RNA STRUCTURE ANALYSIS VIA THE RIGID BLOCK MODEL

by

MAURICIO ESGUERRA NEIRA

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 Chemistry and Chemical Biology

Written under the direction of

Wilma K. Olson

and approved by

New Brunswick, New Jersey

May, 2010

ABSTRACT OF THE DISSERTATION

RNA Structure Analysis via the Rigid Block Model

by Mauricio Esguerra Neira

Dissertation Director: Wilma K. Olson

RNA structure is at the forefront of our understanding of the origin of life, and the mechanisms of life regulation and control. RNA plays a primordial role in some viruses. Our knowledge of the importance of RNA in cellular regulation is relatively new, and this knowledge, along with the detailed structural elucidation of the transcription machine, the ribosome, has propelled interest in understanding RNA to a level which starts to closely resemble that given to proteins and DNA.

In the process of progressively understanding the landscape of functionality of such a complex polymer as RNA, one practical task left to the structural chemist is to understand the details of how structure relates to large-scale polymer processes. With this in mind the fundamental problems which fuel the work described in this thesis are those of the conformations which RNA's assume in nature, and the aim to understand how RNA folds.

The RNA folding problem can be understood as a mechanical problem. Therefore efforts to determine its solution are not foreign to the use of statistical mechanical methods combined with detailed knowledge of atomic level structure. Such methodology is mainly used in this work in a long-term effort to understand the intrinsic structural features of RNA, and how they might relate to its folding.

As a thing among things, each thing is equally insignificant; as a world each one equally significant.

If I have been contemplating the stove, and then am told; but now all you know is the stove, my result does indeed sound trivial. For this represents the matter as if I had studied the stove as one among the many, many things in the world. But if I was contemplating the stove, it was my world, and everything else colorless by contrast with it ...

For it is equally possible to take the bare present image as the worthless momentary picture in the whole temporal world, and as the true world among shadows.

Ludwig Wittgenstein

As a molecule among molecules, each molecule is equally insignificant; as a world each one equally significant.

If I have been contemplating RNA, and then am told; but now all you know is RNA, my result does indeed sound trivial. For this represents the matter as if I had studied RNA as one among the many, many molecules in the world. But if I was contemplating RNA, it was my world, and everything else colorless by contrast with it ...

For it is equally possible to take the bare present image as the worthless momentary picture in the whole temporal world, and as the true world among shadows.

Anonymous Chemist

Acknowledgements

I would first like to give a special thanks to Dr. Yurong Xin, whose patience, help, and collaboration since the very beginning of my joining of the Olson lab have been fundamental for the development of this work. I would like to thank Dr. Olson's extreme patience, and room for freedom on carrying out this research. Finally I thank all colleagues at the Olson lab.

I would like to dedicate this thesis to David and Stella Case, without them these words would not exist.

Table of Contents

Abstract ii

Acknowledgements iv

List of Tables viii

List of Figures ix

1. Introduction 1

1.1. RNA chemistry 1

1.2. RNA folding 3

1.3. Is RNA folding a hard or easy problem? 5

1.4. Experimental folding techniques 7

1.5. RNA simulations 7

1.5.1. Local nucleotide interactions 8

1.5.2. RNA secondary structure algorithms and the lack of tertiary ones 9

1.5.3. RNA overall fold 9

1.5.4. RNA motifs 11

1.6. Overview 12

References 14

2. RNA Base Steps 21

2.1. Consensus Clustering of Single Stranded Base Step Parameters 24

2.1.1. Combining Fourier Averaging Results and Clustering Analysis 24

2.1.2. Selection of a Clustering Methodology 28

References 40

3. RNA Base-Pairing 42

3.1. Canonical and Noncanonical Base-pairs 42

3.2. Clustering of Yurong’s Classification . . . 42
References . . . 44
4. RNA Base Pair Steps . . . 45
4.1. Analysis (Albany Poster) and Django Webserver . . . 45
4.2. Persistence Length of RNA . . . 45
4.3. AMBER: Persistence Length of Base-Pair Step Patterns . . . 46
References . . . 47
5. RNA Motifs . . . 47
5.1. GNRA tetraloop . . . 47
5.1.1. 3DNA-Parser . . . 47
5.1.2. Overlap Scores . . . 48
5.2. Triplets on RNA (comparison to Laing et al.) . . . 48
References . . . 51
6. RNA Helical Regions and Graph Theory . . . 52
Appendix A. Standard reference frame and local parameters . . . 53
A.1. Base-pair and base-step parameters . . . 53
A.2. Local helical parameters . . . 56
References . . . 59
Appendix B. Clustering Analysis (CA) . . . 60
B.1. General Methodology . . . 60
B.2. Hierarchical methods . . . 61
References . . . 65
Appendix C. Dimension Reduction . . . 66
C.1. Principal Component Analysis . . . 66
References . . . 68
Appendix D. Persistence Length . . . 69
D.1. Persistence Lenght Definitions . . . 69

D.2. end-to-end 72

D.3. Models 73

D.3.1. Kuhn - Freely Jointed Chain (FJC) 73

D.3.2. Porod-Kratky - Worm Like Chain (WLC) 73

D.3.3. Olson - Realistic 73

D.4. Suggested Reads 73

References 74

Supplement A. Figure Supplements 75

Curriculum Vitae 72

List of Tables

2.1. Some large RNA structures (>300 bases) elucidated in the last decade. 23

2.2. Number of base-steps with RMSD values less than or equal to 10 Å between the reference base-step vectors from the four groups of non-A-type RNA dinucleotide conformations and all base-step vectors found in the 23S strand of *Haloarcula marismortui*. The percentage is calculated with respect to a total of 2753 base-steps present in the 23S chain of the 50S subunit of the ribosome. 28

2.3. Base step parameters for common DNA and RNA conformations. The base-step parameters are computed for a single-stranded base-step rather than a double-stranded base-pair step. 33

3.1. Classification of RNA Types in Non-Redundant Dataset at less than 3.5 Å (For Base-Pairs in Helices of 3 base-pairs or more). 43

B.1. Example of structures, considered as bidimensional vectors, to be clustered using the average linkage method and the Manhattan distance. 63

D.1. Persistence lengths for some biopolymers with filament structures. 71

List of Figures

1.1. A single strand of RNA drawn in the 5′ to 3′ sense showing the three chemical entities which compose it, base, sugar, and phosphate. The four bases (A, G, C, U) are colored according to the NDB (Nucleic Acid Database) convention [18], the phosphate is colored gray, and the sugars black. The bases G, and C, and the furanose sugar attached to the G are numbered according to the IUPAC rules [19]. This figure is an adaptation of Figure 2.1, in Wolfram Saenger's book, "Principles of Nucleic Acid Structure" [20]. 2

1.2. Saenger base-pairing classes, reproduced from his book, "Principles of Nucleic Acid Structure". [20]. 4

1.3. **Left:** Sugar, and sugar-phosphate backbone torsion angles. **Right:** The most common sugar pucker conformations in RNA, that is, $C_{3'-\mathrm{endo}}$ and $C_{2'-\mathrm{endo}}$, reproduced from Wolfram Saenger's, "Principles of Nucleic Acid Structure". [20]. 5

1.4. Separation of secondary and tertiary interaction in RNA [39]. Double helical secondary structure represented by individual cylinders and tertiary interactions by association of cylinders. Color coding stands for separate helical regions of RNA, and the connecting black strings represent single stranded loop structures. 6

1.5. Ribbon-coil schematic illustraring the fold and intermolecular units of a dimer of prealbumin (PDB_ID:2pab), or transthyretin, taken from Richardson *et al.* [90] 10

1.6. Images of the *Haloharcula marismortui's* large ribosomal subunit NDB_ID:RR0033 (left) and the hammerhead ribozyme (right) NDB_ID:UR0029. The figures were taken directly from the NDB web pages, and show a 3DNA generated [91] ribbon representation of the phosphate backbone, and a block representation for the nucleotide bases. From the figures it's clear that, whereas the ribozyme fold can be clearly understood with this representation, the ribosome fold cannot. 11

2.1. **Left:** Total number of RNA bases added to the PDB database between 2000 and 2010 (Exponential fit line in blue). **Right:** Total number of RNA structures solved yearly by X-Ray crystallography between 2000 and 2010 (Exponential fit line in red). 21

2.2. Frequency of nucleotide bases in RNA molecules found in the PDB classified by the size of RNA molecules. We define the size as the total number of nucleotide bases present per molecule. 23

2.3. Figure taken from Richardson et al. [11] where the blue and green dots in a) mean very accurate van der Waals distances, and in b) the red and orange dots mean steric clashes, that is, distances outside the acceptable van der Waals range. 24

2.4. Dendrogram showing the results of consensus clustering of 20 non-Atype rRNA dinucleotides according to their hexadimensional base-step parameter vectors. 26

2.5. RNA dinucleotide structures organized by clusters obtained from consensus clustering of their hexadimensional base-step parameter vectors. The structures have been centered on the reference frame of the first step, that is, the adenine base, and the minor groove face of the rigid block parameter associated to adenine is facing the viewer. 27

2.6. Root mean square deviation of the main four groups show in Figure 2.5. The color of the histograms is the same as that of the boxes surrounding the structures of Figure 2.5 . . . 29

2.7. Root mean square deviation histograms for the subgroups present in group IV. Since subgroup IVb is composed of A-RNA like conformations we see in the upper left histogram that the highest proportion of small RMSD values belongs to this group. 30

2.8. Rigid block representation of dinucleotide steps. The major groove side of the first nucleotide block is oriented towards the viewer and shaded gray. **Left:** Drawn in blue, the block representing the Group I cluster from Figure 2.5. Superimposed to the Group I cluster are three structures whose step-parameter RMSD's with respect to the Group I cluster are less than or equal to 10 Å. **Right**: With an RMSD less than or equal to 15 Å we "identify" a total of seven structures from the ribosome. We clearly see that three of them (encircled in cyan blobs) are farther apart from the original Group I main structure of Figure 2.5 which is drawn in blue. 31

2.9. Pairs scatterplot for base-step parameters, shift, slide, rise, tilt, roll, and twist, for the non-ARNA dataset colored according to purine-pyrimidine (black), purine-purine (red), pyrimidine-pyrimidine (green), and pyrimidine-purine (blue) steps. 32

2.10. Cluster validity scores for internal measures. Notice how the hierarchical method, labeled as 1 in black color, behaves better for the whole range of Connectivity (smaller values) and Dunn (higher values), and it also outperforms all others after $k = 12$ for Silhoutte (higher values) scores. 34

2.11. Cluster validity scores for stability measures. 35

2.12. RMSD values between base-step parameters of the 23S subunit of ribosomal RNA and the standard base-step parameters derived from Arnott and collaborators [24] work. . . . 36

2.13. Cluster validity scores for the non-ARNA dataset. It can be seen clearly that the optimal method for clustering is the hierarchical one, as measured by lower values in the connectivity scores, and higher values in the Dunn score. The optimal number of clusters given by the dunn score is 67, we also see shoulders at $k = 67$, for the connectivity and silhouette scores. 38

2.14. 17 out of the 67 groups clustered using the hierarchical clustering algorithm are drawn in a photograph contact sheet fashion. Each group is centered on the base reference frame of the adenine block drawn in red. In the lower right corner of the "contact sheet" the full space of 797 reconstructed steps is shown, along with the 20 steps derived from schneider et al. work. Notice how the only "hollow" side of the "onion" formed by the full space of base-step conformations is that corresponding to the watson-crick base-pairing region. 39

5.1. GNRA Tetraloop from *Thermus Thermophilus* 23S Ribosomal RNA PDB-ID:1ffk. 48

5.2. Normalized histograms showing the distribution of overlap values in the 23S subunit or *Thermus Thermophilus* rRNA, PDB-ID:1jjk. In histogram (a) all values are included, but in histogram (b) only values greater than zero are included. Notice the high preponderance of zero values, exactly 897 out of a total of 2705. 49

5.3. Dendrogram for consensus clustering of overlap scores in the ribosome. Zero values filtered out and remaining data normalized. 50

A.1. Standard reference frame of an A-T base-pair [4]. The *y*-axis (dashed green line) is chosen to be parallel to the line connecting the $C1^{'}$ of adenine and the $C1^{'}$ of thymine associated in an ideal Watson-Crick base-pair. The *x*-axis is the perpendicular bisector of the $C1^{'}$ - $C1^{'}$ line, and the origin is located at the intersection of the *x*-axis and the line connecting the C8 atom of adenine and the C6 atom of thymine. The *z*-axis is the cross product of the $\hat{x}$ and $\hat{y}$ unit vectors. 54

A.2. Illustration of base pair and base step parameters [1] 57

B.1. Clustering tree for 5 bidimensional vectors using the Manhattan distance definition and the average linkage clustering method. 64

S1. Non A-RNA Type base steps centered on the standard reference frame of Adenine. Top view with the Minor Groove side of Adenine pointing down the page and the Major Groove pointing up. 76

S2. The total number of structures available in the pdb up to the end of year 2009. The scale of the axis in the left (in black), is ten times that in the right (in green). The black y-axis sets the scale for the number of protein structures available in the PDB up to the end of the year 2009. The green y-axis sets the scale for the number of molecular structures containing, rna only (in red), dna only (in blue), and protein plus nucleic acid (in green). One can clearly see that the total number of protein, rna, and protein plus nucleic acid structures is growing exponentially. It is also clear that the number of DNA structures is perhaps tending toward a constant number, that is, it might not be growing. It is also interesting to see how the number of RNA structures really lifts off in the middle of the nineties, whereas for DNA the growth started earlier and is settling down. 77

Chapter 4

RNA Base Pair Steps

4.1 Analysis (Albany Poster) and Django Webserver

Results shown in Albany and steps part of methods paper.

This gives us the force constant matrices per base-step which are used in the next section.

4.2 Persistence Length of RNA

A quantity commonly used to quantify the stiffness of polymers is the so-called persistence length a. To determine this quantity for DNA or RNA a variety of theoretical and experimental techniques are used. Some common experimental techniques to determine a are Electron Microscopy (EM), gel electrophoresis, sedimentation velocities, electrical birefringence Atomic Force Microscopy (AFM) , Magnetic Tweezers, and Small Angle X-Ray Scattering (SAXS). For reviews of such techniques applied to the determination of RNA persistence length, we refer the reader to Hagerman [1], Abels et al. [2], and Caliskan et al. [3]. We will use their results for comparison with those coming from the "realistic" model developed by Olson and collaborators [4] to describe DNA. The "realistic" model is dependent on high resolution crystallographic data. Initial studies started with small numbers of data for the deformabilities of the ten unique base-pair steps [4]. A more complete picture applied to the study of DNA sequence dependent deformability became available in 1998 [5]. The base-pair step deformability data for DNA has been constantly refined as more high resolution DNA and DNA-protein structures have been added to the Nucleic Acid Database (NDB) [6]. Although such data has been available for DNA since 1998, it had not been so for RNA, until now [7].

A detailed description of the "realistic model" along with the scheme of the C++ code developed by Czapla and Zheng to implement it, and a brief account of various definitions of persistence length and models from which a can be derived are included in Appendix D

4.3 AMBER: Persistence Length of Base-Pair Step Patterns

I guess it needs some input here in order to work on latex compilation.

References

[1] Hagerman, P. J. (1997) Flexibility of RNA. *Annual Review Biophysics Biomolecular Structure,* **26**, 139–156.

[2] Abels, J. A., Moreno-Herrero, F., van derHeijden, T., Dekker, C., and Dekker, N. H. (2005) Single-Molecule Measurements of the Persistence Length of Double-Stranded RNA. *Biophysical Journal,* **88**, 2737–2744.

[3] Caliskan, G., Hyeon, C., Perez-Salas, U., Briber, R. M., Woodson, S. A., and Thirumalai, D. (2005) Persistence Length Changes Dramatically as RNA Folds. *Phys Rev Lett,* **95**, 268303.

[4] Olson, W. K., Babcock, M. S., Gorin, A., Liu, G., Marky, N. L., Martino, J. A., Pedersen, S. C., Srinivasan, A. R., Tobias, I., and Westcott, T. P. (1995) Flexing and Folding Double Helical DNA. *Biophysical Chemistry,* **55**, 7–29.

[5] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proceedings of the National Academy of Sciences,* **95**, 11163–11168.

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

[7] Olson, W. K., Esguerra, M., Xin, Y., and Lu, X.-J. (2009) New Information Content in RNA Base Pairing Deduced from Quantitative Analysis of High-Resolution Structures. *Methods,* **47**, 177–186.

Appendix D
Persistence Length

Nucleic Acids and other polymers can be understood as mechanical objects [1, 2] and therefore engineering approaches can be used for their understanding. Usually the methods followed by the engineering approach consider the polymer as a long continuous rod, and are known as continuum elastic theory. This type of approach leaves little space for taking into account the nature of the subunits which make up the polymer, that is, it is mainly applicable to homopolymers made up of identical subunits. In nucleic acids this is not necesarily the case, a more general approach should take into account the possibility of having different subunits making up the polymer. Olson and collaborators have developed a sequence dependent model, refered to as the "realistic" model [3]. Such model is harmonic and depends on determination of force-constant analogs derived from X-Ray cristallographic data taken from the Nucleic Acid Database (NDB) [4, 5]. Within the context of the "realistic" model Czapla et al. [6] have suggested a gaussian sampling methodology which allows the determination of global polymer properties like the persistence length a following a matrix approach suggested by Flory [7] and expanded upon by Olson et al. [8, 9]

In what follows we summarize various definitions of persistence length and how it's computed using different models.

D.1 Persistence Lenght Definitions

There are two parallel perspectives which can be used to define the persistence length of nucleic acids. One is a more mathematical, or physical one, where it is understood as the resistence to deformation of a curve in space, or thin rod, a physical object. The other one is a stochastic one where it is understood as "a measure of the distance over which the direction of the DNA is maintained [10]" and has the classical formulation by Flory which states that the persistence length is:

"the average sum of the projections of all bonds $j \geq i$ on bond i in an indefinetely long chain. The bond i is taken to be remote from either end of the chain, i.e., $1 \ll i \ll n$". Paul J. Flory, Statistical

Mechanics of Chain Molecules. 1969

This perspective has a more chemically flavored tone, since it assumes some type of bonded connectivity between polymeric units, whether the bond is a "real" one, or a "virtual" one.

Mathematically, a is the mean projection in the limit of infinite chain length of a flexible DNA along its initial direction." [10]

“**Bend-persistence length**:
A length scale beyond which the elastic cost of bending is totally negligible” Philip Nelson, Yearbook of Science and Technology, McGraw Hill 1999

“In a randomly shaken rod any particular point in the rod will be pointing in a random direction, but nearby points will be pointing in roughly the same direction, that is, these nearby points are persistent. Points farther away than the bend-persistence length are said to be uncorrelated.” Philip Nelson, Yearbook of Science and Technology, McGraw Hill 1999

“**twist-persistence length**:
that is, a rubber rod not only resists bending but also twisting”

"basic mechanical property that quantifies stiffness" Abels et al, Biophysical Journal, 2005, 2737-2744

"Classical elasticity tells us that a thin, straight rod that is bent into an arc has a bending energy $E = Bl/2R^2$, where B is the bending elastic constant of the rod, l is the length of the rod and R is the radius of arc. Setting $R = l$ gives us the energy of a 1 radian bend along the rod, and solving for when $E\kappa_B T$ gives us the length of rod along which a thermally excited bend of 1 radian typically occurs: $lB/\kappa_B T$. This is called the persistence length..." John F. Marko and Simona Cocco, Physics World, March 2003

"The persistence length a is a measure of the stiffness of a polymer chain and is related to the limiting value of the characteristic ratio at infinite chain length

$$a = \frac{\nu}{2}(C_\infty + 1) \tag{D.1}$$

"

"length at which the orientation of the sequential bonds which make up a polymer chain, stop being correlated. That is, if you have just two bonds, or a few, they will be correlated, which is the case in most molecules, but, in polymers, you have a long chain of sequential bonds. At some length, bonds will become uncorrelated, but up to that length they were correlated, this is what is meant by persistence length, and, in this context it's obvious that is an exclusive property of polymers." My understanding so far.

Biopolymers can be either rigid or flexible.

They can be classified according to whether their persistence length (a) is greater, smaller, or similar to the contour length (L) of the polymer.

Model Type	Polymer Characteristic	a to L relation
Rigid Rod	Rigid	$a \gg L$
Gaussian chain	Flexible	$a \ll L$
Worm-like chain	Semi-flexible	$a \approx L$

Notice that for $a \gg L$, there is a definition problem, since L has to be large enough to be a good approximation to the definition of persistence length, which is defined for an infinite chain length.

Worm-like-chain = Porod-Kratky = Freely Rotating Chain in limit l=0 and n=infinity

Rigid biopolymers: actin, microtubules

Flexible biopolymers:

Semi-flexible biopolymers: High force extension DNA.

If the persistence length is of the same order of the length of the polymer, then the polymer is classified as semi-flexible

Polymer	a (nm)
α-helix	80-100
coiled-coil	150-300
Ideal DNA	51
Ideal RNA	70-80

Table D.1: Persistence lengths for some biopolymers with filament structures.

Think about the "energy" based perspective of Nicolas, and the stochastic based perspective of Flory and others.

D.2 end-to-end

The end-to-end vector r is the vector which connects the ends of a polymer chain. It can be defined as the sum of the vectors connecting the monomer units in a chain. These connecting vectors are sometimes called virtual bond vectors l.

From the end-to-end vector the quantity which is usually of interest is it's magnitude.

$$r = \sum_{i=1}^{n} l_i \tag{D.2}$$

$$r^2 = r \cdot r = \sum_{i,j} l_i \cdot l_j \tag{D.3}$$

Equation D.2, can also be written:

$$r^2 = \sum_i l_i^2 + 2 \sum_{i \neq j} l_i \cdot l_j \tag{D.4}$$

To describe a polymer it's necessary to think about the various conformations it can adopt due to its flexibility, therefore, it is important to think of polymer related quantities in terms of the average of their possible conformations. For the end-to-end vector the average of its values is denoted as $< r >$, and the average of its norm, also called the second moment of the end-to-end distribution, is denoted by $< r^2 >$:

$$< r^2 >= \sum_i < l_i^2 > +2 \sum_{i<j} < l_i \cdot l_j > \tag{D.5}$$

When there is no correlation between succesive bonds we can write:

$$< l_i \cdot l_j >= 0 \tag{D.6}$$

So that equation D.5 keeps only the bond auto-correlation term:

$$< r^2 >= \sum_i < l_i^2 >= n < l^2 > \tag{D.7}$$

This equation is used to describe a so-called freely-jointed chain.

D.3 Models

Nelson in book says:

$$dE = \frac{1}{2}K_BT[A\beta^2 + Bu^2 + C\omega^2 + 2Du\omega]ds \tag{D.8}$$

A$\kappa\beta$ T = Bend stiffness B$\kappa\beta$ T = Stretch stiffness C$\kappa\beta$ T = Twist stiffness D$\kappa\beta$ T = Twist-stretch coupling

If only the bend stiffness survives then the model is called an inextensible model, also Porod-Kratky, or WLC.

D.3.1 Kuhn - Freely Jointed Chain (FJC)

D.3.2 Porod-Kratky - Worm Like Chain (WLC)

D.3.3 Olson - Realistic

The Hamiltonian for a [11]

D.4 Suggested Reads

From Equilibrium Statistics of Plischke and Bergersen they suggest to read: Des Cloiseaoux and Janik () Rubinstein and Colby (Polymer Physics)

References

[1] Marko, J. F. and Cocco, S. (2003) The Micromechanics of DNA. *Physics World,* **16**, 37–41.

[2] Nelson, P. (2004) Biological Physics: Energy, Information, Life, W. H. Freeman and Company, .

[3] Olson, W. K., Marky, N. L., Jernigan, R. L., and Zhurkin, V. B. (1993) Influence of Fluctuations on DNA Curvature. A Comparison of Flexible and Static Wedge Models of Intrinsically Bent DNA. *Journal of Molecular Biology,* **232**, 530–554.

[4] Go, M. and Go, N. (1976) Fluctuations of an Alpha-Helix. *Biopolymers,* **15**, 1119–1127.

[5] Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M., and Zhurkin, V. B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proceedings of the National Academy of Sciences,* **95**, 11163–11168.

[6] Czapla, L., Swigon, D., and Olson, W. K. (2006) Sequence-dependent effects in the cyclization of short dna. *Journal of Chemical Theory and Computation,* **2**, 685–695.

[7] Flory, P. J. (1969) Statistical Mechanics of Chain Molecules, Interscience Publishers, .

[8] Maroun, R. C. and Olson, W. K. (1988) Base Sequence Effects in Double-Helical DNA. II. Configurational Statistics of Rodlike Chains. *Biopolymers,* **27**, 561–584.

[9] Marky, N. L. and Olson, W. K. (1994) Configurational Statistics of the DNA Duplex: Extended Generator Matrices to Treat the Rotations and Translations of Adjacent Residues. *Biopolymers,* **34**, 109–120.

[10] Olson, W. K., Babcock, M. S., Gorin, A., Liu, G., Marky, N. L., Martino, J. A., Pedersen, S. C., Srinivasan, A. R., Tobias, I., and Westcott, T. P. (1995) Flexing and Folding Double Helical DNA. *Biophysical Chemistry,* **55**, 7–29.

[11] Czapla, L. The Statistical Mechanics of Free and Protein-Bound DNA by Monte Carlo Simulation PhD thesis Rutgers, The State University of New Jersey (2009).