

Table 4.3.2: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the octamer d(G-G-T-A-T-A-C-C) taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps.

Base Pair Step	$Shift^{SP}$(Å)	$Slide^{SP}$(Å)	$Rise^{SP}$(Å)	$Tilt^{SP}$(°)	$Roll^{SP}$(°)	Tw^{SP}(°)	$Kinking^{SP}$(°)	$Shearing^{SP}$(Å)
1	0.26	-1.82	3.51	-0.97	7.29	32.16	7.35	1.83
2	-0.53	-1.34	3.32	-1.70	0.93	34.62	1.93	1.44
3	0.25	-1.29	3.13	0.05	9.94	27.60	9.94	1.31
4	0.31	-1.09	3.16	0.93	3.01	32.95	3.15	1.13
5	-0.54	-1.45	3.27	-0.52	14.75	29.09	14.75	1.54
6	0.51	-1.30	3.16	0.61	3.61	35.06	3.66	1.39
7	-0.52	-1.86	3.46	-1.63	9.68	31.10	9.81	1.93

Table 4.3.3: Base-pair step parameters and the degree of kinking and shearing of each step. Based on the octamer d(G-G-T-A-T-A-C-C) taken from TwiDDL.

Our second structure is representative of B-type DNA. Figure 4.3.3 shows a top-down view of a single helical turn of an ideal B-DNA structure with 10 base-pairs per helical turn. In our comparisons with B-DNA we assume that the structure consists of 10.5 base-pairs per helical turn with an ideal twist of 34.28° (=360°/10.5). We have chosen as our representative B-DNA the X-ray crystal structure of the synthetic DNA dodecamer with a sequence of CGCGAATTCGCG consisting of a little more than a full helical turn (12 base pairs), as determined by Drew et al. [11] (PDB 1BNA). Examination of the 3-D graph provided at the top of the page in TwiDDL's individual TwID page (Figure 4.3.4) shows that all the base-pair steps are color-coded with values relatively close to the twist of relaxed B-DNA. As seen in Figure 4.3.4 and Table 4.3.4 there is greater variation away from the ideal reference at some steps, such as C3-G4 and G4-A5. The set of step parameters describing the structure is given in Table 4.3.5. A structure with small values of $\Delta Tw^{B\text{-}DNA}$ is not considered chiral, meaning it exhibits no handedness in either direction, and because of this one would expect Tw^{SC} to be close to, if not exactly the same as the twist at each step in ideal B-DNA. In this example, the capabilities of TwiDDL to highlight structural variations, such as sequence-dependent impacts on Tw^{SC}, is demonstrated through both 3-D visualizations (Figure 4.3.4) and dynamically color-coded data analysis (Table 4.3.4).

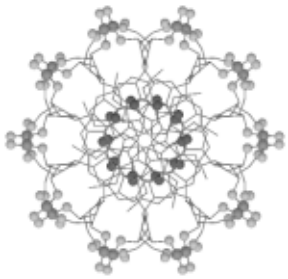


Figure 4.3.3: Top-down view of a single helical turn of an ideal B-DNA fiber with 10 base pairs per helical turn, as illustrated by Andrew Colasanti in the Handbook of Molecular Biophysics: Methods and Applications [9].

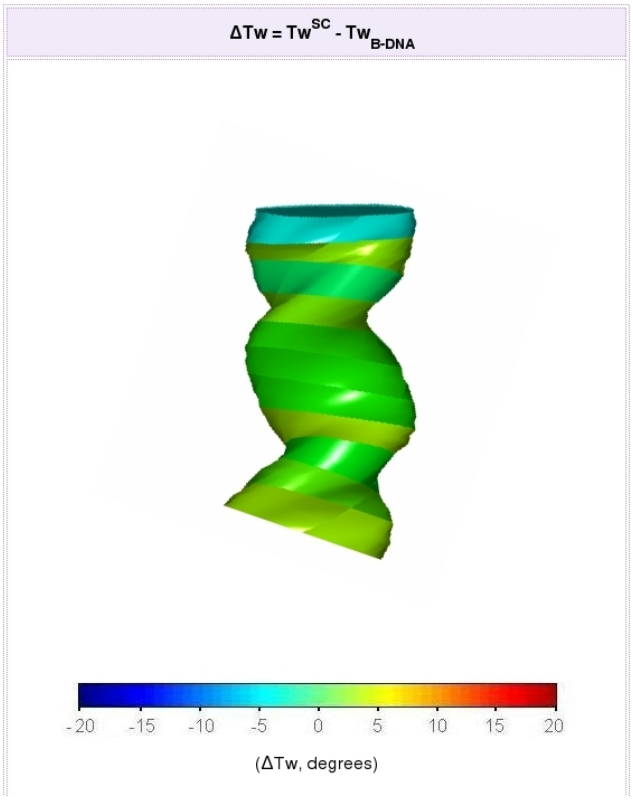


Figure 4.3.4: Representation of the synthetic DNA dodecamer d(CpGpCpGpApApTpTpCpGpCpG) taken from TwiDDL showing the deviation $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$ with color coded base-pair steps.

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°)
1	40.86	6.58
2	37.79	3.51
3	24.78	-9.49
4	41.04	6.76
5	35.06	0.77
6	35.14	0.85
7	35.47	1.19
8	39.23	4.94
9	28.31	-5.97
10	40.07	5.78
11	33.53	-0.74

Table 4.3.4: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the synthetic DNA dodecamer d(CpGpCpGpApApTpTpCpGpCpG) taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps.

Base Pair Step	$Shift^{SP}$(Å)	$Slide^{SP}$(Å)	$Rise^{SP}$(Å)	$Tilt^{SP}$(°)	$Roll^{SP}$(°)	Tw^{SP}(°)	$Kinking^{SP}$(°)	$Shearing^{SP}$(Å)
1	-0.36	0.15	3.52	-3.40	6.42	40.31	7.26	0.39
2	0.50	0.23	3.52	0.80	-4.73	38.15	4.79	0.55
3	-0.32	0.69	3.04	3.63	7.95	24.47	8.73	0.76
4	0.01	0.07	3.36	-2.68	3.16	40.90	4.14	0.07
5	0.10	-0.31	3.32	-0.70	0.95	35.35	1.18	0.32
6	0.33	-0.60	3.34	1.83	-2.75	34.76	3.30	0.68
7	-0.31	-0.18	3.32	2.96	0.73	35.39	3.04	0.35
8	0.02	-0.03	3.39	0.33	-0.05	39.27	0.33	0.03
9	0.38	0.86	3.24	-3.29	3.86	29.40	5.07	0.94
10	-1.30	0.42	3.68	-4.68	-12.20	40.78	13.06	1.36
11	0.77	0.06	3.23	3.14	-3.09	32.62	4.40	0.77

Table 4.3.5: Base-pair step parameters and the degree of kinking and shearing of each step. Based on the synthetic DNA dodecamer d(CpGpCpGpApApTpTpCpGpCpG) taken from TwiDDL.

The third and final structure is representative of Z-DNA. The selected example is the X-ray crystal structure of the complex of spermine and magnesium with the DNA fragment d(CpGpCpGpCpG) determined by Wang et al. [12] (PDB ID 2DCG), corresponding to half a turn of Z-DNA (6 base pairs). The two bound spermines and the magnesium ion found in the complex are shown by the ball-and-stick models in Figure 4.3.6. Unlike A- or B-DNA, Z-type DNA is not known to exist in vivo. We are therefore limited in our selection of structures. As noted above, Z-DNA has two distinct base-pair step parameter values, one for the C-G step and one for the G-C step. Examination of the 3-D graph that is provided at the top of the page in TwiDDL's individual TwID page (Figure 4.3.5) shows that all the base-pair steps are severely undertwisted compared to relaxed B-DNA. The remaining step parameters for the structure are listed in Table 4.3.7. The value of Tw^{SC} is shown in Table 4.3.6 to be some 45 to 91 degrees lower than the value of the B-DNA twist ($Tw^{B\text{-}DNA} = 34.28°$) for each type of step.

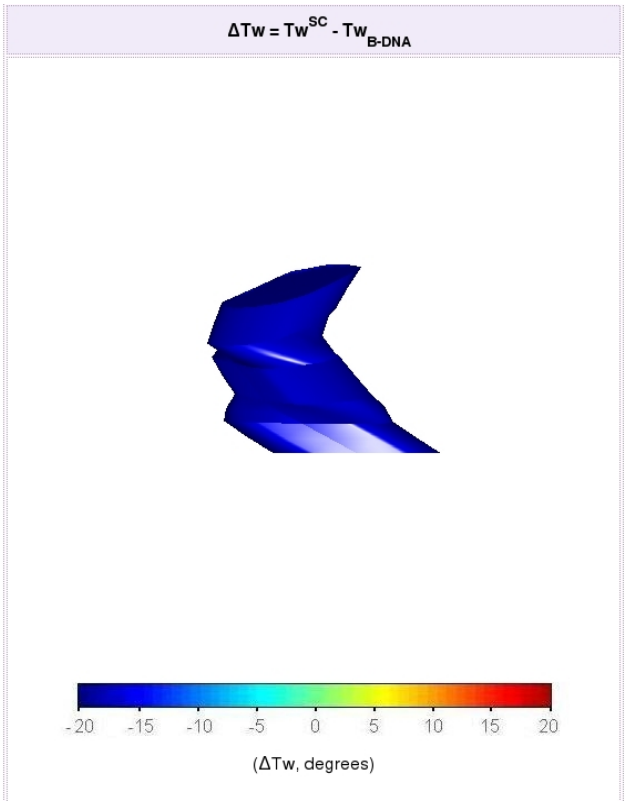


Figure 4.3.5: Representation of the complex of magnesium and spermine with the DNA fragment d(CpGpCpGpCpG) taken from TwiDDL showing the deviation

$\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$ *with color coded base-pair steps.*

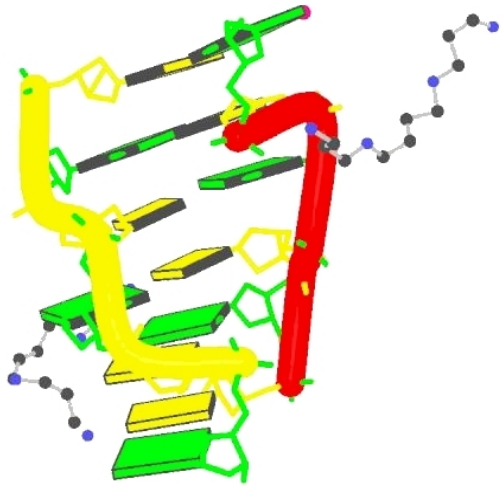


Figure 4.3.6: Representation of the complex of magnesium and spermine with the DNA fragment d(CpGpCpGpCpG) taken from the PDB [12]. The double helical DNA backbones are the thicker curves in red and yellow. The base pairs consist of guanine in green, and cytosine in yellow. The two bound spermines and the magnesium ion are shown by the ball-and-stick models. This image is from the PDB [13].

Base Pair Step	Tw^{SC}(°)	(Tw^{SC}-$Tw^{B\text{-}DNA}$)(°)
1	-12.29	-46.58
2	-56.04	-90.33
3	-13.48	-47.76
4	-56.62	-90.91
5	-11.40	-45.69

Table 4.3.6: Value of Tw^{SC} and the deviation of Tw^{SC} from the twist of B-DNA in the complex of magnesium and spermine with the DNA fragment d(CpGpCpGpCpG) taken from TwiDDL. Red is used to denote over twisted steps and blue to denote under twisted steps.

Base Pair Step	$Shift^{SP}$(Å)	$Slide^{SP}$(Å)	$Rise^{SP}$(Å)	$Tilt^{SP}$(°)	$Roll^{SP}$(°)	Tw^{SP}(°)	$Kinking^{SP}$(°)	$Shearing^{SP}$(Å)
1	0.02	5.34	3.65	0.58	-2.93	-6.46	2.98	5.34
2	-0.15	-0.99	3.27	-0.63	-6.75	-51.38	6.77	1.00
3	0.04	5.47	3.86	-0.57	-2.72	-6.50	2.77	5.47
4	-0.12	-0.86	3.28	1.12	-2.32	-52.69	2.57	0.86
5	0.05	5.10	3.38	1.10	-0.97	-8.32	1.46	5.10

Table 4.3.7: Base-pair step parameters and the degree of kinking and shearing of each step. Based on the complex of magnesium and spermine with the DNA fragment d(CpGpCpGpCpG) taken from TwiDDL..

Type	**A-DNA**	**B-DNA**	**Z-DNA**	
PDB ID	1VJ4	1BNA	2DCG	
Total Tw^{SC}	187.493°	391.486°	–146.249°	
Total Tw^{SP}	222.580°	391.400°	–125.350°	
Total $Turns^{SC}$	0.521	1.087	–0.406	
Average Tw^{SC} Per Step	26.78°	35.59°	–12.39°	–56.33°
Average Tw^{SP} Per Step	31.80°	35.58°	–7.99°	–51.15°
Sequence	GGTATACC	CGCGAATTCGCG	CGCGCG	
Resolution of X-ray Crystal	1.80 Å	1.90 Å	0.90 Å	

Table 4.3.8: Comparisons of the twist in representative structures of the three types of DNA (A, B, and Z) discussed in this chapter [10,11,12].

4.4 Nucleosome Data

When we started examining Tw^{SC} we studied the nucleosome. Formation of the nucleosome is the first of many steps taken to compact the large amounts of DNA (each human cell contains around two meters) inside a eukaryotic cell into a manageable form. The wrapping of double-helical DNA around the nucleosome core protein starts the compaction necessary to contain massive amount of DNA inside one tiny cell. Each nucleosome has about 147 base pairs of DNA wrapped 1.67 helical turns around it. There are currently more than 70 X-ray crystal structures of different nucleosomes stored in the PDB and 33 in TwiDDL. Most of these structures have 146 base pairs but there are some have one more or one less. However, all in all, the nucleosome structures contain the largest number of DNA base-pair steps resolved in any X-ray crystal structure in the PDB.

The motivation for studying the nucleosome is due to it being a major contributor for organizing DNA in eukaryotic cells. This organization requires a lot of distortion the DNA from its ideal B form. There exists so much variety in these steps, we could analyze data in one area of the nucleosome and compare it to a similar step near or far away from it. The nucleosome contains a wealth of information that just needs to be gathered and studied to better understand it.

We have put together a large compilation of nucleosomal data from the structures stored in TwiDDL. We have calculated the Tw^{SC} values for each nucleosome structure.

The challenge here was finding a way to interpret the data from the X-ray crystals. The various X-ray crystal structures were not entered in the same way into the PDB. The content of the PDB files needed to be changed from the original PDB file to correct mislabeled atoms in some structures (a base from one strand would be incorrectly oriented with respect to the sugar-phosphate backbone), and to specify the ordering of base pairs along the strands. 3DNA also needed input that specified the identities of the base pairs. The final corrected files include the Tw^{SC} values, along with the six step parameters and the correctly labeled bases and strands.

The amount of data generated for each nucleosome file is quite large. There are 45 nucleosome structures worked up in total, as shown in Appendix A. These 45 nucleosomes consist of the 33 structures found within TwiDDL combined with 12 additional nucleosomes whose PDB files contain anomalies that require manual corrections and have not been stored in TwiDDL for the reasons documented in Section 3.3.2. In future these 12 additional nucleosome structures will be added to the TwiDDL database. With a large amount of data from all 45 structures, we needed to come up with an efficient way to study what is happening along the sequence in the nucleosomes and how the Tw^{SC} helps to capture it.

In typical B-DNA, we know that one helical turn is 10.5 base pairs long. We also know that each type of DNA (i.e., A, B, Z,...) has its own unique number of base pairs per helical turn. The nucelosomal DNA does not consist of one type of DNA alone. The X-

ray crystal structures studied here are synthetic proteins and sequences. We thought it would be very interesting to see how each type of DNA changes throughout the path of binding over the nucleosome core protein.

To see how the DNA changes over the nucleosome core particle, an average helical turn was calculated. The average helical turn at a particular base-pair step was taken as the average of the Tw^{SC} at that step and the two preceding and two following steps in the sequence. The number is then expressed as the number of base pairs per helical turn by dividing 360° by the average value of Tw^{SC}, i.e., $360° / \langle Tw^{SC}_{i-2->i+2} \rangle$. With this value we can visualize how the helical turns are progressing locally along each sequence. The nucleosomes were evaluated in three separate groups. The first group contained 145 base-pair sequences, the second 146 base-pair sequences, and the third 147 base-pair sequences.

Figure 4.4.1 shows the average helical turns derived by the running average values of Tw^{SC} for each of the 145 base pair nucleosome structures listed in Appendix A. Each structure is labeled by its PDB-ID in the legend on the right. The base-pair numbering on the x-axis is centered at zero, the central base pair which lies on the dyad of the structure. Figure 4.4.2 contains a plot of the composite averages and standard deviation of the number of helical turns at each base-pair step over all of the 145 base-pair nucleosome complex structures. Note that the number of residues per turn shows a regular periodicity.

Figure 4.4.3 shows the average helical turns of the 146 base-pair nucleosome structures listed in Appendix A and Figure 4.4.4 shows the composite averages and standard deviation. There are many more nucleosomes with 146 base pairs of DNA than nucleosomes with 145 and 147 base pairs combined. But even from these two very busy graphs the periodicity that the nucleosome core particle place on the number of residues per turn in double helical DNA bound to it is quite clear. Figures 4.4.5 and 4.4.6 show the corresponding data for the 147 base-pair nucleosome structures listed in Appendix A. These show the periodicity similar to those of the 145 and 146 base-pair structures. Positioning of the DNA base-pair sequence on the histone core particle is crucial for binding to occur. The base-pair fragments centered on the dyad of the core particle in the X-ray crystals generally have a greater number of helical turns compared to B-DNA, i.e., 11 or more versus 10.5 base-pairs per turn.

What this section is telling us is two-fold. The first, and most striking is the clear periodicity that a nucleosome core particle exerts on the number of residues per turn in its DNA bound sequence. This is independent of where this nucleosome is taken from, for example, whether the histones are from frog, fly, human, and so on. However, proteins are sensitive to where on a segment of DNA they prefer to bind. Widom and Lowry have shown that certain sequences bind with varying levels of affinity for the nucleosome [20]. This can be attributed to the topology of the naked sequence and how much energy it takes to complete the deformation around the nucleosome. Widom and

Lowry's experiments found a sequence, referred to as #601, that binds to the nucleosome with higher affinity than a natural sequence [20]. The α-satellite DNA incorporated in many nucleosomes does not bind as tightly as the 601 sequence.

The topology of nucleosomal DNA changes along the sequence along with its overall global picture. The conformation along the sequence includes A, B, or C like forms of DNA. A like DNA is undertwisted with 11 base pairs per turn, B is the ideal relaxed DNA with 10.5 base pairs per turn, and C DNA is overtwisted with 9 base pairs per turn. Seeing as how the Tw^{SC} was developed in a manner that takes into account its neighboring base pairs flanking the step of interest, it is the more practical way to look at the global topology.

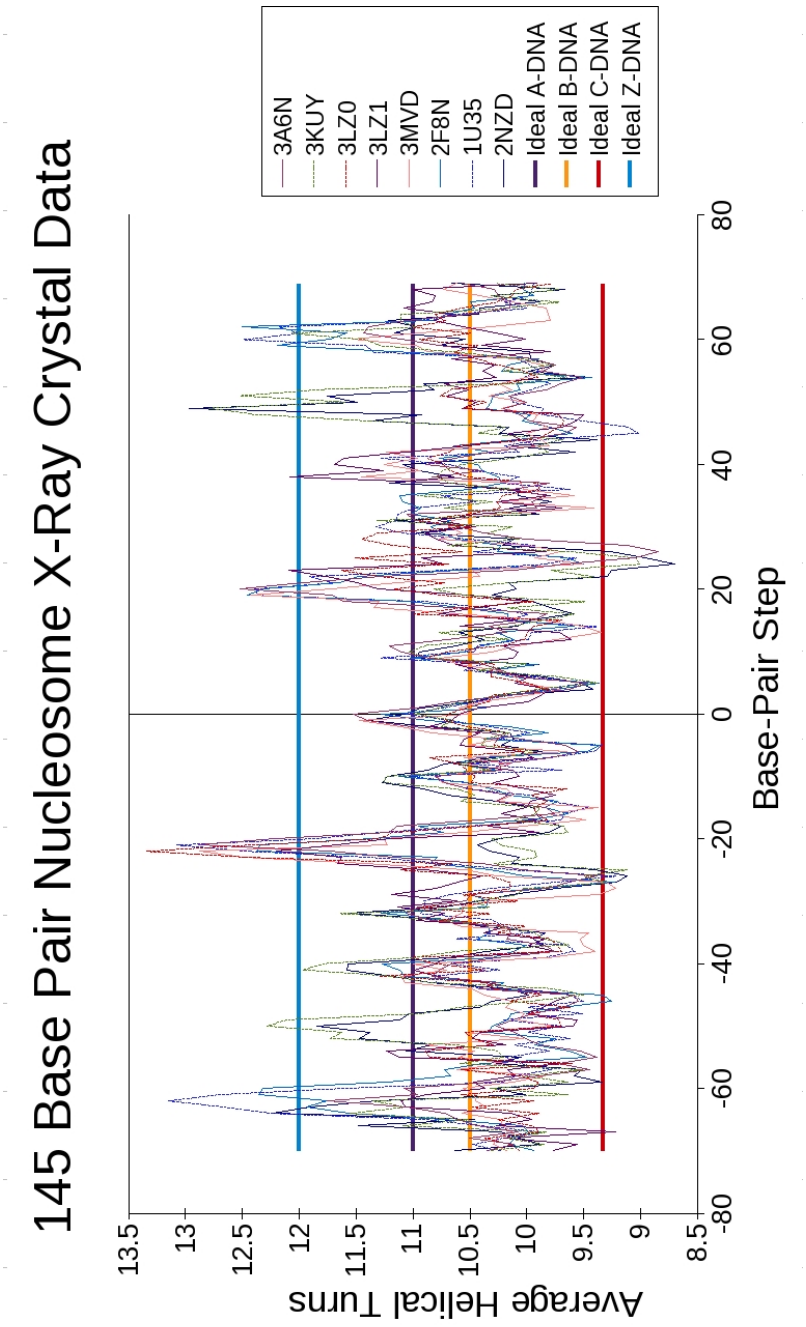


Figure 4.4.1: Variation in the number of helical turns with sequential position in eight 145-base-pair nucleosome complex structures. The PDB-IDs are noted in the legend on the right. The Y-axis shows the average helical turns at each base-pair step based on the running average of Tw^{SC} over five steps centered at the value denoted on the X-axis. Base-pair step locations are expressed with respect to the dyad at zero. Additionally, four bold lines highlight the ideal values for the number of helical turns of the A, B, C, and Z forms of DNA.

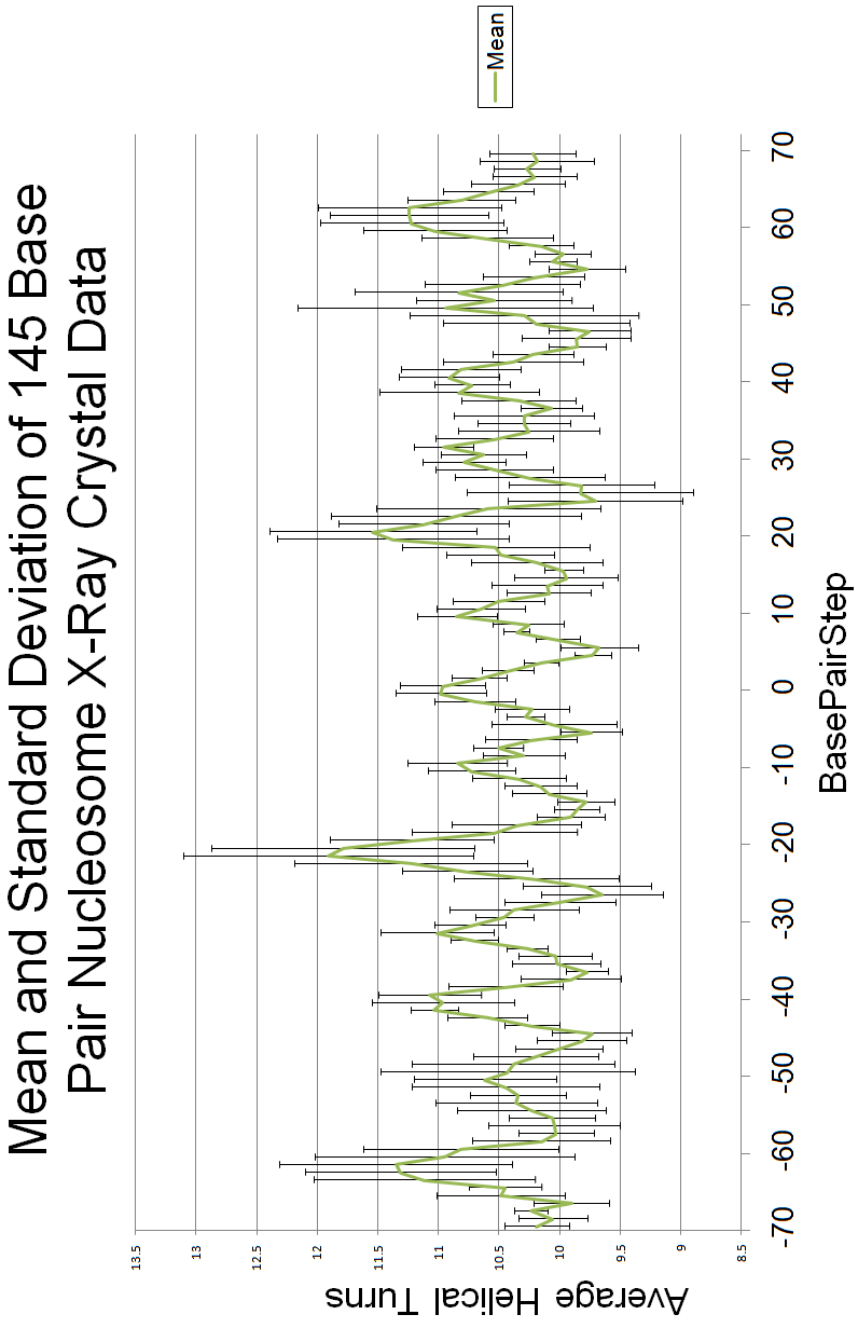


Figure 4.4.2: Composite values of the number of residues per turn in eight 145-base-pair nucleosome complex structures. The data are the averages and standard deviations of corresponding points in the eight curves reported in Figure 4.4.1.

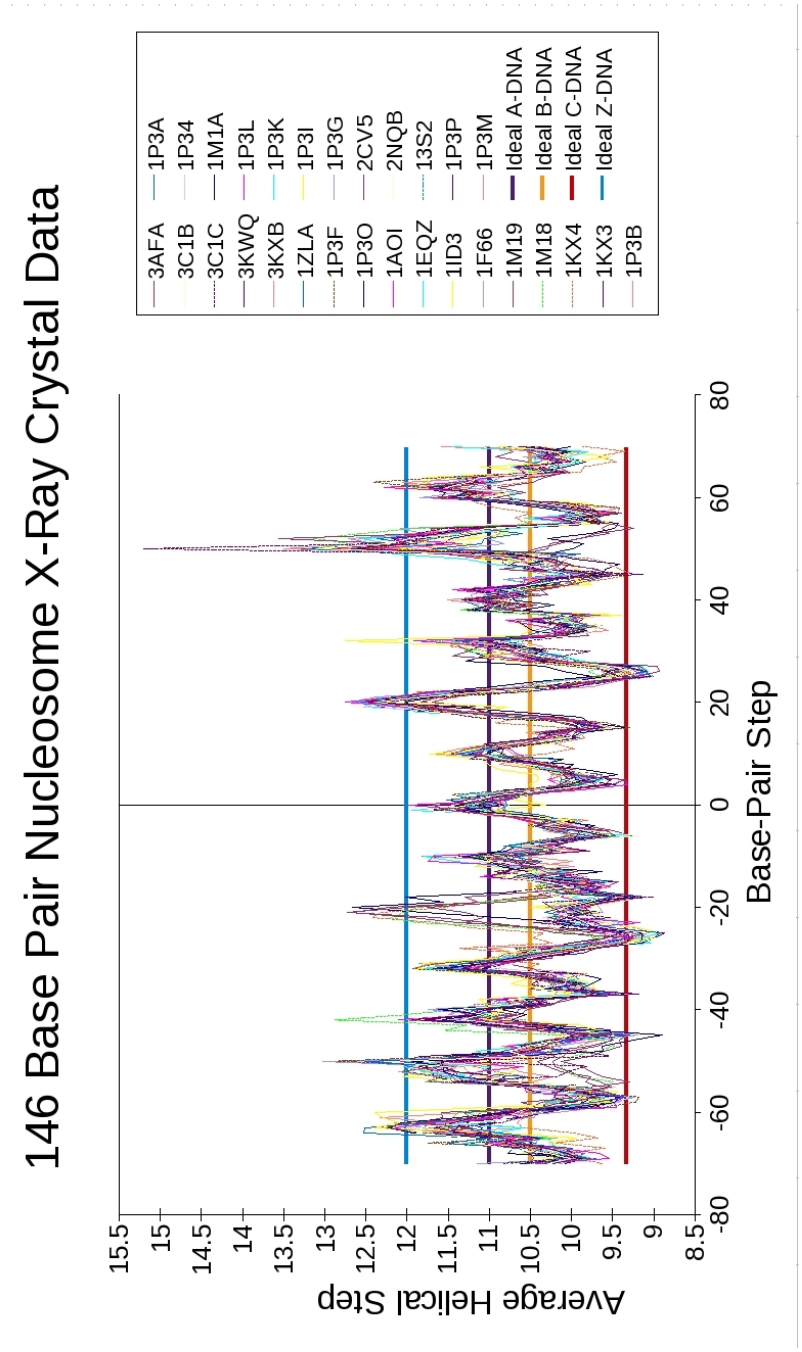


Figure 4.4.3: Variation in the number of helical turns with sequential position in 29 146-base-pair nucleosome complex structures. The PDB-IDs are noted in the legend on the right. The Y-axis shows the average helical turns at each base-pair step based on the running average of Tw^{SC} over five steps centered at the value denoted on the X-axis. Base-pair step locations are expressed with respect to the dyad at zero. Additionally, four bold lines highlight the ideal values for the number of helical turns of the A, B, C, and Z forms of DNA.

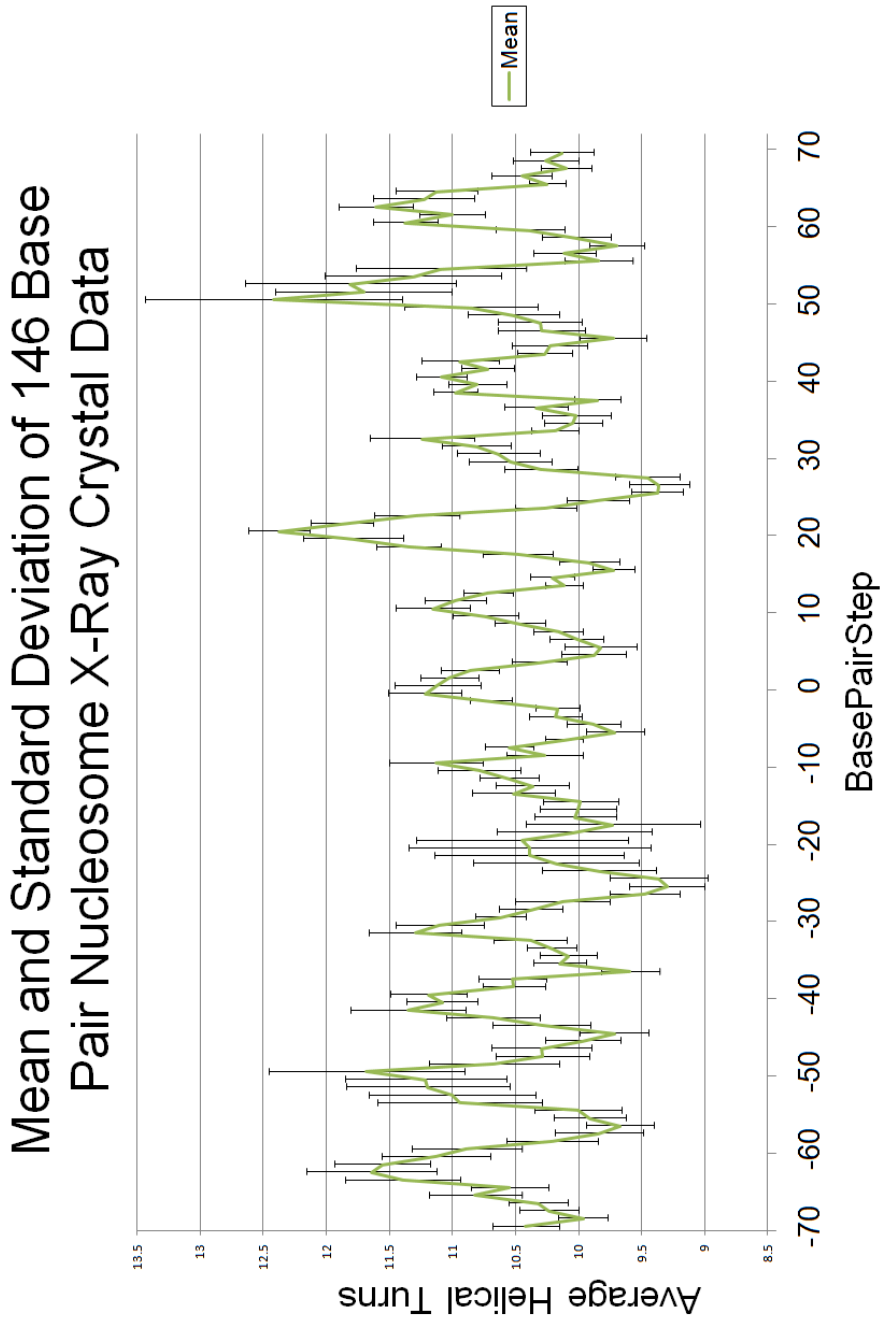


Figure 4.4.4: Composite values of the number of residues per turn in 29 146-base-pair nucleosome complex structures. The data are the averages and standard deviations of corresponding points in the 29 curves reported in Figure 4.4.3.

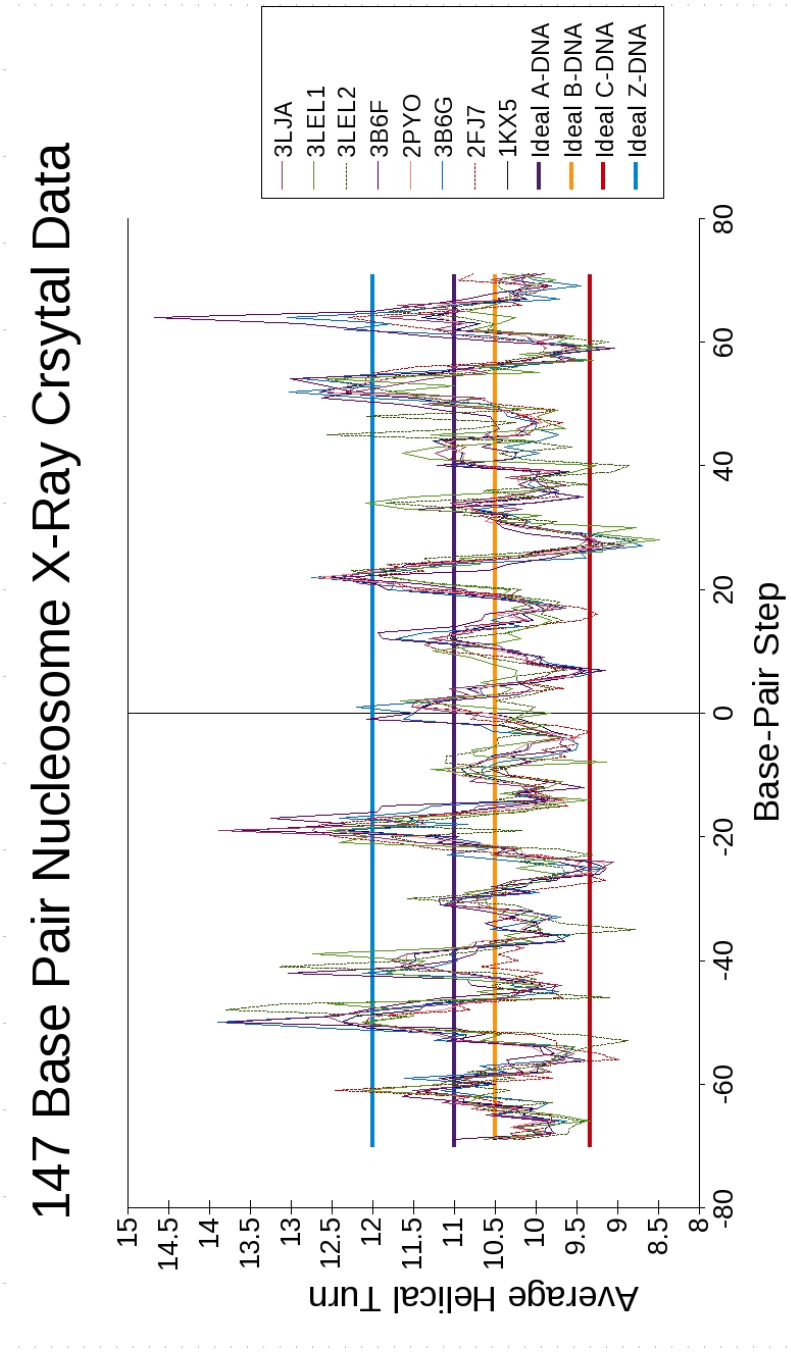


Figure 4.4.5: Variation in the number of helical turns with sequential position in eight 147-base-pair nucleosome complex structures. The PDB-IDs are noted in the legend on the right. The Y-axis shows the average helical turns at each base-pair step based on the running average of Tw^{SC} over five steps centered at the value denoted on the X-axis. Base-pair step locations are expressed with respect to the dyad at zero. Additionally, four bold lines highlight the ideal values for the number of helical turns of the A, B, C, and Z forms of DNA.

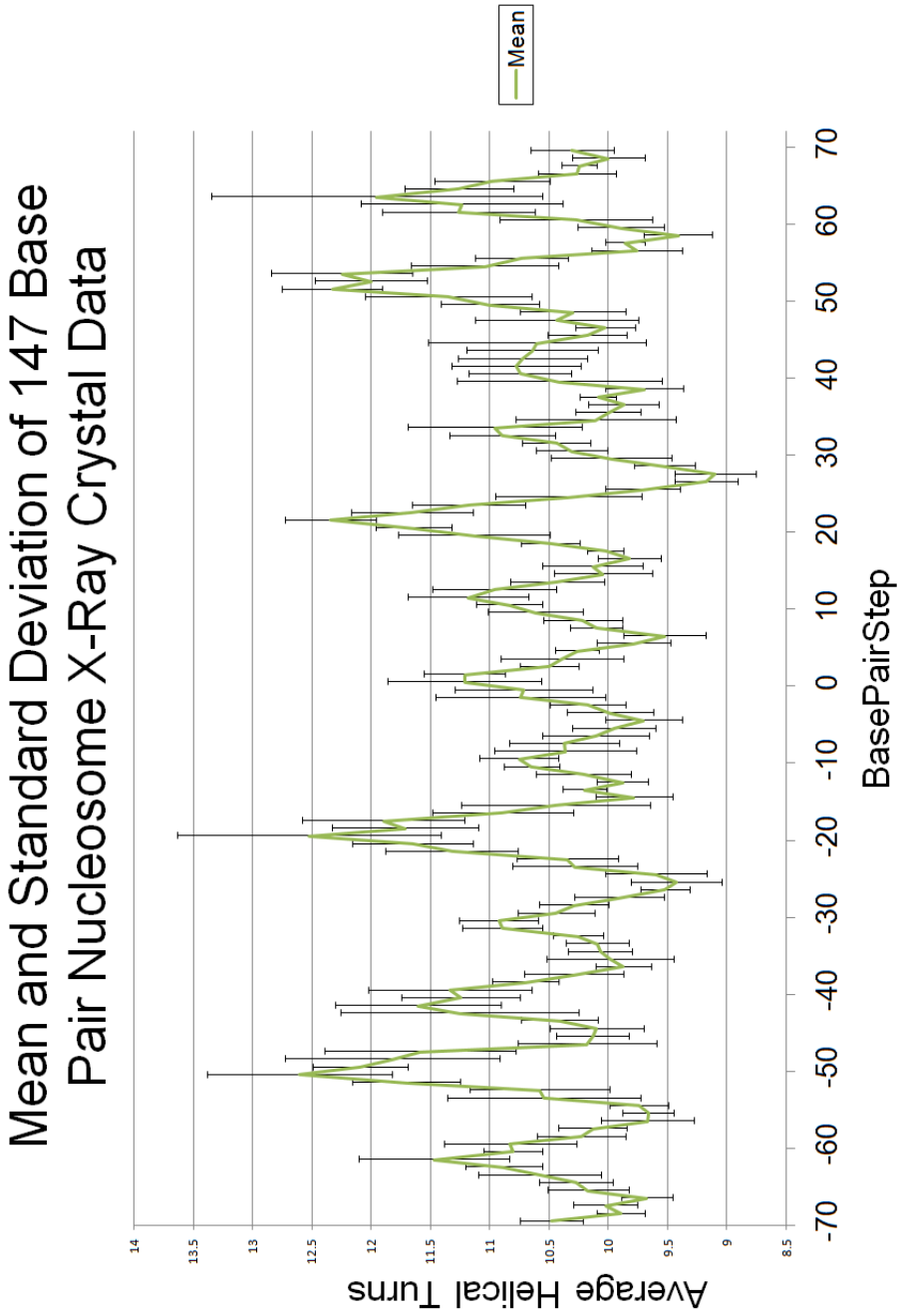


Figure 4.4.6: Composite values of the number of residues per turn in eight 147-base-pair nucleosome complex structures. The data are the averages and standard deviations of corresponding points in the eight curves reported in Figure 4.4.5.

4.5 Effect of Shearing on HU and Nucleosomal Steps

TwiDDL was developed to identify DNA structures on the basis of topology using the twist of supercoiling Tw^{SC} [22]. Chapter 2 discussed the mathematical derevation and observed the importance that the local chirality of the structure has on Tw^{SC}. This means that a base-pair step must have both translation and rotation to change Tw^{SC} from the ideal B-DNA value. Therefore, when we notice that a step has a deviation from ideal B-DNA, whether a positive or negative one, we can automatically tell that we are dealing with a chiral step in the structure. The degree of the deviation lets us know the extent of the chirality.

The surrounding two base pairs sandwiching the step of interest contribute to the deviation since the Tw^{SC} calculations rely on the positions of the origins of those two base pairs. If the base pairs follow the pattern of ideal B-DNA, they do not contribute to the chirality of the step of interest. However, if these surrounding steps do not follow a perfect ideal B-DNA configuration, we know that they are at least partially responsible for the chirality.

You might be asking "so what?" Well, let us think about a simple lock and key analogy. Not every key just has one "tooth" on it to open a lock, and this "tooth" is not always just a rectangular outcropping. Instead, the "tooth" is usually rounded and is not alone on the key. This special code in the key can only open a specific lock and this lock,

in turn, can only be opened by a key with these requirements. Proteins, ligands, and drugs and the interactions they have with DNA seem to show definite preferences in their binding sites [23].

It is not correct to say that only a specific sequence will be the sole one to bind to a certain protein. Rather, there are many sequences that may show similar shapes that can potentially bind a protein or drug. The shape of DNA is readily affected by salinity [24,25], temperature [26], and restrictions placed on the ends of segments of interest. These end conditions are just that, something pulling on the ends of the sequence of interest. Depending on the way the ends are moved, the shape of the DNA will be altered from ideal B-DNA and can be attractive to binding proteins/ligands/drugs. Whether the DNA binds to proteins, ligands, or drugs depends on many factors, including the presence of one protein over another, temperature, and the shape of the binding site and the protein. To show how chirality can change Tw^{SC} and how chirality differs in known protein-decorated DNA segments, here we show two X-ray crystal structures.

During the development of the equations for Tw^{SC} we tested out the theory using a real world example and a simplified model for each incarnation of the process. The simplified model was the simple four base pair tetramer shown in Chapter 2.3. The real world example was a highly deformed step found in the best-resolved nucleosome-core particle structure [17]. We were interested in how this new twist, the Tw^{SC}, detected steps that were not typical B-DNA.

The nucleosome structure is not as simple as the 4 base-pair model. The model only dealt with a B-DNA segment containing a single central kink-and-slide deformation. If the kink and the slide were large enough, the deviation in Tw^{SC} from the step parameter twist would be large. We also noted something very interesting in the nucleosome. Some of the larger differences between the Tw^{SC} and Tw^{SP} occurred when two or more steps in a row (meaning the surrounding two base pairs flanking the step of interest) also had significant bending and shearing. From what we have seen so far, it does not matter that the shearing is relatively small for the steps involved in the calculations of Tw^{SC}, just that the steps occur in conjunction with a chiral step of interest. The flanking steps contribute to a larger difference in Tw^{SC} versus the step parameter twist, than if only the central step were chiral.

Here we look at two DNA/protein complexes used to package DNA. Both structures entail a high degree of DNA bending and translation. The first complex, the currently best resolved nucleosome core particle structure (PDBID 1KX5) resolved by Davey et al. [17], is found in eukaryotes as described. The second complex is an HU/DNA protein complex (PDBID 1P78) which is found in prokaryotes, here the complex with HU from Anabaena resolved by Swinger et al. [18]. In addition to the original DNA/protein complexes, we also looked at idealized structures with the same degree of bending but the shearing removed. Shearing is removed by setting both slide and shift to zero, while keeping the other step parameters at their original values. The

origins and base-pair triads were determined for the reconstructed model and Tw^{SC} was then completed.

The results are shown below in Tables 4.5.1-4.5.4, with the steps of interest highlighted in violet. The non-highlighted rows contain the six step parameters and Tw^{SC} for the central and flanking base-pair steps. The differences between the two types of twists in the case of the HU/DNA are particularly large. In Table 4.5.3, these differences are close to or over 10º. If the shearing is removed the difference in the two twists drop significantly (to less than 0.7º). Although the two step-parameter twists are identical in the presence or absence of the shearing, Tw^{SC} changes from a larger value in the presence of shearing to one that is close to the step-parameter twist when the shearing is removed. When the kink is removed from the step of interest by setting roll and tilt to zero, we also note that the differences in the two twists are reduced significantly.

We focus on the four strongest kink-and-slide steps in the nucleosome. Two of these steps are adjacent to one another, numbers 89 and 90. The differences in Tw^{SC} and Tw^{SP} are relatively small (less than 5º in magnitude). We chose these steps because of their role in DNA compaction, not because of the difference between Tw^{SC} and Tw^{SP}.

The value of Tw^{SC} at each step of interest, listed in Table 4.5.1, is lower than that of Tw^{SP} by –4.40° to –5.86°. The difference in the twist of supercoiling and step parameter twist indicates that shearing and bending are occurring together and, therefore, lending to a chiral structure. If we compare these numbers to those found for a similar

structure with the shearing removed (Table 4.5.2) we can see that the values are much closer to zero. They are not exactly zero since the surrounding base pairs are not ideal B-DNA and some deviation from zero is to be expected. Remember, that the Tw^{SC} relies not only on the two base pairs of the step, but also the base-pair steps immediately surrounding the step of interest. If those extremities differ from the B-like DNA form, they will contribute a non-zero amount to the Tw^{SC} even with the step of interest having zero kinking and/or shearing.

Figures 4.5.1-4.5.2 give visual understanding into what is happening. Figure 4.5.1 shows the superhelical pathway of nucleosomal DNA in the crystal and the wide variation in Tw^{SC} compared to the twist of B-DNA (evident from the color coding). Figure 4.5.2 shows not only the flattening of DNA into a self-intersecting circle that results from the removal of shearing, but also the disappearance of highly over twisted (deep blue) steps at the modified sites. It is clear that the local chirality affects the overall chirality of the DNA. The chain deforms from a left-handed superhelix looking like a spring to a flattened object that cannot exist physically in nature without breaking the laws of physics and having two or more atoms occupying the same space.

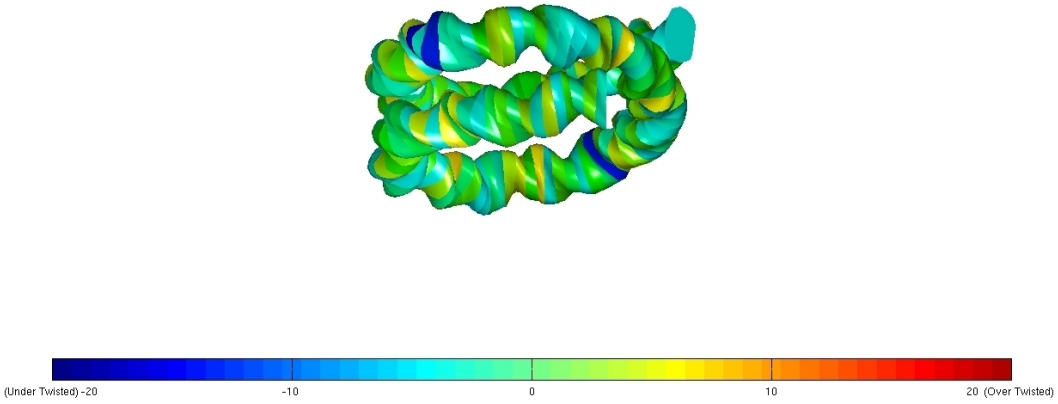


Figure 4.5.1: Screen capture from TwiDDL of the 147 DNA base pairs in the X-ray structure of the nucleosome core particle, NCP147, determined at 1.9Å resolution by Davey et al. [17]. This is a graphical depiction of the difference in the twist of supercoiling compared to the twist of relaxed B-DNA, $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$. Notice the dark blue under twisted steps and the yellow over twisted ones.

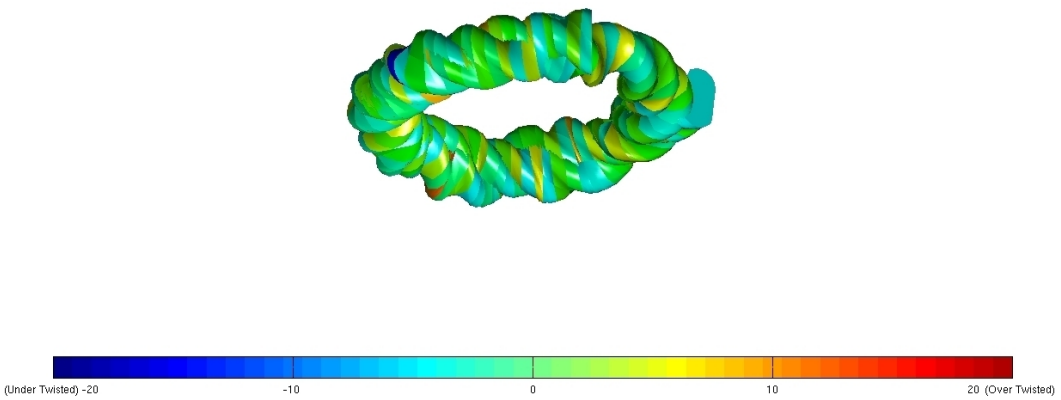


Figure 4.5.2: Screen capture from TwiDDL of DNA in a hypothetical model, based on the X-ray structure of the nucleosome core particle NCP147 by Davey et al. [17], without shearing (shifting and sliding removed) on any of the base-pair steps. This is a graphical depiction of the difference in the twist of supercoiling compared to the twist of relaxed B-DNA, $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$. Notice the dark blue under twisted step and the yellow over twisted ones.

The complex of HU and DNA is a much smaller system than the nucleosome. Here we examine the two steps most severly deformed in the structure. Figure 4.5.3 shows the structure generated with 3DNA and stored in the PDB. Currently we do not have a way to account accurately for the twisting of melted DNA, therefore, the two base-pair steps with the flipped out bases, shown in blue, were removed from our calculations and the step parameters were adjusted to compensate for these losses. The compensation adjustment found the step parameters of the two base pair steps, surrounding the one with the flipped out base, with respect to each other. For instance, the rise between the normal base pair and the one with flipped out base would be close to 3.4Å, however, with the flipped out base pair removed, the two normal base pair steps surrounding the flipped out one has a rise of 7.75Å, which is a little more than double the B-DNA rise. The same idea of getting the other step parameters applies.

Figure 4.5.4 shows the color coded images of the difference, $\Delta Tw^{B\text{-}DNA}$, in the twist of supercoiling compared to the twist of B-DNA. The large over twisted sections, of the structure are immediately obvious from the yellow/orange colored base-pair steps. Table 4.5.3 shows our two steps of interest with large differences between Tw^{SC} and Tw^{SP} (10.98° and 8.48°). Figure 4.5.4 shows how far away from ideal B-DNA the molecule is with large amounts of over and under twisting.

The removal of shearing does not eliminate the over twisting compared to B-DNA. The deviation in the twist of supercoiling compared to relaxed B-DNA in Figure

4.5.5 shows only a slight reduction in the amount of over and under twisting. The colors are only somewhat subdued. The differences in Tw^{SC} compared to Tw^{SP}, however, drop to 0.48° and 0.91° respectively. There is still an amount of bending in the step of interest and the surrounding steps do not resemble ideal B-DNA, so we would not expect the $\Delta Tw^{B\text{-}DNA}$ to be zero but it is still a large change from when shearing was still present.

Base-Pair Step	Shift, Å	Slide, Å	Rise, Å	Tilt, °	Roll, °	T^{SP}, °	T^{SC}, °
25	0.10	0.18	3.42	6.64	7.83	29.68	29.49
26	0.29	2.46	3.00	–2.62	–8.45	43.36	38.87
27	–0.12	0.26	3.55	–2.51	–1.99	36.56	35.49
37	–0.18	–0.18	3.02	–0.30	–0.68	28.61	29.57
38	–0.43	2.58	3.25	–2.30	–18.43	50.04	45.64
39	0.95	0.86	3.47	2.04	7.00	28.54	28.42
88	–1.3	0.52	3.31	–3.00	–7.01	38.71	39.35
89	0.48	1.94	3.66	1.0	–8.32	45.87	41.27
90	–0.36	1.66	3.47	–2.00	–9.28	40.37	34.51
89	0.48	1.94	3.66	1.06	–8.32	45.87	41.27
90	–0.36	1.66	3.47	–2.00	–9.28	40.37	34.51
91	0.09	0.19	3.16	–3.20	3.71	32.37	32.55

Table 4.5.1: Specific steps of the X-ray structure of the nucleosome core particle, NCP147, at 1.9Å resolution which have a high Tw^{SP} . This table holds the values for the step parameters as well as the Tw^{SC}. The steps of interest are the ones highlighted in violet.

Base-Pair Step	Shift, Å	Slide, Å	Rise, Å	Tilt, °	Roll, °	T^{SP}, °	T^{SC}, °
89	0	0	3.66	1.06	–8.32	45.87	45.97
90	0	0	3.47	–1.99	–9.29	40.37	40.35
91	0	0	3.16	–3.20	3.72	32.37	32.30
88	0	0	3.31	–3.00	–7.01	38.71	38.75
89	0	0	3.66	1.06	–8.32	45.87	45.97
90	0	0	3.47	–1.99	–9.29	40.37	40.35
25	0	0	3.42	6.63	7.83	29.68	29.59
26	0	0	3.00	–2.61	–8.45	43.36	43.22
27	0	0	3.55	–2.52	–1.98	36.56	36.56
37	0	0	3.02	–0.31	–0.68	28.61	28.62
38	0	0	3.25	–2.29	–18.44	50.03	49.97
39	0	0	3.47	2.03	7.00	28.54	28.49

Table 4.5.2: The same specific steps of the X-ray structure of the nucleosome core particle, NCP147, at 1.9Å resolution from Table 4.5.1 which have a high Tw^{SP}. These steps have been altered from the original PDB entry. The shearing has been removed and the Tw^{SC} reflects that change. The steps of interest are the ones highlighted in violet.

Figure 4.5.3: Protein/DNA complex of 17 base pairs in the Anabaena HU-DNA cocrystal structure (AHU2). Notice that two thymines (blue bases) are flipped out. These two bases are not taken into consideration in the twist calculation [18]. This image is from the NDB [19].

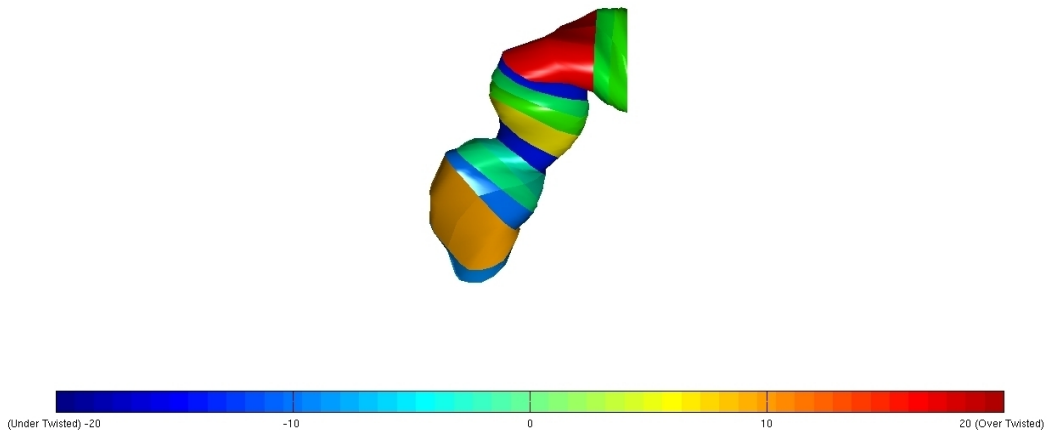


Figure 4.5.4: Screen capture from TwiDDL of the Anabaena HU-DNA cocrystal structure (AHU2), determined at 2.25Å resolution by Swinger et al. [18]. This is a graphical depiction of the difference in the twist of supercoiling compared to the twist of relaxed B-DNA, $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$. *Notice the highly under twisted steps in blue and highly over twisted steps in orange and red.*

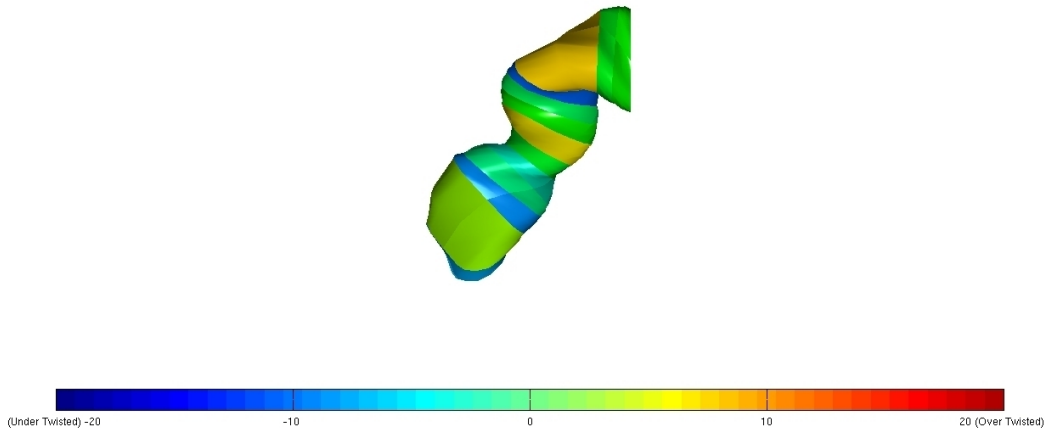


Figure 4.5.5: Screen capture from TwiDDL of DNA in a hypothetical model, based on the Anabaena HU-DNA cocrystal structure (AHU2) by Swinger et al. [18], without shearing (shifting and sliding removed) on any of the base-pair steps. This is a graphical depiction of the difference in the twist of supercoiling compared to the twist of relaxed B-DNA, $\Delta Tw^{B\text{-}DNA} = Tw^{SC} - Tw^{B\text{-}DNA}$.

Base-Pair Step	Shift, Å	Slide, Å	Rise, Å	Tilt, °	Roll, °	T^{SP}, °	T^{SC}, °
2	–0.23	–0.23	3.05	–1.41	1.36	25.31	25.74
3	–2.34	2.40	7.75	24.17	63.03	45.47	56.45
4	0.86	–0.14	2.96	–5.77	4.84	19.72	23.55
11	–0.49	–0.13	2.95	6.51	5.99	17.96	19.71
12	1.41	1.88	8.43	–27.86	64.40	39.31	47.79
13	–0.20	0.41	2.82	–2.04	7.46	17.17	18.54

Table 4.5.3: Specific steps of the HU/DNA complex as taken directly from X-ray crystal data which have a high Tw^{SP}. This table holds the values for the step parameters as well as the Tw^{SC}. The steps of interest are the ones highlighted in violet.

Base-Pair Step	Shift, Å	Slide, Å	Rise, Å	Tilt, °	Roll, °	T^{SP}, °	T^{SC}, °
2	0	0	3.05	–1.40	1.36	25.30	25.23
3	0	0	7.75	24.17	63.03	45.47	45.95
4	0	0	2.96	–5.77	4.84	19.73	20.24
11	0	0	2.95	6.51	5.98	17.96	18.57
12	0	0	8.43	–27.86	64.39	39.31	40.22
13	0	0	2.82	–2.04	7.46	17.17	17.38

Table 4.5.4: Specific steps of the same HU/DNA from Table 4.5.3 which have a high Tw^{SP}. These steps have been altered from the original PDB entry. The shearing has been removed and the Tw^{SC} reflects that change. The steps of interest are the ones highlighted in violet.

4.6 Appendix A - Table of Nucleosomes

PDB ID	Number of Base Pairs	Structure Title	Sequence	Reference
1AOI	146	Complex Between Nucleosome Core Particle (H3,H4,H2A,H2B) and 146 BP Long DNA Fragment	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGAATTCAGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[31]
1EQZ	146	X-ray Structure of the Nucleosome Core Particle at 2.5 A Resolution	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[28]
1F66	146	2.6 A Crystal Structure of a Nucleosome Core Particle Containing the Variant Histone H2A.Z	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[30]
1ID3	146	Crystal Structure of the Yeast Nucleosome Core Particle Reveals Fundamental Differences in Inter-Nucleosome Interactions	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[48]
1KX3	146	X-Ray Structure of the Nucleosome Core Particle, NCP146, at 2.0 A Resolution	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGAATTCAGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[41]
1KX4	146	X-Ray Structure of the Nucleosome Core Particle, NCP146b, at 2.6 A Resolution	ATCTCCAAATATCCCTTGCGGATCGTA GAAAAAGTGTGTCAAACTGCGCTATCA AAGGGAAACTTCAACTGAATTCAGTTG AAGTTTCCCTTTGATAGCGCAGTTTGA CACACTTTTTCTACGATCCGCAAGGGA TATTTGGAGAT	[41]
1KX5	147	X-Ray Structure of the Nucleosome Core Particle, NCP147, at 1.9 A Resolution	ATCAATATCCACCTGCAGATACTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGGAATCCAGCT GAACATGCCTTTTGATGGAGCAGTTTC CAAATACACTTTTGGTAGTATCTGCAG GTGGATATTGAT	[41]

PDB ID	Number of Base Pairs	Structure Title	Sequence	Reference
1M18	146	Ligand Binding Alters the Structure and Dynamics of Nucleosomal DNA	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[34]
1M19	146	Ligand Binding Alters the Structure and Dynamics of Nucleosomal DNA	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[34]
1M1A	146	Ligand Binding Alters the Structure and Dynamics of Nucleosomal DNA	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[34]
1P34	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3A	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3B	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3F	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3G	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]

PDB ID	Number of Base Pairs	Structure Title	Sequence	Reference
1P3I	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3K	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3L	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3M	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3O	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1P3P	146	Crystallographic Studies of Nucleosome Core Particles containing Histone 'Sin' Mutants	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[32]
1S32	146	Molecular Recognition of the Nucleosomal 'Supergroove'	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[37]
1U35	146	Crystal Structure of the Nucleosome Core Particle Containing the Histone Domain of MacroH2A	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[44]

PDB ID	Number of Base Pairs	Structure Title	Sequence	Reference
1ZLA	146	X-ray Structure of a Kaposi's Sarcoma Herpesvirus LANA Peptide Bound to the Nucleosomal Core	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[50]
2CV5	146	Crystal Structure of Human Nucleosome Core Particle	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGAATTCAGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[27]
2F8N	146	2.9 Angstrom X-ray Structure of Hybrid MacroH2A Nucleosomes	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[39]
2FJ7	147	Crystal Structure of Nucleosome Core Particle Containing a Poly (dA.dT) Sequence Element	ATCAATATCCACCTGCACATTCTACCA AAAGTGTCAAAAAAAAAAAAAAAATCA TGATAAGCTAATTTGGCTGACTCAGCT GAACATGCCTTTTGATGGAGCAGTTTC CAAATACACTTTTGGTAGTATCTGCAG GTGGATATTGAT	[38]
2NQB	146	Drosophila Nucleosome Structure	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[29]
2NZD	145	Nucleosome Core Particle Containing 145 bp of DNA	ATCAATATCCACCTGCAGATACTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGAATCAGCTGA ACATGCCTTTTGATGGAGCAGTTTCCA AATACACTTTTGGTAGTATCTGCAGGT GGATATTGAT	[35]
2PYO	147	Drosophila Nucleosome Core	ATCAATATCCACCTGCAGATACTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGGAATCCAGCT GAACATGCCTTTTGATGGAGCAGTTTC CAAATACACTTTTGGTAGTATCTGCAG GTGGATATTGAT	[47]
3A6N	146	The Nucleosome Containing a Testis-Specific Histone Variant, Human H3T	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGAATTCAGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[42]

PDB ID	Number of Base Pairs	Structure Title	Sequence	Reference
3AFA	146	The Human Nucleosome Structure	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGAATTCAGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[42]
3B6F	147	Nucleosome Core Particle Treated with Cisplatin	ATCAATATCCACCTGCAGATACTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGGAATCCAGCT GAACATGCCTTTTGATGGAGCAGTTTC CAAATACACTTTTGGTAGTATCTGCAG GTGGATATTGAT	[40]
3B6G	147	Nucleosome Core Particle Treated with Oxaliplatin	ATCAATATCCACCTGCAGATACTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGGAATCCAGCT GAACATGCCTTTTGATGGAGCAGTTTC CAAATACACTTTTGGTAGTATCTGCAG GTGGATATTGAT	[40]
3C1B	146	The Effect of H3 K79 Dimethylation and H4 K20 Trimethylation on Nucleosome and Chromatin Structure	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[49]
3C1C	146	The Effect of H3 K79 Dimethylation and H4 K20 Trimethylation on Nucleosome and Chromatin Structure	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[49]
3KUY	145	DNA Stretching in the Nucleosome Facilitates Alkylation by an Intercalating Antitumor Agent	ATCAATATCCACCTGCAGATACTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGAATCAGCTGA ACATGCCTTTTGATGGAGCAGTTTCCA AATACACTTTTGGTAGTATCTGCAGGT GGATATTGAT	[36]
3KWQ	146	Structural Characterization of H3K56Q Nucleosomes and Nucleosomal Arrays	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[43]
3KXB	146	Structural Characterization of H3K56Q Nucleosomes and Nucleosomal Arrays	ATCAATATCCACCTGCAGATTCTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCGGAATTCCGCTG AACATGCCTTTTGATGGAGCAGTTTCC AAATACACTTTTGGTAGAATCTGCAGG TGGATATTGAT	[43]

PDB ID	Number of Base Pairs	Structure Title	Sequence	Reference
3LEL	147	Structural Insight into the Sequence-Dependence of Nucleosome Positioning	ATCAATATCCACCTGCAGATACTACCA AAAGTGTATTTGGAAACTGCTCCATCA AATTAAATGTTCTTAAAGGACCTTTAA GAACATTTAATTTGATGGAGCAGTTTC CAAATACACTTTTGGTAGTATCTGCAG GTGGATATTGAT	[45]
3LJA	147	Using Soft X-Rays for a Detailed Picture of Divalent Metal Binding in the Nucleosome	ATCAATATCCACCTGCAGATACTACCA AAAGTGTATTTGGAAACTGCTCCATCA AAAGGCATGTTCAGCTGGAATCCAGCT GAACATGCCTTTTGATGGAGCAGTTTC CAAATACACTTTTGGTAGTATCTGCAG GTGGATATTGAT	[51]
3LZ0	145	Crystal Structures of Nucleosome Core Particle Composed of the Super Strong Positioning '601' Sequence	ATCAGAATCCCGGTGCCGAGGCCGCT CAATTGGTCGTAGACAGCTCTAGCAC CGCTTAAACGCACGTACGCGCTGTCC CCCGCGTTTTAACCGCCAAGGGGATT ACTCCCTAGTCTCCAGGCACGTGTCA GATATATACATCGAT	[33]
3LZ1	145	Crystal Structures of Nucleosome Core Particle Composed of the Super Strong Positioning '601' Sequence	ATCGATGTATATATCTGACACGTGCC TGGAGACTAGGGAGTAATCCCCTTGG CGGTTAAAACGCGGGGGACAGCGCGT ACGTGCGTTTAAGCGGTGCTAGAGCT GTCTACGACCAATTGAGCGGCCTCGG CACCGGGATTCTGAT	[33]
3MVD	147	Crystal Structure of the Chromatin Factor RCC1 in Complex with the Nucleosome Core Particle	ATCGAGAATCCCGGTGCCGAGGCCGC TCAATTGGTCGTAGACAGCTCTAGCA CCGCTTAAACGCACGTACGCGCTGTC CCCCGCGTTTTAACCGCCAAGGGGAT TACTCCCTAGTCTCCAGGCACGTGTC AGATATATACATCCGAT	[46]

4.7 References

[1] B. Alexandrov, N.K. Voulgarakis, K.Ø. Rasmussen, A. Usheva, and A.R. Bishop. (2008). Pre-melting Dynamics of DNA and Its Relation to Specific Functions. *J. Phys.: Condens. Matter,* **21(3)**, 034107.

[2] M. Bleichenbacher, S. Tan, and T.J. Richmond. (2003). Novel Interactions Between the Components of Human and Yeast TFIIA/TBP/DNA Complexes. *J. Mol. Biol*., **332(4)**, 783-793.

[3] M. Bleichenbacher, S. Tan, and T.J. Richmond. (2012). RCSB PDB - Images for 1NVP. Available at *http://www.rcsb.org/pdb/explore/images.do?structureId=1NVP.*

[4] X. Lu, and W. K. Olson. (2003). 3 DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.,* **31(17)**, 5108-5121.

[5] S.E. Stayrook, P. Jaru-Ampornpan, J. Ni, A. Hochschild, and M. Lewis. (2008). Crystal structure of the lambda repressor and a model for pairwise cooperative operator binding. *Nature,* **452(7190)**, 1022-1025.

[6] S.E. Stayrook, P. Jaru-Ampornpan, J. Ni, A. Hochschild, and M. Lewis. (2012). RCSB PDB - Images for 3BDN. Available at *http://www.rcsb.org/pdb/explore/images.do?structureId=3BDN.*

[7] J. Chai, J.W. Wu, N. Yan, J. Massague, N.P. Pavletich, and Y. Shi. (2003). Features of a Smad3 MH1-DNA Complex. Roles of Water and Zinc in DNA Binding. *J. Biol. Chem.,* **278(22)**, 20327-20331.

[8] SJ. Chai, J.W. Wu, N. Yan, J. Massague, N.P. Pavletich, and Y. Shi. (2012). RCSB PDB - Images for 1OZJ. Available at *http://www.rcsb.org/pdb/explore/images.do?structureId=1OZJ.*

[9] W.K. Olson, A.R. Srinivasan, A.V. Colasanti, G. Zhang, and D. Swigon. (2009). *DNA Biomechanics. Handbook of Molecular Biophysics: Methods and Applications (H.G. Bohr, Ed.),* **11**, 361-382.

[10] Z. Shakked, D. Rabinovich, O. Kennard, W.B. Cruse, S.A. Salisbury, and M.A. Viswamitra. (1983). Sequence-dependent Conformation of an A-DNA Double Helix. The Crystal Structure of the Octamer d(G-G-T-A-T-A-C-C). *J. Mol. Biol.,* **166(2)**, 183-201.

[11] H.R. Drew, R.M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R.E. Dickerson. (1981). Structure of a B-DNA Dodecamer: Conformation and Dynamics. *Proc. Natl. Acad. Sci. USA,* **78(4)**, 2179-2183.

[12] A.H. Wang, G.J. Quigley, F.J. Kolpak, J.L. Crawford, J.H. van Boom, G. van der Marel, and A. Rich. (1979). Molecular Structure of a Left-handed Double Helical DNA Fragment at Atomic Resolution. *Nature,* **282(5740)**, 680-686.

[13] A.H. Wang, G.J. Quigley, F.J. Kolpak, J.L. Crawford, J.H. van Boom, G. van der Marel, and A. Rich. (2012). RCSB PDB - Images for 2DCG. Available at *http://www.rcsb.org/pdb/explore/images.do?structureId=2DCG.*

[14] W.K. Olson, M. Bansal, S.K. Burley, R.E. Dickerson, M. Gerstein, S.C. Harvey, U. Heinemann, X.J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.S. Tung, E. Westhof, C. Wolberger, and H.M. Berman. (2001). A Standard Reference Frame for the Description of Nucleic Acid Base-pair Geometry. *J. Mol. Biol.,* **313(1)**, 229-237.

[15] G.M. Blackburn, M.J. Gait, D. Loakes, and D.M. Williams (Eds.). (2006). *Nucleic Acids in Chemistry and Biology, 3rd Edition*. The Royal Society of Chemistry, Cambridge, UK.

[16] P. Cysewski. (2009). The Post-SCF Quantum Chemistry Characteristics of Inter- and Intra-Strand Stacking Interactions in d(CpG) and d(GpC) Steps Found in B-DNA, A-DNA and Z-DNA crystals. *J. Mol. Model,* **15(6)**, 597-606.

[17] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 a Resolution. *J. Mol. Biol.,* **319(5)**, 1097-1113.

[18] K.S. Swinger, K.M. Lemberg, Y. Zhang, and P.A. Rice. (2003). Flexible DNA Bending in HU-DNA Cocrystal Structures. *EMBO J.,* **22(14)**, 3749-3760.

[19] K.S. Swinger, K.M. Lemberg, Y. Zhang, and P.A. Rice. (2012). PD0430: Biological Assembly. Available at *http://ndbserver.rutgers.edu/atlas/xray/structures/P/pd0430/PD0430-biol1.html.*

[20] P.T. Lowary, and J. Widom. (1998). New DNA Sequence Rules for High Affinity Binding to Histone Octamer and Sequence-directed Nucleosome Positioning. *J. Mol. Biol.,* **276(1)**, 19-42.

[21] W.K. Olson, and V.B. Zhurkin. (2011). Working the Kinks Out of Nucleosomal DNA. *Cur. Opin. Struct. Biol.*, **21(3)**, 348-57.

[22] L.A. Britton, W.K. Olson, and I. Tobias. (2009). Two Perspectives on the Twist of DNA. *J. Chem. Phys.,* **131(24)**, 245101.

[23] D. Svozil, J. Kalina, M. Omelka, and B. Schneider. (2008). DNA Conformations and Their Sequence Preferences. *Nucleic Acids Res.,* **36(11)**, 3690–3706.

[24] N.A. Becker, J.D. Kahn, and L. J. Maher III. (2008). Eukaryotic HMGB Proteins as Replacements for HU in E. Coli Repression Loop Formation. *Nucleic Acids Res.,* **36(12)**, 4009–4021.

[25] A. Tsortos, G. Papadakis, and E. Gizeli. (2011). The Intrinsic Viscosity of Linear DNA. *Biopolymers*, **95(12)**, 824–832.

[26] G. Kalosakas, and S. Ares. (2009). Dependence on Temperature and Guanine-Cytosine Content of Bubble Length Distributions in DNA. *J. Chem. Phys*., **130(23)**, 235104.

[27] Y. Tsunaka, N. Kajimura, S. Tate, and K. Morikawa. (2005). Alteration of the Nucleosomal DNA Path in the Crystal Structure of a Human Nucleosome Core Particle. *Nucleic Acids Res.* **33(10)**, 3424-3434.

[28] J.M. Harp, B.L. Hanson, D.E. Timm, and G.J. Bunick. (2000). Asymmetries in the Nucleosome Core Particle at 2.5 A Resolution. *Acta Crystallogr.,* **56(12)**, 1513-1534.

[29] S. Chakravarthy, and K. Luger. Comparative Analysis of Nucleosome Structures from Different Species. *To be Published.*

[30] R.K. Suto, M.J. Clarkson, D.J. Tremethick, and K. Luger. (2000). Crystal Structure of a Nucleosome Core Particle Containing the Variant Histone H2A.Z. *Nat. Struct. Biol.,* **7(12)**, 1121-1124.

[31] K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, and T.J. Richmond. (1997). Crystal Structure of the Nucleosome Core Particle at 2.8 A Resolution. *Nature,* **389(6648)**, 251-260.

[32] U.M. Muthurajan, Y. Bao, L.J. Forsberg, R.S. Edayathumangalam, P.N. Dyer, C.L. White, and K. Luger. (2004). Crystal Structures of Histone Sin Mutant Nucleosomes Reveal Altered Protein-DNA Interactions. *EMBO J.,* **23(2)**, 260-271.

[33] D. Vasudevan, E.Y. Chua, and C.A. Davey. (2010). Crystal Structures of Nucleosome Core Particles Containing the '601' Strong Positioning Sequence. *J. Mol. Biol.,* **403(1)**, 1-10.

[34] R.K. Suto, R.S. Edayathumangalam, C.L. White, C. Melander, J.M. Gottesfeld, P.B. Dervan, and K. Luger. (2003). Crystal Structures of Nucleosome Core Particles in Complex with Minor Groove DNA-binding Ligands. *J. Mol. Biol.,* **326(2)**, 371-380.

[35] M.S. Ong, T.J. Richmond, and C.A. Davey. (2007). DNA Stretching and Extreme Kinking in the Nucleosome Core. *J. Mol. Biol.,* **368(4)**, 1067-1074.

[36] G.E. Davey, B. Wu, Y. Dong, U. Surana, and C.A. Davey. (2010). DNA Stretching in the Nucleosome Facilitates Alkylation by an Intercalating Antitumour Agent. *Nucleic Acids Res.,* **38(6)**, 2081-2088.

[37] R.S. Edayathumangalam, P. Weyermann, J.M. Gottesfeld, P.B. Dervan, and K. Luger. (2004). Molecular Recognition of the Nucleosomal 'Supergroove'. *Proc. Natl. Acad. Sci. USA,* **101(18)**, 6864-6869.

[38] Y. Bao, C.L. White, and K. Luger. (2006). Nucleosome Core Particles Containing a Poly(dA.dT) Sequence Element Exhibit a Locally Distorted DNA

Structure. *J. Mol. Biol.* **361(4)**, 617-624.

[39] S. Chakravarthy, and K. Luger. Nucleosomes Containing the Histone Domain of MacroH2A: In Vitro Possibilities. *To be Published.*

[40] B. Wu, P. Droge, and C.A. Davey. (2008). Site Selectivity of Platinum Anticancer Therapeutics. *Nat. Chem. Biol.,* **4(2)**, 110-112.

[41] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 A Resolution. *J. Mol. Biol.,* **319(5)**, 1097-1113.

[42] H. Tachiwana, W. Kagawa, A. Osakabe, K. Kawaguchi, T. Shiga, Y. Hayashi-Takanaka, H. Kimura, and H. Kurumizaka. (2010). Structural Basis of Instability of the Nucleosome Containing a Testis-specific Histone Variant, Human H3T. *Proc. Natl. Acad. Sci. USA,* **107(23)**, 10454–10459.

[43] S. Watanabe, M. Resch, W. Lilyestrom, N. Clark, J.C. Hansen, C. Peterson, and K. Luger. (2010). Structural Characterization of H3K56Q Nucleosomes and Nucleosomal Arrays. *Biochim. Biophys. Acta.,* **1799(5-6)**, 480-486.

[44] S. Chakravarthy, S.K. Gundimella, C. Caron, P.Y. Perche, J.R. Pehrson, S. Khochbin, and K. Luger. (2005). Structural Characterization of the Histone Variant MacroH2A. *Mol. Cell. Biol.,* **25(17)**, 7616-7624.

[45] B. Wu, K. Mohideen, D. Vasudevan, and C.A. Davey. (2010). Structural Insight into the Sequence Dependence of Nucleosome Positioning. *Structure,* **18(4)**, 528-536.

[46] R.D. Makde, J.R. England, H.P. Yennawar, and S. Tan. (2010). Structure of RCC1 Chromatin Factor Bound to the Nucleosome Core Particle. *Nature,* **467(7315)**, 562-566.

[47] C.R. Clapier, S. Chakravarthy, C. Petosa, C. Fernandez-Tornero, K. Luger, and C.W. Muller. (2007). Structure of the Drosophila Nucleosome Core Particle Highlights Evolutionary Constraints on the H2A-H2B Histone Dimer. *Proteins,* **71(1)**, 1-7.

[48] C.L. White, R.K. Suto, and K. Luger. (2001). Structure of the Yeast Nucleosome Core Particle Reveals Fundamental Changes in Internucleosome Interactions. *EMBO J.* **20(18)**, 5207-5218.

[49] X. Lu, M.D. Simon, J.V. Chodaparambil, J.C. Hansen, K.M. Shokat, and K. Luger. (2008). The Effect of H3K79 Dimethylation and H4K20 Trimethylation on Nucleosome and Chromatin Structure. *Nat. Struct. Mol. Biol.,* **15(10)**, 1122-1124.

[50] A.J. Barbera, J.V. Chodaparambil, B. Kelley-Clarke, V. Joukov, J.C. Walter, K. Luger, and K.M. Kaye. (2006). The Nucleosomal Surface as a Docking Station for Kaposi's Sarcoma Herpesvirus LANA. *Science,* **311(5762)**, 856-861.

[51] B. Wu, and C.A. Davey. (2010). Using Soft X-Rays for a Detailed Picture of Divalent Metal Binding in the Nucleosome. *J. Mol. Biol.,* **398(5),** 633-640.

Chapter 5: 3DNAdesigner®

5.1 Introduction to 3DNAdesigner®

Comprehension of DNA and protein/drug-bound DNA in biological processes, such as transcription, requires an understanding of their structure. When analyzing large DNA systems the complexity of the molecular dynamics in such structures makes it necessary to simplify models of the molecules due to the vast computational requirements of all-atom based simulations. In this thesis, the model consists of two primary simplifications. First, it focuses on the interactions between adjoining base pairs, not all atoms. Second, the base pairs are modeled as rectangular slabs that are subject to small elastic deformations. These assumptions allow the use of a base-pair level theory of DNA elasticity [1].

This chapter describes a user-friendly version of the methodology originally developed by David Swigon to determine the equilibrium configurations of spatially constrained fragments of double-helical DNA [1]. The new program enables a wider audience to build models of superhelical DNA. In order to address this need we identified a core set of input variables needed to generate meaningful models, without requiring the user to modify the source code, and to satisfy a wide range of interests. The resulting set of input data include (i) the sequence of the structure, (ii) the spatial boundary conditions placed on the ends of the molecule, (iii) the decision of whether or

not to bind one or more proteins or drugs to sites along the sequence, and (iv) the identities of the latter proteins or drugs. In order to create the most straightforward and user-friendly experience possible with our new modeling package, we decided to capture the capabilities of our equilibrium configuration application in a graphical user interface, or GUI. The GUI, called 3DNAdesigner, expertly guides the users through the process of inputting the details required to generate a meaningful model by presenting them with a series of wizards that take them through the process step by step.

Once the user has provided the proper input, the model is generated. However, in order to create a structure with the ends held in a particular spatial arrangement, such as a closed circle or protein-mediated loop, moments and forces must be applied to one of the ends. Determination of these moments and forces is challenging since trial-and-error sampling of possible starting values can use up a lot of computer time and/or a reliable 'guesstimate' of a starting structure requires the random sampling of chain configurations.

The output of 3DNAdesigner® includes the coordinates of the constituent base pairs (in the form of the origins and local coordinate axes of each pair) and detailed data on the topology (Tw^{SP}, Tw^{SC}, Wr, Lk), the total elastic energy E_{Total} of the fragment, and the six base-pair-step parameters (Shift, Slide, Rise, Tilt, Roll, Twist). 3DNAdesigner® also provides the user with a three-dimensional visualization of the resulting structure so that (s)he can investigate it in greater detail.

5.1.1 Minimum Energy Configurations

Every configuration generated by 3DNAdesigner is a minimum energy configuration. The energy that we use in our representation of double-helical DNA is an elastic energy. Each minimum-energy configuration generated represents a locally stable configuration. This minimum-energy configuration is only the lowest energy that could be found for the end conditions that are supplied by the user, but it may not be as low an energy configuration as that generated with a different set of end conditions.

Relative base-pair step orientations are characterized using the step parameters described in Section 1.4.1. The rest state parameters, $\overline{\theta}_i$ and $\overline{\rho}_i$, are averaged values of dimers taken from many X-ray crystal structures [2]. The global structure and energies of the molecule are described by the set of θ_i values, the angular parameters, and the set of ρ_i values, the translational parameters, associated with a given configuration. We start from a preexisting configuration and find new step parameters in the search for an equilibrium configuration under certain boundary constraints. The extent to which the step parameters are allowed to change is determined by the elastic moduli.

The mechanical elastic equilibrium is found from an iterative process starting from knowledge of the rest values of the intrinsic variables (step parameter and force constants). When boundary conditions are imposed on the ends of the DNA strands it becomes necessary to account for the forces and moments on the joining ends in order to

find a mechanically stable state. The iterative process that tries to find a mechanical stable state is based on the Newton-Raphson method for finding the roots of the quadratic energy function of Eq. 5.1.1.2 [3]. The initial guesses that are given to the Newton-Raphson function are based on moments and forces that are placed on the ends of the DNA segment to find a closed or anchored boundary condition. If the initial guess provided results in a converged configuration then a locally minimum-energy configuration has been found. If no convergence is achieved, then a local energy minima has not been found. Below are the equations used to calculate the elastic energy which is contingent on the deviation from the rest state intrinsic parameters [4,5]. The total elastic energy Ψ is the sum of all the contributions of energy from each step along the sequence, ψ^n.

$$\Psi = \sum_{n=1}^{N} \psi^n \qquad \textit{Eq. 5.1.1.1}$$

Each energy contribution is a quadratic function of the step parameters θ_i and ρ_j and the elasticity moduli F, G and H.

$$\psi^n = \frac{1}{2} F_{ij}^n (\Delta\theta_i^n)(\Delta\theta_j^n) + G_{ij}^n (\Delta\theta_i^n)(\Delta\rho_j^n) + \frac{1}{2} H_{ij}^n (\Delta\rho_i^n)(\Delta\rho_j^n) \qquad \textit{Eq. 5.1.1.2}$$

The energies are based on the difference between the final configuration values and the intrinsic values of the stress free state, $\bar{\theta}_i^n$ and $\bar{\rho}_j^n$.

$$\Delta\theta_i^n = \theta_i^n - \bar{\theta}_i^n, \ \Delta\rho_i^n = \rho_i^n - \bar{\rho}_i^n \qquad \textit{Eq. 5.1.1.3}$$

5.1.2 User Friendly Code

The capabilities and results of the program have generated greater than anticipated interest from outside users. Therefore, a practical aim was to make the program user friendly. An intuitive and easy to follow graphical user interface has been developed for a non-expert researcher to use the program. After a configuration is found, the results are presented with multiple graphing capabilities. Features of the program include providing ways for the user to input a particular DNA sequence, apply forces and moments at connecting ends of a plasmid, use certain boundary conditions, insert simple intercalating drugs, or add bound proteins to the sequence.

MATLAB is used for all computations. The original program used for the calculation of equilibrium configurations was created by Professor David Swigon, previously a post-doctoral associate in the Olson group [6]. The program has now been improved upon and added to. To make the program useful to researchers it will continuously need to be updated with improved ways to calculate the effects of the moments and forces on the boundaries, energies, and geometric parameters of the whole structure and to address unanticipated structural problems.

At the present time, the intrinsic values and force constants in the sequence-dependent elastic potentials describing the DNA base-pair steps, or dimers, are found by analyzing the arrangements of successive base-pair steps in high-resolution DNA structures in any sequence context. Dr. Andrew Colasanti has also gathered the intrinsic

values and force constants of the dimer steps in all possible tetramer settings. Dimer parameters will be calculated from the averaged values and covariance in the same way as described in Section 1.4.1 and will be incorporated in calculations so as to include the flanking 3´ and 5´ sequences of the 256 possible tetramer sequences only 136 tetrameric sequences are unique due to redundancies in complimentary base-pairs [7].

5.1.2.1 Feature Selection

Determining the required flexibility of the programs was one of the prime tasks of making the software usable for all types of end users, including biologists. Our programs have a huge potential capability, similar to the potential locked up in many people. If one does not know how to access the locked potential, or the best way to retrieve what one wants, it can be big waste. All too often programs are not made to run to their full potential. The equilibrium program that was originally developed by Dr. David Swigon and now adapted by me has tremendous potential. The only way to take advantage of that potential prior to the development of the GUI noted above was to know how to use MATLAB, and to understand the logic of the code and the sometimes detailed and sometimes nonexistent comments. The more options one can give a program, the more flexible this program will be in terms of the end user. This flexibility translates into a more accessible program. This accessibility when paired with a detailed and thorough program yields a very powerful application for the end users.

5.1.2.2 Stabilizing 3DNAdesigner®

The process of converting the compiled program into a GUI required caution. Simply because a program runs without crashing or less significant, generating errors, does not mean that it can be flawlessly ported. Experience shows that this is not the point the rigorous testing should be stopped. The source of problems can come from anywhere. This is why patience and good detective work pay off in program development. The issues that we have come across to date have come from computer software configuration issues, data file input, data file output, and command line data input syntax.

5.2 The Use of MATLAB

5.2.1 The MATLAB compiler

One primary objective in the implementation of 3DNAdesigner® was to take a piece of software that was originally written by David Swigon using MATLAB and to make it into a stand-alone user-friendly software package. To accomplish this goal several alternatives were investigated. Initially the code was reviewed for porting from MATLAB into another language, such as C/C++, that could be easily used to build stand-alone packages. During the review of the code two major discoveries were made. First was the strong dependence on the MATLAB development environment and the MATLAB routines used in the existing code, in particular the routines for visualization.

Second was the discovery of the MATLAB Compiler, a software package designed to enable the building of MATLAB source or m-files into independent software packages [8]. Based on the complexity of porting the code to another language and the simplicity offered by compilation of the code into a stand-alone application, the MATLAB Compiler was chosen as the optimal solution. The MATLAB Compiler takes m-files, which require a working MATLAB package, and turns them into stand-alone applications which are capable of running without requiring the user to have a MATLAB license or package.

The MATLAB Compiler requires inclusion of the MATLAB Compiler Runtime, or in order for a stand-alone program to run without MATLAB. The MCR runtime engine must be installed on the target machine before a stand-alone program can run. The MCR has thus been included in the packages made for distributing 3DNAdesigner®. The MCR is taken from the MATLAB installation on the development machine where 3DNAdesigner® is coded and compiled. Since the end-compiled program relies on the MCR to work, and MATLAB markets their Compiler to generate MATLAB-independent final software, MATLAB provides the MCR royalty-free to the end user. This adds another level of complexity to the final 3DNAdesigner® software package since the MCR needs to be distributed within the final version as well.

When 3DNAdesigner® is installed, the user is prompted by the MATLAB installer to install the MCR and is allowed to select the location of installation. If the

MCR needs to be shared by multiple users on the same machine, there may be permission issues which can be taken care of by the end user having a system administrator install the 3DNAdesigner® package. Once the 3DNAdesigner® installation is completed and after the MCR has been properly placed, the end user's environment needs to be able to access the MCR for 3DNAdesigner® to work. In the Linux operating system, this is handled by the 3DNAdesigner® application taking the path to the MCR as its first argument. This is documented in the README included with the installer, and shows as an error message if 3DNAdesigner® is run without the proper path to the MCR.

Development of 3DNAdesigner® has primarily been done using the Linux operating system. With minor changes, the code has additionally been ported to run as a stand-alone application on Microsoft Windows. This flexibility to easily port from Linux to Windows was accomplished through the use of the MATLAB compiler suite, which allows the same code to compile under each operating system, with very few changes related to file and directory locations that are specific to each operating system. In the future it should be possible to add operating system support for 3DNAdesigner® under the Mac OS as well. For each of the supported operating systems customized build and installation scripts were written to provide users with a software package that is easy to install and use on their machines. The installer packages not only handle the installation of 3DNAdesigner, but also the installation of the supporting 3DNA [9] and MCR software packages.

5.2.2 GUI Creation with MATLAB GUIDE

Another major goal of this research was to make the code developed by Swigon and Britton usable by many scientists in different fields with different levels of computer experience. The final product is thus developed for the most inexperienced user within reason. In the previous section we discussed taking MATLAB code and successfully compiling it with the MATLAB Compiler, creating a MATLAB independent program. This is the first step to providing a wide range of users access to our code. The next major step is to make that code easily accessible and usable. The best way to do this is by not requiring the end user to understand how to run the code from the command line, but rather letting the users have the capability of using a well designed Graphical User Interface, or GUI.

There are two fundamental parts of designing a standard GUI. The first is to make a graphic that has buttons, drop-down menus, input areas, etc. Based on the decision to continue development of 3DNAdesigner® with MATLAB, the module called GUIDE was investigated for development of Graphical User Interfaces based on MATLAB. GUIDE was determined to be an extremely useful and user friendly software module whose purpose is to help the developer create a GUI. It provides the capability to create a backdrop for the GUI, along with inserting buttons, drop-down menus, input areas, scroll bars and the like, which are called widgets. GUIDE provides the ability to group certain objects, change the size, change the fonts, colors, and more very easily.

Each object created using GUIDE can be given a brief description and a name. Writing a name and a description for each object ensures that the developer knows what he/she is dealing with. Without a description of each object within the GUI there is an extra level of difficulty in determining which object is aligned with what function.

The second part to creating a working GUI is to write code to make the objects that are inserted into the GUI have meaning. For instance, if an Enter button is put as an object into the GUI, the program needs to understand what should be done when that button is pressed. That is, should the program read in data or execute a computation? Without the associated code, the Enter button is just a useless button which has the word Enter written upon it. The very useful GUIDE module not only donates the graphics to the utility, but also generates some very basic code, which is exported into an m-file. This m-file is where the new coding takes place to give the widget significance within the overall application.

Still using the Enter button example, generally once all of the choices are made on the user interface, all the input should be collected and applied in the desired program as specified. Therefore, the Enter button code must get the requisite information from the other widgets and pass that information to the desired end program. However, not every user will enter information into the user interface in the proper format. This error may lead to strange results or output that may only confuse the end user. To help reduce the number of such problems, it is best to inform the user in easily understandable terms

what was wrong with their input. The best place to do this is in the m-file generated by GUIDE. Although, it is not possible to anticipate every type of error, thorough testing helped identify many common errors that the 3DNAdesigner® GUI would catch. As with any software package, there should be updates and revisions as issues come forth in the future.

Considerable time and effort were invested in the general look and feel of the 3DNAdesigner® GUI. The initial look of the user interfaces was just a proof of concept. Subsequent changes included the background colors, font, object placement, and addition of various types of widgets. The final step in making a useful GUI was taking an in-depth look at what capabilities the 3DNAdesigner® software package should have, and determining which ones were practical. With the final feature set decided upon, the original MATLAB m-file code was modified so that it accommodated these features and functioned with the GUI.

5.3 Organization of 3DNAdesigner®

5.3.1 DNA Configurations

The 3DNAdesigner® program was designed to aid in the creation of spatially constrained, three-dimensional equilibrium configurations of DNA with information and criteria supplied by the user. The application allows everyone from knowledgeable experts to novice users to generate models of free and spatially anchored DNA of

arbitrary sequence with and without proteins or drugs bound at preselected locations. This is accomplished through a series of wizards, which depend upon the selections made each time 3DNAdesigner® is run, to create the user's DNA configuration. The selections produce a file with the starting conditions for the molecule the user wishes to model, including the sequence selection, the proteins bound, the boundary conditions, and the style in which the molecular structure will be represented. Once a desired configuration has been created successfully, the user is able to reuse it as is, as well as edit it to change some aspects of the molecule that are stored in the chosen configuration.

The main screen of 3DNAdesigner® includes the option to generate a new configuration, edit an old one, or look at some output, as highlighted in Figure 5.3.1.1. 3DNAdesigner® also guides the user along the DNA chain-construction process with a 'Help' button feature that shows how to proceed through most windows in the GUI, as seen in Figure 5.3.1.1(f).

To start with a new configuration, the user simply selects the 'Generate a New Configuration' and clicks the 'Enter' button from Figure 5.3.1.1(a). This will then prompt the user to create a configuration name, which is the file name that will be assigned to their inputs. Next the user will be walked through a series of wizards that collect the needed information to generate a DNA configuration that will be stored in the file. Alternatively, the user may select the 'Reuse an OLD Configuration' option, which allows further work with a previously generated configuration. The user will be prompted to

choose the file that contains the desired configuration. The configuration, if converged, will be displayed with the visualization tool. If the configuration did not converge previously, then the user will be given an error message and will not be able to open that file. The last way the user may start work with a configuration is to select the 'Edit an OLD Configuration' option, which makes it possible to work from an existing configuration, whether it has converged or not. When the user chooses this option, a prompt appears to select the desired configuration file, and to assign a name for the new configuration that will be generated. The user is then walked through the wizards again, but this time the wizards default to the values used to generate the selected configuration. This allows the user to repeat many of the steps in the DNA model building process and to make slight changes to certain conditions in the calculation. For example, the user could choose to keep everything the same about the DNA configuration, but change the protein that binds to it or the site of protein binding. This would enable researchers to see how the selected constraints affect the DNA configuration.

When a DNA configuration has been successfully defined, details about the configuration can be seen in one of two areas highlighted in Figure 5.3.1.1(b) and (c), i.e., text boxes in the middle and lower half of the main screen. The first area, the 'DNA System and Anchoring Conditions' text box, displays the data provided by the user to establish the starting values for the DNA configuration. This includes data such as the configuration file name and location, whether the file contains a new or old

configuration, the nature of the imposed boundary conditions and the initial moments and forces, whether there are proteins bound, and the DNA sequence. The second area, the 'Output Gazer' text box, displays the data generated from the user inputs to create a configuration. This includes the molecular topology, energy scores, protein binding site, quantities that describe the supercoiling, such as twist and writhe, DNA sequence, and whether or not the configuration converges.

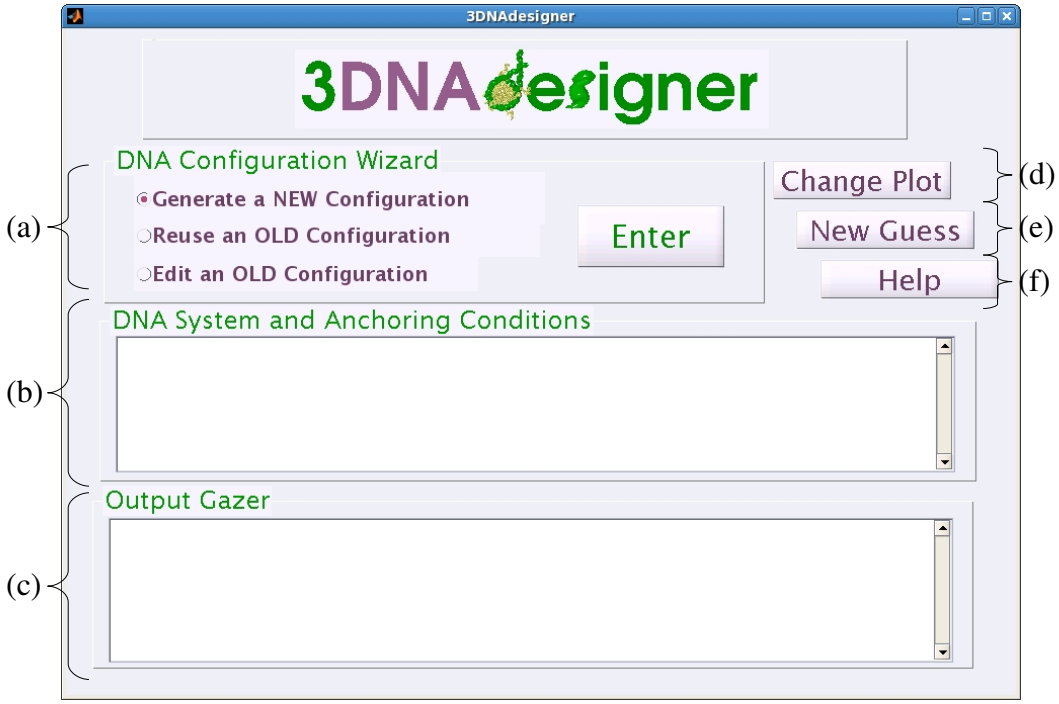


Figure 5.3.1.1: The main 3DNAdesigner® screen allows the user to (a) create and modify existing DNA configurations, (b) view the conditions of the current computation, (c) review the output of running the computation, (d) change the visualization type used to display a configuration, (e) change the guess used for the initial restrictive end condition requirements, and (f) receive help on how to use 3DNAdesigner®.

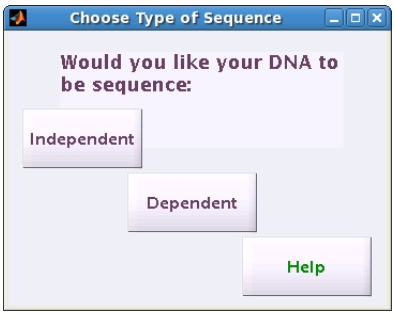
When the computation has converged and the user has succeeded in creating a DNA model, 3DNAdesigner® will display a new window with the type of visualization of the molecule specified by the user. The user may also view the converged configuration using a different visualization method by clicking the 'Change Plot' button shown in Figure 5.3.1.1(d). This simply walks the user through the same wizard originally used to select the current visualization.

In many cases, the computation does not converge, leaving the user with two options: either start anew or change the guess for the moments and forces by clicking the 'New Guess' button shown in Figure 5.3.1.1(e). In some cases when the computation has converged a user may wish to generate a new configuration with a different shape and energy by changing the moments and forces used in the initial guess. To do this the user can again use the 'New Guess' button to change the values from those assigned to the converged configuration.

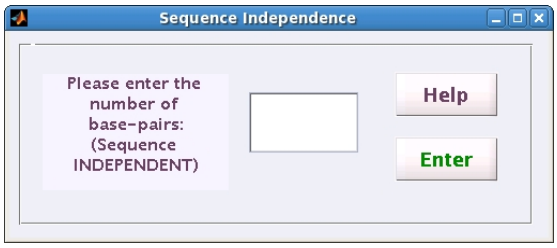
5.3.2 Sequence Dependence

The first wizard allows the user to establish the sequence of base pairs in the model. The sequence can affect the overall shape of the DNA in its minimum-energy state since the potential function depends on the identities of successive base pairs. The user has the option of treating the DNA as sequence-independent or sequence-dependent, as shown in Figure 5.3.2.1(a). The choice of the Independent or Dependent button, prompts the user for information that depends on the selection in order to generate the

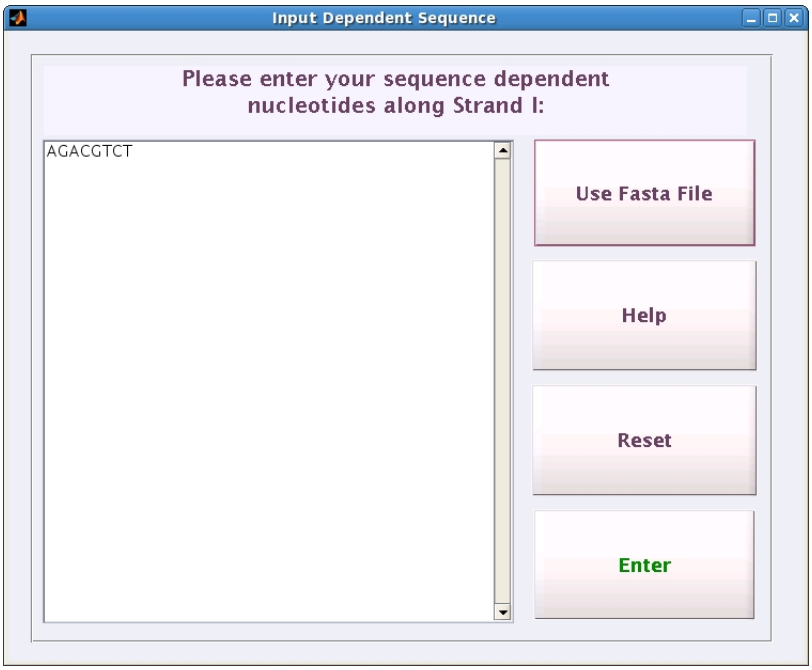
desired configuration.



(a)



(b)



(c)

Figure 5.3.2.1: The sequence type wizard, shown in (a), allows the user to choose between a sequence-independent option (b), which requires entering the number of base pairs, and a sequence-dependent option (c), which allows the user to enter a specific sequence.

If the user chooses to generate a sequence-independent DNA configuration, then a prompt appears in the new dialog shown in Figure 5.3.2.1(b). The sequence-independent DNA selection uses the mean step parameters and force constants of all 16 types of base-pair steps. By definition, the Independent selection assumes that all base pairs are identical, and as such the user only needs to supply how many base pairs comprise the sequence (e.g., 75). Consequently, when the configuration is determined and the details of the molecule are displayed, the Independent DNA sequence will display the letter 'MMMM....', where 'M' represents an averaged base pair, instead of a combination of 'A', 'T', 'G' and 'C'.

However, if the user decides to generate a sequence-dependent DNA configuration, a different prompt appears, as shown in Figure 5.3.2.1(c). Selection of a sequence-dependent DNA configuration requires the user to enter the series of bases in the text box. The sequence must be entered using the capital letters 'A', 'T', 'G', 'C' for the bases adenine, thymine, guanine, and cytosine, respectively. The sequence is currently limited to a maximum length of 500 bases. Additionally, if the user wishes to start over, the 'Reset' button, in the right hand side of the 'Input Dependent Sequence' wizard, can erase anything entered about the sequence in the text box.

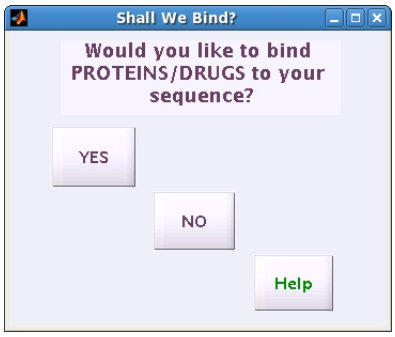
An alternative method for entering a sequence-dependent configuration is for the user to choose the 'Use Fasta File' button. This option allows the user to upload sequence files from a file source. A FASTA file is a standard file format used to represent DNA

sequence. The file consists of a description line prefaced by a '>' and multiple lines containing sequence data encoded in either nucleic acid residue or amino acid residue codes [10]. Once the file is opened, the sequence will appear in the text box of the 'Input Dependent Sequence' window. The user may tailor the uploaded sequence by editing the text displayed in the window.

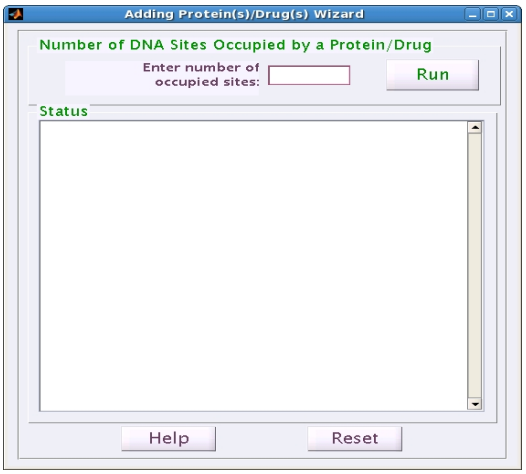
5.3.3 Protein Binding

Once the user has defined the DNA sequence, a prompt appears to choose whether or not to bind a protein or drug to the sequence. The process of binding a protein or drug starts with a simple 'Yes' or 'No', as shown in Figure 5.3.3.1(a). A 'No' moves the user to the next wizard, which requests the choice of plot to display. The configuration generated is that of a naked DNA double helix that does not include any drug or protein. However, 3DNAdesigner® provides several steps related to binding proteins and drugs in order to generate the relevant configurations that help to understand the global effects of these types of interactions and constraints on the DNA sequence.

(a)



(b)



(c)

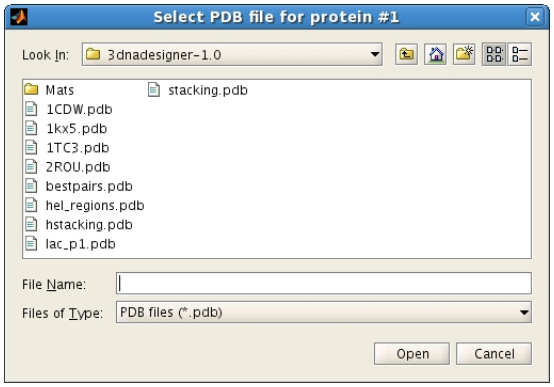
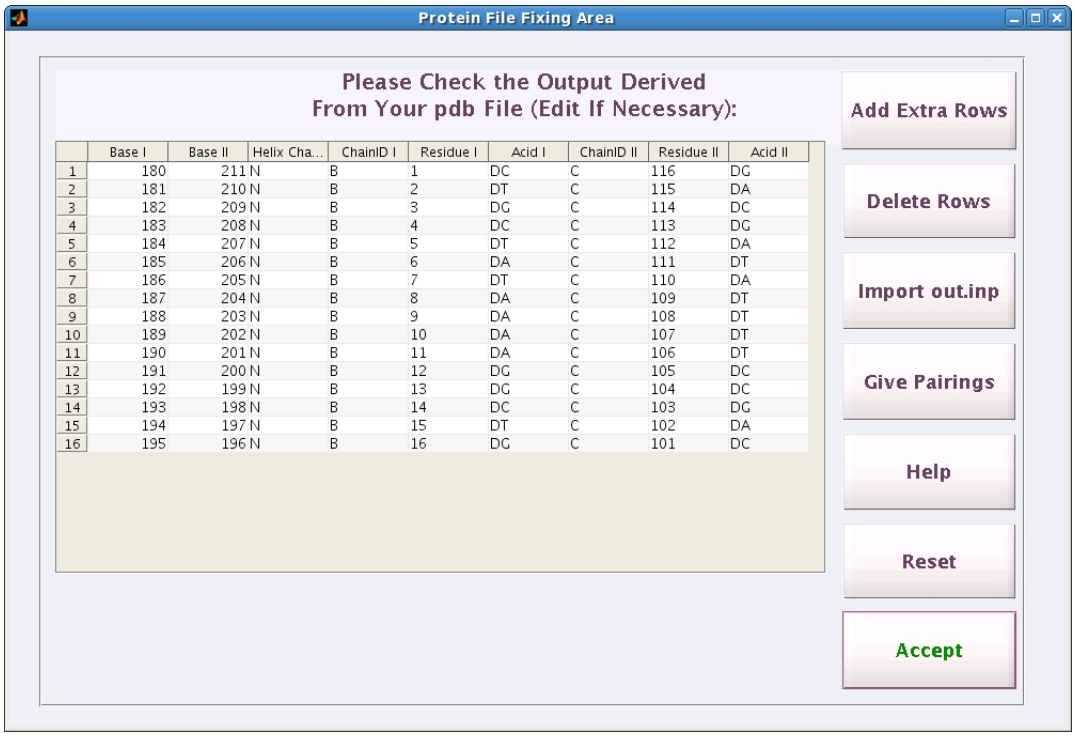


Figure 5.3.3.1: (a) The protein/drug binding wizard allows the user to select whether or not to bind a protein or drug to DNA. When the user chooses "Yes", then additional prompts allow for (b) the quantity and (c) the identities of the protein/drugs to bind. The location of the binding sites are specified through a separate wizard, as seen in Figure 5.3.3.3.

If the user chooses 'Yes' to binding a drug or protein from the dialog in Figure 5.3.3.1(a), the next step is to input the total number of protein(s) and/or drug(s) that will occupy sites on the DNA sequence, and then click the 'Run' button, as seen in Figure 5.3.3.1(b). The 'Adding Protein(s)/Drug(s) Wizard' additionally allows the user to choose which protein or drug they would like to bind to the sequence using the dialog seen in Figure 5.3.3.1(c). Each protein or drug the user intends to bind to the DNA must reside in the 3DNAdesigner® directory in Protein Data Bank (PDB) format [11]. The 'Adding Protein(s)/Drug(s) Wizard' presents the user with a dialog that allows for selection of these files.

In addition to the PDB files included in the 3DNAdesigner® package, the user can download other files from the Protein Data Bank (http://www.rcsb.org/), as well as through the TwiDDL website (http://twiddl.rutgers.edu/search.php). When downloading PDB files not included in 3DNAdesigner®, the protein must have continuous stretches of double-helical DNA. If there is more than one protein or drug chosen to bind to the sequence, the user will be prompted to repeat all of the relevant steps in order for 3DNAdesigner® to bind them properly. This is true even if the same protein or drug is bound repeatedly. The locations of these binding sites are set in a separate wizard, shown in Figure 5.3.3.3, and will be discussed in more detail later in this section.



	Base I	Base II	Helix Cha...	ChainID I	Residue I	Acid I	ChainID II	Residue II	Acid II
1	180	211	N	B	1	DC	C	116	DG
2	181	210	N	B	2	DT	C	115	DA
3	182	209	N	B	3	DG	C	114	DC
4	183	208	N	B	4	DC	C	113	DG
5	184	207	N	B	5	DT	C	112	DA
6	185	206	N	B	6	DA	C	111	DT
7	186	205	N	B	7	DT	C	110	DA
8	187	204	N	B	8	DA	C	109	DT
9	188	203	N	B	9	DA	C	108	DT
10	189	202	N	B	10	DA	C	107	DT
11	190	201	N	B	11	DA	C	106	DT
12	191	200	N	B	12	DG	C	105	DC
13	192	199	N	B	13	DG	C	104	DC
14	193	198	N	B	14	DC	C	103	DG
15	194	197	N	B	15	DT	C	102	DA
16	195	196	N	B	16	DG	C	101	DC

Figure 5.3.3.2: The Protein File Fixing Area wizard allows users to review and edit the 3DNA output. In most cases, the user can click the 'Accept' button without modifications, but in some rare cases changes can be made here before continuing with 3DNAdesigner®.

After the user selects the PDB files of the protein(s) or drug(s) to bind to the DNA, these files will be analyzed with the 3DNA suite of programs [12] to translate the coordinates of the atoms in the PDB file into usable information for 3DNAdesigner®. However, because there can be problems with some PDB files as described below, the user will be asked to review the output from 3DNA in an easy-to-use, table as seen in Figure 5.3.3.2. In most cases, the output displayed in the table will not need to be modified, and the user can simply select 'Accept' to use the data as displayed in the table.

Sometimes the PDB files contain some errors or anomalies, such as mismatched base pairs or missing steps. These issues need to be addressed in order for the protein or

drug to bind to the DNA within 3DNAdesigner®. Some things the user can check for:

(a) Make sure the Base I column (the base numbering of nucleotides along strand I) does not skip numbers

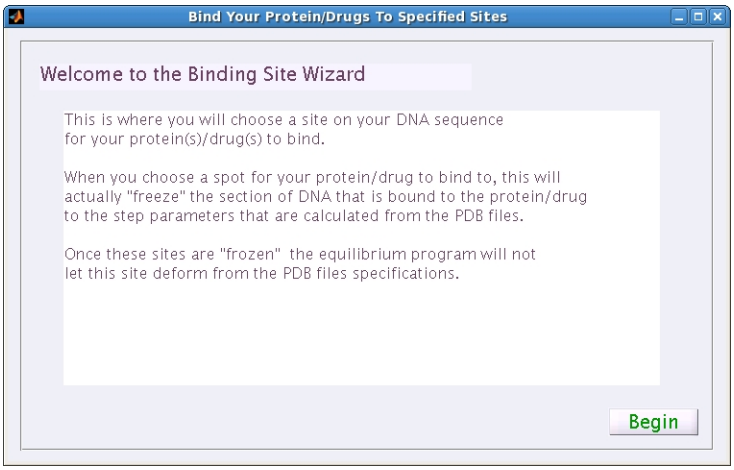
(b) Check the same for the Base II column.

(c) Look at the 'Helix Change' column. If a 'Y' appears anywhere, there may be one or more helices. If so, the user may want to delete all the rows above or the ones below the break in numbers. The box at the bottom of the screen provides a little more help in this regard.

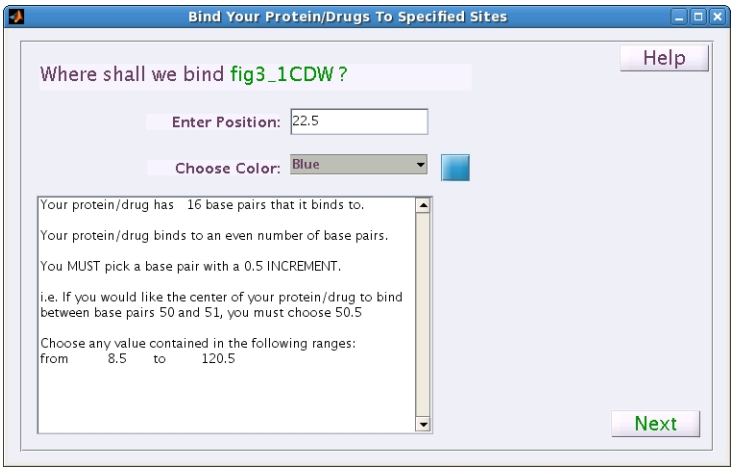
The numbers in the first two columns can also be changed. These two columns can be edited by clicking on the appropriate box to change the number in that box. If something more complicated than changing a single number is needed, the buttons on the right hand side of the wizard can be used. The 'Add Extra Rows' button allows the addition of extra rows anywhere in the table. The 'Delete Rows' button allows deletion of rows in the table. The 'Import out.inp' button is for users familiar with 3DNA who wish to manually generate an out.inp file for use by 3DNAdesigner®. The 'Give Pairings' button allows the user to examine and modify the bases that 3DNA has identified as paired. The 'Reset' button will return the values in the table back to what was originally provided before any changes were made. The 'Accept' button moves to the next step in the processing by accepting whatever is currently displayed in the table as the DNA input for connecting the protein to the user's sequence. Most users will just use the 'Accept'

button without editing anything in the table.

(a)



(b)



(c)

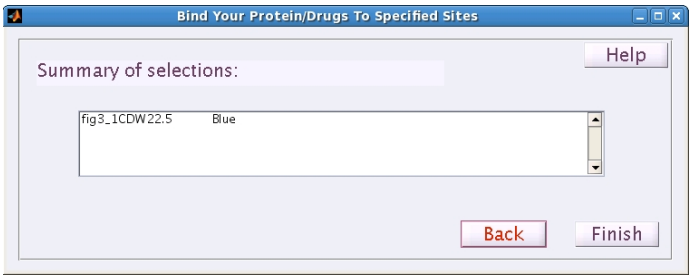


Figure 5.3.3.3: (a) The 'Binding Site' wizard opening screen starts the process of positioning proteins or drugs on DNA with the 'Begin' button. (b) The next screen appears once for each iteration through all of the selected protein/drugs, and allows for selection of both a viable position on the DNA, and a color to highlight the given protein/drug. (c) The final screen of the 'Binding Site' wizard allows the user to review the selected ligand and position, go back to make changes, or continue with these choices.

After the geometry of the protein and drug DNA fragments have been processed for binding, a new wizard appears, which allows for the selection of the binding sites of the proteins or drugs on the DNA sequence. The 'Binding Site Wizard', as seen in Figure 5.3.3.3, takes the specific nucleotides where the proteins or drugs are located. 3DNAdesigner® binds proteins or drugs on the sequence by freezing the part of the DNA that is bound by protein or drug in the exact arrangement found in the specific X-ray crystal or NMR structure. The information is fed to the program in the form of the DNA step parameters obtained by the 3DNA processing of the selected PDB files. Once these sites are frozen, the equilibrium program within 3DNAdesigner® will not let any of the sites deform from the specified values.

As can be seen in Figure 5.3.3.3(b), the binding site for each protein or drug is entered as a numerical position where the center of the DNA bound to the drug or protein should be on the DNA sequence. The position or binding site is specified in terms of the number of base pairs along the sequence. If the DNA bound to the drug or protein contains an odd number of base pairs, the binding site is an integer because one of the base pairs is at the center of the sequence. If the bound DNA contain an even number of base pairs, the center lies on the mid-frame axis half way between the central base pairs, i.e., central base-pair step. As shown in Figure 5.3.3.4(a), a binding site specified as 10 will center the protein or drug at the 10^{th} base pair if the ligand binds an odd number of base pairs. For example, if the drug or protein binds five base pairs in the DNA, two

base pairs lie on either side of the specified numerical site and the central third base pair lies on that site. In contrast, if there were an even number of base pairs, such as the six base pairs shown in Figure 5.3.3.4(b), then the center of the bound DNA lies halfway between the 3rd and 4th base pair (or at 3.5). Selection of a binding site of 10.5 will center the protein or drug between the 10th and 11th base pair of the DNA sequence. Thus, the binding site position may be specified in whole or half base pairs. The text box on the wizard provides the user the appropriate range or ranges where a protein or drug can be bound, so that no two proteins can occupy the same site. Once the binding site is selected by the user, the sequence (shown in green on Figure 5.3.3.4) and the corresponding base pairs are assigned the step parameter of the DNA (shown in purple on Figure 5.3.3.4) found in the selected high-resolution complex with protein or drug.

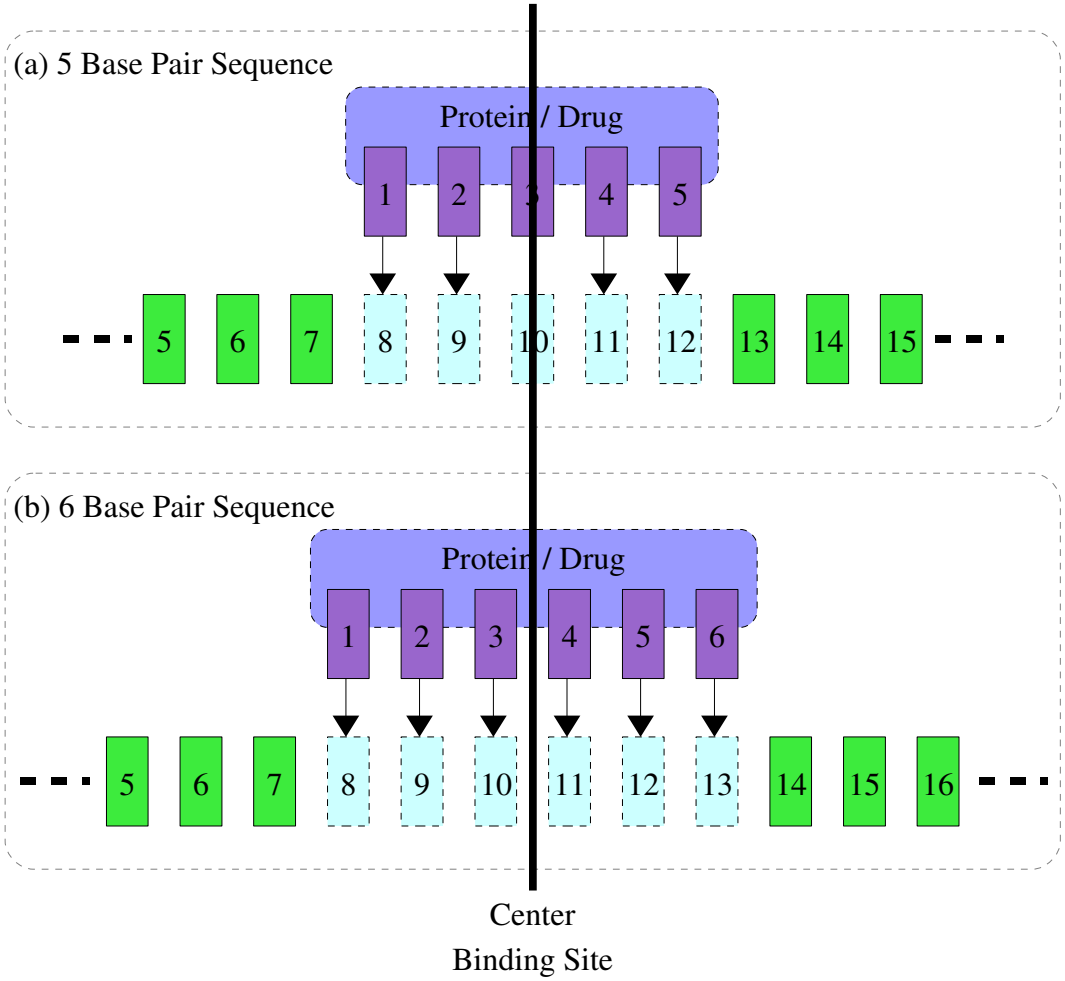


Figure 5.3.3.4: When the user binds a protein/drug to their configuration, the center of the DNA included in the protein/drug PDB file is used by 3DNAdesigner® to align it to the user's sequence. (a) The example of a 5 base-pair sequence shows how an odd number of protein/drug bound base pairs, in purple, can be aligned against the central binding site using a whole number. Here base pair 10 of the DNA sequence, shown in green, would adopt the geometry found at base pair 3 in the known complex. (b) In contrast, the 6 base-pair sequence shows how an even number of protein/drug bound base pairs, in purple, would bind to DNA at half nucleotide values, such as between base-pair steps 10 and 11 at position 10.5 on DNA, shown in green.

In addition to the binding site position, 3DNAdesigner® allows the user to pick one of four colors (yellow, blue, pink, or green) to represent the protein or drug in the final visualization. This comes in handy when there is more than one protein or drug

bound to a single DNA sequence, as it facilitates identifying which ligand is which. The assignment of the position and color is done for each protein or drug identified in the previous wizard, and the user can select the 'Next' or 'Back' button as necessary to change the settings for each. As seen in Figure 5.3.3.3(c) the last step in the 'Binding Site Wizard' presents the users with a summary of these selections, and a 'Finish' button. This option allows the user to check that all of the proteins or drugs are in the correct positions and that the colors are as desired. Once the user chooses the 'Finish' button, all of the steps needed to add proteins or drugs to their DNA configuration are completed.

5.3.4 DNA End Conditions

In addition to the capabilities to create DNA configurations of specified sequences with or without bound proteins or drugs, 3DNAdesigner® also allows the user to specify how the ends of the sequence should be placed in relation to each other. The user may join the ends to form a closed structure, to anchor the ends relative to each other, or to place no restrictions on the ends. These three anchoring conditions correspond to the three major options presented to the user in the 'Boundary Conditions Wizard'; Relaxed State, Create a Closed Structure, and Anchor the Ends, as seen in Figure 5.3.4.1. Since the meaning of closing the structure or placing no restrictions on the ends, thereby giving a relaxed molecule, are self explanatory, this section focuses on the third option, anchoring the ends to points in space, in more detail.

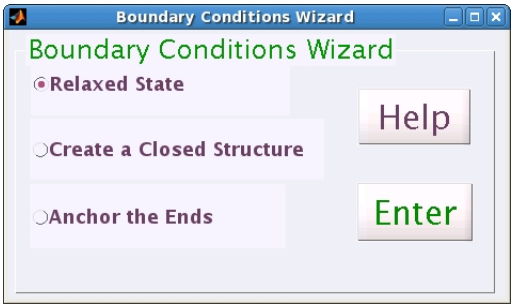


Figure 5.3.4.1: The 'Boundary Conditions Wizard' first allows the user to select one of three options for the constraints to be placed on the ends of the DNA. To make a selection, the user must check one of the three options on the left ('Relaxed State', 'Create a Closed Structure', 'Anchor the Ends'), and then click the 'Enter' button. The user may also select the 'Help' button to learn more details about the boundary conditions they can select from.

Selection of the Relaxed State option creates a molecule exactly as implied. There are no boundary constraints on the ends of the DNA, and accordingly no applied forces or moments. The relaxed state is based on average values of the base-pair-step parameters of high-resolution X-ray crystals found in the Protein Data Bank.

Selection of the Create a Closed Molecule option places the ends of the DNA such that they form a closed (cyclic) structure. This structure can take on many different shapes, not necessarily a circle. The minimum-energy configuration of the closed DNA is based on an elastic model of DNA that uses the data-mined force constants and average values for the six step parameters in the potential energy function [4].

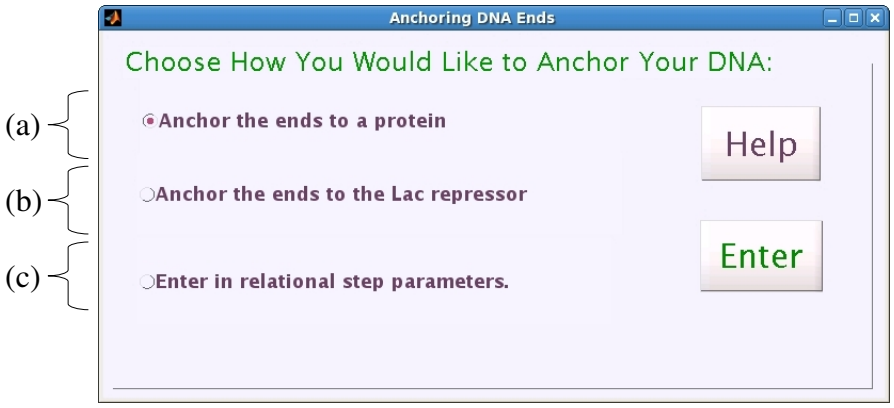


Figure 5.3.4.2: To anchor the ends, the user is presented with three options. (a) Selection of a protein (PDB identifier) with a single binding site that would join the ends of the DNA. (b) Selection of the Lac repressor protein assembly with two binding sites attached to either end of the DNA.. (c) Specification of the six step parameters that describe the displacement and orientation the two ends of the chain will have with respect to each other.

Selection of the Anchor the Ends option gives the user three methods for anchoring the ends of their DNA configuration. The first option, shown in Figure 5.3.4.2(a), allows the user to bind a protein, based on values in a PDB file, such that it forms the chain ends into a protein-bound duplex. The second option in Figure 5.3.4.2(b) allows the user to anchor the ends of the DNA to the Lac repressor protein and form a loop. The third option in Figure 5.3.4.2(c) allows the user to anchor the DNA in an arbitrary spatial arrangement by specifying values for the six step parameters that relate the two ends with respect to each other. Anchoring the ends with any of the three methods leads to determination of a minimum-energy configuration, with the same methodology used to find the configuration of a closed molecule, i.e., with the same elastic model, set of DNA force constants, and rest state for the six step parameters.

5.3.4.1 Anchoring Using Proteins

Two of the three options for anchoring the DNA use proteins as the anchor points. The 'Anchor the ends to a protein' option allows the user to choose any protein-DNA complex, whose coordinates are stored in PDB format in a PDB file located on the user's computer. Anchoring the ends of the DNA configuration to a protein is accomplished by connecting the configuration to the DNA bound to the in from the X-ray crystal structure. The end condition origin values are taken from the PDB file for the protein-DNA complex. To obtain the end conditions 3DNAdesigner® uses the 3DNA software to extract a set of step parameters from the PDB file of the protein the DNA should be anchored to. The anchored DNA will attach to both ends of the DNA that is bound to the protein in the PDB file. The boundary condition to create an anchored structure uses a synthetic closing step. This synthetic closing step geometry comes from the step parameters taken from the ends of the DNA in the PDB file. The PDB file is run through the same screening process discussed in detail in Section 5.3.3. As noted above, the PDB file is limited to those structures that contain a single stretch of protein-bound double-helical DNA.

In addition to being capable of anchoring DNA configurations to many protein-DNA complexes, 3DNAdesigner® handles the special case for the Lac Repressor protein assembly with two disconnected binding sites [5]. The DNA configuration may anchor its ends to the Lac Repressor in any one of the four pictured options presented by the 'Lac

Repressor Orientation' wizard, as can be seen in Figure 5.3.4.1.1. How the DNA can enter and exit the two binding sites is not clear from the palindromic sequence of DNA used in the crystallographic structure [13].

Figure 5.3.4.1.1: The 'Lac Repressor Orientation' wizard allows the user to select between four orientations of DNA on the Lac repressor. The 'A1' or Antiparallel 1 and the 'A2' or Antiparallel 2 form adopt antiparallel direction (arrows) of the DNA on the two binding sites (O1 and O3). Conversely, the 'P1' or Parallel 1 and the 'P2' or Parallel 2 form adopt parallel direction (arrows) of the DNA binding to the two sites.

The anchoring condition for the Lac repressor fall into two major groupings, either antiparallel or parallel. In Figure 5.3.4.1.1, the 'A1' or Antiparallel 1 orientation shows the antiparallel nature of the DNA binding to the protein, coming in from the left of the left-hand side of the V-shaped assembly and going out to the left on the right-hand

side of the complex. The 'A2' or Antiparallel 2 orientation shows a different antiparallel pattern, with the DNA coming in from the right on the left-hand side of the assembly and going out to the right on the right-hand side of the complex. Conversely, the 'P1' or Parallel 1 orientation shows the parallel nature of the DNA binding to the repressor, coming in from the left on the left-hand side and going to the right part on the right-hand side. The 'P2' or Parallel 2 shows a different parallel form, with the DNA coming in from the right of the left-hand side of the assembly and going out to the left on the right-hand side of the complex. Additionally, the 'Lac Repressor Orientation' wizard displays a picture showing the Lac repressor assembly and a simplistic view of the antiparallel and parallel orientations of the ends of the DNA on the two protein binding sites.

5.3.4.2 Anchoring Using Points In Space

The 'Enter in relational step parameters' option (3) of the 'Anchoring DNA Ends' wizard, shown in Figure 5.3.4.2, allows the user to fix the ends in an arbitrary orientation and position by selecting the relative step parameters between the two ends of DNA. Six degrees of freedom are needed to describe how the two base pairs at the ends of the DNA will be positioned in space with respect to each other. The six step parameters used for this purpose are the three rotational parameters (Tilt, Roll, and Twist) and the three translational parameters (Shift, Slide, and Rise). The desired values of the relational step parameters are entered using the 'Anchor To Step Parameters' wizard shown in Figure 5.3.4.2.1. The wizard includes a diagram that graphically displays what each of the step

parameters means in terms of one base pair to another base pair. The translational parameters are given in Ångstrom units and the rotational parameters in degrees. Like the anchoring of DNA to protein, use of the step parameter constraint produces a minimum-energy configuration that is based on an elastic model of DNA with data-mined force constants and rest states in the potential energy function.

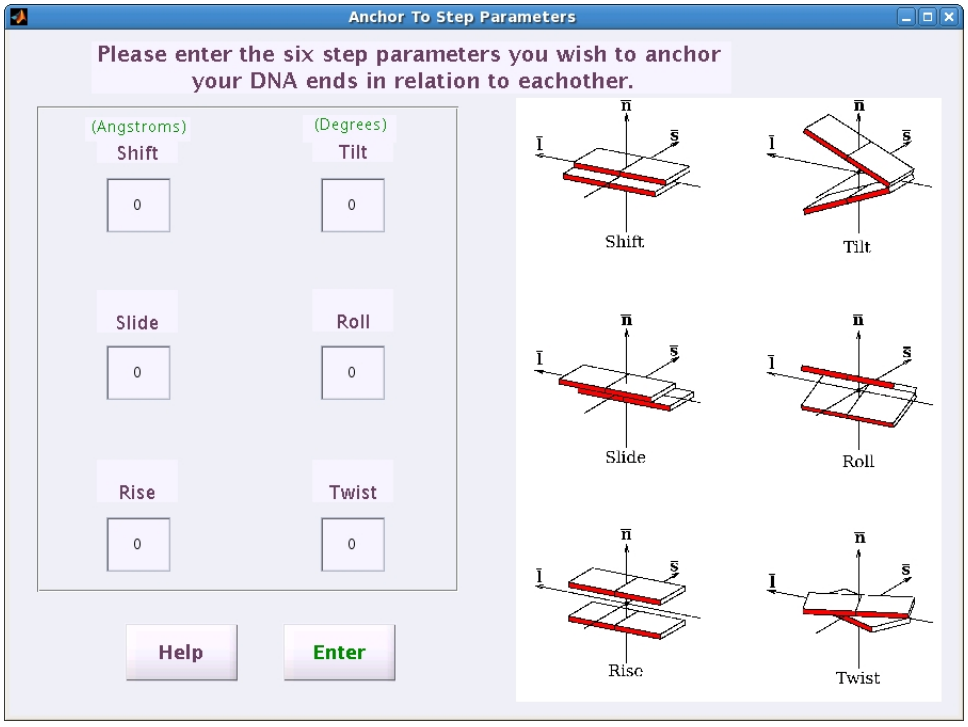


Figure 5.3.4.2.1: The 'Anchor To Step Parameters' wizard allows the user to enter values for the six rigid-body step parameters that relate the two ends of the desired DNA configuration. The translational parameters are given in Ångstrom units and the rotational parameters in degrees. The wizard also includes a diagram that illustrates each of the step parameters in terms of the relation between two base pairs.

5.3.5 Restrictive End Condition Requirements: Moments and Forces

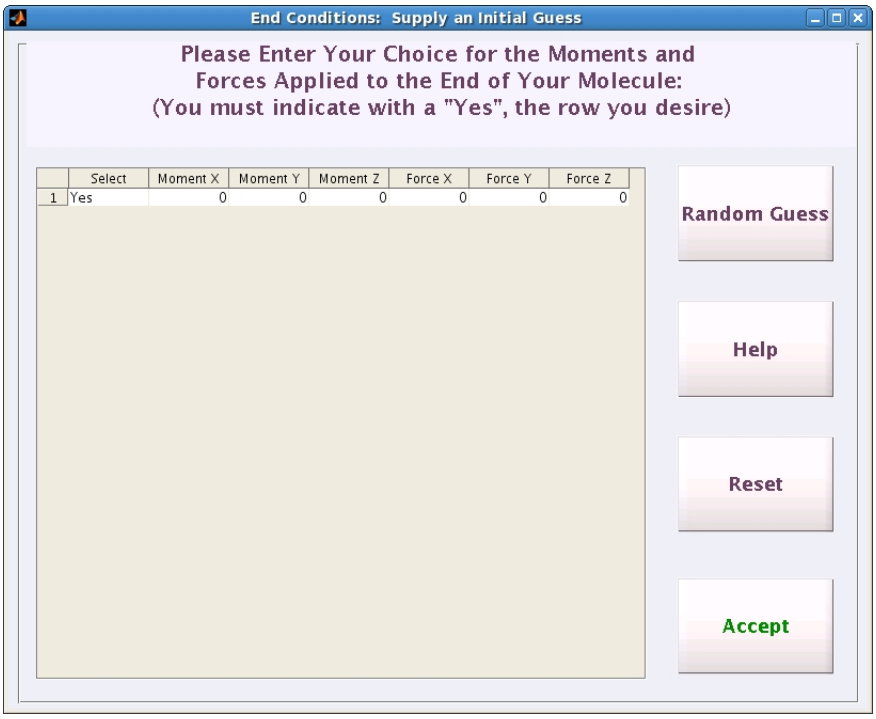
The final step in creating a configuration of a molecule with restricted end conditions is the most complex step in the 3DNAdesigner® software. The 'End

Conditions: Supply an Initial Guess' wizard, as seen in Figure 5.3.5.1, will determine if the DNA configuration created leads a converged state. The determination of whether a configuration converges or not depends on the moments and forces applied to the ends of the DNA. These moments and forces are guesses of the moments and forces at the first step of the molecule, which then propagates through the remaining steps of the molecule as part of an iterative calculation performed by 3DNAdesigner®. This is a critical calculation, which is very sensitive to the values in the initial guess. A change as small as 0.001 Ångstroms or 0.001 degrees can affect whether or not a setup will converge, and if the calculation does converge, the initial moments and forces will have a significant impact on the overall shape and energy of the molecules. The correlation between the initial guesses of the moments and forces and the final outcome of the calculation can seem random.

The moments correspond to the x, y, and z directions of torsions placed on one of the base-pair steps with respect to its neighbor. The forces are the x, y, and z pulls on the base pairs. If the configuration does not converge, the user may give another guess for the moments and forces using the 'New Guess' button (e) from the main 'DNA Configuration' wizard seen in Figure 5.3.1.1. In other words, when a configuration fails to converge the user does not have to go through the setup process again to create the configuration. Instead use of the 'New Guess' button, and entry of new values for the moments and forces in the 'End Conditions: Supply an Initial Guess' wizard allow the

user to attempt the convergence again.

(a)



(b)

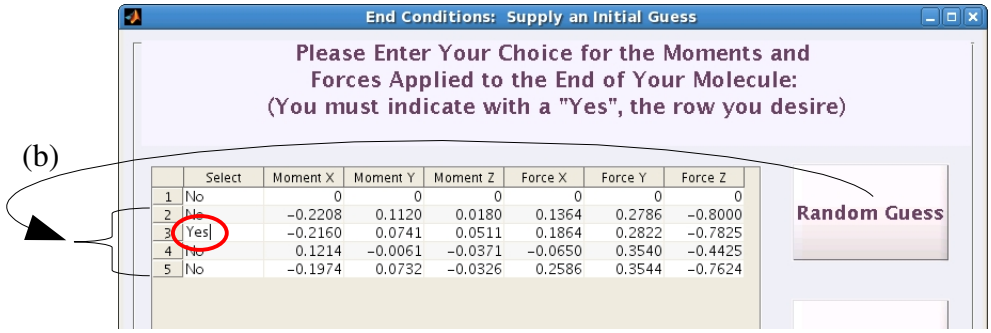


Figure 5.3.5.1: (a) The 'End Conditions: Supply an Initial Guess' wizard provides the user with an interface for generating random guesses for the initial moments and forces applied to one end of the DNA. (b) If the user presses the 'Random Guess' button, five guesses are generated and added to the table automatically by 3DNAdesigner®. The user may type 'Yes' in the first column of the row containing the desired values, and then click the 'Accept' button to see if the computation of the minimum-energy DNA configuration converges with these initial guesses.

The numerical optimization performed in 3DNAdesigner® applies the moments and forces supplied in the 'End Conditions: Supply an Initial Guess' wizard to the step formed by the first and second base pairs. The computation proceeds along successive base-pair steps to see if these numbers make sense for the molecule as a whole. The six relational step parameters are varied and the energy associated with these arrangements are tested along the entire sequence. The test involves a check that the step parameters fall into a conceivable spread for that particular type of step. The lack of convergence after numerous iterations means that the supplied moments and forces were not sufficient, and different ones will need to be entered.

In order to provide initial guesses for the moments and forces, 3DNAdesigner® supplies three tools for entering the values. The first tool is a simple table for manual entry of the moments and forces, as can be seen in Figure 5.3.5.1(a).

The second tool is the 'Random Guess' button, which runs a Metropolis Monte Carlo based calculation to present the user with four new guesses for the moments and forces. These moments and forces will be used as the initial guess to enable the starting configuration to have a greater than average chance of converging to a minimum energy configuration. The Metropolis Monte Carlo calculation accomplishes this by running an iterative process that returns approximations for the step parameters for each base-pair step along the molecule with the given end conditions. It checks the elastic energy with approximations for the base-pair-step parameters and if found to be in an error bar range

for the probability of finding a converged structure when run through the more rigorous Newton-Raphson based process in the main area of the program. These approximations that are deemed in acceptable range are returned in a table in the wizard for the user. The user may choose which set of approximations to use from the list by typing in “Yes” or “No” into the first column. An example of this can be seen in Figure 5.3.5.1(b).

The third tool is the 'Reset' button which brings the user back to a starting point, prior to any manual changes made to the values for the moments and forces. This resets the table to the initial values that the 'Random Guess' button supplied, allowing the user to make changes but still to have an easy way to get back to the supplied guesses.

5.3.6 Choosing a Visualization Method

To facilitate understanding the effects of the DNA sequence on its three-dimensional structure, 3DNAdesigner® provides a visualization tool set to display the overall structure. This feature of the program provides several different types of visual aides to see how the shape is influenced by the sequence and ligands that the user has chosen. The 'Plot Choice Wizard', shown in Figure 5.3.6.1, allows the user to choose the type of visualization.

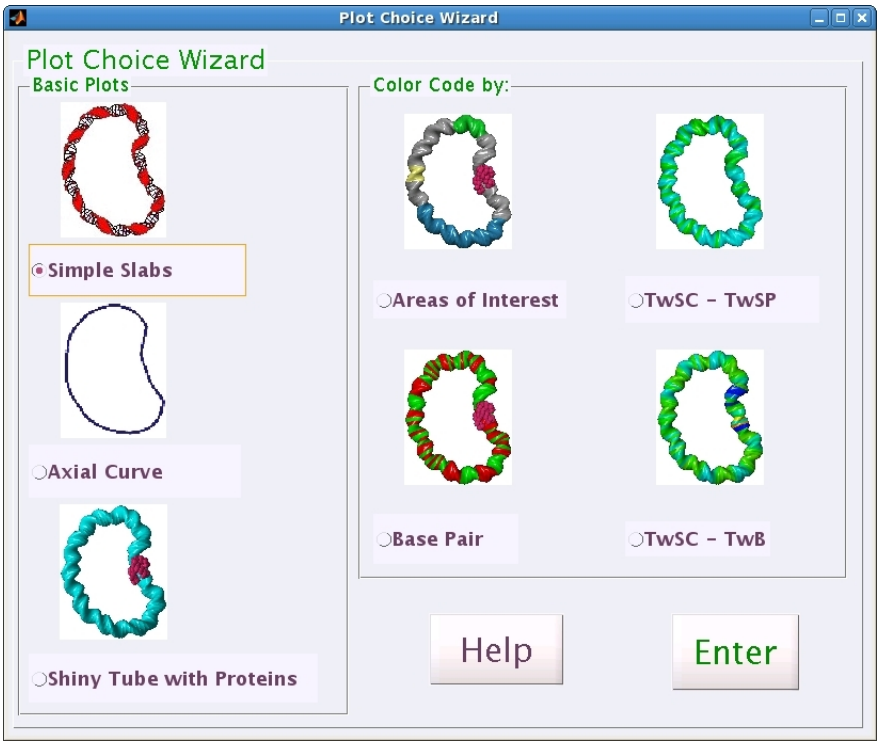


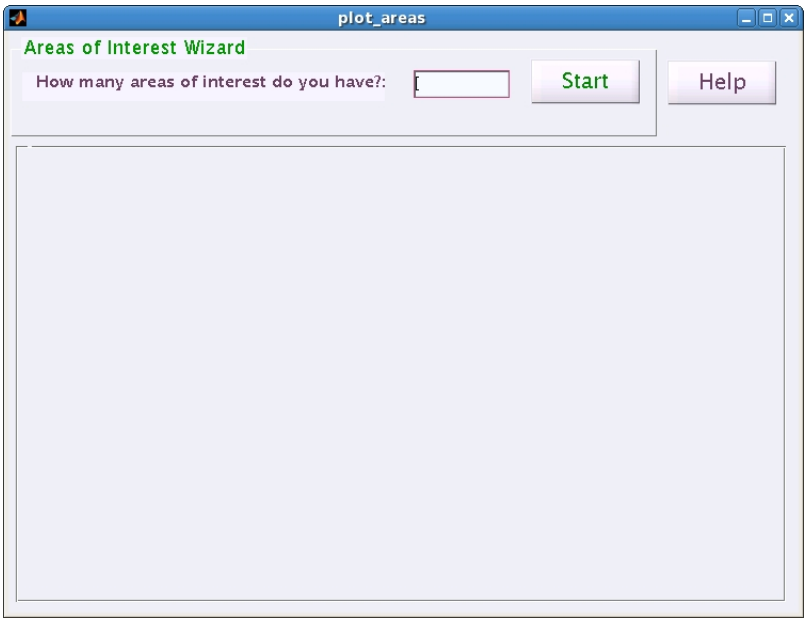
Figure 5.3.6.1: The 'Plot Choice Wizard' allows the user to visualize the optimized DNA configuration in seven different ways. Each plot type is describe by both a sample plot image and a brief description. The 'Help' button provides more details about each visualization option. The plot type is chosen by clicking the check box found under the respective sample plot image and left of the descriptive text. Here the 'Simple Slabs' plot type is selected.

There are seven possible ways to visualize the configurations generated with 3DNAdesigner®. These fall into two major categories. The first category contains three basic plot types: simple base-pair slabs; the axial curve described by the centers of connected base pairs; and shiny protein-bound tubes defined by the base-pair slabs and protein atoms. The rectangular base-pair slabs are color-coded such that the minor groove edge is shaded red. The axial curve plot allows for a better understanding of how the DNA folds in space rather than focusing on its local properties. The protein-bound tube plot creates a more traditional 3D view of DNA and can be downloaded in jpg

format for publication purposes.

The second category contains four color-coded plots to highlight particular features of the DNA configuration. The plots can be color-coded in terms of the segments of interest, the difference between the twist of supercoiling and the step parameter twist (Tw^{SC}–Tw^{SP}), the base pair sequence, or the difference between the twist of supercoiling and that of canonical B-DNA (Tw^{SC}–Tw^{B}). The plot color-coded according to base pairs creates a high quality 3D rendering of the DNA and bound protein or drug molecules and also allows for better visualization of individuals base pairs. In this plot, the A's and T's are shown in red and G's and C's in green. In contrast, the difference in twist values in the plots of Tw^{SC}–Tw^{SP} and Tw^{SC}–Tw^{B} are color-coded in terms of the magnitude of the difference in the twist of supercoiling (Tw^{SC}) in relation to either the step parameter twist (Tw^{SP}) or the twist of B-DNA. The images show only the DNA to highlight the variation in twist. In the plots color-coded by areas of interest plot the user can pick and choose segments within the DNA molecule for examination in the context of the structure as a whole. This option takes the user to the 'Areas of Interest Wizard', as seen in Figure 5.3.6.2, where there are prompts to answer (a) how many regions should be highlighted, and (b) where those regions are located in the DNA molecule.

(a)



(b)

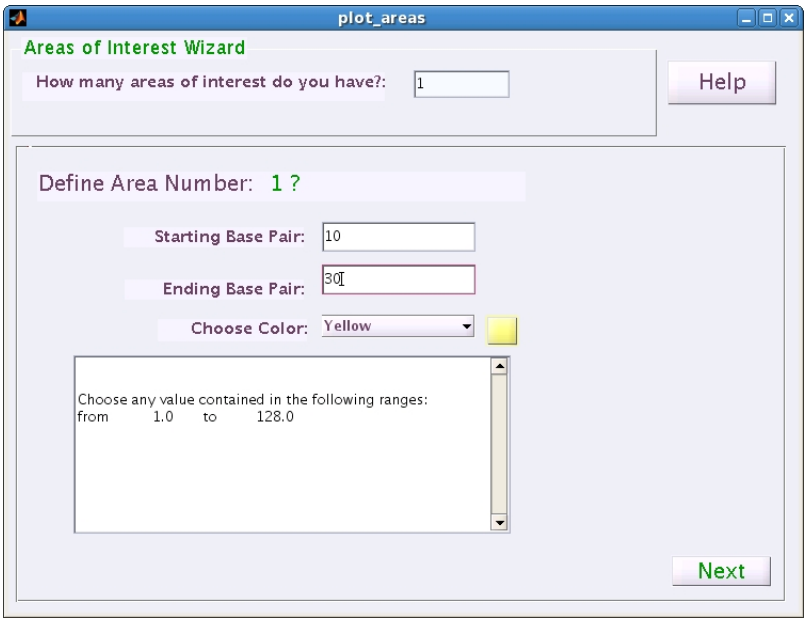


Figure 5.3.6.2: (a) The 'Areas of Interest Wizard' first asks the user for the number of segments to be highlighted, and then for the locations of those segments. (b) The wizard requests the starting and ending base pair for each segment of interest, as well as the color to use to highlight that part of the DNA configuration.

The 'Areas of Interest Wizard' allows the user to choose one or more segments to highlight on the DNA configuration, such as the starting point and the ending point of a successfully closed structure. The user can specify the segments of interest, select a color for each, and specify the base pairs to color. To do this, the user enters the number of segments of interest in the first box of the wizard and clicks the 'Start' button shown in Figure 5.3.6.2(a). The next wizard, in Figure 5.3.6.2(b), allows the user to select the locations of the base pairs at the ends of the first segment of interest, i.e., a 'Starting Base Pair' value and an 'Ending Base Pair' value, and to pick one of four colors (yellow, blue, pink, or green) from the drop-down menu list to represent the selected area. The 'Next' or 'Back' buttons allow the user to go back as necessary and change which base pairs to highlight. To complete the selection process the user pushes the 'Finish' button on the final screen. This screen also allows the user to check that all of the areas of interest are at the correct positions and that the colors are as desired. Once that wizard closes, the user must click ENTER on the plot choice wizard.

5.4 Example Use of 3DNAdesigner®

The three examples presented in this section illustrate how subtle changes in the choice of local features made by the user can affect the global structure of the DNA. The step-by-step walk through shows how to use 3DNAdesigner® to generate the results presented below. All three structural examples bind two nucleosomes (based on the best

resolved structure determined by Davey et. al. [14], 1kx5 from the PDB). The DNA is an ideal homopolymer made up of 454 base pairs and each of the nucleosomes binds to 147 base pairs of DNA. The protein-free DNA linkers between nucleosomes are each 80 base pairs in length. The three examples include a relaxed configuration, a closed configuration, and a configuration anchored by specific distances and angles between terminal base pairs.

5.4.1 Relaxed State

In order to generate a relaxed state configuration, the following procedure is used within 3DNAdesigner®.

1. Start 3DNAdesigner®. Select 'Generate a NEW Configuration' and click 'Enter' as seen in Figure 5.3.1.1.
2. When prompted to enter a filename for the NEW configuration, type in a name and click 'Save'.
3. Click 'Independent' from the 'Choose Type of Sequence' wizard as seen in Figure 5.3.2.1(a) .
4. Enter '454' in the text box for the number of base-pairs and click 'Enter', as seen in Figure 5.3.2.1(b) .
5. Click 'Yes' on the 'Shall We Bind?' wizard as seen in Figure 5.3.3.1(a) .
6. Enter '2' in the text box for the number of occupied sites and click 'Run' as seen in Figure 5.3.3.1(b) .

7. For each of the two prompts that appear in response to the number of bound proteins a PDB file must be selected.
   (a) Each time select the 1kx5.pdb file, and click 'Open'. This will then be processed by 3DNA, which may take some time to run.
   (b) Then click 'Accept' when prompted by the 'Protein File Fixing Area' wizard shown in Figure 5.3.3.2.
8. Click 'Begin' to start the 'Binding Site Wizard' as seen in Figure 5.3.3.3(a).
   (a) Each of the two proteins added through the wizard will take the input of a binding position and color, as seen in Figure 5.3.3.3(b).
   (b) For the first protein enter position '114' and select the color 'Yellow '.
   (c) For the second protein enter position '341' and select the color 'Yellow'.
      i. These positions will place each 147 base pair 1kx5 nucleosome such that there will be 40 base pairs of linker DNA at either end of the configuration with 80 base pairs of linker DNA between the two nucleosomes.
      ii. In this example, both protein assemblies are yellow to make it easier to distinguish them from the DNA, as shown in Figure 5.4.4.1.2. Additionally the common color was selected because the proteins are identical. In some cases it may be beneficial to change the color of the proteins or drugs to help to distinguish one from another in the final configuration.

(d) The user should review the binding site selections before clicking 'Finish' to complete this wizard, as seen in Figure 5.3.3.3(c) .

9. Select 'Areas of Interest' from the 'Plot Choice' wizard, as seen in Figure 5.3.6.1. The user will then be prompted by the 'Areas of Interest' wizard to enter a number of segments, as seen in Figure 5.3.6.2(a) .

(a) Enter '3' in the 'How many areas of interest do you have?' text box and click 'Start '.

(b) For each of the three segments of interest the wizard will prompt the user for the starting and ending base pairs as well as the color. In this example the following values were entered:

i. 1-40 and Pink

ii. 187-267 and Blue

iii. 414-454 and Green

(c) The user should review the selected segments of interest selections before clicking 'Finish' to complete this wizard .

(d) This returns the user to the 'Plot Choice' wizard where the 'Areas of Interest' selection should be highlighted and the user can click 'Enter' .

10. Select 'Relaxed State' and click 'Enter' on the 'Boundary Conditions' wizard, as seen in Figure 5.3.4.1.

11. 3DNAdesigner® will run some calculations and generate a 3D rendering of the

relaxed DNA configuration, as seen in Figure 5.4.4.1.2(a) , as well as a text file and a comma-separated-value (csv) file containing details about the configuration. These results from 3DNAdesigner® are explained in more detail in Section 5.4.4.

5.4.2 Closed Structure

The closed structure differs from the relaxed configuration in that the ends of the DNA are no longer unbounded, but rather connected to form a closed structure. This introduces new constraints on the structure. 3DNAdesigner® provides both visual tools and data to analyze the differences between the relaxed and closed configurations.

Creation of a closed configuration in 3DNAdesigner® entails many of the steps taken to generate the Relaxed State example in Section 5.4.1. The Closed Structure generated in this Section is made up of the same DNA sequence and the same proteins are bound in identical locations along the base-pair sequence.

1. Repeat steps 1-9 of the Relaxed State example in Section 5.4.1.
2. Select 'Create a Closed Structure' and click 'Enter' on the 'Boundary Conditions' wizard, as seen in Figure 5.3.4.1.
3. Click the 'Random Guess' button on the 'End Conditions: Supply an Initial Guess' wizard, as seen in Figure 5.3.5.1(a).
   (a) There will be a dialog box warning the user to be patient as the Monte Carlo calculations done at this point can take a considerable amount of time depending on the desired type of configuration. It is safe to click 'Ok' on this

dialog, but please be patient for the calculations to complete.

(b) The random guess procedure will generate four new entries in the table and each set of values may be selected for use by typing “Yes” into the first column called 'Select' and by clicking 'Accept', as seen in Figure 5.3.5.1(b). This process may be repeated to generate new guesses until one leads to a converged configuration.

(c) For the configuration used in this example the following values for the moments and forces, described in more detail in Section 5.3.5, were used to attain convergence. To use these same values, the user may manually type them into a row and select them by typing “Yes” into the first column of that row, similar to the procedure described in 3(b) above. Note that when manually entering values, the full value will be be used even though the table may only display a rounded portion of the value.

i. Moment X: 0.074973

ii. Moment Y: 0.114814

iii. Moment Z: –0.020312

iv. Force X: 0.066043

v. Force Y: –0.064355

vi. Force Z: 0.053761

4. 3DNAdesigner® performs an iterative calculation of the energy of varying chain

configurations and generates a 3D rendering of the final converged energy-minimal configuration, as seen in Figure 5.4.4.1.2(b) , as well as a text file and comma separated value (csv) file containing details about the configuration. These results from 3DNAdesigner® are explained in more detail in Section 5.4.4.

5.4.3 Anchor the Ends: Points in Space

In this example, the ends of the DNA will neither be relaxed nor connected together. Instead they will be assigned values that position them at fixed locations in space based on the six step parameter values entered. Like the Closed Structure example in Section 5.4.2, the user will repeat several steps from the Relaxed State example in Section 5.4.1 because the primary configurations of the DNA and the locations and identities of the bound proteins are identical.

1. Repeat steps 1-9 of the Relaxed State example in Section 5.4.1.
2. Select 'Anchor the Ends' and click 'Enter' on the 'Boundary Conditions' wizard, as seen in Figure 5.3.4.1.
3. Select 'Enter in relational step parameters' and click 'Enter' on the 'Anchoring DNA Ends' wizard, as seen in Figure 5.3.4.2(c).
4. Enter the following values in the 'Anchor To Step Parameters' wizard and click 'Enter', as seen in Figure 5.3.4.2.1.
   (a) Shift: 4
   (b) Slide: –19

(c) Rise: 76

(d) Tilt: 5

(e) Roll: 19

(f) Twist: –75

5. Click the 'Random Guess' button on the 'End Conditions: Supply an Initial Guess' wizard, as seen in Figure 5.3.5.1(a).

   (a) There will be a dialog box warning the user to be patient as the Monte Carlo calculations done at this point can take a considerable amount of time depending on the desired type of configuration. It is safe to click 'Ok' on this dialog, but please be patient for the calculations to complete.

   (b) The random-guess procedure will generate four new entries in the table and each set of values may be selected for use by typing “Yes” into the first column called 'Select' and by clicking 'Accept', as seen in Figure 5.3.5.1(b). This process may be repeated to generate new guesses until one leads to a converged configuration.

   (c) For the configuration used in this example the following values for the moments and forces, described in more detail in Section 5.3.5, were used to attain convergence. To use these same values, the user may manually type them into a row and select them by typing “Yes” into the first column of that row, similar to the procedure described in 5(b) above. Note that a manually

entering values, the full value will be be used even though the table may only display a rounded portion of the value.

i. Moment X: 0.053866

ii. Moment Y: 0.051762

iii. Moment Z: –0.084272

iv. Force X: 0.004972

v. Force Y: –0.048378

vi. Force Z: 0.078127

6. 3DNAdesigner® performs an iterative calculation of the energy of varying chain configurations and generates a 3D rendering of the final converged energy-minimum configuration, as seen in Figure 5.4.4.1.2(c) , as well as a text file and comma-separated-value (csv) file containing details about the configuration. These results from 3DNAdesigner® are explained in more detail in Section 5.4.4.

5.4.4 Results

A converged configuration created in 3DNAdesigner® generates several types of results that are discussed in more detail in this Section. These results can be used to analyze further the configuration created using the sequential and binding information entered by the user. The results are presented to the user in two forms.

5.4.4.1 Plots

The first form of the results generated by 3DNAdesigner® is the 3D plot selected

by the user on the 'Plot Choice' wizard, shown in Figure 5.3.6.1. This plot can be used to analyze the structure visually by using the tool bar provided with each plot, as seen in Figure 5.4.4.1.1. This tool bar includes options for 3D rotation of the configuration, adjustable lighting, multiple-axis views, positioning on the screen, and zoom in or out. Many of these tools can be used statically or set in motion, such as continuously rotating the structure or varying the lighting around the plotted structure.

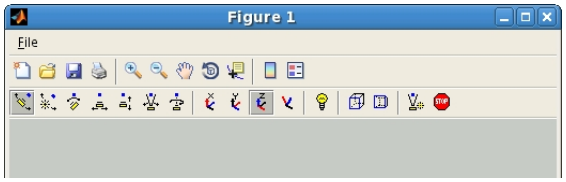


Figure 5.4.4.1.1: The tool bar provided with each plot. This tool bar allows the user to analyze the structure of a configuration visually through a variety of tools, including 3D rotation of the configuration, adjustable lighting, multiple-axis views, positioning on the screen, and zoom in or out options. Many of these tools can be used statically or set in motion, such as continuously rotating the structure or varying the lighting around the plotted structure.

The three examples presented in Sections 5.4.1, 5.4.2, and 5.4.3 generate slightly different configurations . The quickest and most obvious way to distinguish among these structures is to compare plots of the three pathways. Each of the examples is color-coded so that the same three segments of linker DNA are highlighted in the same colors (pink, blue, and green). The DNA bound to the nucleosomes is shown in grey and the histone proteins in yellow. The PDB file used to describe the nucleosome fixes the pathway of the bound DNA and protein in the same arrangement found in the crystal structure.

The similar color-coding of the three plots helps to visualize the differences

between the three configurations by comparing the relative positioning of the pink, green, and blue segments of interest, as well as the nucleosomes. The differences should be clear to even novice users, as shown in Figure 5.4.4.1.2. The relaxed state, displayed in Figure 5.4.4.1.2(a), shows the structure with no forces acting on the ends of the DNA and provides a neutral baseline to compare and contrast how forces acting on the ends of the chain can affect the global structure. The closed structure, displayed in Figure 5.4.4.1.2(b), reveals how the rotation of the pink and green sections of DNA that is needed to close the chain ends also causes the nucleosomes to take on new positions relative to their counterparts in Figure 5.4.4.1.2(a). The anchored configuration, displayed in Figure 5.4.4.1.2(c), shows how fixing the ends in the desired orientation and location rotates the upper nucleosome and places the green section of DNA behind the unbroken linker DNA. Use of the toolbar in Figure 5.4.4.1.1 allows the user to view these structures in an almost infinite number of alternative ways to highlight differences, and to uncover details that warrant further inspection.

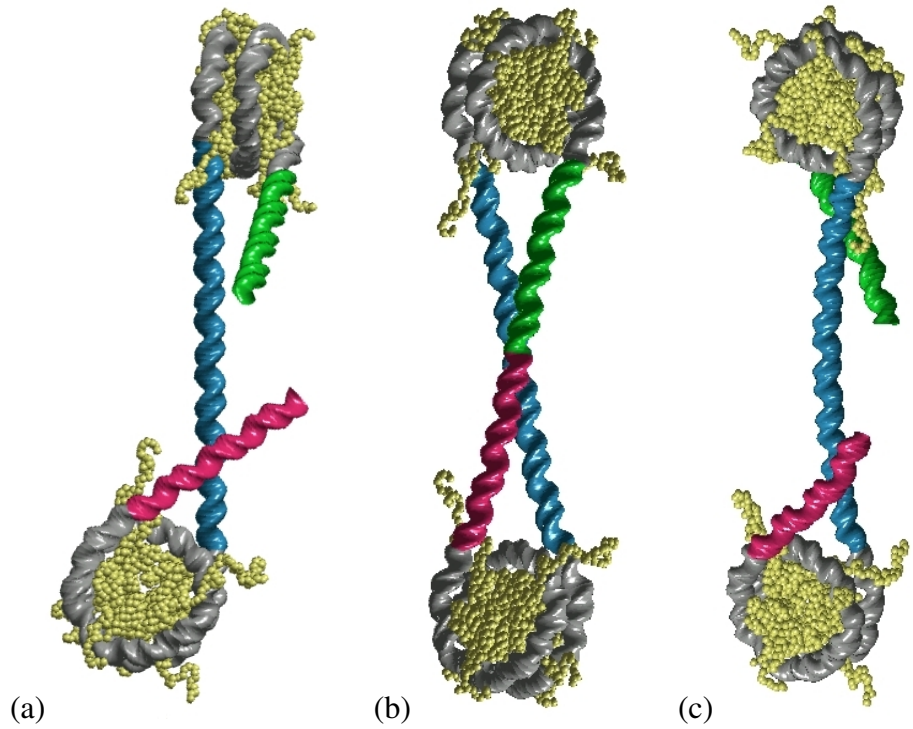


Figure 5.4.4.1.2: Molecular image demonstrating the differences between open, closed, and spatially anchored configurations of a 457 base-pair DNA binding two nucleosomes separated by an unbound 80 base-pair linker. The ends of the DNA are shown in green and pink and the central linker in blue. The nucleosome DNA is shown in gray and the bound histone proteins in yellow. (a) The relaxed state, from Section 5.4.1, has no forces acting on the ends of the DNA. (b) The closed structure, from Section 5.4.2 shows how the rotation of chain ends closes the DNA and rotates the nucleosomes in new relative positions to their counterparts in (a). (c) The anchored state ('Anchor the Ends: Points in Space'), from Section 5.4.3, shows how fixing the ends to points in space has reorientated the upper nucleosome and placed the green end of DNA clearly behind the blue linker DNA.

5.4.4.2 Formatted Files

More advanced users may have greater interest in the data generated by 3DNAdesigner®. The data are stored in three major file formats. The details of the local sequence and ligand binding input are stored in a Matlab .mat file format. The .mat file is less useful on its own but can be reused and edited by 3DNAdesigner®, as seen in Figure 5.3.1.1(a). This information is stored in a file with a syntax of <filename>.mat, where <filename> is the name chosen by the user to assign to a new configuration .

3DNAdesigner® displays two sets of output in the main window. These are highlighted in Figure 5.3.1.1 as (b) the 'DNA System and Anchoring Conditions' and (c) the 'Output Gazer', where the information they contain shows pertinent starting conditions and end results from the configuration generated by the user. This same information is stored in a text file with a syntax of <filename>_StartandEndOutput.txt, where <filename> is the same filename chosen for <filename>.mat.

The third file contains more detailed data about the structure, energy, and topology of the configuration generated by the user. This includes data such as the values of the six step parameters (shift, slide, rise, tilt, roll, twist) at each base-pair step in the structure, the total twist of supercoiling, the total number of helical turns in the structure, the writhe, the linking number, and the coordinates of the origins, normal, short axis, and long axis of each base pair. These data are stored in a comma-separated-value (csv) file format, to make it easy to import into a spreadsheet application for further

analysis. The file is named with a syntax of <filename>_TwSCStepParams.csv, where <filename> is the same filename chosen for <filename>.mat.

5.5 References

[1] B.D. Coleman, W.K. Olson, and D. Swigon. (2003). Theory of Sequence-Dependent DNA Elasticity. *J. Chem. Phys.,* **118(15)**, 7127-7140.

[2] W.K. Olson, A.A. Gorin, X. Lu, L.M. Hock, and V.B. Zhurkin. (1998). DNA Sequence-Dependent Deformability Deduced from Protein–DNA Crystal Complexes. *Proc. Natl. Acad. Sci. USA*, **95(19)**, 11163–11168.

[3] B.D. Coleman, and D. Swigon. (2000). Theory of supercoiled elastic rings with self-contact and its application to DNA plasmids. *J. Elasticity*, **60**, 171–221.

[4] W.K. Olson, D. Swigon, and B. Coleman. (2004). Implications of the Dependence of the Elastic Properties of DNA on Nucleotide Sequence. *Phil. Trans. R. Soc. Lond. A*, **362(1820)**, 1403-1422.

[5] D. Swigon, B.D. Coleman, and W.K. Olson. (2006). Modeling the Lac Repressor-Operator Assembly. I. The Influence of DNA Looping on Lac Repressor Conformation. *Proc. Natl. Acad. Sci. USA*, **103(26)**, 9879–9884.

[6] I. Tobias, D. Swigon, and B.D. Coleman. (2000). Elastic Stability of DNA Configurations. I. General Theory. *Phys. Rev. E.,* **61(1)**, 747-758.

[7] A.V. Colasanti. (2006). Conformational States of Double Helical DNA. *Ph.D. Thesis, Rutgers the State University of N.J. - New Brunswick and U.M.D.N.J.,* 166 pp.

[8] The MathWorks Inc. (2011). MATLAB Compiler – MATLAB. Available at *http://www.mathworks.com/products/compiler/*

[9] X. Lu, and W.K. Olson. (2003). 3 DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.,* **31(17)**, 5108-5121.

[10] National Center for Biotechnology Information (NCBI). (2009). FASTA Format Description. Available at *http://www.ncbi.nlm.nih.gov/blast/fasta.shtml*

[11] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. (2000). The Protein Data Bank. *Nucleic Acids Res.,* **28(1)**, 235-242.

[12] X. Lu, and W.K. Olson. (2003). 3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three-Dimensional Nucleic Acid Structures. *Nucleic Acids Res.,* **31(17)**, 5108-5121.

[13] M. Lewis, G. Chang, N.C. Horton, M.A. Kercher, H.C. Pace, M.A. Schumacher, R.G. Brennan, and P. Lu. (1996). Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. *Science*, **271(5253)**, 1247-1254.

[14] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 a Resolution. *J. Mol. Biol.,* **319(5)**, 1097-1113.

Chapter 6: Naked and Protein-bound DNA Equilibrium Structures

6.1 Significance of 3DNAdesigner to Real World Problems

With the development of 3DNAdesigner™ it has become very easy and convenient to construct three-dimensional models of spatially constrained DNA molecules. It does not matter whether that molecule is plain naked DNA, or DNA bound to a protein/drug/ligand. The limitations on creating a model of a distinct DNA molecule are only affected by the numbers of base pairs. The greater the number of base pairs, the more computer-processing power required. The optimization procedure used to construct these models is also not valid for long chains, which may undergo significant structural fluctuations. Therefore, 3DNAdesiger™ only handles structures with a maximum of 500 base pairs. With the user-friendly design of 3DNAdesigner, biologists can now have a reliable and effective tool to visualize a molecule of interest to them, as well as to estimate the relative energy costs. Chapter 5 discussed how to download, run, and use 3DNAdesigner™ to create a model of a DNA molecule, as well as how to examine the calculated results. This chapter will provide some results we have gathered from using 3DNAdesigner™ and an additional back-end Linux script, which runs multiple DNA

configurations through the MATLAB code serially.

3DNAdesigner™ simulations can help show the effect of a bound protein on the overall fold of the DNA. The ability to study a multitude of sequences with bound proteins in 3DNAdesigner™ allows it also to be used as a tool to suggest new sequences worthy of study in the lab. The stabilities of all converged configurations are calculated and quantified through the examination of the elastic energies.

6.2 Naked 339 Base Pair Baylor Sequence

During the summer of 2005 a collaboration was formed between the Zechiedrich lab in the Department of Molecular Virology and Microbiology at Baylor College of Medicine and the Olson group at Rutgers University. Considering that all of the research in this thesis is based *in silico*, it was compelling to form a partnership with a lab that worked with *in vitro* studies. Such a collaboration allows for a comprehensible exchange of ideas and a way for each group to qualify and quantify our interconnected studies.

The Zechiedrich lab studies closed, superhelical DNA structures. One sequence that they are interested in is a closed, 339 base-pair structure that is made from a larger piece of supercoiled DNA that was shortened by the use of cutting enzymes. The Zechiedrich lab has the capability to make very short mini-circles in very high yield with unprecedented levels of supercoiling using their novel enzymatic approach. Figure 6.2.1

shows this base-pair sequence with sections of interest highlighted. The highlighted area in yellow is a section included in the sequence that will bind to an enzyme from E. coli called restriction endonuclease V, or EcoRV [4]. The other two highlighted areas are remnants of the method used to make the supercoiled loop. These are referred to as the attb and attp sites. In the future, we will look at how changing the unhighlighted areas, by deleting parts of them, can alter the energy landscapes of the closed molecule with 3DNAdesigner. The highlighted areas, however, are to remain as is, untouched.

TTATACTAA CTTGAGCGAA ACGGGAAGGG TTTTCACCGA TATCACCGAA

ACGCGCGAGG CAGCTGTATG GCGAAATGAA AGAGTTCTTC CCGGAAAACG

CGGTGGAATA TTTCGTTTCC TACTACGACT ACTATCAGCC GGAAGCCTAT

GTACCGAGTT CCGACACTTT CATTGAGAAA GATGCCTCAG CTCTGTTACA

GGTCACTAAT ACCATCTAAG TAGTTGATTC ATAGTGACTG CATATGTTGT

GTTTTACAGT ATTATGTAGT CTGTTTTTTA TGCAAAATCT AATTTAATAT

ATTGATATTT ATATCATTTT ACGTTTCTCG TTCAGCTTT

Figure 6.2.1: 339 base-pair Baylor sequence with a binding site for EcoRV, a restriction endonuclease, highlighted in yellow. The starting site (a remnant of how this sequence is formed) is in blue and the ending site (a remnant of how this sequence is formed) is in pink. All three highlighted parts of the sequence may not be changed in molecular redesign. A, T, G, C bases are color-coded red, blue, green, and cyan, respectively.

3DNAdesigner can show the nature of the rest state when the DNA is open and in its lowest energy form. Figure 6.2.2 gives an idea as to what the sequence in Figure 6.2.1 looks like without imposing forces on the ends or requiring it to be closed. The three-

dimensional pathway reflects the assumed sequence-dependent elastic potential of DNA.

Figure 6.2.2: Image of the naked 339 base-pair Baylor oligomer without imposing forces on the ends or requiring it to be closed. The yellow highlighted area near the left end of the structure points out the EcoRV binding site, highlighted in Figure 6.2.1.

A cyclic structure results if the ends of the sequence can be forced to meet. In order to close the structure in 3DNAdesigner™, the moments and forces at the joining steps must be optimized. To find the stable closed structures it is necessary to vary one component of the end moment or force at a time at minute increments. Often it can take many attempts to find optimal moments and forces for creating a closed structure. Therefore, it can be extremely time consuming to find a new shape that will meet the requirements of an *in vitro* experiment. The Zechiedrich group is interested in looking at configurations with negative ΔLk compared to the relaxed cyclic form, where the linking number $Lk = 32$. This is also the Lk of the lowest energy "circle" that was found in our preliminary calculations. It should be noted that the configuration is not a smooth circle, but rather resembles the shape of the letter D (Figure 6.2.3).

Since the location of the binding site for EcoRV is known, we can suggest alterations of the original 339 base-pair sequence that may be more receptive for binding the EcoRV protein. Rearrangement of the sequence, i.e., placement of the EcoRV

binding site in different parts of the structure, may identify areas of the DNA where a protein may or may not bind more easily in terms of the energy costs of overall DNA deformation, the topology of the molecule as a whole, and the possible interferences with the binding area. Figure 6.2.3 shows a small sample of all possible minimum-energy shapes that could be found with $\Delta Lk \leq 0$, and Table 6.2.1 shows all pertinent information about the shapes in Figure 6.2.3. The sequence is color-coded according to the base pairs on the leading strand (A = red, T = blue, G = green, C = cyan). The yellow area is the segment that binds specifically to the EcoRV enzyme. The Zechiedrich group has seen shapes that resemble (a) and (b). They can produce closed structures with *Lk* of 32 or less [14]. Some of the shapes that were found using the 339 base-pair Baylor sequence, have been seen under atomic force microscopy (AFM) [14].

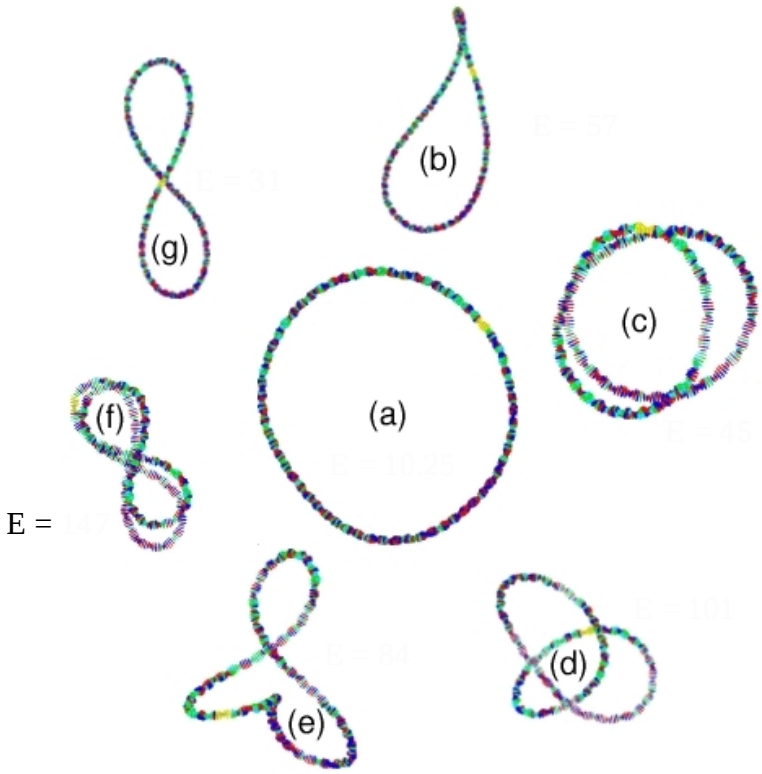


Figure 6.2.3: Predicted locally stable equilibrium configurations of a naked closed 339 base-pair sequence. While only the very lowest minimum-energy configuration is likely to occur, some of these examples can never occur due to self penetration. Our code does not exclude configurations that have self intersection and while not physically possible, it can give a sense of what is happening to the shape and energy as the program iterates and tries to find more minimum-energy configurations. The EcoRV binding site, shown in yellow, contains the GATATC sequence. (a)-(e) refer to entries in . (A = red, T = blue, G = green, C = cyan)

Figure 6.2.3	*Shape*	*Energy (kT)*	*Tw^{SC} (turns)*	*Wr*	*Lk*	*ΔLk*
(a)	Circle	10.25	31.79	0.21	32	0
(b)	Open Figure 8*	57.00	33.62	0.38	34	2
(c)	Doubly Wound Circle	45.00	31.75	–0.75	31	–1
(d)	Knot	101.00	29.88	–2.88	27	–5
(e)	Three Lobes	84.00	31.38	–1.38	30	–2
(f)	Dragonfly**	147.00	31.70	1.30	33	1
(g)	Figure 8	31.00	31.76	–0.76	31	–1

*Table 6.2.1: Description of the computed shape, the associated elastic energy, and topological parameters – twist (Tw^{SC}), writhing number (Wr), linking number (Lk), and difference in linking number versus that of the relaxed cyclic form (ΔLk) – for the selected shapes depicted in Figure 6.2.3. * Similar shape seen by Zechiedrich . ** Depicted since it is the first time that this shape has ever been seen.*

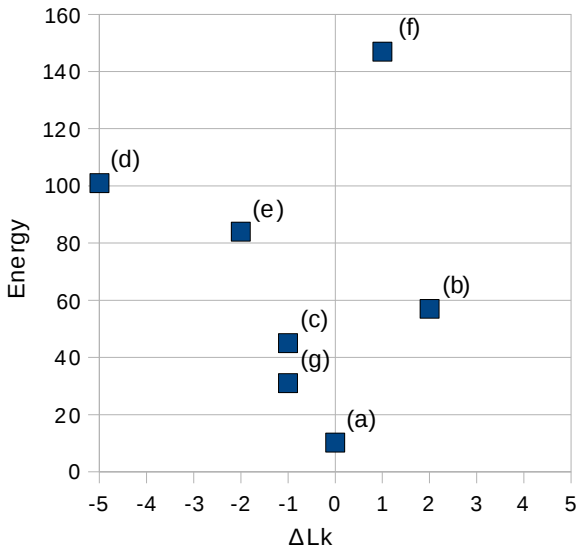


Figure 6.2.4: Plot of elastic energies versus difference in linking number compared to the relaxed cyclic form (ΔLk) from Table 6.2.1.

6.2.1 Results for the Naked 339 Base-Pair Baylor Sequence

Minimum-energy configurations were found using 3DNAdesigner for the naked closed circular 339 base-pair Baylor sequence. To obtain a huge number of minimum-energy configurations a perl script was written to run 3DNAdesigner continuously for a period of three months. The script made it possible to perform new runs as soon as the last one would finish and this made the most efficient use of available computing power. Each run would have a minutely changed guess of initial values for the moments and forces input to 3DNAdesigner. Since there are no correlations between the initial guess for the moments and forces and whether or not convergence is achieved, we are not assured of a converged minimum-energy configuration for each initial guess input by the script.

During this period of time, numerous converged minimum-energy configurations were found for this closed structure. There were many identical structures found within this set of data despite the varied moments and forces input. This cannot be helped since we have no way of telling what initial values of moments and forces create which configurations. In addition, different starting values for the same initial guess for the moments and forces can, indeed, lead to the same output. Therefore, even though thousands of results came back that had converged, only 117 were unique. However, just because the program ran for a long time, it does not mean that these are the only

configurations that can exist. The automated program slowly increments one of the six moments or forces at a time from a the supplied starting values of moments and forces. 3DNAdesigner™ then attempts to find converged configurations. It is entirely possible and probable to find more converged structures with different initial starting guesses.

Figure 6.2.5 is a plot of the energies and linking numbers of all 117 converged structures. The value of *Lk* ranges between 26 to 37. The values for Tw^{SC}, energy, linking number, and writhe for each of the 117 converged structures can be seen in Section 6.5.1. The shear volume of results allows us to see a lot about the configurations of this molecule. For example, we can easily see that the lowest energy structure does, in fact, have a linking number *Lk* of 32, as expected from the Zechiedrich group's *in vitro* experiments. As also expected, the next lowest energy states occur for linking numbers that differ by +/–1 from $Lk = 32$ and the energy increases with further change in ΔLk. The highest energy value also lies on the $Lk = 32$ line.

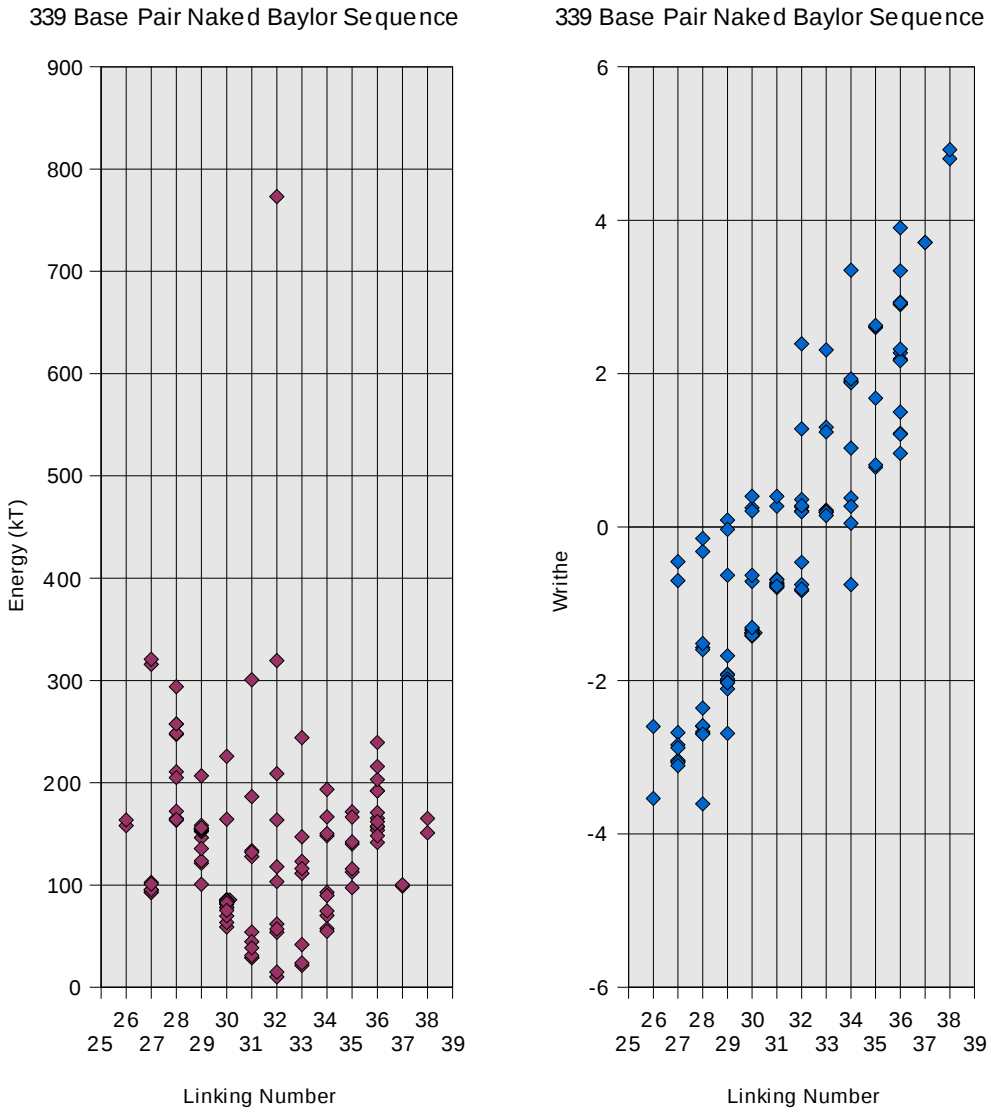
The lowest energy structure, shown in Figure 6.2.6(i), is almost an ideal circle, while the highest energy structure, shown in Figure 6.2.6(ii), is a multi-wound circle with a lot of supercoiling that creates a compact state and shows how much the compaction of DNA can influence the energy. Comparison of the two structures in Figure 6.2.6 shows how the amount of supercoiling can influence the compaction of DNA. Just looking at the scale of the tubular widths can give such an appreciation. It takes much more energy

to make structure (ii). The energy for the low energy structure is 10.25kT while the energy for the highest energy structure is 773.06kT. The latter figure does not include terms associated with the close contact or overlap of different parts of the chain model.

What makes 3DNAdesigner so powerful is its ability to show a 3-D plot of the DNA. If we just compared the energy, *Wr*, *Lk*, and Tw^{SC} for the molecules, we would miss part of the story since values other than the energy are so similar for various structures. With the depictions in Figures 6.2.6 and 6.2.7 it is crystal clear that there is a huge difference between the two structures.

To look closer at how the energies can indicate an extensively different structure, we present Figure 6.2.8. In this figure we show the sequential variation in the twist of supercoiling, Tw^{SC}, along the contour of the highest and lowest energy structures. The *Lk* for both of these configurations happens to be 32. This figure seems quite fascinating. We are dealing with the same sequence for both structures, but we can clearly see how the highest energy structure has Tw^{SC} values at each step that vary widely from those of the lowest energy structure. As we trace our way along the base pairs of the molecule, we can also see how the Tw^{SC} values for both of these structures are relatively consistent. If the ΔTw^{SC} between steps increases for one structure it generally is echoed on the other, but of course not with the same magnitude. Biologists may have an interest in these types of comparison graphs if they were especially interested in where or if a protein/ligand/drug

could bind to a specific area. The Tw^{SC} could help identify those areas since, in addition to specific interactions with the DNA bases, many proteins prefer either an over- or under-twisted spot to bind. Many proteins, of course, also look for specific interactions with the DNA bases.



(i) *(ii)*

Figure 6.2.5: Plots of the variation of (i) energy and (ii) writhe with the linking numbers of the 117 unique configurations of the 339 base-pair naked Baylor sequence found using 3DNAdesigner™. The value for each configuration is marked with a diamond. The lowest energy structures have a Lk of 32. A 339 base-pair structure of ideal B-DNA would also have a Lk about 32 and a writhe of 0. The plot shows, due to the large spread of data points, that there is a large range of Lk numbers possible for this molecule and that a few have low enough energies that a biologist may be interested in trying to interpret his/her data in terms of these supercoiled structures.

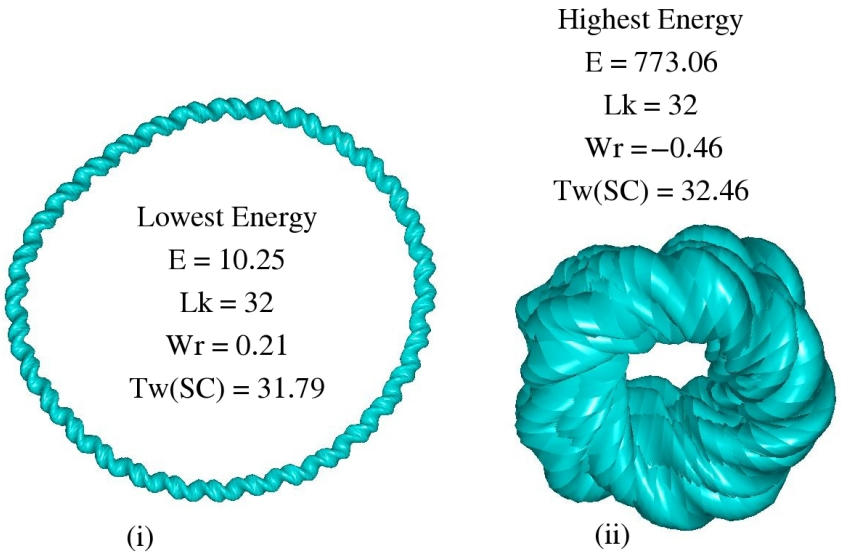
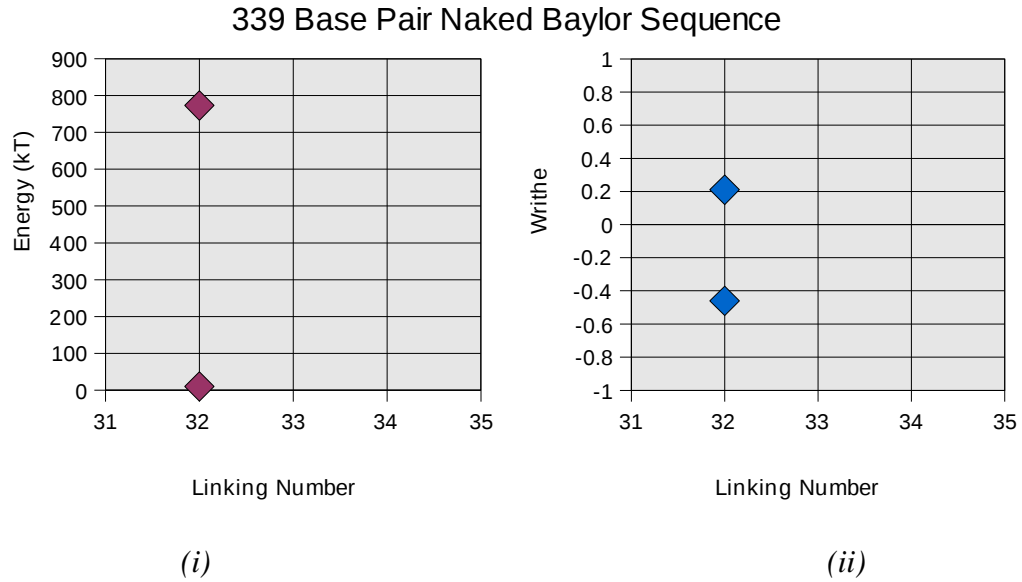
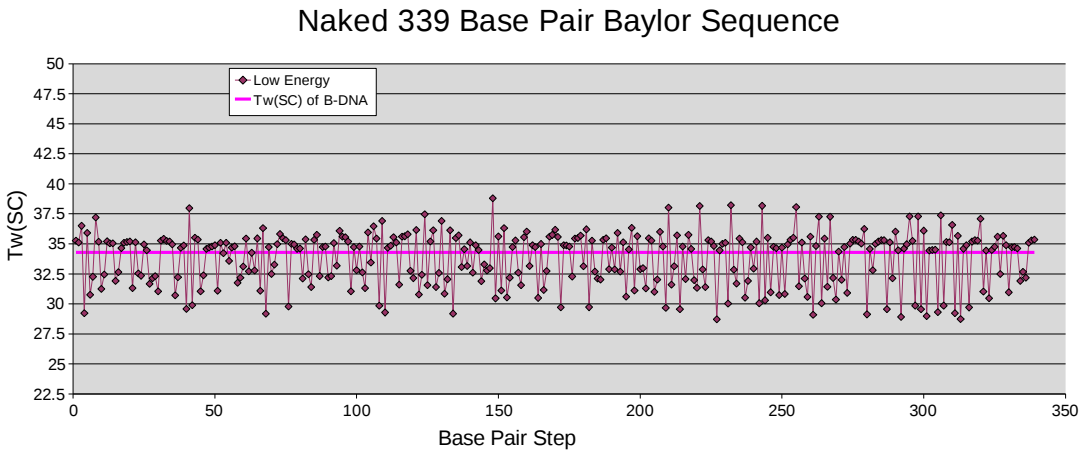


Figure 6.2.6: Images of the topology for the (i) lowest and (ii) highest energy structures for the naked 339 base-pair Baylor closed sequence. Energy is in kT. The structure in (ii) includes areas of self contact. This structure will not ever exist in vivo or vitro and is only to be used as a point of reference for energy calculations versus topology.

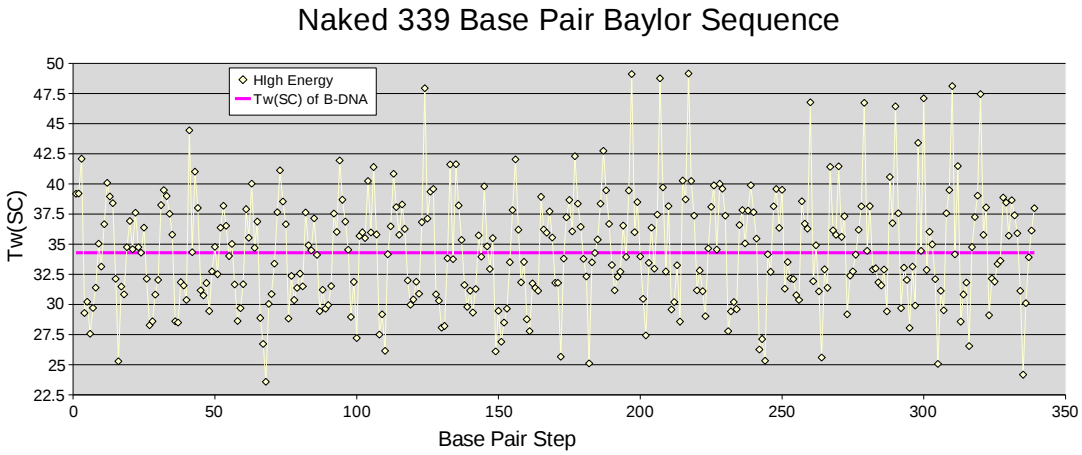


(i) *(ii)*

Figure 6.2.7: Plots of the (i) energy and (ii) writhe with respect to the linking number for both the lowest and highest energy structures for the naked 339 base-pair Baylor closed sequence. Energy is in kT.



(i)



(ii)

Figure 6.2.8: Sequential variation of the twist of supercoiling along the chain contour of the (i) lowest and (ii) highest energy states of the 339 base-pair Baylor sequence found with 3DNAdesigner. The Tw^{SC} values of the lowest energy structure stay much closer to the value of relaxed B-DNA, depicted by the pink line, than those of the highest energy structure. However, note in general that the steps with higher Tw^{SC} coincide in the two configurations, although the amplitude is dampened in the lower energy state.

6.3 EcoRV Bound to the 339 Base Pair Baylor Sequence

The Zechiedrich group has found that the binding of EcoRV (Figure 6.3.1) to closed circular DNA is sensitive to the degree of supercoiling. This is noteworthy because until their work it was believed that supercoiling had no effect on binding protein to DNA. Previously only very long segments were used in studies of the DNA binding properties of this enzyme, which meant that all the slack due to the length of DNA negated the effects of supercoiling. The Zechiedrich group is very enthusiastic to know, in advance, the effects that DNA sequence-dependent topological properties might have on EcoRV and other DNA-binding proteins. Predicting the shape of a circular DNA sequence can help to determine if a protein can directly bind to the DNA at a specified binding site [1]. The residues near the binding site are affected by the overall shape and flexibility of the DNA as a whole [2, 3]. Figure 6.3.2 shows the Baylor sequence in its most relaxed form, with no end conditions placed on it, when the EcoRV protein has been bound at the pre-prescribed segment. In comparison to Figure 6.2.2 the sequence can be seen to have a distinct bend in the structure at the binding site of the EcoRV protein.

Figure 6.3.1: EcoRV, an endonuclease, bends DNA approximately 50° into the major groove at the TA step [4]. The purple ribbon is the enzyme, and the red, blue, green, and yellow rectangular slabs are respectively the A, T, G, and C bases. The green and cyan smooth cylindrical shapes are the backbones for the DNA double helix. This image is taken from the PDB [5].

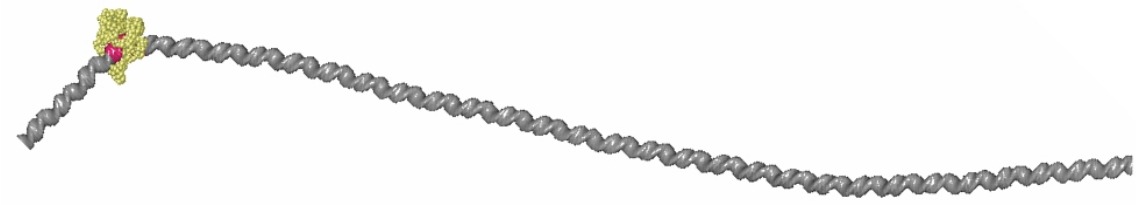


Figure 6.3.2: The 339 base-pair Baylor sequence discussed in the previous section without any forces acting upon it aside from the EcoRV protein that is bound on the left-hand-side of this figure. In comparison to Figure 6.2.2 the sequence can be seen to have a distinct bend in the structure at the binding site of the EcoRV protein. EcoRV is depicted with yellow spheres. The gray shiny tubular areas are plain DNA and the pink shiny tubular area is the part of the DNA sequence that is bound to the EcoRV protein. Notice how adding a protein bends this segment of DNA.

Figure 6.3.3 shows a closed configuration of the 339 base-pair Baylor sequence without bound EcoRV. The portions that are of interest are the attb and attp sites, highlighted in purple and yellow respectively, which are relics from the method of creating the 339 base-pair mini-circle in the wet lab. The third and final area, highlighted in bright blue, is the EcoRV protein binding site. This figure-8 has the lowest energy for $Lk = 31$ in Figure 6.2.5(i). The location of the EcoRV binding site in this low energy configuration, denoted by the bright-blue area, lies at a point of high curvature (at the bottom right). This is noteworthy since there is a high degree of bending where the EcoRV and DNA are joined.

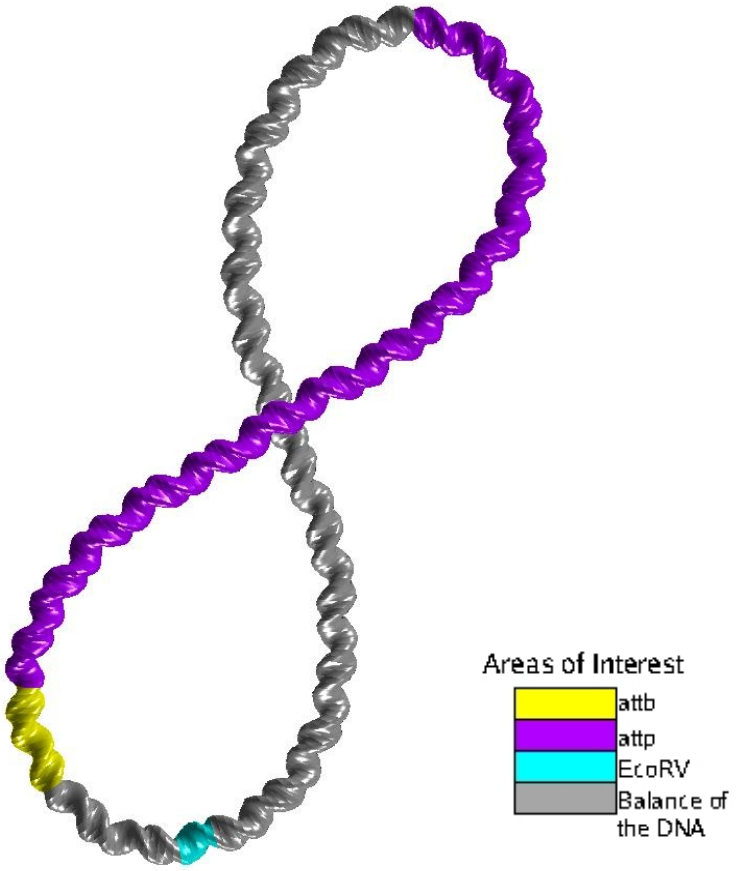


Figure 6.3.3: A figure 8 conformation of the Baylor sequence showing the areas of interest. This configuration has an elastic energy of 28.90 kT, Lk of 31, Wr of –0.73, and Tw^{SC} of 31.73. This configuration has the lowest minimum-energy for a structure with a Lk of 31. The attb and attp sites, highlighted in purple and yellow respectively, are relics from the method of creating the 339 base-pair mini-circle in the wet lab. The EcoRV binding site is highlighted in bright blue. The grey area is the rest of the 339 base-pair molecule.

6.3.1 Results for the 339 Base-Pair Baylor Sequence Bound to EcoRV

Closed structures of the 339 base-pair Baylor sequence with EcoRV bound to it were created in the same way as the closed structure of the naked 339 base-pair Baylor sequence. The ends were joined and minimum-energy converged configurations were sought. These were found using an automated iterative process that varies the initial guesses for the moments and forces. One at a time, one of the three components of the moments or the forces is adjusted minutely and then used by 3DNAdesigner until a converged configuration is found. The results presented in this Section were obtained using an automated script that put in the changing initial guesses over a period of two months. What came back were thousands of structures, of which only 45 were unique. The values for Tw^{SC}, energy, linking number, and writhe for each of the 45 converged structures can be seen in Section 6.5.2.

All of the unique structures are shown in Figures 6.3.4-6.3.7. In 3DNAdesigner these structures are shown with a plotting tool that can view and rotate the structure in 3-D, but in this thesis we can only display depictions in a 2-D manner. Therefore only one 2-D plot is shown for each structure. To get a better understanding of how the molecule really looks, one needs to view the structure with 3DNAdesigner's plotting program. Each structure shown displays the energy, *Lk*, and *Wr*. The figure is laid out over four pages so the most detailed views are possible. The figures are arranged from

the lowest to the highest energies as indicated with the arrows and the listed energies. Therefore, the lowest-energy structure is on the first page and located on the top left of Figure 6.3.4. Following this logic, the highest energy structure is located on the fourth page and located on the far right of Figure 6.3.7. The 2-D plots are displayed in a fashion to get the best angle and magnification that most clearly depicts the molecule while having the 2-D limitation. When one of the structures with a larger tubular area of DNA is shown, it is due to it being magnified to a point where the intricacies of the supercoiling can be better viewed.

The variation of the elastic energy with the values of *Lk* from the structures found in Figures 6.3.4-6.3.7 is shown in Figure 6.3.8. Here we see that the structure with the lowest energy has a linking number of 33, and that the structure with the highest energy has a linking number of 30. We also do not find any structures of relatively low energy with a linking number of 32, even though that was the linking number for most of the low energy structures in the naked 339 base-pair Baylor molecule. The energies of the configurations are computed using the same method as that for the naked DNA configurations with one exception. The base pairs on which a protein is bound contribute to the overall energy slightly differently. The energy for those "occupied" base pairs is taken to be that of "relaxed" DNA, as described in Section 5.1.2. This is because we do not know how much of the energy is relieved by the bound protein and how much is contributed by the DNA. So we make the assumption that when a protein is bound to

DNA, even though the DNA can be supercoiled beyond the relaxed state, that the combination of protein/DNA binding creates a minimum-energy state. We also note that the lowest energy structure seen here has an energy of 20.06kT, whereas the lowest energy arrangement of the naked structure from the previous section had an energy of 10.25kT. However, this could be due to not having good enough initial guesses for the moments and forces used as input and a lower energy structure could definitely be possible. What this does imply is that the binding of EcoRV to DNA does change the overall Tw^{SC} of the molecule and that, in turn, changes the linking number of the minimum-energy configuration from 32 to 33. This difference is clear from the plot of energy versus the linking number for the closed configurations of the Baylor sequence without EcoRV in Figure 6.2.5 versus those with bound EcoRV in Figure 6.3.8.

As in the previous section with the naked 339 base-pair Baylor molecule, Figure 6.3.9 shows the sequential variation of Tw^{SC} in the lowest and highest energy configurations that were obtained with 3DNAdesigner. We can see where the EcoRV protein is bound to the sequence from the sharp jump in Tw^{SC} near base pairs 37 to 46. The low-energy structure has an energy of 20.06kT with a *Lk* of 33 and the high-energy one has an energy of 332.35kT with a *Lk* of 30. As with Figure 6.2.8, the high-energy structure has values of Tw^{SC} that span a larger range of values than those of the low-energy structure. Also, just like the plot for naked DNA, the trend between the high- and

low-energy Tw^{SC} values generally follows the same pattern, but is not of the same magnitude.

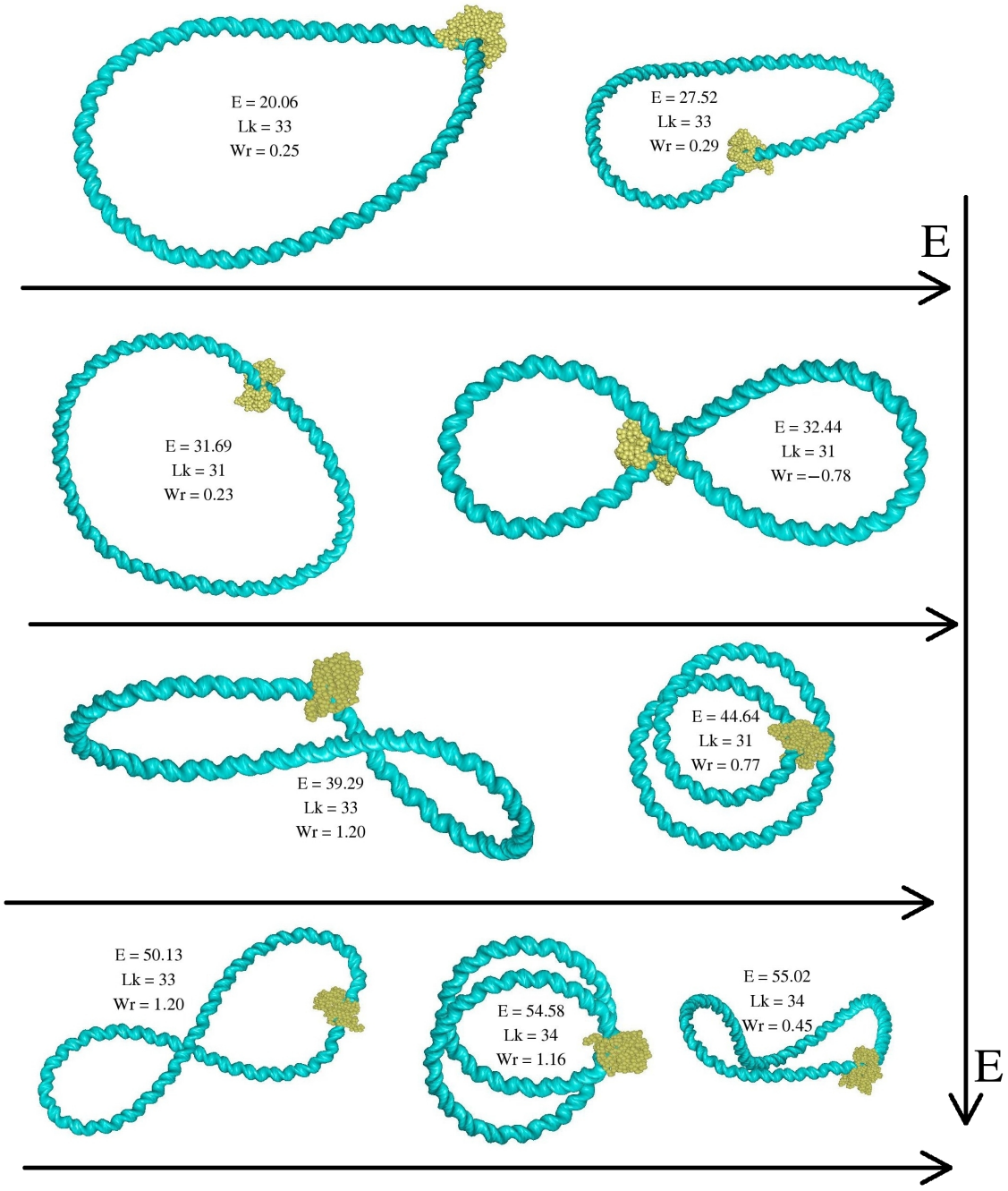


Figure 6.3.4: Part 1 of 4 (description is located on part 4 of 4).

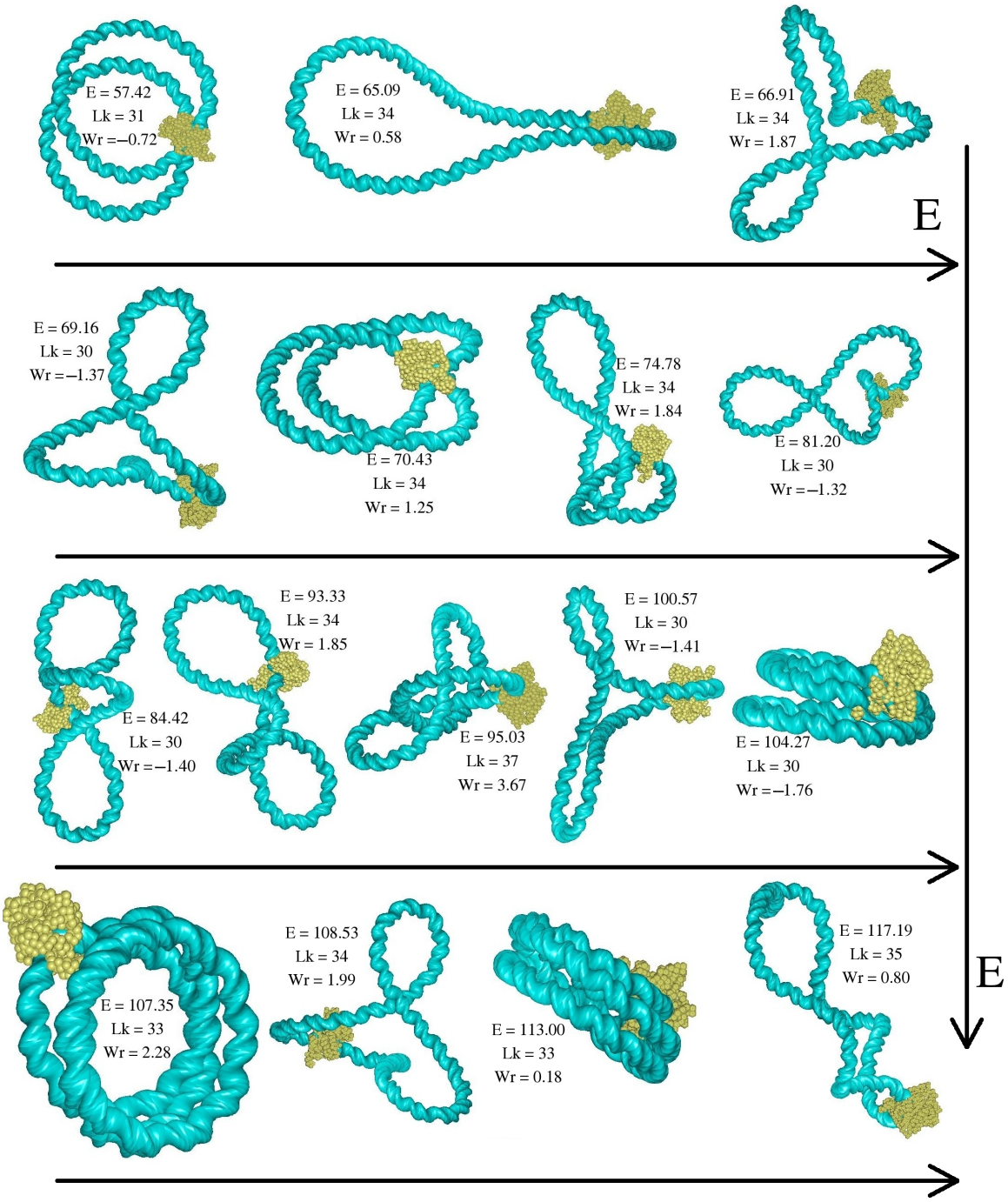


Figure 6.3.5: Part 2 of 4 (description is located on part 4 of 4).

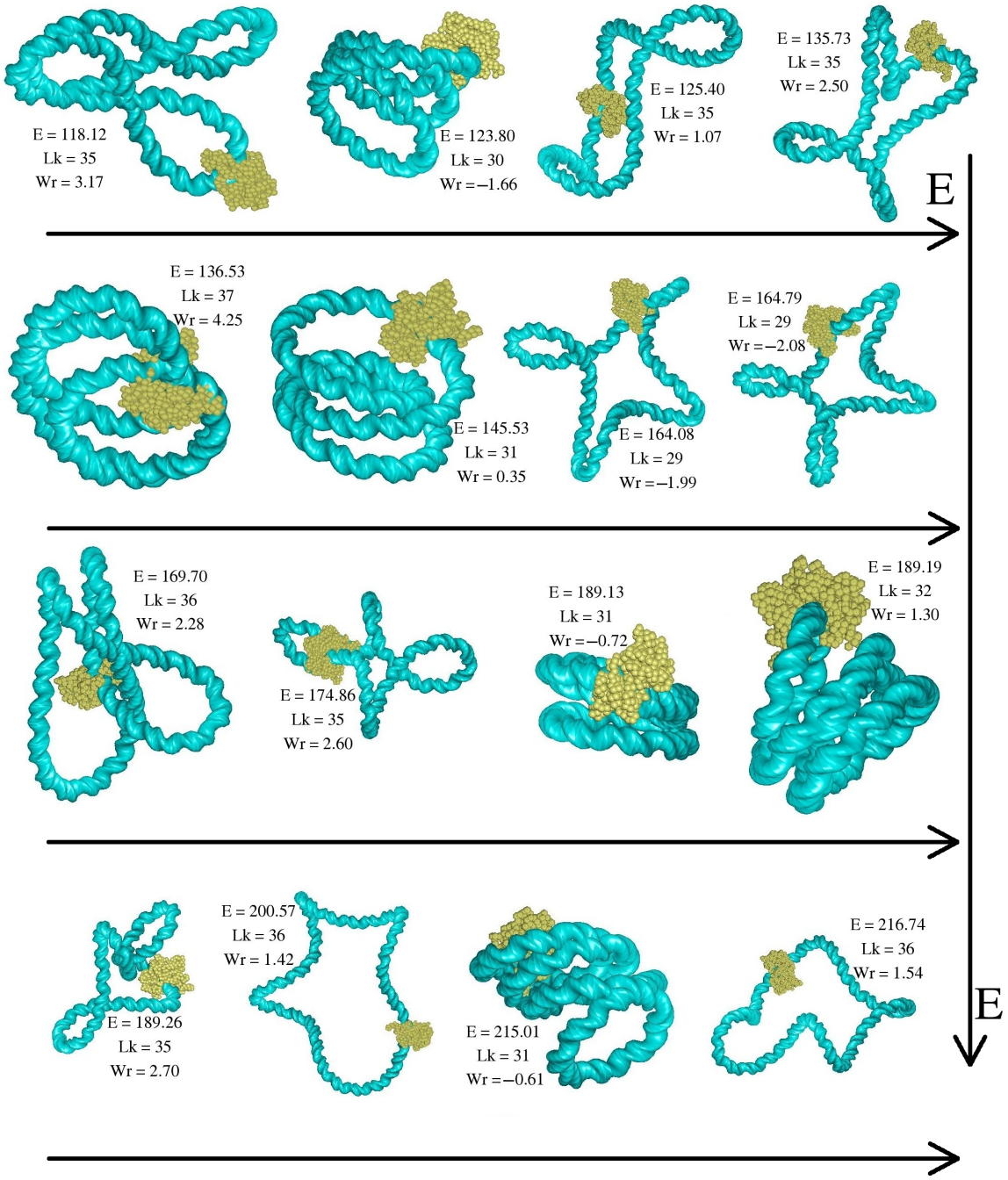


Figure 6.3.6: Part 3 of 4 (description is located on part 4 of 4).

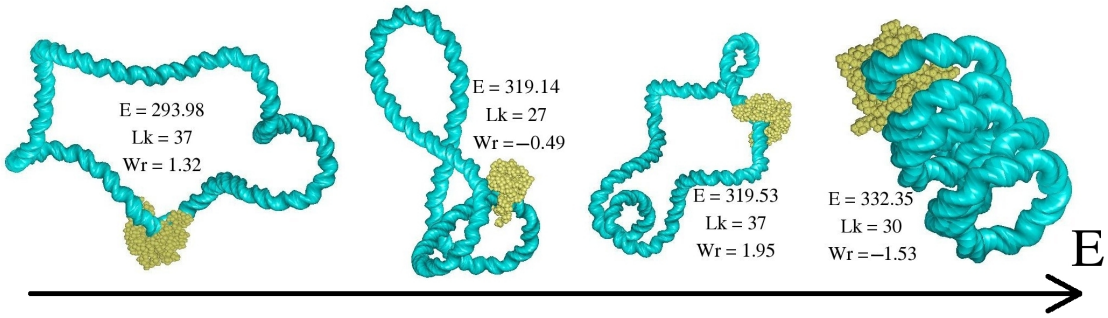
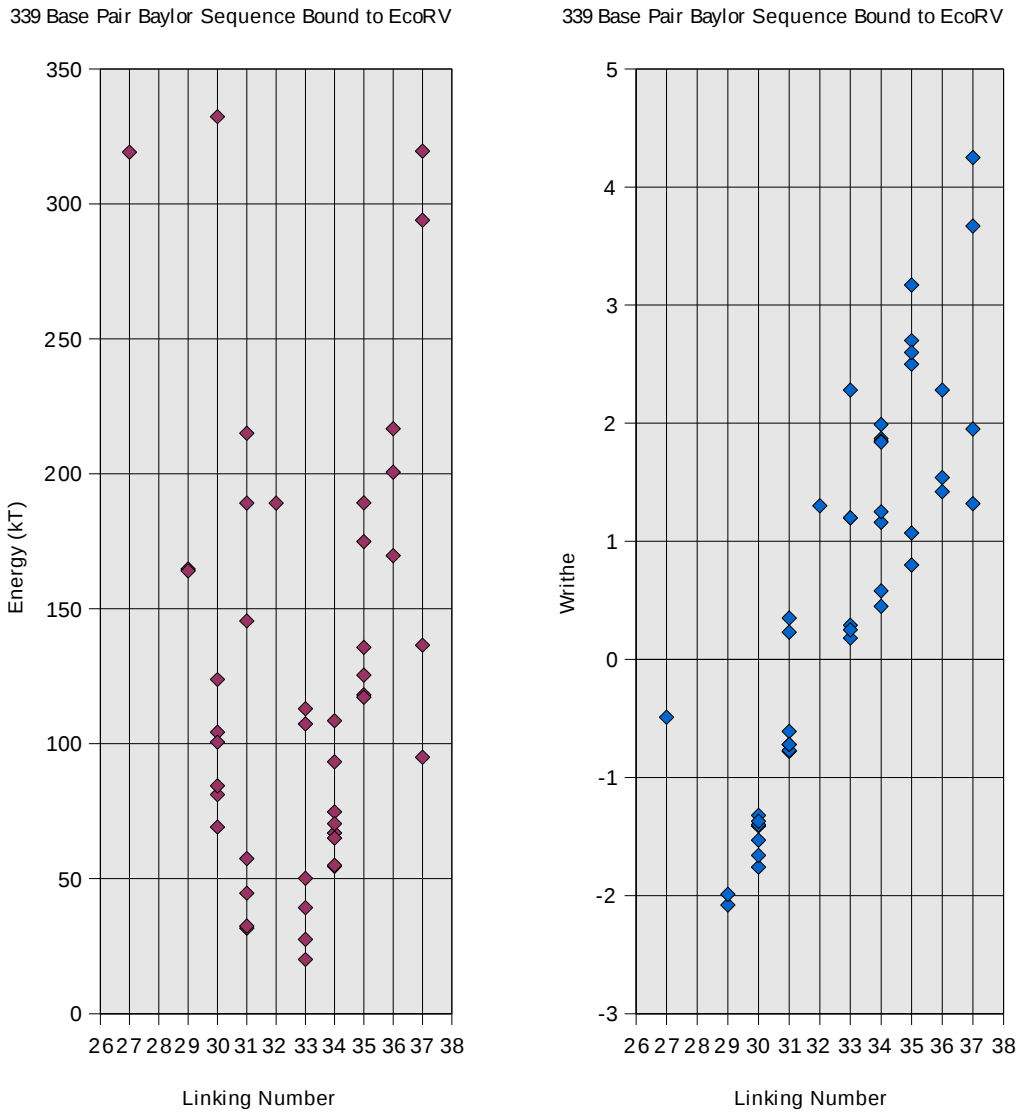
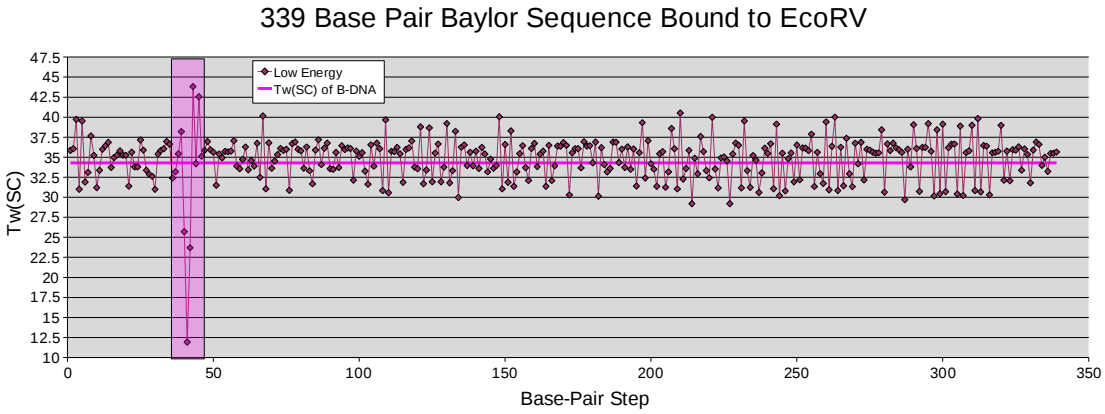


Figure 6.3.7: Part 4 of 4. This is a four-part figure depicting all of the unique minimum-energy configurations found while running 3DNAdesigner over a period of a few months, on the 339 base-pair Baylor sequence with the EcoRV bound to it and with the ends of the DNA joined together to form a closed structure. Each of the figures is labeled with its energy (E) in kT, Lk, and Wr. The configurations are sorted from left to right and top down in order of increasing energies.

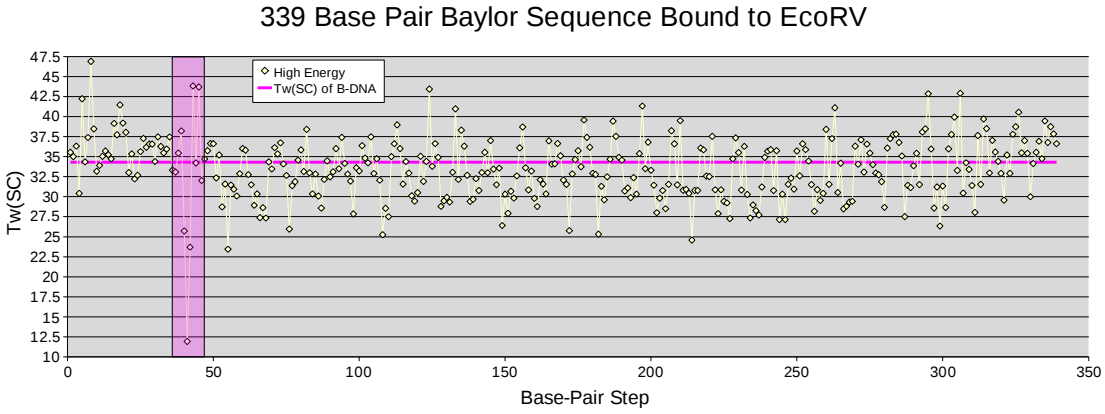


(i) *(ii)*

Figure 6.3.8: Plots of the variation of (i) energy and (ii) writhe with linking number of the 45 unique configurations of the 339 base-pair Baylor sequence bound to EcoRV found using 3DNAdesigner™. The value for each configuration is marked with a diamond. The lowest energy structures have a Lk of 33, as opposed to the Lk of 32 in the lowest energy structures of the naked 339 base-pair Baylor sequence in Figure 6.2.5. The plot shows, due to the large spread of data points, that there is a large range of Lk numbers possible for this molecule and that a few have low enough energies that a biologist may be interested in trying to interpret his/her data in terms of these supercoiled structures.



(i)



(ii)

Figure 6.3.9: Sequential variation of the twist of supercoiling along the chain contour of the (i) lowest and (ii) highest energy states of the 339 base-pair Baylor sequence bound to EcoRV found with 3DNAdesigner . As compared with the Tw^{SC} values for the naked sequence in Figure 6.2.8, we can tell that both of the configurations represented have lower twist fluctuations, in general, all the way around the structure. The EcoRV protein is bound to the DNA sequence between bases 37 to 46 (highlighted in pink).

6.4 16 Examples of Binding a 79 Base-Pair Sequence to the Lac Repressor

In this section we will show the effects of another E. coli protein, called the Lac repressor, on a small loop of DNA. In vivo, the Lac repressor binds to a fragment of DNA called the *lac* operon [6]. The *lac* operon is a piece of DNA that controls when RNA polymerase will be transcribed to make proteins involved in the metabolism of lactose. However, if lactose is not present, there is not a need to have those metabolites. In that case, the Lac repressor protein assembly binds to the *lac* operon so that RNA polymerase is not transcribed. RNA polymerase is thought to be unable to bind to DNA when the Lac repressor binds the *lac* operon. It is also thought that the amount of supercoiling that the Lac repressor imparts on the *lac* operon may control whether RNA polymerase attaches itself to the DNA. Therefore, the more tightly bound the Lac repressor is, the less chance that the transcription by polymerase will happen [7].

The Lac repressor protein binds to the DNA at two different, widely spaced locations and if bound to both sites produces DNA loops with four different orientations. This is discussed in Section 5.3.4.1 and can be seen in Figure 5.3.4.1.1. The orientation, the way the DNA enters and exits the protein, is described by the following labels: (i) A1, anti-parallel type 1; (ii) A2, anti-parallel type 2; (iii) P1, parallel type 1; and (iv) P2, parallel type 2. 3DNAdesigner's use of these orientations is shown in Chapter 5 and illustrated in Figure 5.3.4.1.1. All data discussed in this Section will be grouped by the

orientation of the DNA loop. In order to save time in generating the various DNA loops for the different types of molecules studied in this Section, 16 in all, each simulation was limited to three days of computer time to run. This approximation reduced the amount of data that could be collected and analyzed. If a researcher were more deeply interested in one or more of these molecules, running more iterations with a wider variety of moments and forces requiring longer computing time would be advisable.

In this study, each DNA segment bound to the Lac repressor in each of the four orientations consists of 79 base pairs. These base pairs correspond to the free bases found between the DNA operators bound to the Lac repressor in wild-type E. coli. The 79 base-pair sequence was treated as both a sequence-dependent molecule, meaning the sequence was a 'real' sequence consisting of A, C, G, and Ts, and a sequence-independent molecule, meaning the DNA was ideal and its energy deformations were independent of sequence. Each of these two sequences was studied both in its naked form and with an additional protein bound to it.

The sequence of the 79 base-pair loop is the same as that used by Swigon et al. [8,9] in modeling the Lac repressor-operator assembly. Here that sequence is AATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTT CCGGCTCGTATGTTGTGTGGAATT. This sequence does not include the DNA that is bound to either headpiece of the Lac repressor. In 3DNAdesigner™ the DNA that is

bound to (i.e., touches) the heads of the Lac repressor is taken from Swigon and Olson [8]. The model of the Lac repressor, determined by Swigon and Olson, is a composition of currently available X-ray crystal structures, including the low-resolution structure of the DNA-bound tetramer described by Lewis et al. [10], and the Lac repressor dimer complex with the O_{sym} operator [11]. The latter crystal structure only had one of the two head pieces bound to DNA. Swigon and Olson made the assumption that both head pieces would be similar in binding DNA so they superimposed one head piece on the other and calculated the positions of the DNA bound to the originally unbound Lac repressor head piece. 3DNAdesigner™ constrains the ends of the 79 base-pair sequence to fit against the segments of DNA that were contained in this model of the Lac repressor.

A sequence-independent molecule in 3DNAdesigner means the base-pair steps are found using average values for the base-pair-step parameters. The two types of sequence treatments will give further insight into how a particular DNA can guide a molecule into a lower or higher minimum-energy state just by the order and makeup of its base pairs.

6.4.1 Lac Repressor: Naked

Figure 6.4.1 shows the four types of orientations in which DNA can bind to the Lac repressor. In this figure the sequence used is the naked 79 base-pair sequence-independent DNA. Each of the four loops shown is representative of the lowest energy configuration that was found for that orientation. The dark blue spheres make up the Lac

repressor protein. This figure makes use of 3DNAdesigner's ability to highlight areas of interest. Of interest here are the beginning, or 5´-end, of the sequence, shown in green, and the terminal, or or 3´-end, of the sequence, shown in pink. This coloring shows how the sequence is threaded through the Lac repressor protein.

While the highlighted areas of interest in green and pink are the ends of our 79 base-pair sequence, they are not the true ends of the DNA shown in the figure. This is not a cause for concern. The gray areas preceding and trailing the areas of interest, that lead to the sheared-off section, are just the part of the DNA that is bound to the Lac repressor protein. 3DNAdesigner groups the Lac repressor protein into the category of a DNA anchored to points in space. In this case, the points in space are the terminal base pairs of the DNA that is included in the pdb file of the Lac repressor. The 79 base-pair sequence does not count or include in any way those gray areas that are actually touching the Lac repressor. That is why the Lac repressor is under the anchoring wizard in 3DNAdesigner and is unique among the proteins that can bind to DNA with this program. This is because the Lac repressor has two binding spots, and therefore, is treated differently. There are other proteins besides the Lac repressor that bind DNA at two or more widely spaced sites. The configurations of the intervening DNA can also be determined with 3DNAdesiger if the anchoring conditions of the bound DNA are known.

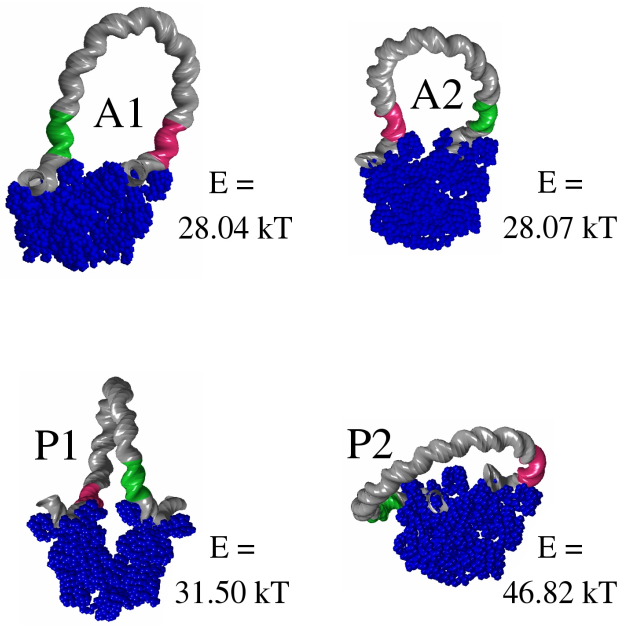


Figure 6.4.1: Depiction of the lowest energy structures for the 79 base-pair sequence-independent DNA segment bound to the Lac repressor protein in each of the four possible orientations; A1, A2, P1, P2. The blue spheres represent the Lac repressor, the gray shiny tube is the DNA, the green shiny tube is the start of the 79 base-pair sequence and the pink shiny tube is the end of the sequence.

6.4.2 Lac Repressor: DNA Bound with HU

Figure 6.4.3 shows the lowest energy structures of a 79 base-pair sequence-dependent DNA segment anchored in each of the four orientations possible to the Lac repressor and also bound to the architectural prokaryotic protein called HU. The HU protein-DNA complex used in the following calculations was obtained from the cocrystal structure of the *Anabaena* HU (AHU2) protein and DNA determined by Swinger et al. (PDBID 1P78) [12]. HU is a prokaryotic analog of the nucleosome studied in Chapter 4 and is responsible for the packaging of DNA into a compact state, or nucleoid, in prokaryotes. The protein creates bends of 105-140 degrees in DNA [12]. A rendering of one of the four known crystal structures of HU bound to DNA is found in Figure 6.4.2.

HU is a nonspecific binding protein that can bind to any site on DNA, but has a possible preference for A-T rich DNA [15,16]. We have HU bound to the DNA sequence in one or two places depending on the orientation of the Lac repressor. When DNA is bound in the A1 or A2 orientations, only one HU is bound, at position 48 or 31, respectively. The positions refer to the location of the center of the HU-bound fragment on the 79 base-pair DNA segment anchored to the Lac repressor. The DNA found in P1 and P2 loops binds to two HU proteins. The centers of the HUs on the P1 loops bind to positions 23 and 57 on the DNA sequence, and those in the P2 loops bind to positions 19 and 70. In Figure 6.4.3 the dark blue assembly of spheres is the Lac repressor, the yellow

and pink assemblies are the HU proteins, and the shiny cyan tube is the DNA , which includes the base pairs that are bound to the Lac repressor and HU.

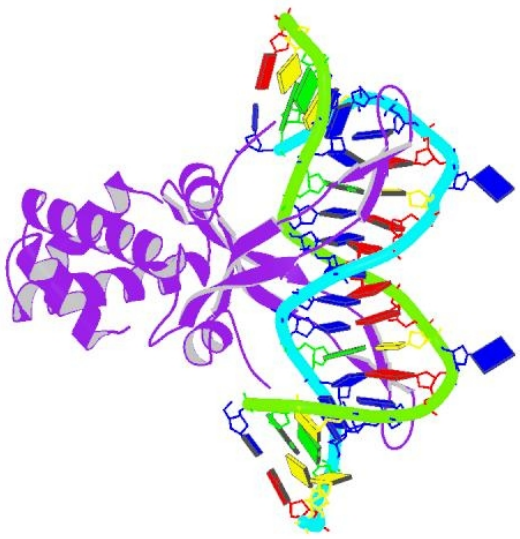


Figure 6.4.2: Image showing the HU protein bound to a DNA oligomer [12]. The DNA bends sharply around the HU protein. The protein is represented by the ribbon like curves, lines, and arrows in purple. The double helical DNA backbones are the thicker curves in cyan and green. The base pairs consist of adenine in red, thymine in blue, guanine in green, and cytosine in yellow. This image is taken from the NDB [13].

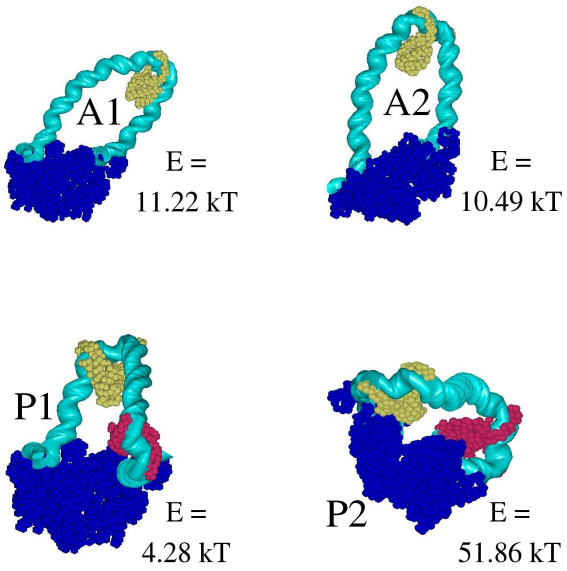


Figure 6.4.3: Depiction of the lowest energy structures for the 79 base-pair sequence-dependent DNA segment in the presence of the HU protein and bound to the Lac repressor protein in each of the four possible orientations; A1, A2, P1, P2. The blue spheres represent the Lac repressor, the yellow and pink spheres are the HU proteins, and the blue shiny tube is the DNA.

6.4.3 Results for the Lac Repressor

In the following sections we discuss the results for all 16 combinations of DNA orientations, base-pair-step energy treatments, and HU presence or absence on the wild-type 79 base-pair DNA sequences, anchored to the Lac repressor. Tables 6.4.1 - 6.4.2 contain the energies and topological parameters for the minimum-energy structures found for each unique combination. Figures 6.4.4 - 6.4.7 depict the results for every unique structure found. The figures group the data according to the Lac repressor orientation. The Lac repressor presents a small hurdle to our calculation of the *Wr* and *Lk*. To get an integral *Lk,* closed DNA segments are required. DNA bound to the Lac repressor head pieces do not have their ends connected to each other. To compensate for the "missing" closing step we have added a "fake" closing step. This closing step is created by joining the origin of the first base pair of the 79 base-pair sequence to that of the last base pair of the sequence.

Table 6.4.1 contains the energies and topological parameters of all eight combinations of DNA orientation and HU presence or absence found for a sequence-independent model of DNA and Table 6.4.2 has all the corresponding data found for a sequence-dependent model. Each table is grouped by the orientation of DNA on the Lac repressor (A1, A2, P1, and P2) and is further separated by whether the sequence is naked or contains bound HU. For both types of DNA, we see that adding HU consistently

lowers the elastic energy of the entire structure. Some of this lowering comes from the shortened length of the deformable DNA in the HU-bound versus free DNA loops. In the case of the HU-bound A1 and A2 loops the proteins are at the sites of highest curvature at the apexes of the loops and contribute substantially to the reduction in energy. An especially drastic change in energy, however, happens when HU is added to DNA in a P1 orientation. This difference stems from the straightening of DNA between the bound HU molecules. We can also see that the P2 loops are considerably higher in energy than the other loops. Comparison of the two tables also shows that the sequence-dependent models of DNA are consistently higher in energy than their sequence independent counterparts.

These results can be compared to the data obtained in another similar study. For a collation of the minimum-energy results from this Section and those by the similar studies of Swigon et. al. [8,9], please refer to Table 6.4.3. From this table it is clear that the trend for the lowest energy DNA-Lac repressor combinations, in the absence of HU, would most likely be antiparallel. This finding matches what we can conclude by looking at the energies presented in Tables 6.4.1 - 6.4.3. It should be noted that the “fake” closing step used to determine these topological parameters, as discussed above, differs from that of Swigon et. al. In the study by Swigon et. al., the closing connection is made through the two halves of the protein structure, rather than between the first and last base pairs of the sequence.

Lac Repressor with 79 Base Pair Independent Sequence

		Energy	Lk	Wr	Tw(SC)
A1	Naked	28.0374	7	0.0933	6.9067
	HU	9.9244	7	–0.0086	7.0086
A2	Naked	28.0692	7	0.0926	6.9074
	HU	9.1963	7	–0.0033	7.0033
P1	Naked	31.5049	7	0.0437	6.9563
	HU	2.5234	7	–0.0687	7.0687
P2	Naked	46.8203	8	0.1518	7.8482
	HU	41.0067	8	0.4420	7.5580

Table 6.4.1: Elastic energy and topological parameters of a 79 base-pair DNA loop, modeled as an ideal sequence-independent chain and anchored in each of four orientations on the Lac repressor with and without bound HU. These data correspond to configurations of lowest energy found using 3DNAdesigner. The results are grouped by the Lac repressor orientation and whether or not the HU protein was bound on the DNA sequence. The values of the energy, Lk, Wr, and Tw^{SC} are presented.

Lac Repressor with 79 Base Pair Dependent Sequence

		Energy	Lk	Wr	Tw(SC)
A1	Naked	29.2723	8	0.0742	7.9258
	HU	11.2232	7	–0.0314	7.0314
A2	Naked	32.8730	7	0.0390	6.9610
	HU	10.4865	7	–0.0634	7.0634
P1	Naked	38.6244	7	0.0267	6.9733
	HU	4.2831	7	–0.0780	7.0780
P2	Naked	53.4904	7	0.1851	6.8149
	HU	51.8574	8	0.4832	7.5168

Table 6.4.2: Elastic energy and topological parameters of a 79 base-pair DNA loop, modeled as an ideal sequence-dependent chain and anchored in each of four orientations on the Lac repressor with and without bound HU. These data correspond to configurations of lowest energy found using 3DNAdesigner. The results are grouped by the Lac repressor orientation and whether or not the HU protein was bound on the DNA sequence. The values of the energy, Lk, Wr, and Tw^{SC} are presented.

	Data Source	E (kT)	Lk
A1	**(a)**	29.3	8
	(b)	31.8	8
	(c)	32.1	9
A2	**(a)**	32.9	7
	(b)	32.9	8
	(c)	33.1	8
P1	**(a)**	38.6	7
	(b)	38.6	9
	(c)	38.8	9
P2	**(a)**	53.5	7
	(b)	45.5	9
	(c)	45.7	10

Table 6.4.3: This table holds the values for the elastic energy, E, and Lk. These are the reported values for the minimum-energy configurations found for the naked wild-type DNA (sequence-dependent) 79 base-pair sequence. The data source indicates (a) findings by Lauren A. Britton reported in this Chapter, (b) findings by Swigon et al in Proc. Natl. Acad. Sci. USA [9], and (c) findings by Swigon and Olson in Int. J. Non-linear Mech. [8].

6.4.3.1 Lac Repressor with A1 and A2 Orientations

Figure 6.4.4 depicts the variation in elastic energy with linking number of DNA looped in an A1 orientation on the Lac repressor for the four possible combinations of DNA model and HU presented. Here it is seen that the lowest energy structures have a linking number of either 7 or 8, where the value depends on the specific combination of DNA model and HU. Figure 6.4.5 contains similar data for DNA looped to the Lac repressor in the A2 orientation. Here we see the same trend as with A1. The energy is lowest when *Lk* is either 7 or 8, depending on the particular combination of DNA model and HU. It is interesting to note that the lowest energy state of the A1 and A2 loops have the same *Lk* for the same combination of DNA model and HU. For example, the linking number is 7 for the minimum-energy configuration of A1 and A2 loops of the sequence-independent DNA with bound HU.

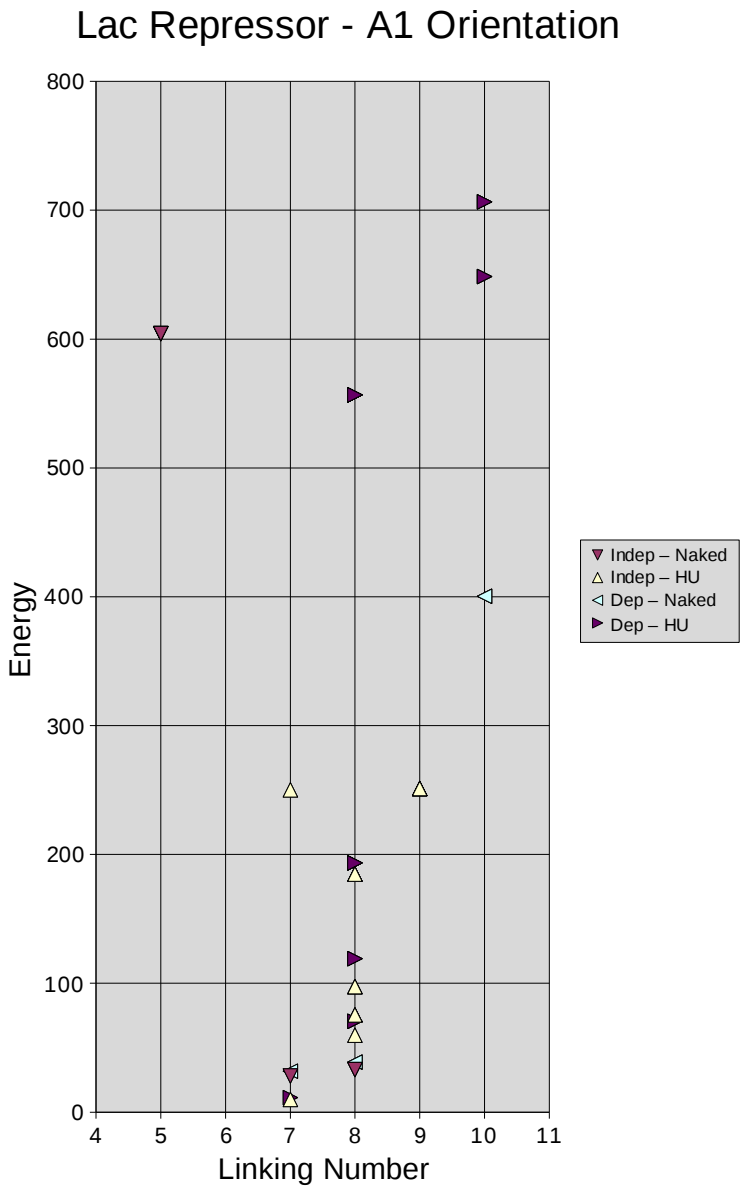


Figure 6.4.4: Plot of the variation of the energy with the linking number for four configurations of the 79 base-pair DNA loop in the A1 orientation found using 3DNAdesigner™. The maroon downward pointing triangles describe the naked sequence-independent DNA, the yellow upward triangles the sequence-independent DNA with HU bound to it, the cyan leftward triangles the naked sequence-dependent DNA, and the purple right pointing triangles the sequence-dependent DNA with HU bound to it.

Lac Repressor - A1 Orientation Linking Numbers sorted by Energy				
Energy	**Indep – Naked**	**Indep – HU**	**Dep – Naked**	**Dep – HU**
9.92		7		
11.22				7
28.04	7			
31.84			7	
33.12	8			
38.91			8	
59.81		8		
70.52				8
75.42		8		
97.41		8		
119.17				8
184.75		8		
193.46				8
250.24		7		
251.08		9		
400.41			10	
556.64				8
604.23	5			
648.53				10
706.5				10

Table 6.4.4: Data used to plot Figure 6.4.4 showing the variation of the energy with the linking numbers for four configurations of the 79 base-pair DNA loop in the A1 orientation found using 3DNAdesigner™. The data show all unique combinations of DNA energy modeled and HU presence or absence, sorted by the energy.

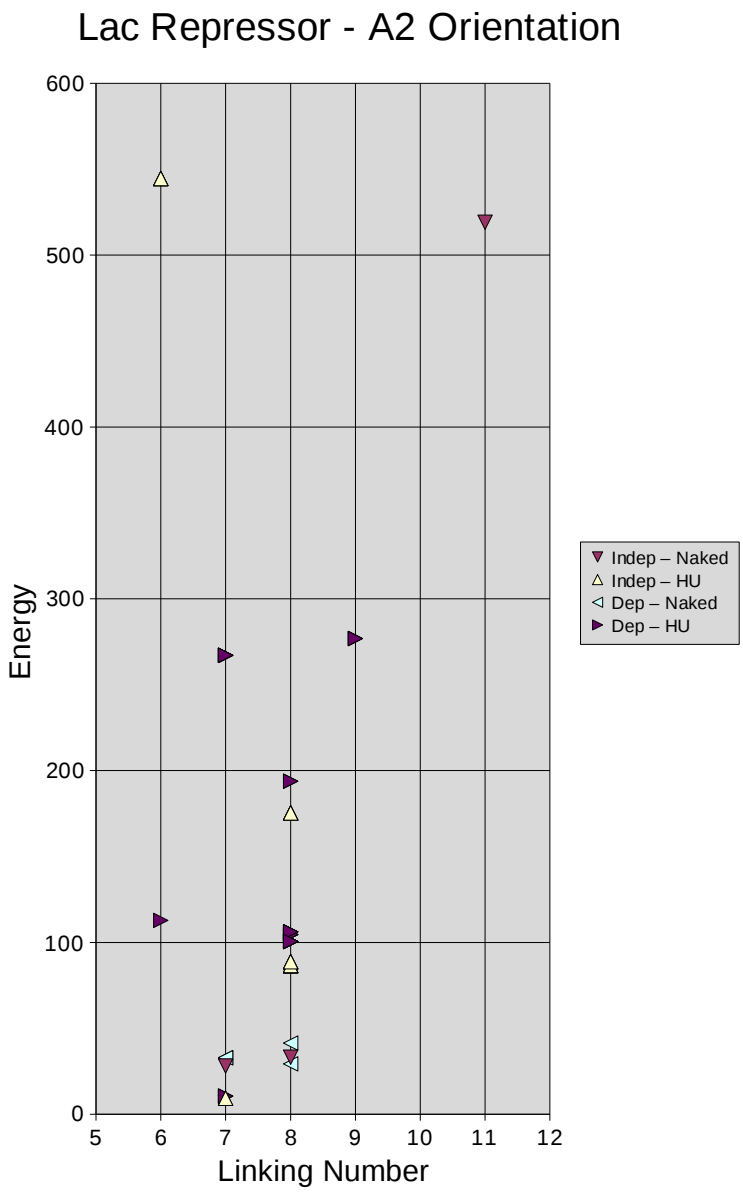


Figure 6.4.5: Plot of the variation of the energy with the linking number for four configurations of the 79 base-pair DNA loop in the A2 orientation found using 3DNAdesigner™. The maroon downward pointing triangles describe the naked sequence-independent DNA, the yellow upward triangles the sequence-independent DNA with HU bound to it, the cyan leftward triangles the naked sequence-dependent DNA, and the purple right pointing triangles the sequence-dependent DNA with HU bound to it.

Lac Repressor - A2 Orientation Linking Numbers sorted by Energy				
Energy	**Indep – Naked**	**Indep – HU**	**Dep – Naked**	**Dep – HU**
9.2		7		
10.49				7
28.07	7			
29.27			8	
32.87			7	
33.19	8			
41.38			8	
86.44		8		
86.65		8		
88.7		8		
100.54				8
104.52				8
106.17				8
112.73				6
175.38		8		
193.87				8
267.13				7
276.84				9
519.16	11			
544.68		6		

Table 6.4.5: Data used to plot Figure 6.4.5 showing the variation of the energy with the linking numbers for four configurations of the 79 base-pair DNA loop in the A2 orientation found using 3DNAdesigner™. The data show all unique combinations of DNA energy modeled and HU presence or absence, sorted by the energy.

6.4.3.2 Lac Repressor with P1 and P2 Orientations

Figures 6.4.6 and 6.4.7 hold the data for the DNA loops oriented respectively in P1 and P2 arrangements on the Lac repressor. The lowest energy configurations occur for loops with a *Lk* of 7 or 8, just like A1 and A2. However, unlike the anti-parallel orientations, the *Lk* associated with the lowest energy states differs in the parallel orientations. For example, the linking number of the P1 loop made up of the sequence-independent DNA and bound HU is 7, but for the P2 loop under the same conditions is 8.

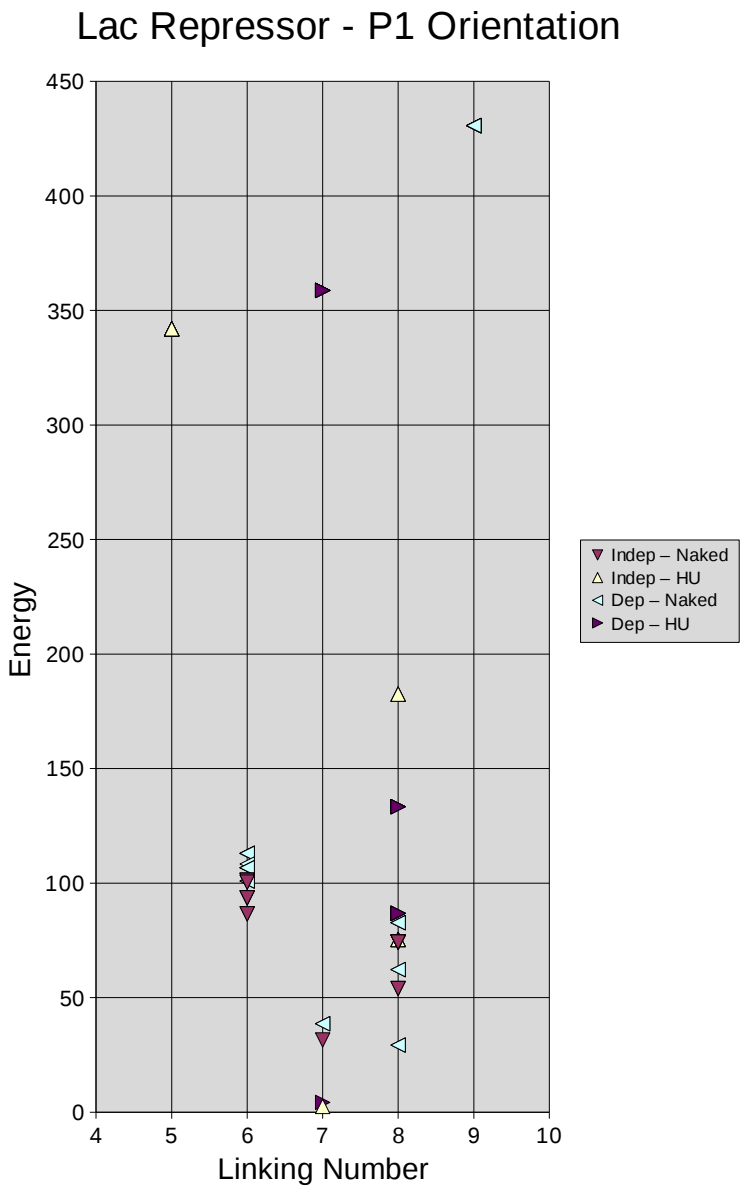


Figure 6.4.6: Plot of the variation of the energy with the linking number for four configurations of the 79 base-pair DNA loop in the P1 orientation found using 3DNAdesigner™. The maroon downward pointing triangles describe the naked sequence-independent DNA, the yellow upward triangles the sequence-independent DNA with HU bound to it, the cyan leftward triangles the naked sequence-dependent DNA, and the purple right pointing triangles the sequence-dependent DNA with HU bound to it.

Lac Repressor - P1 Orientation Linking Numbers sorted by Energy				
Energy	**Indep – Naked**	**Indep – HU**	**Dep – Naked**	**Dep – HU**
2.52		7		
4.28				7
29.27			8	
31.5	7			
38.62			7	
53.97	8			
62.28			8	
74.2	8			
75.36		8		
82.71			8	
86.56	6			
86.81				8
93.59	6			
100.47	6			
100.93			6	
101.23	6			
106.78			6	
108.32			6	
112.96			6	
133.35				8
182.56		8		
342.01		5		
358.75				7
430.74			9	

Table 6.4.6: Data used to plot Figure 6.4.6 showing the variation of the energy with the linking numbers for four configurations of the 79 base-pair DNA loop in the A2 orientation found using 3DNAdesigner™. The data show all unique combinations of DNA energy modeled and HU presence or absence, sorted by the energy.

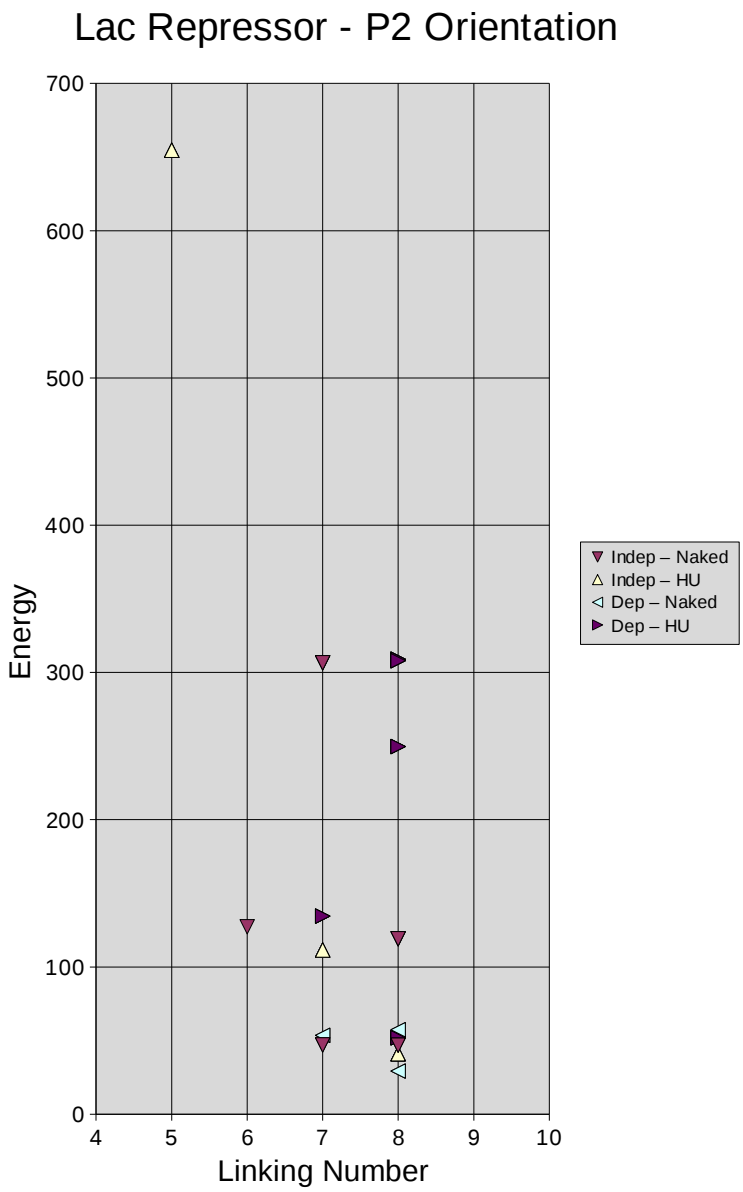


Figure 6.4.7: Plot of the variation of the energy with the linking number for four configurations of the 79 base-pair DNA loop in the P2 orientation found using 3DNAdesigner™. The maroon downward pointing triangles describe the naked sequence-independent DNA, the yellow upward triangles the sequence-independent DNA with HU bound to it, the cyan leftward triangles the naked sequence-dependent DNA, and the purple right pointing triangles the sequence-dependent DNA with HU bound to it.

Lac Repressor - P2 Orientation Linking Numbers sorted by Energy				
Energy	**Indep – Naked**	**Indep – HU**	**Dep – Naked**	**Dep – HU**
29.27			8	
41.01		8		
46.82	8			
47.09	7			
51.86				8
53.49			7	
57.3			8	
111.47		7		
119.07	8			
127.22	6			
134.43				7
249.6				8
306.58	7			
307.98				8
308.94				8
654.53		5		

Table 6.4.7: Data used to plot Figure 6.4.7 showing the variation of the energy with the linking numbers for four configurations of the 79 base-pair DNA loop in the A2 orientation found using 3DNAdesigner™. The data show all unique combinations of DNA energy modeled and HU presence or absence, sorted by the energy.

6.5 Appendix A – Data

6.5.1 Data for the 117 unique configurations of the naked 339 base-pair Baylor sequence

Tw^{SC}	Energy (kT)	Lk	Wr
34.75	148.14	34	-0.75
32.1	70.17	34	1.9
31.38	85.92	30	-1.38
32.4	171.56	35	2.6
32.1	92.79	34	1.9
31.77	54.1	31	-0.77
29.58	165.14	28	-1.57
30.36	172.15	28	-2.36
33.95	150.29	34	0.05
33.2	151	38	4.8
28.32	210.66	28	-0.32
33.09	165.08	38	4.92
28.91	121.23	29	0.09
31.69	206.82	29	-2.69
29.63	100.64	29	-0.63
30.92	146.45	29	-1.92
31.73	28.9	31	-0.73
31.5	85.44	30.12	-1.38
29.75	58.91	30	0.25
31.61	293.98	28	-3.61
30.06	95.23	27	-3.06
28.6	158.03	26	-2.6
31.74	117.85	32	0.26
31.41	81.05	30	-1.41
28.15	205.12	28	-0.15
33.29	98.74	37	3.71

Tw^{SC}	Energy (kT)	Lk	Wr
32.78	123.22	33	0.22
33.32	97.41	35	1.68
30.69	111.41	33	2.31
31.35	77.94	30	-1.34
31.79	127.75	31	-0.79
33.08	165.28	36	2.92
30.04	102.64	27	-3.04
31.73	133.51	31	-0.73
30.69	257.48	28	-2.69
31.79	10.25	32	0.21
35.04	192.31	36	0.96
29.6	163.29	28	-1.6
30.73	29.02	31	0.27
31.11	158.44	29	-2.11
31.39	83.56	30	-1.39
31.8	14.98	32	0.2
30.93	152.25	29	-1.93
31	153.37	29	-2
31.76	31.02	31	-0.76
30.67	247.95	28	-2.67
30.72	53.72	32	1.28
31.71	147.14	33	1.3
29.79	63.52	30	0.21
31.01	154.33	29	-2.01
31.75	44.62	31	-0.75
31.38	84.44	30	-1.38
32.12	89.75	34	1.88
29.54	163.66	26	-3.54
30.6	247.14	28	-2.6
29.61	163.63	32	2.39
33.62	57.29	34	0.38
33.29	100.3	37	3.71

Tw^{SC}	Energy (kT)	Lk	Wr
30.07	95.23	27	-3.07
34.78	141.62	36	1.22
31.76	41.74	33	1.24
31.64	319.41	32	0.36
32.81	21.31	33	0.19
31.01	154.49	29	-2.01
31.35	77.94	30	-1.35
30.6	300.58	31	0.4
32.07	74.77	34	1.93
31.68	186.58	31	-0.68
27.45	315.86	27	-0.45
34.22	112.61	35	0.78
31.03	155.47	29	-2.03
31.42	82.55	30	-1.42
29.03	123.59	29	-0.03
29.68	92.79	27	-2.68
31.32	85.18	30	-1.32
30.68	135.56	29	-1.68
29.6	164.27	30	0.4
30.71	69.81	30	-0.71
32.66	239.31	36	3.34
33.73	170.92	36	2.27
32.39	140.16	35	2.61
33.73	54.85	34	0.27
33.07	156.65	36	2.93
30.65	193.59	34	3.35
34.79	192.03	36	1.21
34.5	203.04	36	1.5
34.19	115.8	35	0.81
30.59	248.26	28	-2.59
32.46	773.06	32	-0.46
32.79	116.22	33	0.21

Tw^{SC}	Energy (kT)	Lk	Wr
30.63	225.89	30	-0.63
32.37	142.21	35	2.63
32.37	166.39	35	2.63
33.1	157.26	36	2.9
30.7	257.52	28	-2.7
27.7	320.69	27	-0.7
29.52	164	28	-1.52
32.1	215.95	36	3.9
32.83	61.8	32	-0.83
30.98	154.86	29	-1.98
31.38	85.44	30	-1.38
31.02	155.11	29	-2.02
32.81	23.73	33	0.19
29.84	92.91	27	-2.84
31.38	84.45	30	-1.38
31.41	82.53	30	-1.41
31.69	131.92	31	-0.69
31.72	103.41	32	0.28
32.75	208.94	32	-0.75
32.81	56.84	32	-0.81
33.81	153.67	36	2.19
32.97	166.71	34	1.03
33.08	157.63	36	2.92
32.85	244.01	33	0.15
33.83	148.22	36	2.17
33.68	162.06	36	2.32
31.31	75.19	30	-1.31
31.03	156.03	29	-2.03
30.11	95.37	27	-3.11
29.88	101.06	27	-2.88
31.77	38.53	31	-0.77

6.5.2 Data for the 45 unique configurations of the 339 base-pair Baylor sequence bound to EcoRV

Tw^{SC}	Energy (kT)	Lk	Wr
31.83	118.12	35	3.17
30.77	31.69	31	0.23
31.78	32.44	31	-0.78
32.5	135.73	35	2.5
31.8	39.29	33	1.2
31.76	104.27	30	-1.76
31.41	100.57	30	-1.41
32.13	66.91	34	1.87
33.34	95.03	37	3.67
31.77	44.64	31	-0.77
32.82	113	33	0.18
31.53	332.35	30	-1.53
34.46	216.74	36	1.54
32.72	27.52	33	0.29
33.93	125.4	35	1.07
31.32	81.2	30	-1.32
31.61	215.01	31	-0.61
32.3	189.26	35	2.7
31.8	50.13	33	1.2
33.42	65.09	34	0.58
31.72	57.42	31	-0.72
32.75	136.53	37	4.25
33.72	169.7	36	2.28
32.75	20.06	33	0.25
32.84	54.58	34	1.16
34.58	200.57	36	1.42
31.72	189.13	31	-0.72
31.4	84.42	30	-1.4

Tw^{SC}	Energy (kT)	Lk	Wr
32.15	93.33	34	1.85
32.4	174.86	35	2.6
31.08	164.79	29	-2.08
32.16	74.78	34	1.84
27.49	319.14	27	-0.49
35.68	293.98	37	1.32
34.2	117.19	35	0.8
32.75	70.43	34	1.25
32.01	108.53	34	1.99
35.05	319.53	37	1.95
30.72	107.35	33	2.28
30.7	189.19	32	1.3
33.55	55.02	34	0.45
31.66	123.8	30	-1.66
31.37	69.16	30	-1.37
30.65	145.53	31	0.35
30.99	164.08	29	-1.99

6.6 References

[1] Y. Zhang, Z. Xi, R.S. Hegde, Z. Shakked, and D.M. Crothers. (2004). Predicting Indirect Readout Effects in Protein-DNA Interactions. *Proc. Natl. Acad. Sci. USA*, **101(22)**, 8337-8341.

[2] S.A. Mauro and G.B. Koudelka. (2004). Monovalent Cations Regulate DNA Sequence Recognition by 434 Repressor. *J. Mol. Biol.*, **340(3)**, 445-457.

[3] S.A. Mauro, D. Pawlowski, and G.B. Koudelka. (2003). The Role of the Minor Groove Substituents in Indirect Readout of DNA Sequence by 434 Repressor. *J. Biol. Chem.*, **278(15)**, 12955-12960.

[4] N.C. Horton, and J.J. Perona. (2000). Crystallographic Snapshots Along a Protein Induced DNA-bending Pathway. *Proc. Natl. Acad. Sci. USA*, **97(11)**, 5729-5734.

[5] N.C. Horton, and J.J. Perona. (2012). RCSB PDB - Images for 1EOP. Available at *http://www.rcsb.org/pdb/explore/images.do?structureId=1EOP.*

[6] B. Müller-Hill. (1996). *The lac Operon*. Walter De Gruyter & Co. Berlin, Germany.

[7] C.L. Lawson, D. Swigon, K. Murakami, S.A. Darst, H.M. Ebright, and R.H. Ebright. (2004). Catabolite Activator Protein (CAP): DNA Binding and Transcription Activation. *Curr. Opin. Struct. Bio.*, **14(1)**, 1–11.

[8] D. Swigon, and W.K. Olson. (2008). Mesoscale Modeling of Multi-Protein–DNA Assemblies: The Role of the Catabolic Activator Protein in Lac-Repressor-Mediated Looping. *Int. J. Non-linear Mech.*, **43(10)**, 1082-1093.

[9] D. Swigon, B.D. Coleman, and W.K. Olson. (2006). Modeling the Lac Repressor-Operator Assembly: The Influence of DNA Looping on Lac Repressor Conformation. *Proc. Natl. Acad. Sci. USA*, **103(26)**, 9879–9884.

[10] M. Lewis, G. Chang, N.C. Horton, M.A. Kercher, H.C. Pace, M.A. Schumacher, R.G. Brennan, and P. Lu. (1996). Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. *Science*, **271(5253)**, 1247-1254.

[11] C.E. Bell, and M. Lewis. (2000). A Closer View of the Conformation of the Lac Repressor Bound to Operator. *Nat. Struct. Biol.*, **7(3)**, 209-214.

[12] K.S. Swinger, K.M. Lemberg, Y. Zhang, and P.A. Rice. (2003). Flexible DNA Bending in HU-DNA Cocrystal Structures. *EMBO J.*, **22(14),** 3749-3760.

[13] K.S. Swinger, K.M. Lemberg, Y. Zhang, and P.A. Rice. (2012). PD0430: Biological Assembly. Available at *http://ndbserver.rutgers.edu/atlas/xray/structures/P/pd0430/PD0430-biol1.html.*

[14] J.M. Fogg, N. Kolmakova, I. Rees, S. Magonov, H. Hansma, J.J. Perona, and E.L. Zechiedrich. (2006). Exploring Writhe in Supercoiled Minicircle DNA. *J. Phys. Condens. Matter.*, **18(14),** S145-S159.

[15] S. Semsey, K. Virnik and S. Adhya. (2005). A Gamut of Loops: Meandering DNA. *Trends Biochem. Sci.*, **30(6)**, 334-341.

[16] M.Y. Tolstorukov, K.M. Virnik, S. Adhya, and V.B. Zhurkin. (2005). A-Tract Clusters May Facilitate DNA Packaging in Bacterial Nucleoid. *Nucleic Acids Res.*, **33(12)**, 3907-3918.

Chapter 7: Recommendations for Future Work

7.1 Introduction

Through the course of this thesis many challenging problems were solved as part of the tasks set forth in the objectives discussed in Section 1.7. The first specific aim was to construct a model of the twist of DNA base-pair steps as it relates to the overall DNA molecule and the effects that bound proteins have on DNA twist. In collaboration with Dr. Tobias and Dr. Olson, the twist of supercoiling or Tw^{SC}, a new contribution to the field of DNA topology, was developed. The Tw^{SC} was discussed in detail in Chapter 2 where we presented the background, the methodology as published in the *Journal of Chemical Physics* manuscript "Two Perspectives on the Twist of DNA", the 3DNATwSC software developed during this thesis to calculate the Tw^{SC}, and an evaluation of the new Tw^{SC} calculation and the 3DNATwSC software based on seven simplified ideal structures. Chapter 2 established the value of the Tw^{SC} as an effective gage for the topological landscape of DNA due to its sensitivity to changes in chirality. The following two specific aims further established the biological significance of Tw^{SC} by developing two new software tools which could be used to evaluate Tw^{SC} in real world structures.

The second specific aim of this thesis resulted in the creation of TwiDDL, a web accessible interface used to showcase the impact of the Tw^{SC} on DNA/protein complexes

taken from NMR and X-ray crystal structures. In Chapters 3 and 4, TwiDDL was presented as a web interface for accessing and maintaining relevant data on the Tw^{SC}, and as a way to leverage that interface to evaluate biologically relevant data about the Tw^{SC}. It was demonstrated that TwiDDL can effectively be used to search for and find structures that are either under- or over-twisted when compared to the ideal B-DNA step parameter twist, $Tw^{B\text{-}DNA}$, to Tw^{SC}. The visualization capabilities of TwiDDL, such as 2D-plots, color-coded tables, and 3D-plots showing comparisons of Tw^{SP} versus Tw^{SC}, were used to analyze the meaningfulness of Tw^{SC} in a variety of structures in Chapter 4. The structures that were selected included under- and over-twisted DNA, representations of A-, B-, and Z-DNA, as well as a collection of 45 nucleosomes. Further, we demonstrated the effect of shearing on the Tw^{SC}. In section 4.5, we show that the Tw^{SC} changes from a greater $|\Delta Tw| = |Tw^{SC} - Tw^{SP}|$ in the presence of shearing to one that was close to the Tw^{SP} when the shearing was removed from both a nucleosome core particle [1], and the *Anabaena* HU-DNA cocrystal structure (AHU2) [2].

The third and final specific aim, designed a user-friendly graphical user interface for biologists to create minimum-energy DNA/configurations of open linear and spatially constrained supercoiled DNA molecules called 3DNAdesigner. In Chapter 5, the background to the design of 3DNAdesigner was given, followed by a detailed description of the features that were developed, and finally three step-by-step examples were given to

show the capabilities of 3DNAdesigner. In Chapter 6, 3DNAdesigner was used to demonstrate the real world value of our user-friendly application which creates minimum-energy configurations of DNA and protein/DNA complexes. The evaluation of a sequence, provided by the Zechiedrich lab [3], in both a naked and EcoRV bound configuration showed that 3DNAdesigner can effectively be used to generate *in silico* stable minimum-energy configurations which would be of interest to a lab doing *in vitro* studies of the same structure. 3DNAdesigner was also able to find numerous locally minimum-energy configurations of a wild-type 79 base-pair DNA sequence bound to the Lac repressor in four orientations with and with out bound *Anabaena* HU (AHU2) protein, as determined by Swinger et al. [2], with results that were comparable to Swigon et al. [4,5].

While addressing these objectives many trade offs needed to be made and several opportunities presented themselves for future work. In this Chapter we will discuss both the recommendations for how to mitigate some of the trade offs made, as well as identify key areas for future work and enhancements.

7.2 TwiDDL

Maintaining the TwiDDL database to keep it relevant and usable will be an ongoing endeavor. The following subsections discuss possible areas for future work

related to TwiDDL.

7.2.1 New Structures

Since more and more data will be entering the Protein Data Bank (PDB) and other reputable repositories, TwiDDL will need to be updated at regular intervals. The updated data that are stored in TwiDDL will be taken from the PDB and NDB (Nucleic Acids Database). This new data will need to be closely watched and maintained by the owner of TwiDDL (Lauren A. Britton), as discussed in Section 3.5. Additionally new methods to calculate the twist of supercoiling developed for other structures, as discussed in Section 7.3.1, should be incorporated into TwiDDL as an enhancement that would allow the handling of many new types of structures not present in TwiDDL today.

7.2.2 Web Advancements

As web technology changes, the TwiDDL user interface and database will also need to be upgraded to comply with new requirements. This may include minute changes or a complete overhaul depending on the compatibility that future web technology has with the current version of TwiDDL.

7.2.3 Keeping Data Relevant

In addition to TwiDDL upgrades related to new structural data or changes in web

browsers, TwiDDL may also need to be updated when new variables are defined to describe the topology of DNA. These types of changes will need to be included in both the user interface and the MySQL database that TwiDDL runs off of. There are also many current structures that TwiDDL does not handle because the underlying code is unable to process the PDB files. These include structures with nicked backbones or mispaired bases, for example. The code for TwiDDL should be reevaluated to address such structures without requiring the manual intervention that is needed to address these structures today. Improvements and maintenance such as this will allow TwiDDL to always remain relevant.

7.2.4 New Features

There are three features that were never developed for TwiDDL, but would prove very useful in future administration of the database. The first feature would be a user interface designed for the administrator to add a new structure in TwiDDL based on a custom PDB file that they specify, for example one that is modified to address issues discussed in Section 3.3.2. Currently TwiDDL only automates input of the PDB or NDB structures based on the files found directly from the PDB database, but we have found using this method that there are many relevant structures that require manual corrections of the PDB file. I would envision this as an extension of the existing administration interface discussed in Section 3.5. This new feature would be added to the interface in

Figure 3.5.1 and allow the administrator to specify a custom PDB file generate a structure from. This would significantly improve the ease of administration of TwiDDL in the cases where automated data generation fails due to limitations in the code's ability to deal with unexpected structural anomalies or poorly formated PDB files.

The second feature would allow the administrator to review the data of all structures with anomalies that were flagged by automation as discussed in Section 3.3.2. Currently the user interface only allows the administrator to view the data that are not flagged, just like any other user who visits TwiDDL. The only way an administrator can view the flagged data today is through the command line interface of the MySQL database, which is not ideal and can be cumbersome. This feature would ideally be based on the existing TwiDDL user interface, but allow the administrator the special privilege to see the flagged data.

The third feature is to create a user interface to allow the administrator to manually edit the TwiDDL database entries. Currently maintaining and updating the data within TwiDDL requires an understanding of MySQL or Perl. In some cases there have been simple edits (like correcting a spelling error or enter a missing NDB ID, for example) and this feature would enable these edits to be done much more quickly than through other means. This feature would ideally be based on the existing TwiDDL user interface, but would provide the administrator the ability to do something like select a

TwID and click a button to enter an edit mode. Once in the edit mode, the administrator could simply type in any changes through a web browser, save these changes, and that would in turn update both the database and the files available for download about the TwID.

7.3 3DNAdesigner

3DNAdesigner should have yearly update releases to remain relevant. These releases should fix any bugs that are found and include relevant upgrades to the software. The following subsections detail some possible enhancements that can be incorporated in future releases when the code and/or theory is completed.

7.3.1 Twist of Supercoiling

Double-helical Watson-Crick DNA is not the only type of nucleic acid structure that exists in vitro or in X-ray crystal structures. It would be very useful to be able to define the Tw^{SC} of structures such as Holliday junctions, melted DNA, and even single-stranded RNA and DNA. There is definite interest in defining melted DNA and denaturation from Dr. Sarah Harris at the University of Leeds. As new algorithms are developed to model the Tw^{SC} for these structures they should be added to 3DNAdesigner, as well as TwiDDL. These new algorithms can easily be incorporated into the 3DNAdesigner GUI as another wizard that would allow the user to specify what type of

structure they wished to include in their configuration, and the appropriate algorithm, or even multiple algorithms, could be used to model it.

7.3.2 Electrostatic Interactions

The structure of DNA, a polyanion, is sensitive to the ionic environment. It is very important to understand the contributions of electrostatic forces to global shape [6, 7]. Electrostatic forces are currently not used in 3DNAdesigner in the search for a stable configuration and therefore the calculations do not account for DNA self-penetration [8]. The electrostatics are calculated, however, after a mechanically stable configuration has been reached.

There are several other ways to account for the electrostatics during the optimization process. One way, developed by Biton and Coleman [9], computes the electrostatic energy while finding a minimum energy configuration, but the method is more time intensive than computing the electrostatic energy after an elastic minimum has been found. In the case of short DNA circles, electrostatic minimization is not expected to have a significant effect on the shape of DNA if the only charges are those on the phosphate groups. The electrostatics associated with base sequence can potentially perturb the global shape. The partial charges are not uniformly distributed in the bases. The approach of Petrella and Karplus can find a minimum electrostatic energy equilibrium relatively quickly [10]. At a future date one or both of these methods may be

worth investigating to correct for DNA self-penetration and to check the assumption that small circles are not affected by electrostatics.

It would also be of interest to simulate the work by Maher and coworkers [11,12], who chemically neutralized selected phosphates on the DNA backbone in a linear oligomer and saw a large effect on overall shape when the oligomers were incorporated as multimers in a long DNA segment. The changes in shape were detected using gel electrophoresis. Simulating the likely shapes of the neutralized DNA *in silico* would be a great extension of 3DNAdesigner. It would be able to reproduce, using simulation, the apparent changes in DNA curvature obtained when one face of the double helix on a linear oligomer is neutralized, per the experimental observation, as well as to apply a method that accounts for electrostatic interaction to a plasmid of the same sequence. Kosikov et al. [6,13] performed atomic-level simulations of very short DNA fragments that accounted for the curvature of the neutralized DNAs structure defined by the Maher group.

Based on this background, it is recommended that electrostatic interactions within the structures modeled by 3DNAdesigner be explicitly addressed. While electrostatics plays a key role in DNA topology, our current calculations have electrostatics implicitly accounted for in the rest states and force constants that govern the sequence-dependent deformation of DNA step parameters. Ideally these electrostatic interactions can be

directly calculated and utilized in the models generated by 3DNAdesigner.

7.4 References

[1] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 a Resolution. *J. Mol. Biol.,* **319(5)**, 1097-1113.

[2] K.S. Swinger, K.M. Lemberg, Y. Zhang, and P.A. Rice. (2003). Flexible DNA Bending in HU-DNA Cocrystal Structures. *EMBO J*., **22(14),** 3749-3760.

[3] J.M. Fogg, N. Kolmakova, I. Rees, S. Magonov, H. Hansma, J.J. Perona, and E.L. Zechiedrich. (2006). Exploring Writhe in Supercoiled Minicircle DNA. *J. Phys. Condens. Matter*., **18(14),** S145-S159.

[4] D. Swigon, and W.K. Olson. (2008). Mesoscale Modeling of Multi-Protein–DNA Assemblies: The Role of the Catabolic Activator Protein in Lac-Repressor-Mediated Looping. *Int. J. Non-linear Mech.*, **43(10)**, 1082-1093.

[5] D. Swigon, B.D. Coleman, and W.K. Olson. (2006). Modeling the Lac Repressor-Operator Assembly: The Influence of DNA Looping on Lac Repressor Conformation. *Proc. Natl. Acad. Sci. USA*, **103(26)**, 9879–9884.

[6] K.M. Kosikov, A.A. Gorin, X. Lu, W.K. Olson, and G.S. Manning. (2002). Bending of DNA by Asymmetric Charge Neutralization: All-Atom Energy Simulations. *J. Am. Chem. Soc.*, **124(17)**, 4838-4847.

[7] G.S. Manning. (2003). Comments on Selected Aspects of Nucleic Acid Electrostatics. *Biopolymers*, **69(1)**, 137-143.

[8] T.P. Westcott, I. Tobias, and W.K. Olson. (1997). Modeling Self-Contact Forces in the Elastic Theory of DNA Supercoiling. *J. Chem. Phys.*, **107(10)**, 3967-3980.

[9] D. Jeulin and S. Forest. (2008). *Continuum Models and Discrete Systems CMDS 11: Proceedings of the International Symposium held in Paris, July 30th-*

August 3rd 2007. Les Presses de l'Ecole des mines de Paris, Paris, France.

[10] R.J. Petrella, and M. Karplus. (2005). Electrostatic Energies and Forces Computed Without Explicit Interparticle Interactions: A Linear Time Complexity Formulation. *J. Comput. Chem.,* **26(8)**, 755-787.

[11] E.D. Ross, P.R. Hardwidge, and L.J. Maher III. (2001). HMG Proteins and DNA Flexibility in Transcription Activation. *Mol. Cell Biol.*, **21(19)**, 6598-6605.

[12] L.D. Williams, and L.J. Maher III, (2000). Electrostatic Mechanisms of DNA Deformation. *Annu. Rev. Biophys. Struct.*, **29**, 497-521.

[13] K.M. Kosikov, A.A. Gorin, V.B. Zhurkin, and W.K. Olson. (1999). DNA Stretching and Compression: Large-Scale Simulations of Double Helical Structures. *J. Mol. Biol.,* **289(5)**, 1301-1326.