

## ABSTRACT

SETHAPHONG, LATSAVONGSAKDA. Molecular Recognition and Structural Influences on Function in Bio-nanosystems of Nucleic Acids and Proteins. (Under the direction of Yaroslava G. Yingling.)

This work examines smart material properties of rational self-assembly and molecular recognition found in nano-biosystems. Exploiting the sequence and structural information encoded within nucleic acids and proteins will permit programmed synthesis of nanomaterials and help create molecular machines that may carry out new roles involving chemical catalysis and bioenergy.

Responsive to different ionic environments thru self-reorganization, nucleic acids (NA) are nature's signature smart material; organisms such as viruses and bacteria use features of NAs to react to their environment and orchestrate their lifecycle. Furthermore, nucleic acid systems (both RNA and DNA) are currently exploited as scaffolds; recent applications have been showcased to build bioelectronics and biotemplated nanostructures via directed assembly of multidimensional nanoelectronic devices<sup>1</sup>. Since the most stable and rudimentary structure of nucleic acids is the helical duplex, these were modeled in order to examine the influence of the microenvironment, sequence, and cation-dependent perturbations of their canonical forms. Due to their negatively charged phosphate backbone, NA's rely on counterions to overcome the inherent repulsive forces that arise from the assembly of two complementary strands. As a realistic model system, we chose the HIV-TAR helix (PDB ID: 397D) to study specific sequence motifs on cation sequestration. At physiologically relevant concentrations of sodium and potassium ions, we observed sequence based effects where purine stretches were adept in retaining high residency cations. The

transitional space between adenine and guanosine nucleotides (ApG step) in a sequence proved the most favorable. This work was the first to directly show these subtle interactions of sequence based cationic sequestration and may be useful for controlling metallization of nucleic acids in conductive nanowires. Extending the study further, we explored the degree to which the structure of NA duplexes alone interacted with cations distinct from a specific sequence. Under physiologically relevant conditions, a duplex of RNA polyguanine-polycytidine was highly responsive and able to sequester cations to the middle of the purine stretches. The least responsive structure was a DNA polyadenine-polythymine duplex. A random sequence DNA duplex contorted into an RNA-like helix resulted in cationic dynamics similar to RNA systems. These studies showed that cation diffusive binding events in nucleic acid duplex structures are sequence specific and heavily influenced by structural aspects helical forms to account for much of the differences observed.

Although structural information in nucleic acids is encoded within their sequence, linking amino acid sequence to protein structure is murkier; the structural information within proteins is encoded by the folding process itself: a complex phenomenon driven toward the equilibrium state of the active conformation. Upwards of two thirds of a protein's sequence can be substituted with similar amino acids without significantly perturbing its function; conserved residues of about 10% seem to be vital; since evolutionary selection pressure in proteins operates 3-dimensionally, a linear sequence is partially informative. We explored this problem by folding *de-novo* the cytosolic portion of the membrane protein, cellulose synthase, CESA1 from upland cotton, *Gossypium hirsutum* (Ghcesa1). The cytoplasmic region was generated by homology modeling and refined with molecular dynamics. These

mutations impair local structural flexibility which likely results in cellulose that is produced at a lower rate and is less crystalline. Additional modeling of fragments of cellulose synthases from the model plant, *Arabidopsis thaliana*, offered novel insights into the function of conserved cytosolic domains within plant cellulose synthases. Transport mechanisms related to the transmembrane region revealed significant differences between plants and a bacterial complex. These studies generated possible mutations that may allow for the creation of new synthases and identified other avenues of research in order to develop technologies that may alter the crystallinity and other useful properties of cellulose.

© Copyright 2013 by Latsavongsakda Sethaphong

All Rights Reserved

Molecular Recognition and Structural Influences on Function in Bio-nanosystems of Nucleic  
Acids and Proteins

by  
Latsavongsakda Sethaphong

A dissertation submitted to the Graduate Faculty of  
North Carolina State University  
in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

Materials Science and Engineering

Raleigh, North Carolina

2013

APPROVED BY:

---

Dr. Yaroslava G. Yingling  
Committee Chair

---

Dr. Donald Brenner

---

Dr. Thomas LaBean

---

Dr. Steffen Heber

## **DEDICATION**

*To my family and friends.*

*In memoriam, CAPT(SEL) Kurt William Juengling, United States Navy.*

## BIOGRAPHY

A refugee from Laos, Latsavongsakda Sethaphong grew up in Nashville, TN. He was somewhat of a polymath in high school and thought that he might someday become a classicist; however, he was never too keen to focus merely on one subject. Fate intervened when he read Richard Feynman's unique biographical sketches "Surely You're Joking, Mr. Feynman!": Adventures of a Curious Character and "What Do You Care What Other People Think?": Further Adventures of a Curious Character. However, his choice of college emerged randomly after reading the novels of Robert Heinlein. Ultimately, he "sort of" studied physics and a few other things at Harvard College. He found work in California as an engineer, but returned home to Nashville and immersed himself in biophysics at Vanderbilt. He broke off that pursuit in mid-course, joined the Navy and returned to engineering for a brief spell. Answering an advertisement led him to NCSU's Materials Science and Engineering graduate program. His time was interrupted by a year-long tour in Afghanistan where he met new people and had interesting experiences. In his wanderings, he has managed to preserve an appreciation for uncertainty and remains coy on any future plans: *Parcis profugus, non viris.*

## ACKNOWLEDGMENTS

Firstly, I thank my advisor, Dr. Yaroslava G. Yingling, who has guided me and others thru this arduous process; she amply provided the material and motivational requisites for this research. I thank Bill Becklean, Ruth Harris, Marge Lynch, Dr. Kathleen Hawke, and especially Dr. Sam Wells who have served as my mentors in the journey of learning. I thank “The Smackers” -- John, Uttam, Dave, Gaurav, and Sean -- who kept me mostly focused while punctuating the years with random revelry since college; their intellectual brilliance and humanity inspired me to do more. I thank my friend Charles Starks who taught me the virtues of learning even when life gets hard. I thank my friend Brendan who kept me in line with reality. I thank my best friend, Aaron, who always took me in from the cold and gave encouragement whenever I needed it. I thank the women in my life: my friend, Rebecca Frydenger, my sister, Thatsada, and my long suffering mother, Syda, who has patiently waited for her overly romantic son to find his footing as he wandered the globe and various deserts, never quite knowing where else he might end up; I’ve always told her that wherever I am is where I should be until I moved on.

## TABLE OF CONTENTS

<b>LIST OF TABLES</b> .....	viii
<b>LIST OF FIGURES</b> .....	x
<b>Chapter 1</b> .....	1
1.0 Introduction.....	2
1.2 Background.....	2
1.3 Biomaterials .....	5
1.3.1 Molecular Recognition.....	5
1.3.2 Self Assembly .....	10
1.4 Methods.....	16
1.4.1 Molecular Dynamics.....	17
1.4.2 Ab-initio Protein Structure Prediction .....	21
References.....	24
<b>Chapter 2</b> .....	30
2.1 Introduction.....	31
2.2 Methods.....	39
2.3 Results and Discussions.....	42
2.4 Conclusions.....	52
References.....	53
<b>Chapter 3</b> .....	61
3.1 Introduction.....	63
3.2. Materials and Methods.....	65
3.3 Results and Discussion .....	72
3.3.1. Non random sequences .....	72
3.3.2. Polyguanine RNA and DNA duplexes .....	73
3.3.3. Polyadenine RNA and DNA duplexes.....	79
3.3.4 Helices with random sequences.....	85
3.3.5. Na <sup>+</sup> and random helices of RNA versus DNA.....	86
3.3.6 K <sup>+</sup> and random helices of RNA versus DNA .....	88
3.3.7 Effect of geometry: Fixed A-form random helical DNA.....	90
3.4. Conclusions.....	93

References.....	95
<b>Chapter 4</b> .....	103
4.1 Introduction.....	105
4.2 Results and Discussion .....	109
4.2.1 An in silico predicted structure of the GhCESA1 cytosolic region.....	109
4.2.2 Comparison between bacterial and plant cellulose synthases.....	113
4.2.3 Genetic mutations demonstrate functional nodes within plant CESA structure..	115
4.3 Methods and Materials.....	126
4.3.1 Simulations and Modeling .....	126
4.3.2 Novel mutations in Arabidopsis CESAs and phenotypes of mutant plants.....	127
References.....	129
<b>Chapter 5</b> .....	134
5.1 Introduction.....	135
5.2 Results.....	139
5.2.1 Decoy Production Analysis.....	139
5.2.2 Plant Conserved Region.....	140
5.2.3 Class Specific Region .....	144
5.2.4 Protein Motifs .....	146
5.2.5 MD Analysis .....	149
5.3 Discussion.....	149
5.3.1 Rosette Formation and Hierarchy of Terminal Complex Assembly .....	149
5.3.2 Achieving material control of cellulose.....	152
5.3.3 Resolving Paradoxes.....	155
5.4 Methods.....	159
5.5 Conclusion .....	161
References.....	161
<b>Chapter 6</b> .....	168
6.1 Background.....	169
6.2 Methods.....	172
6.3 Results.....	173
6.3.1 Folding distinctions between wild-type and mutant TMH56 .....	176

6.3.2 Predicted BcsA vs true Structure .....	178
6.3.3 Predicted Transmembrane Assembly of a Cellulose Synthase.....	181
6.4 Discussion .....	186
6.5 Conclusion .....	187
References.....	188
<b>Chapter 7</b> .....	192
7.0 Future Directions .....	193
7.1 Nucleic Acids as Nano-materials.....	193
7.2 Synthases and Synthetic Biology at the Membrane Interface .....	194
7.3 Ascertaining mechanism of Isoxaben resistance .....	195
7.4 Role for distal N-terminus membrane helices .....	198
References.....	199
APPENDICES .....	202
Appendix A Chapter 4 Supplemental.....	203
Appendix B Chapter 5 Supplemental .....	230
Appendix C Ion Counting Perl Script.....	239

## LIST OF TABLES

Table 1.1: Representative Mechanical Strength of Materials versus Nanocellulose. ArboraNano Inc.....	13
Table 2.1: Simulations of TAR and TAR-like Sequences Performed under Solvent Conditions as Indicated.....	38
Table 2.2: Cationic Occupancy Statistics Calculated from 20 ns MD Simulations.....	48
Table 3.1: Major and minor groove parameters of simulated duplexes and X-ray structures.....	75
Table 3.2: Interaction energies of high residency ions with specific residues of RNA and DNA after truncating the first 2ns and analyzing the remaining 18ns.....	77
Table 3.3: Occupancy and lifetimes of Na <sup>+</sup> hydrogen bonding to RNA and DNA polypurine duplexes.....	78
Table 3.4: Occupancy and lifetimes of Na <sup>+</sup> hydrogen bonding to RNA and DNA polypurine.....	81
Table 3.5: Occupancy and lifetimes of K <sup>+</sup> and Na <sup>+</sup> minor groove binding to identical random sequences of RNA and DNA after truncating the first 2ns and analyzing the remaining 18ns.....	82
Table 3.6: Occupancy and lifetimes of K <sup>+</sup> and Na <sup>+</sup> major groove hydrogen bonding to identical random sequences of duplex RNA and DNA (cut off distance is 3.8 Å for K <sup>+</sup> ; 3.2 Å for Na <sup>+</sup> ).....	87
Table 4.1: Plant phenotypes for new CESA mutations. Significantly different compared with wild-type (LER or Col-0) as determined by <i>t</i> test: * P< 0.001; **P=0.009; ***P<0.01. RCI, Relative crystallinity index. ....	115
Table 6.1: Folding Statistics. Clusters were evaluated on the top 10% best scoring subpopulations with the maximum entropy approach of the Durandal algorithm. ....	174

Table 4S.1: The PDB identification numbers, E-values, and snapshots of structures used in predicting the structure of the $\beta$ -sheet region of the GhCESA1 cytosolic region using Hidden Markov chain modeling. During the selection of the top models, the SAM-T08 generates pairwise alignments of the target sequence and the best-scoring templates, which are adjudicated by E-value representing how many sequences would score this well in the database. Structures with E-values of less than about 1.0E-5 are very likely to have a domain of the same fold as the target. Structures with E-values of larger than about 0.1 are very speculative.....	211
Table 4S.2: Structure quality scores.....	214
Table 4S.3: Identity and locations of Gh506 structural features. Entries are in order of appearance in the GhCESA1 cytosolic sequence that was used to generate the Gh506 structure (Fig. 1B). Five of these $\alpha$ -helices are designated “core $\alpha$ -helices” because they co-align in the superimposed GT-2 domain of BcsA and the predicted Gh506 structure. Amino acid residue numbers are relative to full-length GhCESA1 (NCBI Accession P93155) or BcsA (NCBI Accession Q3J125; PDB ID 4HG6). Functions ascribed to BcsA are from Ref. [25]. .....	217
Table 4S.4: Summary stability measurements measured as root mean square deviation (RMSD) from the initial structure on the whole structure and on key secondary structure elements of the CESA as a result of mutations over a window of 10 ns. ....	225
Table 5S.1: Genbank sequences.....	234
Table 5S.2: Top 1K Cut Decoy Results – Plant Conserved Region .....	235
Table 5S.3: Top 1K Cut Decoy Results – Class Specific Region.....	235
Table 5S.4: MM/GBSA Calculations: Free energy of best centroid cut from the top scoring models form the last 2ns of MD in explicit solvent. ....	237
Table 5S.5: TM-Align Scoring of Minimized Structures .....	238
Table 5S.6: RMSD Stability of structures in solvent.....	238

## LIST OF FIGURES

Figure 1.1: Typical helical duplexes of Nucleic acids. A-form helix of RNA. (b) B-form helix of DNA.....	7
Figure 1.2: Sugar pucker: (a) 2'Endo as seen in DNA. (b) 3' Endo as seen in RNA. The lack of a 2'OH group renders DNA more flexible. ....	7
Figure 1.3: Depiction of cellulose from the plant wall to its monofilament. (Adapted from Cranston, ED. (2012) "Mechanical Testing of Thin Film Nanocellulose Materials.").....	11
Figure 1.4: Nanocellulose as a biocompatible tissue scaffold. Bodin et al. 2007 <sup>33</sup> .....	12
Figure 1.5: General topology of cellulose synthase glycosyltransferases. PCR is the plant conserved region, an CSR is the class specific region. There are 8 transmembrane helices in total. Adapted from Roberts et al. <sup>44</sup> . D1-3 are conserved aspartic acid motifs. The QxxRW motif is thought to hold the glucan chain as it grows, and is found in all processive glycosyltransferases including chitin synthase. The zinc finger is at the N terminus and thought to allow other protein-protein interactions. ....	15
Figure 1.6: Cellulose microfibrils formed polymerized by rosettes (Neville 1993). ....	16
Figure 1.7: Diagram of Basic Molecular Dynamics Methodology.....	19
Figure 1.8: SAM T08 Secondary Structure Prediction where E labels correspond to beta strands and H labels are alpha helices. Everything unlabeled is a random coil. ....	22
Figure 1.9: ROSETTA starts from (a) fragment libraries with sequence-dependent ( $\zeta$ and $\psi$ ) angles that capture the local conformational space accessible to a sequence. (b) Combining fragments from the libraries, ROSETTA folds the protein by optimizing non-local contacts. A low-resolution energy function smoothes the energy surface, to funnel the structure toward a native conformation as denoted by "N" (c). Metropolis Monte Carlo minimization drives the structure toward the global minimum. Reproduced from Kaufman et al. 2010 <sup>60</sup> .....	23

Figure 2.1: (a) Sequence of the helical stem used in this study with a star denoting the location of the removed bulge. (b) Sequence of TAR RNA duplex (PDB id 397D) with boxed nucleotides important for Tat binding. (c) Superimposition of TAR RNA crystal structure (grey) with four bonded calcium ions (yellow) and a snapshot from MD simulations of a TAR-like RNA helix (orange) and the highest occupancy sodium ion (green) and potassium ion (blue). .....	36
Figure 2.2: Pearson correlation coefficient calculated for ionic occupancies by an atom between two symmetric halves of a polyG/polyC helix as a function of time. Diamonds indicate MD results from 0 to 10ns. The solid line indicates an exponential fit of the data.....	41
Figure 2.3: Sodium ion interactions with polyA/polyU duplexes and polyG/polyC duplexes. (a) Cationic association bins per residue for polyA/polyU and polyG/polyC duplexes. Vertical line denotes the strand break. (b) RDFs of sodium ions with guanine, adenine, cytosine, and uracil. (c,d) Surface map of the sodium occupancy for (c) polyG/polyC and (d) polyA/polyU duplexes calculated for rotational angle $\theta$ around the helical axis. ....	45
Figure 2.4: Sodium and potassium ion interactions with TAR-like sequence. (a) Cationic association bins per residue in different solvents. (b) $\text{Na}^+$ -- purine N7 RDF in mixed solvent solution. The inset represents the observed movement of the cation around A7. (c,d,e) Surface map of the cationic occupancy within the TAR-like duplex calculations: (c) $\text{K}^+$ in 0.1 KCl simulations, (d) $\text{Na}^+$ in 0.1 NaCl simulations, and (e) cations in mixed 0.1 NaCl and KCl solutions. ....	47
Figure 3.1: Convergence of the Pearson correlation coefficient for DNA and RNA helix systems of a random sequence thru 10ns. (a) B-form DNA duplex systems of a random sequence with 0.1M counterions of KCl and NaCl converged to a $0.76 \pm 0.042$ Pearson correlation coefficient by 10 ns; by 10 ns, a 0.76 Pearson correlation coefficient between a NaCl system versus mixed KCl and NaCl was reached for cationic occupancies at residue resolution. (b) A-form RNA duplex systems of a random sequence with 0.1M counterions of KCl and NaCl converged to a $0.79 \pm 0.038$ Pearson correlation coefficient by 7 ns; by 10 ns, a 0.92 Pearson correlation coefficient between a NaCl system versus mixed KCl and NaCl was reached for cationic occupancies at residue resolution.....	68

Figure 3.2: Ion interaction frequency in arbitrary count every 20ps of the last 8ns in the polypurine sequence trajectories for DNA and RNA in 0.1M sodium.....	70
Figure 3.3: Ion interaction frequency in arbitrary count every 20ps of the last 18ns in the random sequence trajectories for DNA and RNA. ....	71
Figure 3.4: Modeled RNA and DNA helices (a) Polyguanine RNA sequence. (b) 3'Endo Sugar Pucker of RNA's backbone. (c) Polyguanine DNA sequence. (d) 2' Endo Sugar Pucker of DNA's backbone.....	78
Figure 3.5: Heat maps of cumulative diffusive residencies of individually tracked sodium cations within 5 Å of a nucleobase in polyguanine duplexes of RNA and DNA.....	79
Figure 3.6: Modeled RNA and DNA helices (a) Polyadenine RNA sequence. (b) Chemical structure of the nucleobase Uracil. (c) Polyadenine DNA sequence. (d) Chemical structure of the nucleobase Thymine. ....	81
Figure 3.7: Heat maps of cumulative diffusive residencies of individually tracked sodium cations within 5 Å of a nucleobase in polyadenine duplexes of RNA and DNA.....	84
Figure 3.8: Cumulative residencies of cations in the cation-duplex modeled systems: Sodium cation diffusive interaction events within 5Å to non alternating tracts of purine nucleotides in RNA and DNA duplexes simulated to 10ns. On the x-axis, residues 1-11 are the purines, and residues 12-22 are the pyrimidines. ....	85
Figure 3.9: Modeled RNA and DNA duplexes (a) RNA random sequence. (b) A-RNA helical structure colored to show electrostatic surface, red is negatively charged. (c) DNA random sequence. (d) B-DNA helical structure for the random sequence; red is negatively charged.....	85
Figure 3.10: Heat maps of cumulative diffusive residencies of individually tracked sodium cations within 5 Å of a nucleobase for duplexes of RNA and DNA of identical random sequences. ....	88
Figure 3.11: Heat maps of cumulative diffusive residencies of individually tracked potassium cations within 5 Å of a nucleobase for RNA and DNA duplexes of identical random sequence. ....	89
Figure 3.12: Heat map of cumulative diffusive residencies of individually tracked sodium cations within 5 Å of a nucleobase for A-DNA of a random sequence.....	92

Figure 3.13: Cumulative residencies of cations in the cation-duplex modeled systems: Sodium and potassium ion interacting within 5 Å of a nucleobase in random sequence duplexes of RNA and DNA. On the x-axis, residues 1-11 are the purines, and residues 12-22 are the pyrimidines. .... 93

Figure 4.1: Predicted structure of the large cytosolic region of GhCESA1. (A) Diagram of GhCESA1 showing 8 predicted TMH and the large cytosolic loop between TMH2 and TMH3. Labels within the cytosolic loop indicate: the approximate locations of the four conserved motifs; the P-CR region; the CSR region; and the analogous locations for missense mutations in Arabidopsis CESAs that perturb cellulose synthesis (black and red type for published or newly reported mutations, respectively). (B) Snapshot of the Gh506 structure. The catalytic core is grey, the P-CR is pink, and the CSR is light blue. The catalytic core contains a  $\beta$ -sheet with six strands in yellow ( $\beta$ -1 S287-S291;  $\beta$ -2 D253-S257;  $\beta$ -3 F454-D459;  $\beta$ -4 C532-N535;  $\beta$ -5 Y488-F491; and  $\beta$ -6 S686-C689). Green highlights DD, DCD, ED (directly behind DCD), and the position of QVLRW within core  $\alpha$ -13. The five  $\alpha$ -helices that are part of the GT core are labeled: [ $\alpha$ -2 L267-A278;  $\alpha$ -6 H433-V448;  $\alpha$ -7 N466-D479;  $\alpha$ -8 N508-K517; and  $\alpha$ -13 S705-R725 (containing QVLRW)]. All predicted  $\alpha$ -helices are numbered in order within the primary sequence. (C) Diagram of the secondary structure showing a total of 6  $\beta$ -strands (yellow arrows) and 13 major  $\alpha$ -helices (shown as barrels) in three regions: catalytic core (red outlines); P-CR (pink fill); and CSR (blue fill). Possible additional shorter helical regions are indicated as unnumbered small barrels. (D) UDP-Glc docked into the catalytic site above DCD. .... 107

Figure 4.2: Possible oligomeric assemblies of the Gh506 cytosolic structure under (A) C2, (B) C3, (C) C4, and (D) C6 crystallization symmetries. The catalytic region is grey, the CSR is light blue, the P-CR is pink, QVLRW is yellow and the site of *fra6* mutations is red. (D) Bottom, (E) top, and (F) side view of the hexameric Gh506 assembly. .... 112

Figure 4.3: Comparison of the Gh506 model with the structure of bacterial cellulose synthase. (A) Surface representation of the *Rhodobacter sphaeroides* (Rs) cellulose synthase (PDB ID 4HG6) superimposed with the Gh506 structure. The model aligns well with the central  $\beta$ -sheet of the bacterial GT-domain, and the P-CR and CSR domains point away from the membrane-spanning region. UDP and the translocating glucan of the Rs cellulose synthase highlight the active site and the TM-channel and are colored violet and cyan, respectively. The BcsA subunit of the Rs cellulose synthase complex is colored gray and green, respectively, BcsB is colored wheat. (B) Superimposition of the Rs GT-domain with the Gh506 model by secondary structure of the central  $\beta$ -sheet. The GT-domain is colored green for RsBcsA and gray and yellow for the GhCESA1 model, respectively. The GhCESA1 P-CR and CSR domains are colored pink and light blue, UDP and the glucan are shown as spheres. (C) Conserved sequence motifs that form the binding site for UDP and the acceptor glucan are compared with the corresponding residues in RsBcsA. The Rs and Gh sequence alignments of the motifs are shown and the depicted residues are indicated in bold (black: Rs, blue: Gh). RsBcsA is shown in gray and the UDP and glucan bound to Rs BcsA are colored violet and cyan, respectively. Residues from Gh506 are colored blue. The ED motif was omitted for clarity. Horizontal bars indicate the membrane boundaries..... 114

Figure 4.4: Previously known missense mutations (green) in Arabidopsis CESAs mapped onto the predicted GhCESA1 cytosolic structure. New mutations are blue with the DD, DCD, ED, and QVLRW motifs in red. Two opposite rotations (a) and (b) provide a view of where the mutations map relative to the catalytic region of the Gh506 structure. The equivalent Ghcesa1 amino acid position precedes the mapped mutation: R351 (*Atcesa8*<sup>R362K</sup>, *fra6*); A436 (*Atcesa3*<sup>A522V</sup>, *eli1-2*); A447 (*Atcesal*<sup>A549V</sup>, *rsw1-1*); D459 (*Atcesa7*<sup>D524N</sup>, *irx3-5*); P492 (*Atcesa7*<sup>P557T</sup>, *fra5* and *Atcesa3*<sup>P578S</sup>, *thanatos*); G529 (*Atcesal*<sup>G631S</sup>, *rsw1-2*); G531 (*Atcesa3*<sup>G617E</sup>, *cev1*); S668 (*Atcesa3*<sup>S679L</sup>, *irx1-2*); E671 (*Atcesal*<sup>E779K</sup>, *rsw1-45*); D672 (*Atcesa3*<sup>D683N</sup>, *irx1-1* and *Atcesal*<sup>D780N</sup>, *rsw1-20*); and H680 (*Atcesa7*<sup>H734Y</sup>, *mur10-2*). The conserved residues with no known mutations are red..... 116

Figure 4.5: The locations of novel missense mutations in the predicted structure helped to support the existence of new functionally important regions within CESA. Conserved residues are shown in red. (A) S291 (teal) just below DD is the analog of the novel *Atcesa3*<sup>S377F</sup>, *ixr1-6*, mutation. In the predicted structure, it contacts L442 (rust ball and stick residue) within  $\alpha$ -6 (rust), which has the analogs of *Atcesa3*<sup>A522V</sup> (*eli1-2*; brown) and *Atcesal*<sup>A549V</sup> (*rsw1-1*; tan) at either end. (B) The P492-G518 loop (brown) contains native aspartates (green ball and stick residues) near QVLRW. At its base are G518 (blue; the analog of the novel *Atcesal*<sup>G620E</sup>, *lycos*, mutation) and P492 (rust; the analog of *Atcesa7*<sup>P557T</sup>, *fra5*), where they may putatively act as hinge points. (C) Cross correlation of atomic fluctuations at four mutation sites over all simulations by residue. The peaks shown had at least 97% correlation, indicating distant effects of the mutations analogous to *Atcesa3*<sup>S377F</sup> (*ixr1-6*; blue bars), *Atcesal*<sup>G620E</sup> (*lycos*; green bars), and *Atcesa7*<sup>P557T</sup> (*fra5*; red bars). ..... 123

Figure 5.1: Putative homomeric-hexamer rosette assembly. (a) Class Specific Region b) Plant Conserved Region (FRA6 mutation amino acid position is in red) c) A putative hexameric assembly of a rosette, for scale, cryo-em image of a Terminal Complex comprised of six rosettes is shown from Kimura 1999. .... 137

Figure 5.2: Representative Folding Funnels of Primary and Secondary PCR's and CSR's. The y-axis is calculated as RMSD values from the lowest scoring decoy. The x-axis is the score in terms of Rosetta Energy Units (R.E.U.). (a) PCR3 density plot (b) PCR3 folding funnel (cluster of interest is in red) (c) CSR3 density plot (d) CSR3 folding plot (cluster of interest is in red) (e) PCR7 density plot (f) PCR7 folding plot (cluster of interest is in red) (g) CSR7 density plot (h) CSR7 folding plot (cluster of interest is in red). .... 139

Figure 5.3: Primary (a-c) and Secondary (d-f) PCR's. (a) PCR1 (b) PCR3 (c) PCR6 (d) PCR4 (e) PCR7 (f) PCR8 (g) Secondary structure block diagram labeled according to each PCR. The N-terminus is to the left for the block diagram. .... 141

Figure 5.4: Comparison of Wild Type PCR of *Atcesa8* and the FRA6 mutant. (a) Folding funnel of PCR8 (b) Folding funnel of FRA6 mutant (c) Secondary structure comparison with tertiary alignment. An asterisk over the top block diagram is the lysine mutant of FRA6. The bottom block diagram is that of PCR8. Helices are in red with disordered regions as a green line. .... 142

Figure 5.5: Electrostatic surface potential of the Wild Type Atcesa8 PCR (a) and FRA6 (b). .....	144
Figure 5.6: Primary and Secondary CSR's: (a) CESA1 (b) CESA3 (c) CESA6 (d) CESA4 (e) CESA7 (d) CESA8 (g) Secondary Structure layout of a-f. The N-terminus is to the left for the block diagram.....	145
Figure 5.7: Electrostatic surface potentials of the primary versus secondary wall associated class specific region folded fragments. CSR1 (a), CSR3 (b), CSR6 (c), CSR4 (d), CSR7 (e), and CSR8 (f).....	147
Figure 5.8: Unified Trimer of Dimers CSC model to deconflict the results of Carroll et al. <sup>25</sup> and Wightman et al. <sup>52</sup> .....	155
Figure 5.9: Walker P Loop (GXXXGK[T/S]) of CSR1 highlighted in magenta with the serine and arginine in line form. The loop is a binding motif for the phosphate of nucleotides. ....	157
Figure 6.1: General topology of plant cellulose synthases (a) adapted from Harris et al. 2012 <sup>1</sup> . Eight membrane spanning helices are enumerated as shown. Transmembrane 5 and 6 are connected by a very long outer membrane loop with a unique amphiphatic structure. Psipred alignment and secondary structure prediction(b). Underlined in green are the regions the OCTOPUS topology server predicted to reside within the plasma membrane. Helices are in red with beta strands in yellow. The length of each segment is annotated to the right.....	175
Figure 6.2: Folding Funnels for TMH5 & 6: Atcesa1 (a); Atcesa3 (b); Atcesa6 (c); Atcesa3 <sup>T942I</sup> (d). Highest densities are in red. The biphasic folding patterns of the wild type sequences (a-c) contrast greatly with the multiple subpopulations seen in T942I (d). ....	176
Figure 6.3: Representative of best centroid decoys for the transmembrane helices 5 and 6 from all the CESAs and the mutant T942I. Atcesa1 (a), Atcesa3 (b), Atcesa6 (c), Atcesa3 <sup>T942I</sup> (d).....	177
Figure 6.4: Best scoring decoys for the transmembrane helices 5 and 6 from the primary wall associated CESAs and the mutant T942I. Atcesa1 (a), Atcesa3 (b), Atcesa6 (c), Atcesa3 <sup>T942I</sup> (d).....	178

- Figure 6.5: Prediction of a membrane associated helix pair in BCSA corresponding to helices 5 and 6 of CESAs in Arabidopsis. The folding funnel plot does not indicate a biphasic folding path with over 21K decoys generated (a). The difference between the centroid of the largest cluster (b) and the best scoring decoy (c) are very slight..... 178
- Figure 6.6: Comparing the Rosetta predicted structure with that of the actual crystal structure corresponding to TMH56 of plant CESA (blue). The recently solved structure of BcsA (PDB ID: 4HG6) (a). RMSD fit to the crystal structure: 9.899 Å for the best cluster centroid (orange) with 59 residues aligned (b); 8.734 Å for the lowest energy structure (red) with 58 aligned residues (c). Since the lowest energy structure also bears close resemblance to the centroid of the largest cluster such that the fittings are almost the same..... 180
- Figure 6.7: Membrane organization of helices 3-8 of Ghcesa1 from composite predictive runs of four helix bundles aligned on TMH 5 and 6. TMH3 (magenta); TMH4 (wheat); TMH5(grey); TMH6(orange); TMH7 (red); TMH8 (blue). Inner cytosolic region is down. The linker between TMH 5 and 6 display bistability between the best scoring decoys (a) and the best centroid (b) alignments. The beta sheet from of the linker is in yellow (a) and assumes a conformation of random coils and small helices in cyan (b). A semblance of a pore is visible in (c) where the cavity is taken up by the unusual beta sheet formation (yellow) of the linker between TMH 5 and 6. .... 182
- Figure 6.8: Membrane organization of helices 3-8 of Ghcesa1 with the Atcesa3<sup>T942I</sup> amino acid mutation from composite predictive runs. TMH3 (magenta); TMH4 (wheat); TMH5(grey); TMH6(orange); TMH7 (red); TMH8 (blue). The alignments are rather poor for the best scoring decoys (a). The centroids of the largest clusters also do not show a real pattern (b)..... 185
- Figure 7.1: Isoxaben (a) targets CESA3 and CESA6, and ethyl-methanesulfonate (b)..... 196
- Figure 7.2: Membrane organization of helices 3-8 of Ghcesa1 from composite predictive runs of four helix bundles aligned on TMH 5 and 6. TMH 3 (magenta) and TMH 8 (blue). Inner cytosolic region is down. The linker between TMH 5 and 6 display bistability between the best scoring decoys (a) and the best centroid (b) alignments. The beta sheet from of the linker is in yellow (a) and assumes a conformation of random coils and small helices in cyan (b)..... 197

Figure 7.3: The N-terminal helices (TMH 1 and 2)with the first C-terminal transmembrane helix (TMH 3) from Ghcesa1: TMH1 (orange); TMH 2(red and pink); TMH 3 with artificial linker (cyan).....	198
Figure 4S.1: Residues from GhCESA1 that were included in the Gh506 structure are aligned with the same regions of Arabidopsis CESAs with missense mutations. Numbering is relative to residue position in full length GhCESA1. Pink and blue lines indicate the positions of the P-CR and CSR plant-specific regions, respectively. Red and yellow rectangles indicate $\alpha$ -helices and $\beta$ -sheets, respectively. By comparison to the structure of RsBcsA (see the main text), $\alpha$ -2,6,7,8,13 and $\beta$ -1–6 are predicted to be in the core GT domain. Light purple vertical highlights show the position of selected conserved domains. Large green letters indicate sites of missense mutation in the AtCESA indicated.....	209
Figure 4S.2: (A-C) Aligned structures used in model prediction as listed in Table 4S.1. Side of the $\beta$ -strands (A) colored by individual structure and (B) colored by secondary structure. (C) View of the slice to expose the $\beta$ -sheet region, with individual $\beta$ -strands numbered $\beta$ 1– $\beta$ 6. (D)The snapshot of the starting structure of the Gh506 cytosolic domain from the SAM-T08 HMM structure prediction server. (E) The predicted structure after molecular dynamics refinement with six $\beta$ -strands in yellow and DD, DCD, and ED in green. The $\alpha$ -helices dispersed throughout the structure are red.....	210
Figure 4S.3: A similar prediction of the secondary structure was obtained using PSIPRED v3.3. Highest confidence prediction areas are given a value of 9, lowest 0. The DCD putative UDP binding motif, reidues 459-461, are at the tip of a $\beta$ -strand. The $\beta$ -strands highlighted in red were lost with MD refinement. In contrast, a low coil confidence area highlighted magenta (686-689) became a $\beta$ -strand. Greyed boxes did not appear in the final structure due to their low confidence levels. Yellow highlights the retained $\beta$ -strands. Green boxes highlight the DD, DCD, ED, and QVLRW motifs. ....	213

Figure 4S.4: Comparison of the quality of the Gh506 structure to experimentally solved structures. (A) Pro-SA Z scores for various stages of GhCESA1 structure prediction (labeled green, black, and red dots) compared to scores of solved structures from the PDB databank (dense blue dots). The initial Z score of the predicted GhCESA1 cytosolic structure (-3.4, green dot) was improved to -5.56 (black dot) after about 4 ns of MD refinement and reached -6.09 (red dot) after a series of MD simulations followed by a short minimization. (B) ERRATv2 analysis of the predicted GhCESA1 cytosolic structure (i) and the solved structures of three other GT-2 enzymes used as templates [ii, K4CP domains A and B (PDB: 2Z86); iii, SpsA (PDB: 1QG8); and iv, a putative glycosyltransferase from *Bacteriodes fragilis* (PDB: 3BCV)]. The histograms show the error value of residues, and the band in the middle of the graph indicates the difference between the lower 95% and the upper 99% value. Of the three crystal structures, the 218 amino acid structure of 3BCV from *Bacteriodes fragilis* exhibited the best score with only B chain residue 40 showing significant error. Areas possibly in need of further refinement in the GhCESA1 predicted structure include residues that either have high local mobility or are deeply buried: (1) N457-V464; (2) D253-V256 that form a  $\beta$ -strand adjacent to the putative UDP binding motif, DCD, in the catalytic core; (3) solvent-exposed P327-I335 that fold back into residues V347-R355 within the P-CR region; (4) P492-G518 that appear to form a loop beside the catalytic site that abuts the QVLRW motif. Even for the SpsA structure, similarly buried residues are nearly impossible to refine fully. For K4CP, core residues around the UDP binding motif of domain “B” shows the greatest error values, probably because they are more mobile and solvent accessible. Similarly, a small region near the UDP-binding motif of SpsA (residues 130-135) also exhibits error values greater than 95% as exemplified by the filled in black bars. .... 215

Figure 4S.5: Interaction of manganese uridine diphosphate glucose (MnUDP-G) complex with residues of the modeled CESA. The positions of the “D” residues were taken from the CESA structure generated in this study and all atomic positions were allowed to relax to minimum energy positions determined by our DFT methodology. Mn-O distances to carboxylate group of the D residues and to the diphosphate moiety of UDP are given in Angstroms. H=white; C=grey, O=red, N=blue, P=orange, Mn=green. This geometry was used to dock the UDP-Glc into the Gh506 structure in Fig. .... 220

- Figure 4S.6: Three loops (upper left corner of the image) in the vicinity of the UDP-Glc binding site of the Gh506 structure that may help to control catalysis through modulation of local accessibility to key residues: (1) T258–L267 at the end of  $\beta$ -2 (green); (2) A294-F300, just after DDG and leading into  $\alpha$ 3 of the PCR (orange); and (3) Y421-H432, leading from  $\alpha$ 5 into core  $\alpha$ 6 (aqua). The conserved motifs, DD, DCD, ED, and QLVRW are highlighted red, the  $\beta$  sheet is yellow, the P-CR is pink, and the CSR is blue..... 221
- Figure 4S.7: Correlated residue motions via atomic fluctuations. The CSR region, residues Y540-W658, shows the greatest motion correlation to itself as expected. The P-CR region, residues A295-V420, shows a self-correlation as well, but not as strong since it is less ordered. .... 222
- Figure 4S.8: Top left view: A possible hexameric assembly of one CESA cytosolic domain isoform (the predicted structure from GhCESA1). One monomer is shown in the ribbon diagram at the top, showing the location of the barely visible  $\beta$ -sheets (yellow) below motifs with conserved D residues (green). The catalytic regions of the other monomers are shown in aqua, magenta, yellow, orange, and dark blue. The light blue and pink regions are the CSR and the P-CR regions, respectively, for all monomers. Top right view: Possible packing of hexameric assemblies into an orthorhombic unit cell of space group  $P2_12_12_1$  (red box). Note that this theoretical possibility for crystallization of hexamers of the predicted GhCESA1 cytosolic region does not imply any preference for hexameric subunits of the rosette CSC *in vivo*. The number of CESAs in the rosette CSC remains an open question. .... 223
- Figure 4S.9: A comparison between Gh506 and RsBcsA sequences and structures. (A) A sequence alignment of a cytosolic region for GhCesA1, AtCesA1, and RsBcsA. (B,C) A top and bottom view of structural alignment between Gh506 (red) and RsBcsA (blue) without the regions that do not have a template and the region shaded blue. This reduced both structures to the GT-A core and the structures align with an overall RMSD of 3.9 Å. .... 224
- Figure 4S.10: Hydrogen bonding of P492T to Y688. The distance cut off is 3.5 Å. The strongest interactions during this time interval for *Ghcesa*<sup>P492T</sup> is before the 4 ns mark. This interaction may serve to stabilize the P492-G518 loop..... 226

Figure 5S.1: Sequence alignment of PCR's: Ranked by descending similarity. Consensus regions are underscored with black boxes. Two regions of high variability are boxed in red. ....	231
Figure 5S.2: Sequence Similarity of Arabidopsis PCR's (no. 11 corresponds to Ghcesa1) .....	232
Figure 5S.3: Sequence alignment of CSR's: Ranked by descending similarity. Consensus regions are underscored with black boxes. Two regions of high variability are boxed in red. ....	233
Figure5S.4: Sequence Similarity of Arabidopsis CSR (no. 11 corresponds to Ghcesa1) .....	234
Figure5S.5: Folding plots of primary and secondary associated PCRs: PCR1 (a), PCR3(b), PCR6(c), PCR4(d), PCR7(e), PCR8(f).....	236
Figure 5S.6: Folding plots of primary and secondary associated CSRs: CSR1 (a), CSR3(b), CSR6(c), CSR4(d), CSR7(e), CSR8(f). ....	237

# Chapter 1

## Introduction

## **1.0 Introduction**

As the ability to manipulate matter at the nanoscale improves, it becomes increasingly important to understand the laws and principles that permit an efficient (economic) molecular assembly of new materials and engineer properties that fit a functional application. These may be the design of sensors, electronic and photonic devices, and structural composites. Nature has solved the assembly problem in bio-nanosystems in order to generate biopolymers such as nucleic acids, proteins and polysaccharides such as cellulose. The synthesis of new materials in biological systems is largely accomplished by protein synthases that are able to selectively recognize their substrates and alter the physical properties of their product. Bio-nanotechnology is an emergent field that promises to deliver significant advances for industries reliant on natural materials as well as engendering new technologies for applications in green chemistry, therapeutics, sensors, hybrid nano-composites and construction. Despite this array of potential applications and development, there still remains a paucity of fundamental insights into the basic mechanisms that nature exploits in order to achieve programmable self-assembly of molecules at the nanoscale resulting in macroscopic structures.

## **1.2 Background**

Demand for renewable resources is increasing as world economic development consumes finite resources of hydrocarbon, rare earth elements, and other raw materials. Economic growth generates industrial pollution and green house gasses since the world is still mainly reliant on fossil fuels; a troubling matter for international security, the majority of

hydrocarbon resources are located in geopolitically unstable areas. Furthermore, worldwide consumption of building materials largely focuses on wood. In order to meet the challenges of discovering renewable energy resources and replacing functional materials derived from finite resources, bio-inspired nanotechnologies offer limitless potential. Operating at the nanoscale in dimensions of 1 to 100nm, nanotechnology can meet the need for clean energy, green chemistry, advanced therapeutics, and atomic level fabrication of novel electronics.

A class of biomaterials of considerable interest are nucleic acids due to their intrinsic properties as polyelectrolytes, molecular recognition capabilities, and ease of incorporation into hybrid polymers<sup>1</sup> and functionalized hydrogels<sup>2,3</sup>. As a renewable resource, nucleic acids are produced by every living organism. Nucleic acids can be given additional functions by being conjugated to other materials or be used as a scaffold in order to create patterned 3-D structures<sup>4</sup>. Of these applications, DNA based templating of nanostructures is also showing promise. The mechanical, electrical and optical properties of these materials are tunable<sup>1</sup> as evidenced by using DNA in a copolymer doped for lasing applications<sup>5,6</sup>. Since these applications rely on nucleic acid base-pairing and molecular recognition, understanding how to finely tune these properties is vital toward creating an all organic engineered system with the desired properties suitable for its intended purpose.

The second biomaterial examined in this work is cellulose, specifically, the molecular machinery that is responsible for its synthesis and assembly into nanofibrils. Cellulose, long glucan chains of cellulobiose, is also an abundant and renewable resource. In the form of nanocellulose sized between 1 to 20nm, there are broad applications as a novel therapeutic drug delivery system. Like nucleic acids, nanocellulose can be incorporated into hybrid

nanocomposites<sup>7</sup>. Unlike most nanomaterials, cellulose is both cheap to manufacture and relatively nontoxic; carbon nanotubes (CNT) cost between \$50 to \$500 per kilogram<sup>8</sup> and generally act like asbestos fibers<sup>9</sup> with debate remaining on single walled CNTs<sup>10</sup>; synthetic DNA still costs several thousand dollars per gram (Oligofactory Inc.); in contrast, the US Forrest Service expects to sell nanocellulose for a few dollars per kilogram.

But it is in the field of renewable energy where much of the interest in cellulose lies. The first challenge for is to develop economical ways to free cellulosic fibrils from source material<sup>11</sup>. Cellulosic ethanol is a desirable alternative to corn-based ethanol since it will not compete with the food supply. Ethanol use in gasoline fuels reduces smog since it is 36% by weight oxygen, helping to provided complete combustion and avoid the use of toxic chemicals based on benzene such as *Methyl tert-butyl ether* (MTBE). As a structural material, cellulose's strength is correlated to its crystallinity that renders cellulose resistant to chemical attack, microbial ingestion, and confers significant tensile strength and rigidity. For construction materials, these mechanical properties are highly desirable; yet, these same properties hinder biofuel production. The mechanisms that govern crystallinity lie in the proteins, cellulose synthases, and the protein complexes that polymerize its production. Being able to control cellulose production at the point of synthesis would represent a significant advance. In order to accomplish this goal, this work uses the tools of structural biology and computational molecular dynamics to examine new cellulose synthase mutants that result in lower crystallinity and altered saccharification. However, the structure of cellulose synthases have not been solved experimentally due to problems in transgenic expression and isolation of protein in sufficient yield<sup>12</sup>. To overcome this problem, we used

ab-initio structure prediction methods to generate a likely 3-D structure. After assessing its quality, we mapped the single point mutations to this structure and subjected each mutant to computational analysis. By examining the structure of these mutants relative to the normal wild type, it will be possible to design new proteins that will allow for the tuning of cellulose to have defined properties. Developing economical cellulosic ethanol will reduce the current reliance on corn based ethanol which competes as a food source; when these dual use crops are affected by adverse harvests, then result is higher food prices.

### **1.3 Biomaterials**

#### **1.3.1 Molecular Recognition**

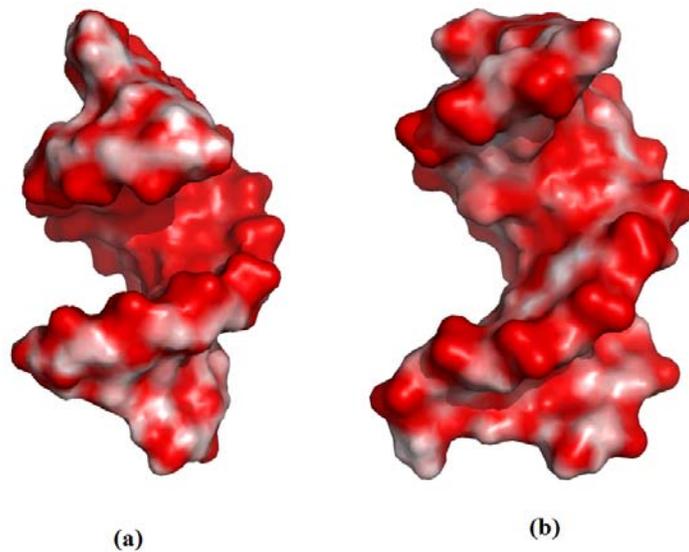
Molecular recognition in biology is a central requirement in order to achieve specificity of reaction and the correct function of signaling pathways in the crowded macromolecular environment of living cells. These interactions are non-bonded and typically rely on hydrogen bonding, solvent effects, van der Waals forces,  $\pi$ - $\pi$  interactions, and electrostatic effects. From these sets of forces, biological materials, like nucleic acids, can selectively bind everything from large macromolecules to small cations.

##### **1.3.1.1 Nucleic Acids**

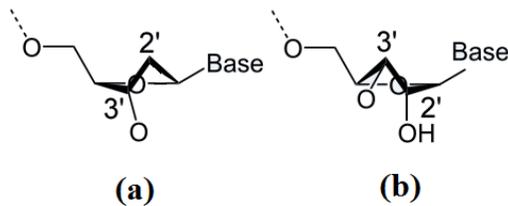
The most error-sensitive biological processes involve genetic transcription and genomic repair. Nucleic acids follow a simple rule based method of molecular recognition through complementary base pairing. Because of this trait, nucleic acids have been extensively studied as model systems for understanding molecular recognition. Nucleic acids

are of two varieties, ribonucleic acid (RNA) or deoxyribonucleic acid (DNA). Nucleic acids encode the genetic information of all living cells. RNA's sugar contains a 2'OH group unlike DNA. Both exist as polymers of nucleotides that contain a five-membered ribose sugar as a backbone and a nucleotide base. They share the same purine bases of adenine (A) and guanine (G), one pyrimidine, cytosine (C); however, they differ with the complement to adenine where it is thymine (T) for DNA and uracil (U) for RNA. Consequently, the Watson-Crick base pairing rules are A-T/U and G-C. This canonical base pairing is what confers unto nucleic the ability to form helical duplexes, recognize complementary pairs, and give rise to the structural variety of bulges, loops, and hairpin turns. One key distinction is that due to the lack of a 2'OH group, DNA generally favors to be a duplex (Figure 1.1). The two most prevalent structures are A-form and B-form. They are characterized by their helical parameters of twist, base pairs per turn, and the rise per base pair, and pitch per turn. Because RNA has a 2'OH group, its duplexes are of the relatively stiff A-form characterized by a tight 11 base pairs per turn and 2.4Å rise. The A-form is also on average of a wider diameter 24Å. The B-form helix which is more typical for DNA is narrower at 20Å with 10.4 base pairs per turn. Consequently, the helix backbone forms grooves of where the backbones are far apart, major, and when they are closer, minor. Another consequence of the 2'OH group is that it forces the sugar pucker of RNA into a C3-endo, and that of DNA into C2'endo (Figure 1.2). This also makes DNA more stable than RNA such that it is an ideal structure for storing genetic information. RNA, on the other hand, can exist as a duplex in short sequences, but mostly assumes a disordered structure consisting of loops and hairpin turns. RNA's 2'OH group prevents efficient packing into duplexes and also may undergo

spontaneous cleavage. This structural disorder and affinity for self cleavage allows RNA to take part in biochemical reactions. Because of its relative stability, DNA is the nucleic acid of choice for genetic storage and it can be wound tightly around histone proteins to provide a compact structure that is further assembled into chromatin. This self regulating nature of nucleic acids arises from their structural motifs and charged surface landscape associated with these duplex structures.



**Figure 1.1:** Typical helical duplexes of Nucleic acids. A-form helix of RNA. (b) B-form helix of DNA



**Figure 1.2:** Sugar pucker: (a) 2' Endo as seen in DNA. (b) 3' Endo as seen in RNA. The lack of a 2'OH group renders DNA more flexible.

### 1.3.1.2 Nucleic Acid Secondary Structures as Recognition Motifs

In nucleic acids, the number of possible secondary structures is approximately  $1.8^N$ , where N is the number of nucleotides<sup>13</sup>. Despite this large conformational space of possibilities, there are three main structural recognition motifs in nucleic acids, bulges, loops, hairpins, and duplexes. Bulges are formed when one or more consecutive bases in a duplex do not pair to bases on the complementary strand. Loops are long stretches of unpaired bases typically greater than 4. Hairpin loops result from two regions of the same strand fold back onto itself and forms a duplex of a complementary sequence. Interior loops and junction loops occur when bases on both sides of a duplex cannot form base pairs; junction loops result when at least two double stranded regions converged and form a closed structure. These two latter two structures are typical for RNA, but can occur in DNA. With the two most common helical duplex forms of the A-form helix and the B-form helix, the tendency to transition from one form to another is largely dependent upon both the nucleic acid sequence and the molar salinity of the microenvironment. Variations in these canonical structures are detectable by regulatory proteins that perform the task of recognizing sequences which determine the local geometry of the duplex. The major groove of B-form DNA helices serve as the binding site of proteins; in contrast, the geometry of A-form RNA helices are poor targets of proteins<sup>14</sup>. Consequently, proteins recognize RNA mostly by its loop and bulge structures<sup>15,16</sup>.

### **1.3.1.3 Nucleic Acid Structure Influence by Small Monovalent Cations**

In order to assume their secondary structural forms, both DNA and RNA must rely on metal cations to provide charge screening of their highly negative phosphate backbones. The bases themselves also exhibit small dipole moments that may sequester small cations and thereby perturb the secondary structure<sup>17</sup>. Consequently, these structures are also amendable to influence by the ionic environment such that some bacteriophage viruses utilize sequence dependent cation recognition motifs to regulate genetic expression; and thereby, reduce delay in the transcription process when cations are used as a control signal<sup>18,19</sup>. The susceptibility toward discernable structural changes of the duplexes induced by the ionic environment for these kind of regulatory processes appear to be sequence dependent<sup>20</sup>. Moreover, monovalent cations are of interest since both RNA and DNA can be highly sensitive to their cytosolic concentration and species resulting in adaptive recognition<sup>18,21-23</sup>.

### **1.3.1.4 Materials Science of Nucleic Acids**

Conjugated nucleic acids with nanomaterials are unusually stable compared to proteins<sup>24</sup>. For this reason, they make ideal platforms for sensor development where their complementary Watson-Crick type hydrogen bonding, stacking interactions and electrostatics can be used in systemic evolution to recognize key ligands by exponential enrichment (SELEX). When nucleic acids have been designed to provide scaffolding for single walled nanotubes, the helical duplex is preferred for its relative rigidity and stability, and allows for uniformity of the desired pattern<sup>25</sup>. The tendency for particular molecules to intercalate in certain modalities of nucleic acid helical duplexes, make them useful as a

medium to control the dispersion of photoactive substrates as well as metallization in applications involving optoelectronics, photonics and as organic semiconductors. Even with the influence of sequence on structure, nucleic acids must first rely on cationic charge compensation which is a first order effect. For this reason, we investigated the role cations have on structure which will ultimately allow for the fine tuning of the secondary structures and their recognition properties of interest that emerging nanotechnologies depend upon.

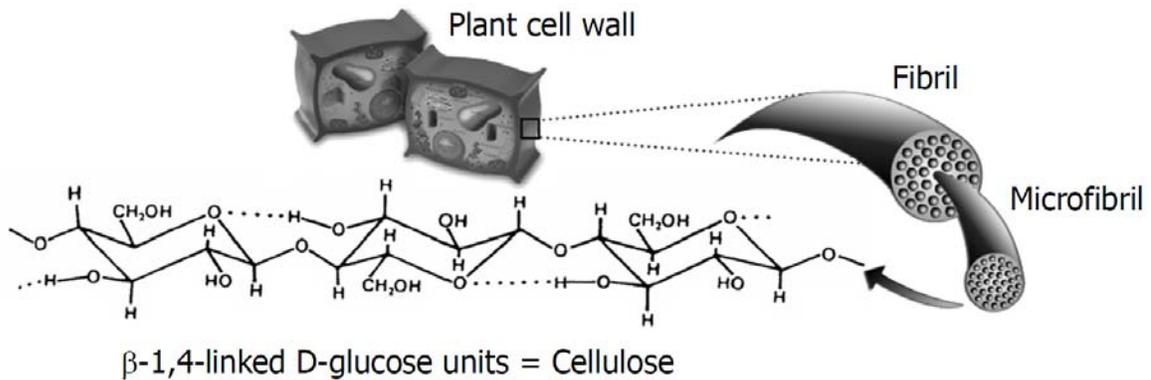
### **1.3.2 Self Assembly**

Self assembly is the process in which molecular nanostructures self-organize. The forces that drive this process occur between individual molecules, between the atoms that make up the molecule and the microenvironment. These materials spontaneously create increasingly complex structures given suitable conditions. All natural biomaterials are formed in this manner.

#### **1.3.2.1 Materials Science of Cellulose**

Cellulose is the most abundant renewable biomaterial; approximately  $10^{11}$  metric tons is produced annually on earth. Cellulose exists as a linear polysaccharide polymer of semicrystalline fibrils (Figure 1.3). Each microfibril is estimated to be 3nm thick and made up of 36 crystalline chains<sup>26</sup>. The highly crystalline nature of cellulose protects it from chemical attack and confers considerable strength on the order of 128 GPa<sup>27</sup>. At approximately 100-150nm in length, these microfibrils contain an amorphous region weaker than the rest of the fiber<sup>28</sup>. The glucose units that constitute  $\beta$ -1,4 glucan are rotated 180°

from each other to assume a non-planar orientation. The length of these polymeric chains is estimated to be upwards to 15,000 glucans. The average crystallinity of cellulose for plants ranges widely depending on the tissue, plant species and environmental growth conditions such as the stress from high winds; for Arabidopsis, it is at most 55% in the stem tissue and can be as low as 38% for leaves<sup>28</sup>.



**Figure 1.3:** Depiction of cellulose from the plant wall to its monofilament. (Adapted from Cranston, ED. (2012) “Mechanical Testing of Thin Film Nanocellulose Materials.”)

With the emergence of cellulose nanofibrils, cellulose fiber-reinforced composites have been made to be relatively strong with the advantage of being transparent and biodegradable<sup>29</sup>. Composites of cellulose often rely on chemical<sup>30</sup> and mechanical<sup>31</sup> treatments to optimize the size and purity of the nanofibrils. Being able to lower the cellulose crystallinity, in this case, would be helpful in simplifying processing and obtaining uniform raw materials.

Perhaps the greatest interest in cellulose structure is in improving the production of cellulosic ethanol for biofuels. World energy consumption is expected to increase

significantly by 2025 as human populations grow, resulting in more demand for biomass based fuel. Cellulosic ethanol is relatively environmentally friendly and may provide renewable transportation fuel. As one advantage over corn ethanol, cellulosic biofuels do not require fertilizers, pesticides, significant energy, and water to produce. Ethanol as an oxidizer in gasoline fuel, reduces smog, since it is 35% oxygen by weight and significantly less toxic than benzene-based oxygenates or Methyl Tertiary Butyl Ether (MTBE) that was proven to be highly carcinogenic. The Department of Energy estimates that approximately 130,000 barrels per day of ethanol will be needed to meet the demand created by refiner decisions to replace MTBE<sup>32</sup>.



**Figure 1.4:** Nanocellulose as a biocompatible tissue scaffold. Bodin et al. 2007<sup>33</sup>.

**Table 1.1:** Representative Mechanical Strength of Materials versus Nanocellulose. ArboraNano Inc.

<b>Material</b>	<b>Density (g/cm<sup>3</sup>)</b>	<b>Tensile Strength (MPa)</b>	<b>Young's modulus (GPa)</b>	<b>Elongation at break (%)</b>
Nanocellulose	1.5	10000	150	6.7
SWCNT	1.2	30000	1054	6
MWCNT	2.6	30000	1000-1280	12.5
Carbon	1.7	4000	230-240	1.4-1.8
Kevlar 29	1.44	2800	183	4
Aramid	1.4	3000-3150	63-67	3.3-3.7
302 Stainless Steel	7.7-8.1	1280	210	
Cotton	1.5-1.6	287-800	5.5-12.6	~7-8

Newer applications for cellulose involve nanocellulose, the actual fibril element, to creating novel hydrogels<sup>34</sup>; transparent composites of cellulose have favorable characteristics can useful in many fields, such as flexible electrodes, flexible display devices<sup>35</sup>, bio-sensors, as a platform substrate to study the effect of electrical signals on cell activities, and to direct desirable cell function for tissue engineering applications (Figure 1.4). The mechanical properties of nanocellulose are comparable to Kevlar (Table 1). These properties of nanocellulose are expected to produce an industry that will be worth up to 600 billion dollars annually by 2020 (US Forrest Service, Department of Agriculture).

### 1.3.2.2 Cellulose Synthase

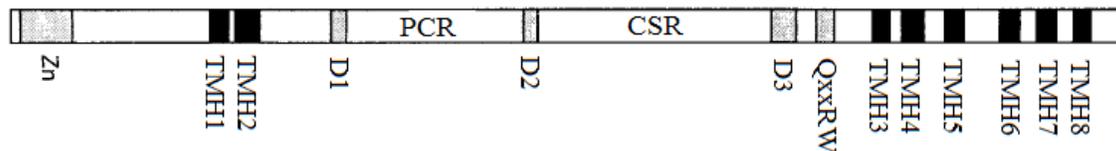
Rosettes of enzymes that synthesized cellulose fibrils were first observed by Mueller et al. arranged into rafts that aggregated into larger clusters<sup>36</sup>. Each fibril (~5nm dia)

produced by these rosettes eventually combine into larger microfibrils of ~20nm. These larger fibrils strands are thought to form entirely through self-assembly<sup>37</sup>. Current theory suggests that the crystallographic makeup and orientation of these fibrils depend entirely with the proteins that polymerize  $\beta$ -1,4 glucan into cellulose<sup>38</sup>. Examining the protein machinery itself will give some insight into the self assembly of cellulose into microfibrils. Furthermore, if these proteins can be manipulated, the cellulose that is produced can also be altered.

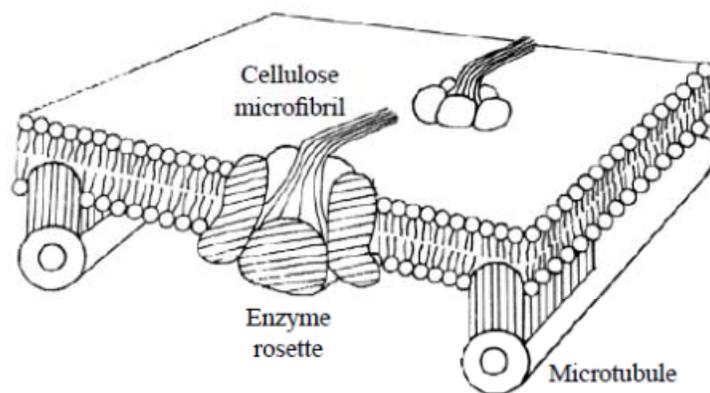
Cellulose is created at the molecular level by glycosyltransferases known as cellulose synthases, CESA's, that catalyze the polymerization of  $\beta$ -1,4 glucan chains from cellulobiose. CESAs are transmembrane proteins comprised of a large transmembrane domain of eight alpha helices and two major portions located in the cytoplasm. The monomer that is polymerized into  $\beta$ -1,4 glucan is a uridine diphosphate bound glucose (UDP-Glc) molecule of which there is debate on whether there are two or one binding sites.

Based upon a similar protein structure solved from a bacterium, *Bacillus subtilis*, it is assumed that CESA's have a Rossman fold with a beta sheet core comprised of six to seven beta strands<sup>39</sup>. This fold motif is characteristic of nucleotide binding proteins which is necessary if CESA's catalyze reactions with UDP-Glc. For all CESAs, there is a conserved amino acid motif comprised of three catalytic aspartic residues D,D,D and a glucan chain binding motif of QxxRW (Figure 1.5); the glucan chain binding motif of glutamine separated from the arginine and tryptophan by two positions<sup>40</sup>. Cellulose synthases are organized into larger protein hetero-oligomers known as rosettes. There is debate as to how many CESA's constitute one rosette, but the prevailing theory is 36<sup>41</sup>. These rosettes are first thought to be

formed in the golgi complex and exported to the cell membrane<sup>41</sup>. These rosettes self aggregate at the cellular membrane into higher level complexes and produce the cellulose microfibrils (Figure 1.6). Plant cellulose synthases are classified into six major isoforms along the lines of primary and secondary cell walls using the model plant *Arabidopsis thaliana* (mouse ear cress) which is related to cotton. Primary CESAs are those that produce a pliant wall in a plant as its cells grow. Secondary CESAs are responsible for forming the thicker walls when cells stop growing, and it is here where the majority of plant biomass is stored. With only a half life of less than 30 minutes, experimental structure determination of CESA's has been difficult<sup>42</sup>. Delivery of cellulose synthase to the plasma membrane occurs at a rate of 4.8 delivery events  $\mu\text{m}^{-2} \text{h}^{-1}$ <sup>43</sup>. The turnover rate is much faster than for most membrane proteins. This recalcitrant character of CESA proteins has stymied the field for well over 40 years.



**Figure 1.5:** General topology of cellulose synthase glycosyltransferases. PCR is the plant conserved region, an CSR is the class specific region. There are 8 transmembrane helices in total. Adapted from Roberts et al.<sup>44</sup>. D1-3 are conserved aspartic acid motifs. The QxxRW motif is thought to hold the glucan chain as it grows, and is found in all processive glycosyltransferases including chitin synthase. The zinc finger is at the N terminus and thought to allow other protein-protein interactions.



**Figure 1.6:** Cellulose microfibrils formed polymerized by rosettes (Neville 1993<sup>45</sup>).

CESA mutants arising from single amino acid substitutions have shown altered synthesis of cellulose<sup>46</sup>. Understanding how these specific mutations affect the structure of the CESA and ultimately the cellulose that is produced can now leverage recent gains in bioinformatics<sup>47</sup>. This work makes use of computational modeling of the structures from ab-initio prediction to all-atom physical simulations. Glycosyltransferases such as cellulose synthase are responsible for catalyzing the vast array of polysaccharide based biopolymers found in nature. Being able to alter their substrate affinity and enzymatic activity will allow for the creation of new biomaterials beyond those just based upon cellulose.

## 1.4 Methods

Computational experiments via molecular dynamics simulations were used to explore the behavior of nucleic acid structures with small cations, and to relax predicted protein structures. Ab-initio protein structure prediction relied on two different methods based on the complexity of the structure of interest. For the large structures beyond 150 amino acids

such as the case for the catalytic core of the cellulose synthase, a homology model from the SAM T-08 server<sup>48</sup> was employed and subsequently refined with molecular dynamics. Once the final structure was within an acceptable quality, we introduced one known mutation, Atcesa7<sup>P557T</sup> and two novel ones, Atcesa1<sup>S377F</sup> and Atcesa3<sup>G620E</sup>, into the modeled Ghcesa1 catalytic core. ROSETTA was subsequently used to examine the folding of three fragments that show relevance to cellulose synthesis and protein-protein interactions, one structure is a fragment from the transmembrane region consisting of two helices and a connecting loop, and the two others were small domains under 100 amino acids within the catalytic core itself.

### 1.4.1 Molecular Dynamics

Macroscopic observables are intrinsically related to microscopic behavior at the atomic scale. The time dependent (and independent) microscopic behavior of a molecule can be calculated. The most accurate method of examining atomic behavior is molecular dynamics which models the atomic interactions using empirical force-fields. These force-fields approximate the interactions in the system using simplified models. In practice, these models typically include only those features that are necessary to describe the system:

$$U = \sum_{\text{Bonds}} \frac{1}{2} K_b (b - b_0)^2 + \sum_{\text{Angles}} \frac{1}{2} K_\theta (\theta - \theta_0)^2 \quad (1)$$

Bonds Angles

$$\sum K_\phi [1 - \cos(n\phi + \delta)] \quad (2)$$

Torsions

$$\sum \epsilon \left[ \left( \frac{r_0}{r} \right)^{12} - 2 \left( \frac{r_0}{r} \right)^6 \right] \quad (3)$$

Non-bonded

$$\sum \frac{q_i q_j}{r} \quad (4)$$

Electrostatic

Once the system is adequately described, the next step is to solve Newton's Equations of motion.

The most common time integration method is based on the Verlet algorithm<sup>49</sup>. It is an implicit method that is a third order Taylor expansion of the positions  $r(t)$  into a forward and backward time component such that the accelerations,  $a(t)$ , and third time derivative result in a simplified expression with fourth order errors of the time step,  $\Delta t$ :

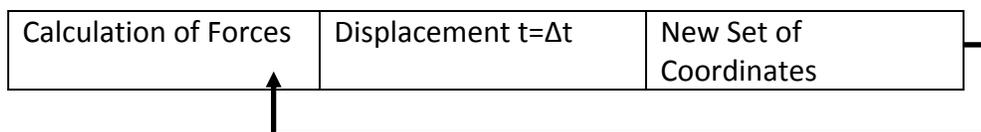
$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + a(t)\Delta t^2 + O(\Delta t^4) \quad (5)$$

Forces can be computed from the acceleration:

$$a(t) = -\left(\frac{1}{m}\right) \nabla(r(t)) \quad (6)$$

Given that the choice of time step,  $\Delta t$ , must be very small to capture the fastest motions a system, this requirement translates into chosen time steps being on the order of 1 femtosecond. This high precision makes MD the principal tool for modeling proteins, nucleic acids and other types of soft matter at atomistic detail. However, this high precision

also limits the amount of time that can be sampled to a few hundred nanoseconds depending on the numerical precision of the computer system. Very long simulations eventually accrue numerical errors as to render the later times of the simulation unrealistic. The schematic below shows the MD methodology (Figure 1.7):



**Figure 1.7** Diagram of Basic Molecular Dynamics Methodology

In this work, nucleic acid helices were generated using the Nucleic Acid Builder module of the molecular dynamics software Amber 9.0. Nucleic acid all atom simulations utilized the updated Cornell force field for nucleic acids<sup>50,51</sup>. K<sup>+</sup>, Na<sup>+</sup>, and Cl<sup>-</sup> ions are modeled as point charges with van der Waals spheres without either polarization or charge transfer effects. This methodology represents monovalent ions faithfully. However, ad hoc adaptation of the Åqvist parameters for the AMBER-99 force field led to artifacts in long simulations of biomolecules in salt solutions resulting in salt crystal formation below their solubility limit according to Chen et al.<sup>52</sup>. Subsequently, Joung and Cheatham reparameterized the Lennard-Jones potential for ions and specific rigid water models<sup>53,54</sup>. All NA structures were subjected to conjugate gradient energy minimization for 5000 steps. Minimized NA structures were then neutralized with Na<sup>+</sup> ions and immersed in a water box with at least 10 Å deep solvation shell using the TIP3P water model<sup>55</sup>. Additional Na<sup>+</sup> and Cl<sup>-</sup> ions or K<sup>+</sup> and Cl<sup>-</sup> ions were added to represent a 0.1 M effective salt concentration around a given NA helix. The equilibration of each NA sample was carried out in 11 stages

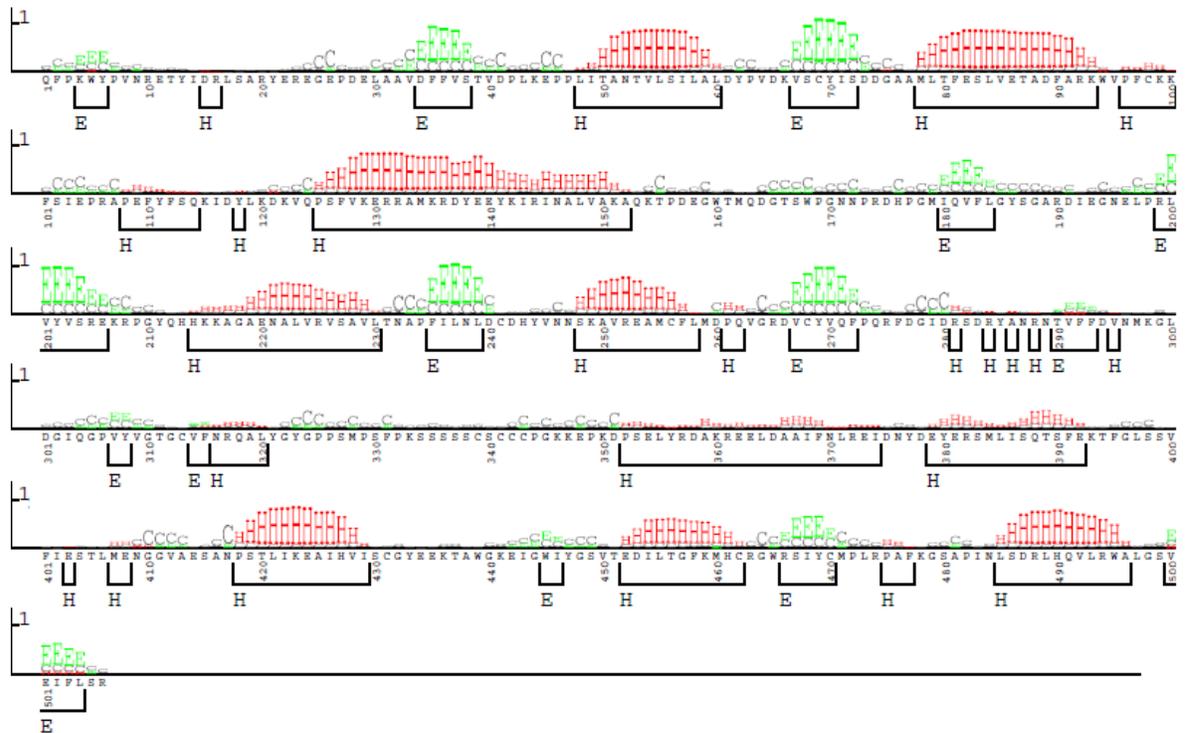
starting from the solvent minimization for 10000 steps and keeping the duplex restrained for 200 Kcal/mol. The system was heated to 300K in 40 ps while imposing a 200kcal/mol constraint on the duplex. A brief NPT MD run was performed for 200 ps with a duplex restrained maintained at 200 kcal/mol. Another constrained minimization step follows with the restraint of 25 kcal/mol for 10000 steps. A second NPT MD run was performed at 25 kcal/mol restraint for 20 ps. Subsequently four additional 1000 cycle minimization steps were performed while relaxing the positional constraint from 20 kcal/mol to 5 kcal/mol in 5 kcal/mol increments. A final unconstrained minimization stage of 1000 cycles was performed before reheating the system to 300K at constant volume within 40 ps. Subsequently, NPT equilibrations were performed to ensure uniformity in solvent density. Long range electrostatic interactions were calculated by Particle Mesh Ewald summation (PME)<sup>56</sup> and the non-bonded interactions were truncated at 9 Å cutoff along with a 0.00001 tolerance of Ewald convergence. A Berendsen thermostat maintained temperature at 300 K<sup>57</sup>. The SHAKE algorithm was used to constrain the position of hydrogen atoms<sup>58</sup>. The production simulations were performed for an NVT ensemble. Each production simulation was performed for 20 ns with a 2 fs time step.

Protein all atom molecular dynamics simulations for the modeled catalytic region of a cellulose synthase differed from the nucleic systems by having 0.3M NaCl and endured a slower initial heating step of 100ps versus 40ps to attain 300K. Each production run was executed for a minimum of 10ns. Mutants were generated from the last “good” structure of the wild type by mutating the residue of interest with the Amber tool TLEAP. Analysis of all trajectories was performed using the Amber utility, PTRAJ.

## **1.4.2 Ab-initio Protein Structure Prediction**

### **1.4.2.1 Homology Modeling**

A low resolution homology based model catalytic core region utilized the SAM-T08 server from the Karplus lab<sup>48</sup>. After submission of the protein FASTA sequence, the server performs three iterations to find homologs from the NCBI non-redundant database of protein sequences. The resulting multiple sequence alignments (MSA) are used by neural networks to predict the local structure properties. Each of the MSA's are used to provide a set of probability vectors corresponding to each residue of whether its secondary structure is a beta strand, helix, or loop. The secondary structure predictions (Figure 7) are then fed into the Sequence and Alignment Modeling software<sup>59</sup> using a template library of pre-computed Hidden Markov Models (HMM) that are used to generate the 3-D structure.

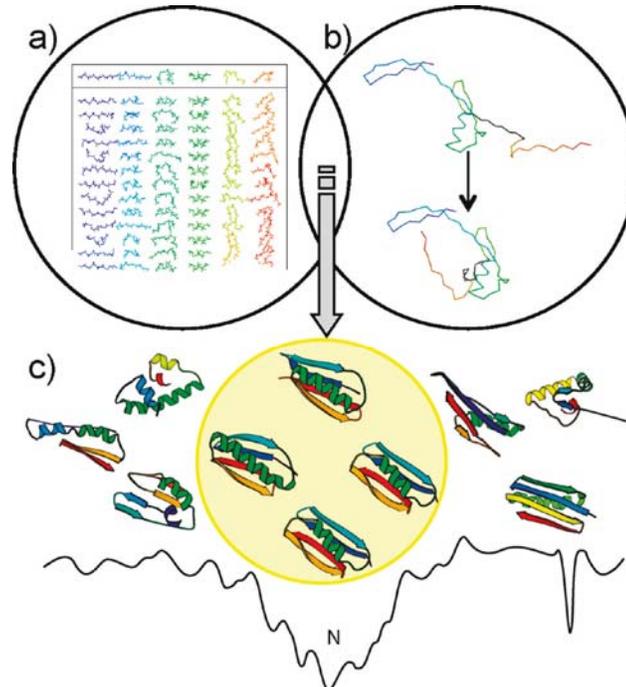


**Figure 1.8:** SAM T08 Secondary Structure Prediction where E labels correspond to beta strands and H labels are alpha helices. Everything unlabeled is a random coil.

#### 1.4.2.2 Fragment Based Assembly

Ab-initio prediction of fragments from the catalytic core and a transmembrane helices relied on the ROSETTA program. Starting with the FASTA file of single letter protein codes, a secondary structure prediction was accomplished using the PSIPRED program that performs a position specific iterated BLAST search. BLAST is the Basic Local Alignment Search Tool. The results of this search is then used by the ROSETTA script, `make_fragments.pl`, to generate the 3 and 9 amino acid fragment library prior to starting the main ROSETTA algorithm that performs ab-initio folding. ROSETTA employs Bayesian statistics to derive a structure from short fragments; these short residue fragments (< 15) have

a greater correlation between local sequence and structure than for longer sequences. Subsequently, fragment assembly occurs by a Monte Carlo procedure; the process begins with a random positioning in the fully extended protein with either a 3 or 9 residue fragment. This first fragment is chosen randomly from the top 25 structures in the libraries.



**Figure 1.9:** ROSETTA starts from (a) fragment libraries with sequence-dependent ( $\zeta$  and  $\psi$ ) angles that capture the local conformational space accessible to a sequence. (b) Combining fragments from the libraries, ROSETTA folds the protein by optimizing non-local contacts. A low-resolution energy function smoothes the energy surface, to funnel the structure toward a native conformation as denoted by “N” (c). Metropolis Monte Carlo minimization drives the structure toward the global minimum. Reproduced from Kaufman et al. 2010<sup>60</sup>.

During each iteration, torsion angles in the protein segment are replaced with ones from the new fragment. The resulting conformation’s energy is calculated using scoring functions; a few hundred of the most likely angles for the fragments are derived from X-ray resolved structures. The fragments are matched against sequences present in the NCBI non redundant database via a PSIBLAST search. The final step ranks the matches by minimal steric overlap

to construct the fragment library from the favorable torsion angles and compatibility with secondary structure prediction as determined by an algorithm like PSIPRED or JUF0<sup>61</sup>.

Prediction of membrane proteins relied on the OCTOPUS server for the membrane topology<sup>62</sup>. This topology file is used to generate a lipophilicity profile to orient individual residues during the fragment insertion method within ROSETTA's implicit membrane model.

## References

1. Mentovich ED, Livanov K, Prusty DK, Sowwan M, Richter S. DNA-nanoparticle assemblies go organic: Macroscopic polymeric materials with nanosized features. *J Nanobiotechnol* 2012;10.
2. Helwa Y, Dave N, Froidevaux R, Samadi A, Liu JW. Aptamer-Functionalized Hydrogel Microparticles for Fast Visual Detection of Mercury(II) and Adenosine. *Acs Appl Mater Inter* 2012;4(4):2228-2233.
3. Nishikawa M, Mizuno Y, Mohri K, Matsuoka N, Rattanakit S, Takahashi Y, Funabashi H, Luo D, Takakura Y. Biodegradable CpG DNA hydrogels for sustained delivery of doxorubicin and immunostimulatory signals in tumor-bearing mice. *Biomaterials* 2011;32(2):488-494.
4. Saccà B, Siebers B, Meyer R, Bayer M, Niemeyer CM. Nanolattices of Switchable DNA-Based Motors. *Small* 2012;8(19):3000-3008.
5. Mysliwiec J, Sznitko L, Sobolewska A, Bartkiewicz S, Miniewicz A. Lasing effect in a hybrid dye-doped biopolymer and photochromic polymer system. *Appl Phys Lett* 2010;96(14).
6. Kawabe Y, Wang L, Nakamura T, Ogata N. Thin-film lasers based on dye-deoxyribonucleic acid-lipid complexes. *Appl Phys Lett* 2002;81(8):1372-1374.
7. Khan F, Dahman Y. A Novel Approach for the Utilization of Biocellulose Nanofibres in Polyurethane Nanocomposites for Potential Applications in Bone Tissue Implants. *Des Monomers Polym* 2012;15(1):1-29.

8. Wijnhoven SWP, Dekkers S, Hagens WI, de Jong WH. Exposure to nanomaterials in consumer products. RIVM-letter report 340370001/2009 2009.
9. Nagai H, Toyokuni S. Differences and similarities between carbon nanotubes and asbestos fibers during mesothelia carcinogenesis: Shedding light on fiber entry mechanism. *Cancer Science* 2012;103(8):1378-1390.
10. Hitoshi K, Katoh M, Suzuki T, Ando Y, Nadai M. Differential effects of single-walled carbon nanotubes on cell viability of human lung and pharynx carcinoma cell lines. *J Toxicol Sci* 2011;36(3):379-387.
11. Hubbe MA, Rojas OJ, Lucia LA, Sain M. Cellulosic Nanocomposites: A Review. *Bioresources* 2008;3(3):929-980.
12. Brown RM, Saxena I. Cellulose Synthesizing Complexes in Vascular Plants and Prokaryotes. DOE-ER15396 2009.
13. SantaLucia J, Hicks D. The thermodynamics of DNA structural motifs. *Annu Rev Bioph Biom* 2004;33:415-440.
14. Hermann T, Westhof E. Non-Watson-Crick base pairs in RNA-protein recognition. *Chem Biol* 1999;6(12):R335-R343.
15. Hermann T, Patel DJ. RNA bulges as architectural and recognition motifs. *Struct Fold Des* 2000;8(3):R47-R54.
16. McCann MD, Lim GFS, Manni ML, Estes J, Klapac KA, Frattini GD, Knarr RJ, Gratton JL, Serra MJ. Non-nearest-neighbor dependence of the stability for RNA group II single-nucleotide bulge loops. *Rna* 2011;17(1):108-119.
17. Sponer J, Leszczynski J, Hobza P. Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers* 2001;61(1):3-31.
18. Shkilnyj P, Koudelka GB. Effect of salt shock on stability of lambda(imm434) lysogens. *J Bacteriol* 2007;189(8):3115-3123.
19. Bushman FD, Ptashne M. Activation of Transcription by the Bacteriophage-434 Repressor. *P Natl Acad Sci USA* 1986;83(24):9353-9357.
20. Mauro SA, Koudelka GB. Monovalent cations regulate DNA sequence recognition by 434 repressor. *J Mol Biol* 2004;340(3):445-457.
21. Stellwagen E, Muse JM, Stellwagen NC. Monovalent Cation Size and DNA Conformational Stability. *Biochemistry-U S* 2011;50(15):3084-3094.

22. Casiano-Negrone A, Sun XY, Al-Hashimi HM. Probing Na<sup>+</sup>-Induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: New insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry-US* 2007;46(22):6525-6535.
23. Leulliot N, Varani G. Current topics in RNA-protein recognition: Control of specificity and biological function through induced fit and conformational capture. *Biochemistry-US* 2001;40(27):7947-7956.
24. Wang H, Yang RH, Yang L, Tan WH. Nucleic Acid Conjugated Nanomaterials for Enhanced Molecular Recognition. *ACS Nano* 2009;3(9):2451-2460.
25. Han SP, Maune HT, Barish RD, Bockrath M, Goddard WA. DNA-Linker-Induced Surface Assembly of Ultra Dense Parallel Single Walled Carbon Nanotube Arrays. *Nano Lett* 2012;12(3):1129-1135.
26. Somerville C. Cellulose synthesis in higher plants. *Annu Rev Cell Dev Bi* 2006;22:53-78.
27. Nishino T, Takano K, Nakamae K. Elastic-Modulus of the Crystalline Regions of Cellulose Polymorphs. *J Polym Sci Pol Phys* 1995;33(11):1647-1651.
28. Harris D, DeBolt S. Relative Crystallinity of Plant Biomass: Studies on Assembly, Adaptation and Acclimation. *Plos One* 2008;3(8).
29. Gindl W, Keckes J. All-cellulose nanocomposite. *Polymer* 2005;46(23):10221-10225.
30. Eichhorn SJ, Baillie CA, Zafeiropoulos N, Mwaikambo LY, Ansell MP, Dufresne A, Entwistle KM, Herrera-Franco PJ, Escamilla GC, Groom L, Hughes M, Hill C, Rials TG, Wild PM. Review: Current international research into cellulosic fibres and composites. *J Mater Sci* 2001;36(9):2107-2131.
31. Zimmermann T, Pohler E, Geiger T. Cellulose fibrils for polymer reinforcement. *Adv Eng Mater* 2004;6(9):754-761.
32. EIA. Renewable Fuels Legislation Impact Analysis. 2005.
33. Bodin A, Backdahl H, Risberg B, Gatenholm P. CELL 107-Nanocellulose as scaffolds for tissue engineering and organ regeneration. *Abstr Pap Am Chem S* 2007;233:701-701.
34. Cha RT, He ZB, Ni YH. Preparation and characterization of thermal/pH-sensitive hydrogel from carboxylated nanocrystalline cellulose. *Carbohydr Polym* 2012;88(2):713-718.

35. Okahisa Y, Yoshida A, Miyaguchi S, Yano H. Optically transparent wood-cellulose nanocomposite as a base substrate for flexible organic light-emitting diode displays. *Compos Sci Technol* 2009;69(11-12):1958-1961.
36. Mueller SC, Brown RM. Evidence for an Intramembrane Component Associated with a Cellulose Microfibril-Synthesizing Complex in Higher-Plants. *J Cell Biol* 1980;84(2):315-326.
37. Vincent JFV. From cellulose to cell. *J Exp Biol* 1999;202(23):3263-3268.
38. Ruel K, Nishiyama Y, Joseleau JP. Crystalline and amorphous cellulose in the secondary walls of Arabidopsis. *Plant Sci* 2012;193:48-61.
39. Egelund J, Skjot M, Geshi N, Ulvskov P, Petersen BL. A complementary bioinformatics approach to identify potential plant cell wall glycosyltransferase-encoding genes. *Plant Physiol* 2004;136(1):2609-2620.
40. Endler A, Persson S. Cellulose Synthases and Synthesis in Arabidopsis. *Mol Plant* 2011;4(2):199-211.
41. Mutwil M, Debolt S, Persson S. Cellulose synthesis: a complex complex. *Curr Opin Plant Biol* 2008;11(3):252-257.
42. Jacob-Wilk D, Kurek I, Hogan P, Delmer DP. The cotton fiber zinc-binding domain of cellulose synthase A1 from *Gossypium hirsutum* displays rapid turnover in vitro and in vivo. *P Natl Acad Sci USA* 2006;103(32):12191-12196.
43. Wightman R, Turner S. Trafficking of the Plant Cellulose Synthase Complex. *Plant Physiol* 2010;153(2):427-432.
44. Roberts AW, Roberts EM, Delmer DP. Cellulose synthase (CesA) genes in the green alga *Mesotaenium caldariorum*. *Eukaryot Cell* 2002;1(6):847-855.
45. Neville AC. *Biology of Fibrous Composites; Development Beyond the Cell Membrane*. Cambridge: University Press; 1993.
46. Harris D, Stork J, Debolt S. Genetic modification in cellulose-synthase reduces crystallinity and improves biochemical conversion to fermentable sugar. *Gcb Bioenergy* 2009;1(1):51-61.
47. Carroll A, Specht CD. Understanding Plant Cellulose Synthases Through a Comprehensive Investigation of the Cellulose Synthase Family Sequences. *Frontiers in Plant Science* 2011;2(5).
48. Karplus K. SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res* 2009;37:W492-W497.

49. Verlet L. Computer Experiments on Classical Fluids .I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys Rev* 1967;159(1):98-&.
50. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J Am Chem Soc* 1995;117(19):5179-5197.
51. Cieplak P, Caldwell J, Kollman P. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: Aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J Comput Chem* 2001;22(10):1048-1057.
52. Chen AA, Pappu RV. Parameters of monovalent ions in the AMBER-99 forcefield: Assessment of inaccuracies and proposed improvements. *Journal of Physical Chemistry B* 2007;111(41):11884-11887.
53. Joung IS, Cheatham TE. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *Journal of Physical Chemistry B* 2008;112(30):9020-9041.
54. Joung IS, Cheatham TE. Molecular Dynamics Simulations of the Dynamic and Energetic Properties of Alkali and Halide Ions Using Water-Model-Specific Ion Parameters. *Journal of Physical Chemistry B* 2009;113(40):13279-13290.
55. William L. Jorgensen JC, and Jeffry D. Madura. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;79.
56. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* 1995;103(19):8577-8593.
57. Berendsen HJC, Postma JPM, Gunsteren WFv, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* 1984;81(8):3684-3690.
58. Ryckaert JP, Ciccotti, G, Berendsen, H.J.C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* 1977;23(3):327-341.
59. Hughey R, Karplus K, Krogh A. SAM: sequence alignment and modeling software system, version 3. Santa Cruz: UC Santra Cruz; 1999.

60. Kaufmann KW, Lemmon GH, DeLuca SL, Sheehan JH, Meiler J. Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You. *Biochemistry-US* 2010;49(14):2987-2998.
61. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. *P Natl Acad Sci USA* 2003;100(21):12105-12110.
62. Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 2008;24(15):1662-1668.

# Chapter 2

## **The Sequence of HIV-1 TAR RNA Helix Controls Cationic Distribution**

Published in the Journal of Physical Chemistry B (2010)

### **The Sequence of HIV-1 TAR RNA Helix Controls Cationic Distribution**

Latsavongsakda Sethaphong, Abhishek Singh, Ashley E. Marlowe, Yaroslava G. Yingling\*

Department of Materials Science and Engineering, NNC State University, Raleigh, NC

#### **Abstract**

Sequence dependency of metal ion aggregation around RNA structures is known to be involved in critical functions ranging from processes of molecular recognition to enzymatic chemistry. Ion interactions with an HIV-1 TAR RNA core helix were examined with explicit solvent molecular dynamics simulations. The results have shown that there is a sequence-dependent cationic localization toward the purine-rich run within the TAR helix and other purine-rich duplexes. The behavior is independent of ionic species or a presence of a bulge. A region of high ion affinity agrees very well with the position of the X-ray determined divalent cations within a fragment from the HIV-1 TAR RNA.

KEYWORDS. RNA helix, cations, HIV-1 TAR RNA, molecular dynamics simulations.

#### **2.1 Introduction**

RNA is a highly electronegative nucleic acid biopolymer that participates in the storage, expression and control of genetic information along with various other biochemical reactions. In order to perform its evolutionary developed functions, RNA must overcome the inherent self-repulsive forces stemming from the electronegative landscape created by its nucleic acid bases. Therefore, RNA must rely on counterion interactions in various roles in order to fold into tertiary structures necessary for function and molecular recognition<sup>1-3</sup>. The

charged electronegative surface potential of the RNA molecule is sequence dependent wherein even a single base substitution can fundamentally alter the reactivity of the entire complex<sup>4</sup>. Even though the role of cationic interactions had been appreciated early on, characterizing and quantifying their influence on nucleic acid structure and function has only been a recently renewed effort<sup>5,6</sup>.

Interactions of nucleic acids with cations are complex processes involving charge compensation, hydration free energy, coordination geometry and coordinate bond forming capacity<sup>7,8</sup>. Previous observations have been made about the role of specific nucleotide sequence stretches in their affinity for monovalent and divalent cations<sup>9</sup>. The culminated evidence from many years of various theoretical and experimental studies have hinted at a direct interplay of base sequence effects and metal ion sequestration around purine rich regions of nucleic acids<sup>10-14</sup>. From the stabilizing effect of metal binding to the purine N7 site and prevalence of purine dependent biological function, sequence dependent organization of the nonspecific cationic binding is compelling<sup>15,16</sup>. Crystallographic studies of metal binding in HIV-1 RNA duplexes by 13 different cationic species by Ennifar et al. concluded that divalent species such as  $Mg^{2+}$  often prefer the Hoogsteen sites of guanine residues, specifically N7 and O6<sup>17</sup>. Interweaving these sets of observations may elucidate some mechanisms behind phenomena in the structural studies of viral Polypurine Tracts<sup>18,19</sup>. Yet, the problem remains of concretely connecting nucleic acid sequence order with identifiable and predictable ion generalized dynamics. The behavior of monovalent cations is immensely difficult to elucidate experimentally since monovalent ions interact weakly with nucleic acids via principally electrostatic attraction. NMR studies have been applied to cationic effects on

nucleic acid structure, but the results often lead to more ambiguous questions<sup>20</sup>. The core problem is that diffusive binding is a relatively rare event compared to the NMR correlation times 20-40 ns<sup>21</sup>. Our own computational experiments have shown that an ion can bind to phosphates on the order of hundreds of picoseconds. Even if the dynamics of cationic binding to nucleic acid tracts can only be inferred, it has not extinguished experimental pursuit of a more substantive understanding of specific ion interactions<sup>22</sup>.

Computational models have been successful in characterizing and reaching close agreement with experimental studies on cation and RNA interactions<sup>23-25</sup>. Molecular dynamics simulations (MD) can observe the localization, diffusive motions, and structural effects of the counterions in atomic details. For example, MD studies we used provide a detailed information for preferred ion binding sites, such as Na<sup>+</sup> localization at the ApT step, i.e. the position between an adenine base and a thymine base in the 5' to 3' direction, shows a region of unique high electronegative potential within the DNA minor groove<sup>26</sup>. Pronounced polymorphism of nucleic acids was observed under varying ionic strengths where X-ray crystal of a benchmark decamer d(CCAACGTTGG)<sub>2</sub> duplex showed A-form to B-form transitions under Amber force fields upon MD convergence<sup>27</sup>. However, the reliability of MD simulations depends on accurate and representative force fields for both nucleic acids and solvent.

Despite the experimental difficulty in studying monovalent cationic interactions with RNA, many have investigated this integral relationship toward RNA stability, folding, and enabling self interactions<sup>28,29</sup>. The major contributing factor that defines this relationship is the electrostatic potential surface that surrounds RNA. This potential surface arises from a

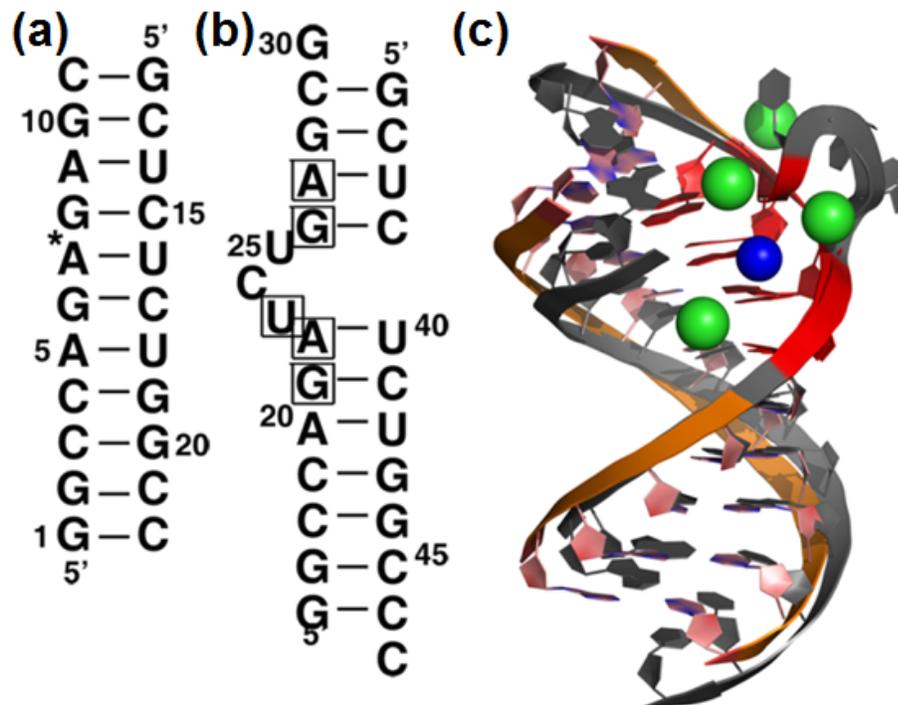
sequence dependency that is strongly associated with highly electronegative bases and large dipole moments<sup>30</sup>. The phenomena can be seen in crystal structures of helical DNA where monovalent cations are observed to cluster in the major groove near G:C pairs and the minor groove of A:T pairs<sup>9</sup>. However, the more pronounced dynamic behavior involve cations with intricate RNA tertiary structures comprised of loops, helices, bulges, and kinks where unpaired bases are exposed to solvent.

Many studies have focused on metal ion binding to DNA and RNA helices since this structural motif is the most prevalent. Much of the different dynamics of cation binding between DNA and RNA are due to preferred helical conformations<sup>31</sup>. The 2'-OH of the RNA sugar backbone promotes a rigid and open A-form helix as opposed to DNA which adopts the more compact B-form helix. Therefore charge screening is more efficient with ions more closely bound in RNA than DNA. RNA's deeper major groove allows ions to penetrate further near the central helical axis. In the presence of a divalent cation such as  $Mg^{2+}$ , RNA can form more compacted structures<sup>32</sup>. For RNA, a lower bulk ionic strength of  $Mg^{2+}$  is required to induce neutralization when compared to DNA. These slight structural differences of the sugar backbone are what make RNA cation interactions more dynamic compared to those for DNA.

Helices often have unpaired nucleotides that protrude into bulges. Cationic binding to these areas is less diffusive contrasting with the intact helices which has been shown to be quite different even for similar structures<sup>17</sup>. In retroviruses such as HIV, a transactivation responsive element (TAR) forms a loop-loop complex that is critical for RNA dimerization to subsequently increase viral replication a hundred fold. The TAR RNA element is located

at the 5'-end of all HIV-1 mRNAs and controls the trans-activation of a viral transcription via interactions with the viral transactivator protein (Tat)<sup>33</sup>. Specifically, Tat protein binds in the region of a three-base bulge and recognizes the adjacent helical sequence (Figure 2.1). Consequently, the TAR RNA element is a primary target for developing anti-HIV therapeutics<sup>34,35</sup>.

Moreover, metal ions are required for nucleocapsid protein-transactivation response TAR RNA interactions and putatively assist in Tat recognition of TAR *in vivo* due to the large conformational changes associated with cationic binding to the bulge structure. There have been several studies devoted to the understanding of the interactions of TAR RNA with metal ions<sup>36-41</sup>. Experimental studies suggest that interactions with diffusible counterions drive the TAR's conformational transitions, however, sodium and magnesium ions may associate with TAR in distinct modes<sup>41</sup>.



**Figure 2.1:** (a) Sequence of the helical stem used in this study with a star denoting the location of the removed bulge. (b) Sequence of TAR RNA duplex (PDB id 397D) with boxed nucleotides important for Tat binding. (c) Superimposition of TAR RNA crystal structure (grey) with four bonded calcium ions (yellow) and a snapshot from MD simulations of a TAR-like RNA helix (orange) and the highest occupancy sodium ion (green) and potassium ion (blue).

Investigations of screening cationic interactions which have utilized the HIV-transactivation response element (Tar) have focused on modeling cationic dependent loop-loop complexation. A rationally designed Tar complementary loop sequence, Tar\*, has also been studied extensively to reveal that cation binding stabilization is governed by chloride ion exclusion in an HIV Tar-Tar\* complex when subjected to alkali-chloride salts, ( $\text{CsCl} < \text{KCl} < \text{NaCl}$ )<sup>42</sup>. Likewise, the 3-nucleotide bulge structure of HIV-TAR also displays unusual cationic dependent behaviors. Moreover, bulge structures are well known metal cation binding sites<sup>37</sup>. They have garnered additional scrutiny for study since they are catalytically

active regions, vital for protein-RNA recognition, intra-RNA interactions, and are potential drug binding sites<sup>43</sup>. For HIV-TAR complexation, the complete mechanism of Tat recognition to the 3-nt bulge remains unclear since site specificity may rely on structural as well as sequence complementarity which ultimately drive cation dependent effects. The bulge region of the TAR RNA binds divalent cations in a weak, in the millimolar range, but highly specific manner (Figure 2.1)<sup>37,38</sup>. Buck et al. observed that this bulge region exhibited binding to dinuclear ruthenium(II) complexes was much stronger than previously reported. They attributed their observations to a possible cleft structure of the minor groove. The bulge sequence itself does affect the TAR RNA interaction with various divalent cations as shown by Carter-O'Connell et al. Although their study examined pyrimidine substitutions, the effect of changing uracil to cytosine depended on the position such that the altered divalent cationic interaction was not uniform among  $Mg^{2+}$  and  $Ca^{2+}$  species; specifically, divalent interactions between U23 and C24 are quite different and quantifiable. Hence, the caution is that base modification experiments regarding non-helical and dynamic regions as identifiable metal ion binding sites are not straightforward<sup>40</sup>. Given this admonition, deducing the interplay of specific base sequence and cationic localization would require that we remove the effect of the bulge entirely before introducing single base mutations.

Another factor of concern is the influence of cationic species and their impact on the structural dynamics of RNA. Parameters such as charge density, diffusivity, and hydration shells of metal ions have been shown to affect stability<sup>28 44</sup>. The folding landscape induced by a monovalent cation versus a divalent cation do follow different kinetics.  $Na^+$  versus  $Mg^{2+}$  initiated folding using the Tetrahymena ribozyme as a model, revealed that  $Mg^{2+}$

induction led to fewer more stable structural intermediates for mixed conditions<sup>45</sup>. Although ion valency affects different steps of the folding process, folding in the presence of monovalent ions alone was faster and direct. In general, Na<sup>+</sup> induced folding follows cooperative binding as in the case of an RNA kink motif<sup>46</sup>. Despite their differing kinetics, there is close agreement experimentally on the structural binding sites of monovalent ions which often occupy divalent cation-binding regions<sup>2</sup>. Some specific metal-ion interactions are attributable to coordination geometry and may explain why Na<sup>+</sup> stabilize tertiary contacts while Mg<sup>2+</sup> is preferentially bound at catalytic sites of certain ribozymes<sup>47</sup>. Underlying this relationship between cations and tertiary structures is the base sequence.

In this work, we examine the influence of base sequence arrangement on cationic distributions of sodium and potassium ions in helical conformations of RNA via molecular dynamic simulations. We probed the sequence dependent RNA-ion interactions on a fragment of human immunodeficiency type 1 virus (HIV-1) TAR core duplex and compared it to experiments of two mutated fragments without the bulge, the fragment with the bulged; polyG/polyC and polyA/polyU helical duplexes were used as controls (Table 2.1).

**Table 2.1:** Simulations of TAR and TAR-like Sequences Performed under Solvent Conditions as Indicated

<b>Simulations Performed</b>		
<b>Bulged-TAR Helix</b>	<b>TAR Helix</b>	<b>Helix Controls</b>
0.1M NaCl	0.1M NaCl	TAR A7G, 0.1M NaCl
0.1M NaCl + 4Mg	0.1M NaCl + 1Mg	TAR A9G, 0.1M NaCl
	0.1M KCl	Poly A-U, 0.1M NaCl
	0.1M NaCl + KCl	Poly G-C, 0.1M NaCl

## 2.2 Methods

The coordinates of TAR-like duplex were established by removing three bulged residues from (PDB id 397d). The starting coordinates of 11bp polyG-polyC and polyA-polyU RNA helices were built using Nucleic Acid Builder software.

All simulation runs were performed using the ff99 Cornell force field for RNA<sup>48</sup> which is a reliable force field for nucleic acids, and the molecular dynamics software Amber 9.0<sup>49</sup>. Simulations were performed in sodium chloride and potassium chloride solutions. Divalent cations were not modeled as they are not well characterized by pair additive force fields and possess very slow diffusion constants incomparable with the possible simulation time length<sup>50,51</sup>.

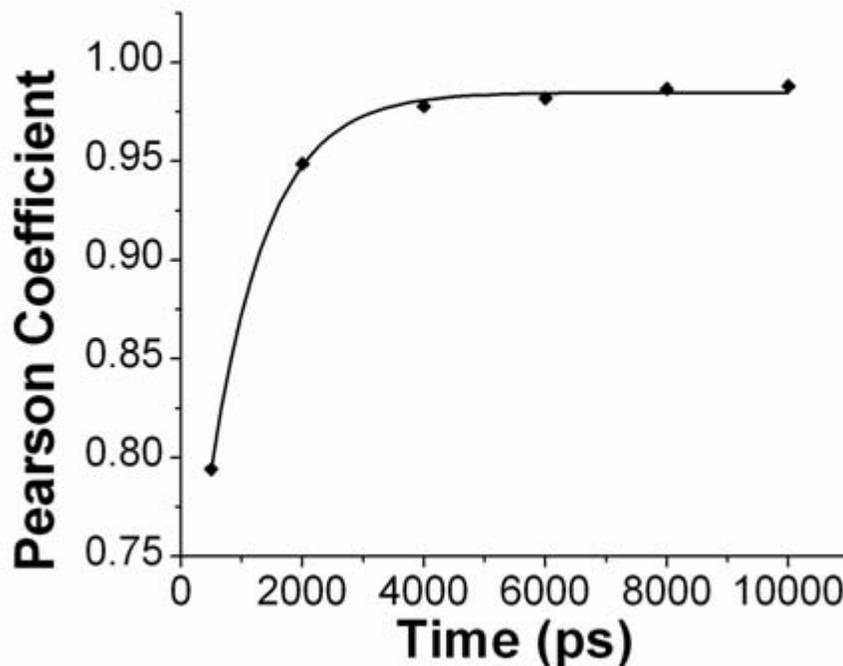
RNA polyA/polyU and polyG/polyC duplexes simulations were performed in 0.1M of sodium chloride solvent. Simulations of RNA TAR-like helices were performed in sodium chloride, potassium chloride and mixed sodium and potassium chloride solutions at 0.1M concentration. Each RNA structure was first subjected to conjugate gradient energy minimization for 5000 steps, then neutralized with 20 Na<sup>+</sup> or K<sup>+</sup> ions and immersed in a water box with 10 Å thick solvation shell using TIP3P model for water. Additional Na<sup>+</sup> and Cl<sup>-</sup> ions or K<sup>+</sup> and Cl<sup>-</sup> ions were added to represent a 0.1 M relative salt concentration. The long range electrostatic interactions were calculated by Particle Mesh Ewald summation (PME)<sup>52</sup> and the non-bonded interactions were truncated at 9Å cutoff along with a 0.00001 tolerance of Ewald convergence. The explicit solvent method<sup>53</sup> used in our simulations produce an accurate modeling of solvation effects and provide important information on the direct interactions of water and ions with nucleic acids. The system equilibration protocol

was previously described in <sup>54</sup>. Briefly, the system was minimized by constraining the solute, heated to 300 K, constraining the RNA then the solvent, and finally equilibrated by gradual release of constraints. SHAKE was applied to all hydrogen bonds in the system. Pressure was maintained at 1.0 Pa using the Berendsen algorithm<sup>55</sup>, and a periodic boundary conditions in all directions were imposed. A production simulation was performed for 10 ns with a 2 fs time step. Molecular dynamics trajectories were processed using in house PERL scripts along with the standard tool suite accompanying Amber9.0.

MD runs need to be extended until the system converges reasonably toward a lower global energy minima. One way to estimate convergence in our samples was to observe a high correlation of ion occupancy around structurally similar constructs. Pearson rank correlation coefficient (*r*) is a bivariate analysis to measure the strength of association between two variables; mathematically, *r* is evaluated using expression 1. In our calculations, two variables are the ionic occupancy measured on two symmetric parts of the RNA molecule. A high correlation coefficient indicates that the parameterized force field is suitable to converge the system dynamics. Convergence is plotted in Figure 2.2 as a trajectory time evolution of the Pearson rank coefficient. Ion occupancy was calculated on an atomic basis using hbond analysis of the ptraj module (AmberTools). Ionic occupancy calculation refers to the fraction of trajectory time the chosen ion was within 5 Å from either the oxygen or nitrogen atoms of guanine and cytosine.

$$r = \frac{\sum XY - (\sum X \sum Y)/N}{\sqrt{(\sum X^2 - (\sum X)^2/N)(\sum Y^2 - (\sum Y)^2/N)}} \dots\dots\dots (1)$$

The PolyGC system showed the rank coefficient converged to 0.98 at the end of 10 ns. Earlier studies showed a rank coefficient of 0.69 by 14 ns for palindromic DNA<sup>11</sup>. Even though our simulations indicated a convergence at 10ns, we decided to extend the simulation up to 20 ns due to statistical considerations.



**Figure 2.2:** Pearson correlation coefficient calculated for ionic occupancies by an atom between two symmetric halves of a polyG/polyC helix as a function of time. Diamonds indicate MD results from 0 to 10ns. The solid line indicates an exponential fit of the data.

*Spatial region analysis.* Ion positional associations were mapped by constructing a cylindrical space for analysis relative to backbone phosphates. Backbone phosphates were paired to create connecting lines whose medians were chosen for their relative orientations along the nucleic acid helix such that an imaginary ray passing through the medians traversed a path approximately parallel to the helical direction. The point midway between the

medians was chosen as the center. Ions extending beyond 8 angstroms of the constructed cylindrical axis were discounted from analysis as well as ions extending from +/- 25 Angstroms beyond the midpoint of the cylindrical volume. The vector direction perpendicular to the cylindrical axis and in the direction of the first phosphate atom served as an orienting origin for the theta angular determination. Theta is measured in a counter clockwise manner relative to the orienting origin. Following construction of this spatial region for analysis, pairwise distance matrix calculations between the ion and each atom of the RNA within 5 Å was performed. This spatial calculation was performed for all generated snapshots derived from the trajectory data.

The surface map of the cationic occupancy plots were generated by binning ion counts that were within a wedge shaped bin that is calculated within 8 Å from the cylinder axis in 5 degree increments and a step of 0.5 Å along the axis. The cationic occupancy was calculated at the nearest proximal residue for each ion of no greater than 5 Å. Performing this calculation generated 7200 bins which were then mapped with TECPLOT into the surfaces of the respective figures.

### **2.3 Results and Discussions**

The critical residues for Tat binding in a TAR-sequence are localized around three-base bulge (Figure 2.1). Within TAR core helix this crucial sequence is a short alternating poly-purine run (GA)<sub>n</sub>. Other features of interest that are manifested in the TAR helical sequence are the alternating complement r(UC)<sub>n</sub> or d(TC)<sub>n</sub>; alternating pyrimidine-purine tract (CpG)<sub>n</sub> and a homo-purine doublet (GG). Tracts of d(GA) are more often associated with DNA synthesis arrest along with d(TC); dysregulation in these regions of the genetic code can

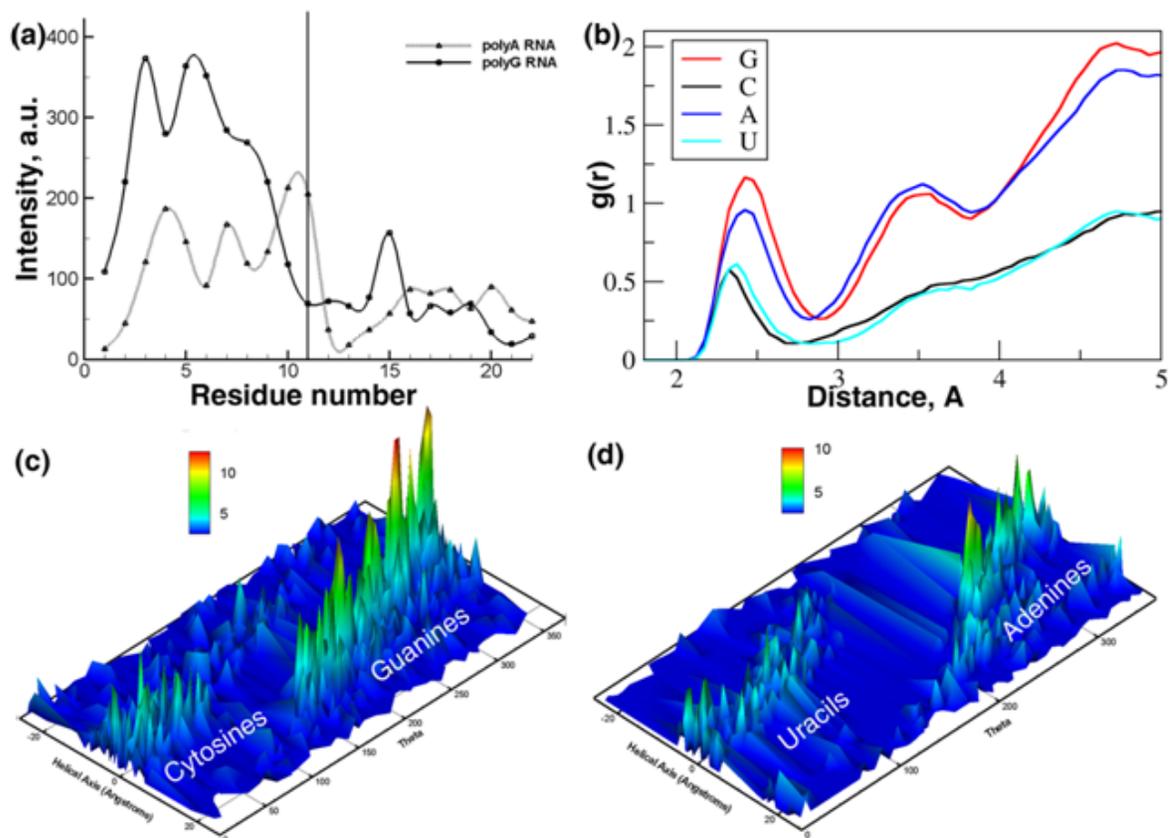
account for trait overexpression and impaired gene silencing<sup>56,57</sup>. Doublet stacking repeats of (CpG) are typically associated with severe human diseases and susceptibilities e.g cancer and systemic lupus<sup>58,59</sup>. The only recent study combining experimental and *ab initio* calculations of stacking effects concluded that GG/CC dinucleotide stacking was the least stable while the interlaced GC/CG form<sup>60</sup> is the most stable. The significance of (GG) doublets also arises from long range DNA oxidative damage mostly likely a result of the longitudinal polarizability of DNA and effects of base pairing<sup>61</sup>. It is highly likely that these two phenomena are interrelated. Thus, an examination into the role of cationic interactions may shed light onto the basic mechanisms that give rise to the thermodynamic stabilities seen and the specific contributions from ion dynamics in and around the helical duplex.

TAR RNA motif has been shown to bind four calcium ions around three-base bulge<sup>38</sup> (Figure 2.1). The three calcium ions arranged in a distorted pentagonal bipyramid configuration are stabilizing interactions via outer shell ligand interactions with the sugar backbone. Another calcium ion makes no coordination with TAR-RNA and appears to be only a stabilizing interaction similar to group I introns with A-rich 5-nt bulges<sup>62</sup>. Binding of ions also led to straightening of a TAR bulged duplex<sup>63</sup>. Molecular dynamics simulations show that the TAR core helical stem overlays nicely with the crystal structure of TAR bulged RNA (Figure 2.1). Also MD simulations indicate that the strongest binding of monovalent cations to TAR core helix is located within close proximity of a divalent ion position (Figure 2.1).

To examine the role of purine-rich runs on sequestering ions we performed a set of control simulations of polyG/polyC and polyA/polyU duplexes in 0.1 M of NaCl solvent. A

comparison of the polyA and polyG duplexes (Figure 2.3a) shows that there is a strong sequestration of cations toward the purine side of the respective duplex tract. To determine the correlation between the cations and specific nucleotides we calculated radial distribution function (RDF) averaged over second to tenth nucleotide on each strand (Figure 2.3b). RDF indicates that guanines have the highest propensity for sodium ion binding, followed by adenines, and both pyrimidines. Sodium ions tend to associate with non-anionic oxygens as inner shell ligands<sup>64</sup>. For guanine, O6 atom is the most probable cationic binding site, followed by N7 atom with a highly favorable electrostatic potential. Guanine itself has a very large dipole moment ( $\sim 7$  Debye) to sequester any cation toward it<sup>65</sup>. Adenine has three possible metal binding sites consisting of nitrogen atoms, N1, N3, and N7. The N7 atom is preferred in double stranded nucleic acid helices since the highest affinity site N1 participates in Watson Crick base pairing. Only when the N1 and N7 sites are blocked will there be a possibility of cationic binding to N3 atom. Overall, metal ion interacts strongly with the N7 atom of purines through water molecules in their coordination shell<sup>66,67</sup>. Metal ions can possibly interact with N3 and O2 atom of cytosine and uracil in the minor groove of nucleic acids, however, thermodynamic stability of both binding coordinates is low<sup>68,69</sup>. Also cation association with guanine in the major groove is stabilizing; DNA duplexes experience up to a 30% increase due to hydrated divalent cation binding to (G)N7<sup>70,71</sup>. Electronic structure calculations observed that the stability of GC Watson-Crick base pair was enhanced by 20-30% due to the coordination of the hydrated cation<sup>72</sup>. Accumulations of the partial charges in purine-runs will lead to the observation of the highest propensity to metal ion interactions somewhere in the middle of the purine stem. On Figure 2.2b,c the surface maps of ionic

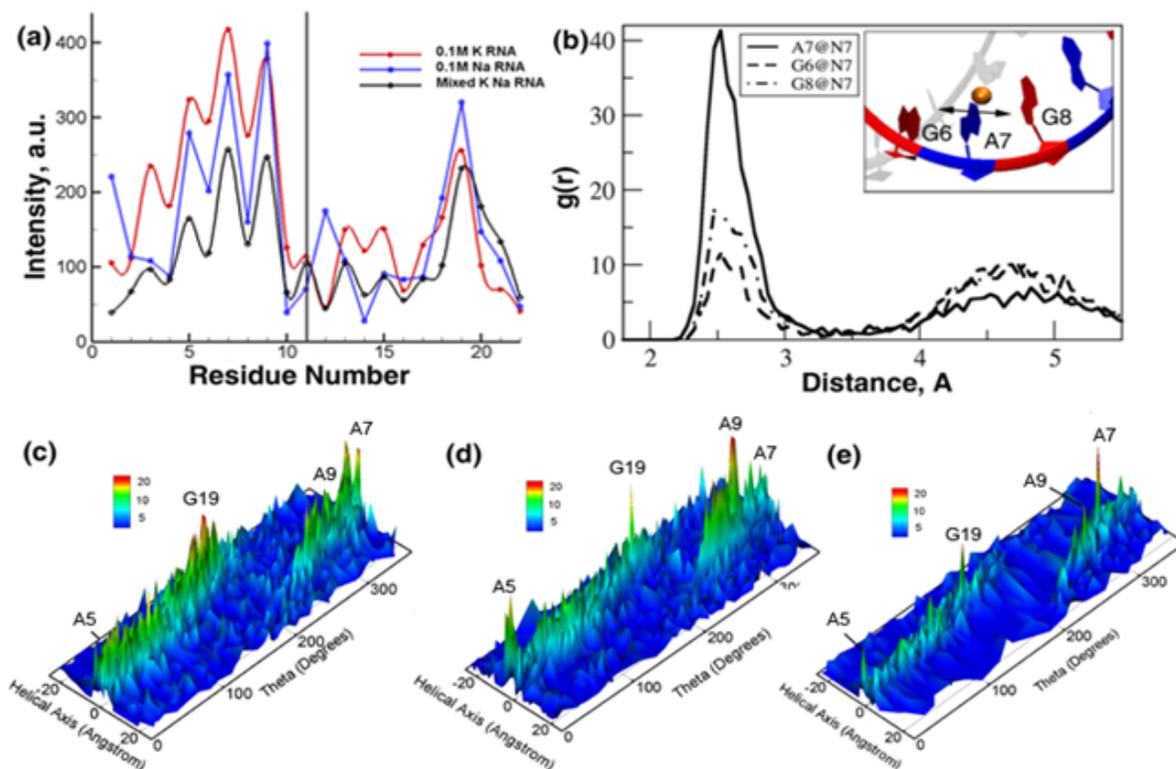
occupancy along the helical axis versus rotational angle  $\theta$  contrast ionic preference in detail. Figure 2.3b,c show that the ions interact strongly with purine residues indicated by the sharp peaks located at approximately the same distance from the center of the helix. More diffusive motion represented by smaller scattered peaks is observed for ion association with pyrimidines.



**Figure 2.3:** Sodium ion interactions with polyA/polyU duplexes and polyG/polyC duplexes. (a) Cationic association bins per residue for polyA/polyU and polyG/polyC duplexes. Vertical line denotes the strand break. (b) RDFs of sodium ions with guanine, adenine, cytosine, and uracil. (c,d) Surface map of the sodium occupancy for (c) polyG/polyC and (d) polyA/polyU duplexes calculated for rotational angle  $\theta$  around the helical axis.

To determine the solvent-dependent interactions of monovalent cations with the sequence of interest, TAR RNA core helix, molecular dynamics simulations were performed in three

different solutions; 0.1M of NaCl, 0.1 M of KCl, and 0.1 M of mixed NaCl and KCl. On Figure 2.4a cationic association is plotted as a function of residue. Figure 2.4a shows that the local maximum of the cationic intensity is located within AGAG-run. AGAG sequence in TAR core helix is critical for Tat protein recognition. This trend is common to all examined monovalent ions. In simulations with 0.1M KCl, potassium ions seemed to form the most favorable association. However, in mixed solvent sodium ions out-compete potassium ions (Figure 2.4a). Interestingly, adenine in the position 7 has the highest propensity to bind cations as shown in Figure 2.4b. A7 is flanked by two guanines in the middle of a six-purine stretch. The base stacking effect for AG stretches are expected to have a net dipole toward the purine stem. When stacked, the directionality of the guanine dipole is oblique to the adenine dipole; the weaker net dipole direction could be shifted toward N7. This effect is explained by our observation of the oscillating motion of a single ion between two neighboring guanines, G6 and G8 (insert in Figure 2.4b). Experimental observation of cationic binding to r(GCGUUUGAAACGC) RNA duplex indicated a strong preference of ion association with GAAA/UUU sequences<sup>66</sup>. However, experiment suggested that a specific site for sodium ions located at the UU side of the sequence. Our simulations indicate the opposite site for ionic association in both polyA/polyU duplex and TAR core helix.



**Figure 2.4:** Sodium and potassium ion interactions with TAR-like sequence. (a) Cationic association bins per residue in different solvents. (b)  $\text{Na}^+$  -- purine N7 RDF in mixed solvent solution. The inset represents the observed movement of the cation around A7. (c,d,e) Surface map of the cationic occupancy within the TAR-like duplex calculations: (c)  $\text{K}^+$  in 0.1 KCl simulations, (d)  $\text{Na}^+$  in 0.1 NaCl simulations, and (e) cations in mixed 0.1 NaCl and KCl solutions.

**Table 2.2:** Cationic Occupancy Statistics Calculated from 20 ns MD Simulations

Simulation	Bulged TAR Helix				TAR Helix								A7G TAR Helix,		A9G TAR Helix,	
	0.1M NaCl		(4Mg) + 0.1M NaCl		0.1M NaCl		0.1M KCl		0.1M (KCl) + NaCl		(1Mg) + 0.1 NaCl		0.1M NaCl		0.1M NaCl	
Position	t <sub>i</sub> , ps	Occ., %	t <sub>i</sub> , ps	Occ., %	t <sub>l</sub> , ps	Occ., %	t <sub>l</sub> , ps	Occ., %								
A9 N7	(A12)170	26.0	(22)	10.7	190	27.3	117	19.2	155 (112)	13.1 (18.7)	(20000)	(100)	113	33.3		(G) 14.4
A7 N7	145	14.6	31	1.4	204	22.7	95	21.9	122 (173)	7.3 (19.0)	254	22.5	142	(G) 22.3	217	27.9
A5 N7	214	20.0	324	35.0	148	17.4	132	21.1	(151)	(13.4)	191	12.5	245	21.2	155	13.4
G6 N7	131	16.9	170	13.4	140	14.9	85	25.6	104	3.2			134	17.9	85	10.5
G8 N7	(G11)132	14.6	(19)	13.1	92	11.3	145	20.0	74	10.9	103	39.6	110	11.6	84	14.6
G6 O6	54	4.2			101	2.8	63	9.9	(95)	(9.7)			101	10.2	37	1.3
G8 O6	(G11)108	5.9			75	1.6	56	8.6	50 (76)	3.1 (7.7)	211	32	57	5.6	112	7.5

<sup>a</sup>Only atoms with the highest occupancy are presented. Occupancy was calculated within 3.2 Å distance from the RNA atoms for Na<sup>+</sup>, within 3.8 Å for K<sup>+</sup>, and within 4.25 Å for Mg<sup>2+</sup>.

The effect of single base mutations which substitute guanosine for adenosine in two key positions (7 and 9) were examined for their effect on modulating the region specific residency near the excised bulge area. For both, the higher accumulation of ion affinity shifted toward the N7 of the neighboring adenosines. This is reflected in the calculated cationic occupancy from an Hydrogen bonding analysis performed using the Amber PTRAJ module (Table 2.2) and a cation distribution plot. The mechanism for the change in the cationic distribution is unclear but may be related to the flexibility of an ApG step versus a GpG as is the case for DNA <sup>71</sup>. The ApG step is the major groove edge of the guanine and is often the preferred binding of divalent metal ions. An interesting permutation on the triplet sequence has shown biological relevance; continuous GGA triplet repeats are an RNA aptamer for an anti-bovine protein <sup>73</sup>. AGG triplet repeats, although prominent in disease like Huntingtons, are more associated with destabilizing CGG hairpin formations in a positional dependent manner than anything unique to themselves. <sup>74</sup>. Experimental evidence for ApG step transitions in polypurine stretches as ion binding sites have been shown for divalent cations <sup>75</sup>. The effect seems to stem from favorable coordination of metal ions between the adjacent N7 of adenine and the N7 and O6 of guanosine. Increased cationic occupancy for the neighboring adenine N7s bear this out with each guanosine substitution.

Distortions of the RNA backbone at bulge sites create optimal binding pockets with exposed negatively charged phosphate groups <sup>43</sup>. The original 1998 crystallographic structure of Ippolito and Steitz showed cations bound to three phosphates at the bulge of HIV-TAR. While Na<sup>+</sup> and Mg<sup>2+</sup> compete for the same cation binding sites, Mg<sup>2+</sup> compensates for its slower diffusion rate with a higher charge density and smaller ionic

radius that allow for strong binding. Normally,  $Mg^{2+}$  cations will temporarily bind to the phosphate backbone before exchanging back in to bulk. Simulations involving  $Mg^{2+}$  were conducted in the presence of sodium to characterize this interaction. Using starting coordinates from the PDB structure 397D, the calcium ions were replaced with  $Mg^{2+}$  and simulated with 0.1M NaCl in explicit water for 20ns. Two additional simulations with only one  $Mg^{2+}$  but no bulge followed the same protocol of counterions and solvent, but were individually placed at positions revealed to be cationic binding sites from previous NaCl-only simulations. Our results are similar to previous characterizations of  $Mg^{2+}$  versus monovalent cation binding<sup>76</sup>. Using the crystal structure 397D as a set of starting coordinates and ionic species agreed with the observations made by Auffinger and their study of the 354D structure for that of a 5 S ribosomal RNA subunit. The three  $Mg^{2+}$  ions seen in 397D that strongly interacted with phosphates did not deviate significantly in position over the simulation time course of 20ns. The hydrogen bonding analysis showed only limited interaction with the N7 sites of the bases comprising the bulge. The lone  $Mg^{2+}$  coordinated between freely diffused through the central axis of the helix. Two simulations placing a single magnesium at an optimal coordination geometry within its second hydration shell at two different GpA steps (A7 and A9) of the TAR core helix produces similar results. Additional simulations placing a single magnesium within its first hydration shell (2.25 Å) to N7 of the A7 and A9 residues showed early drifting from the initial position. To compensate, the last equilibration stage and final heating included constraints of 0.5 and 0.1 Kcal/mol respectively. For both simulations, a single  $Mg^{2+}$  in excess  $Na^+$ , the  $Mg^{2+}$  is eventually outcompeted. Only for the cases as seen in the crystallographic image where  $Mg^{2+}$  is coordinated with an oxygen does it

remain throughout a simulation time course. This is seen for a close placement of a  $Mg^{2+}$  close to the A9 residue where the ion remains pentahydrated and is coordinated between the phosphate oxygen and N7 of A9, and site N7 of G10 is within the second hydration shell. Hbond analysis with PTRAJ showed a 100% occupancy for N7 of A7, 25.26% for N9 of A7, and 15.06% for N7 of G10. A mean distance of 2.350 Angstroms to the N7 of A7 is within the first hydration shell. The close distance with site N9 of A7 is indicative of close position with the phosphate backbone oxygen. Another feature of this coordination geometry is a flexed apical section of the Tar- helix facilitating access to N7 of G10. This may explain why it is more difficult to attain long-term close binding at the N7 position of A7 under excess  $Na^+$  since it is near the helical center which does not allow sufficient flexion. But the single  $Mg^{2+}$  in excess  $Na^+$  simulations are not expected to model fully the behavior of  $Mg^{2+}$  due to the aforementioned limitations in divalent cation modeling. Our 20ns simulations are not on the timescale of microseconds required to capture the desolvation of hexahydrated  $Mg^{2+}$ .<sup>76</sup> When the  $Mg^{2+}$  ions are tightly bound within the first hydration shell as in the simulation starting with the crystal structure,  $Na^+$  ions are displaced toward the N7 sites of A5 and A7 and are unable to approach the bulge area during the simulation. Extracting a single water molecule from  $Mg^{2+}$  is energetically costly; the enthalpy of hydration of  $Mg^{2+}$  at 1926 kJ/mol is approximately five times that of  $Na^+$ . When  $Mg^{2+}$  is bound with its first hydration shell, it is unlikely to relinquish it. Our simulations agree with an earlier 2.6 ns MD simulation by Golebiowski et al.<sup>77</sup> of the binding behavior of  $Na^+$  versus  $Mg^{2+}$  counterions on a fragment of hepatitis C RNA. They observed that the torsion of the A-form helix allows ion binding pockets to be created by bringing N7, O2, and the backbone

phosphate oxygen close together. This structural feature is present in both ApG and GpG steps, of which the latter likely accounts for our third binding site at positions 19 and 20.

In contrast, the competition for binding sites between  $K^+$  and  $Na^+$  follow a different dynamic.

Surface maps of ionic occupancy along the helical axis versus rotational angle  $\theta$  (Figure 2.4c,d,e,) indicate that even though the potassium has the highest propensity for ionic interactions, it also has higher diffusive motion around the helix as evidenced by multiple scattered peaks (Figure 2.4c). Sodium ions have higher affinity for specific locations as shown by fewer scattered peaks on the surface maps. Interestingly, in mixed potassium/sodium solvent the binding profile follows the pure sodium features and induces even more specific localization of cations, e.g. around residues A7 and A9. This can be attributed to the smaller radius of a sodium ion and its higher charge density. Smaller ions can approach electronegative atoms of purines more closely to increase electrostatic free energy<sup>25</sup>. Our observations agree with the fundamental properties of group I ions which dictate that the ion distribution is most strongly influenced by the strong charge density of the RNA major groove; ion location within the groove is biased by electronegative contributions from purines; and waters more weakly hydrate larger monovalent ions, such as  $K^+$ <sup>25</sup>.

## 2.4 Conclusions

In this work, interactions of various metal ions ( $Na^+$ ,  $K^+$ , and  $Mg^{2+}$ ) with HIV-1 TAR RNA core helix and other purine-rich duplexes were examined with a series of explicit solvent MD simulations. We observed that cations strongly prefer to interact with continuous purine-runs within helices. Moreover, the strongest preference is observed for alternating

adenine and guanine sequences, such as GAG sequence of TAR RNA which is important for Tat protein recognition, and is independent of cation typ. In TAR RNA all simulations indicate the same most probable binding site which are the A7 and A9 residue positions in the purine-rich run of a stem. Similar propensities for ionic binding were observed in the structure with the bulge. Our results of ion binding to a TAR core helix agrees well with the X-ray determined ion location in the TAR RNA bulged helix fragment. Thus we propose that accumulation of cation is triggered by the sequence of TAR RNA.

**Acknowledgements.** This work was supported by start-up funds from North Carolina State University. The computer support was provided by the High Performance Computing Center at North Carolina State University.

## References

1. Vieregg J, Cheng W, Bustamante C, Tinoco I, Jr. Measurement of the effect of monovalent cations on RNA hairpin stability. *J Am Chem Soc* 2007;129(48):14966-14973.
2. Ke A, Ding F, Batchelor JD, Doudna JA. Structural roles of monovalent cations in the HDV ribozyme. *Structure* 2007;15(3):281-287.
3. Summers JS, Shimko J, Freedman FL, Badger CT, Sturgess M. Displacement of Mn<sup>2+</sup> from RNA by K<sup>+</sup>, Mg<sup>2+</sup>, neomycin B, and an arginine-rich peptide: indirect detection of nucleic acid/ligand interactions using phosphorus relaxation enhancement. *J Am Chem Soc* 2002;124(50):14934-14939.
4. Perrotta AT, Been MD. A single nucleotide linked to a switch in metal ion reactivity preference in the HDV ribozymes. *Biochemistry* 2007;46(17):5124-5130.

5. Anwander EH, Probst MM, Rode BM. The influence of Li<sup>+</sup>, Na<sup>+</sup>, Mg<sup>2+</sup>, Ca<sup>2+</sup>, and Zn<sup>2+</sup> ions on the hydrogen bonds of the Watson-Crick base pairs. *Biopolymers* 1990;29(4-5):757-769.
6. Pyle AM. Metal ions in the structure and function of RNA. *J Biol Inorg Chem* 2002;7(7-8):679-690.
7. Draper DE. A guide to ions and RNA structure. *Rna* 2004;10(3):335-343.
8. Varma S, Rempe SB. Coordination numbers of alkali metal ions in aqueous solutions. *Biophys Chem* 2006;124(3):192-199.
9. Gold B, Marky LM, Stone MP, Williams LD. A review of the role of the sequence-dependent electrostatic landscape in DNA alkylation patterns. *Chem Res Toxicol* 2006;19(11):1402-1414.
10. Mocci F, Saba G. Molecular dynamics simulations of A. T-rich oligomers: sequence-specific binding of Na<sup>+</sup> in the minor groove of B-DNA. *Biopolymers* 2003;68(4):471-485.
11. Ponomarev SY, Thayer KM, Beveridge DL. Ion motions in molecular dynamics simulations on DNA. *Proc Natl Acad Sci U S A* 2004;101(41):14771-14775.
12. Stellwagen E, Dong Q, Stellwagen NC. Quantitative analysis of monovalent counterion binding to random-sequence, double-stranded DNA using the replacement ion method. *Biochemistry* 2007;46(7):2050-2058.
13. Dong Q, Stellwagen E, Stellwagen NC. Monovalent cation binding in the minor groove of DNA A-tracts. *Biochemistry* 2009;48(5):1047-1055.
14. Heddi B, Foloppe N, Hantz E, Hartmann B. The DNA structure responds differently to physiological concentrations of K(+) or Na(+). *J Mol Biol* 2007;368(5):1403-1411.
15. Šponer J, Sabat M, Gorb L, Leszczynski J, Lippert B, Hobza P. The Effect of Metal Binding to the N7 Site of Purine Nucleotides on Their Structure, Energy, and Involvement in Base Pairing. *The Journal of Physical Chemistry B* 2000;104(31):7535-7544.
16. Heide C, Feltens R, Hartmann RK. Purine N7 groups that are crucial to the interaction of Escherichia coli rnae P RNA with tRNA. *Rna* 2001;7(7):958-968.
17. Ennifar E, Walter P, Dumas P. A crystallographic study of the binding of 13 metal ions to two related RNA duplexes. *Nucleic Acids Res* 2003;31(10):2671-2682.

18. Chang KW, Oh J, Alvord WG, Hughes SH. The effects of alternate polypurine tracts (PPTs) and mutations of sequences adjacent to the PPT on viral replication and cleavage specificity of the Rous sarcoma virus reverse transcriptase. *J Virol* 2008;82(17):8592-8604.
19. Fitzgerald ME, Drohat AC. Structural Studies of RNA/DNA Polypurine Tracts. *Chemistry and Biology* 2003;15:203-204.
20. Casiano-Negrone A, Sun X, Al-Hashimi HM. Probing Na(+)-induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: new insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry* 2007;46(22):6525-6535.
21. Denisov VP, Halle B. Sequence-specific binding of counterions to B-DNA. *Proc Natl Acad Sci U S A* 2000;97(2):629-633.
22. Cesare Marincola F, Denisov VP, Halle B. Competitive Na(+) and Rb(+) binding in the minor groove of DNA. *J Am Chem Soc* 2004;126(21):6739-6750.
23. Krasovska MV, Sefcikova J, Reblova K, Schneider B, Walter NG, Sponer J. Cations and hydration in catalytic RNA: Molecular dynamics of the hepatitis delta virus ribozyme. *Biophysical Journal* 2006;91(2):626-638.
24. McDowell SE, Spackova N, Sponer J, Walter NG. Molecular dynamics simulations of RNA: An in silico single molecule approach. *Biopolymers* 2007;85(2):169-184.
25. Draper DE, Grilley D, Soto AM. Ions and RNA folding. *Annual Review of Biophysics and Biomolecular Structure* 2005;34:221-243.
26. Young MA, Jayaram B, Beveridge DL. Intrusion of counterions into the spine of hydration in the minor groove of B-DNA: fractional occupancy of electronegative pockets. *Biophys J* 1995;68:634-647.
27. Cheatham TE, Crowley MF, Fox T, Kollman PA. A molecular level picture of the stabilization of A-DNA in mixed ethanol-water solutions. *Proceedings of the National Academy of Sciences of the United States of America* 1997;94(18):9626-9630.
28. Koculi E, Thirumalai D, Woodson SA. Counterion charge density determines the position and plasticity of RNA folding transition states. *J Mol Biol* 2006;359(2):446-454.
29. Lambert D, Leipply D, Shiman R, Draper DE. The Influence of Monovalent Cation Size on the Stability of RNA Tertiary Structures. *J Mol Biol* 2009;390(4):791-804.

30. Sponer J, Leszczynski J, Hobza P. Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers* 2001;61(1):3-31.
31. Pabit SA, Qiu XY, Lamb JS, Li L, Meisburger SP, Pollack L. Both helix topology and counterion distribution contribute to the more effective charge screening in dsRNA compared with dsDNA. *Nucleic Acids Res* 2009;37(12):3887-3896.
32. Koculi E, Hyeon C, Thirumalai D, Woodson SA. Charge density of divalent metal cations determines RNA stability. *J Am Chem Soc* 2007;129(9):2676-2682.
33. Jones KA, Peterlin BM. Control of Rna Initiation and Elongation at the Hiv-1 Promoter. *Annual Review of Biochemistry* 1994;63:717-743.
34. Bennasser Y, Yeung ML, Jeang KT. RNAi therapy for HIV infection - Principles and practicalities. *Biodrugs* 2007;21(1):17-22.
35. Mayer M, Lang PT, Gerber S, Madrid PB, Pinto IG, Guy RK, James TL. Synthesis and testing of a focused phenothiazine library for binding to HIV-1 TAR RNA. *Chemistry & Biology* 2006;13(9):993-1000.
36. Zapata L, Bathany K, Schmitter JM, Moreau S. Metal-assisted hybridization of oligonucleotides, evaluation of circular 2'-O-Me RNA as ligands for the TAR RNA target. *European Journal of Organic Chemistry* 2003(6):1022-1028.
37. Olejniczak M, Gdaniec Z, Fischer A, Grabarkiewicz T, Bielecki L, Adamiak RW. The bulge region of HIV-1 TAR RNA binds metal ions in solution. *Nucleic Acids Research* 2002;30(19):4241-4249.
38. Ippolito JA, Steitz TA. A 1.3-angstrom resolution crystal structure of the HIV-1 trans-activation response region RNA stem reveals a metal ion-dependent bulge conformation. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95(17):9819-9824.
39. Al-Hashimi HM, Pitt SW, Majumdar A, Xu WJ, Patel DJ. Mg<sup>2+</sup>-induced variations in the conformation and dynamics of HIV-1 TAR RNA probed using NMR residual dipolar couplings. *Journal of Molecular Biology* 2003;329(5):867-873.
40. Carter-O'Connell I, Booth D, Eason B, Grover N. Thermodynamic examination of trinucleotide bulged RNA in the context of HIV-1 TAR RNA. *Rna-a Publication of the Rna Society* 2008;14(12):2550-2556.
41. Casiano-Negrone A, Sun XY, Al-Hashimi HM. Probing Na<sup>+</sup>-Induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: New insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry* 2007;46(22):6525-6535.

42. Chen AA, Draper DE, Pappu RV. Mechanism Of Monovalent Counterion Specificity In A RNA Kissing Loop Complex. *Biophysical Journal* 2009;96(3S1):576-576.
43. Hermann T, Patel DJ. RNA bulges as architectural and recognition motifs. *Structure* 2000;8(3):R47-R54.
44. Vieregg J, Cheng W, Bustamante C, Tinoco I. Measurement of the effect of monovalent cations on RNA hairpin stability. *J Am Chem Soc* 2007;129(48):14966-14973.
45. Shcherbakova I, Mitra S, Laederach A, Brenowitz M. Energy barriers, pathways, and dynamics during folding of large, multidomain RNAs. *Curr Opin Chem Biol* 2008;12(6):655-666.
46. Schroeder KT, Lilley DM. Ion-induced folding of a kink turn that departs from the conventional sequence. *Nucleic Acids Res* 2009.
47. Woodson SA. Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr Opin Chem Biol* 2005;9(2):104-109.
48. Wang JMC, P. Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* 2000;21(12):1049-1074.
49. Case DA, Darden TA, Cheatham I, T.E. , Simmerling CL, Wang J, R.E. Duke, R.Luo, K.M. Merz, D.A. Pearlman, M. Crowley, R.C. Walker, W. Zhang, B. Wang, S.Hayik, A. Roitberg, G. Seabra, K.F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H.Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D.H. Mathews, C.Schafmeister, W.S. Ross, Kollman PA. AMBER 9. University of California, San Francisco 2006.
50. Hashem Y, Auffinger P. A short guide for molecular dynamics simulations of RNA systems. *Methods* 2009;47(3):187-197.
51. Draper DE, Grilley, D. and Soto, A. M. ions and RNA folding *Annual Review of Biophysics and Biomolecular Structure* 2005; 34 221-243.
52. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A Smooth Particle Mesh Ewald Method. *Journal of Chemical Physics* 1995;103(19):8577-8593.
53. Auffinger PW, E. Simulations of the molecular dynamics of nucleic acids. *Current opinion in structural biology* 1998;8(2):227-236.

54. Yingling YG, Shapiro BA. Dynamic behavior of the telomerase RNA hairpin structure and its relationship to dyskeratosis congenita. *Journal of Molecular Biology* 2005;348(1):27-42.
55. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, Haak JR. Molecular-Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* 1984;81(8):3684-3690.
56. Baran N, Lapidot A, Manor H. Formation of DNA triplexes accounts for arrests of DNA synthesis at d(TC)<sub>n</sub> and d(GA)<sub>n</sub> tracts. *Proceedings of the National Academy of Sciences of the United States of America* 1991;88(2):507-511.
57. Hodgson JW, Argiropoulos B, Brock HW. Site-specific recognition of a 70-base-pair element containing d(GA)<sub>n</sub> repeats mediates bithoraxoid polycomb group response element-dependent silencing. *Molecular and cellular biology* 2001;21(14):4528-4543.
58. Nicholas J. Evolutionary aspects of oncogenic herpesviruses. *Mol Pathol* 2000;53(5):222-237.
59. Zorro S, Arias M, Riano F, Paris S, Ramirez L, Uribe O, Garcia L, Vasquez G. Response to ODN-CpG by B Cells from patients with systemic lupus erythematosus correlates with disease activity. *Lupus* 2009;18(8):718-726.
60. Alexandrov BS, Gelev V, Monisova Y, Alexandrov LB, Bishop AR, Rasmussen KO, Usheva A. A nonlinear dynamic model of DNA with a sequence-dependent stacking term. *Nucleic Acids Res* 2009;37(7):2405-2410.
61. Williams TT, Barton JK. The effect of varied ion distributions on long-range DNA charge transport. *J Am Chem Soc* 2002;124(9):1840-1841.
62. Luebke KJ, Landry SM, Tinoco I. Solution conformation of a five-nucleotide RNA bulge loop from a group I intron. *Biochemistry* 1997;36(33):10246-10255.
63. Zacharias M, Hagerman PJ. The Bend in Rna Created by the Transactivation Response Element Bulge of Human-Immunodeficiency-Virus Is Straightened by Arginine and by Tat-Derived Peptide. *Proceedings of the National Academy of Sciences of the United States of America* 1995;92(13):6052-6056.
64. Hsiao C, Tannenbaum M, VanDeusen H, Herskovitz E, Perng G, Tannenbaum A, Williams LD. Complexes of Nucleic Acids with Group I and Group II Cations. In: Hud NV, editor. *Nucleic Acid Metal Ion Interactions*. London: The Royal Society of Chemistry 2008. p 1-35.

65. Sponer J, Leszczynski J, Hobza P. Structures and energies of hydrogen-bonded DNA base pairs. A nonempirical study with inclusion of electron correlation. *Journal of Physical Chemistry* 1996;100(5):1965-1974.
66. Timsit Y, Bombard S. The 1.3 angstrom resolution structure of the RNA tridecamer r(GCGUUUGAAACGC): Metal ion binding correlates with base unstacking and groove contraction. *Rna-a Publication of the Rna Society* 2007;13(12):2098-2107.
67. Mayer-Jung C, Moras D, Timsit Y. Hydration and recognition of methylated CpG steps in DNA. *Embo Journal* 1998;17(9):2709-2718.
68. Lippert B. Multiplicity of Metal Ion Binding Patterns to Nucleobases. *Coordination Chemistry Reviews* 2000;200-202:487-516.
69. Rodgers MT, Armentrout PB. Noncovalent metal-ligand bond energies as studied by threshold collision-induced dissociation. *Mass Spectrom Rev* 2000;19(4):215-247.
70. Howerton SB, Sines CC, VanDerveer D, Williams LD. Locating monovalent cations in the grooves of B-DNA. *Biochemistry* 2001;40(34):10023-10031.
71. Hud NV, Polak M. DNA-cation interactions: The major and minor grooves are flexible ionophores. *Curr Opin Struct Biol* 2001;11(3):293-301.
72. Sponer J, Burda JV, Sabat M, Leszczynski J, Hobza P. Interaction between the guanine-cytosine Watson-Crick DNA base pair and hydrated group IIa (Mg<sup>2+</sup>, Ca<sup>2+</sup>, Sr<sup>2+</sup>, Ba<sup>2+</sup>) and group IIb (Zn<sup>2+</sup>, Cd<sup>2+</sup>, Hg<sup>2+</sup>) metal cations. *Journal of Physical Chemistry A* 1998;102(29):5951-5957.
73. Murakami K, Nishikawa F, Noda K, Yokoyama T, Nishikawa S. Anti-bovine prion protein RNA aptamer containing tandem GGA repeat interacts both with recombinant bovine prion protein and its beta isoform with high affinity. *Prion* 2008;2(2):73-80.
74. Sobczak K, Krzyzosiak WJ. CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. *J Biol Chem* 2005;280(5):3898-3910.
75. Hofmann HP, Limmer S, Hornung V, Sprinzl M. Ni<sup>2+</sup>-binding RNA motifs with an asymmetric purine-rich internal loop and a G-A base pair. *Rna* 1997;3(11):1289-1300.
76. Auffinger P, Bielecki L, Westhof E. Symmetric K<sup>+</sup> and Mg<sup>2+</sup> ion-binding sites in the 5 S rRNA loop E inferred from molecular dynamics simulations. *J Mol Biol* 2004;335(2):555-571.

77. Golebiowski J, Antonczak S, Di-Giorgio A, Condom R, Cabrol-Bass D. Molecular dynamics simulation of hepatitis C virus IRES IIIId domain: structural behavior, electrostatic and energetic analysis. *J Mol Model* 2004;10(1):60-68.

# Chapter 3

## **Decoupling Structure from Sequence Effects in Cation and Nucleic Acid Helix Interactions**

## **Decoupling Structure from Sequence Effects in Cation and Nucleic Acid Helix Interactions**

Latsavongsakda Sethaphong and Yaroslava G. Yingling

Department of Materials Science and Engineering, NC State University, Raleigh, NC

### **Abstract**

Genomic control of genetic expression is highly complex. One fundamental mechanism involves protein recognition of nucleic acid duplexes which can be highly sensitive to even minute variations of physiological cation concentrations. Recent studies have pointed at intrinsic sequence dependent variations in the electrostatic landscape arising from the unique local geometry that can be influenced by cations. In this work, we conducted molecular dynamics simulations of DNA and RNA duplexes that examine the subtleties of cationic interactions with these structures. Despite small differences in the chemical moieties of DNA and RNA, stark contrasts in counterion interactions occurred mostly due to intrinsic sequence dependent structural differences. Not all sequences are equally sensitive to cationic interactions; sodium and potassium cations interact more strongly with RNA than with DNA helices for all modeled sequences. In the presence of either 0.1M Na<sup>+</sup> or 0.1M K<sup>+</sup>, ion dynamics were not significantly altered for a particular duplex. PolyAT DNA helices were the least sensitive to changes in the ionic environment while polyGC RNA helices were most affected. Finally, modeling with a fixed helical geometry that replicated observed cationic dynamics and diffusive ion interactions revealed that structure influenced the electrostatics more than the differences in chemical moieties alone.

**Keywords:** Molecular Dynamics Simulations; RNA and DNA helices; Cations; Helical Geometry; Molecular Recognition

### 3.1 Introduction

RNA's catalysis of biochemical reactions and self-regulating functions distinguishes it from DNA that has the singular biological purpose of storing genetic information. RNA's ability to adopt complex tertiary structures permit a specificity of interaction with ligand molecules; thus giving rise to enzymatic functionality, direct control of transcription, viral replication and self-processing<sup>1</sup>. Of these tertiary structures, the helical duplex is most identifiable with both nucleic acids due to complementary base-pairing. These helices are classified by turn regularity and the major and minor groove dimensions; the long and narrow B-form helix which is most prevalent for DNA and the wider and more twisted A-form helix which is common for RNA. The fundamental distinction between RNA and DNA derives from a difference in the backbone: a 2'-hydroxyl group of RNA's ribose sugar clashes with an adjacent nucleotide's phosphate group imposing an A-form double helix. However, the more flexible DNA helix may undergo solvent-induced transitions between the slim and elongated B-form and wide and stubby A-form helix<sup>2-4</sup>.

Both the shape and sequence of nucleic acid helices factor in protein binding specificity, since proteins may recognize nucleic acids through direct (sequence) and indirect (shape) readouts<sup>5</sup>. However, helical conformations are polymorphic and heavily influenced by base stacking, inter-base effects<sup>6-8</sup> and interactions with metal cations<sup>9</sup>. Moreover, the ability to recognize specific helical conformations with changes in the ionic environment is used as a genetic control by some regulatory proteins<sup>10</sup>. For example, repressor binding affinity of the bacteriophage 434 site falls dramatically with increased metal cation

concentration in a sequence dependent manner<sup>11</sup>. Since a variety of protein families identify nucleic acid substrates by these intrinsic features, helical conformations are proving to be as critical as base-specific interactions<sup>12-14</sup>.

The fundamental importance of helical conformation recognition can also be illustrated through the mechanisms that maintain genetic integrity and regulate expression. In order to carry out maintenance functions, the structures of the DNA and RNA ligases discriminate between polynucleotides based upon their inherent geometries<sup>15,16</sup>. Recently solved structures show that both DNA and RNA ligases utilize a multi-domain protein architecture where proper alignment of nucleotide ends require an RNA-like conformation of the substrate near nucleotide breaks. Consequently, DNA ligases must alter the dimensions of the minor groove to enforce an RNA shape onto their DNA substrate.

This complexity of cation mediated protein recognition of nucleic acids has necessitated multiple approaches to examine sequence and structure dependent mechanisms<sup>17-19</sup>. Experimental methods have encountered difficulty in measuring specific interactions with metal cations in order to assess their functional roles; monovalent cations such as sodium and potassium ions often are indistinguishable from water in X-ray crystallographic structures<sup>20</sup>. NMR studies usually require isotopic labeling of individual residues or isotope substitutions with proton dense cations such as thallium<sup>21,22</sup>. Consequently, molecular dynamics (MD) simulations have become the tool for investigating the interactions of monovalent cations with nucleic acids. Simulations have been successful in yielding atomistic details of conformation, ion-specific interactions and aggregation due to nucleobase

modifications, protonation, and solvent motion while care must still be taken to correctly set up the system<sup>23-28</sup>.

The helical duplex, as the most common stable secondary structure, provides a good model in which to examine the differences in cationic interactions between DNA and RNA. For example, Cheng et al. were able to investigate the effect of monovalent cation hydration on a condensed DNA duplex structure to reveal distinctive characteristics between Na<sup>+</sup> and K<sup>+</sup><sup>29</sup>. Conformational switching between related helical structures have been investigated to reveal a strong correlation with specific sequences where DNA is more mutable<sup>30</sup>. Two early studies by Auffinger and Westhof sought to compare the behavior of potassium cations on complementary strands of alternating guanine and cytosine DNA and RNA duplexes<sup>31,32</sup>. Previously, Feig et al. had observed greater A-form DNA interactions with monovalent cations than B-form DNA under high salt conditions, ~1M<sup>33,34</sup>. However, there were no studies that directly compare the roles of RNA and DNA sequence and geometry. Here, we perform a comprehensive examination of the effect of sequence and geometry on the ability of nucleic acid duplexes to be recognized by ions. We compare the effect of sodium versus potassium cationic interaction to complementary non alternating idealized polyGC, polyAT/U, and random DNA and RNA duplexes under physiological salinity of 100mM; we also examine the extent to which helical geometry alone contributes to these cationic interactions.

### **3.2. Materials and Methods**

Nucleic acid helices were generated using the Nucleic Acid Builder module of the molecular dynamics software Amber 9.0<sup>35</sup>. Random helix coordinates were established by

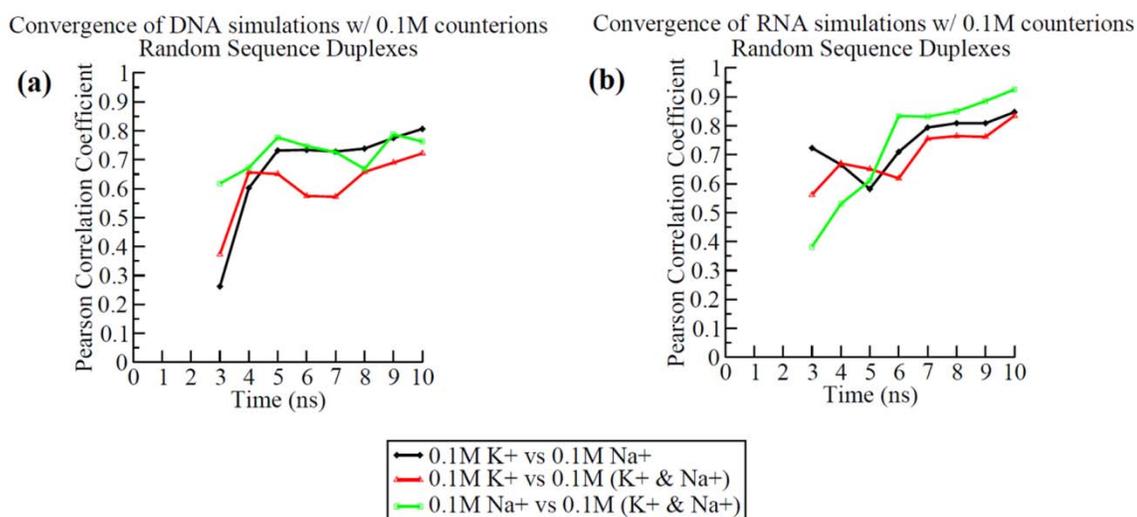
removing three bulged residues from the native crystal structure (Protein Data Bank (PDB) ID 397D)<sup>36</sup>. Nine different simulation conditions based on seven structures and two counterion environments were generated for a total of 140ns.

All atom simulations utilized the updated Cornell force field for nucleic acids<sup>37,38</sup>. Ions  $K^+$ ,  $Na^+$ , and  $Cl^-$  ions are modeled as point charges with van der Waals spheres without either polarization or charge transfer effects. This methodology represents monovalent ions faithfully. However, ad hoc adaptation of the Åqvist parameters for the AMBER-99 force field led to artifacts in long simulations (more than 50 ns) of biomolecules in high concentration salt solutions resulting in salt crystal formation below their solubility limit according to Chen et al.<sup>39</sup>. Subsequently, Joung and Cheatham reparameterized the Lennard-Jones potential for ions and specific rigid water models<sup>40,41</sup>. Validation of the new force field modifications was further verified by Zhang et al. utilizing GROMACS molecular dynamics package<sup>42</sup>. Although fixed charge representations restrict model flexibility and ignore polarization effects, these forcefields perform well for moderately long simulations<sup>43</sup>. Generally, biological macromolecules are less polarizable than their aqueous solvent. Conformational effects, which have been the focus of forcefield refinements<sup>44,45</sup> are observed for very long simulations (more than 50 ns). Other force field refinements include problems with high salt concentration that constitute dense systems<sup>46</sup>. Despite these concerns, the AMBER forcefield has proven to be physically meaningful for moderate simulations and has addressed successfully various questions related to mechanical properties, folding, and inter-molecular interactions of nucleic acids<sup>47-49</sup>. Therefore, we

believe that our experiments involving low salt concentrations of 100 mM with single duplexes were sufficiently calibrated to mitigate known artifacts<sup>50</sup>.

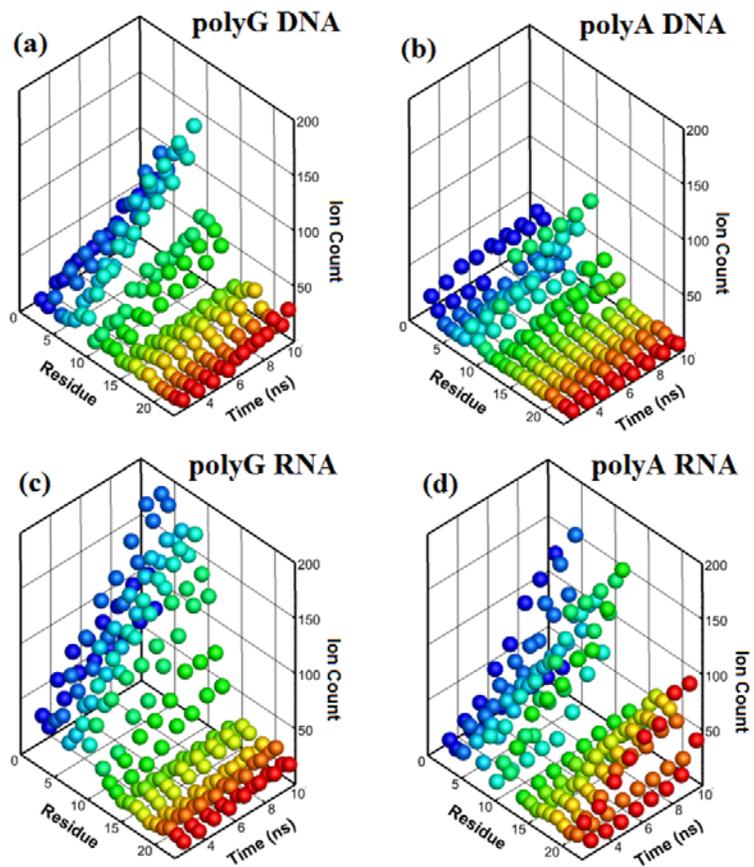
All nucleic acid structures were subjected to conjugate gradient energy minimization for 5000 steps. Minimized NA structures were then neutralized with Na<sup>+</sup> ions and immersed in a water box with at least 10 Å deep solvation shell using the TIP3P water model<sup>51</sup>. Additional Na<sup>+</sup> and Cl<sup>-</sup> ions or K<sup>+</sup> and Cl<sup>-</sup> ions were added to represent a 0.1 M effective salt concentration around a given NA helix. The equilibration of each NA sample was carried out in 11 stages starting from the solvent minimization for 10000 steps and keeping the duplex restrained for 200 Kcal/mol. The system was heated to 300K in 40 ps while imposing a 200kcal/mol constraint on the duplex. A brief NPT MD run was performed for 200 ps with a duplex restrained maintained at 200 kcal/mol. Another constrained minimization step follows with the restraint of 25 kcal/mol for 10000 steps. A second NPT MD run was performed at 25 kcal/mol restraint for 20 ps. Subsequently four additional 1000 cycle minimization steps were performed while relaxing the positional constraint from 20 kcal/mol to 5 kcal/mol in 5 kcal/mol increments. A final unconstrained minimization stage of 1000 cycles was performed before reheating the system to 300K at constant volume within 40 ps. Subsequently, NPT equilibrations were performed to ensure uniformity in solvent density. Long range electrostatic interactions were calculated by Particle Mesh Ewald summation (PME)<sup>52</sup> and the non-bonded interactions were truncated at 9 Å cutoff along with a 0.00001 tolerance of Ewald convergence. A Berendsen thermostat maintained temperature at 300 K<sup>53</sup>. The SHAKE algorithm was used to constrain the position of hydrogen atoms<sup>54</sup>. The production simulations were performed for an NVT ensemble. Each production simulation was

performed for 20 ns with a 2 fs time step. The simulation time was chosen in order to allow convergence of the simulations; we have shown previously that simulations of helices in solvent converged to a 0.98 Pearson correlation coefficient by 10 ns<sup>55</sup>. In this work, A-form RNA duplex systems of a random sequence with 0.1M counterions of KCl and NaCl converged to a  $0.79 \pm 0.038$  Pearson correlation coefficient by 7 ns. By 10 ns, a 0.92 Pearson correlation coefficient between a NaCl system versus mixed KCl and NaCl was reached for cationic occupancies at residue resolution (Figure 3.1).

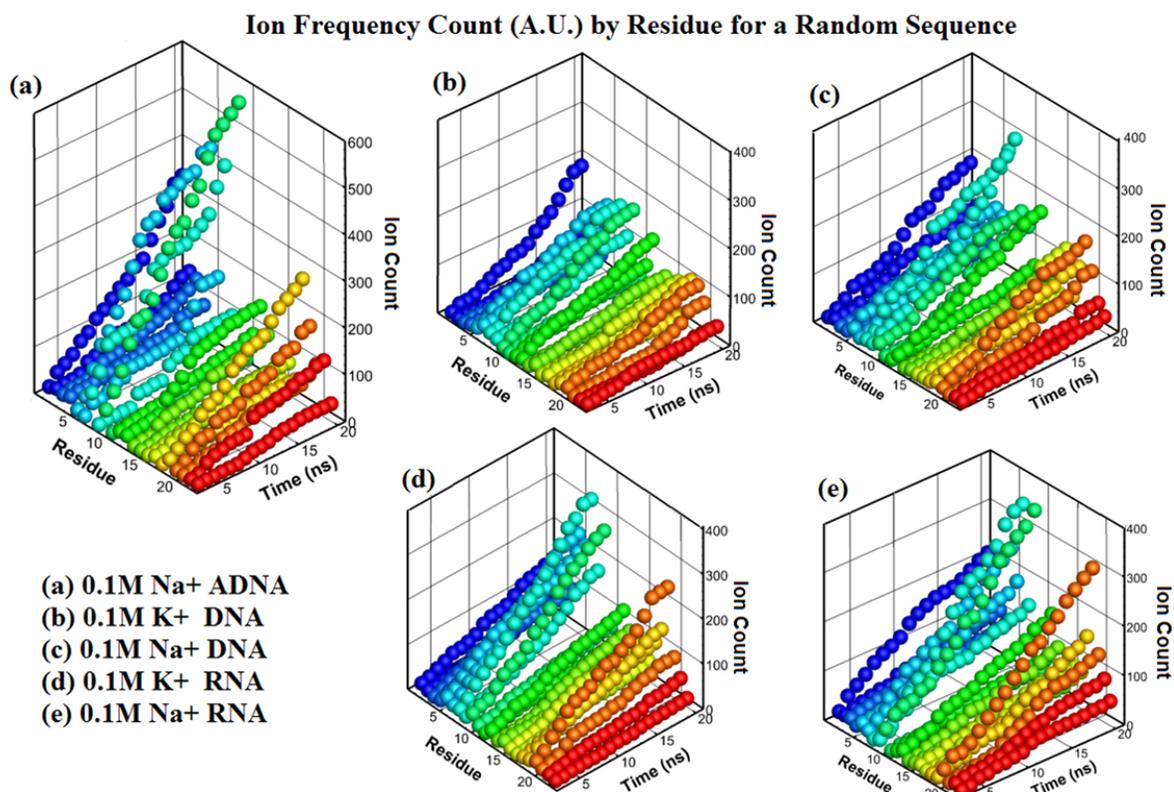


**Figure 3.1:** Convergence of the Pearson correlation coefficient for DNA and RNA helix systems of a random sequence thru 10ns. (a) B-form DNA duplex systems of a random sequence with 0.1M counterions of KCl and NaCl converged to a  $0.76 \pm 0.042$  Pearson correlation coefficient by 10 ns; by 10 ns, a 0.76 Pearson correlation coefficient between a NaCl system versus mixed KCl and NaCl was reached for cationic occupancies at residue resolution. (b) A-form RNA duplex systems of a random sequence with 0.1M counterions of KCl and NaCl converged to a  $0.79 \pm 0.038$  Pearson correlation coefficient by 7 ns; by 10 ns, a 0.92 Pearson correlation coefficient between a NaCl system versus mixed KCl and NaCl was reached for cationic occupancies at residue resolution.

Ion frequency counts used to calculate the Pearson correlation coefficients for unmixed salt systems are shown in Figure 3.2 and Figure 3.3. The fixed helical geometry simulation of an A-form DNA duplex involving a random sequence followed the same equilibration protocol as the other NA simulations. The key difference involved the last heating stage and subsequent production phase where only the solvent and cations were allowed to freely diffuse via belly dynamics<sup>35</sup>. All molecular dynamics trajectories were processed using in-house scripts along with the standard tool suite accompanying Amber 9.0. Likewise, the PTRAJ module was used to perform hydrogen bonding analysis and residency times with a distance cut off of 3.2 Å.



**Figure 3.2** Ion interaction frequency in arbitrary count every 20ps of the last 8ns in the polypurine sequence trajectories for DNA and RNA in 0.1M sodium.



**Figure 3.3:** Ion interaction frequency in arbitrary count every 20ps of the last 18ns in the random sequence trajectories for DNA and RNA.

Identification of the ions and residue association was performed by an in house PERL script that processes a user defined sequence of snapshots taken from a trajectory to output all specific ion distances to the nearest residue by time step given a distance cutoff. This serial positional and time data is used to calculate the residency times from contiguous periods. This script then bins the data in order to generate the intensity maps that depict ion residency versus residue number following further processing with Tecplot. High residency ions were chosen if they had occupancy times of at least 200 ps with the same residue. Calculation of the interaction energies between high residency ions and nucleic acids was performed by the ANAL program of Amber 8 which utilizes a shell script to iteratively process multiple

snapshots from the trajectory corresponding to the time segments with long residency interaction events. This shell script generates the input files for the ANAL program and extracts the final energies into a table file. The electrostatic and van der Waals interaction energies were calculated between an individual ion and the entire residue inclusive of the backbone phosphate, sugar, and nucleobase during the time periods identified as contiguous interaction events. Measurement of the helical parameters of the structures from the last one hundred picoseconds of a trajectory was performed using the program CURVES 5.2<sup>56</sup>. Internally developed shell scripts processed the output to generate statistics. The Adaptive Poisson Boltzman Solver (APBS) plug-in utility for Pymol (Schrodinger Inc)<sup>57</sup> was used to calculate the electrostatic surface of the representative duplexes.

### **3.3 Results and Discussion**

#### **3.3.1. Non random sequences**

Polypurine nucleic acid helical duplexes of polyGC and polyAT/U present themselves as good models for studying cationic interactions by isolating the structural influences from sequence specific effects. Biologically, polypurine stretches are often found in the promoter regions of genes where they act as binding sites for transcription factors<sup>58,59</sup>. Short dinucleotides of polypurines also appear to regulate the positional packing of DNA into nucleosomes which are then folded into chromosomes<sup>60</sup>. The position of these nucleosomes ultimately regulates how the genetic information is used. Consequently, these simple motifs give rise to some of the complex mechanisms for genetic regulation and gives impetus for studying closely the intimate interactions between cations, structure and sequence in these polypurine duplexes.

### 3.3.2. Polyguanine RNA and DNA duplexes

We begin by examining the polyGC duplexes of DNA and RNA (Figure 3.4) wherein only the sugar pucker distinguishes the two. A subtle difference in the sugar pucker can be seen in Figure 3.4b,d where the 2' carbon is displaced from the plane resulting in the characteristic pucker of DNA, and RNA's 2'hydroxyl group of the ribosefuranose pushes out its 3' sugar producing the 3'endo pucker. Both nucleic acid duplexes carry a net uncompensated charge differing minutely in magnitude. The B-form helix of dsDNA possess a linear charge density of  $\sim -0.96$  nC/m that is slightly smaller than dsRNAs  $\sim -1.23$  nC/m<sup>61</sup>. To examine the difference in cationic interactions with RNA and DNA duplexes we calculated individual cation residency time versus sequence (Figure 3.5). It is clear that RNA displays higher intensities and more frequent cationic interaction events as shown by higher intensity and larger patches of red on Figure 2. The higher uncompensated charge for RNA may likely contribute to the greater number of ions interacting with the polyGC RNA duplex in Figure 3.5.

From a structural perspective, the deeper minor groove of DNA would appear to better accommodate cations. Analysis using the CURVES program showed that the average width of the polyGC DNA major groove of 16.79 Å was over twice that of the polyGC RNA (Table 3.1). The values for polyGC RNA of 9.8 Å and 7.4 Å for the minor and major groove, respectively, qualitatively agree with the crystal structure 1QCU<sup>62</sup> as measured by CURVES (Table 3.1). Excluding phosphate backbone dominant interactions and focusing on nucleobase interaction, highest cationic occupancies occur at the N<sub>7</sub> atom within the major groove of the most favored residues, G5-7 and 9 (Table 3.3). Figure 2 shows that the longest

residency cations occupy the middle of the both duplexes where the major groove is the deepest.

**Table 3.1:**Major and minor groove parameters of simulated duplexes and X-ray structures

	Minor Groove		Major Groove	
	Width(Å)	Depth(Å)	Width(Å)	Depth(Å)
polyGC-DNA	7.47±0.42	3.31±0.27	16.79±1.82	8.30±2.36
polyGC-RNA	9.8±0.2	0.61±0.11	7.4±0.06	11.71±0.15
polyAT-DNA	5.95±1.4	4.18±1.14	14.77±0.56	4.42±3.0
polyAU-RNA	9.81±0.34	0.9±0.13	8.34±2.16	9.26±1.95

	Minor Groove		Major Groove	
	Width(Å)	Depth(Å)	Width(Å)	Depth(Å)
<b>0.1M NaCl</b>				
DNA	7.09±1.61	4.02±0.88	14.09±2.17	6.2±1.54
RNA	9.44±0.24	0.98±0.16	10.61±0.26	11.78±0.19
A-DNA	10.27±0.061	0.80±0.16	3.60±0.02	10.49±0.07
<b>0.1M KCl</b>				
DNA	6.69±0.98	4.09±0.63	13.49±0.67	8.45±0.54
RNA	9.59±0.26	0.90±0.13	10.61±3.69	8.17±3.97

	Minor Groove		Major Groove	
	Width(Å)	Depth(Å)	Width(Å)	Depth(Å)
<b>1QCU</b>				
polyGC-AB	10.14±0.28	0.51±.30	5.28±0.41	9.42±0.04
polyGC-CD	10.14±0.28	0.46±.28	5.30±.41	9.49±0.08
<b>1H1K</b>				
polyAU-18	9.90±0.93	-0.34±0.69	6.65±1.73	9.36±1.55
polyAU-33	10.15±1.19	0.45±1.31	5.75±3.18	8.18±1.18

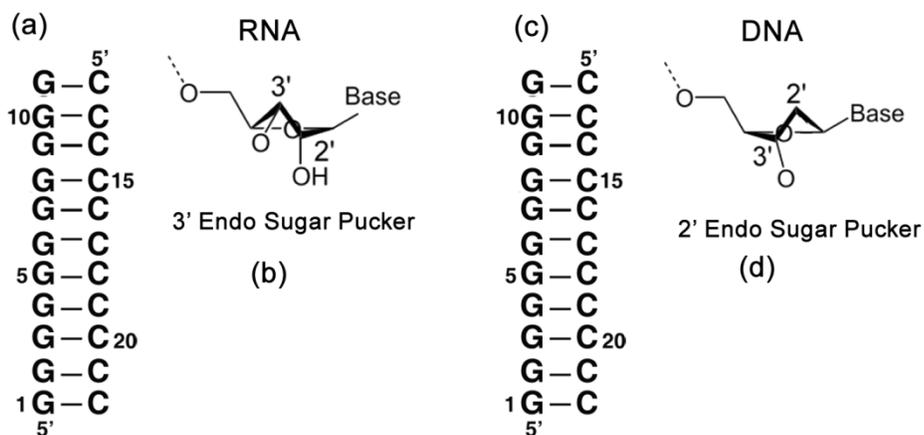
In order to determine electrostatic contributions of nucleic acids we compared the interaction energies of ions (Table 3.2). Table 3.2 lists the strongest and the weakest interacting high residency ions for all duplexes. A comparison of the electrostatic dominant interaction energies of the highest residency ions is very similar; the ion-RNA energies are higher than those of ion-DNA. Moreover, we observe that the interaction of cations to either polyGC DNA or RNA is purine dominated, and longer in duration for RNA than for DNA. The highest interaction energy for the polyGC RNA system is -90.237 kcal/mol, due to the difference in electrostatic landscape. The highest interaction energy for polyGC DNA is -76.446 kcal/mol, which is still much lower than the lowest value for the RNA system at -85.123 kcal/mol despite a similar mean interaction distance of  $\sim 2.59\text{\AA}$ . Thus, the difference in sugar pucker between RNA and DNA results in RNA having stronger cationic interactions, due to combination of structural helical geometry and electrostatics differences. Sugar puckering was also suggested to affect sequence-dependent cation interaction by Nakano et al.; they found it impossible to distinguish among duplex species simply by linearly plotting either the enthalpy or the free energy of duplex formation versus  $T\Delta S$  in an examination of the role of cations in duplex formation<sup>63</sup>.

**Table 3.2:** Interaction energies of high residency ions with specific residues of RNA and DNA after truncating the first 2ns and analyzing the remaining 18ns.

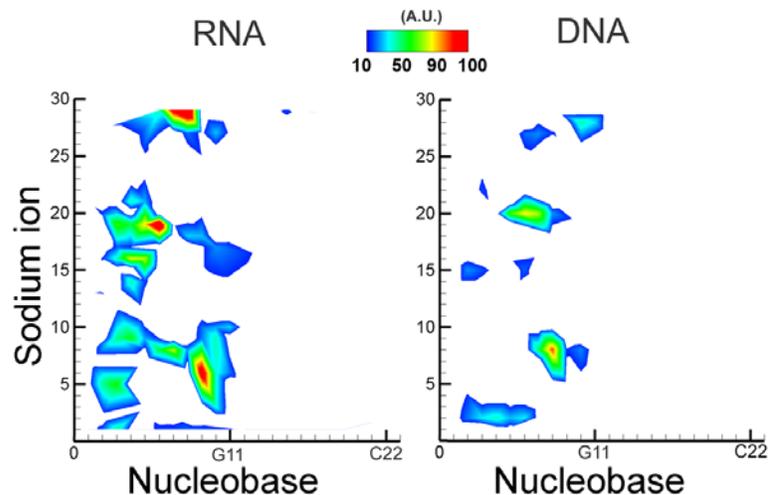
<b>polyGC RNA in 0.1 M NaCl</b>					
residue	cation no.	Mean interaction distance, Å	$\Delta E^{\text{Electro}}$ , kcal/mol	$\Delta E^{\text{vdw}}$ , kcal/mol	$\Delta E^{\text{interaction}}$ , kcal/mol
G3	9	2.594±0.159	-86.352±6.782	1.228±0.962	-85.123±6.529
G6	29	2.619±0.162	-91.516±3.905	1.280±0.945	-90.237±3.736
<b>polyGC DNA in 0.1 M NaCl</b>					
G2	15	2.521±0.124	-63.663±5.228	1.448±0.974	-62.216±4.685
G7	8	2.569±0.189	-77.946±8.020	1.499±1.222	-76.446±7.701
<b>polyAU RNA in 0.1 M NaCl</b>					
A3	17	2.497±0.110	-78.368±5.150	1.258±1.194	-77.110±4.573
A10	24	2.529±0.151	-56.252±10.373	1.623±1.293	-54.630±9.707
<b>polyAT DNA in 0.1 M NaCl</b>					
A2	23	2.594±0.172	-48.873±7.123	1.349±1.180	-47.524±6.330
A6	25	2.621±0.167	-69.023±7.642	1.728±1.510	-67.295±7.169
<b>Random RNA in 0.1 M NaCl</b>					
A7	16	2.619±0.196	-71.291±8.650	1.262±1.367	-70.029±7.950
G19	23	2.657±0.217	-86.810±5.013	1.244±1.188	-85.566±4.471
<b>Random DNA in 0.1 M NaCl</b>					
A7	17	2.578±0.187	-49.533±13.077	1.377±1.286	-48.155±12.23
G8	2	2.435±0.127	-72.783±7.069	1.722±1.229	-71.062±6.468
G19	17	2.605±0.192	-55.002±13.332	0.699±1.181	-54.304±12.555
<b>Random RNA in 0.1 KCl</b>					
A7	34	2.932±0.229	-66.574±7.596	1.311±1.476	-65.263±7.024
G19	19	2.814±0.176	-78.453±8.128	1.711±1.458	-76.741±7.503
<b>Random DNA in 0.1 KCl</b>					
A5	12	2.931±0.215	-48.252±5.560	1.529±1.483	-46.723±4.913
G8	26	2.749±0.127	-67.157±4.998	2.210±1.265	-64.947±4.179
G20	31	2.804±0.244	-62.778±6.177	1.376±0.886	-61.401±5.771
<b>Random A-DNA in 0.1 M NaCl</b>					
A7	15	2.560±0.157	-68.566±7.900	1.529±1.306	-67.037±7.137
G19	25	2.566±0.148	-85.643±6.120	1.475±1.194	-84.168±5.902

**Table 3. 3:** Occupancy and lifetimes of Na<sup>+</sup> hydrogen bonding to RNA and DNA polypurine duplexes.

	polyGC RNA		polyGC DNA	
	% occ.	max occ. (ps)	% occ.	max occ. (ps)
Major groove				
G4@N7	24.1	87	5.2	33
G5@N7	17.6	112	7.4	35
G6@N7	21.9	140	13.7	46
G7@N7	16.5	82	11.6	56
G8@N7	22.9	107	12.1	51
G9@N7	24.2	96	4.7	38
Minor groove				
C18@O2	0.4	14		
C22@O2			0.8	23



**Figure 3.4:** Modeled RNA and DNA helices (a) Polyguanine RNA sequence. (b) 3'Endo Sugar Pucker of RNA's backbone. (c) Polyguanine DNA sequence. (d) 2' Endo Sugar Pucker of DNA's backbone.



**Figure 3.5:** Heat maps of cumulative diffusive residencies of individually tracked sodium cations within 5 Å of a nucleobase in polyguanine duplexes of RNA and DNA.

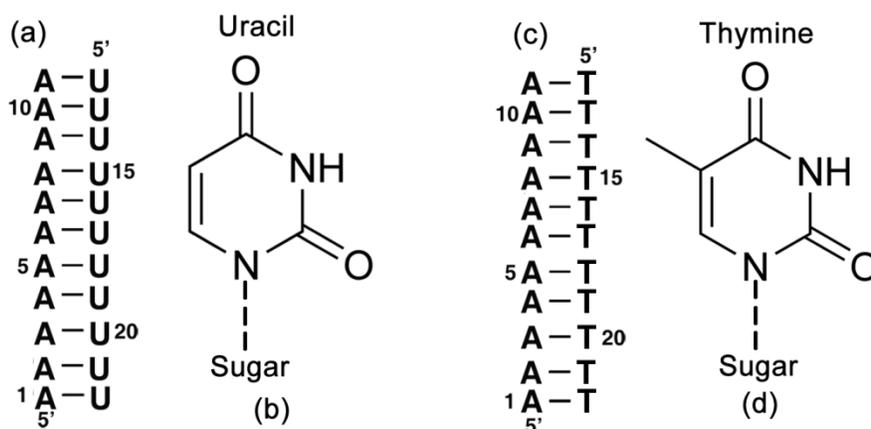
### 3.3.3. Polyadenine RNA and DNA duplexes

Next, we examined the combined effect of sugar-pucker differences and nucleoside changes with uracil in RNA versus thymine in DNA (Figure 3.5). Similar as seen in polyGC duplexes, cations interact most favorably with the purine N<sub>7</sub> chelation site of polyAT DNA and polyAU RNA duplexes. The cumulative ion intensity map (Figure 3.7) shows that cation interaction is more favored in the RNA polyAU helix. Occupancies for polyAT/U (Table 2) display a propensity for cationic association for the interior residues. The cumulative difference is put into stark detail in Figure 3.8 which shows that the per residue affinity of the polyAT DNA to be the least amongst the four polypurine duplexes studied. This is the first direct comparison of A-tract monovalent cation interaction affinities between duplexes of DNA and RNA incorporating a full helical turn. It is notable that a single sodium

cation diffused amongst adjacent residues A8, A9, A10 and A11 of the RNA helix; yet, no such long residence and directed diffusion was detected in polyAT DNA (Table 3.2). This could be due to the short size of the 11-mer duplex in which a sufficiently long minor groove is lacking. These findings agree with the Hud-Plavec model of sequence dependent DNA curvature that can influence cationic aggregation such that A-form DNA may have a narrow major groove rich in counterions<sup>33,64</sup>. However, the sequence directed structure of DNA in relation to whether cations influence the groove geometry or vice versa remains a conundrum<sup>65</sup>. In the case of RNA, strong cationic interaction as seen in the trajectories clearly distort the helical structure, making it more compact, further enhancing interaction and reduced charge screening, which leads to the attraction of even more cations.

**Table 3.4:** Occupancy and lifetimes of Na<sup>+</sup> hydrogen bonding to RNA and DNA polypurine

	polyAU RNA		polyAT DNA	
	%	max	%	max
	occ.	occ.	occ.	occ.
		(ps)		(ps)
Major groove				
A3@N <sub>7</sub>	56.0	569	1.9	105
A4@N <sub>7</sub>	35.9	1064	0	0
A6@N <sub>7</sub>	20.2	215	10.9	237
A9@N <sub>7</sub>	21.7	176	17.6	283
Minor groove				
A2@N <sub>1</sub>	1.8	41		
A11@N <sub>3</sub>			5.6	319



**Figure 3.6:** Modeled RNA and DNA helices (a) Polyadenine RNA sequence. (b) Chemical structure of the nucleobase Uracil. (c) Polyadenine DNA sequence. (d) Chemical structure of the nucleobase Thymine.

Analysis of the major and minor grooves by CURVES in Table 3.1 shows that our simulated RNA system agrees closely with a segment of a long viral RNA (PDB 1H1K) polyAU-18<sup>62,66-68</sup>. The large variance in the minor groove depth results from bases protruding beyond the phosphodiester backbone into a concave geometry. Interaction in the minor groove is relatively insignificant, but DNA does show higher residencies than RNA

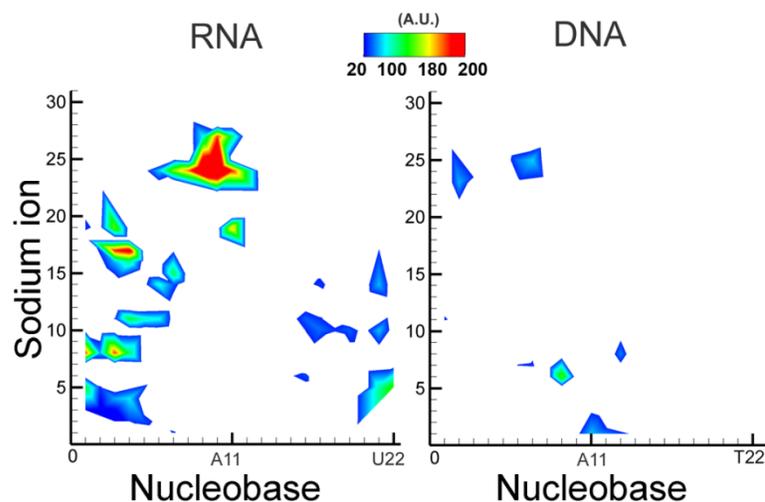
(Table 3.5). This same trend has been noted earlier for polyGC RNA, but is now greatly enhanced. Variances seen in the groove widths may result from cationic interactions inducing compaction and compensate for incomplete hydration; their effects on duplex structure have been examined elsewhere <sup>69-72</sup>. Since this is largely an electrostatic phenomena, the energetics should correlate with residency times in association with specific residues. Indeed, this case is seen for both polyAU RNA and polyAT DNA duplexes. Adenines of RNA have the highest interaction energies with cations at -77 kcal/mol (Table 3.2).

**Table 3.5:** Occupancy and lifetimes of K<sup>+</sup> and Na<sup>+</sup> minor groove binding to identical random sequences of RNA and DNA after truncating the first 2ns and analyzing the remaining 18ns.

	% occu	max occu (ps)
<b>Na<sup>+</sup> in 0.1M NaCl</b>		
<b>RNA</b> A5@N3	0.1	15
<b>DNA</b> T16@O2	7.72	356
<b>A-DNA</b> C15@O2	0.23	24
<b>K<sup>+</sup> in 0.1 M KCl</b>		
<b>RNA</b> C4@O2	0.35	38
<b>DNA</b> T18@O2	2.71	217

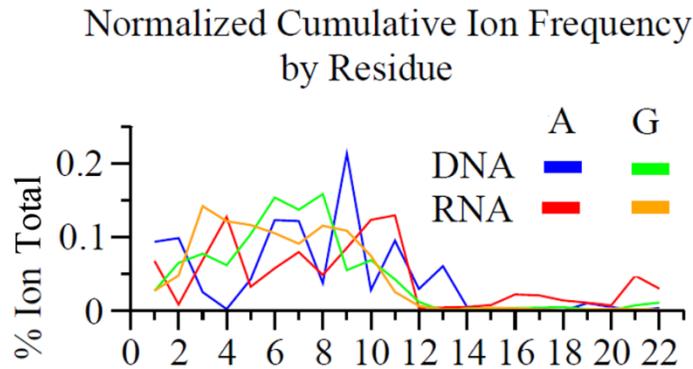
The cumulative ion residency by residue (Figure 3.7) reveals that the polyAU RNA duplex has the highest association with cations. The distributions for all the homopurine systems clearly favor the purine strand. The intermediate accumulations are associated with the polyGC DNA and RNA duplexes. The most unusual feature is that the polyAT DNA has

the least frequent interaction with cations. This may account for the desensitization of cationic control with lengthened polyAT tracts reported in the bacteriophage repressor experiments of Mauro and Koudelka <sup>11</sup>. For polyAT DNA and polyAU RNA duplexes, an additional difference occurs with their complementary bases to adenine. DNA's thymine nucleobase is essentially uracil with a methylated C<sub>5</sub> atom. Biologically, methylation of uracil into thymine protects DNA by rendering it unrecognizable to nucleases. While thymine's conformation is generally constrained within a helix, uracil is free to rotate and can even pair with guanine in loops and knot structures. The polyAT DNA duplex also has other peculiar structural properties which have been previously investigated <sup>73,74</sup>. PolyAT DNA has 10 base pairs per turn which is shorter than the average 10.4 for a typical B-DNA; it is seemingly insensitive to changes in cationic species or salt concentration; it adopts a rigid structure *in vitro*, such that its narrow minor groove assembles a multilayer spine of hydration <sup>30,75-77</sup>. Moreover, it has been shown that polyAT DNA duplexes exhibit higher thermal stability over polyAU RNA <sup>78</sup>. However, it is still unclear if the enhanced stability arises from the differences in helical structure, counterion uptake/release, or states of counterion hydration. It has been postulated that RNA duplexes and polyAT DNA duplexes generally form more intimate contacts with their counterions and have weaker hydration <sup>79</sup>.



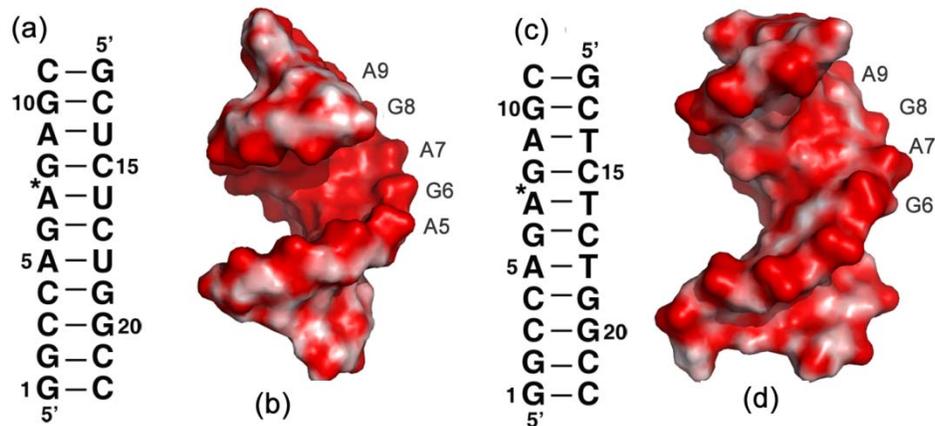
**Figure 3.7:** Heat maps of cumulative diffusive residencies of individually tracked sodium cations within 5 Å of a nucleobase in polyadenine duplexes of RNA and DNA.

We observed that the cationic distributions for all the homopurine systems clearly favor the purine strand comprising of residues 1 to 11 (Figure 3.8). Previous studies, have shown that sequence dependent divalent cationic interactions have been associated with polypurine tracts of both adenine and guanine<sup>55,74,80</sup>. Nastasjivevic et al. observed that a polyadenosine tract of four base pairs was sufficient to give DNA and RNA molecules a strong affinity for  $\text{Ni}^{2+}$ <sup>81</sup>. They suggest that this adenine specificity pointed to an aspect of the nucleic acid secondary or tertiary structure since both adenine and guanine have isosteric  $\text{N}_7$  and  $\text{N}_3$  ring nitrogens. A corollary for monovalent cations has not been established since monovalent cations may also bind to the same divalent cation interacting sequence. These cationic interactions, predominantly electrostatic in nature, may arise more from the geometry of nucleotide sequences than any specific chemical subgroup.



**Figure 3.8:** Cumulative residencies of cations in the cation-duplex modeled systems: Sodium cation diffusive interaction events within 5Å to non alternating tracts of purine nucleotides in RNA and DNA duplexes simulated to 10ns. On the x-axis, residues 1-11 are the purines, and residues 12-22 are the pyrimidines.

### 3.3.4 Helices with random sequences



**Figure 3.9:** Modeled RNA and DNA duplexes (a) RNA random sequence. (b) A-RNA helical structure colored to show electrostatic surface, red is negatively charged. (c) DNA random sequence. (d) B-DNA helical structure for the random sequence; red is negatively charged.

In order to examine the effect of cationic properties such as size and charge density on RNA and DNA duplexes formed by a random sequence, potassium was compared with the sodium simulations. The random sequence is that of a fragment of HIV-TAR helix (PDB ID 397D) possessing prominent G-C steps and an isolated G-G step near the end of the

second strand (Figure 3.9). This location has been shown to be favorable for ion localization<sup>82</sup>. Also included in the sequence is an alternating polypurine tract of GpA steps, previously shown to attract and bind cations<sup>80,83</sup>. The RNA helix (Figure 3.9b) shows the classic A-form structure with a deep and narrow major groove. Areas colored in blue are less electronegative than those of the most electronegative red; white is more neutral. The DNA helix (Figure 3.9d) has a major groove completely open and slightly less electronegative. It is this openness that largely accounts for greater hydration but less cation affinity as compared to RNA. Further probing of both helices with two different cations has provided additional insights into the effect of valency and ionic charge density on these structures.

### **3.3.5. Na<sup>+</sup> and random helices of RNA versus DNA**

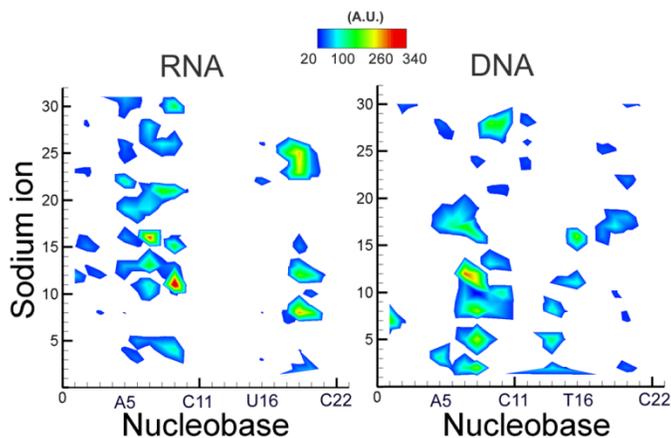
With an ionic radius of 0.95 Å, sodium has a high charge density, low enthalpy of formation, -239.7 kJ/mol, and high enthalpy of hydration, 405 kJ/mol, in aqueous solutions. The differences at the preferred chelation sites of the purine atoms N<sub>7</sub> and O<sub>6</sub> indicate that RNA has significantly longer interaction with the sodium ions (Table 3.6). For both, the A7 and G8 step is a region of high affinity for ions possibly due to its greater flexibility. Surprisingly the occupancy at the O<sub>6</sub> position of G8 is significantly higher with a very long ion residency of 625 ps. Atom N<sub>7</sub> of adenines A7 and A9 are the highest for RNA while G8 and A7 are the highest for DNA. G8 and A7 of DNA is at the edge of helical groove and highly accessible to solvent as well as any cations (Figure 3.9d). The dominance of the adenine N<sub>7</sub> chelation site for RNA remains similar to polyGC and polyAU. Minor groove interaction, again, is more prominent for DNA than for RNA with a significant occupancy for

thymine. The presence of competing interaction sites (O<sub>6</sub> and N<sub>7</sub>) in guanine residues could explain the difference in occupancies of G6 and G8 of RNA (Table 4)<sup>84</sup>. The G6 and G8 in DNA show a remarkably similar occupancy between the N<sub>7</sub> and O<sub>6</sub> positions.

**Table 3.6:** Occupancy and lifetimes of K<sup>+</sup> and Na<sup>+</sup> major groove hydrogen bonding to identical random sequences of duplex RNA and DNA (cut off distance is 3.8 Å for K<sup>+</sup>; 3.2 Å for Na<sup>+</sup>).

	Na <sup>+</sup> in 0.1 M NaCl				K <sup>+</sup> in 0.1 M KCl				Na <sup>+</sup> in 0.1M NaCl	
	RNA		DNA		RNA		DNA		A-DNA	
	% occ.	max occ. (ps)	% occ.	max occ. (ps)	% occ.	max occ. (ps)	% occ.	max occ. (ps)	% occ.	max occ. (ps)
A5@N7	16.8	148	5.3	256	22.5	132	13.4	151	44.5	884
A7@N7	24.5	204	20.1	412	22.6	90	12.6	172	35.6	681
A9@N7	26.7	190	13.97	187	19.59	117	11.0	156	52.3	639
G6@N7	14.2	140	2.9	79	26.4	85	8.9	90	13.3	251
G6@O6	3.1	101	1.8	114	10.5	63	7.4	128	3.3	315
G8@N7	10.1	76	11.8	116	21.4	145	10.4	67	4.7	109
G8@O6	1.9	75	18.0	625	8.6	56	8.5	87	3.6	86

The energies and interaction distances of high residency ions with RNA are greater than that of the DNA (Table 3.2). RNA's A-tract and G19 show the highest interaction energies with their cations while G8 of DNA has the greatest interaction. For DNA, the highest residency ions associated with G8 is due to a favorable geometry and proximity. The lone RNA G19, however, has a longer average interaction distance of 2.657 Å and large interaction energy of -85.566 kcal/mol than that with adenines. As expected, the van der Waals energy is larger for cation-DNA than cation-RNA interactions. The combination of the stronger electrostatic energy and small van der Waals reflects in the localized ion accumulation as seen in Figure 3.10.



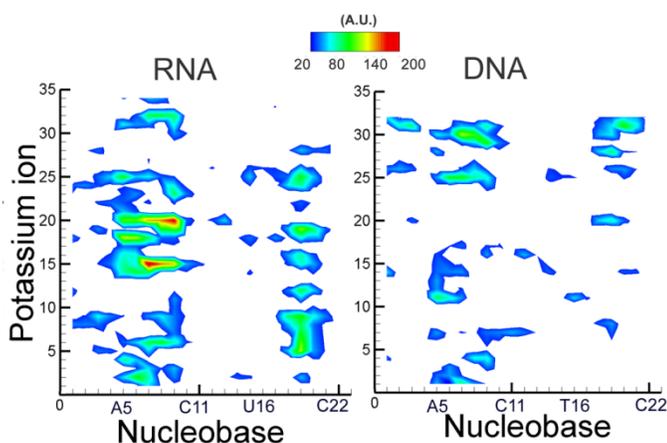
**Figure 3.10:** Heat maps of cumulative diffusive residencies of individually tracked sodium cations within 5 Å of a nucleobase for duplexes of RNA and DNA of identical random sequences.

Measurements of the groove width parameters show that RNA adopts the standard A-form, while DNA adopts the B-form. The major groove of DNA had a less uniform width and depth as evidenced by the large standard deviation (Table 3.1) when compared to RNA. Surprisingly, the cumulative cationic distributions for both RNA and DNA with sodium show a similar pattern in Figure 3.10. The interactions are mainly confined to the interior purine bases and the guanine platform. The greater uniformity of RNA compared to DNA would imply that sodium cations can better shield and compact RNA into a stable geometry.

### 3.3.6 K<sup>+</sup> and random helices of RNA versus DNA

Contrasting with sodium, potassium is larger with a 1.33 Å ionic radius. Consequently, it has a higher enthalpy of formation, -251.2 kJ/mol, and lower enthalpy of hydration, 314 kJ/mol, in aqueous solutions. This contributes to potassium's watershell being more easily stripped than sodium's. Experimentally, Rodgers et al. found that dissociation of nucleobase chelated Na<sup>+</sup> from either adenine or uracil was greater than that

for  $K^+$ <sup>85</sup>. The major groove cationic residency of potassium in the highest occupied purine positions strongly resembles those of sodium (Figure 3.11). The N<sub>7</sub> potassium occupancies in RNA are just as uniform but over twice that in some instances than those of DNA (Table 3.5). Interestingly, the occupancies of the N<sub>7</sub> and O<sub>6</sub> positions in the guanines are nearly coequal. The low but near uniformity of all the high interaction sites in DNA are more attributable to its greater hydration and ease of potassium to lose part of its water shell. Occupancies in the minor groove for DNA are still more prominent than RNA; the occupancy of DNA's O<sub>2</sub> of T18 is comparable to a major groove interaction and almost a full order greater than that of RNA (Table 3.3).



**Figure 3.11:** Heat maps of cumulative diffusive residencies of individually tracked potassium cations within 5 Å of a nucleobase for RNA and DNA duplexes of identical random sequence.

The trend is mirrored in the energetics of nucleobase interactions with long resident ions showing longer interaction distances and lower energies (Table 3.2). The lower charge density of potassium largely accounts for this drop. The energies for RNA ion interaction events remain larger than their DNA counter parts. The guanine ion interaction events in

DNA are comparable to those of adenine in RNA; albeit, they are still lower on average. The major grooves of RNA and DNA have comparable depths under potassium, but RNA is narrower. The major groove of RNA is also less uniform than under sodium; in contrast, DNA's major groove is now more uniform with potassium (Table 3.1). Overall, the helical groove structures of RNA and DNA seem relatively invariant under either sodium or potassium. The pattern of potassium accumulation by residue (Figure 3.11) shows an overlap with the sodium. Potassium interacts less specifically with RNA residues than sodium. The DNA ion accumulation is a muted version of the RNA pattern, showing a similar pattern of affinity for the stretch of alternating purines and the G19/G20 guanine stack.

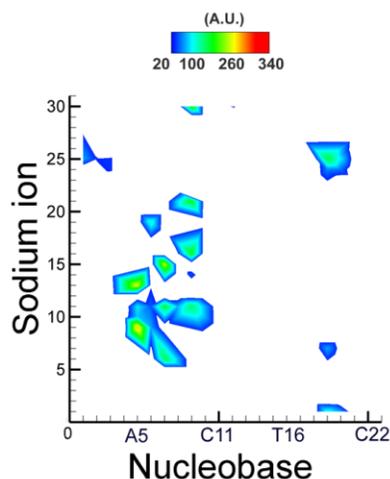
These differences between sodium versus potassium interaction to each of the RNA and DNA duplexes appear relatively minor. Substituting potassium for sodium in RNA duplexes led to a mild disruption of the major groove and more diffusive interaction. The same substitution for DNA duplexes led to an enhanced uniformity of the major groove and more diffusive interaction. However, the difference between the behaviors of each ion with RNA duplexes versus DNA duplexes was more substantial.

### **3.3.7 Effect of geometry: Fixed A-form random helical DNA**

To uncouple the effect of helical geometry from the sequence we examined A-form RNA and DNA helices. The groove widths of A-DNA (Table 3.1) show a minor groove comparable to RNA with a deep and narrow major groove. The depth of the major groove of

10.49 Å is comparable to RNA's 11.78 Å. These groove parameters are also very similar to the measurements of the X-ray crystal structures of 1QCU and 1H1K<sup>62,86</sup>.

An indication that the A-DNA conformation has a dramatic effect on ion dynamics is shown in the hydrogen bonding statistics (Table 3.5). For position N<sub>7</sub> of the purine tract comprised of A5 thru A9, the occupancies are much higher for A-DNA's adenines as compared to either the RNA or DNA duplexes. The longest residency times are also higher and reflect strong interactions, which imply a slow exchange of cations with the N<sub>7</sub> chelation site. For comparison, the N<sub>7</sub> site of adenine A7 in RNA experiences 24.53% occupancy with the longest period of interaction at 204 ps, whereas in A-DNA, the same site is 35.62% occupied, but the longest interaction event is over quadruple, 884 ps (Table 3.5). This is also much higher than the DNA duplex's 412 ps maximum interaction time. A-DNA's longest interaction events are mostly higher than either RNA or DNA implying that there is low exchange and competition for a given site. Minor groove interaction events mimic RNA more closely than DNA with insignificant occupancies (Table 3.3). Figure 9 shows that the distribution of interaction is clearly more localized for A-form DNA helix than either RNA or DNA and a clustering toward the polypurine tract of A5 thru G10. The interaction with the G19/G20 platform is also visible.

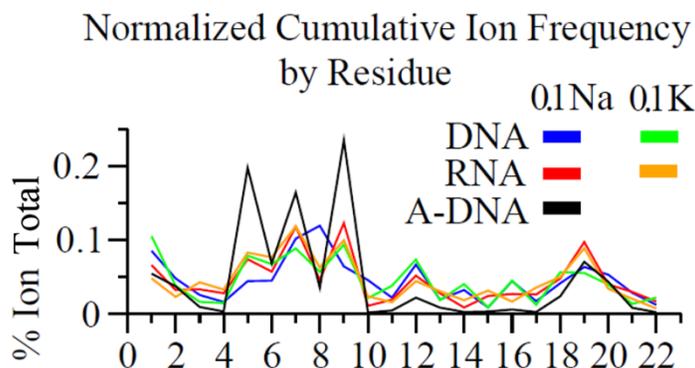


**Figure 3.12:** Heat map of cumulative diffusive residencies of individually tracked sodium cations within 5 Å of a nucleobase for A-DNA of a random sequence.

The electrostatic interaction of sodium cations with A-DNA shows a greater resemblance to RNA duplexes than that of DNA duplexes. Although position A5 and A9 display diffusive ion interaction, the energies are in the range of -71.275 kcal/mol to -73.049 kcal/mol which is nearly identical to the range of RNA-ion interactions ranging from -71.615 kcal/mol to -73.739 kcal/mol (Table 3.2). The interaction energy between G19 and sodium is -85.566 kcal/mol for RNA and -84.168 kcal/mol for A-DNA. When compared with each other, the cumulative ion occupancies of the duplexes display progressive cation localizations toward greater purine affinity in the following order: A-DNA > RNA > DNA (Figure 3.13).

Previous modeling by Lee et al.<sup>87</sup> also suggest that specific motions may enhance the effect of helical geometry, but it is still modulated by various factors such as the electrolytic environment. Those results complement those obtained by Mazur et al. where interactions of cation interaction to the minor groove of the Dickerson-Drew dodecamer appeared to be

more structure specific than sequence specific<sup>88</sup>. The results from our computational experiment also strongly implicate geometry as a major determinate of cationic interactions with duplexes.



**Figure 3.13:** Cumulative residencies of cations in the cation-duplex modeled systems: Sodium and potassium ion interacting within 5 Å of a nucleobase in random sequence duplexes of RNA and DNA. On the x-axis, residues 1-11 are the purines, and residues 12-22 are the pyrimidines.

### 3.4. Conclusions

We investigated monovalent cationic interactions with canonical helical duplexes of RNA and DNA in explicit solvent MD simulations; we observed that monovalent cations alone alter less drastically the helical conformation of RNA and DNA duplexes at low physiologic concentration. A greater variance in cationic interactions occurred between specific duplex conformations: PolyAT DNA duplexes were the weakest in interacting with sodium cations; the strongest interaction was observed with polyGC RNA. For helices of a random sequence, these specific interactions occurred in an alternating polypurine tract. Interaction energies of sodium with a fixed A-form DNA agreed more closely with A-form

RNA than B-form DNA. The chemical makeup differentiating RNA from DNA alone does not fully characterize these phenomena arising from their allowed conformations. Based on these observations, helical conformation exerts a greater influence on cationic interactions around helical duplexes and better explains diffusive interaction events of high residency cations.

### **Acknowledgments**

This work was supported by grant CMMI-1150682 from the National Science Foundation. The computer support was provided by the High Performance Computing Center at North Carolina State University.

## References

1. Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through RNA structure. *Nat Rev Genet* 2011;12(9):641-655.
2. Jose D, Porschke D. Dynamics of the B-A transition of DNA double helices. *Nucleic Acids Research* 2004;32(7):2251-2258.
3. Pastor N. The B- to A-DNA Transition and the Reorganization of Solvent at the DNA Surface. *Biophysical Journal* 2005;88(5):3262-3275.
4. Hackl EV, Blagoi YP. Effect of ethanol on structural transitions of DNA and polyphosphates under Ca<sup>2+</sup> ions action in mixed solutions. *Acta Biochimica Polonica* 2000;47(1):103-112.
5. Rohs R, Jin XS, West SM, Joshi R, Honig B, Mann RS. Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem* 2010;79:233-269.
6. Chou SH, Chin KH, Wang AHJ. Unusual DNA duplex and hairpin motifs. *Nucleic Acids Res* 2003;31(10):2461-2474.
7. Maehigashi T, Hsiao C, Kruger Woods K, Moulai T, Hud NV, Dean Williams L. B-DNA structure is intrinsically polymorphic: even at the level of base pair positions.
8. Hunt RA, Munde M, Kumar A, Ismail MA, Farahat AA, Arafa RK, Say M, Batista-Parra A, Tevis D, Boykin DW, Wilson WD. Induced topological changes in DNA complexes: influence of DNA sequences and small molecule structures. *Nucleic Acids Res* 2011;39(10):4265-4274.
9. Moulai T, Maehigashi T, Lountos GT, Komeda S, Watkins D, Stone MP, Marky LA, Li JS, Gold B, Williams LD. Structure of B-DNA with cations tethered in the major groove. *Biochemistry* 2005;44(20):7458-7468.
10. Hase CC, Mekalanos JJ. Effects of changes in membrane sodium flux on virulence gene expression in *Vibrio cholerae*. *P Natl Acad Sci USA* 1999;96(6):3183-3187.
11. Mauro SA, Koudelka GB. Monovalent cations regulate DNA sequence recognition by 434 repressor. *Journal of Molecular Biology* 2004;340(3):445-457.
12. Landt SG, Tipton AR, Frankel AD. Localized influence of 2'-hydroxyl groups and helix geometry on protein recognition in the RNA major groove. *Biochemistry* 2005;44(17):6547-6558.

13. Carlson CB, Stephens OM, Beal PA. Recognition of double-stranded RNA by proteins and small molecules. *Biopolymers* 2003;70(1):86-102.
14. Leung DW, Basler CF, Amarasinghe GK. Molecular mechanisms of viral inhibitors of RIG-I-like receptors. *Trends Microbiol* 2012;20(3):139-146.
15. Pascal JM. DNA and RNA ligases: structural variations and shared mechanisms. *Current Opinion in Structural Biology* 2008;18(1):96-105.
16. Pascal JM, O'Brien PJ, Tomkinson AE, Ellenberger T. Human DNA ligase I completely encircles and partially unwinds nicked DNA. *Nature* 2004;432(7016):473-478.
17. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein-DNA recognition. *Nature* 2009;461(7268):1248-U1281.
18. Rohs R, West SM, Liu P, Honig B. Nuance in the double-helix and its role in protein-DNA recognition. *Curr Opin Struc Biol* 2009;19(2):171-177.
19. Hobaika Z, Zargarian L, Boulard Y, Maroun RG, Mauffret O, Femandjian S. Specificity of LTR DNA recognition by a peptide mimicking the HIV-1 integrase alpha 4 helix. *Nucleic Acids Res* 2009;37(22):7691-7700.
20. Howerton SB, Sines CC, VanDerveer D, Williams LD. Locating monovalent cations in the grooves of B-DNA. *Biochemistry* 2001;40(34):10023-10031.
21. Lu K, Miyazaki Y, Summers MF. Isotope labeling strategies for NMR studies of RNA. *Journal of Biomolecular Nmr* 2010;46(1):113-125.
22. Basu S, Szewczak AA, Cocco M, Strobel SA. Direct detection of monovalent metal ion binding to a DNA G-quartet by T1-205 NMR. *Journal of the American Chemical Society* 2000;122(13):3240-3241.
23. Sponer J, Cang X, Cheatham Iii TE. Molecular dynamics simulations of G-DNA and perspectives on the simulation of nucleic acid structures. *Methods* 2012.
24. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. *Bmc Biol* 2011;9.
25. Sim AYL, Minary P, Levitt M. Modeling nucleic acids. *Curr Opin Struc Biol* 2012;22(3):273-278.
26. Perez A, Luque FJ, Orozco M. Frontiers in Molecular Dynamics Simulations of DNA. *Accounts Chem Res* 2012;45(2):196-205.

27. Auffinger P, Leontis N, Westhof E. Ions in Molecular Dynamics Simulations of RNA Systems. In: Bujnicki JM, editor. RNA 3D Structure Analysis and Prediction. Volume 27, Nucleic Acids and Molecular Biology: Springer Berlin Heidelberg; 2012. p 299-318.
28. Draper D. A guide to ions and RNA structure. RNA 2004;10(3):335-343.
29. Cheng YH, Korolev N, Nordenskiöld L. Similarities and differences in interaction of K<sup>+</sup> and Na<sup>+</sup> with condensed ordered DNA. A molecular dynamics computer simulation study. Nucleic Acids Research 2006;34(2):686-696.
30. Beveridge DL, Dixit SB, Barreiro G, Thayer KM. Molecular dynamics simulations of DNA curvature and flexibility: Helix phasing and premelting. Biopolymers 2004;73(3):380-403.
31. Auffinger P, Westhof E. Water and ion binding around RNA and DNA (C,G) oligomers. Journal of Molecular Biology 2000;300(5):1113-1131.
32. Auffinger P, Westhof E. Water and ion binding around r(UpA)(12) and d(TpA)(12) oligomers - Comparison with RNA and DNA (CpG)(12) duplexes. Journal of Molecular Biology 2001;305(5):1057-1072.
33. Hud NV, Polak M. DNA-cation interactions: the major and minor grooves are flexible ionophores. Current Opinion in Structural Biology 2001;11(3):293-301.
34. Feig M, Pettitt BM. Sodium and chlorine ions as part of the DNA solvation shell. Biophysical Journal 1999;77(4):1769-1781.
35. D. A. Case TAD, Cheatham, C. L. Simmerling, J. Wang, R. E., Duke RL, K. M. Merz, D. A. Pearlman, M. Crowley, R. C. Walker,, W. Zhang BW, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong,, F. Paesani XW, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan,, J. Mongan VH, G. Cui, P. Beroza, D. H. Mathews,, C. Schafmeister WSR, and P. A. Kollman. . AMBER 9. University of California: San Francisco 2006.
36. Ippolito JA, Steitz TA. A 1.3-Å resolution crystal structure of the HIV-1 trans-activation response region RNA stem reveals a metal ion-dependent bulge conformation. Proc Natl Acad Sci U S A 1998;95(17):9819-9824.
37. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. J Am Chem Soc 1995;117(19):5179-5197.

38. Cieplak P, Caldwell J, Kollman P. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: Aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J Comput Chem* 2001;22(10):1048-1057.
39. Chen AA, Pappu RV. Parameters of monovalent ions in the AMBER-99 forcefield: Assessment of inaccuracies and proposed improvements. *Journal of Physical Chemistry B* 2007;111(41):11884-11887.
40. Joung IS, Cheatham TE. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *Journal of Physical Chemistry B* 2008;112(30):9020-9041.
41. Joung IS, Cheatham TE. Molecular Dynamics Simulations of the Dynamic and Energetic Properties of Alkali and Halide Ions Using Water-Model-Specific Ion Parameters. *Journal of Physical Chemistry B* 2009;113(40):13279-13290.
42. Zhang C, Raugei S, Eisenberg B, Carloni P. Molecular Dynamics in Physiological Solutions: Force Fields, Alkali Metal Ions, and Ionic Strength. *J Chem Theory Comput* 2010;6(7):2167-2175.
43. Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE, Laughton CA, Orozco M. Refinement of the AMBER force field for nucleic acids: Improving the description of alpha/gamma conformers. *Biophysical Journal* 2007;92(11):3817-3829.
44. Lankas F, Spackova N, Moakher M, Enkhbayar P, Sponer J. A measure of bending in nucleic acids structures applied to A-tract DNA. *Nucleic Acids Res* 2010;38(10):3414-3422.
45. Yildirim I, Stern HA, Tubbs JD, Kennedy SD, Turner DH. Benchmarking AMBER Force Fields for RNA: Comparisons to NMR Spectra for Single-Stranded r(GACC) Are Improved by Revised chi Torsions. *J Phys Chem B* 2011;115(29):9261-9270.
46. Yoo JJ, Aksimentiev A. Improved Parametrization of Li<sup>+</sup>, Na<sup>+</sup>, K<sup>+</sup>, and Mg<sup>2+</sup> Ions for All-Atom Molecular Dynamics Simulations of Nucleic Acid Systems. *J Phys Chem Lett* 2012;3(1):45-50.
47. Bomble YJ, Case DA. Multiscale modeling of nucleic acids: Insights into DNA flexibility. *Biopolymers* 2008;89(9):722-731.
48. Sorin EJ, Engelhardt MA, Herschlag D, Pande VS. RNA simulations: Probing hairpin unfolding and the dynamics of a GNRA tetraloop. *Journal of Molecular Biology* 2002;317(4):493-506.

49. Ding S, Kropachev K, Cai YQ, Kolbanovskiy M, Durandina SA, Liu Z, Shafirovich V, Brody S, Geacintov NE. Structural, energetic and dynamic properties of guanine(C8)-thymine(N3) cross-links in DNA provide insights on susceptibility to nucleotide excision repair. *Nucleic Acids Res* 2012;40(6):2506-2517.
50. Besseova I, Otyepka M, Reblova K, Sponer J. Dependence of A-RNA simulations on the choice of the force field and salt strength. *Phys Chem Chem Phys* 2009;11(45):10701-10711.
51. William L. Jorgensen JC, and Jeffrey D. Madura. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;79.
52. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* 1995;103(19):8577-8593.
53. Berendsen HJC, Postma JPM, Gunsteren WFv, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* 1984;81(8):3684-3690.
54. Ryckaert JP, Ciccotti, G, Berendsen, H.J.C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* 1977;23(3):327-341.
55. Sethaphong L, Singh A, Marlowe AE, Yingling YG. The Sequence of HIV-1 TAR RNA Helix Controls Cationic Distribution. *Journal of Physical Chemistry C* 2010;114(12):5506-5512.
56. Lavery R, Sklenar H. Defining the Structure of Irregular Nucleic-Acids - Conventions and Principles. *Journal of Biomolecular Structure & Dynamics* 1989;6(4):655-667.
57. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: Application to microtubules and the ribosome. *P Natl Acad Sci USA* 2001;98(18):10037-10041.
58. Rustighi A, Tessari MA, Vascotto F, Sgarra R, Giancotti V, Manfioletti G. A polypyrimidine/polypurine tract within the Hmga2 minimal promoter: A common feature of many growth-related genes. *Biochemistry* 2002;41(4):1229-1240.
59. Anderson JD, Widom J. Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol Cell Biol* 2001;21(11):3830-3839.
60. Ercan S, Lieb JD. New evidence that DNA encodes its packaging. *Nat Genet* 2006;38(10):1104-1105.

61. Michiel van den Hout IDV, Susanne Hage, Nynke H. Dekker. Direct Force Measurements on Double-Stranded RNA in Solid State Nanopores. *NANO Letters* 2010;10:701-707.
62. Klosterman PS, Shah SA, Steitz TA. Crystal structures of two plasmid copy control related RNA duplexes: An 18 base pair duplex at 1.20 Å resolution and a 19 base pair duplex at 1.55 Å resolution. *Biochemistry-U.S.* 1999;38(45):14784-14792.
63. Nakano S, Fujimoto M, Hara H, Sugimoto N. Nucleic acid duplex stability: influence of base composition on cation effects. *Nucleic Acids Research* 1999;27(14):2957-2965.
64. Hud NV, Plavec J. A unified model for the origin of DNA sequence-directed curvature. *Biopolymers* 2003;69(1):144-159.
65. Justin P. Peters LJM. DNA Curvature and Flexibility in vitro and in vivo. *Quarterly Reviews of Biophysics* 2010.
66. Priyakumar UD, MacKerell AD. Computational approaches for investigating base flipping in oligonucleotides. *Chem Rev* 2006;106(2):489-505.
67. Trikha J, Filman DJ, Hogle JM. Crystal structure of a 14 bp RNA duplex with non-symmetrical tandem G center dot U wobble base pairs. *Nucleic Acids Research* 1999;27(7):1728-1739.
68. Tanaka Y, Fujii S, Hiroaki H, Sakata T, Tanaka T, Uesugi S, Tomita K, Kyogoku Y. A  $\lambda$ -form RNA double helix in the single crystal structure of r(UGAGCUUCGGCUC). *Nucleic Acids Research* 1999;27(4):949-955.
69. Chen YZ, Prohofsky EW. The Role of a Minor Groove Spine of Hydration in Stabilizing Poly(Da) Poly(Dt) against Fluctuational Interbase H-Bond Disruption in the Premelting Temperature Regime. *Nucleic Acids Research* 1992;20(3):415-419.
70. Buckin V, Tran H, Morozov V, Marky LA. Hydration effects accompanying the substitution of counterions in the ionic atmosphere of poly(rA)center dot poly(rU) and poly(rA)center dot 2poly(rU) helices. *Journal of the American Chemical Society* 1996;118(30):7033-7039.
71. Siegfried NA, Kierzek R, Bevilacqua PC. Role of Unsatisfied Hydrogen Bond Acceptors in RNA Energetics and Specificity. *Journal of the American Chemical Society* 2010;132(15):5342-+.
72. Brovchenko I, Krukau A, Oleinikova A, Mazur AK. Ion dynamics and water percolation effects in DNA polymorphism. *Journal of the American Chemical Society* 2008;130(1):121-131.

73. Fritsch V, Westhof E. 3-Center Hydrogen-Bonds in DNA - Molecular-Dynamics of Poly(Da).Poly(Dt). *Journal of the American Chemical Society* 1991;113(22):8271-8277.
74. Cote ML, Pflomm M, Georgiadis MM. Staying straight with A-tracts: A DNA analog of the HIV-1 polypurine tract. *Journal of Molecular Biology* 2003;330(1):57-74.
75. McConnell KJ, Beveridge DL. Molecular dynamics simulations of B'-DNA: Sequence effects on A-tract-induced bending and flexibility. *Journal of Molecular Biology* 2001;314(1):23-40.
76. Alexeev DG, Lipanov AA, Skuratovskii IY. Poly(Da).Poly(Dt) Is a B-Type Double Helix with a Distinctively Narrow Minor Groove. *Nature* 1987;325(6107):821-823.
77. Aymami J, Coll M, Frederick CA, Wang AHJ, Rich A. The Propeller DNA Conformation of Poly(Da).Poly(Dt). *Nucleic Acids Research* 1989;17(8):3229-3245.
78. Tikhomirova A, Taulier N, Chalikian TV. Energetics of nucleic acid stability: The effect of Delta C-P. *J Am Chem Soc* 2004;126(50):16387-16394.
79. Tikhomirova A, Chalikian TV. Probing hydration of monovalent cations condensed around polymeric nucleic acids. *J Mol Biol* 2004;341(2):551-563.
80. Timsit Y, Bombard S. The 1.3 angstrom resolution structure of the RNA tridecamer r(GCGUUUGAAACGC): Metal ion binding correlates with base unstacking and groove contraction. *Rna-a Publication of the Rna Society* 2007;13(12):2098-2107.
81. Nastasijevic B, Becker NA, Wurster SE, Maher LJ. Sequence-specific binding of DNA and RNA to immobilized nickel ions. *Biochemical and Biophysical Research Communications* 2008;366(2):420-425.
82. Wild K, Weichenrieder O, Leonard GA, Cusack S. The 2 angstrom structure of helix 6 of the human signal recognition particle RNA. *Struct Fold Des* 1999;7(11):1345-1352.
83. Anastassopoulou J. Metal-DNA interactions. *J Mol Struct* 2003;651:19-26.
84. Burda JV, Sponer J, Hobza P. Ab Initio study of the interaction of guanine and adenine with various mono- and bivalent metal cations (Li<sup>+</sup>, Na<sup>+</sup>, K<sup>+</sup>, Rb<sup>+</sup>, Cs<sup>+</sup>; Cu<sup>+</sup>, Ag<sup>+</sup>, Au<sup>+</sup>; Mg<sup>2+</sup>, Ca<sup>2+</sup>, Sr<sup>2+</sup>, Ba<sup>2+</sup>; Zn<sup>2+</sup>, Cd<sup>2+</sup>, and Hg<sup>2+</sup>). *J Phys Chem-U*s 1996;100(17):7250-7255.
85. Rodgers MT, Armentrout PB. Noncovalent interactions of nucleic acid bases (uracil, thymine, and adenine) with alkali metal ions. Threshold collision-induced dissociation and theoretical studies. *J Am Chem Soc* 2000;122(35):8548-8558.

86. Diprose JM, Grimes JM, Sutton GC, Burroughs JN, Meyer A, Maan S, Mertens PPC, Stuart DI. The core of bluetongue virus binds double-stranded RNA. *Journal of Virology* 2002;76(18):9533-9536.
87. Lee DJ, Wynveen A, Kornyshev AA, Leikin S. Undulations Enhance the Effect of Helical Structure on DNA Interactions. *J Phys Chem B* 2010;114(35):11668-11680.
88. Mazur AK. DNA dynamics in a water drop without counterions. *J Am Chem Soc* 2002;124(49):14707-14715.

# Chapter 4

## **A Tertiary Structure of a Cellulose Synthase**

Published in Proceedings of the National Academy of Sciences (2013)

### **A Tertiary Structure of a Cellulose Synthase**

Latsavongsakda Sethaphong<sup>1</sup>, Candace H. Haigler<sup>2</sup>, James D. Kubicki<sup>3</sup>, Jochen Zimmer<sup>4</sup>,  
Dario Bonetta<sup>5</sup>, Seth DeBolt<sup>6</sup>, Yaroslava G. Yingling<sup>1\*</sup>

<sup>1</sup>911 Partners Way, Materials Science and Engineering, North Carolina State University, Raleigh, NC 27695

<sup>2</sup> Dept. of Crop Science and Dept. of Plant Biology, Campus Box 7620, North Carolina State University, Raleigh, NC 27695

<sup>3</sup> Dept. of Geosciences and the Earth & Environmental Systems Institute, The Pennsylvania State University, University Park PA 16802

<sup>4</sup>Center for Membrane Biology, Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908

<sup>5</sup> Faculty of Science, University of Ontario Institute of Technology, Oshawa, ON Canada

<sup>6</sup> Department of Horticulture, University of Kentucky, Lexington, KY 40546

#### **Abstract**

A three-dimensional atomistic model of a plant cellulose synthase (CESA) has remained elusive despite over forty years of experimental effort. Here we report a computationally predicted three-dimensional structure of 506 amino acids of cotton CESA within the cytosolic region. Comparison of the predicted plant CESA structure with the solved structure of a bacterial cellulose-synthesizing protein validates the overall fold of the modeled glycosyltransferase (GT) domain. The co-aligned plant and bacterial GT domains share a six-stranded  $\beta$ -sheet, five  $\alpha$ -helices, and conserved motifs similar to those required for catalysis in other GT-2 glycosyltransferases. Extending beyond the cross-kingdom similarities related to cellulose polymerization, the predicted structure of cotton CESA reveals that plant specific modules ('plant conserved region', P-CR, and 'class specific

region', CSR) fold into distinct subdomains on the periphery of the catalytic region. Computational results support the importance of the P-CR and/or CSR in CESA oligomerization to form the multimeric cellulose-synthesis complexes that are characteristic of plants. Relatively high sequence conservation between plant CESAs allowed mapping of known mutations and two novel mutations that perturb cellulose synthesis in *Arabidopsis thaliana* to their analogous positions in the modeled structure. Most of these mutations sites are near the predicted catalytic region, and the confluence of other mutation sites supports the existence of previously undefined functional nodes within the catalytic core of CESA. Overall, the predicted tertiary structure provides a platform for the biochemical engineering of plant CESAs.

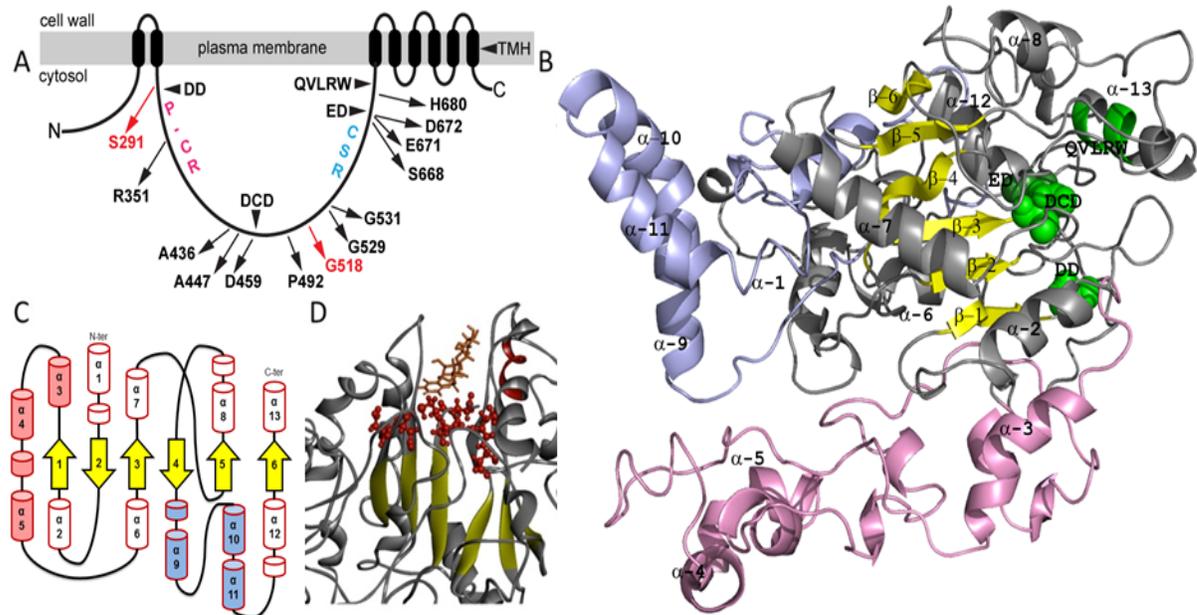
#### **4.1 Introduction**

Cellulose fibrils within plant cell walls provide the foundation for plant structure and are renewable biomaterials that account for most of the world's biomass. Despite the importance of plant cellulose to nature and industry, we have little insight into the three dimensional structure of proteins required for plant cellulose biosynthesis. This deficiency arose due to experimental barriers in purification of active enzyme, recombinant expression, and crystallization of any plant cellulose synthase (CESA). However, manipulating the physical properties of cellulose through biochemical engineering of CESA structure offers many prospects for improved biomaterials. For example, moderate reduction of cellulose crystallinity increases the efficiency of saccharification<sup>1</sup>, a process important for biofuels production from lignocellulosic biomass. However, the capacity for directed enzyme design requires an understanding of CESA protein structure/function relationships.

CESA is a membrane-bound Glycosyltransferase Family 2 (GT-2) enzyme <sup>2</sup> that catalyzes  $\beta$ -1,4-glucan (cellulose) chain polymerization using UDP-glucose as substrate <sup>3</sup>. Although CESA proteins typically arrange themselves into multimeric cellulose synthase complexes (CSC), which are required for the production of multi-chain cellulose microfibrils, the CSCs of land plants and related algae are uniquely organized as six-lobed circular “rosettes” containing a still-undefined number (e.g. 18-36) of CESAs <sup>3</sup>. In contrast, bacteria, other algae, and tunicates have linear CSCs that correlate with synthesis of cellulose fibrils with different physical structures <sup>4</sup>. Accordingly, there are differences in CSC organization and the resulting properties of cellulose fibrils between, for example, bacteria and plants.

Plant CESA has a transmembrane region with eight predicted transmembrane helices (TMH) and a large (~500 amino acid) cytosolic region. The cytosolic region of plant CESAs has four characteristic conserved motifs containing DD, DCD, ED and QVLRW residues <sup>3,5</sup> (**Figure 4.1A**) that were predicted to be involved in substrate and/or acceptor binding, a “plant conserved region” (P-CR), and a “class specific region” (CSR). For GhCESA1 from *Gossypium hirsutum* (cotton), deletion of the first conserved region containing DD abolished UDP-glucose binding *in vitro* <sup>6</sup>, and four missense mutations causing cellulose deficiency occur in the conserved DCD or ED residues of CESAs in the model plant *Arabidopsis thaliana* (called hereafter *Arabidopsis*) <sup>7-9</sup>. A few amino acids (D, D, D, QxxRW) within the plant CESA conserved motifs are more broadly conserved and required for catalysis in other GT-2 enzymes such as hyaluronan and chitin synthases <sup>10-12</sup>. In GT-2 enzymes a conserved DxD motif is usually part of a GT-A fold, as shown in solved structures of SpsA from

*Bacillus subtilis*<sup>13</sup>, chondroitin polymerase from *Escherichia coli* K4<sup>14</sup> and most recently *Rhodobacter sphaeroides* cellulose synthase (BcsA)<sup>15</sup>. Plant CesAs will likely to have a similar fold due to conservation of the cellulose polymerization mechanism, but experimental evidence is lacking.



**Figure 4.1** Predicted structure of the large cytosolic region of GhCESA1. (A) Diagram of GhCESA1 showing 8 predicted TMH and the large cytosolic loop between TMH2 and TMH3. Labels within the cytosolic loop indicate: the approximate locations of the four conserved motifs; the P-CR region; the CSR region; and the analogous locations for missense mutations in *Arabidopsis* CESAs that perturb cellulose synthesis (black and red type for published or newly reported mutations, respectively). (B) Snapshot of the Gh506 structure. The catalytic core is grey, the P-CR is pink, and the CSR is light blue. The catalytic core contains a  $\beta$ -sheet with six strands in yellow ( $\beta$ -1 S287-S291;  $\beta$ -2 D253-S257;  $\beta$ -3 F454-D459;  $\beta$ -4 C532-N535;  $\beta$ -5 Y488-F491; and  $\beta$ -6 S686-C689). Green highlights DD, DCD, ED (directly behind DCD), and the position of QVLRW within core  $\alpha$ -13. The five  $\alpha$ -helices that are part of the GT core are labeled: [ $\alpha$ -2 L267-A278;  $\alpha$ -6 H433-V448;  $\alpha$ -7 N466-D479;  $\alpha$ -8 N508-K517; and  $\alpha$ -13 S705-R725 (containing QVLRW)]. All predicted  $\alpha$ -helices are numbered in order within the primary sequence. (C) Diagram of the secondary structure showing a total of 6  $\beta$ -strands (yellow arrows) and 13 major  $\alpha$ -helices (shown as barrels) in three regions: catalytic core (red outlines); P-CR (pink fill); and CSR (blue fill). Possible additional shorter helical regions are indicated as unnumbered small barrels. (D) UDP-Glc docked into the catalytic site above DCD.

In contrast to bacteria, the plant CESA cytosolic region contains large insertions specific to plants only, namely the P-CR and the CSR<sup>5,6,16</sup>. Although their exact functions are unknown, the CSR and P-CR are hypothesized to mediate aspects of cellulose synthesis unique to plants, such as the formation of rosette-like CSCs that move through the plasma membrane producing cellulose fibrils through the coupling of  $\beta$ -1,4-glucan polymerization and crystallization<sup>1,17,18</sup>. However, no insight into the structure, folding, and putative role in CSC assembly of the plant-specific CESA regions has been reported.

To fill in the gaps in our understanding about the tertiary structure of plant CESAs, we generated a model of the three dimensional structure of 506 amino acids from a cytosolic region of cotton GhCESA1 (GenBank Accession P93155)<sup>6</sup>, called hereafter the ‘Gh506’ structure. GhCESA1 is an apparent ortholog of AtCESA8 from Arabidopsis, and its gene is highly expressed during cotton fiber secondary wall thickening<sup>6,19</sup>. Structural co-alignment of selected regions of the recently solved bacterial CESA (BcsA)<sup>15</sup> with the plant Gh506 model revealed numerous structural commonalities within the GT-2 domains despite poor sequence similarity over the cytosolic region. This result supports the veracity of the plant CESA model, given that BcsA was not used as homolog for Gh506 prediction. Moreover, the Gh506 structure reveals how plant specific P-CR and CSR domains are interfaced with the GT domain, and showed possibilities for how they may participate in CESA oligomerization to generate plant-specific CSCs.

Taking advantage of the high conservation between seed plant CESA sequences, we mapped Arabidopsis CESA missense mutations that alter cellular morphogenesis via effects on cellulose synthesis onto the Gh506 structure. The confluence of some of the point mutations

allows us to propose the existence of previously unidentified functional nodes within CESA. These new insights into structure/function relationships in plant CESAs may have importance for optimization of the properties of renewable biomass.

## **4.2 Results and Discussion**

### **4.2.1 An in silico predicted structure of the GhCESA1 cytosolic region**

A rough initial model of 506 amino acids of the GhCESA1 cytosolic region (**Figure 4.1A, Figure 4S.1**) was generated by the SAM-T08 server using twenty solved protein structures (**Figure 4S.3A**). The template structures were selected via multipass Blastp search for putative homologs in the NCBI non-redundant protein database (**Table 4S.1, Figure 4S.2**). Two of the top selected structures were from the bacterial protein templates of SpsA and K4CP that have been extensively used to examine the molecular basis for catalysis and substrate recognition of glycosyltransferases<sup>13,14</sup>. Note that recently solved structure of BscA was not included in the prediction of Gh506, as it was not available at the time. After refinement with MD simulations, the Gh506 structure (**Figure 4.1B, Figure 4S.3B, Figure 4S.7**) had a Pro-SA Z score of -6.09 and an ERRAT2 quality factor of 86.9 %, which is the percentage of the protein where the calculated conformational error falls below the 95% rejection limit. The overall quality of the Gh506 structure is consistent with solved structures of three other GT-2 enzymes obtained from crystallography with 2 Å resolution (**Table 4S.2, Figure 4S.4, Figure 4S.5**). The regions with conformation errors either have high local mobility or are deeply buried. Similar difficulties in full refinement arise for some regions within solved crystal structures (**Figure 4S.4B**).

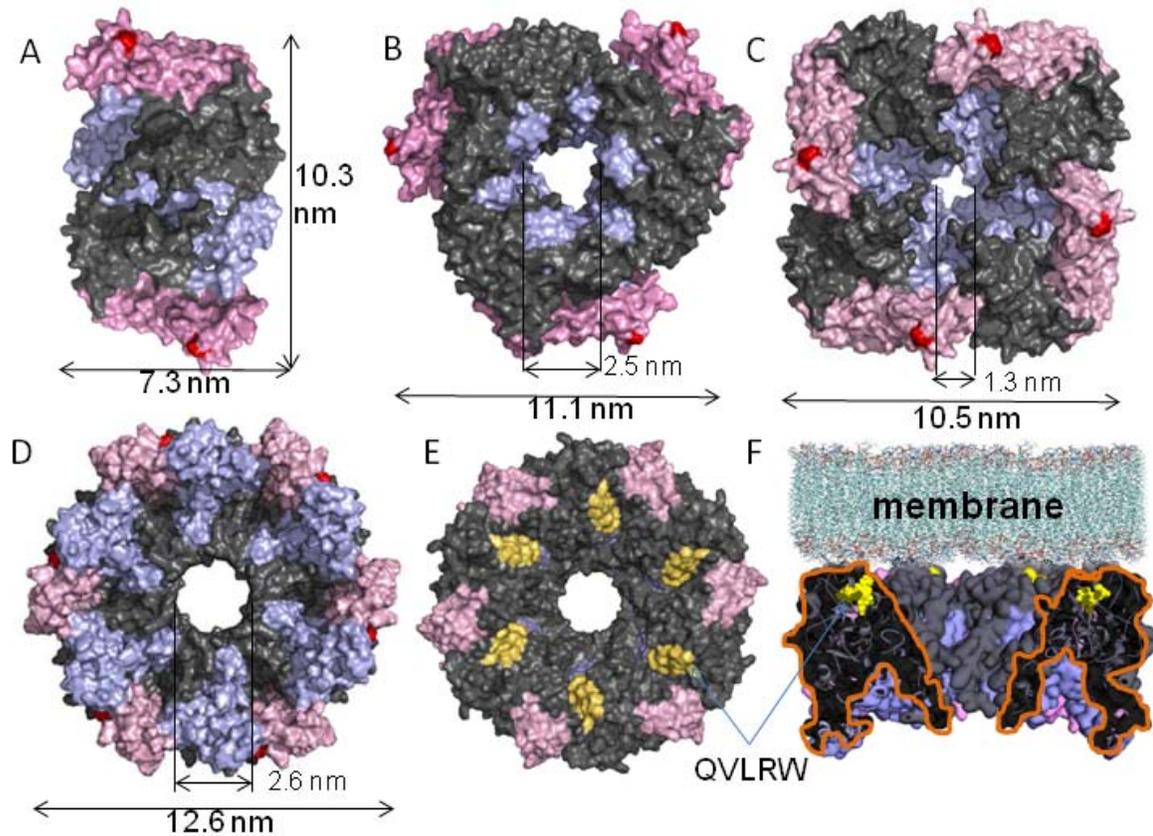
The Gh506 structure contains 13  $\alpha$ -helices and 6  $\beta$ -strands, that are organized into a  $\beta$ -sheet near the catalytic site where UDP-glucose binds, forming a GT-A domain with a canonical Rossmann fold (**Figure 4.1C,D, Figure 4S.2G; Table 4S.3**) similar to bacterial GT-2 enzymes, such as SpsA and BcsA<sup>13,15,20</sup>. In this core GT-2 domain, the structural elements include five core  $\alpha$ -helices ( $\alpha$ -2,-6,-7,-8,-13) and the  $\beta$ -sheet (6  $\beta$ -strands) that helps to stabilize the catalytic residues (**Figure 4S.2; Table 4S.3**). The catalytic pocket of the Gh506 structure contained the closely arranged conserved motifs. The QVLRW motif might interact with the newly polymerized cellulose with its tryptophan residue<sup>21</sup>, and it is located in the center of  $\alpha$ -13 above a pocket with linearly arranged DD, DCD, and ED motifs (**Figure 4.1B,D**), matching their proximal locations in early CESA diagrams<sup>22</sup>. By analogy to BcsA that contains a co-crystallized glucan chain and a UDP molecule, we can postulate the functions of the classical conserved motifs in plant CESA: (a) to coordinate UDP (D292 of DD, D459 and D461 of DCD, R713 of QVLRW); (b) to act as the catalytic base (D672); and (c) to stabilize the acceptor glucan (W714 of QVLRW). The catalytic site of the Gh506 structure was supported by docking UDP-glucose into its solvent-exposed catalytic pocket in proximity to DCD (**Figure 4.1D**). Density Functional Theory calculations supported the coordination of UDP via a divalent cation interacting with D459 and D461 of the Gh506 structure (**Figure 4S.13**), similar to other glycosyltransferases<sup>23,24</sup>. In addition, we identified three loops in the vicinity of the UDP-glucose binding site in the Gh506 structure that may control catalysis through modulation of local accessibility to key residues: (1) T258–L267 located at the end of  $\beta$ -2; (2) A294–F300, just after DD and leading to  $\alpha$ -3 of the PCR; and (3) Y421–H432, between  $\beta$ -5 and  $\beta$ -6 (**Figure 4S.8**). The function of these loops can be

further explored in future experiments. Overall, the predicted Gh506 structure shows a highly conserved single active site for coordinating the donor and acceptor sugars for cellulose synthesis.

Regions unique to plant CESAs, the P-CR and CSR domains, extend away from the GT-domain of Gh506 towards the cytosol where they may feasibly regulate other aspects of plant CESA function including the assembly of rosette CSCs (**Figure 4.1B**). The relatively high structural independence of these plant-specific regions was indicated by cross-correlated atomic fluctuations (**Figure 4S.9**). Based on the Gh506 structure, we propose that these regions partake in the oligomerization of CESAs to form the rosette CSCs that are found in land plants and their close relatives. To examine possible roles of the CSR and P-CR in assembly of CESA homo-oligomers, we used the Rosetta Symmetry docking protocol to show possible dimers, trimers, tetramers, and hexamers of the Gh506 structure (**Figure 4.2, Figure 4S.10**). The assemblies show that the CSR and P-CR regions are located at the interfaces of the monomers, supporting the possibility that these regions may help to stabilize CESA assembly through non-covalent interactions. Interestingly, the CSR region is more important for assemblies of dimers and trimers, whereas both the CSR and P-CR participate in assemblies of tetramers and hexamers. Future computational and laboratory experiments can be designed to test how the N-terminal zinc finger region, which is also unique to plant CESAs but not included in the Gh506 structure, may help to modulate CESA assembly through dimerization as shown previously for GhCESA1<sup>25</sup>.

No known missense mutation exists in the CSR, but one does occur in the P-CR: *Atcesa* $\delta$ <sup>R362K</sup> (*fra6*), which was reported to cause reduced cellulose content when

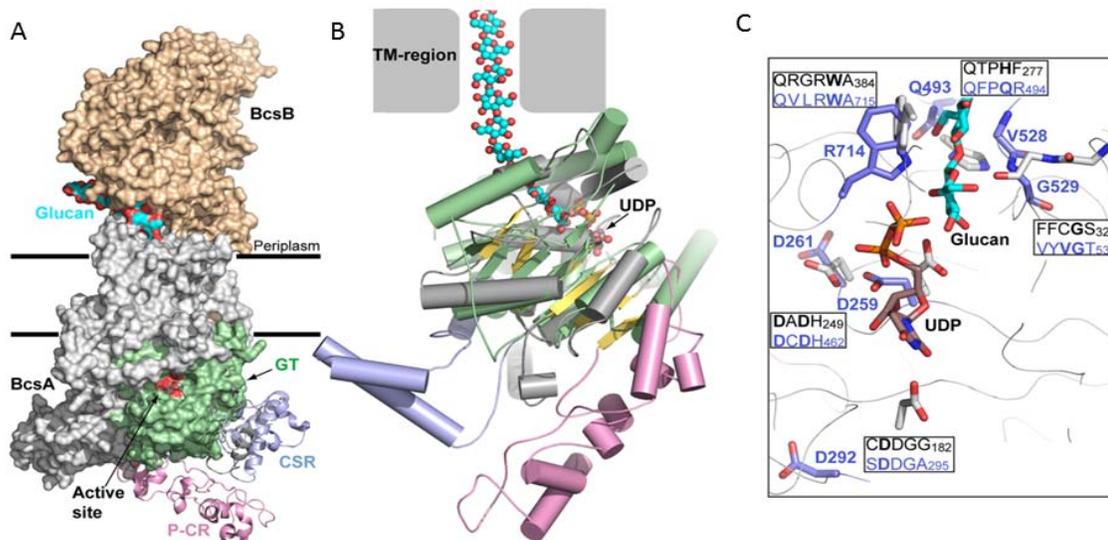
homozygous in *Arabidopsis*. However, no phenotype resulted from over-expression of the mutant gene in wild-type plants<sup>26</sup>. In our tetrameric model, the *fra6* mutation is located at the surface, which potentially could affect the assembly of oligomers into rosette CSCs (**Figure 4.2C**). In our hexameric assembly, *fra6* is located at the interface between the CESA monomers, which could disrupt the assembly of oligomers (**Figure 4.2D**). This result suggests that the affected arginine residue may be important for CESA oligomerization within the rosette CSC.



**Figure 4.2:** Possible oligomeric assemblies of the Gh506 cytosolic structure under (A) C2, (B) C3, (C) C4, and (D) C6 crystallization symmetries. The catalytic region is grey, the CSR is light blue, the P-CR is pink, QVLRW is yellow and the site of *fra6* mutations is red. (D) Bottom, (E) top, and (F) side view of the hexameric Gh506 assembly.

#### 4.2.2 Comparison between bacterial and plant cellulose synthases

The inherent differences in CSC formation and resultant cellulose fibril properties between bacteria and plant CESAs must arise from differences in their protein sequences and, thus, structures. For example, a sequence comparison of the cytosolic region responsible for cellulose synthesis between bacterial BcsA (276 amino acids from GenBank Accession Q3J125) and plant GhCESA1 (506 amino acids) show 17.5% identity, 26.1% similarity, and 49.9% gaps. The plant CESA cytosolic region is longer mostly due to the presence of P-CR and CSR insertions specific to plants<sup>5,6,16</sup>. Even with omission of the P-CR and CSR regions, the co-aligned sequences of the edited bacterial and plant cytosolic regions (240 or 259 residues, respectively) showed 28% identity, 44% similarity, and 10% gaps. However, a structural alignment retaining only homologous sequences between BcsA (solved at 3.25 Å resolution) and Gh506 resulted in a 3.9 Å RMSD overall (**Figure 4.3A, Figure 4S.6**).



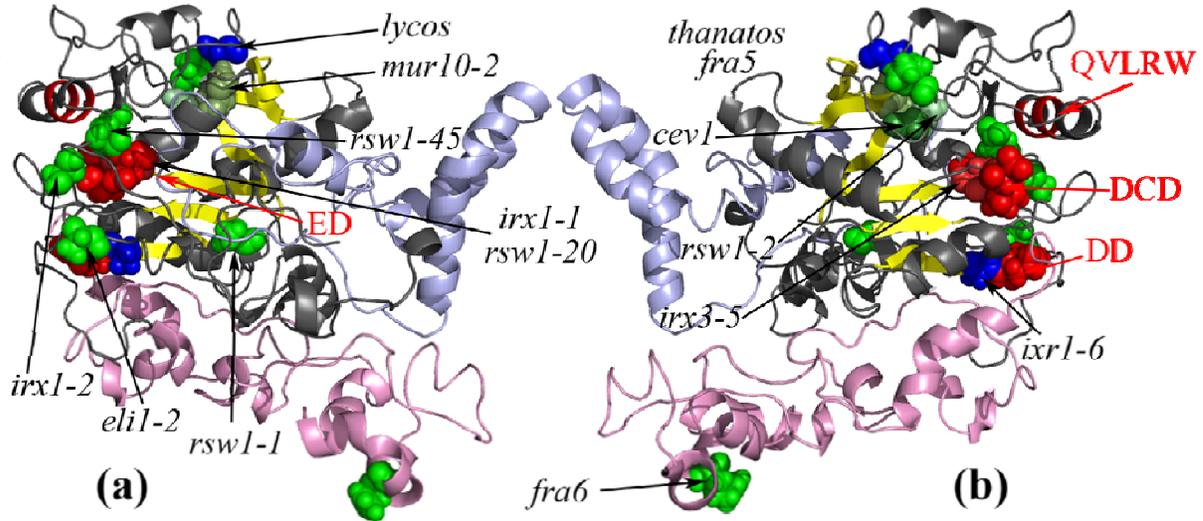
**Figure 4.3:** Comparison of the Gh506 model with the structure of bacterial cellulose synthase. (A) Surface representation of the *Rhodobacter sphaeroides* (Rs) cellulose synthase (PDB ID 4HG6) superimposed with the Gh506 structure. The model aligns well with the central  $\beta$ -sheet of the bacterial GT-domain, and the P-CR and CSR domains point away from the membrane-spanning region. UDP and the translocating glucan of the Rs cellulose synthase highlight the active site and the TM-channel and are colored violet and cyan, respectively. The BcsA subunit of the Rs cellulose synthase complex is colored gray and green, respectively, BcsB is colored wheat. (B) Superimposition of the Rs GT-domain with the Gh506 model by secondary structure of the central  $\beta$ -sheet. The GT-domain is colored green for RsBcsA and gray and yellow for the GhCESA1 model, respectively. The GhCESA1 P-CR and CSR domains are colored pink and light blue, UDP and the glucan are shown as spheres. (C) Conserved sequence motifs that form the binding site for UDP and the acceptor glucan are compared with the corresponding residues in RsBcsA. The Rs and Gh sequence alignments of the motifs are shown and the depicted residues are indicated in bold (black: Rs, blue: Gh). RsBcsA is shown in gray and the UDP and glucan bound to Rs BcsA are colored violet and cyan, respectively. Residues from Gh506 are colored blue. The ED motif was omitted for clarity. Horizontal bars indicate the membrane boundaries.

The bacterial BcsA GT-domain adopts a GT-A fold consisting of a mixed 7-stranded  $\beta$ -sheet surrounded by 7  $\alpha$ -helices<sup>15</sup>. Our Gh506 model aligns well with the BcsA GT-domain, particularly with its central  $\beta$ -sheet and 4 of the surrounding  $\alpha$ -helices (**Figure 4.3AB**), and the Gh506 structure shares 5 of 7  $\alpha$ -helices and 6 of 7  $\beta$ -strands as found in the GT-A fold of BcsA. Other structural features that are likely to function similarly in plant

CESAs and BcsA are noted in **Table 4S.3**. **Figure 4.3C** shows that the invariant DD, DCD, and QVLRW motifs of our Gh506 model align well with the bacterial cellulose synthase structure. This comparison importantly confirms a high degree of structural similarity between the catalytic sites of eukaryotic and prokaryotic cellulose-synthesizing proteins, which indicates a conserved mechanism of cellulose polymerization. Moreover, the structural alignment orients the P-CR and CSR domains of Gh506 towards the cytosol (**Figure 4.3A**).

#### 4.2.3 Genetic mutations demonstrate functional nodes within plant CESA structure

The relatively high sequence conservation between seed plant CESAs (**Figure 4S.1**) allowed Arabidopsis CESA missense mutations to be mapped onto the analogous residue in the Gh506 structure (**Figure 4.4; Table 4S.3**). Based on the primary sequence, several previously identified Arabidopsis CESA missense mutations coincide with the conserved ED motif [*Atcesa1*<sup>E779K</sup> (*rsw1-45*), *Atcesa8*<sup>D683N</sup> (*irx1-1*), and *Atcesa1*<sup>D780N</sup> (*rsw1-20*)] or the first D in the DCD motif (*Atcesa7*<sup>D524N</sup>)<sup>7-9</sup>. However, other missense mutations are dispersed throughout the cytosolic region of CESAs at locations with no known function. Interestingly, in our Gh506 model the mutated residues primarily converged in a spatially discrete cluster around the catalytic site even though the residues were dispersed throughout the sequence (**Figure 4.4**). A plausible interpretation for this result is that the catalytic region retains an overarching tertiary structure across plant CESAs and that most of the currently known missense mutations that lead to reduced cellulose content cluster around this core domain.



**Figure 4.4:** Previously known missense mutations (green) in Arabidopsis CESAs mapped onto the predicted GhCESA1 cytosolic structure. New mutations are blue with the DD, DCD, ED, and QVLRW motifs in red. Two opposite rotations (a) and (b) provide a view of where the mutations map relative to the catalytic region of the Gh506 structure. The equivalent Ghcesa1 amino acid position precedes the mapped mutation: R351 (*Atcesa8*<sup>R362K</sup>, *fra6*); A436 (*Atcesa3*<sup>A522V</sup>, *eli1-2*); A447 (*Atcesa1*<sup>A549V</sup>, *rsw1-1*); D459 (*Atcesa7*<sup>D524N</sup>, *irx3-5*); P492 (*Atcesa7*<sup>P557T</sup>, *fra5* and *Atcesa3*<sup>P578S</sup>, *thanatos*); G529 (*Atcesa1*<sup>G631S</sup>, *rsw1-2*); G531 (*Atcesa3*<sup>G617E</sup>, *cev1*); S668 (*Atcesa3*<sup>S679L</sup>, *irx1-2*); E671 (*Atcesa1*<sup>E779K</sup>, *rsw1-45*); D672 (*Atcesa3*<sup>D683N</sup>, *irx1-1* and *Atcesa1*<sup>D780N</sup>, *rsw1-20*); and H680 (*Atcesa7*<sup>H734Y</sup>, *mur10-2*). The conserved residues with no known mutations are red.

In addition, the location of missense mutations in the Gh506 structure provided new insights about putative functionally important nodes within CESA. For example, the analog of the *Atcesa7*<sup>H734Y</sup> (*mur10-2*) mutation<sup>27</sup> located after TED in  $\alpha$ -12 makes contact with  $\beta$ -5 and  $\beta$ -6, as well as the site of the *Atcesa1*<sup>G631S</sup> (*rsw1-2*) mutation<sup>28</sup> even though these mutated histidine and glycine are separated by ~150 residues in the sequence.

The *Atcesa7*<sup>H734Y</sup> plants have dwarfed shoots and cellulose-deficient xylem secondary walls<sup>27</sup>. The *Atcesa1*<sup>G631S</sup> mutant seedlings have ~75% less crystalline cellulose and swollen organs<sup>28</sup>. The mapped sites of the *Atcesa1*<sup>G631S</sup> and *Atcesa3*<sup>G617E</sup> (*cev1*) mutations are

separated by one amino acid, and *Atcesa3*<sup>G617E</sup> mutant plants were dwarfs with radial cell swelling and cellulose deficiency compared to wild-type<sup>29</sup>. The *Atcesal*<sup>G631S</sup> and *Atcesa3*<sup>G617E</sup> mutation sites lie at the end of  $\beta$ -4 in a VYVGTG motif, which structurally aligns with the FFCGS motif of BcsA in the core GT domain. The perturbation of a  $\beta$ -sheet structure may affect the structure of the catalytic site and substrate binding. In BcsA, FFCGS binds the terminal disaccharide of the glucan acceptor on the opposite side as compared to QRGRW<sup>15</sup>. Therefore, the *Atcesa7*<sup>H734Y</sup>, *Atcesal*<sup>G631S</sup>, and *Atcesa3*<sup>G617E</sup> mutation sites may represent a functional node that controls the acceptor glucan placement or conformation within the active site (**Figure 4.4**).

To further investigate the effect of mutations, we mapped and cloned two new missense mutations to regions of interest within Arabidopsis CESAs. The first of these, *Atcesa3*<sup>S377F</sup> confers resistance to the cellulose synthesis inhibitor, isoxaben, and the mutant Arabidopsis plants showed reduced growth and a lower relative crystallinity although the cellulose content was not statistically different than in the control. According to our Gh506 model the analogous affected residue, S291 in GhCESA1, resides within a water-accessible pocket at the end of  $\beta$ -1 and before the A294-F300 loop containing DD (**Figure 4.5A**). Interestingly, this mutation disturbs the crystallization process with little impact on cellulose catalysis despite proximity to the conserved DD residue. To explore how this mutation disrupts CESA structure, MD simulations were performed on the mutated model CESA. In **Figure 4.5C** each peak represents a highly motile residue that is often solvent accessible, whereas the valleys are largely populated by buried amino acids. For example, mapped peaks are the analogs of *Atcesa8*<sup>R362K</sup> (*fra6*) located in the P-CR and *Atcesa*<sup>S679L</sup> (*irx1-2*) three

residues below the conserved ED motif. This analysis showed that S291 in GhCESA1 is tightly coupled to the conserved ED motif by the short T258–L267 loop and to residues S572-R580 in  $\alpha$ -9 within the CSR. As previously explained based on analogy to BcsA, the ED residues are likely to affect catalysis directly as well as interact with glucose when it is bound to UDP. Disturbance of glucose positioning in the active site could affect glucan chain conformation and/or the rate of catalysis, which could affect cellulose crystallization. Possible effects arising through coupling to the CSR are not easily defined given the unknown function of this domain.

In general, a single point mutation of a key residue may affect CESA function in multiple ways. For example, in the Gh506 structure, the S291 residue that is analogous to *Atcesa3*<sup>S377F</sup> contacts L442 within core  $\alpha$ -6 (**Figure 4.5A**), and MD simulations of mutated *GhCesa1*<sup>S291F</sup> revealed a larger distance between these residues compared to wild-type (**Figure 4S.11Q, Table 4S.4**). Notably, the analog of the A residue in the temperature-sensitive *Atcesal*<sup>A549V</sup> (*rsw1-1*) mutation is at the base of core  $\alpha$ -6 (**Figure 4.5A**). *Arabidopsis Atcesal*<sup>A549V</sup> mutants grown at the restrictive temperature showed severely impaired crystalline cellulose deposition and seedling growth<sup>30</sup>, effects that are similar to the *Atcesa3*<sup>S377F</sup> mutation. The analog of the *Atcesa3*<sup>A522V</sup> (*eli1-2*) mutation, which also caused similar phenotypes<sup>31,32</sup>, is 10 residues away from the *Atcesal*<sup>A549V</sup> mutation site at the other end of core  $\alpha$ -6 within a HKKAGA motif (**Figure 4.5A**) that is co-aligned with the HAKAGN motif of BcsA. In BcsA, the A225 and K226 residues of HAKAGN lie on the other side of the pocket that may accommodate glucose when bound to UDP<sup>15</sup>. Taken together, these results suggest that core  $\alpha$ -6, predicted to be in the interior of CESA behind  $\beta$ -

1–3, may help to control the positioning of the donor glucose in the catalytic site, which in turn may modulate the organization of glucan chains into crystalline cellulose fibrils.

**Table 4.1:** Plant phenotypes for new CESA mutations. Significantly different compared with wild-type (LER or Col-0) as determined by *t* test: \* P< 0.001; \*\*P=0.009; \*\*\*P<0.01. RCI, Relative crystallinity index.

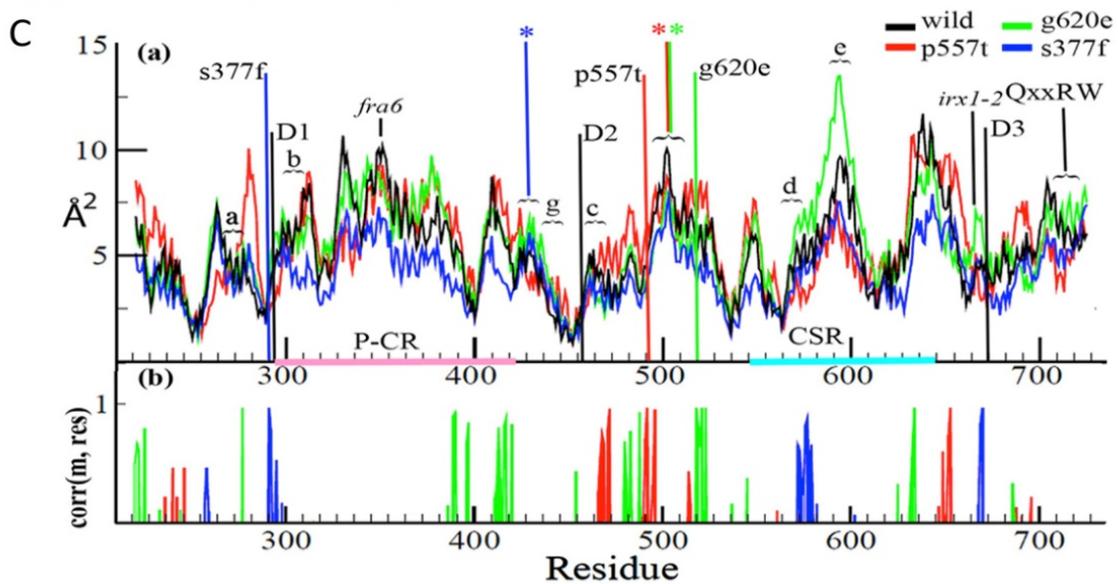
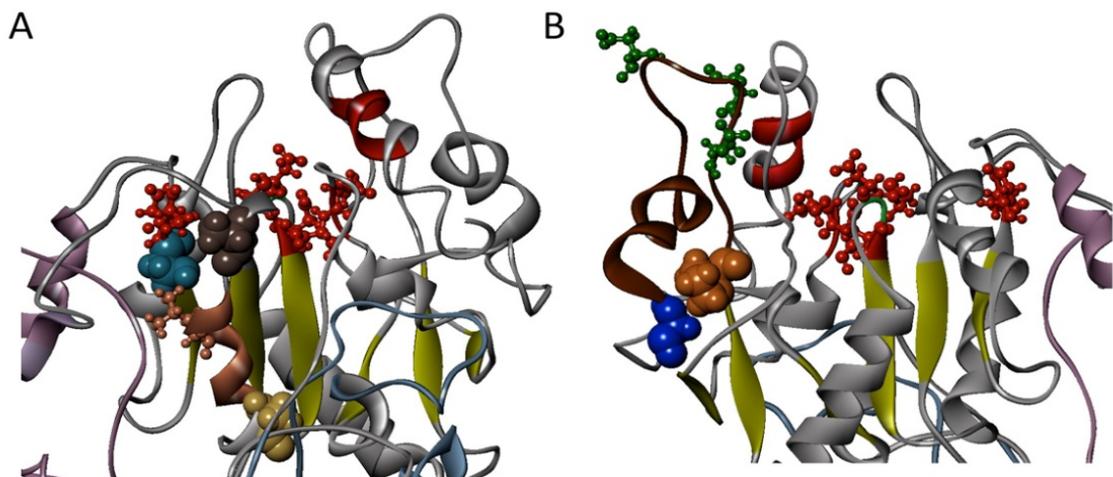
Allele/genotype	Dark-grown hypocotyls length, %w/t	Height, mature stem, cm (SE)	Cellulose content, mature stem, %wt	RCI, mature stem (SE)
Atcesa3 <sup>S377F</sup> ( <i>ixr1-6</i> ) LER background	45.6*	20.8 (0.5)*	87.5	32.8 (4.7)**
Atcesa1G620E ( <i>lycos</i> ) Col-0 background	100	13.4 (0.5)*	61.7*	41.9 (1.0)***
Wild type (LER)	100	30.7 (3.1)	100	48.4 (1.1)
Wild type (Col-0)	100	39.1 (0.8)	100	49.2 (2.1)

A second new Arabidopsis missense mutation, *Atcesa1*<sup>G620E</sup>, conferred resistance to the cellulose synthesis inhibitor quinoxiphen and also caused reductions in stem height, relative crystallinity, and cellulose content (**Table 4.1**). Its analogous residue in GhCESA1, G518, helped to support the functional importance of the solvent-accessible P492-G518 loop that lies between  $\beta$ -4 and  $\beta$ -5 and behind  $\alpha$ -13/QVLRW in the Gh506 structure (**Figure 4.5B**). The G518 residue is predicted to sit adjacent to the P492 residue, which is analogous to the site of the *Atcesa7*<sup>P557T</sup> (*fra5*) and *Atcesa3*<sup>P578S</sup> (*thanatos*) mutations<sup>26,33</sup>. The P492 and G518 residues appear to act as hinge points for the loop between them. The range of motion for this loop was established from the MD simulation trajectory and mutated residues caused the change in range of motion (**Figure 4S.11, Table 4S.4**). The tip of the loop contains three aspartic acid residues (**Figure 4.5B**), and it is able to contact the QVLRW motif and may potentially modulate its interaction with the newly forming  $\beta$ -1,4-glucan chain. Thus, changing the dynamic behavior of the loop by mutations at its base may adversely affect QVLRW interaction with the cellulose product.

Several computational experiments showed putative effects of altering the predicted hinge points of the P492-G518 loop through: (a) substitution at P492 of threonine (analogous to *Atcesa7*<sup>P557T</sup>) or serine (analogous to *Atcesa3*<sup>P578S</sup>); and (b) substitution of glutamic acid at G518 (analogous to *Atcesa1*<sup>G620E</sup>). The dynamic behavior of the three D residues at the tip of the loop was reduced for mutant E518 compared to wild-type G518 in MD simulations based on the Gh506 structure (**Table 4S.4c**). Reasonably, the larger glutamic acid residue could cause constrained movement, steric clash, and changed hydrophobicity of the solvent-exposed loop. However, the expected reduced local rigidity for mutant T492 compared to

wild type was not observed (**Table 4S.4, Figure 4S.11**), possibly due to the effects of intermittent hydrogen bonding interactions between T492 and Y688 of  $\beta$ -6 observed in the MD simulations. This hydrogen bonding interaction may serve to stabilize the mutant T492-G518 loop (**Figure 4S.12**). Similarly, E518 showed intermittent hydrogen bonding with the adjacent L517 residue.

**Figure 4.5:** The locations of novel missense mutations in the predicted structure helped to support the existence of new functionally important regions within CESA. Conserved residues are shown in red. (A) S291 (teal) just below DD is the analog of the novel *Atcesa3*<sup>S377F</sup>, *ixr1-6*, mutation. In the predicted structure, it contacts L442 (rust ball and stick residue) within  $\alpha$ -6 (rust), which has the analogs of *Atcesa3*<sup>A522V</sup> (*elil-2*; brown) and *Atcesal*<sup>A549V</sup> (*rsw1-1*; tan) at either end. (B) The P492-G518 loop (brown) contains native aspartates (green ball and stick residues) near QVLRW. At its base are G518 (blue; the analog of the novel *Atcesal*<sup>G620E</sup>, *lycos*, mutation) and P492 (rust; the analog of *Atcesa7*<sup>P557T</sup>, *fra5*), where they may putatively act as hinge points. (C) Cross correlation of atomic fluctuations at four mutation sites over all simulations by residue. The peaks shown had at least 97% correlation, indicating distant effects of the mutations analogous to *Atcesa3*<sup>S377F</sup> (*ixr1-6*; blue bars), *Atcesal*<sup>G620E</sup> (*lycos*; green bars), and *Atcesa7*<sup>P557T</sup> (*fra5*; red bars).



Mutations at these putative hinge points also had widely distributed effects on the Gh506 structure as determined through correlations of residue fluctuations. The position of mutant E518 strongly couples to atomic motions in  $\beta$ -3,  $\beta$ -6, and part of the P-CR and CSR regions. The mutant T492 position couples to  $\beta$ -2,  $\beta$ -5,  $\beta$ -6, and F696 near the QVLRW motif. Therefore, both of these mutations are likely to perturb the  $\beta$ -sheet, which can affect catalysis and substrate binding<sup>34</sup>. Previous modeling of 185 amino acids with the HMMSTR/Rosetta Server suggested that the catalytic domain structure was altered by *Atcesa3*<sup>P578S</sup> (*thanatos*) mutation<sup>35</sup>. However, we could not reproduce this result with *de novo* modeling using the SAM-T08 server for the 506 amino acid long GhCESA1 cytosolic region containing the analogous mutation (data not shown).

Commonalities in the *Atcesa7*<sup>P557T</sup> (*fra5*) mutation and the *Atcesa3*<sup>P578S</sup> (*thanatos*) mutation can now be explained through effects on the same functional loop. Both are semi-dominant: *Atcesa3*<sup>P578S</sup> causes reduced primary wall cellulose synthesis<sup>35</sup> and *Atcesa7*<sup>P557T</sup> causes reduced cellulose content of fiber cells<sup>26</sup>. Both mutations exert dominant negative effects when over-expressed in wild-type, which has been described only for these two Arabidopsis CESA missense mutations<sup>26,35</sup>. Therefore, the mutant proteins must compete effectively for entry into the rosette CSC, which is logical given the location of the analogous residues near the catalytic region of the Gh506 structure. Together the data presented here illustrate the utility of the predicted tertiary structure of the GhCESA1 cytosolic region to provide insight into mechanisms of cellulose polymerization in plants, help systematize data on CESA missense mutations, and illuminate possible new structure/function relationships that are broadly conserved among plant CESAs.

Overall, we were able to predict a complex three dimensional structure of plant cellulose synthase from *Gossypium hirsutum* using a molecular modeling approach. Our model is in close agreement with the core region of the recently solved structure of the bacterial BcsA cellulose synthase<sup>15</sup> despite substantial differences in the plant and bacterial sequences. Given that BcsA was not used as structural homolog for model prediction, this structural convergence supports a conserved mechanism for cellulose polymerization. The clustering of most Arabidopsis missense mutations around the structurally conserved catalytic site further supports the similarity of the cellulose catalytic mechanism across Kingdoms. Moreover, unique regions to plant CESAs, the CSR and P-CR, were revealed to fold into distinguishable subdomains within the cytosolic region, and these regions can be explored further for how they potentially control the assembly of plant CSCs, other regulatory aspects of plant cellulose synthesis, and, consequently, the unique material properties of plant cellulose.

## **4.3 Methods and Materials**

### **4.3.1 Simulations and Modeling**

We used secondary structure prediction bioinformatics tools such as PSI-PRED<sup>36</sup>, in order to isolate the cytosolic region of GhCESA1 (P93155) (**Figure 4S.1**). Almost the entire region was modeled (506 amino acids; Q220–R725) beginning just after transmembrane helix 2 (TMH2). Only a small loosely conserved linker between the C-terminal region including QVLRW and TMH3 was excluded to reduce computational complexity.

Successful *de novo* structure modeling is predicated on an accurate energy function, an efficient search method, and selection of appropriate models from the ensemble. Heuristic

HMM protein structure prediction approaches, along with fragment based assembly algorithms such as ROSETTA (<http://www.rosettacommons.org/>), have proven to be most successful<sup>37,38</sup>. However, due to the intensive computational time required, successful *de novo* folding with ROSETTA has generally been limited to 100–150 amino acids<sup>39,40</sup>. To overcome this limitation, we used the protein structure prediction server of SAM-T08<sup>41</sup> to generate an initial homology model. The SAM-T08 server relies on the construction of HMM and multiple sequence alignments to generate structural homologs for parts of the target structure. The initial fragmented homology based structure was manually refined and subjected to series of MD simulations to explore the conformational space and develop the final modeled structure. After preliminary structural quality checks, the modeled structure was analyzed comprehensively using the Protein Structure Validation Software suite (PSVS)<sup>42</sup> and its quality score was determined with ERRAT<sup>43</sup>. The UDP-glucose was docked into the structure with the help of Density Functional Theory carried out in Gaussian 03<sup>44</sup> with the B3LYP/6-311+G(d,p) method and D residues constrained. Mutants based on the Gh506 structure were generated using the TLEAP tool of Amber 11<sup>45</sup>, subjected to MD simulations and the resultant structural flexibility was assessed using cross correlation analysis<sup>46</sup>. Putative homomeric assemblies of the Gh506 structure were generated using the symmetric docking protocol of Rosetta<sup>47</sup>. Additional details are available in the supplemental methods.

### **4.3.2 Novel mutations in Arabidopsis CESAs and phenotypes of mutant plants**

Approximately forty-five thousand *A. thaliana* ecotype Landsberg (LER) and Columbia-0 (Col-0) seeds were mutagenized by ethyl methane sulfonate (EMS) by immersing seeds in a solution of 0.3% EMS (M1) for 16 h, extensively washed with distilled

water (12 h) and sown into soil to generate M2 seeds. M2 seed were surface sterilized and one million M2 seeds were plated on 0.5X strength Murashige and Skoog (MS) agar plates supplemented with 20 nM isoxaben (LER screen) or 5  $\mu$ M quinoxyphen (Col-0 screen). Seed were stored at 4°C for 4 days to synchronize germination and then exposed to 100  $\mu$ E/m<sup>2</sup>/s white light at room temperature until seeds germinated and cotyledons had expanded. Resistant mutants grew above the surface of the agar while non-resistant plants did not. Resistant plants from the M2 generation were retested in the M3 generation to confirm heritability of the resistance phenotype. The novel isoxaben resistant (*ixr*) allele in AtCESA3 discussed here was named *ixr1-6*, and the novel quinoxyphen resistance allele in AtCESA1 discussed here was named *lycos*. For clarity, the mutants are referred to in the text as Atcesa3<sup>S377F</sup> and Atcesa1<sup>G620E</sup>, respectively. Methods for assessing phenotypes were as described previously<sup>1</sup>.

### **Acknowledgements**

Work by L.S, A.S., J.K., C.H.H. and Y.G.Y. was supported as part of The Center for LignoCellulose Structure and Formation, Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Science under Award Number DE-SC0001090. Work by SD was supported by the National Science Foundation award IOS- 0922947. Work by JZ was support by the NIH grant 1R01GM101001 and start-up funds from the University of Virginia School of Medicine. Work by DB was supported by the National Science and Engineering Research Council of Canada (NSERC).

## References

1. Harris DM, Corbin K, Wang T, Gutierrez R, Bertolo AL, Petti C, Smilgies DM, Estevez JM, Bonetta D, Urbanowicz BR, Ehrhardt DW, Somerville CR, Rose JKC, Hong M, DeBolt S. Cellulose microfibril crystallinity is reduced by mutating C-terminal transmembrane region residues CESA1(A903V) and CESA3(T942I) of cellulose synthase. *Proceedings of the National Academy of Sciences of the United States of America* 2012;109(11):4098-4103.
2. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* 2009;37(Database issue):D233-238.
3. Somerville C. Cellulose synthesis in higher plants. *Annu Rev Cell Dev Bi* 2006;22:53-78.
4. Nishiyama Y. Structure and properties of the cellulose microfibril. *J Wood Sci* 2009;55:241-249.
5. Roberts E, Roberts AW. A Cellulose Synthase (Cesa) Gene from the Red Alga *Porphyra Yezoensis* (Rhodophyta). *J Phycol* 2009;45(1):203-212.
6. Pear JR, Kawagoe Y, Schreckengost WE, Delmer DP, Stalker DM. Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. *Proceedings of the National Academy of Sciences of the United States of America* 1996;93(22):12637-12642.
7. Taylor NG, Laurie S, Turner SR. Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in *Arabidopsis*. *Plant Cell* 2000;12(12):2529-2539.
8. Beeckman T, Przemeck GKH, Stamatiou G, Lau R, Terryn N, De Rycke R, Inze D, Berleth T. Genetic complexity of cellulose synthase A gene function in *Arabidopsis* embryogenesis. *Plant Physiology* 2002;130(4):1883-1893.
9. Liang YK, Xie X, Lindsay SE, Wang YB, Masle J, Williamson L, Leyser O, Hetherington AM. Cell wall composition contributes to the control of transpiration efficiency in *Arabidopsis thaliana*. *Plant J* 2010;64(4):679-686.
10. Saxena IM, Brown RM. Identification of cellulose synthase(s) in higher plants: Sequence analysis of processive beta-glycosyltransferases with the common motif 'D, D, D35Q(R,Q)XRW'. *Cellulose* 1997;4(1):33-49.

11. Yoshida M, Itano N, Yamada Y, Kimata K. In vitro synthesis of hyaluronan by a single protein derived from mouse HAS1 gene and characterization of amino acid residues essential for the activity. *Journal of Biological Chemistry* 2000;275(1):497-506.
12. Nagahashi S, Sudoh M, Ono N, Sawada R, Yamaguchi E, Uchida Y, Mio T, Takagi M, Arisawa M, Yamadaokabe H. Characterization of Chitin Synthase-2 of *Saccharomyces-Cerevisiae* - Implication of 2 Highly Conserved Domains as Possible Catalytic Sites. *Journal of Biological Chemistry* 1995;270(23):13961-13967.
13. Charnock SJ, Davies GJ. Structure of the Nucleotide-Diphospho-Sugar Transferase, SpsA from *Bacillus subtilis*, in Native and Nucleotide-Complexed Forms. *Biochemistry* 1999;38(20):6380-6385.
14. Sobhany M, Kakuta Y, Sugiura N, Kimata K, Negishi M. The Chondroitin Polymerase K4CP and the Molecular Mechanism of Selective Bindings of Donor Substrates to Two Active Sites. *Journal of Biological Chemistry* 2008;283(47):32328-32333.
15. Morgan JLW, Strumillo J, Zimmer J. Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature* 2012(doi:10.1038/nature11744).
16. Carpita NC. Update on Mechanisms of Plant Cell Wall Biosynthesis: How Plants Make Cellulose and Other (1→4)-β-d-Glycans. *Plant Physiology* 2011;155(1):171-184.
17. Guerriero G, Fugelstad J, Bulone V. What Do We Really Know about Cellulose Biosynthesis in Higher Plants? *J Integr Plant Biol* 2010;52(2):161-175.
18. Diotallevi F, Mulder B. The cellulose synthase complex: A polymerization driven supramolecular motor. *Biophys J* 2007;92(8):2666-2673.
19. Betancur L, Singh B, Rapp RA, Wendel JF, Marks MD, Roberts AW, Haigler CH. Phylogenetically distinct cellulose synthase genes support secondary wall thickening in arabidopsis shoot trichomes and cotton fiber. *J Integr Plant Biol* 2010;52(2):205-220.
20. Breton C, Snajdrova L, Jeanneau C, Koca J, Imberty A. Structures and mechanisms of glycosyltransferases. *Glycobiology* 2006;16(2):29r-37r.
21. Saxena IM, Brown RM. Cellulose biosynthesis: Current views and evolving concepts. *Ann Bot-London* 2005;96(1):9-21.
22. Delmer DP. Cellulose biosynthesis: Exciting times for a difficult field of study. *Annu Rev Plant Phys* 1999;50:245-276.

23. Hashimoto K, Madej T, Bryant SH, Panchenko AR. Functional States of Homooligomers: Insights from the Evolution of Glycosyltransferases. *Journal of Molecular Biology* 2010;399(1):196-206.
24. Wiggins CAR, Munro S. Activity of the yeast MNN1 alpha-1,3-mannosyltransferase requires a motif conserved in many other families of glycosyltransferases. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95(14):7945-7950.
25. Kurek I, Kawagoe Y, Jacob-Wilk D, Doblin M, Delmer D. Dimerization of cotton fiber cellulose synthase catalytic subunits occurs via oxidation of the zinc-binding domains. *Proceedings of the National Academy of Sciences of the United States of America* 2002;99(17):11109-11114.
26. Zhong RQ, Morrison WH, Freshour GD, Hahn MG, Ye ZH. Expression of a mutant form of cellulose synthase AtCesA7 causes dominant negative effect on cellulose biosynthesis. *Plant Physiology* 2003;132(2):786-795.
27. Bosca S, Barton CJ, Taylor NG, Ryden P, Neumetzler L, Pauly M, Roberts K, Seifert GJ. Interactions between MUR10/CesA7-dependent secondary cellulose biosynthesis and primary cell wall structure. *Plant Physiology* 2006;142(4):1353-1363.
28. Gillmor CS, Poindexter P, Lorieau J, Palcic MM, Somerville C. alpha-glucosidase I is required for cellulose biosynthesis and morphogenesis in Arabidopsis. *Journal of Cell Biology* 2002;156(6):1003-1013.
29. Ellis C, Karafyllidis I, Wasternack C, Turner JG. The Arabidopsis mutant *cev1* links cell wall signaling to jasmonate and ethylene responses. *Plant Cell* 2002;14(7):1557-1566.
30. Arioli T, Peng L, Betzner AS, Burn J, Wittke W, Herth W, Camilleri C, Hofte H, Plazinski J, Birch R, Cork A, Glover J, Redmond J, Williamson RE. Molecular analysis of cellulose biosynthesis in Arabidopsis. *Science* 1998;279(5351):717-720.
31. Cano-Delgado A, Penfield S, Smith C, Catley M, Bevan M. Reduced cellulose synthesis invokes lignification and defense responses in Arabidopsis thaliana. *Plant J* 2003;34(3):351-362.
32. Pysh L, Alexander N, Swatzyna L, Harbert R. Four alleles of AtCESA3 form an allelic series with respect to root phenotype in Arabidopsis thaliana. *Physiol Plantarum* 2012;144(4):369-381.
33. Daras G, Rigas S, Penning B, Milioni D, McCann CM, Fasseas C, Hatzopoulos P. Thanatos mutation in Cesa3 gene exhibits a nonconditional semidominant-negative phenotype on Arabidopsis primary cell wall formation. *Febs J* 2008;275:361-361.

34. Reynolds KA, McLaughlin RN, Ranganathan R. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* 2011;147(7):1564-1575.
35. Daras G, Rigas S, Penning B, Milioni D, McCann MC, Carpita NC, Fasseas C, Hatzopoulos P. The thanatos mutation in *Arabidopsis thaliana* cellulose synthase 3 (AtCesA3) has a dominant-negative effect on cellulose synthesis and plant growth. *New Phytol* 2009;184(1):114-126.
36. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 1999;292(2):195-202.
37. Fleishman SJ, Corn JE, Strauch EM, Whitehead TA, Andre I, Thompson J, Havranek JJ, Das R, Bradley P, Baker D. Rosetta in CAPRI rounds 13-19. *Proteins-Structure Function and Bioinformatics* 2010;78(15):3212-3218.
38. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. *Proteins* 2011;79:37-58.
39. Lee J, Wu S, Zhang Y. *Ab Initio* Protein Structure Prediction. *From Protein Structure to Function with Bioinformatics* 2009:3-25.
40. Drew K, Winters P, Butterfoss GL, Berstis V, Uplinger K, Armstrong J, Riffle M, Schweighofer E, Bovermann B, Goodlett DR, Davis TN, Shasha D, Malmstrom L, Bonneau R. The Proteome Folding Project: Proteome-scale prediction of structure and function. *Genome Research* 2011;21(11):1981-1994.
41. Karplus K. SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Research* 2009;37:W492-W497.
42. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. *Proteins-Structure Function and Bioinformatics* 2007;66(4):778-795.
43. Colovos C, Yeates TO. Verification of Protein Structures - Patterns of Nonbonded Atomic Interactions. *Protein Science* 1993;2(9):1511-1519.

44. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery J, J. A., Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 03. C.02. Wallingford CT: Gaussian, Inc.; 2004.
45. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, C. Tan, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, C. Sagui, Babin V, Kollman. PA. Amber 11: University of California, San Francisco; 2010.
46. Kormos BL, Baranger AM, Beveridge DL. A study of collective atomic fluctuations and cooperativity in the U1A-RNA complex based on molecular dynamics simulations. *J Struct Biol* 2007;157(3):500-513.
47. Andre I, Bradley P, Wang C, Baker D. Prediction of the structure of symmetrical protein assemblies. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104(45):17656-17661.

# Chapter 5

## **Subdomains that Modulate Cellulose Synthase Complex Formation and Terminal Complexes**

# **Subdomains that Modulate Cellulose Synthase Complex Formation and Terminal Complexes**

Latsavongsakda Sethaphong and Yarosava G. Yingling

911 Partners Way, Materials Science and Engineering, North Carolina State University,  
Raleigh, NC 27695

## **Abstract**

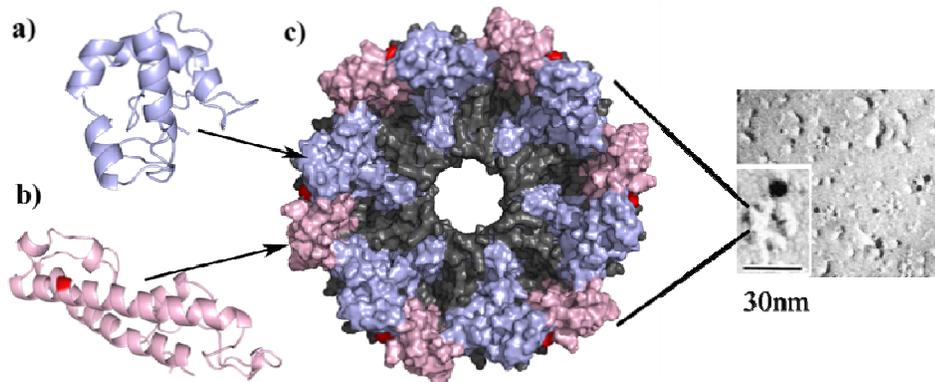
Higher plants express six classes of cellulose synthase (CesA) which are differentially expressed in specific tissues and stages of plant growth. They are further grouped according to their association with primary or secondary cell wall formation. While CesAs in higher plants originated from cyanobacteria<sup>1</sup>, they have evolved additional functionality to gain greater control in the material properties of cellulose. Plant CesA's possess distinct areas in their catalytic domain that differentiates them from their prokaryotic forebears: These are the Plant Conserved Region (CR-P) and a Class Specific Region (CSR). In the absence of an experimentally determined model of plant CESAs, we used structure prediction and molecular modeling techniques to survey the structural motifs of these domains in order to further understand their evolutionary relationships and their biochemical roles in controlling cellulose production. The CSR's distinguish paralogous classes from each other. These insights into plant control of cellulose's physical properties may aid in engineering useful polysaccharide important to pharmaceuticals and industrial applications.

## **5.1 Introduction**

CESAs share a common topology of eight transmembrane helices that form a pore over a globular cytosolic domain that polymerizes cellulose from UDP-bound glucose. In

general, the cellulose synthase family can be divided according to their organismal origin<sup>1-3</sup>. CESA's from different phyla share a common catalytic mechanism as represented by their conserved amino acid motifs: These are the D, D, D and QxxRW amino-acid sequence motif of which the latter is a hallmark of all processive glycosyltransferases<sup>4-7</sup>. These signature residues were also proven to be essential for chitin synthase activity in yeast<sup>8</sup>. Hence, catalysis of polysaccharides appears to follow evolutionarily conserved mechanisms.

Although the a dimensional structure of a bacterial CESA has been recently solved<sup>9</sup>, there has yet to be an experimentally determined plant CESA. Cellulose synthesis in plants is a more complex process than found in bacteria since specialized plant cells must orchestrate the production of cell walls in specific tissues (e.g. xylem, sclerenchyma, phloem). In the model plant *Arabidopsis thaliana*, there are 10 cellulose synthase proteins that perform the task of forming the primary and secondary cell walls. These CESA proteins embed in the plasma membrane and arrange into hexameric arrays termed rosettes which have been found to exist in only land plants and certain algae classified as streptophytes<sup>10-12</sup>. The remainder of other cellulose producing organisms form linear arrays of CESAs as typified by chlorophyta<sup>13</sup>. This differentiation between rosette versus linear complex forming cellulose producing organisms has been previously examined at a supramolecular level<sup>13,14</sup>. Molecular studies suggest that there are significant differences between the catalytic domains of rosette forming CESA's versus those that form linear complexes<sup>15</sup>.



**Figure 5.1:** Putative homomeric-hexamer rosette assembly. (a) Class Specific Region (b) Plant Conserved Region (FRA6 mutation amino acid position is in red) (c) A putative hexameric assembly of a rosette, for scale, cryo-em image of a Terminal Complex comprised of six rosettes is shown from Kimura 1999.

In our previous work, we modeled GhCESA1 which is an analogue to AtCESA8. The results of that effort, summarized in **Figure 5.1**, showed that the PCR and CSR regions do in fact form distinctive subregions around the cytosolic catalytic region (**Figure 5.1a,b**). A putative hexameric assembly using this early model compared favorably with the scale and geometry of cryo-em images of rosette terminal complexes (**Figure 5.1c**). Hence, the quintessential distinguishing feature of plant CESA's is the existence of these two unique subdomains in the globular cytosolic region: a Plant Conserved Region (PCR) and a Class Specific Region (CSR). The PCR extends from the first D motif to just before the second D motif; this region is highly conserved in all plant CESA's. The CSR region, however, is not conserved among the plant CESA orthologs and serves to distinguish between those associated with the primary (CESA 1, 3, 6) and secondary (CESA 4,7,8) cell wall formation. The CSR lies between the second D motif and before the third D motif<sup>16</sup>. CESA 2, 5 and 9 are related to CESA6 and are expressed in specific tissues and stages of development such as

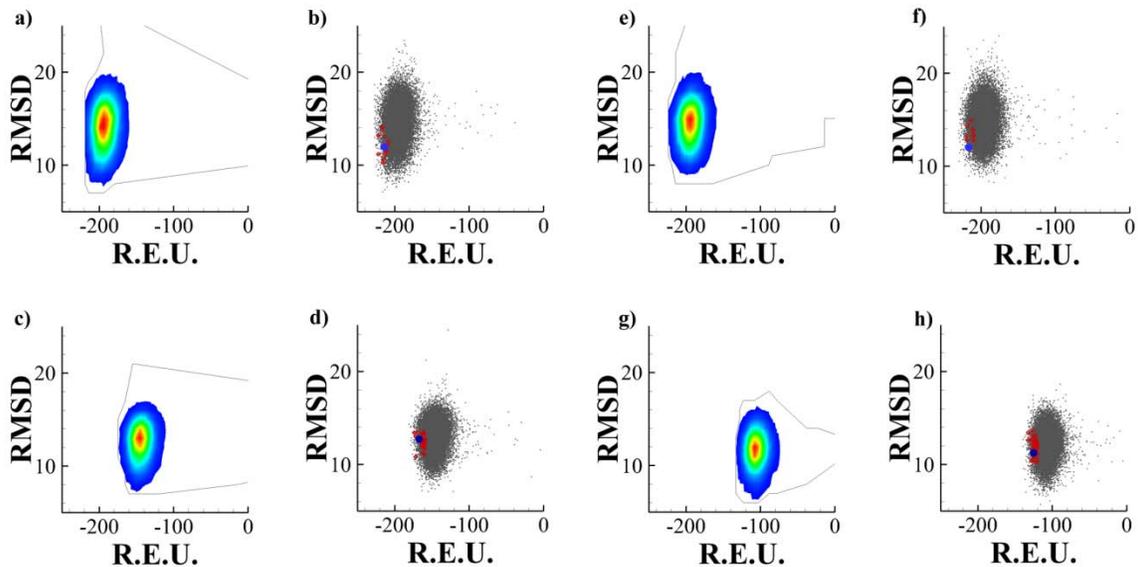
embryogenesis and seed coat formation<sup>17-22</sup>. Closely related to CESA1, CESA10 has been observed in plant ovules and the petiole of rosette leaves<sup>23</sup>; however, its function has yet been fully determined<sup>19,24</sup>. Moreover, a recent study by Carroll et al. has shown that CESAs can broadly interact with each other<sup>25</sup>. The mixed rosettes that were formed functioned abnormally, a result suggesting that the complexes of the primary and secondary rosettes share a similar assembly; yet individual CESA's were not fully interchangeable. The region with the most significant differences among the CESAs is the CSR which differ in sequence and length; in contrast, the PCR regions differ in sequence only. Similarity of the CSR's of CESA's among their paralogs within same species is much lower than their orthologs from different species<sup>26</sup>. Cellulose like proteins that lack the Plant Conserved Region are responsible for synthesizing non-cellulosic polysaccharides<sup>3</sup>.

The importance of the CSR and PCR in plant cellulose production can be shown by the lack of mutational phenotypes traced to amino acid substitutions occurring within these regions. Only the FRA6 mutation in CESA8 has been shown to occur in the PCR<sup>27</sup>. To date, no mutations with observable phenotypes have been found to reside in the CSR. Artificial mutation of residues in CESA1 are phosphorylation sites<sup>28</sup>. CESA10 . CESA2. Chen et al. have postulated that these regions might have evolved to support or regulate CESA-microtubule interactions. However, it is also likely that these regions are what permit the formation of rosette TC's in Streptophytes and Charophyceae but noticeably absent in Chlorophyta linear TC forming. CESA diversification was necessary in the evolution of the rosette.

Without experimentally determined 3-D models of these features of plant CESAs, our ability to postulate mechanisms of function and gain insights into how these might foster rosette formation/interactions and affect cellulose synthesis is limited. To surmount this problem, we used ab-initio structure prediction via ROSETTA to provide three 3-D models of these PCR and CSR sequence fragments. We follow the sequence delineation of Roberts et al. 2002<sup>29</sup>.

## 5.2 Results

### 5.2.1 Decoy Production Analysis



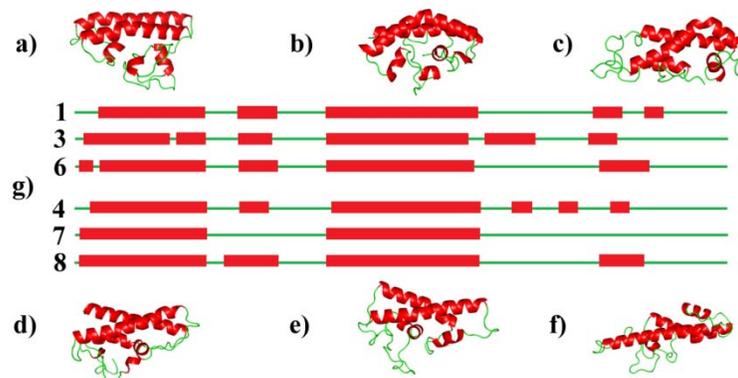
**Figure 5.2:** Representative Folding Funnels of Primary and Secondary PCR's and CSR's. The y-axis is calculated as RMSD values from the lowest scoring decoy. The x-axis is the score in terms of Rosetta Energy Units (R.E.U.). (a) PCR3 density plot (b) PCR3 folding funnel (cluster of interest is in red) (c) CSR3 density plot (d) CSR3 folding plot (cluster of interest is in red) (e) PCR7 density plot (f) PCR7 folding plot (cluster of interest is in red) (g) CSR7 density plot (h) CSR7 folding plot (cluster of interest is in red).

Since ROSETTA algorithm's *ab initio* structure prediction uses a Monte Carlo search with simulated annealing to generate putative decoys, the resulting ensemble resembles a 3-D Gaussian distribution; in Figure 5.2, we see that no natural funnel is apparent from the production runs. This problem led us to pursue a clustering analysis with respect to the best scoring structure under the premise that ROSETTA does provide decoys with features of the native conformation. CSR3 and PCR3 from AtCesa3 is representative of the primary wall associated CESA's. Likewise CSR7 and PCR7 from AtCesa7 represents the CESA's associated with secondary wall formation. The scoring function of ROSETTA is a cumulative function such that the longer sequences of the PCR's skew the ensemble scores to be lower than that observed for the CSR's. Typically, scores are comparative for equivalent length sequences. The RMSD component values are calculated relative to the best scoring decoy within the ensemble. A minimum production run of 10k decoys is generally acceptable. Some sequences generated acceptable models after only 13k generated structures. Others, such as PCR8, was remarkably difficult to fold even under the accommodative clustering threshold of 8 Å RMSD (**Table S2**). The most difficult structure to obtain acceptable structures of interest was CSR4 due mostly to its length of 132 amino acids which represents the extreme of the ROSETTA algorithm's ability under modest computational resources. We generated over half a million decoys in this effort. Free energy analysis required an additional 480 ns of aggregate molecular dynamics simulation time.

### **5.2.2 Plant Conserved Region**

The Plant Conserved Region is not present in other glycosyltransferase 2 family members. Occurring between the U1 and U2 motifs, the number of amino acids is the same

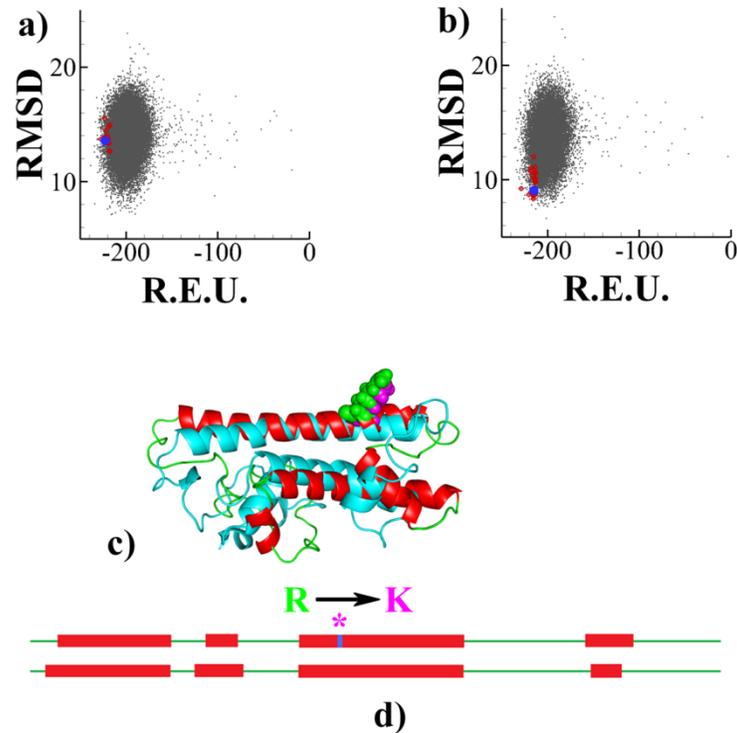
for all the CESA isoforms of Arabidopsis. Sequence alignment shows high conservation among PCR regions (Figure 5.S1, S2). However, little is known about the functional role of this region and the reasons for its high conservation. The predicted structures from ROSETTA *ab initio* show common structural features that persist among all the CESA fragments as seen in **Figure 5.3**. The folded tertiary structure of the PCR's are dominated by two helices separated by a smaller helix and random coil. The C-terminus after the second significant helix is random and shows a lower degree of sequence conservation. Hence, there is a plurality of small helices and random coils. The lack of conservation might suggest that this section of the PCR is less important for CESA function.



**Figure 5.3:** Primary (a-c) and Secondary (d-f) PCR's. (a) PCR1 (b) PCR3 (c) PCR6 (d) PCR4 (e) PCR7 (f) PCR8 (g) Secondary structure block diagram labeled according to each PCR. The N-terminus is to the left for the block diagram.

The only experimental evidence of a functional importance for the PCR has been observed in the FRA6 mutation within *Atcesa8*<sup>27</sup>. A single purine nucleotide mutation (guanine to adenine) results in an amino acid change of arginine to lysine (R362K). The

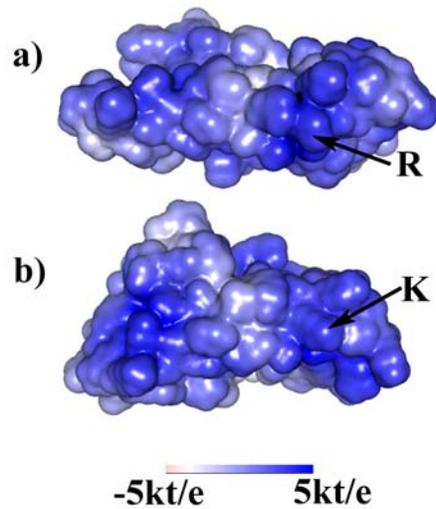
structure of the FRA6 mutant showed no appreciable deviation from the wildtype sequence of PCR8 (**Figure 5.4**).



**Figure 5.4:** Comparison of Wild Type PCR of Atcesa8 and the FRA6 mutant. (a) Folding funnel of PCR8 (b) Folding funnel of FRA6 mutant (c) Secondary structure comparison with tertiary alignment. An asterisk over the top block diagram is the lysine mutant of FRA6. The bottom block diagram is that of PCR8. Helices are in red with disordered regions as a green line.

However, arginine to lysine mutations aren't always neutral. When calculating the electrostatics surface potential, it can be seen that the FRA6 is less positive since lysine carries only a +1 charge compared to arginine's guanidinium group. This suggests that the PCR structure is more likely to be involved as a protein-protein interaction site (**Figure 5.5**). Indeed, the mutation occurs at the second arginine of a very long helix. The amino acid sequence of RRAMKR (fra6 position underlined), is invariant in all of the PCR fragments.

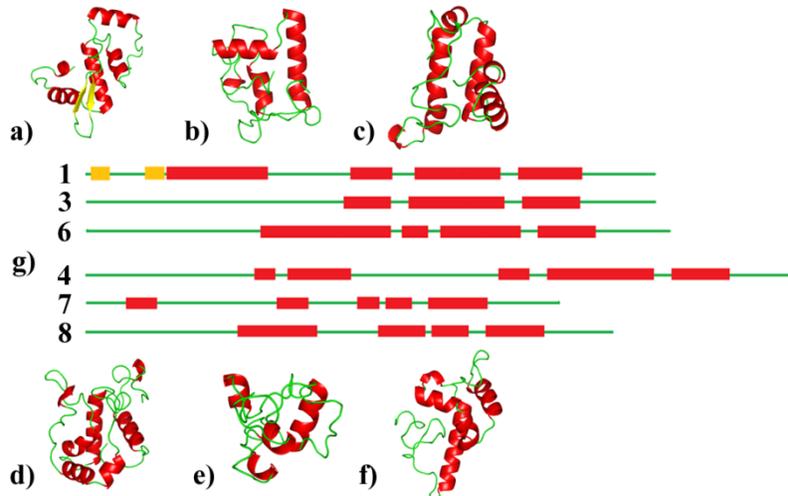
Homozygous *fra6* mutants have reduced secondary cellulose and lower fiber all thickness<sup>27</sup>. Since overexpression of the *fra6* mutant in a wild type plant also suggests that this mutation doesn't affect rosette formation and catalytic activity. Invariant arginines and lysines have been shown to be critical for protein structure and stability. These charged residues are facing outward in the same direction toward solvent. This localized region of positive charge forms an ideal recognition motif for interaction with anionic molecules or phosphorylated proteins. High affinity interactions of arginine rich motifs with phosphorylated serines have been shown to approach near covalent stability<sup>30</sup>. Electrostatics play a key role in protein-protein interaction and contribute significantly to complex stabilization and long range interactions<sup>31</sup>. The high conservation of this area of the PCR where the residues are solvent exposed would further indicate that it is a recognition motif for intermolecular interaction. The recent discovery of the CESA Interacting Protein (CSI1) could prove to be a likely candidate<sup>32</sup>. CSI is the first non-CESA protein to be associated with cellulose synthase complexes that are responsible for fibril formation. In other plants, there is more evidence that there is more to cellulose formation than the CSC alone<sup>33</sup>.



**Figure 5.5:** Electrostatic surface potential of the Wild Type Atcesa8 PCR (a) and FRA6 (b).

### 5.2.3 Class Specific Region

Recent experimental evidence have shown that in primary cell walls, the CSC of CESA1, CESA3, and CESA 6 form the hexameric complex<sup>18,34</sup>. However, in secondary cell walls formed by CESA4, CESA7, and CESA8 oligomers of dimers, tetramers, and hexamers were identified in *Arabidopsis* xylem tissue<sup>35</sup>. This presents is the stoichiometry problem of how many subunits do each of the CESA's contribute to the rosette. Moreover, it is within the second large cytosolic domain of CESA region where there is both variability in amino acid sequence and length; this second hypervariable region more recently classified by Vegara and Carpita as the Class Specific Region (CSR) was previously conjectured to be important for CESA-CESA interactions<sup>16</sup>.



**Figure 5.6:** Primary and Secondary CSR's: (a) CESA1 (b) CESA3 (c) CESA6 (d) CESA4 (e) CESA7 (f) CESA8 (g) Secondary Structure layout of a-f. The N-terminus is to the left for the block diagram.

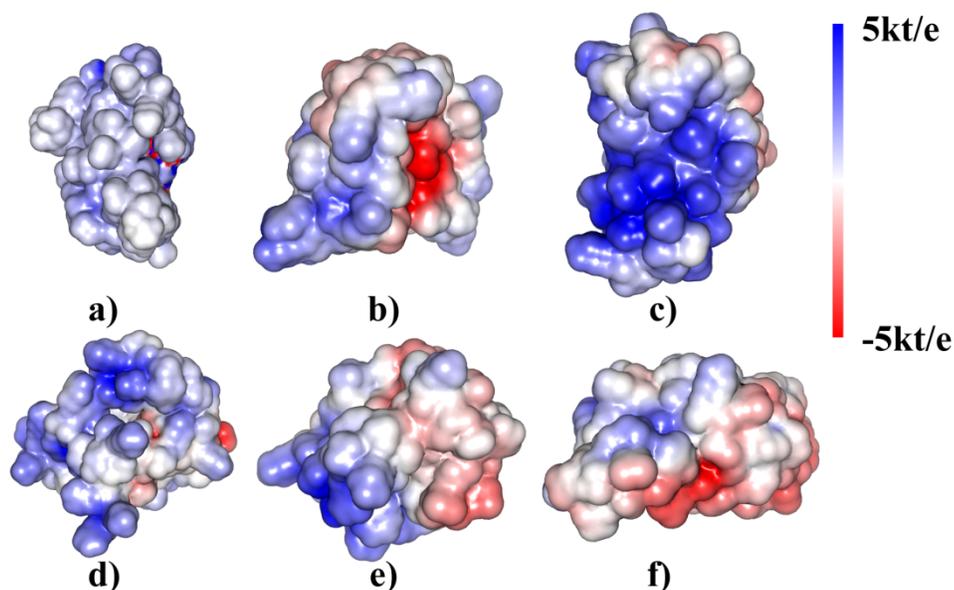
The final CSR models, **Figure 5.6**, show significant conformational variability.

**Figure 5.6(a-c)** are the primary CSR's. **Figure 5.6(g)** depicts a block diagram of the secondary structures with the N-terminus to the left. The secondary structures are referenced by enumeration from left to right. CSR1 is unusual where the best scoring decoys show the presence of two beta strands at the N-terminus and characterized by three prominent helices. The two C-terminus helices are conserved in both CSR3 and CSR6. In CSR3, its three helices are arranged in a near coplanar conformation. CSR6 is dominated by a long helix that aligns almost parallel to the third helix where the second helix is a short linking structure. The fourth helix lies with an axis nearly normal to the plane formed by the first and third helix. CSR6 also has the greatest composition of basic residues, 23.1%. The secondary wall associated CSR's, **Figure 5.6(d-f)**, are diverse in both conformation and amino acid length. The secondary cell wall CESA's are thought to have evolved later than

the primary cell wall. Overall, the CSR's show very low conservation among the CESA clades (**Figure 5.S4**). CSR4, **Figure 5.6(d)**, is notable for being the longest of the CSR fragments. It required the most decoys to be generated in order to obtain a good sampling and modeled structure from the ROSETTA algorithm. The CSR4 structure is divisible into four parts, an N-terminal disordered region, a short section of two helices, another stretch of random coils, and two helices. CSR7 is comprised of five relatively short helices and is the smallest of the CSR fragments. CSR8 is the second shortest fragment and features a prominent helix after a N-terminus disordered region. The remaining three helices are arranged near the C-terminal end of the longest helix. CSR8 contains the same total number of basic residues as CSR7.

#### **5.2.4 Protein Motifs**

Although the sequence and structure conservation is low, a conserved amino acid sequence consisting of EKxFGxS (**Figure 5.S3**) is apparent. The C-terminus of the CSR's are moderately conserved from this motif onward. The function of this motif at the base of a conserved helix remains unknown. The amount of basic residues as a percent of the fragment length varies as well. The most basic structure is CSR6 with 23% of its composition being either arginine or lysine. CSR6 is the most arginine rich. The least basic is CSR8 at 10.2%. The hydrophilic lysines and arginines are mostly solvent exposed and surround a core of polar and hydrophobic residues that may indicate that these regions of CESA are likely involved with protein-protein interactions. A majority of cytosolic proteins that associate with cellular membranes are also lysine rich.



**Figure 5.7:** Electrostatic surface potentials of the primary versus secondary wall associated class specific region folded fragments. CSR1 (a), CSR3 (b), CSR6 (c), CSR4 (d), CSR7 (e), and CSR8 (f).

CSR1 shows a classic Walker A motif  $GxxxxGK(T/S)$  which has been implicated in nucleotide binding; when the last amino acid is serine, the motif preferentially binds phosphate<sup>36</sup>. The motif is highlighted in **Figure 5.7a**. In CSR1, the Walker motif fits the definition as it is preceded by a  $\beta$ -strand and followed by an alpha helix. CSR1 is also the only structure known to shown to have two phosphorylation sites that results in altered trafficking of the rosette<sup>28</sup>. In the modeled structure, both of the serines, S686 and S688, are solvent exposed at the tip of a loop.

Within CSR3, two main helical structures are expected to be solvent exposed and are juxtaposed against an intrinsically disordered region of the N-terminus; the basic lysines and one arginine cluster into a basic patch. This region of unusual charge density most likely is a binding motif since. All of the CSR's display regions of general disorder that seems to arise

from their high concentration of lysine and arginine composition. In between the two parallel helices is an electronegative pocket formed by the acidic residues of aspartic and glutamic acid, **Figure 5.7b**.

CSR6 forms a significant cluster of basic lysine residues, **Figure 5.7c**, and presents the most oblique motif of the CSRs modeled. Additionally, CSR6 presents the most concentrated basic patch as evidenced by the charge distribution compared to that of the other CSRs.

Although the secondary wall associated CSRs vary significantly in amino acid length, they present less polarized conformations than the primary cell wall CSRs. CSR4, **Figure 5.7d**, is the largest fragment and least efficiently packed likely due to under sampling even with the 40,000 decoys generated. The significant length of CSR4 compared to the other CSRs would suggest that it evolved much later in the history of plant as evidenced by the lack of orthologs in the nonvascular plant *Physcomitrella patens*<sup>37</sup>. *Atcesa7* and *Atcesa8* also do not have orthologs in *P. patens*. Consequently, CSR7 (**Figure 5.7e**) and CSR8 (**Figure 5.7f**) show markedly different structures than CSR3 and CSR6. Carroll et al. showed that *Atcesa7* and *Atcesa3* were only partially interchangeable and the charge distribution of the CSR folds is suggestive of a similarity of charge distributions; interestingly, CSR3's surface potential has the same polarity CSR7 albeit with a more pronounced electronegative pocket.

In contrast to the other CSRs, CSR8 is the least basic. It also has the highest concentration of polar residues that is near the N-terminus and forms a stretch of eight residues before encountering a triplet of lysines. This arrangement of residues is suggestive

of CSR8 being a key interacting region with another as yet unidentified protein that is involved in cellulose synthesis.

### **5.2.5 MD Analysis**

Due to the limitations with the conformational sampling method employed ROSETTA, we subjected the centroid structures to molecular dynamics simulations in order to ascertain their relative stability in an explicit solvent environment. The forcefield and energy scoring methodology employed by ROSETTA is not directly portable to normal units of kcal/mol. In Table S3, the free energies of the PCR and CSR structures are tabulated on a per residue and total basis. The most energetic structure in absolute terms is PCR4 while the least is that of CSR7. CSR7 also has the lowest calculated solvation free energy. In contrast, the structure expected to be the most soluble is CSR6 possibly due to its abundance of lysine. Additionally, a brief examination of the retention of their structural changes was explored with a conformational analysis, **Table S5**. PCR8 displays the most conformational immutability reaching 3.03 Å RMSD relative to its initial structure after 18ns. The most unstable structure seems to be CSR4 due to its bulk by reaching 5.82 Å RMSD after a mere 12ns of simulation.

## **5.3 Discussion**

### **5.3.1 Rosette Formation and Hierarchy of Terminal Complex Assembly**

For the vast majority of cellulose producing organisms, there is only one type of cellulose synthase in use. But for plants, there are six isoforms. The evolutionary drive to have more than one type of cellulose synthase in plants is likely due to the development of

specialized tissues for organs such as roots, stems, and leaves. Although hemicellulose, pectin, cellulose and lignin comprise the cell wall, it is the latter two that are the majority constituents. Cellulose comprises between 40 to 50% of xylem tissue alone. Cellulose from primary and secondary cell walls differ in degrees of polymerization (DP) and microfibril crystallinity; the more supple primary wall cellulose varies in DP from 2,000 to 6,000 while that of secondary cellulose is at minimum 15,000<sup>38,39</sup>. Mutations that affect cellulose synthesis have been observed in 6 of the 10 CESAs of Arabidopsis<sup>38</sup>. The only fundamental distinction between the CESAs are their class specific regions. Since the isoforms that constitute the primary and secondary cell wall associated rosettes are not fully interchangeable, the observations to date would signify that the molecular machinery involved with cellulose synthesis is made up of as yet many other protein actors. While all land plants have rosette terminal complexes, only vasculature forming plants have the secondary wall associated CESA orthologs<sup>40</sup>; it is clear that the control of CESA assembly is important in generating tissue specific cellulose as seen in the slime mold *D. discoideum* whose terminal complexes may form either a close clustered packing or linear arrays<sup>41</sup>. The structure of a bacterial cellulose synthase that has been recently solved confirms that the absence of these regions<sup>9</sup>.

In folding the PCR and CSR fragments, we hoped that their structure would give insight into how each might contribute to rosette formation and terminal complex aggregation; from our results, we believe that the CSR is principally important in the control of terminal complex formation from rosettes wherein it was previously suggested that both the PCR and CSR are involved<sup>37</sup>. Since the PCR regions are not particularly different amongst

the isoforms in structure and sequence, it is unlikely that they serve to modulate TCs. It is possible that the PCR region accounts for rosette formation. Given that the FRA6 mutation is only detrimental in homozygous genotypes, it is more likely that it serves as region that interacts with non-CESA proteins, perhaps CSI1<sup>32</sup>. However, CSI1 has only been shown to interact with primary wall associated CESA<sup>32,42</sup>. What

It had been postulated that perhaps thiol sensitive amino acid sequences might be involved in multimeric assembly of the CESAs<sup>43</sup>. At first, it may seem that perhaps either the PCR or CSR regions may have a part in CESA oligomerization<sup>16</sup>. In Arabidopsis, both the PCR and CSR regions are highly basic due to the presence of arginine and lysine residues. In contrast, the region in Dictyostelium CESA that corresponds to the plant CSR has a bias toward acidic residues. Of what significance that this might influence terminal complex assembly remains to be investigated. However, arginines are prevalent at dimer interfaces (13.3%), but lysines (6.29%) are not<sup>44,45</sup>. Although predominant, stretches of polylysine have been shown to be important in the oligomerization of other proteins such as K-RasB, HSP90B. Lysine rich subdomains are more often sites of interaction with other proteins as well as for membrane association. GPI-anchored proteins also have these lysine rich subdomains. The mechanism of CESA subunit association still remains an open question; given the putative structures resulting from this work, it is probable that noncovalent interactions and disulfide bonds stabilize the CESA subunits in the rosettes as postulated by Atanassov et al. 2009<sup>35</sup>.

If the PCR and CSR are not responsible for CESA subunit oligomerization since non-plant CESAs can still self-associate even if only as linear arrays or close packed clusters,

what are their putative roles in cellulose production of higher plants? A possible explanation might be to control the polymerization and organization of cellulose by restructuring the nature of the terminal complex assembly.

### **5.3.2 Achieving material control of cellulose**

The essential material properties that distinguish cellulose from primary and secondary cell walls is its crystallinity and the degree of polymerization. Early freeze fracture images of *Micrasterias denticulata* primary and secondary cell wall formation show that the aggregation of the TC's account for the macrofibrillar differences between the two processes<sup>46</sup>. Recent studies have demonstrated the coupling of increased saccharification to lower crystallinity cellulose as well as overall lower cellulose production<sup>47</sup>. Early freeze fracture images of the rosette terminal complexes likely imaged only the transmembrane portion of the protein complex<sup>48</sup>. However, existing models of rosette construction have assumed a CESA-only phenomena and largely derived from observation of the primary wall system. These models of CESA interaction have largely been confined to CSC assembly and not specifically to the terminal complex. The Scheible and Doblin model envisioned tertiary interactions among two general types of occupying specific locations within the rosette. Subsequent models of rosette formation remain variations on slight modifications of this original hypothesis. It had been postulated that secondary cell wall CesA isoforms randomly assemble<sup>49</sup>; however it would imply identical functionality which is clearly not the case, so the random interaction model must be rejected. A random insertion model of rosette assembly been considered earlier, but the specificity and co-dependency of the CesA isoforms for the proper formation of cellulose signified greater molecular control<sup>24</sup>.

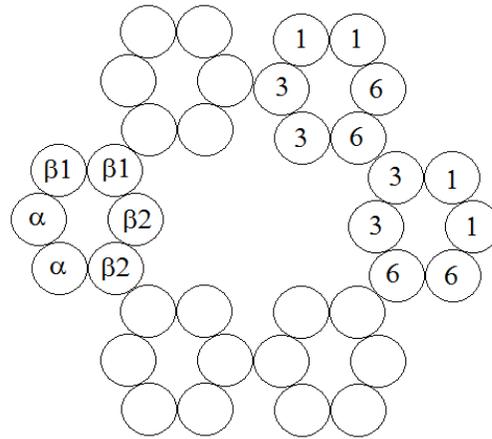
Recent *in vitro* and *in planta* data show in essence that primary and secondary CESAs may interact via heterodimerization<sup>25</sup>. In a mixed expression, Cesa7 could partially rescue Cesa3 in a primary wall CSC. Cesa1 could partially rescue a Cesa8 knock out, but not the reverse. This experiment had been proposed earlier by Taylor et al REF but only completed just recently. Wang et al had previously shown that the CSR region encoded the positional arrangement of Cesa's into the rosette. The difficulty with in interpreting association studies have been interpreting what reduces to a symmetry problem. Within both the primary and secondary wall CSC, there is non-compensation of one isoform position for another. With the fragment folding of this work, we can further break the symmetry.

Why should the rosette structure matter and why are there six classes of CesAs? In order to attain a hexameric assembly, at least one out of the three CesAs must not be allowed to self-associate in *planta*. A hexameric fiber is stiffer than a dimer or tetramer<sup>50</sup>. The polymerization rate is theoretically same for both primary and secondary cell wall CSC's if one assumes that 700 glucose units per chain per minute<sup>51</sup>. With rosette lifetimes of approximately 20 minutes, it is conceivable to attain a chain 14,000 units long which is the typical "DP" for secondary cell wall cellulose and relatively consistent. Hence, in order to obtain smaller fibrils with primary cell wall CSCs, an editing mechanism to create shorter fibers is required. The synthesis of cellulose for the primary cell wall needs an "interrupt" mechanism. Since Cesa6 is the most varied isoform, it is conceivable that position in the primary cell wall rosette performs this function. Since Cesa1 can partially rescue Cesa8 knockouts in a coexpressed system and Cesa7 may do the same for a Cesa3 knockout, this is conceivable<sup>25</sup>. The DP for primary cell walls ranges between 2000 to 6000; this definitely

results from all the various combinations CSCs incorporating the CesaA6 class of CesaA's as well as CesaA10 which is a near identical copy of CesaA1. Interestingly there is a biphasic pool of primary cellulose DP into those less than 500 and those ranging between approximately 2500 to 4500 REF. Wightman reports a significantly higher rate of deposition (17,000 units/sec) in newly forming secondary wall xylem<sup>52</sup>. Paredez had reported 16 units/sec<sup>51</sup> in primary wall formation visually tracking at YFP-CESA6 construct. Until there is a greater understanding of the key differences in secondary versus primary wall synthesis, these stark contrasts in cellulose production remains a conundrum.

In vitro demonstrations of interactions between CESAs have shown that the zinc binding domains can confer the ability to homo and hetero dimerize<sup>53</sup>. CESA oxidation was shown to be a prerequisite prior to interaction (Peng et al. 2001, Kurek 2002)<sup>54,55</sup>. However, the strength of the zinc binding domains alone has been suggested as being insufficient to confer specific protein-protein interactions. By elimination, the remaining distinct regions of consequence are the PCR and CSR. There may also be a role for the C-terminal sequence after the eight transmembrane helix in a conserved proline as evidenced by the difficulty of the rsw5 CesaA3 mutant inserting itself into a CSC. In spite of this, the CSR appears to be the best candidate to provide specific and strong interactions since the PCR is relatively undifferentiated amongst the CESA classes. Since the positioning in the primary wall CSC seems to be unimportant, but that is more controlled in secondary wall CSC. We conjecture a simple trimer of homodimers architecture based on the knockout rescues and the supposition of independent homodimerization prior to CSC assembly, **Figure 5.8**, and bears some resemblance to a model proposed by Robert et al. 2004<sup>56</sup>. The homodimers are

expected to be more condensed such there is no “inside” stray CESA subunit. This model of rosette structure is also consistent with was described by Anatassanov et al. 2009<sup>35</sup>. A preliminary dimer assembly using a predicted cytosolic model of Ghcesa1 shows that this may be possible where the CSR and PCR regions are facing outward.



**Figure 5.8:** Unified Trimer of Dimers CSC model to deconflict the results of Carroll et al.<sup>25</sup> and Wightman et al.<sup>52</sup>

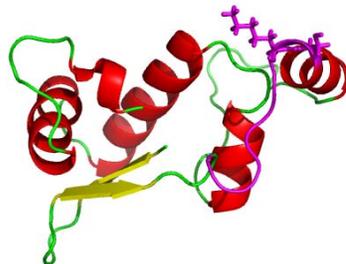
### 5.3.3 Resolving Paradoxes

The key distinction between primary wall and secondary wall associated CSCs is that the secondary wall CSC’s form very large TC arrays and are very dense. However, they evolved from primary wall CESAs. Through correspondence, a CESA in one CSC system that can partially recover the function in another system essentially “accesses” the same relative location of the other CSC system. However, the fact that not all dualities were functional indicates that there is less symmetry in one system – namely the primary cell wall CSC. CSR1 is the only CSR that has been proven to have two phosphorylation sites and

therefore, must be solvent accessible and turned away from the center of the CSC. The homodimers of CESA3 and CESA6 then are able to interact. If one translates this same model to the secondary cell wall system, then it would partially explain some of the experimental results seen most recently. Then the apparent randomness of the secondary cell wall CESA associations would indicate that all the CESAs of that system can interact with each other; therefore, there is some rotational freedom in the association of the rosette formed by CESA4, CESA7, and CESA8<sup>49</sup>. If one looks at CSR7 and CSR3, the charge distribution look fairly similar although not in the clustered intensity, **Figure 5.7e and 7b**, respectively. If the CSR region is important for CSC formation through electrostatic interactions, then a dimer of CESA7 could reasonably stabilize a primary wall rosette. Why CESA3 could not rescue a CESA7 knockout would imply that CESA8 and CESA3 do not interact in vivo since CESA8 could not recover a CESA1 knockout. Interestingly, the common feature of these two CSC's is that they both contain CESA7 with CESA1. CESA6 and CESA4 in their assumed positions would seem the key distinguishing mechanics of primary versus secondary CSCs. This observation would further support a unified CSC correspondence model.

This is consistent with the assumption that CESA6 and CESA4 occupy corresponding positions within the CSC. CESA4 has been shown to be important for the high DP of secondary cell wall cellulose and crystallinity. Equivalently, the CESA6-like CESAs, 2, 5, and 9 are substituted into the CESA6 position during various stages of plant growth. Given the difficulty in explaining how differences in the CSR may account for the rosette topologies observed in primary and secondary cell wall growth, it is likely that the CSR

regions themselves are binding motifs for as yet unknown proteins that interact with the terminal complex. Bowling et al. imaged rather large cytosolic hexameric assemblies of a 50nm mean diameter and approximately 30-35nm tall<sup>48</sup>. The hexameric model of with the main catalytic region of Ghcesa1, **Figure 5.1**, only measures a mean diameter of 12.5nm; translating this dimension to encompass at TC would amount to at least 37 to 40 nm at best. Even the inclusion of the N-terminus domain preceding the first transmembrane helix would be insufficient to span the difference. The greatest indication that additional proteins are involved in the TC formation is the height of the granules seen. The recent discovery of additional proteins critical to cellulose biosynthesis such as CSI1, MUR10, PttMAP20, KOR, and eEF-1B $\beta$ 1 also likely indicates that CESAs are only part of a more intricate system<sup>32,57-60</sup>. There will likely be more proteins discovered to be involved in cellulose biosynthesis.



**Figure 5.9:** Walker P Loop (GXXXGK[T/S]) of CSR1 highlighted in magenta with the serine and arginine in line form. The loop is a binding motif for the phosphate of nucleotides.

If we ignore the differences in the CSRs among the isoforms, there is essentially only one type of CESA with a conserved catalytic mechanism. This interpretation does not contradict the results of Wang et al. 2006<sup>61</sup>. Hence, the CSR and PCR regions result from a

requirement to “fit” a functional catalytic system on top of a protein complex that likely regulates substrate generation and fibril length. The lower level of variance seen with the PCRs would further suggest that the CSCs of primary and secondary wall systems share common protein interactions. However, the Walker A motif in CSR1, **Figure 5.9**, which also is also the only CSR with two phosphorylation sites can possibly bind to a nucleotide derivative like cyclic-di-GMP which is known to regulate post-translational cellulose production in bacteria<sup>62-64</sup>. Another possibility is that the CSR regions regulate trafficking of the CESAs since CESA7 insertion into a primary CSC increased the velocity of observed complexes<sup>25</sup>. If these small pieces of evidence might be a guide, then the CSRs of CESAs serve a regulatory role in cellulose synthesis and allow plants to gain fine control over the mechanical properties of the cell wall and, thus, cell morphology. The reason for having a hexameric engine for cellulose generation might ultimately be able to control the directionality of the fibrils being produced. Granted, microtubule association for alignment has been implicated, it may not be sufficient alone to mold the cell wall architecture. A hexameric assembly allows for omni-directional control advantages over a tetramer based system given the need to also control microfibrillar angle in secondary wall formation. Ultimately, more needs to be done to uncover all the actors involved with cellulose biogenesis in plants.

## 5.4 Methods

The ROSETTA ab initio modeling algorithm was used to generate a minimum of 20,000 decoys for each sequence and up to 40,000 decoys for those proving difficult to fold as was the case for the class specific region of Atcesa4. Ab initio modeling is used when no known comparative homologs exist for the protein of interest. The aim is then to be able to take the top 10 percent best scoring structures such that the entropy based exact clustering with Durandal<sup>65</sup> would result in a “good” cluster size of at least 10 structures within a cut off value of at least 8 Å. The exact clustering algorithm of Durandal smooths the free energy function in ROSETTA by searching for the structure with the maximum number of neighbors given a cutoff value; subsequently, this cluster is removed from the remaining pool of structures. This process is repeated until no set of structures remains in the pool of structures. The largest cluster from this process is the one most likely to contain the nearest native like structure. The centroid that represents the average inter-similarity of the cluster is taken as the structure of interest. When a folding funnel is present, the model with the most ‘native’ like features is often the lowest energy structure. The optimal structure is then relaxed using the ROSETTA abinitio relax method using a weak constraint. Our installation of the ROSETTA fragment library database was made using the 2010 NCBI non redundant protein database. Tecplot was used to generate the folding graphs. PyMOL was used to capture and render the protein structures<sup>66</sup>.

Fully solvated MD simulations were conducted with the AMBER 11 software suite with the FF10 force field<sup>67</sup>. Minimized protein structures were then neutralized with Na<sup>+</sup> ions and immersed in a water box with at least 10 Å deep solvation shell using the TIP3P

water model<sup>68</sup>. Additional Na<sup>+</sup> and Cl<sup>-</sup> ions were added to represent a 0.3 M effective salt concentration around a given NA helix. The equilibration of each NA sample was carried out in 11 stages starting from the solvent minimization for 10000 steps and keeping the duplex restrained for 200 Kcal/mol. The system was heated to 300K in 40 ps while imposing a 200kcal/mol constraint on the duplex. A brief NPT MD run was performed for 200 ps with a duplex restrained maintained at 200 kcal/mol. Another constrained minimization step follows with the restraint of 25 kcal/mol for 10000 steps. A second NPT MD run was performed at 25 kcal/mol restraint for 20 ps. Subsequently four additional 1000 cycle minimization steps were performed while relaxing the positional constraint from 20 kcal/mol to 5 kcal/mol in 5 kcal/mol increments. A final unconstrained minimization stage of 1000 cycles was performed before reheating the system to 300K at constant volume within 40 ps. Subsequently, NPT equilibrations were performed to ensure uniformity in solvent density. Long range electrostatic interactions were calculated by Particle Mesh Ewald summation (PME)<sup>69</sup> and the non-bonded interactions were truncated at 9 Å cutoff along with a 0.00001 tolerance of Ewald convergence. A Berendsen thermostat maintained temperature at 300 K<sup>70</sup>. The SHAKE algorithm was used to constrain the position of hydrogen atoms<sup>71</sup>. The production simulations were performed for an NVT ensemble. Each production simulation was performed for 20 ns with a 2 fs time step. Free Energy was performed with the MMPBSA tool of AMBER 11 software suite.

## 5.5 Conclusion

The modeling performed highly suggest that the Plant Conserved Region is a candidate for intermolecular interaction. The *fra6* mutant form of the PCR from AtCESA8 does not alter the global conformation of the folded fragment. The localization of basic residues in the region of the *fra6* mutation site is consistent with known patterns of recognition specific arginine/lysine epitopes and suggestive of unknown protein-protein interactions. The folding of the distinctive class specific regions into largely unconserved conformations lends to a theory that they are responsible for CESA insertion and positioning over a larger protein complex responsible for regulating cellulose synthesis. The presence of a Walker A motif within CSR1 implicates it is as a likely regulatory domain.

## Acknowledgements

This work was supported as part of The Center for LignoCellulose Structure and Formation, Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Science under Award Number DE-SC0001090.

## References

1. Nobles DR, Romanovicz DK, Brown RM. Cellulose in cyanobacteria. Origin of vascular plant cellulose synthase? *Plant Physiology* 2001;127(2):529-542.
2. Yin YB, Huang JL, Xu Y. The cellulose synthase superfamily in fully sequenced plants and algae. *Bmc Plant Biology* 2009;9.
3. Richmond TA, Somerville CR. The cellulose synthase superfamily. *Plant Physiology* 2000;124(2):495-498.
4. Williamson RE, Burn JE, Hocart CH. Cellulose synthesis: mutational analysis and genomic perspectives using *Arabidopsis thaliana*. *Cellular and Molecular Life Sciences* 2001;58(10):1475-1490.

5. Saxena IM, Brown RM, Dandekar T. Structure-function characterization of cellulose synthase: relationship to other glycosyltransferases. *Phytochemistry* 2001;57(7):1135-1148.
6. Nobles DR, Brown RM. The pivotal role of cyanobacteria in the evolution of cellulose synthases and cellulose synthase-like proteins. *Cellulose* 2004;11(3-4):437-448.
7. Saxena IM, Brown RM. Cellulose synthases and related enzymes. *Current Opinion in Plant Biology* 2000;3(6):523-531.
8. Nagahashi S, Sudoh M, Ono N, Sawada R, Yamaguchi E, Uchida Y, Mio T, Takagi M, Arisawa M, Yamadaokabe H. Characterization of Chitin Synthase-2 of *Saccharomyces-Cerevisiae* - Implication of 2 Highly Conserved Domains as Possible Catalytic Sites. *Journal of Biological Chemistry* 1995;270(23):13961-13967.
9. Morgan J, Strumillo J, Zimmer J. Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature*. Volume advanced online publication; 2012.
10. Popper ZA, Tuohy MG. Beyond the Green: Understanding the Evolutionary Puzzle of Plant and Algal Cell Walls. *Plant Physiology* 2010;153(2):373-383.
11. Roberts E, Roberts AW. A Cellulose Synthase (Cesa) Gene from the Red Alga *Porphyra Yezoensis* (Rhodophyta). *Journal of Phycology* 2009;45(1):203-212.
12. Sorensen I, Domozych D, Willats WGT. How Have Plant Cell Walls Evolved? *Plant Physiology* 2010;153(2):366-372.
13. Tsekos I. The sites of cellulose synthesis in algae: Diversity and evolution of cellulose-synthesizing enzyme complexes. *Journal of Phycology* 1999;35(4):635-655.
14. Kimura S, Chen HP, Saxena IM, Brown RM, Itoh T. Localization of c-di-GMP-Binding protein with the linear terminal complexes of *Acetobacter xylinum*. *Journal of Bacteriology* 2001;183(19):5668-5674.
15. Nakashima J, Heathman A, Brown RM. Antibodies against a *Gossypium hirsutum* recombinant cellulose synthase (Ces A) specifically label cellulose synthase in *Micrasterias denticulata*. *Cellulose* 2006;13(2):181-190.
16. Vergara CE, Carpita NC. beta-D-Glycan synthases and the CesaA gene family: lessons to be learned from the mixed-linkage (1 -> 3),(1 -> 4)beta-D-glucan synthase. *Plant Molecular Biology* 2001;47(1-2):145-160.

17. Sullivan S, Ralet MC, Berger A, Diatloff E, Bischoff V, Gonneau M, Marion-Poll A, North HM. CESA5 Is Required for the Synthesis of Cellulose with a Role in Structuring the Adherent Mucilage of Arabidopsis Seeds. *Plant Physiology* 2011;156(4):1725-1739.
18. Desprez T, Juraniec M, Crowell EF, Jouy H, Pochylova Z, Parcy F, Hofte H, Gonneau M, Vernhettes S. Organization of cellulose synthase complexes involved in primary cell wall synthesis in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104(39):15572-15577.
19. Scheible WR, Pauly M. Glycosyltransferases and cell wall biosynthesis: novel players and insights. *Current Opinion in Plant Biology* 2004;7(3):285-295.
20. Mendu V, Griffiths JS, Persson S, Stork J, Downie AB, Voiniciuc C, Haughn GW, DeBolt S. Subfunctionalization of Cellulose Synthases in Seed Coat Epidermal Cells Mediates Secondary Radial Wall Synthesis and Mucilage Attachment. *Plant Physiology* 2011;157(1):441-453.
21. Chu ZQ, Chen H, Zhang YY, Zhang ZH, Zheng NY, Yin BJ, Yan HY, Zhu L, Zhao XY, Yuan M, Zhang XS, Xie Q. Knockout of the AtCESA2 gene affects microtubule orientation and causes abnormal cell expansion in Arabidopsis. *Plant Physiology* 2007;143(1):213-224.
22. Stork J, Harris D, Griffiths J, Williams B, Beisson F, Li-Beisson Y, Mendu V, Haughn G, DeBolt S. CELLULOSE SYNTHASE9 Serves a Nonredundant Role in Secondary Cell Wall Synthesis in Arabidopsis Epidermal Testa Cells. *Plant Physiology* 2010;153(2):580-589.
23. Betancur L, Singh B, Rapp RA, Wendel JF, Marks MD, Roberts AW, Haigler CH. Phylogenetically Distinct Cellulose Synthase Genes Support Secondary Wall Thickening in Arabidopsis Shoot Trichomes and Cotton Fiber. *Journal of Integrative Plant Biology* 2010;52(2):205-220.
24. Doblin MS, Kurek I, Jacob-Wilk D, Delmer DP. Cellulose biosynthesis in plants: from genes to rosettes. *Plant and Cell Physiology* 2002;43(12):1407-1420.
25. Carroll A, Mansoori N, Li SD, Lei L, Vernhettes S, Visser RGF, Somerville C, Gu Y, Trindade LM. Complexes with Mixed Primary and Secondary Cellulose Synthases Are Functional in Arabidopsis Plants. *Plant Physiology* 2012;160(2):726-737.
26. Ranik M, Myburg AA. Six new cellulose synthase genes from Eucalyptus are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* 2006;26(5):545-556.

27. Zhong RQ, Morrison WH, Freshour GD, Hahn MG, Ye ZH. Expression of a mutant form of cellulose synthase AtCesA7 causes dominant negative effect on cellulose biosynthesis. *Plant Physiology* 2003;132(2):786-795.
28. Chen SL, Ehrhardt DW, Somerville CR. Mutations of cellulose synthase (CESA1) phosphorylation sites modulate anisotropic cell expansion and bidirectional mobility of cellulose synthase. *Proceedings of the National Academy of Sciences of the United States of America* 2010;107(40):17188-17193.
29. Roberts AW, Roberts EM, Delmer DP. Cellulose synthase (CesA) genes in the green alga *Mesotaenium caldariorum*. *Eukaryotic Cell* 2002;1(6):847-855.
30. Woods AS, Ferre S. Amazing stability of the arginine-phosphate electrostatic interaction. *Journal of Proteome Research* 2005;4(4):1397-1402.
31. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. *Current Opinion in Structural Biology* 2000;10(2):153-159.
32. Gu Y, Kaplinsky N, Bringmann M, Cobb A, Carroll A, Sampathkumar A, Baskin TI, Persson S, Somerville CR. Identification of a cellulose synthase-associated protein required for cellulose biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 2010;107(29):12866-12871.
33. Song DL, Shen JH, Li LG. Characterization of cellulose synthase complexes in *Populus* xylem differentiation. *New Phytol* 2010;187(3):777-790.
34. Wang J, Elliott JE, Williamson RE. Features of the primary wall CESA complex in wild type and cellulose-deficient mutants of *Arabidopsis thaliana*. *Journal of Experimental Botany* 2008;59(10):2627-2637.
35. Atanassov II, Pittman JK, Turner SR. Elucidating the Mechanisms of Assembly and Subunit Interaction of the Cellulose Synthase Complex of *Arabidopsis* Secondary Cell Walls. *Journal of Biological Chemistry* 2009;284(6):3833-3841.
36. Kinoshita K, Sadanami K, Kidera A, Go N. Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. *Protein Engineering* 1999;12(1):11-14.
37. Roberts AW, Roberts E. Cellulose synthase (CesA) genes in algae and seedless plants. *Cellulose* 2004;11(3-4):419-435.
38. Somerville C. Cellulose synthesis in higher plants. *Annu Rev Cell Dev Bi* 2006;22:53-78.

39. Harris D, DeBolt S. Relative Crystallinity of Plant Biomass: Studies on Assembly, Adaptation and Acclimation. *Plos One* 2008;3(8).
40. Roberts AW, Bushoven JT. The cellulose synthase (CESA) gene superfamily of the moss *Physcomitrella patens*. *Plant Molecular Biology* 2007;63(2):207-219.
41. Grimson MJ, Haigler CH, Blanton RL. Cellulose microfibrils, cell motility, and plasma membrane protein organization change in parallel during culmination in *Dictyostelium discoideum*. *J Cell Sci* 1996;109:3079-3087.
42. Li SD, Lei L, Somerville CR, Gu Y. Cellulose synthase interactive protein 1 (CS11) links microtubules and cellulose synthase complexes. *Proceedings of the National Academy of Sciences of the United States of America* 2012;109(1):185-190.
43. Carpita NC. Update on Mechanisms of Plant Cell Wall Biosynthesis: How Plants Make Cellulose and Other (1 → 4)-beta-D-Glycans. *Plant Physiology* 2011;155(1):171-184.
44. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol* 1998;280(1):1-9.
45. Moreira IS, Fernandes PA, Ramos MJ. Hot spots-A review of the protein-protein interface determinant amino-acid residues. *Proteins* 2007;68(4):803-812.
46. Giddings TH, Brower DL, Staehelin LA. Visualization of Particle Complexes in the Plasma-Membrane of *Micrasterias-Denticulata* Associated with the Formation of Cellulose Fibrils in Primary and Secondary Cell-Walls. *Journal of Cell Biology* 1980;84(2):327-339.
47. Harris DM, Corbin K, Wang T, Gutierrez R, Bertolo AL, Petti C, Smilgies DM, Estevez JM, Bonetta D, Urbanowicz BR, Ehrhardt DW, Somerville CR, Rose JKC, Hong M, DeBolt S. Cellulose microfibril crystallinity is reduced by mutating C-terminal transmembrane region residues CESA1(A903V) and CESA3(T942I) of cellulose synthase. *Proceedings of the National Academy of Sciences of the United States of America* 2012;109(11):4098-4103.
48. Bowling AJ, Brown RM. The cytoplasmic domain of the cellulose-synthesizing complex in vascular plants. *Protoplasma* 2008;233(1-2):115-127.
49. McDonnell L. Investigating the Role of Cellulose Synthases in the Biosynthesis and Properties of Cellulose in Secondary Cell Walls [Doctoral dissertation]. Vancouver: University of British Columbia; 2010. 198 p.

50. Shen T, Langan P, French AD, Johnson GP, Ganankaran S. Shift in conformational flexibility of soluble cellulose oligomers with increasing chain length. *J Am Chem Soc* 2009;131(41):14786-14794.
51. Paredez AR, Somerville CR, Ehrhardt DW. Visualization of cellulose synthase demonstrates functional association with microtubules. *Science* 2006;312(5779):1491-1495.
52. Wightman R, Marshall R, Turner SR. A Cellulose Synthase-Containing Compartment Moves Rapidly Beneath Sites of Secondary Wall Synthesis. *Plant and Cell Physiology* 2009;50(3):584-594.
53. Xu F, Joshi CP. In vitro demonstration of interactions among zinc-binding domains of cellulose synthases in Arabidopsis and aspen. *Advances in Biosciences and Biotechnology* 2010;1:152-161.
54. Kurek I, Kawagoe Y, Jacob-Wilk D, Doblin M, Delmer D. Dimerization of cotton fiber cellulose synthase catalytic subunits occurs via oxidation of the zinc-binding domains. *Proceedings of the National Academy of Sciences of the United States of America* 2002;99(17):11109-11114.
55. Peng LC, Xiang F, Roberts E, Kawagoe Y, Greve LC, Kreuz K, Delmer DP. The experimental herbicide CGA 325'615 inhibits synthesis of crystalline cellulose and causes accumulation of non-crystalline beta-1,4-glucan associated with CesaA protein. *Plant Physiology* 2001;126(3):981-992.
56. Robert S, Mouille G, Hofte H. The mechanism and regulation of cellulose synthesis in primary walls: lessons from cellulose-deficient Arabidopsis mutants. *Cellulose* 2004;11(3-4):351-364.
57. Bosca S, Barton CJ, Taylor NG, Ryden P, Neumetzler L, Pauly M, Roberts K, Seifert GJ. Interactions between MUR10/CesA7-dependent secondary cellulose biosynthesis and primary cell wall structure. *Plant Physiology* 2006;142(4):1353-1363.
58. Rajangam AS, Kumar M, Aspeborg H, Guerriero G, Arvestad L, Pansri P, Brown CJL, Hober S, Blomqvist K, Divne C, Ezcurra I, Mellerowicz E, Sundberg B, Bulone V, Teeri TT. MAP20, a Microtubule-Associated Protein in the Secondary Cell Walls of Hybrid Aspen, Is a Target of the Cellulose Synthesis Inhibitor 2,6-Dichlorobenzonitrile. *Plant Physiology* 2008;148(3):1283-1294.
59. Sato S, Kato T, Kakegawa K, Ishii T, Liu YG, Awano T, Takabe K, Nishiyama Y, Kuga S, Sato S, Nakamura Y, Tabata S, Shibata D. Role of the putative membrane-bound endo-1,4-beta-glucanase KORRIGAN in cell elongation and cellulose synthesis in Arabidopsis thaliana. *Plant and Cell Physiology* 2001;42(3):251-263.

60. Hossain Z, Amyot L, McGarvey B, Gruber M, Jung J, Hannoufa A. The Translation Elongation Factor eEF-1B beta 1 Is Involved in Cell Wall Biosynthesis and Plant Development in *Arabidopsis thaliana*. *Plos One* 2012;7(1).
61. Wang J, Howles PA, Cork AH, Birch RJ, Williamson RE. Chimeric proteins suggest that the catalytic and/or C-terminal domains give CesA1 and CesA3 access to their specific sites in the cellulose synthase of primary walls. *Plant Physiology* 2006;142(2):685-695.
62. Brown RM, Saxena IM. Cellulose biosynthesis: A model for understanding the assembly of biopolymers. *Plant Physiol Bioch* 2000;38(1-2):57-67.
63. Jenal U, Malone J. Mechanisms of cyclic-di-GMP signaling in bacteria. *Annu Rev Genet* 2006;40:385-407.
64. Le Quere B, Ghigo JM. BcsQ is an essential component of the *Escherichia coli* cellulose biosynthesis apparatus that localizes at the bacterial cell pole. *Mol Microbiol* 2009;72(3):724-740.
65. Berenger F, Shrestha R, Zhou Y, Simoncini D, Zhang KYJ. Durandal: Fast exact clustering of protein decoys. *J Comput Chem* 2012;33(4):471-474.
66. The PyMOL Molecular Graphics System. 1.2r3pre: Schrodinger, LLC.
67. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, C. Tan, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, C. Sagui, Babin V, Kollman. PA. Amber 11: University of California, San Francisco; 2010.
68. William L. Jorgensen JC, and Jeffry D. Madura. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;79.
69. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* 1995;103(19):8577-8593.
70. Berendsen HJC, Postma JPM, Gunsteren WFv, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* 1984;81(8):3684-3690.
71. Ryckaert JP, Ciccotti, G, Berendsen, H.J.C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* 1977;23(3):327-341.

# Chapter 6

## **Conformational Entropy of a Re-entrant Loop Influencing Cellulose Synthesis**

## Conformational Entropy of a Re-entrant Loop Influencing Cellulose Synthesis

Latsavongsakda Sethaphong, Joshua Amick and Yaroslava G. Yingling

911 Partners Way, Materials Science and Engineering, North Carolina State University,  
Raleigh, NC 27695

### Abstract

Molecular transport of cytosolic synthesized molecules plays an important role in biology. The means of accomplishing this require egress thru the plasma membrane. For cells that actively produce polymerized sugars like chitin and cellulose, the mechanisms involved with this transport are complex. We modeled a segment of the transmembrane region from the primary wall associated cellulose synthases of Arabidopsis; this region was shown to influence the rate of saccharification and cellulose crystallinity by a mutant cellulose synthase of Arabidopsis, *Atcesa3<sup>T942I</sup>*, resulting from a single amino acid substitution<sup>1</sup>. We compared the predicted conformations using the ROSETTA algorithm against a consensus region from a recently solved bacterial cellulose synthase structure, *BcsA*, from *rhodobacter*<sup>2</sup>. The differences in this region imply fundamentally distinct mechanisms intended to organize the glucan strand while the essential structures for the catalytic polymerization of cellulose seems conserved within prokaryotes and eukaryotes. Our results suggest that the transmembrane region in plant cellulose synthases is highly dynamic and may be a fruitful target of genetic manipulation.

### 6.1 Background

Cellulose is the most abundant biopolymer in the world and is a valuable economic resource as raw industrial material used for fuel, construction and evolving applications<sup>3,4</sup>. Despite four decades of research, the mechanisms by which  $\beta$ -1,4-cellulose is produced by

plants in their cell membrane has not been fully resolved. Steric hindrance with the rotation of the individual glucan molecules in a developing cellulose remains an intractable problem to provide experimentally testable theoretical mechanisms<sup>5</sup>. Even if the holistic view of cellulose biosynthesis remains out of reach, specific steps of the process may be more accessible for study. The two broad classification of plant cellulose synthases group those that are involved with primary wall synthesis versus those associated with secondary wall synthesis. The primary wall is the most pliant and formed as a plant cell is in its growth phase. The secondary wall is formed after the plant cell has matured. It is in the secondary cell wall where the bulk of cellulose is deposited in plants.

Terrestrial plant cellulose synthase (CesA) genes are remarkably well conserved with those of other phyla<sup>6</sup>. This conservation of specific protein architectures suggest a common evolutionary ancestry as well as fundamental mechanisms that remain vital<sup>7</sup>. For all plant cellulose synthase proteins, the conserved structures are one zinc binding domain at the N terminus, two hypervariable regions, one cytoplasmic catalytic domain bridging between transmembrane helices (TMH) 2 and 3; between these helices is the structural motif D,D,D,QXXRW which is indicative of processive glycosyltransferases (**Figure 6.1a**). Recent mutational evidence suggests a role for the transmembrane region inclusive of TMH 5 and 6 with organizing the nascent cellulose fibril of plants<sup>1,8</sup>. The linker between TMH 5 and 6 is one of the longest and might possibly guide the emergent fibril or perhaps be important in pore formation. The hydrophobic segment is well conserved but the acidic side is less so where aspartic acid may be substituted for glutamic acid. A role for the lysine residue conserved in between these segments has not been explored; although it would seem

likely that may be interact with cellulobiose given its central location in the loop. Isoxaben resistance was originally used to identify mutants occurring in this transmembrane region in primary wall associated cellulose synthases<sup>9,10</sup>. For the formation of plant cellulose fibrils, three CesA classes must form a hexameric rosette structure<sup>5</sup>. In this work, we consider the primary wall associated CesAs; specifically, we examine the region in TMH 5 and 6 of CesA1, CesA3, and CesA6. Often, re-entrant loops within the transmembrane region of plasma membrane bound proteins are associated with transport mechanisms or pore formation<sup>11</sup>. It is likely that the structure formed by the linker and these two transmembrane helices serves either one or both of these purposes.

Although abundant in the genomes of organisms, membrane bound proteins represent a minority of solved structures in the protein databank<sup>12</sup>. With the advances in computational prediction methods such as ROSETTA used in this work, investigations into the transmembrane region are greatly facilitated<sup>13,14</sup>. This is largely due to the conformational restrictions within the membrane environment, helical arrangement and general packing order resulting in a lower diversity of structures. Since the effect of a single amino acid change on the overall structure should be readily evaluated as in the case of *Atcesa3*<sup>T942I</sup> where there was a marked reduction in overall cellulose production and crystallinity. Comparing the predicted structures from wild type CesA1, CesA3, and CesA6 might give unique insights on what the role of TMH 5 and 6's unusual linker. We expand our study to a recently solved structure of a bacterial cellulose synthase from rhodobacter (PDB ID 4HG6) by taking an amino acid sequence that corresponds to the region in plants CesAs. The ROSETTA fragment database used for the prediction of the BcsA fragment is prior to the

advent of the BcsA structure. Therefore, it serves as a means test on the accuracy and suitability of computationally predicted structures to derive insights in absence of experimentally solved structures.

## 6.2 Methods

Psipred from the David Jones Lab was used to generate the secondary structure prediction file<sup>15</sup>. The OCTOPUS membrane topology prediction server was used to generate inputs for the ROSETTA membrane prediction algorithm<sup>16</sup>. The membrane protein folding algorithm of ROSETTA was used to predict that tertiary structures of the fragments<sup>13</sup>. ROSETTA was used to generate a minimum of 30,000 decoys for each sequence and up to 30,000 decoys for those proving difficult to fold. Selection of decoys was performed using the maximum entropy based clustering on the top 10 percent of the generated structures with the Durandal algorithm<sup>17</sup>. However only 10,000 was generated for Atcesa3<sup>T942I</sup> after sizeable clusters were detected with Durandal and it represented a sufficient production run. Ab initio modeling is often used when no known comparative homologs exist for the protein of interest. The aim is then to be able to take the top 10 percent best scoring structures such that the entropy based exact clustering with Durandal<sup>17</sup> would result in a “good” cluster size of at least 10 structures within a cut off value of at least 8 Å. The exact clustering algorithm of Durandal smooths the free energy function in ROSETTA by searching for the structure with the maximum number of neighbors given a cutoff value; subsequently, this cluster is removed from the remaining pool of structures. This process is repeated until no set of structures remains in the pool of structures. The largest cluster from this process is the one

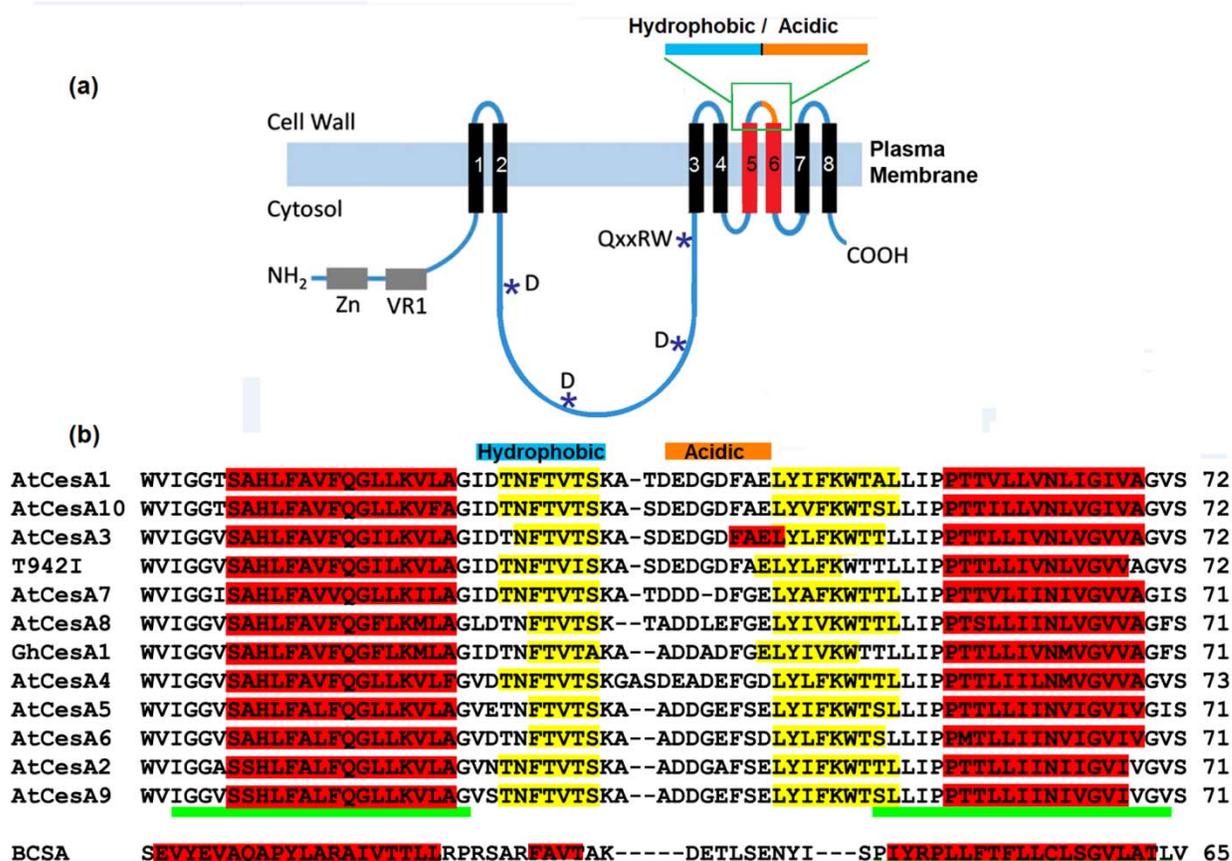
most likely to contain the nearest native like structure. The centroid that represents the average inter-similarity of the cluster is taken as the structure of interest. When a folding funnel is present, the model with the most 'native' like features is often the lowest energy structure. The optimal structure is then relaxed using the ROSETTA abinitio relax method under a weak constraint. Our installation of the ROSETTA fragment library database was made using the 2010 NCBI non redundant protein database. Automated scripts generated the selection of the best decoys and initial analysis of the folding funnels. Tecplot was used to generate the folding graphs. PyMOL was used to capture and render the protein structures<sup>18</sup>.

### 6.3 Results

The bacterial cellulose synthase of rhodobacter seemingly possesses a conserved amino acid sequence with that of the primary wall CesAs of Arabidopsis(**Figure 6.1b**). However, it would appear to have a distinct topology based on prediction from the OCTOPUS server. Since the membrane environment represents an additional constraint on the conformation, the resulting structures of transmembrane helices arguably present less conformational entropy. The folding statistics from ROSETTA show a fair consensus of structures sampled with RMSD's < 5 Å (**Table 6.1**). The looser RMSD cutoff for Atcesa3<sup>T942I</sup> is possibly due to a lower amount of sampling but its centroid scores better than that of CesA6.

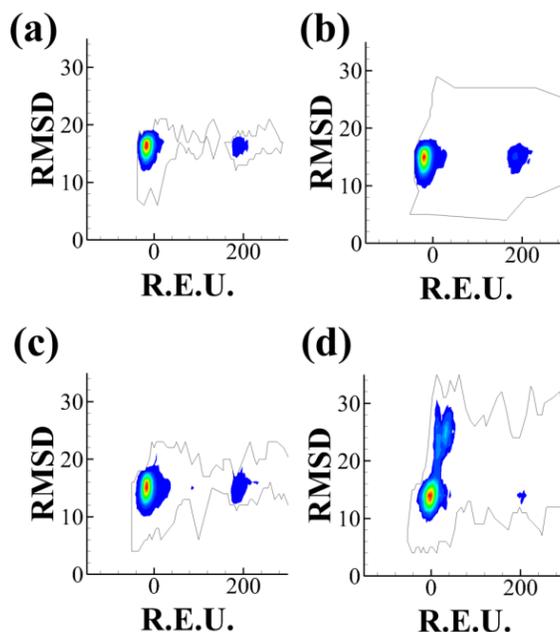
**Table 6.1:** Folding Statistics. Clusters were evaluated on the top 10% best scoring subpopulations with the maximum entropy approach of the Durandal algorithm.

Protein	Decoys	Best R.E.U.	Largest Cluster	RMSD cut	Centroid R.E.U
AtCesA1	30000	-68.661	200	3.332 Å	-37.688
AtCesA3	29261	-59.880	183	3.125 Å	-38.993
AtCesA6	30000	-67.640	179	4.000 Å	-27.642
AtCesA3 <sup>T942I</sup>	10000	-59.928	52	4.798 Å	-33.496
BCSA	21863	-60.236	93	3.524 Å	-42.791



**Figure 6.1:** General topology of plant cellulose synthases (a) adapted from Harris et al. 2012<sup>1</sup>. Eight membrane spanning helices are enumerated as shown. Transmembrane 5 and 6 are connected by a very long outer membrane loop with a unique amphiphatic structure. Psipred alignment and secondary structure prediction(b). Underlined in green are the regions the OCTOPUS topology server predicted to reside within the plasma membrane. Helices are in red with beta strands in yellow. The length of each segment is annotated to the right.

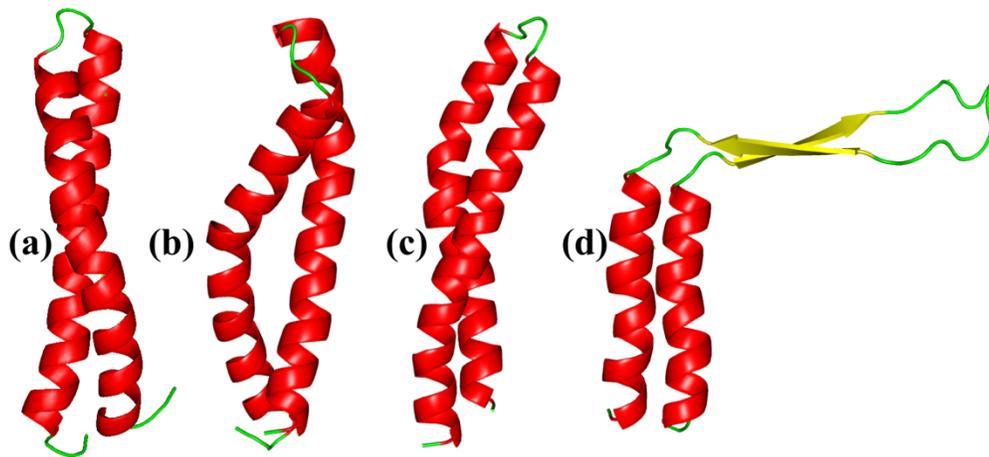
### 6.3.1 Folding distinctions between wild-type and mutant TMH56



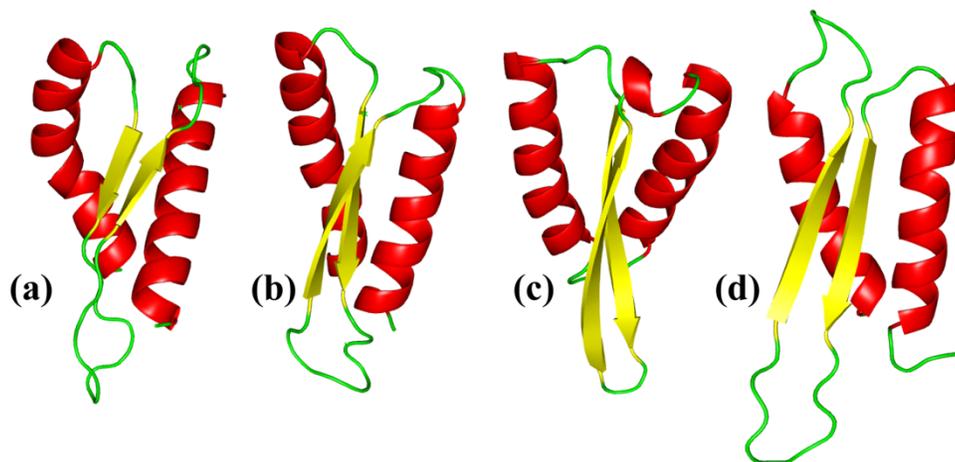
**Figure 6.2:** Folding Funnels for TMH5 & 6: Atcesa1 (a); Atcesa3 (b); Atcesa6 (c); Atcesa3<sup>T942I</sup> (d). Highest densities are in red. The biphasic folding patterns of the wild type sequences (a-c) contrast greatly with the multiple subpopulations seen in T942I (d).

The folding pathways of the wild type sequences show remarkable similarity and a biphasic sampling pool (**Figure 6.2**). The largest clustering of decoys occurs below zero Rosetta Energy Units (R.E.U) which corresponds to about 0.5 kcal/mol<sup>19</sup>. Overall, the very long linker loop between the two transmembrane helices displays two very stable states (**Figure 6.3** vs **Figure 6.4**). The dominant form of the wild type seems to be an extended helical structure that would extend from the membrane surface (**Figure 6.3a-c**). The mutant Atcesa3<sup>T942I</sup> segment displays a conformation that would essentially rest on the surface of the plasma membrane. As a transport mechanism, it would be within the realm of possibility that it serves to rearrange the plasma membrane. The presence of a beta sheet in the most

stable conformation for all the structures is unusual for transmembrane helix linkers but has been observed before in membrane vesicle fusing proteins; however, these domains were suggested to facilitate oligomerization of proteins and was concentration dependent<sup>20</sup>. The existence of a beta-sheet structure (**Figure 6.3a-c**) would seem implausible since the helix is the most stable structure within the hydrophobic interior of the lipid bilayer<sup>21</sup>. A majority of reentrant loops are indeed short  $\alpha$ -helices with some reorienting their transmembrane helices in order to achieve a stable topology<sup>22,23</sup>. The effect of a single amino acid change in the topology of transmembrane helices had previously been seen observed to drastically rearrange the insertion of multispinning proteins<sup>24</sup>. The T942I mutation is essentially substituting a polar residue for one that is bulkier and more hydrophobic. The loss of threonine for isoleucine is also seen as a helix destabilizing influence<sup>25</sup>.

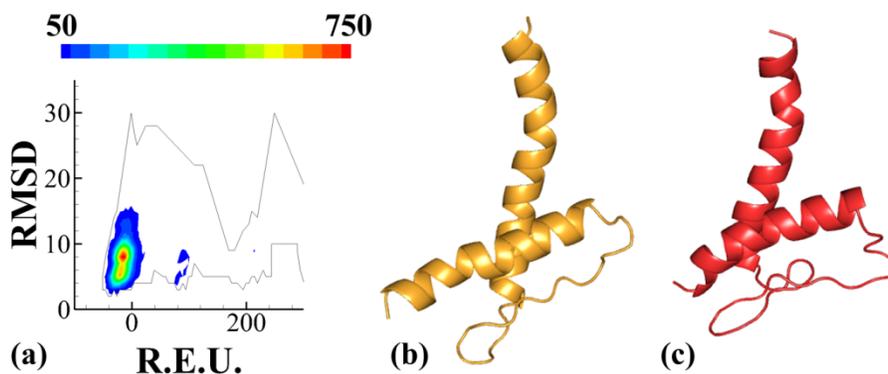


**Figure 6.3:** Representative of best centroid decoys for the transmembrane helices 5 and 6 from all the CESAs and the mutant T942I. Atcesa1 (a), Atcesa3 (b), Atcesa6 (c), Atcesa3<sup>T942I</sup> (d).



**Figure 6.4:** Best scoring decoys for the transmembrane helices 5 and 6 from the primary wall associated CESAs and the mutant T942I. Atcesa1 (a), Atcesa3 (b), Atcesa6 (c), Atcesa3<sup>T942I</sup> (d).

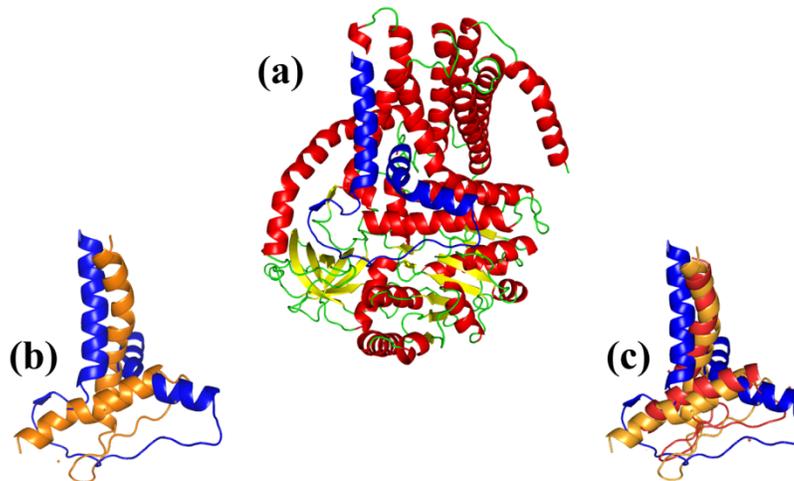
### 6.3.2 Predicted BcsA vs true Structure



**Figure 6.5:** Prediction of a membrane associated helix pair in BCSA corresponding to helices 5 and 6 of CESAs in Arabidopsis. The folding funnel plot does not indicate a biphasic folding path with over 21K decoys generated (a). The difference between the centroid of the largest cluster (b) and the best scoring decoy (c) are very slight.

The predicted structure of BcsA shows one alpha helix inserted into the cell membrane with a long N-terminus of an unstructured coil. This topology is seen in the actual

crystal structure. Folding fragments can be fraught with many sources of error; the greatest source of error would be the loss of local influences on structure from neighboring helices. The best scoring decoy and the centroid of the largest cluster from the ROSETTA algorithm show very little difference in global conformation (**Figure 6.5**). The folding pathway does not display the biphasic nature (**Figure 6.5a**) that was evident with the Arabidopsis primary CESA TMH fragments. The N-terminus region of the BcsA fragment is folded as a periplasmic helix with a short random coil connecting it to a C-terminal transmembrane helix structure. Surprisingly, the periplasmic helix shows little difference between the best decoy and that of the centroid. In the Arabidopsis TMH fragments, the conformational entropy was distinct and may have functional implications. Ultimately, this result further scores the differences between plant cellulose synthesis versus that of bacterial (prokaryotic) cellulose synthesis and organization. BcsA actually is part of a dimeric superstructure with BcsB guiding the glucan chain across the periplasmic area<sup>2</sup>.



**Figure 6.6:** Comparing the Rosetta predicted structure with that of the actual crystal structure corresponding to TMH56 of plant CESA (blue). The recently solved structure of BcsA (PDB ID: 4HG6) (a). RMSD fit to the crystal structure: 9.899 Å for the best cluster centroid (orange) with 59 residues aligned (b); 8.734 Å for the lowest energy structure (red) with 58 aligned residues (c). Since the lowest energy structure also bears close resemblance to the centroid of the largest cluster such that the fittings are almost the same.

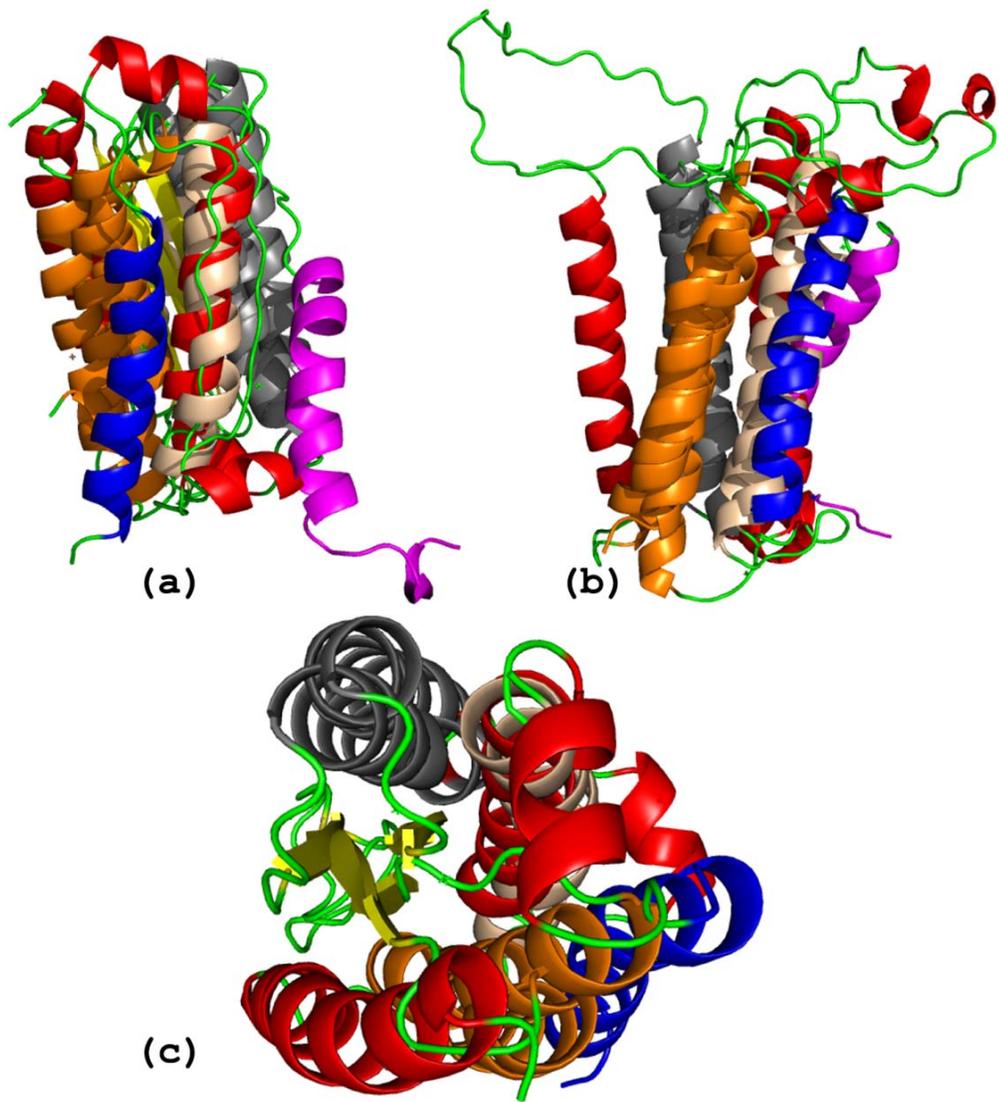
Unlike the *Arabidopsis* CESA structures, BcsA does have a recently solved crystal structure (PDB ID: 4HG6). In **Figure 6.6a**, the blue highlighted region is the selection of amino acids corresponding the analogous TMH 5 and 6 section of the plant CESAs. It consists of a transmembrane C-terminal helix with a periplasmic N-terminus helix structure with a linker consisting of a random coil. Aligning this solved structure to the best centroid (**Figure 6.6b**), the fit is skewed by the poor fit of the C-terminal helix orientation. Likewise, the same results for aligning the best scoring decoy (**Figure 6.6c**). Yet, the overall orientation and general structure is in close agreement between the predicted and experimentally determined fragment. This result bodes well for the utility of the CESA

fragments since the restriction of two helix domains to the the lipid bilayer drastically limits the conformational space.

### **6.3.3 Predicted Transmembrane Assembly of a Cellulose Synthase**

Homology modeling typically outperforms most methods of prediction, even for membrane proteins<sup>26</sup>. However, the existence of solved protein structures of similar sequences would be necessary to obtain the template models. In the absence of these data, one must still revert back to ab initio methods. Despite the limitations inherent with an implicit membrane model, useful insights may still be garnered with a mosaic method of approaching the problem of modeling more than two transmembrane helices. Complex systems may easily be broken down to their constituent parts. Unlike bacterial cellulose which manufacture single strands of glucan chains that are highly crystalline and pure, the plant must have an architecture that manufactures a product with greater control. Bacterial cellulose synthases lack both the plant conserved regions and the class specific regions. Consequently, we should expect that the architecture of the transmembrane domain also be different. There are ten transmembrane helices in the BcsA and BcsB complex which contrasts with the eight associated with plant cellulose synthase.

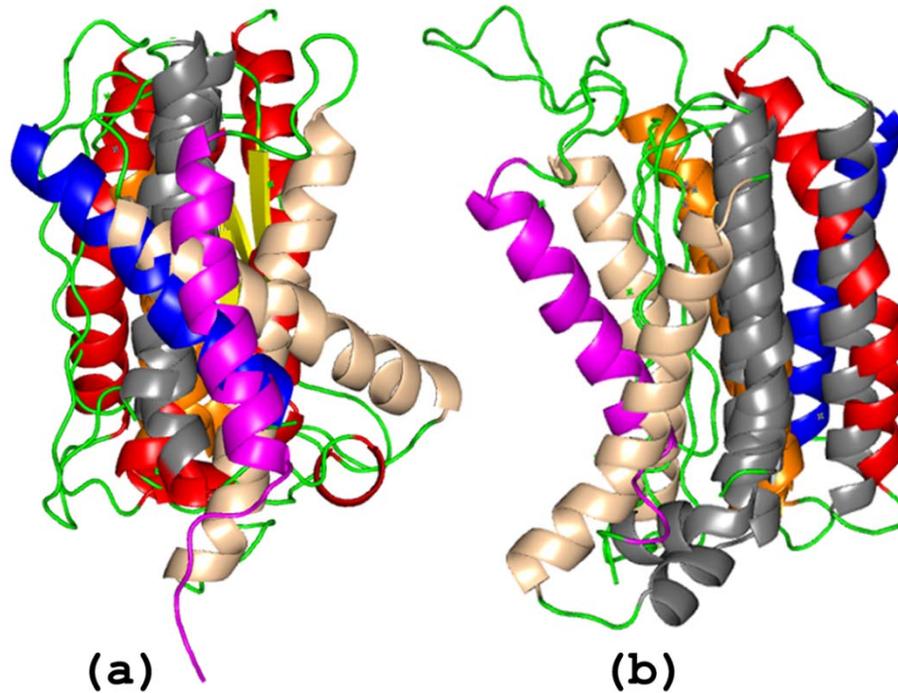
**Figure 6.7:** Membrane organization of helices 3-8 of Ghcesa1 from composite predictive runs of four helix bundles aligned on TMH 5 and 6. TMH3 (magenta); TMH4 (wheat); TMH5(grey); TMH6(orange); TMH7 (red); TMH8 (blue). Inner cytosolic region is down. The linker between TMH 5 and 6 display bistability between the best scoring decoys (a) and the best centroid (b) alignments. The beta sheet from of the linker is in yellow (a) and assumes a conformation of random coils and small helices in cyan (b). A semblance of a pore is visible in (c) where the cavity is taken up by the unusual beta sheet formation (yellow) of the linker between TMH 5 and 6.



An attempt to fold four membrane segments at a time that essentially scans the conformational space of transmembrane three thru eight of Ghcesa1 resulted in the ensemble of structures seen in Figure 6.7a-c. In Figure 6.7a, the alignments using TMH 5 and 6 results in a tight grouping of the structures evocative of BcsA's helix packing structure (Figure 6.6a). Although these results represent a short production run of 15K structures, the implicit membrane limits the conformational search space such that useful information about helix packing is extractable. Similar to the packing TMH 5 and 6 alone from the Arabidopsis primary CESAs, Ghcesa1's linker loop shows a bistable conformation of an "out of pore" configuration (Figure 6.7b) that is more disorganized than what was shown in Figure 6.3a-c. Overall, the folding of four transmembrane helices versus only two resulted in similar dynamics of an "in pore" versus an "out of pore" configuration (Figure 6.7a-b).

Given the length of the linker between TMH 5 and 6, it should be able to form a space between the transmembrane helices sufficiently large enough for the emergence of a glucan chain. When the essential helices are aligned, a semicircle of a putative channel becomes evident (Figure 6.7c). The unusual feature of this packing arrangement is that TMH 8 is pushed out with only TMH 4, 5, 6 and 7 interfacing the channel. TMH 3 was excluded since it seemed to show poor stability. It is interesting that TMH4 appears to be putatively involved in pore forming since Atcesa1<sup>A903V</sup> affects the middle of TMH4<sup>1</sup>. However, the features of that mutation will require additional exploratory work since it unclear whether that mutation changes helical packing or impedes pore formation by constricting the pore channel itself. However, the Atcesa3<sup>T942I</sup> seems to be the more dynamic mutation that has perplexed the cellulose field shows a dramatic effect on the organization of the

transmembrane structure formed by these helices<sup>9</sup>. While a two helix folding shows dramatic differences between wild type and mutant, folding four helices should reveal more.



**Figure 6.8:** Membrane organization of helices 3-8 of Ghcesa1 with the Atcesa3<sup>T942I</sup> amino acid mutation from composite predictive runs. TMH3 (magenta); TMH4 (wheat); TMH5(grey); TMH6(orange); TMH7 (red); TMH8 (blue). The alignments are rather poor for the best scoring decoys (a). The centroids of the largest clusters also do not show a real pattern (b).

Mutating the highly conserved threonine to isoleucine in the Ghcesa1 TMH region results in a more chaotic folded structure (Figure 6.8a). Alignment toward a more uniform vertical structure as would be desired proves difficult. The isoleucine residue which is more highly branched has a propensity to contort TMH 5 and 6 outward and shows poor packing within the putative channel region. The bistable conformation is retained (Figure 6.8b) but there appears to be more random coiling in the “out of pore” configuration as compared to

the presence of small helical loops as with the wild type (Figure 6.7b). Even with this limited production run, the isoleucine mutation for the conserved threonine is a destabilizing mutation that disrupts membrane helix packing as well as the coil interactions with the membrane surface.

## 6.4 Discussion

Cellulose biosynthesis is thought to retain conserved catalytic mechanisms across all phyla stemming from a common ancestry<sup>7</sup>. However, the distinction between the cellulose from plants and prokaryotes such as rhodobacter and gluconacetobacter are quite significant since the latter is often of a greater crystallinity and higher degree of polymerization<sup>27,28</sup>. Primary wall cellulose from plants typically have chains of between 500 to 3000 glucans in length<sup>29</sup>. This variability in the degree of polymerization is evidence of fine control by plants since the cell wall ultimately dictates the shape of their cells. Being able to control cellulose production and the mechanical properties of cellulose is important in organ formation. In contrast, the properties of bacterial cellulose are uniform<sup>30</sup>.

The bistable conformation of the linker between TMH 5 and 6 would suggest that it may serve as a regulator of pore formation. More recently a third mutation, *mre1/Atcesa3*<sup>G916E</sup>, maps to this region of the transmembrane domain, specifically the base of TMH 5, and further indicates that this region likely forms a pore<sup>31</sup>. The majority of re-entrant loops studied extensively have largely been confined to ion and synaptic transporters<sup>32,33,34</sup>. This is the first evidence of a membrane transport regulatory mechanism involving the transmembrane architecture of a protein outside of small molecules. Reentrant

loops have also been hypothesized to be critical for the function of hyaluronic synthases since they may either participate in forming or modulating the excretory channel<sup>35,36</sup>. The two stranded  $\beta$ -sheet structure predicted for the best scoring decoys represents a highly unusual topology and raises some suspicion. However, evidence of a mixed transmembrane topology for pore formation has been experimentally ambiguous but not entirely exclusionary<sup>37-39</sup>. Once again, cellulose biosynthesis in plants refuses to give up its secrets easily.

## 6.5 Conclusion

The reentrant loop between the transmembrane helices of number 5 and 6 of plant CESA are unusually long and can span the length of the cell membrane in a predicted structure. The single amino acid mutant in *Atcesa3*<sup>T942I</sup> results in a distinct conformational state from the wild type structure. A comparison between the predicted tertiary structure of a corresponding sequence in *BcsA* and its experimentally solved structure in *BcsA* showed good agreement with the transmembrane helix despite deviation with the periplasmic domain. The restrictive biophysical environment of the lipid bilayer enables greater accuracy in studying membrane proteins in absence of an experimentally determined structure. It is likely that the linker sequence between TMH 5 and 6 of plant cellulose synthases participates in both pore formation and regulating protein activity.

## Acknowledgements

This work was supported as part of The Center for LignoCellulose Structure and Formation, Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Science under Award Number DE-SC0001090.

## References

1. Harris DM, Corbin K, Wang T, Gutierrez R, Bertolo AL, Petti C, Smilgies DM, Estevez JM, Bonetta D, Urbanowicz BR, Ehrhardt DW, Somerville CR, Rose JKC, Hong M, DeBolt S. Cellulose microfibril crystallinity is reduced by mutating C-terminal transmembrane region residues CESA1(A903V) and CESA3(T942I) of cellulose synthase. *P Natl Acad Sci USA* 2012;109(11):4098-4103.
2. Morgan JLW, Strumillo J, Zimmer J. Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature* 2013;493(7431):181-U192.
3. Shahsavarani H, Hasegawa D, Yokota D, Sugiyama M, Kaneko Y, Boonchird C, Harashima S. Enhanced bio-ethanol production from cellulosic materials by semi-simultaneous saccharification and fermentation using high temperature resistant *Saccharomyces cerevisiae* TJ14. *J Biosci Bioeng* 2013;115(1):20-23.
4. Hassan ML, Moorefield CM, Elbatal HS, Newkome GR, Modarelli DA, Romano NC. Fluorescent cellulose nanocrystals via supramolecular assembly of terpyridine-modified cellulose nanocrystals and terpyridine-modified perylene. *Mater Sci Eng B-Adv* 2012;177(4):350-358.
5. Carpita NC. Update on Mechanisms of Plant Cell Wall Biosynthesis: How Plants Make Cellulose and Other (1 -> 4)-beta-D-Glycans. *Plant Physiology* 2011;155(1):171-184.
6. Nobles DR, Romanovicz DK, Brown RM. Cellulose in cyanobacteria. Origin of vascular plant cellulose synthase? *Plant Physiology* 2001;127(2):529-542.
7. Roberts AW, Roberts EM, Delmer DP. Cellulose synthase (CesA) genes in the green alga *Mesotaenium caldariorum*. *Eukaryot Cell* 2002;1(6):847-855.
8. Zhang BC, Deng LW, Qian Q, Xiong GY, Zeng DL, Li R, Guo LB, Li JY, Zhou YH. A missense mutation in the transmembrane domain of CESA4 affects protein abundance in the plasma membrane and results in abnormal cell wall biosynthesis in rice. *Plant Mol Biol* 2009;71(4-5):509-524.

9. Scheible WR, Eshed R, Richmond T, Delmer D, Somerville C. Modifications of cellulose synthase confer resistance to isoxaben and thiazolidinone herbicides in Arabidopsis Ixr1 mutants. *P Natl Acad Sci USA* 2001;98(18):10079-10084.
10. Desprez T, Vernhettes S, Fagard M, Refregier G, Desnos T, Aletti E, Py N, Pelletier S, Hofte H. Resistance against herbicide isoxaben and cellulose deficiency caused by distinct mutations in same cellulose synthase isoform CESA6. *Plant Physiology* 2002;128(2):482-490.
11. Yan CH, Luo JR. An Analysis of Reentrant Loops. *Protein Journal* 2010;29(5):350-354.
12. Lehnert U, Xia Y, Royce TE, Goh CS, Liu Y, Senes A, Yu HY, Zhang ZL, Engelman DM, Gerstein M. Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions. *Q Rev Biophys* 2004;37(2):121-146.
13. Yarov-Yarovoy V, Schonbrun J, Baker D. Multipass Membrane Protein Structure Prediction Using Rosetta. *Proteins* 2006;62:1010-1025.
14. Vinothkumar KR, Henderson R. Structures of membrane proteins. *Q Rev Biophys* 2010;43(1):65-158.
15. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 1999;292(2):195-202.
16. Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 2008;24(15):1662-1668.
17. Berenger F, Shrestha R, Zhou Y, Simoncini D, Zhang KYJ. Durandal: Fast exact clustering of protein decoys. *J Comput Chem* 2012;33(4):471-474.
18. The PyMOL Molecular Graphics System. 1.2r3pre: Schrodinger, LLC.
19. Das R. Four Small Puzzles That Rosetta Doesn't Solve. *Plos One* 2011;6(5).
20. Yassine W, Taib N, Federman S, Milochau A, Castano S, Sbi W, Manigand C, Laguerre M, Desbat B, Oda R, Lang J. Reversible transition between alpha-helix and beta-sheet conformation of a transmembrane domain. *Bba-Biomembranes* 2009;1788(9):1722-1730.
21. Engelman DM, Steitz TA, Goldman A. Identifying Nonpolar Transbilayer Helices in Amino-Acid-Sequences of Membrane-Proteins. *Annu Rev Biophys Bio* 1986;15:321-353.

22. von Heijne G. Membrane-protein topology. *Nature Reviews, Molecular Cell Biology* 2006;7:909-918.
23. Viklund H, Granseth E, Elofsson A. Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: Application to complete genomes. *J Mol Biol* 2006;361(3):591-603.
24. Seppala S, Slusky JS, Lloris-Garcera P, Rapp M, von Heijne G. Control of Membrane Protein Topology by a Single C-Terminal Residue. *Science* 2010;328(5986):1698-1700.
25. Lyu PC, Sherman JC, Chen A, Kallenbach NR. Alpha-Helix Stabilization by Natural and Unnatural Amino-Acids with Alkyl Side-Chains. *P Natl Acad Sci USA* 1991;88(12):5317-5320.
26. Kelm S, Shi JY, Deane CM. MEDELLER: homology-based coordinate generation for membrane proteins. *Bioinformatics* 2010;26(22):2833-2840.
27. Hornung M, Ludwig M, Gerrard AM, Schmauder HP. Optimizing the production of bacterial cellulose in surface culture: Evaluation of substrate mass transfer influences on the bioreaction (Part 1). *Eng Life Sci* 2006;6(6):537-545.
28. Iguchi M, Yamanaka S, Budhiono A. Bacterial cellulose - a masterpiece of nature's arts. *J Mater Sci* 2000;35(2):261-270.
29. Blaschek W, Koehler H, Semler U, Franz G. Molecular weight distribution of cellulose in primary cell walls. *Planta* 1982;154(6):550-555.
30. Sani A, Dahman Y. Improvements in the production of bacterial synthesized biocellulose nanofibres using different culture methods. *J Chem Technol Biot* 2010;85(2):151-164.
31. Pysh L, Alexander N, Swatzyna L, Harbert R. Four alleles of AtCESA3 form an allelic series with respect to root phenotype in *Arabidopsis thaliana*. *Physiologia Plantarum* 2012;144(4):369-381.
32. Dobrowolski A, Lolkema JS. Functional Importance of GGXG Sequence Motifs in Putative Reentrant Loops of 2HCT and ESS Transport Proteins. *Biochemistry-Us* 2009;48(31):7448-7456.
33. Cao Y, Jin XS, Huang H, Derebe MG, Levin EJ, Kabaleeswaran V, Pan YP, Punta M, Love J, Weng J, Quick M, Ye S, Kloss B, Bruni R, Martinez-Hackert E, Hendrickson WA, Rost B, Javitch JA, Rajashankar KR, Jiang YX, Zhou M. Crystal structure of a potassium ion transporter, TrkH. *Nature* 2011;471(7338):336-+.

34. Tzingounis AV, Wadiche JI. Glutamate transporters: confining runaway excitation by shaping synaptic transmission. *Nat Rev Neurosci* 2007;8(12):935-947.
35. Heldermon C, DeAngelis PL, Weigel PH. Topological organization of the hyaluronan synthase from *Streptococcus pyogenes*. *J Biol Chem* 2001;276(3):2037-2046.
36. Kumari K, Weigel PH. Molecular cloning, expression, and characterization of the authentic hyaluronan synthase from group C *Streptococcus equisimilis*. *J Biol Chem* 1997;272(51):32539-32546.
37. Gornetschelnokow U, Naumann D, Weise C, Hucho F. Secondary Structure and Temperature Behavior of Acetylcholinesterase - Studies by Fourier-Transform Infrared-Spectroscopy. *Eur J Biochem* 1993;213(3):1235-1242.
38. Methot N, Ritchie BD, Blanton MP, Baenziger JE. Structure of the pore-forming transmembrane domain of a ligand-gated ion channel. *J Biol Chem* 2001;276(26):23726-23732.
39. Ortells MO, Lunt GG. A mixed helix-beta-sheet model of the transmembrane region of the nicotinic acetylcholine receptor. *Protein Eng* 1996;9(1):51-59.

# Chapter 7

## Future Directions

## **7.0 Future Directions**

Studying atomic interactions is fraught with issues of model integrity. Empirically force field models that are derived from experiments will be the key to generate simulations of a realistic enough nature to drive further research or when experimental approaches break down. More often, the case is that experimental results require a theoretical basis for understanding. For time dependent processes such as self assembly and dynamic phenomena, computational simulations become a verification tool to fit theory to observables.

### **7.1 Nucleic Acids as Nano-materials**

Investigating the natural assembly of nucleic acids in order to generate scaffolds for nanoelectronics and drug delivery remains a daunting task not only due to the prohibitive economics of generating tailored sequences of significant length, but also the lack of design tools. Biosensors development seems to be more advanced since the amount of required material is relatively small and the design tools are developed from known techniques<sup>1,2</sup>. Controlling the self assembly of engineered sequences of nucleic acids that direct nanoparticle morphology requires additional characterization although there has been some progress<sup>3</sup>. Molecular self-recognition and assembly ultimately implies that there is a structural code that can be programmed and designed in order to generate any desired material. Simulating larger scale systems while being able to limit the types of interactions in order to control assembly appears to be the key toward building scaffold architectures and nano-patterning with nucleic acids. This regimented approach toward investigating

molecular self assembly is both tedious and possibly expensive with experimental resources. As computing power grows inversely to its economic barriers, *in silico* methods should make it feasible to bridge the gap between atomistic time scales and that of the bench researcher.

## 7.2 Synthases and Synthetic Biology at the Membrane Interface

Computational determination of protein structures has progressed significantly since the manual experimental methods first used to develop a three dimensional structure of haemoglobin and myoglobin by Perutz and Kendrew<sup>4-6</sup>. Modern technology greatly assists in purifying, crystallizing, and model building, but a structure alone remains only the first step in ascertaining the functionality of novel protein structures. With computational simulations, we can now develop virtual experiments to develop theoretical ligands and study the molecular effects with *in silico* mutagenesis<sup>7-9</sup>. But these methods have largely focused on the cytosolic domains of proteins. The most complex phenomena occur at interfacial surfaces such as that between the cytosol and the plasma membrane.

Currently, the problem of solving the full tertiary structure of membrane proteins remains challenging. The cost of obtaining a crystal structure remains prohibitively high<sup>10,11</sup>. For plant cellulose synthases, the most significant obstacle is the inability to obtain enough protein in order to perform even conventional structure determination<sup>12</sup>. Computationally predicting the structure of soluble proteins has become much easier with new bioinformatics software<sup>13,14</sup>. However, prediction of the packing of transmembrane domain greater than two remains challenging due to a lack of suitable force-fields and knowledge about how proteins pack within the lipid bilayer<sup>15,16</sup>. There is a general lack of understanding about the roles of

pores, regulatory loops, and structural dynamics of proteins at the interface of the plasma membrane and cytosol. The biophysics of plasma membranes is also not fully understood<sup>17</sup>.

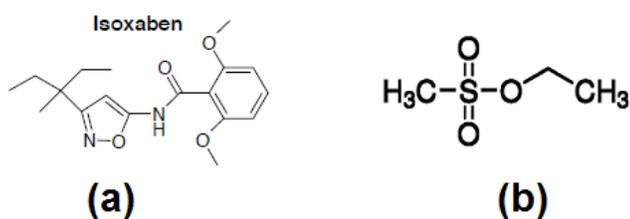
The final assembly of cellulose synthases in the plasma membrane remain unresolved. It is unclear why the reentrant loop of transmembrane helices 5 and 6 are unusually long. The organization of the transmembrane membrane region of cellulose synthase might offer a unique perspective on a common tasks in biology: how to export molecules in order to build structure, send messages, sense the microenvironment, and exert atomic control outside of the molecular machinery readily available within the cytosol.

### 7.3 Ascertaining mechanism of Isoxaben resistance

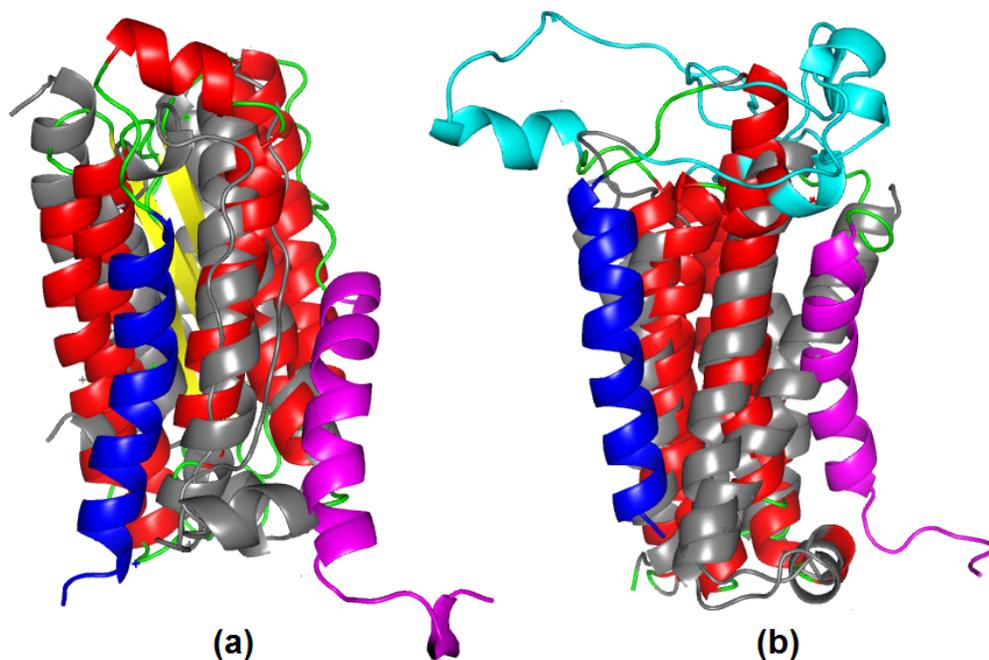
Isoxaben resistance in Arabidopsis is highly peculiar in that all the single amino acid mutations known to confer it occur in various positions specific to the transmembrane helices. To date, four are known: (aegeus) Atcesa1<sup>A903V</sup>, Atcesa3<sup>G998D</sup>, Atcesa3<sup>T942I</sup>, and Atcesa6<sup>R1064W</sup>

<sup>18-20</sup>. A fifth mutant in this region was discovered by screening under ethylmethanesulfonate, (mre1)Atcesa3<sup>G916E</sup>, but was seen to affect root cell shape and have a high correlation with overall cellulose content<sup>21</sup>. How two seeming different chemicals may affect the outcome of cellulose activity is perplexing (Figure 7.1). Isoxaben resistance has largely been confined to transmembrane helices 4, 5, and 8. A unifying theory would be that these mutations affect the pore of the cellulose synthases in question. Although the preliminary results of Chapter 6 are promising, a more detailed study needs to be performed.

Full scale *ab-initio* folding of all 8 transmembrane helices will be required to generate a geometry profile of how these helical structures pack and form a pore (Figure 7.2). The current protocol as implemented in ROSETTA faces limitations beyond four helices despite the advantages of conformational restrictions. The current force field assumes an implicit bilayer that stands in the place of the lipid membrane consisting of positional penalties and a continuum dielectric<sup>16,22</sup>.



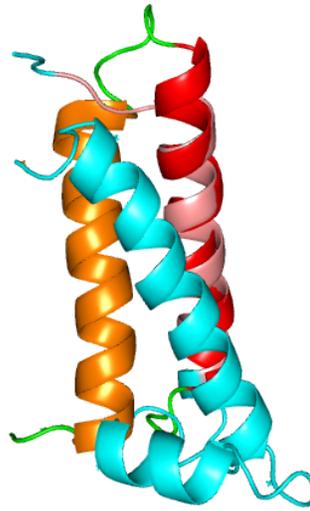
**Figure 7.1:** Isoxaben (a) targets CESA3 and CESA6, and ethyl-methanesulfonate (b)



**Figure 7.2:** Membrane organization of helices 3-8 of Ghcesa1 from composite predictive runs of four helix bundles aligned on TMH 5 and 6. TMH 3 (magenta) and TMH 8 (blue). Inner cytosolic region is down. The linker between TMH 5 and 6 display bistability between the best scoring decoys (a) and the best centroid (b) alignments. The beta sheet from of the linker is in yellow (a) and assumes a conformation of random coils and small helices in cyan (b).

The next stage of development would be to generate refined models of four helix packing which has a reasonable plausibility of being close to actuality. Thereafter, assembling the larger structure will require minimizing it using more advanced forcefields<sup>23</sup>. This region of the cellulose synthase complex offers a more viable target of affecting synthesis since it is putatively coupled to the polymerization process<sup>19</sup>. Molecular dynamics simulations of the region in an explicit lipid membrane might offer a means to query how isoxaben interacts with the transmembrane region or how the helices alter the membrane and indicate whether membrane dynamics have any effect on cellulose crystallization and microfibrillar assembly.

#### 7.4 Role for distal N-terminus membrane helices



**Figure 7.3:** The N-terminal helices (TMH 1 and 2) with the first C-terminal transmembrane helix (TMH 3) from Ghcesa1: TMH1 (orange); TMH 2 (red and pink); TMH 3 with artificial linker (cyan).

For cellulose synthases, the N-terminal helices (TMH 1 and 2) and TMH3 that emerges after the major cytosolic domain are the least studied. Additional modeling will be required for to understand how these N-terminal helices may or may not contribute to forming the pore channel. The lack of viable mutant with phenotype in these helices might indicate that they are crucial to the proper function and formation of the pore channel. Unfortunately, it is difficult to study these structures in isolation since spanning the distance between TMH2 and TMH3 is the very large catalytic domain lying in the cytosol.

## Acknowledgements

This work was supported as part of The Center for LignoCellulose Structure and Formation, Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Science under Award Number DE-SC0001090.

## References

1. Zhang LB, Guo SJ, Dong SJ, Wang EK. Pd Nanowires as New Biosensing Materials for Magnified Fluorescent Detection of Nucleic Add. *Anal Chem* 2012;84(8):3568-3573.
2. Ermini ML, Scarano S, Bini R, Banchelli M, Berti D, Mascini M, Minunni M. A rational approach in probe design for nucleic acid-based biosensing. *Biosens Bioelectron* 2011;26(12):4785-4790.
3. Wang ZD, Tang LH, Tan LH, Li JH, Lu Y. Discovery of the DNA "Genetic Code" for Abiological Gold Nanoparticle Morphologies. *Angew Chem Int Edit* 2012;51(36):9078-9082.
4. Bodo G, Dintzis HM, Kendrew JC, Wyckoff HW. The Crystal Structure of Myoglobin .5. A Low-Resolution 3-Dimensional Fourier Synthesis of Sperm-Whale Myoglobin Crystals. *Proc R Soc Lon Ser-A* 1959;253(1272):70-&.
5. Strandberg B, Dickerson RE, Rossmann MG. 50 Years of Protein Structure Analysis. *J Mol Biol* 2009;392(1):2-32.
6. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT. Structure of Haemoglobin - 3-Dimensional Fourier Synthesis at 5.5-Å Resolution, Obtained by X-Ray Analysis. *Nature* 1960;185(4711):416-422.
7. Laine E, de Beauchene IC, Perahia D, Auclair C, Tchertanov L. Mutation D816V Alters the Internal Structure and Dynamics of c-KIT Receptor Cytoplasmic Region: Implications for Dimerization and Activation Mechanisms. *Plos Comput Biol* 2011;7(6).
8. Marrone TJ, Briggs JM, McCammon JA. Structure-based drug design: Computational advances. *Annu Rev Pharmacol* 1997;37:71-90.

9. Dixit A, Yi L, Gowthaman R, Torkamani A, Schork NJ, Verkhivker GM. Sequence and Structure Signatures of Cancer Mutation Hotspots in Protein Kinases. *Plos One* 2009;4(10).
10. Stevens RC. Long live structural biology. *Nat Struct Mol Biol* 2004;11(4):293-295.
11. Terwilliger TC, Stuart D, Yokoyama S. Lessons from Structural Genomics. *Annu Rev Biophys* 2009;38:371-383.
12. Somerville C. Cellulose synthesis in higher plants. *Annu Rev Cell Dev Bi* 2006;22:53-78.
13. Zhang Y. I-TASSER server for protein 3D structure prediction. *Bmc Bioinformatics* 2008;9.
14. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol* 2011;487.
15. Nugent T, Jones DT. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *P Natl Acad Sci USA* 2012;109(24):E1540-E1547.
16. Yarov-Yarovoy V, Schonbrun J, Baker D. Multipass Membrane Protein Structure Prediction Using Rosetta. *Proteins* 2006;62:1010-1025.
17. Phillips R, Ursell T, Wiggins P, Sens P. Emerging roles for lipids in shaping membrane-protein function. *Nature* 2009;459(7245):379-385.
18. Desprez T, Vernhettes S, Fagard M, Refregier G, Desnos T, Aletti E, Py N, Pelletier S, Hofte H. Resistance against herbicide isoxaben and cellulose deficiency caused by distinct mutations in same cellulose synthase isoform CESA6. *Plant Physiol* 2002;128(2):482-490.
19. Harris DM, Corbin K, Wang T, Gutierrez R, Bertolo AL, Petti C, Smilgies DM, Estevez JM, Bonetta D, Urbanowicz BR, Ehrhardt DW, Somerville CR, Rose JKC, Hong M, DeBolt S. Cellulose microfibril crystallinity is reduced by mutating C-terminal transmembrane region residues CESA1(A903V) and CESA3(T942I) of cellulose synthase. *P Natl Acad Sci USA* 2012;109(11):4098-4103.

20. Scheible WR, Eshed R, Richmond T, Delmer D, Somerville C. Modifications of cellulose synthase confer resistance to isoxaben and thiazolidinone herbicides in *Arabidopsis Ixr1* mutants. *P Natl Acad Sci USA* 2001;98(18):10079-10084.
21. Pysh L, Alexander N, Swatzyna L, Harbert R. Four alleles of *AtCESA3* form an allelic series with respect to root phenotype in *Arabidopsis thaliana*. *Physiol Plantarum* 2012;144(4):369-381.
22. Senes A. Computational design of membrane proteins. *Curr Opin Struc Biol* 2011;21(4):460-466.
23. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic Validation of Protein Force Fields against Experimental Data. *Plos One* 2012;7(2).

## APPENDICES

# Appendix A

Supporting Information

For

## Chapter 4: Tertiary model of a plant cellulose synthase

Latsavongsakda Sethaphong<sup>1</sup>, James D. Kubicki<sup>2</sup>, Jochen Zimmer<sup>6</sup>, Dario Bonetta<sup>4</sup>, Seth DeBolt<sup>5</sup>, Candace H. Haigler<sup>3</sup>, Yaroslava G. Yingling<sup>1\*</sup>

<sup>1</sup>911 Partners Way, Materials Science and Engineering, North Carolina State University, Raleigh, NC 27695

<sup>2</sup> Dept. of Geosciences and the Earth & Environmental Systems Institute, The Pennsylvania State University, University Park PA 16802

<sup>3</sup> Dept. of Crop Science and Dept. of Plant Biology, Campus Box 7620, North Carolina State University, Raleigh, NC 27695

<sup>4</sup> Faculty of Science, University of Ontario Institute of Technology, Oshawa, ON Canada

<sup>5</sup> Department of Horticulture, University of Kentucky, Lexington, KY 40546

<sup>6</sup> Center for Membrane Biology, Department of Molecular Physiology and Biological Physics, University of Virginia

## Materials and Methods

### Simulations and Modeling

Approaches to computational structure prediction fall under the spectrum of knowledge-based algorithms spanning template-based modeling to use of physical force-fields (or de novo modeling) when no highly similar structures are available. Knowledge based three dimensional structure prediction from a linear protein sequence was accomplished with the SAM-T08 prediction server of the Karplus Laboratory [1]. A FASTA file of the putative cytosolic domain amino acid sequence was submitted to the prediction server: [http://compbio.soe.ucsc.edu/SAM\\_T08/T08-query.html](http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html). This method has exhibited good performances across diverse proteins, and high quality structures result when there is a good match between the target and available templates [1, 2]. Two of the top selected structures were from the bacterial protein templates of SpsA and K4CP that have been extensively used to examine the molecular basis for catalysis and substrate recognition of glycosyltransferases [3-5]. SpsA is a glycosyltransferase involved in producing the *B. subtilis* spore coat that co-crystallized with Mg<sup>++</sup>- or Mn<sup>++</sup>-UDP. K4CP catalyzes alternative transfers of glucuronic acid and N-acetylgalactosamine to form chondroitin (glycosaminoglycan) in *E. coli* [3].

Since the resulting homology model, **Fig. S2A**, is fragmentary in form, it was initially manually refined with DS Visualizer from Accelrys to correct for steric clashes and breakages. An Amber molecular dynamics package with the force field FF99SB and TIP3P water model was used for relaxing this structure [6, 7]. Atom types were converted into

Amber acceptable format via an in-house script prior to equilibration and subsequent MD production run.

All structures were subjected to conjugate gradient energy minimization for 5000 steps. Minimized protein structures were then neutralized with Na<sup>+</sup> ions and immersed in a water box with at least 10 Å deep solvation shell using the TIP3P water model [7]. Additional Na<sup>+</sup> and Cl<sup>-</sup> ions were added to represent a 0.3 M effective salt concentration. The equilibration of each system was carried out in 11 stages starting from the solvent minimization for 10000 steps and keeping the protein restrained for 200 Kcal/mol. The system was heated to 300K in 100 ps while imposing a 200kcal/mol constraint on the structure. A brief NPT MD run was performed for 40 ps with the protein restraint maintained at 200 kcal/mol. Another constrained minimization step follows with the restraint of 25 kcal/mol for 10000 steps. A second NPT MD run was performed at 25 kcal/mol restraint for 20 ps. Subsequently four additional 1000 cycle minimization steps were performed while relaxing the positional constraint from 20 kcal/mol to 5 kcal/mol in 5 kcal/mol increments. A final unconstrained minimization stage of 1000 cycles was performed before reheating the system to 300K at constant volume within 40 ps. Subsequently, NPT equilibrations were performed to ensure uniformity in solvent density. Long range electrostatic interactions were calculated by Particle Mesh Ewald summation (PME)[8] and the non-bonded interactions were truncated at 9 Å cutoff along with a 0.00001 tolerance of Ewald convergence. A Berendsen thermostat maintained temperature at 300 K [9]. The SHAKE algorithm was used to constrain the position of hydrogen atoms [10]. The production simulations were performed

for an NVT ensemble. Each production simulation was performed for 10 ns with a 2 fs time step.

Intermediate structures were evaluated for quality; gross mis-fold errors were unfolded using a protocol starting directed MD with a harmonic force followed by free Langevin self-guided dynamics. Several series of such MD simulations were performed (for more than 150ns simulations time) until a reasonable z-score was reached. The final structure from the MD simulations was energy minimized for 10,000 cycles with a convergence criterion less than a drms of  $1.0E-4$  kcal/mole Å.

Initial evaluation of the final predicted structure of the native GhCESA1 cytosolic region was performed using Pro-SA (<https://prosa.services.came.sbg.ac.at/prosa.php>) [11]. Two characteristics of the structure were derived: the z-score and a graphic of the residue energies. The z-score measures the deviation of the total energy from an energy distribution of random conformations, and an acceptable z-score of the computed structure must fall within the distribution of those derived from experimentally determined structures. High energy residues contribute to poor z-scores are likely areas that need further refinement or may have intrinsically high conformational entropy. The stereochemical quality of the intermediate and final structures was analyzed by PROCHECK (<http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/>) [12]. WhatCheck, another protein verification tool, was also used (<http://swift.cmbi.ru.nl/gv/whatcheck/>). The final structure was analyzed comprehensively using the protein structure validation software suite (PSVS; [http://psvs-1\\_4-dev.nesg.org/](http://psvs-1_4-dev.nesg.org/)), which integrates the analyses performed by PROCHECK, MolProbity, Verify3D, Prosa II, and the PDB validation software [13].

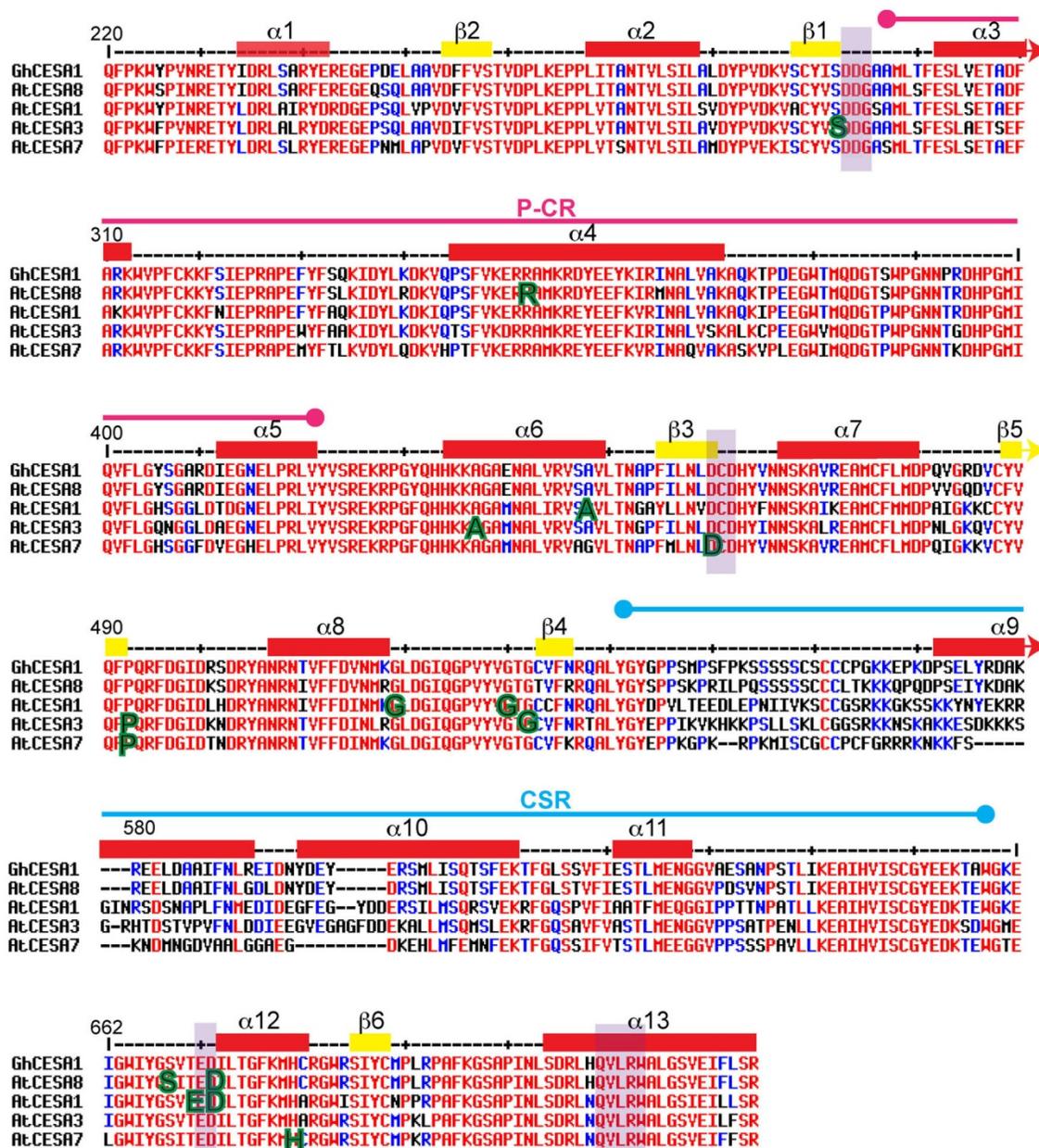
Additional validation of our protein model was performed using ERRAT[14] (<http://nihserver.mbi.ucla.edu/ERRATv2/>), which is a protein structure verification algorithm mainly used to assess crystallographic models where a 9-residue sliding window is used to generate the value of the error function: ERRAT2 (Quality Factor). Earlier approaches that coupled a *de novo* prediction with further refinement under molecular dynamics simulations have not shown additive improvements [15]. In this work, we achieved appreciable gains in structure quality over time (**Figure S2b**).

The symmetric docking protocol of Rosetta 3.4 was used generate homooligomeric assemblies[16]; the algorithm allows translation occurring on the plane connecting the center of mass for the monomers. A slide degree of freedom is randomly chosen and subunits are translated into contact. An optimization of the rigid body orientation proceeds with a Monte Carlo search under a low energy resolution function followed by a high resolution optimization of side-chain and rigid body conformation via Monte Carlo Minimization.

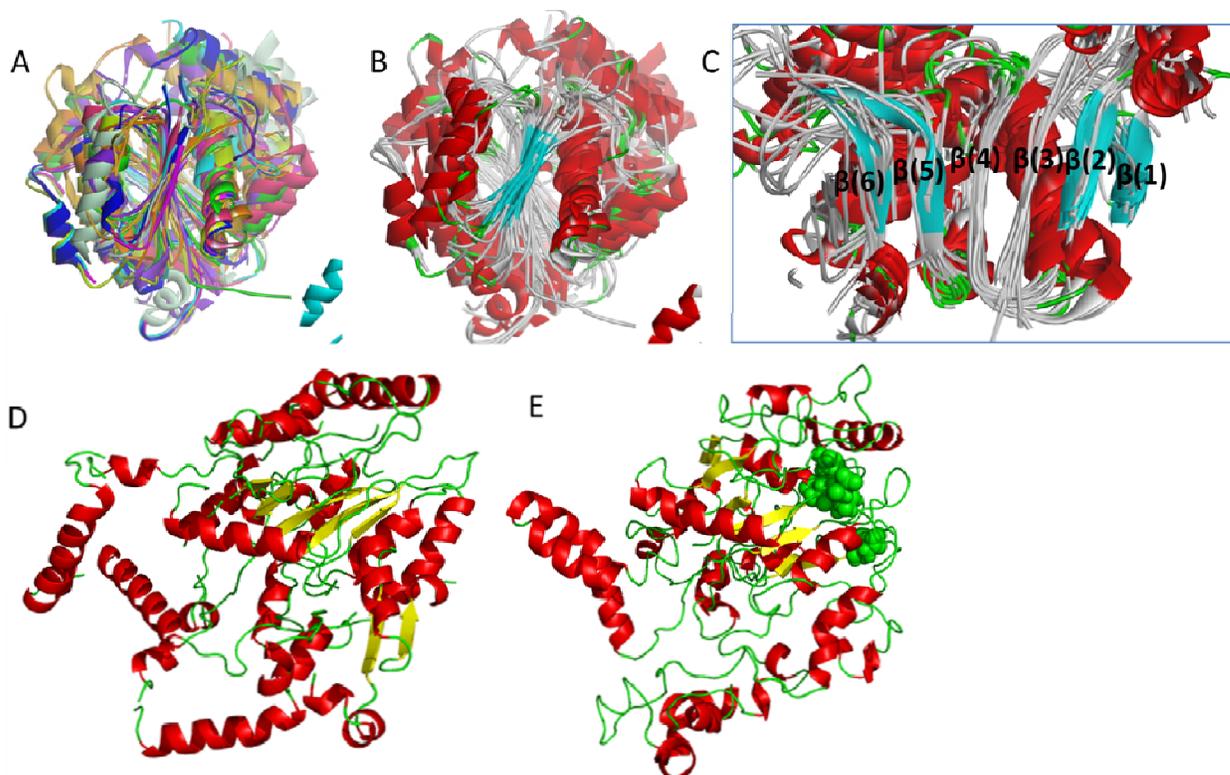
Related to docking UDP-Glc into the catalytic site, Density Functional Theory (DFT) calculations were carried out on the  $Mn^{2+}$ - and  $Mg^{2+}$ -UDP-Glc + DxD models using the B3LYP [17, 18] exchange and correlation functionals and the 6-311+G(d,p) basis set [19, 20] using Gaussian 03 program [21]. All atoms were allowed to relax without constraint or symmetry. After energy minimization, frequency analyses were performed to ensure an energy minimum had been found.

For the native predicted structure as well as three mutant structures, the flexibility of each residue was assessed using molecular dynamic simulations [22]. Each residue position was used as a variable in four simulations to generate four observations for each residue,

allowing cross correlation analysis for coupled motions to be derived from the fluctuation data. The total atomic fluctuation data was calculated using the PTRAJ tool of Amber 11 [23] and then imported into MATLAB (R2011a, MathWorks) with an in-house script to generate the correlation matrix. The input 4x506 matrix was constructed such that the rows corresponded to each simulation with columns corresponding to individual residues.

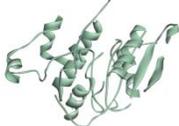
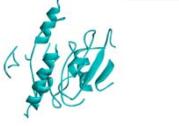
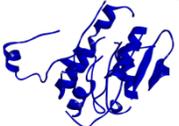
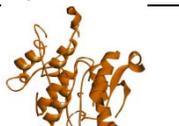
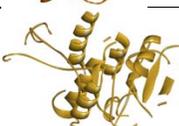


**Figure 4S.1:** Residues from GhCESA1 that were included in the Gh506 structure are aligned with the same regions of Arabidopsis CESAs with missense mutations. Numbering is relative to residue position in full length GhCESA1. Pink and blue lines indicate the positions of the P-CR and CSR plant-specific regions, respectively. Red and yellow rectangles indicate  $\alpha$ -helices and  $\beta$ -sheets, respectively. By comparison to the structure of RsBcsA (see the main text),  $\alpha$ -2,6,7,8,13 and  $\beta$ -1–6 are predicted to be in the core GT domain. Light purple vertical highlights show the position of selected conserved domains. Large green letters indicate sites of missense mutation in the AtCESA indicated.

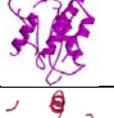
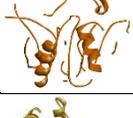


**Figure 4S.2:** (A-C) Aligned structures used in model prediction as listed in Table 4S.1. Side of the  $\beta$ -strands (A) colored by individual structure and (B) colored by secondary structure. (C) View of the slice to expose the  $\beta$ -sheet region, with individual  $\beta$ -strands numbered  $\beta$ 1– $\beta$ 6. (D) The snapshot of the starting structure of the Gh506 cytosolic domain from the SAM-T08 HMM structure prediction server. (E) The predicted structure after molecular dynamics refinement with six  $\beta$ -strands in yellow and DD, DCD, and ED in green. The  $\alpha$ -helices dispersed throughout the structure are red.

**Table 4S.1:** . The PDB identification numbers, E-values, and snapshots of structures used in predicting the structure of the  $\beta$ -sheet region of the GhCESA1 cytosolic region using Hidden Markov chain modeling. During the selection of the top models, the SAM-T08 generates pairwise alignments of the target sequence and the best-scoring templates, which are adjudicated by E-value representing how many sequences would score this well in the database. Structures with E-values of less than about  $1.0E-5$  are very likely to have a domain of the same fold as the target. Structures with E-values of larger than about 0.1 are very speculative.

<b>№</b>	<b>PDB ID</b>	<b>Description</b>	<b>E-value</b>	<b>Snapshot of part of the structure used for prediction</b>
<b>1</b>	1xhb	Crystal structure of UDP-GalNAc:polypeptide alpha-N-acetylgalactosaminyltransferase-T1	1.1661e-21	
<b>2</b>	2z86	Crystal structure of chondroitin polymerase from <i>Escherichia coli</i> strain K4 (K4CP) complexed with UDP-GlcUA and UDP	1.6825e-20	
<b>3</b>	2ffu	Dynamic association between the catalytic and lectin domains of human UDP-GalNAc:polypeptide alpha-N-acetylgalactosaminyltransferase-2	4.7760e-20	
<b>4</b>	3ckj	Essential GT (MAP2569c) from <i>Mycobacterium avium</i> subsp. paratuberculosis	1.6086e-18	
<b>5</b>	3bcv	Putative glycosyltransferase from <i>Bacteroides fragilis</i>	4.9831e-18	
<b>6</b>	1qg8	SpsA from <i>Bacillus subtilis</i>	6.8435e-18	
<b>7</b>	2bo4	Mannosylglycerate Synthase	1.5988e-17	
<b>8</b>	1omz	Alpha 1,4-N-acetylhexosaminyltransferase (EXTL2)	2.5389e-16	

**Table 4S.1: Continued**

9	1fo8	rabbit N-acetylglucosaminyltransferase I	5.2404e-14	
10	2nvx	RNA polymerase II (pol II)	7.9916e-14	
11	2zu9	mannosyl-3-phosphoglycerate synthase from <i>Pyrococcus horikoshii</i>	1.5238e-12	
12	1yro	bovine beta-1,4-galactosyltransferase I	3.6727e-08	
13	2fy7	beta-1,4-galactosyltransferase-I	2.7044e-06	
14	1i52	4-diphosphocytidyl-2-C-methylerythritol synthetase	2.4158e-01	
15	1fgx	bovine beta-4-galactosyltransferase catalytic domain	2.7667e-01	
16	2vsh	CDP-activated ribitol for teichoic acid precursors in <i>Streptococcus pneumoniae</i>	4.0782e-01	
17	2px7	2-C-methyl-D-erythritol 4-phosphate cytidyltransferase from <i>Thermus thermophilus</i> HB8	1.5287e+00	
18	3cgx	putative Nucleotide-diphospho-sugar Transferase (YP_389115.1) from <i>Desulfovibrio desulfuricans</i> G20	1.7882e+00	
19	1pzt	beta 1,4-galactosyltransferase-I	2.2884e+00	
20	1ezi	sialic acid-activating synthetase, CMP-acylneuraminate synthetase in the presence and absence of CDP	4.3372e+00	

```

# PSIPRED HFORMAT (PSIPRED V3.3)

Conf: 9996312224322667877432599999987540998179999999859899989998612
Pred: CCCCCCCCCCCHHHHHHHHC CCCCCCCCCC EEE EC CCCCCCHHHHHHHHHHHHC
AA: QPFWYPVNRETYIDRLSARYERECEPDELAAVDFVSTVDP LKRFPLI TANTVLSI LAL
      230      240      250      260      270      279

Conf: 599005009975090404569999900643100110200200700090001120344447
Pred: CCCCC EEE EC CCCCCCHHHHHHHHHHHHC CCHHHHHHC CCCCCCCCCC CCCCC
AA: DYPVDRKVSICYIS D GAAML TPE SLVETAD FARKWVPF CKKFS IEPIAPE FYFSQKIDY LK
      290      300      310      320      330      339

Conf: 98890569999999753899999999887324589664223799667899999985525
Pred: CCCCCCHHHHHHHHHHHHHHHHHHHHC CCCCCCCCCC CCCCCCCCCC CCCCC
AA: DKVQPSFVKE PRAMKPDYE EYKIRINALVAKAQKTPD EGWTMQD GTSWPGNNPRDHP GMI
      350      360      370      380      390      399

Conf: 885048987665467768079994148999711011011002334401246986799722
Pred: EE CCCCCCCCCC C EEE C CCCCCCHHHHC CCHHHHHHC CCCCC EEE EC
AA: QVFLGYS GARDIEGNE LPRLVYVS REKRP CYQHKKAGA ENALVRVSAVLTNAP FLLNL
      410      420      430      440      450      459

Conf: 88756915999844411377899740364489422599987945665520000111464
Pred: CCCCCCHHHHHHHHHHC CCCCC EEE C CCCCCCCCCC CCCCCCCCCC CCCCC
AA: CHYVWNSFAVREAMCF LMDPQVGRDVCYVQF PQRFD GIDRSDRYANRNTVVF DVMNGL
      470      480      490      500      510      519

Conf: 57787402213803603222244999988988899987665789988999850233445
Pred: CCCCC RRR C RR C CCCCCCCCCC CCCCCCCCCC CCCCCCCCCC CCCCCCHHHHHH
AA: DGIQGPVYVGTGCVFN RQALYCYGPPSMP SFPKS SSSSC SCCCP GRKRPKDPSEL YRDAK
      530      540      550      560      570      579

Conf: 44556764200014651145454532000011117872567765553189987899158
Pred: HHHHHHHHC CCCCCCHHHHHHHHHHC CCHHHHC CCHHHHHHHHC CCCCCCHH
AA: REELDAAIFNLREIDNYDE YERSMLISQTSFKT FGLSSVFI ESTIMRNGCVAS SANPST
      590      600      610      620      630      639

Conf: 999864500011000143444112120467871565689974598300258887654346
Pred: HHHHHHHHC CCCCCC C C C C E C C C C C H H H H H H H H C C C C C C C C C
AA: LIKRAIHVIS CGYEBKTAJCKEIGWIYGSVT ED ILTCFKMHC RGWES IYCMPLRPAFRGS
      650      660      670      680      690      699

Conf: 78873232032000134532553119
Pred: C C C C H H H H H H H C C C C C E E E E E C C
AA: APINLSDRHQVLRWALGSVEIFLSR
      710      720

```

**Figure 4S.3:** A similar prediction of the secondary structure was obtained using PSIPRED v3.3. Highest confidence prediction areas are given a value of 9, lowest 0. The DCD putative UDP binding motif, residues 459-461, are at the tip of a  $\beta$ -strand. The  $\beta$ -strands highlighted in red were lost with MD refinement. In contrast, a low coil confidence area highlighted magenta (686-689) became a  $\beta$ -strand. Greyed boxes did not appear in the final structure due to their low confidence levels. Yellow highlights the retained  $\beta$ -strands. Green boxes highlight the DD, DCD, ED, and QVLRW motifs.

**Table 4S.2:** Structure quality scores

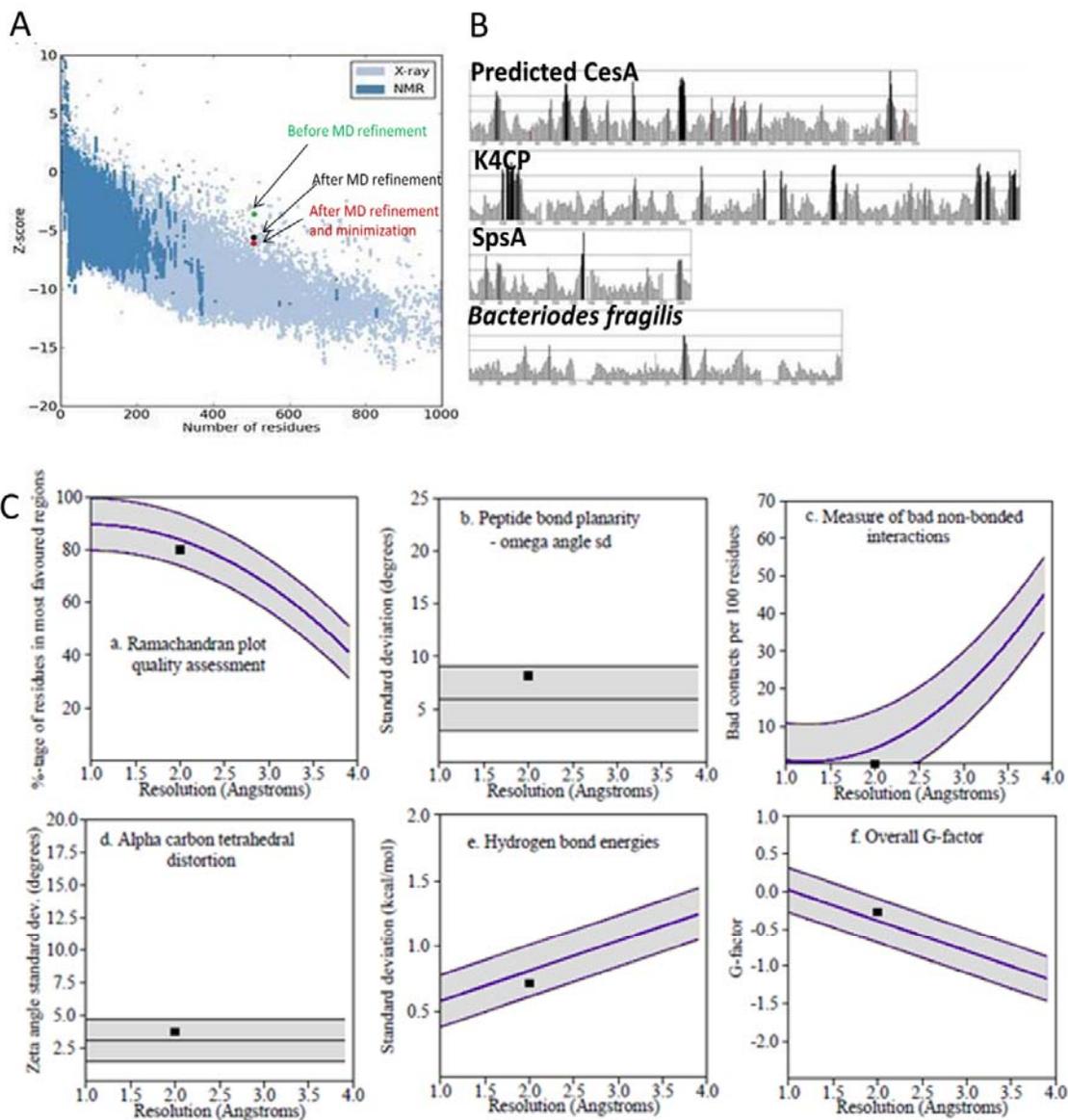
<b>Structure</b>	<b>ProSA Z-Score</b>	<b>Quality Factor (ERRAT2)</b>	<b>AA Length</b>
GhCESA1	-6.09	86.875%	504
SpsA (1qg8)	-7.8	92.411%	241
K4CP (2z86)	-9.16	86.067%	580
(3BCV)	-6.98	98.082%	196

**Figure 4S.4:** Comparison of the quality of the Gh506 structure to experimentally solved structures.

**(A)** Pro-SA Z scores for various stages of GhCESA1 structure prediction (labeled green, black, and red dots) compared to scores of solved structures from the PDB databank (dense blue dots). The initial Z score of the predicted GhCESA1 cytosolic structure (-3.4, green dot) was improved to -5.56 (black dot) after about 4 ns of MD refinement and reached -6.09 (red dot) after a series of MD simulations followed by a short minimization.

**(B)** ERRATv2 analysis of the predicted GhCESA1 cytosolic structure (i) and the solved structures of three other GT-2 enzymes used as templates [ii, K4CP domains A and B (PDB: 2Z86); iii, SpsA (PDB: 1QG8); and iv, a putative glycosyltransferase from *Bacteriodes fragilis* (PDB: 3BCV)]. The histograms show the error value of residues, and the band in the middle of the graph indicates the difference between the lower 95% and the upper 99% value. Of the three crystal structures, the 218 amino acid structure of 3BCV from *Bacteriodes fragilis* exhibited the best score with only B chain residue 40 showing significant error. Areas possibly in need of further refinement in the GhCESA1 predicted structure include residues that either have high local mobility or are deeply buried: (1) N457-V464; (2) D253-V256 that form a  $\beta$ -strand adjacent to the putative UDP binding motif, DCD, in the catalytic core; (3) solvent-exposed P327-I335 that fold back into residues V347-R355 within the P-CR region; (4) P492-G518 that appear to form a loop beside the catalytic site that abuts the QVLRW motif. Even for the SpsA structure, similarly buried residues are nearly impossible to refine fully. For K4CP, core residues around the UDP binding motif of domain “B” shows the greatest error values, probably because they are more mobile and solvent accessible. Similarly, a small region near the UDP-binding motif of SpsA (residues 130-135) also exhibits error values greater than 95% as exemplified by the filled in black bars.

**(C)** Resolution of main chain parameters of Gh506 compared to solved crystallographic structures assessed by ProCheck. In the graphs, the value for the predicted GhCESA1 cytosolic structure is shown by the black square relative to values typical for solved structures (grey band): (a) Ramachandran plot quality is the percentage of the residues in the most favored regions of the Ramachandran plot where a high quality structure is well over 90%, but becomes less at lower resolutions; (b) Peptide bond planarity is a measure of the structure's  $\omega$ -torsion angle where a tight clustering around the ideal  $180^\circ$  represents a planar peptide bond; (c) Bad non-bonded interactions is defined by the number of bad contacts less than or equal to  $2.6 \text{ \AA}$  per 100 residues; (d) C-alpha tetrahedral distortion measures the standard deviation of the zeta torsion angle defined by C- $\alpha$ , N, C and C- $\epsilon$  atoms of a given residue; (e) Main-chain hydrogen bond energy is derived from the measured standard deviation of the hydrogen bond energies in the main chain by the method of Kabsch and Sanders (1983)[24]; (f) Overall G-factor measures the overall normality of the structure as an average of all the different G-factors for each residue.



Stereochemical parameter	No. of data pts	Parameter value	Comparison values		No. of band widths from mean
			Typical value	Band width	
a. %-tage residues in A, B, L	440	80.0	83.8	10.0	-0.4 Inside
b. Omega angle st dev	494	8.2	6.0	3.0	0.7 Inside
c. Bad contacts / 100 residues	0	0.0	4.2	10.0	-0.4 Inside
d. Zeta angle st dev	474	3.8	3.1	1.6	0.4 Inside
e. H-bond energy st dev	250	0.7	0.8	0.2	-0.5 Inside
f. Overall G-factor	506	-0.3	-0.4	0.3	0.4 Inside

**Table 4S.3:** Identity and locations of Gh506 structural features. Entries are in order of appearance in the GhCESA1 cytosolic sequence that was used to generate the Gh506 structure (Fig. 1B). Five of these  $\alpha$ -helices are designated “core  $\alpha$ -helices” because they co-align in the superimposed GT-2 domain of BcsA and the predicted Gh506 structure. Amino acid residue numbers are relative to full-length GhCESA1 (NCBI Accession P93155) or BcsA (NCBI Accession Q3J125; PDB ID 4HG6). Functions ascribed to BcsA are from Ref. [25].

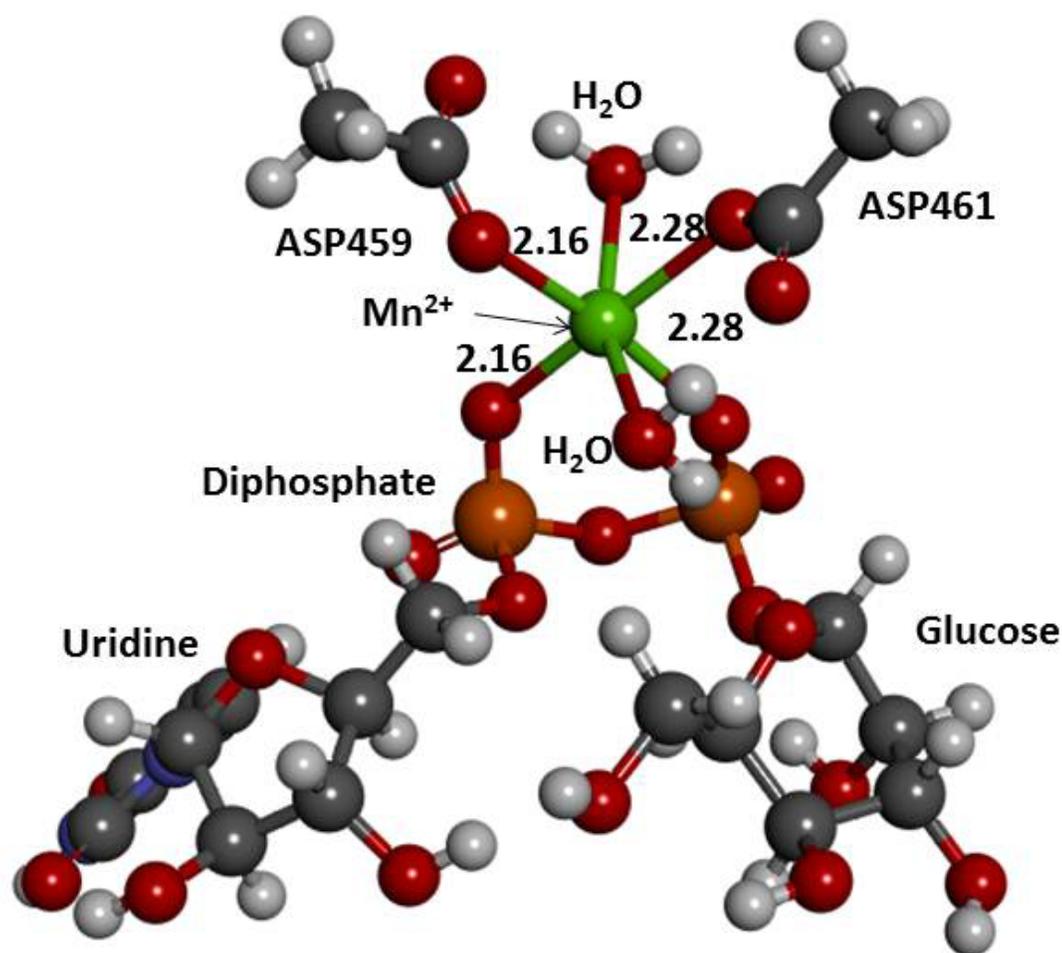
Gh506 major secondary structure elements	Position in GhCESA1, including additional key motifs	Amino Acid Sequence in GhCESA1 of major secondary structure elements and additional key motifs	GhCESA1 residues analogous to Arabidopsis CESA mutations	Structurally co-aligned motifs in the BcsA and the Gh506 GT-2 domains
$\alpha$ -1	I233–E241	IDRLSARYE		
core $\beta$ -2	D253–S257	DFFVS		VDILVPS148
core $\alpha$ -2	L267–A278	LITANTVLSIL		ADMLSVTLAAAKN165
core $\beta$ -1	S287–S291	SCYIS	S291: <i>Atcesa3</i> <sup>S377F</sup> <i>ixr1-6</i> (this paper)	LRTVVLCD179
	D292–G294	DDG		DDG181; D179 coordinates UDP
$\alpha$ -3	E301–K312	ESLVETADFARK		
$\alpha$ -4	P344–K370	PSFVKERRAMKRDYEEYKI RINALVAK, in the P-CR	R351: <i>Atcesa8</i> <sup>R362K</sup> <i>fra6</i> [26]	
$\alpha$ -5	I411–V420	IEGNELPRLV, ending the P-CR		
core $\alpha$ -6	H433–V448	HKKAGAENALVRVSAV; the HKKAGA motif is near DDG.	A436: <i>Atcesa3</i> <sup>A522V</sup> <i>eli1-2</i> [27] A447: <i>Atcesal</i> <sup>A549V</sup> <i>rsw1-1</i> [28]	HAKAGN229; A225 and K226 lie on the other side of the pocket that may accommodate Glc when bound to UDP.
core $\beta$ -3	F454–D459	FILNLD; including the first D of DCD		LVVVF245
	D459–D461	DCD	D459: <i>Atcesa7</i> <sup>D524N</sup> <i>irx3-5</i> [29]	DADH249; D246 coordinates UDP
core $\alpha$ -7	N466–D479	NSKAVREAMCFLMD; crosses several $\beta$ -strands leading toward DCD		FLARTVGY262

**Table S4.3: Continued**

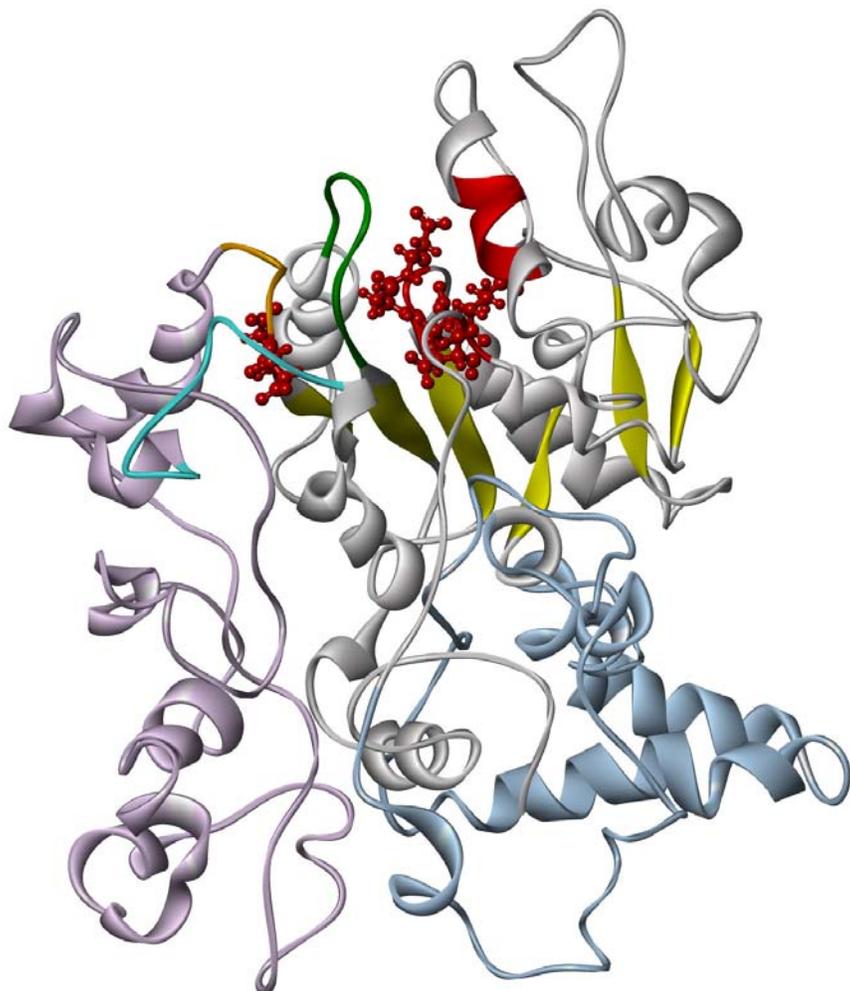
Gh506 major secondary structure elements	Position in GhCESA1, including additional key motifs	Amino Acid Sequence in GhCESA1 of major secondary structure elements and additional key motifs	GhCESA1 residues analogous to Arabidopsis CESA mutations	Structurally co-aligned motifs in the BcsA and the Gh506 GT-2 domains
core $\alpha$ -8	N508–K517	NTVFFDVNMK, within the P492–G518 loop. A longer sequence N508-I521, (NTVFFDVNMKGLDGI), shares sequence conservation of N..F.....GLD.. with Rs BcsA.		Interfacial Helix 1, N298-W312: NEMFYGKIHRLDRW312,
	V525–G531	VYVGTG531, at the end of $\beta$ -4	G529: <i>Atcesa1</i> <sup>G631S</sup> <i>rsw1-2</i> [31] G531: <i>Atcesa3</i> <sup>G617E</sup> <i>cevl</i> [32]	FFCGS320, binds the terminal disaccharide of the glucan acceptor on the opposite side as compared to QRGRW
core $\beta$ -4	C532–N535	CVFN, just before the CSR		AVLR325
core $\beta$ -5	Y488–F491	YVQF		LVQT274
	P492–G518	PQRFDGIDRSDRYANRNTV FFDVNMKG (loop between $\beta$ -4,5 and behind QVLRW), contains core $\alpha$ -8	P492: <i>Atcesa7</i> <sup>P557T</sup> <i>fra5</i> and <i>thanatos</i> [26, 30] G518: <i>Atcesa1</i> <sup>G620E</sup> <i>lycos</i> (this paper)	
$\alpha$ -9	P571–R591	PSELYRDAKREELDAAIFN LR, in the CSR		
core $\beta$ -6	S686–C689	SIYC		SLYI360
core $\alpha$ -13	S705–R725	SDRLHQVLRWALGSVEIFL SR, containing QVLRW		Interfacial Helix 2, F373-R395: FASFIQRGRWATGMMQMLLLK. Contains QRGRW383. R382 coordinates UDP and W383 interacts with the penultimate glucose at the acceptor site

**Table S4.3: Continued**

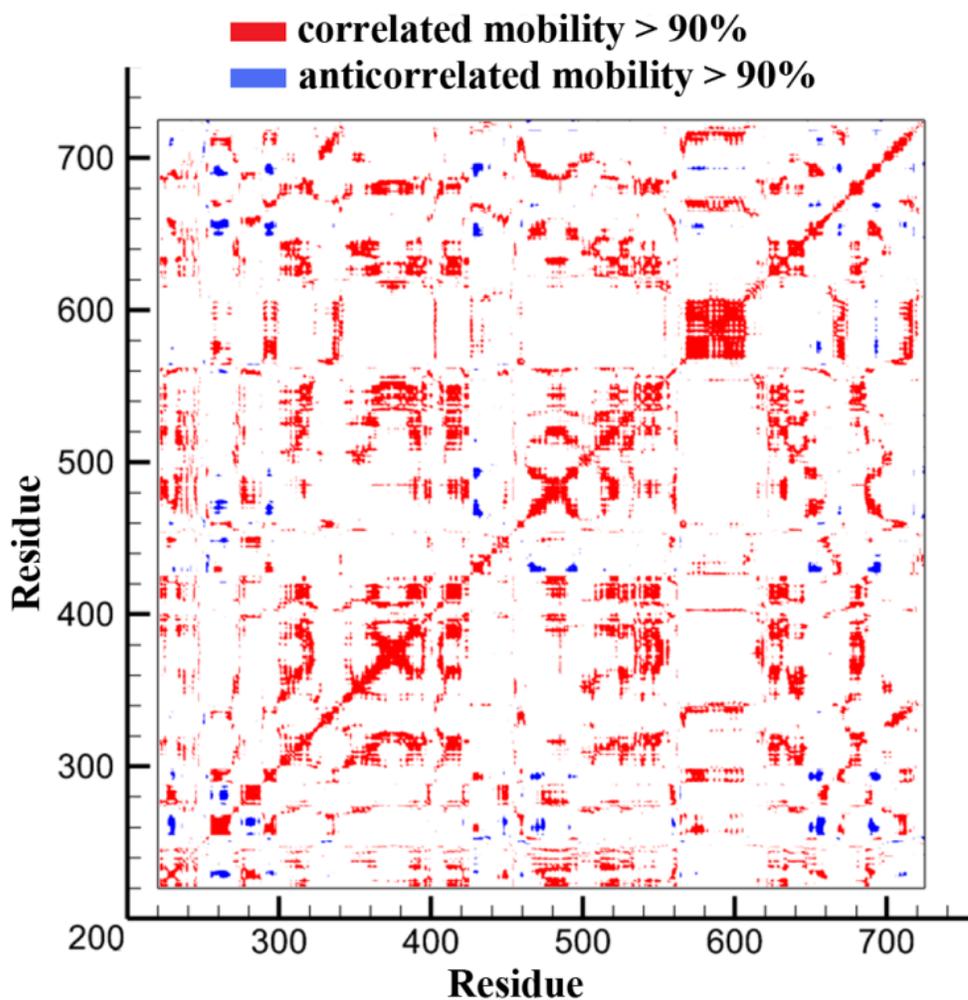
Gh506 major secondary structure elements	Position in GhCESA1, including additional key motifs	Amino Acid Sequence in GhCESA1 of major secondary structure elements and additional key motifs	GhCESA1 residues analogous to Arabidopsis CESA mutations	Structurally co-aligned motifs in the BcsA and the Gh506 GT-2 domains
$\alpha$ -10	Y596–K612	YDEYERSMLISQTSFEK, in the CSR		
$\alpha$ -11	E622–G629	ESTLMENG, in the CSR		
			S668: <i>Atcesa8</i> <sup>S679L</sup> <i>irx1-2</i> [33]	
	T670–D672	TED	E671: <i>Atcesa1</i> <sup>E779K</sup> <i>rsw1-45</i> [34] D672: <i>Atcesa8</i> <sup>D683N</sup> <i>irx1-1</i> and <i>Atcesa1</i> <sup>D780N</sup> <i>rsw1-20</i> [33, 34]	TED343, near the glucan terminus with D343 likely to be the catalytic base. E342 lies on one side of a pocket that may accommodate Glc when bound to UDP
$\alpha$ -12	I673–C681	ILTGFKMHC	H680: <i>Atcesa7</i> <sup>H734Y</sup> <i>mur10-2</i> [35]	



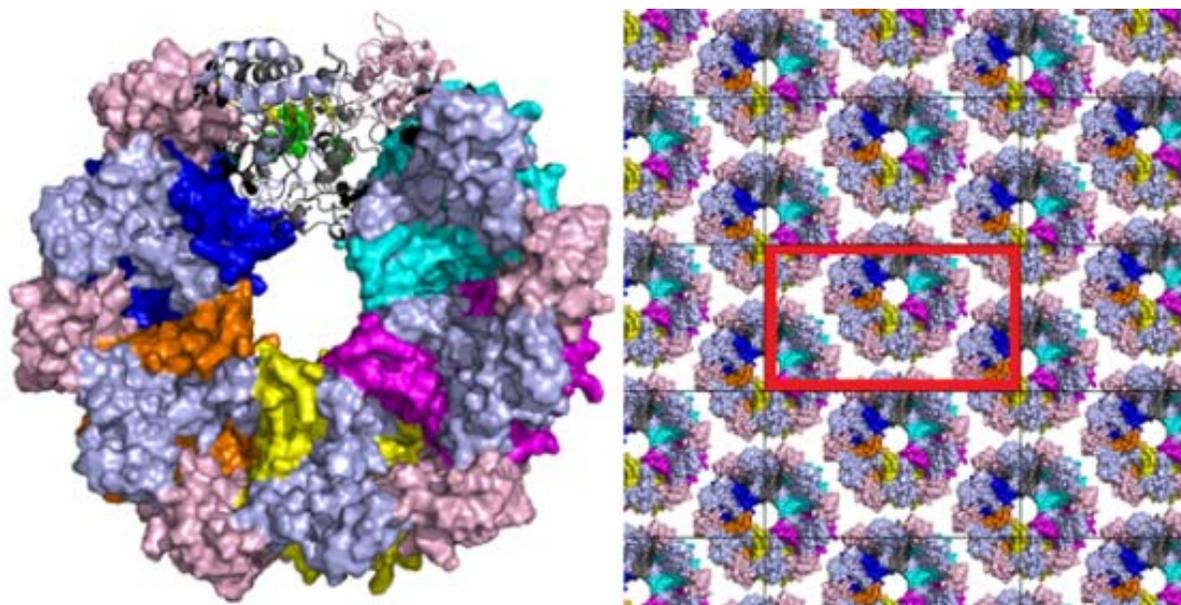
**Figure 4S.5:** Interaction of manganese uridine diphosphate glucose (MnUDP-G) complex with residues of the modeled CESA. The positions of the “D” residues were taken from the CESA structure generated in this study and all atomic positions were allowed to relax to minimum energy positions determined by our DFT methodology. Mn-O distances to carboxylate group of the D residues and to the diphosphate moiety of UDP are given in Angstroms. H=white; C=grey, O=red, N=blue, P=orange, Mn=green. This geometry was used to dock the UDP-Glc into the Gh506 structure in Fig.



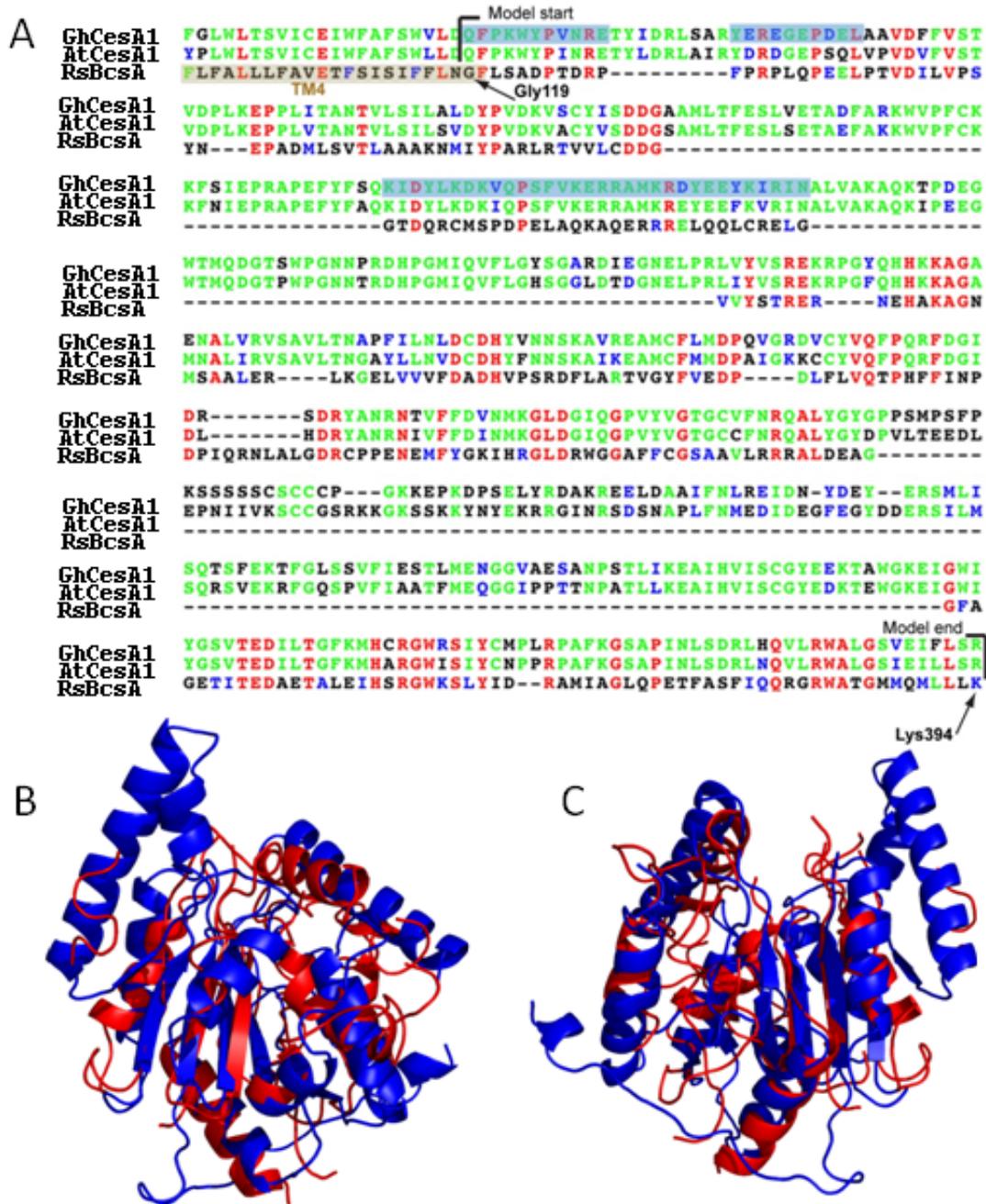
**Figure 4S.6:** Three loops (upper left corner of the image) in the vicinity of the UDP-Glc binding site of the Gh506 structure that may help to control catalysis through modulation of local accessibility to key residues: (1) T258–L267 at the end of  $\beta$ -2 (green); (2) A294–F300, just after DDG and leading into  $\alpha$ 3 of the PCR (orange); and (3) Y421–H432, leading from  $\alpha$ 5 into core  $\alpha$ 6 (aqua). The conserved motifs, DD, DCD, ED, and QLVRW are highlighted red, the  $\beta$  sheet is yellow, the P-CR is pink, and the CSR is blue.



**Figure 4S.7:** Correlated residue motions via atomic fluctuations. The CSR region, residues Y540-W658, shows the greatest motion correlation to itself as expected. The P-CR region, residues A295-V420, shows a self-correlation as well, but not as strong since it is less ordered.



**Figure 4S.8:** Top left view: A possible hexameric assembly of one CESA cytosolic domain isoform (the predicted structure from GhCESA1). One monomer is shown in the ribbon diagram at the top, showing the location of the barely visible  $\beta$ -sheets (yellow) below motifs with conserved D residues (green). The catalytic regions of the other monomers are shown in aqua, magenta, yellow, orange, and dark blue. The light blue and pink regions are the CSR and the P-CR regions, respectively, for all monomers. Top right view: Possible packing of hexameric assemblies into an orthorhombic unit cell of space group  $P2_12_12_1$  (red box). Note that this theoretical possibility for crystallization of hexamers of the predicted GhCESA1 cytosolic region does not imply any preference for hexameric subunits of the rosette CSC *in vivo*. The number of CESAs in the rosette CSC remains an open question.



**Figure 4S.9:** A comparison between Gh506 and RsBcsA sequences and structures. (A) A sequence alignment of a cytosolic region for GhCesA1, AtCesA1, and RsBcsA. (B,C) A top and bottom view of structural alignment between Gh506 (red) and RsBcsA (blue) without the regions that do not have a template and the region shaded blue. This reduced both structures to the GT-A core and the structures align with an overall RMSD of 3.9 Å.

**Table 4S.4:** Summary stability measurements measured as root mean square deviation (RMSD) from the initial structure on the whole structure and on key secondary structure elements of the CESA as a result of mutations over a window of 10 ns.

RMSD of structure during molecular dynamics simulations referenced to the starting structure for each system.

Structure	RMSD, Å
Gh506	2.69 ± 0.55
P557T	3.14 ± 0.53
G620E	3.01 ± 0.70
S377F	2.62 ± 0.41

a) RMSD of  $\alpha$ -helices, Å

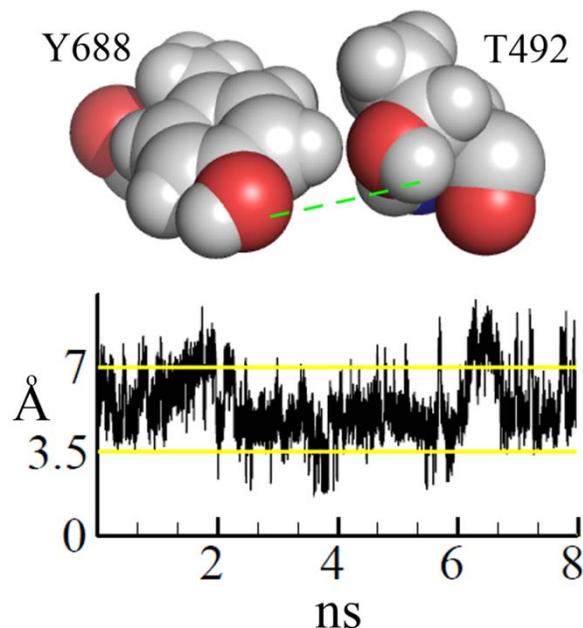
Structure	$\alpha$ -2 helix	$\alpha$ -3 helix	$\alpha$ -7 helix	$\alpha$ -9 helix	$\alpha$ -11 helix	$\alpha$ -13 helix	$\alpha$ -6 helix
Gh506	0.70 ± 0.14	1.21 ± 0.51	0.89 ± 0.27	0.57 ± 0.23	0.42 ± 0.09	0.81 ± 0.14	0.81 ± 0.14
P557T	0.88 ± 0.15	0.93 ± 0.35	0.93 ± 0.27	0.50 ± 0.12	0.44 ± 0.10	1.14 ± 0.31	0.44 ± 0.10
G620E	0.62 ± 0.25	0.63 ± 0.33	0.60 ± 0.10	0.77 ± 0.32	0.34 ± 0.10	0.53 ± 0.12	0.46 ± 0.11
S377F	0.72 ± 0.24	1.00 ± 0.24	0.79 ± 0.21	0.62 ± 0.15	0.42 ± 0.10	0.60 ± 0.20	0.60 ± 0.20

b) RMSD of selected loops, Å

Structure	Loop P492-G518	Loop S257-P266	Loop Y430 -N440
Gh506	1.65±0.30	0.52±0.10	1.10±0.18
P557T	1.15±0.14	0.50±0.10	0.65±0.17
G620E	1.37±0.31	0.45±0.12	0.85±0.17
S377F	1.55±0.54	0.60±.13	1.18±0.21

c) stability of the  $\alpha$ -helical region of the CSR

Structure	Angle formed by residues 598, 608, and 572 with 608 at the vertex (degrees)
Gh506	79.26±6.59
P557T	76.14±4.80
G620E	80.32±12.38
S377F	81.56±6.68



**Figure 4S.10:** Hydrogen bonding of P492T to Y688. The distance cut off is 3.5 Å. The strongest interactions during this time interval for *Ghcesa*<sup>P492T</sup> is before the 4 ns mark. This interaction may serve to stabilize the P492-G518 loop.

## References

1. Karplus, K., *SAM-T08, HMM-based protein structure prediction*. Nucleic Acids Research, 2009. **37**: p. W492-W497.
2. Das, R. and D. Baker, *Macromolecular modeling with Rosetta*. Annual Review of Biochemistry, 2008. **77**: p. 363-382.
3. Sobhany, M., et al., *The Chondroitin Polymerase K4CP and the Molecular Mechanism of Selective Bindings of Donor Substrates to Two Active Sites*. Journal of Biological Chemistry, 2008. **283**(47): p. 32328-32333.
4. Keenleyside, W.J., A.J. Clarke, and C. Whitfield, *Identification of residues involved in catalytic activity of the inverting glycosyl transferase WbbE from Salmonella enterica serovar borreze*. Journal of Bacteriology, 2001. **183**(1): p. 77-85.
5. Urresti, S., et al., *Mechanistic Insights into the Retaining Glucosyl-3-phosphoglycerate Synthase from Mycobacteria*. Journal of Biological Chemistry, 2012. **287**: p. 24649-24661.

6. Duan, Y., et al., *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations*. Journal of Computational Chemistry, 2003. **24**(16): p. 1999-2012.
7. Jorgensen, W.L., et al., *Comparison of Simple Potential Functions for Simulating Liquid Water*. Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
8. Essmann, U., et al., *A smooth particle mesh Ewald method*. The Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.
9. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath*. The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
10. Ryckaert, J.P., Ciccotti, G, Berendsen, H.J.C, *Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*. Journal of Computational Physics, 1977. **23**(3): p. 327-341.
11. Wiederstein, M. and M.J. Sippl, *ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins*. Nucleic Acids Research, 2007. **35**: p. W407-W410.
12. Laskowski, R.A., et al., *Procheck - a Program to Check the Stereochemical Quality of Protein Structures*. Journal of Applied Crystallography, 1993. **26**: p. 283-291.
13. Bhattacharya, A., R. Tejero, and G.T. Montelione, *Evaluating protein structures determined by structural genomics consortia*. Proteins-Structure Function and Bioinformatics, 2007. **66**(4): p. 778-795.
14. Colovos, C. and T.O. Yeates, *Verification of Protein Structures - Patterns of Nonbonded Atomic Interactions*. Protein Science, 1993. **2**(9): p. 1511-1519.
15. Lee, J., S. Wu, and Y. Zhang, *Ab Initio Protein Structure Prediction*. From Protein Structure to Function with Bioinformatics, 2009: p. 3-25.
16. Andre, I., et al., *Prediction of the structure of symmetrical protein assemblies*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(45): p. 17656-17661.
17. Becke, A.D., *A New Mixing of Hartree-Fock and Local Density-Functional Theories*. Journal of Chemical Physics, 1993. **98**(2): p. 1372-1377.
18. Lee, C.T., W.T. Yang, and R.G. Parr, *Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density*. Physical Review B, 1988. **37**(2): p. 785-789.

19. Krishnan, R., et al., *Self-Consistent Molecular-Orbital Methods .20. Basis Set for Correlated Wave-Functions*. Journal of Chemical Physics, 1980. **72**(1): p. 650-654.
20. Mclean, A.D. and G.S. Chandler, *Contracted Gaussian-Basis Sets for Molecular Calculations .1. 2nd Row Atoms, Z=11-18*. Journal of Chemical Physics, 1980. **72**(10): p. 5639-5648.
21. Frisch, M.J., et al., *Gaussian 03*. 2004, Gaussian, Inc.: Wallingford CT.
22. Kormos, B.L., A.M. Baranger, and D.L. Beveridge, *A study of collective atomic fluctuations and cooperativity in the UIA-RNA complex based on molecular dynamics simulations*. Journal of Structural Biology, 2007. **157**(3): p. 500-513.
23. Case, D.A., et al., *Amber 11*. 2010, University of California, San Francisco.
24. Kabsch, W. and C. Sander, *Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features*. Biopolymers, 1983. **22**(12): p. 2577-2637.
25. Morgan, J.L.W., J. Strumillo, and J. Zimmer, *Crystallographic snapshot of cellulose synthesis and membrane translocation*. Nature, 2012(doi:10.1038/nature11744).
26. Zhong, R.Q., et al., *Expression of a mutant form of cellulose synthase AtCesA7 causes dominant negative effect on cellulose biosynthesis*. Plant Physiology, 2003. **132**(2): p. 786-795.
27. Cano-Delgado, A., et al., *Reduced cellulose synthesis invokes lignification and defense responses in Arabidopsis thaliana*. Plant J, 2003. **34**(3): p. 351-62.
28. Arioli, T., et al., *Molecular analysis of cellulose biosynthesis in Arabidopsis*. Science, 1998. **279**(5351): p. 717-20.
29. Liang, Y.K., et al., *Cell wall composition contributes to the control of transpiration efficiency in Arabidopsis thaliana*. Plant J, 2010. **64**(4): p. 679-86.
30. Daras, G., et al., *The thanatos mutation in Arabidopsis thaliana cellulose synthase 3 (AtCesA3) has a dominant-negative effect on cellulose synthesis and plant growth*. New Phytologist, 2009. **184**(1): p. 114-126.
31. Gillmor, C.S., et al., *alpha-glucosidase I is required for cellulose biosynthesis and morphogenesis in Arabidopsis*. Journal of Cell Biology, 2002. **156**(6): p. 1003-1013.
32. Ellis, C., et al., *The Arabidopsis mutant cev1 links cell wall signaling to jasmonate and ethylene responses*. Plant Cell, 2002. **14**(7): p. 1557-1566.

33. Taylor, N.G., S. Laurie, and S.R. Turner, *Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in Arabidopsis*. *Plant Cell*, 2000. **12**(12): p. 2529-2539.
34. Beeckman, T., et al., *Genetic complexity of cellulose synthase A gene function in Arabidopsis embryogenesis*. *Plant Physiology*, 2002. **130**(4): p. 1883-1893.
35. Bosca, S., et al., *Interactions between MUR10/CesA7-dependent secondary cellulose biosynthesis and primary cell wall structure*. *Plant Physiology*, 2006. **142**(4): p. 1353-1363.

# Appendix B

Supporting Information

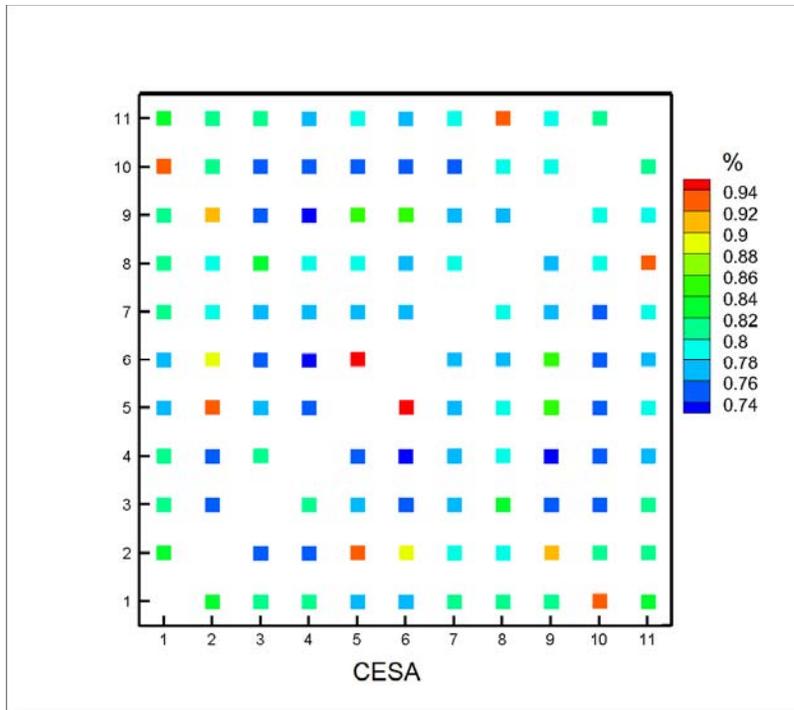
For

## **Chapter 5: Subdomains that Modulate Cellulose Synthase Complex Formation and Terminal Complexes**

Latsavongsakda Sethaphong and Yarosava G. Yingling

911 Partners Way, Materials Science and Engineering, North Carolina State University,  
Raleigh, NC 27695





**Figure 5S.2:** Sequence Similarity of Arabidopsis PCR's (no. 11 corresponds to Ghcesa1)

```

AtCESA2      APKKKKPPGKTCNCWPKWCCLCCG-----LRKK-----SWTKAKD-KKT 38
AtCESA9      APKKKQPPGRTCNCWPKWCCLCCG-----MRKK-----KTGKVKDNQRK 39
AtCESA5      APKKKKTKRMTTCNCWPKWCLFCCG-----LRKN-----RKSKIT--DKK 37
AtCESA6      APKKKKGPRKTCNCWPKWCLLFCG-----SRKN-----RKAKTVAADKK 39
AtCESA4      PPVSEKRKMTCDWPSWICCCCGGGNRNHSKSDSSKKKSGIKSLFSKLNKKTKKKSDDKT 60
AtCESA1      --PVLTEEDLEPNIIIVKSCCGSRKKG-----KSSKKNYNEKR 35
AtCESA10     --PVLTEEDLEPNIIIVKSCFGSRKKG-----KSRKIPNYEDN 35
AtCESA3      --PPIKVKHKKPSLLSKLGGSSRKK-----NSKAKKESDKK 34
AtCESA8      ---PPSKPRILPQSSSSSC-CCLTK-----KKQPQDPSEIY 32
GhCesA1     ---PPSMP-SFPKSSSSSCSCCCPG-----KKEPKDPSELY 32
AtCESA7      ---PKGP-KRPKMISCGCCPCFGR-----RRKNKKFS--- 28

          .
          .

AtCESA2      NTK----ETSKQIHALENVDEGVIVPVSN-VEKRSEATQLKLEKKFGQSPVAVASAVLQ 92
AtCESA9      KFK----ETSKQIHALEHIEEG--LQVIN-AENNSETAQLKLEKKFGQSPVLVASTLLL 91
AtCESA5      KKNR---EASKQIHALEHIEEG--TKGTNDAAKSPEAAQLKLEKKFGQSPVAVASAGME 91
AtCESA6      KKNR---EASKQIHALEHIEEGRVTKGSN-VEQSTEAMQMKLEKKFGQSPVAVASARME 94
AtCESA4      MSSYSRKRSSSTEAFDLEDIEEG--LEGYDELEKSSLSMQKNFEKRFGMSPVFIASITLME 118
AtCESA1      RGIN--RSDSNAPLFNMEDI DEGFEG--YD-DERSILMSQRSVEKRFQSPVFIATFME 90
AtCESA10     RSIK--RSDSNVPLFMEDI DEDEVEG--YE-DEMSLLVSQKRLKRFQSPVFIATFME 90
AtCESA3      KSGR--HTDSTVPVFNLDIEEGVEGAGFD-DEKALLMSQMSLEKRFQSAVFVASTLME 91
AtCESA8      KDAK--REELDAAIFNLGDLDNVDEY----DRSMLISQTSFEKTFGLSTVFIESTLME 84
GhCesA1     RDAK--REELDAAIFNLREIDNVDEY----ERSMLISQTSFEKTFGLSSVFIESTLME 84
AtCESA7     -----KNDMNGDVAALGGAEG-----DKEHLMSEMNFECTFGQSSIFVITSTLME 72

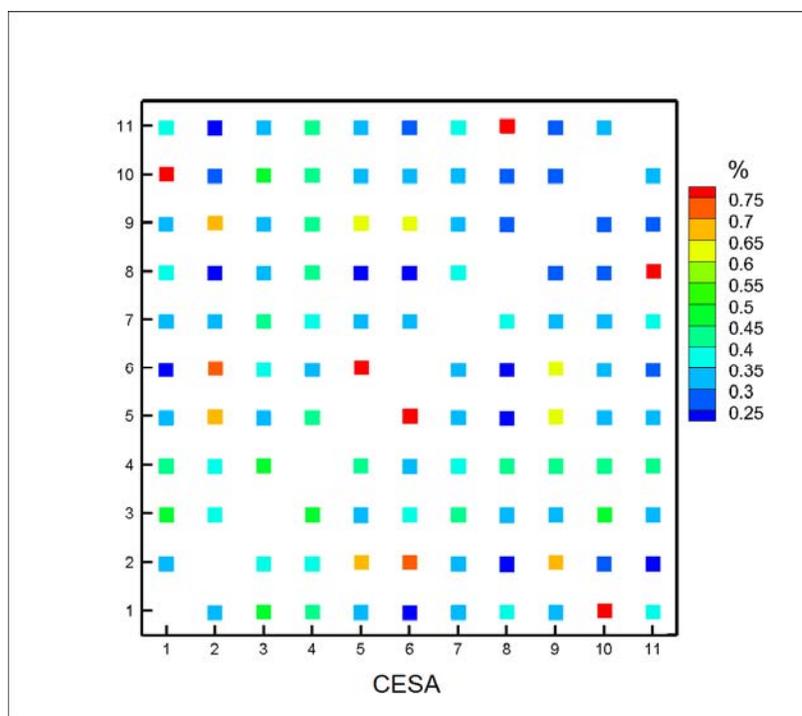
          .      :      :      :      :      :      :      :      :      :      :
          .      :      :      :      :      :      :      :      :      :      :

AtCESA2      NGGVPRNASPACLL 106
AtCESA9      NGGVPSNVNPASLL 105
AtCESA5      NGGLARNASPASLL 105
AtCESA6      NGGMARNASPACLL 108
AtCESA4      NGGLPEATNTSSLI 132
AtCESA1      QGGIPPTINPATLL 104
AtCESA10     QGGLPSTINPLILL 104
AtCESA3      NGGVPPSATPENLL 105
AtCESA8      NGGVPSVNPSTLI 98
GhCesA1     NGGVAESANPSTLI 98
AtCESA7      EGGVPPSSSPAVLL 86

          :      :      :      :      :      :      :      :      :      :
          :      :      :      :      :      :      :      :      :      :

```

**Figure 5S.3:** Sequence alignment of CSR's: Ranked by descending similarity. Consensus regions are underscored with black boxes. Two regions of high variability are boxed in red.



**Figure 5S.4:** Sequence Similarity of Arabidopsis CSR (no. 11 corresponds to Ghcesa1)

**Table 5S.1:** Genbank sequences

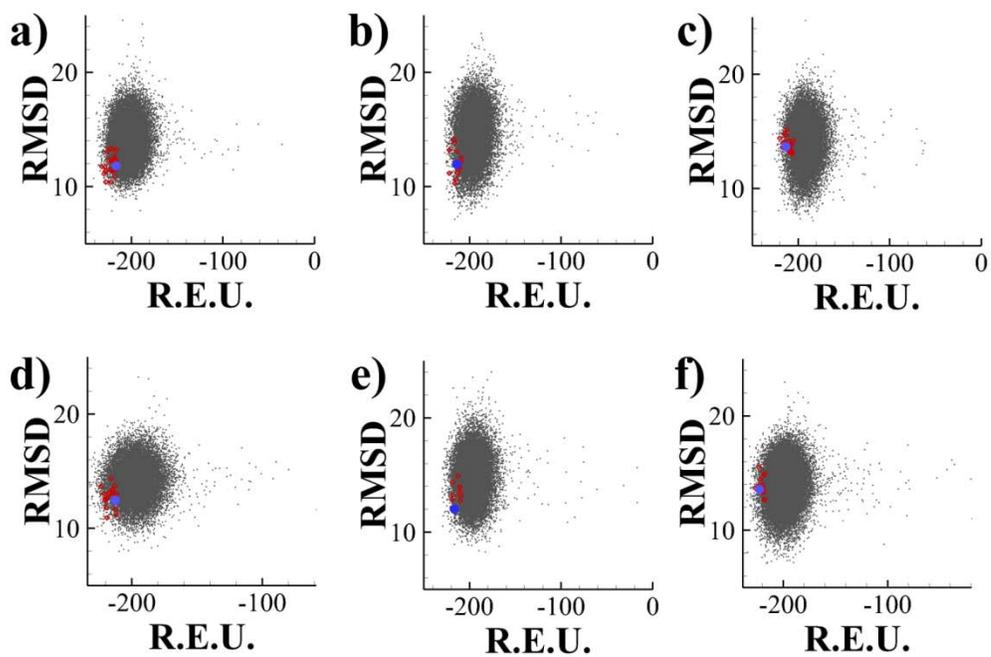
Cesa	Gene	SwissProt Nr. ID	NCBI GI	AA-L	Date
Atcesa1	At4g32410	SwissProt Nr. O48946		1081	
Atcesa2	At4g39350	SwissProt Nr. O48947		1071	
Atcesa3	At5g50170	SwissProt Nr. Q941L0	NCBI GI: 9759258	1065	2004
Atcesa4	At5g44030	SwissProt Nr. Q84JA6 (N-T)*	NCBI GI: 9758562	1043	2004
Atcesa5	At5g09870	SwissProt Nr. Q8L778	NCBI GI: 9758965	1069	2004
Atcesa6	At5g64740	SwissProt Nr. Q94JQ6	NCBI GI: 10177205	1084	2004
Atcesa7	At5g17420	SwissProt Nr. Q9SWW6		1026	
Atcesa8	At4g18780	SwissProt Nr. Q8LPK5		985	
Atcesa9	At2g21770	SwissProt Nr. Q9SJ22		1088	
Atcesa10	At2g25540	SwissProt Nr. Q9SKJ5		1065	
Ghcesa1	Aat64028	SwissProt Nr. P93155	NCBI GI: 49333389	974	

**Table 5S.2:** Top 1K Cut Decoy Results – Plant Conserved Region

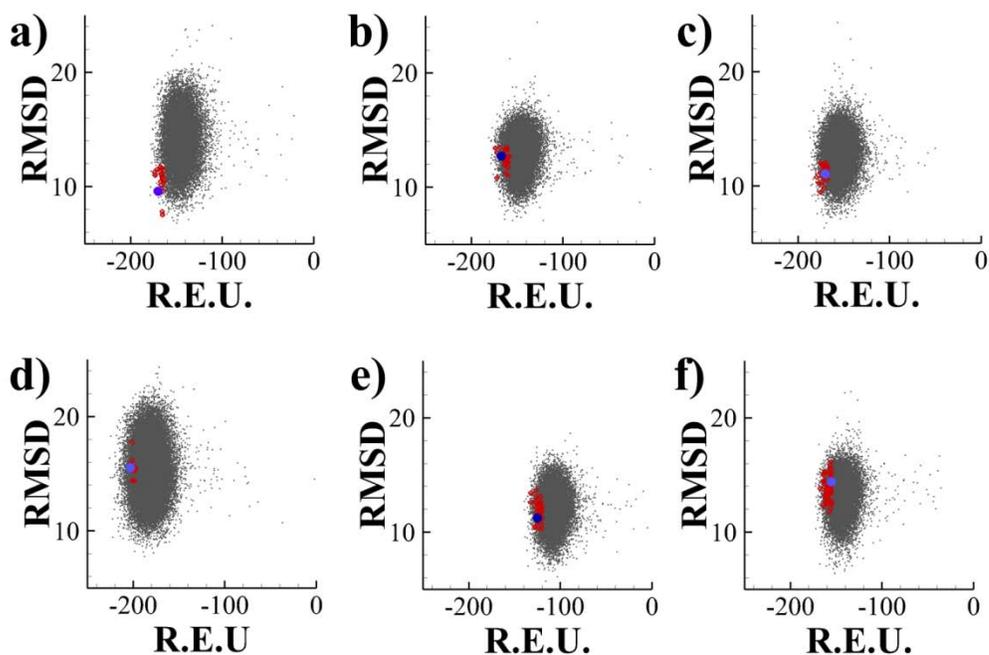
Cesa	Cluster Size	RMSD Cut	REU Score	REU Best	Z-score ProsA	Potential Energy (Tinker)	No. Decoys
Atcesa1	15	8 Å	-181.110	-236.119	-4.65	-803.0604 Kcal/mol	24666
Atcesa2	15	7.5 Å	-218.608	-238.259	-5.4		24385
Atcesa3	10	8 Å	-213.894	-228.969	-3.28		20000
Atcesa4	15	8 Å	-214.09	-240.606	-4.49		16596
Atcesa5	10	8 Å	-213.348	-236.244	-4.94		25942
Atcesa6	13	8 Å	-213.573	-225.775	-5.58		20000
Atcesa7	11	8 Å	-217.011	-233.377	-4.96		23415
Atcesa8	14	8.5 Å	-222.851	-239.66	-5.02		30000
Atcesa9	11	8 Å	-213.204	-229.801	-4.79		20000
Atcesa10	17	8 Å	-214.833	-246.804	-4.37		18365
Ghcesa1	15	8 Å	-217.002	-239.660	-5.28		23255
GhFra6	11	7.5 Å	-216.424	-233.144	-4.8		15531
Fra6	14	8 Å	-220.331	-238.365	-5.03		20000

**Table 5S.3:** Top 1K Cut Decoy Results – Class Specific Region

Cesa	Cluster Size	RMSD Cut	REU Score	REU Best	Z-score ProsA	Potential Energy (Tinker)	No. Decoys
Atcesa1	12	7.82 Å	-169.862	-182.404	-5.65		30000
Atcesa2	19	8 Å	-158.421	-163.873	-3.69		29401
Atcesa3	18	7.5 Å	-162.313	-180.871	-4.7		20000
Atcesa4	13	8.5 Å	-202.898	-225.696			40000
Atcesa5	19	8 Å	-161.394	-177.835	-4.48		20000
Atcesa6	14	8 Å	-175.835	-190.952	-5.02		20000
Atcesa7	51	7.97 Å	-127.297	-146.102	-4.88		20000
Atcesa8	18	7.5 Å	-157.876	-173.244	-6.04		20000
Atcesa9	13	7.5	-157.422	-177.292	-3.23		20000
Atcesa10	16	8 Å	-173.546	-181.419	-3.97		25328
Ghcesa1	17	7 Å	-160.478	-177.762	-4.9		23255



**Figure5S.5:** Folding plots of primary and secondary associated PCRs: PCR1 (a), PCR3(b), PCR6(c), PCR4(d), PCR7(e), PCR8(f).



**Figure 5S.6:** Folding plots of primary and secondary associated CSRs: CSR1 (a), CSR3(b), CSR6(c), CSR4(d), CSR7(e), CSR8(f).

**Table 5S.4:** MM/GBSA Calculations: Free energy of best centroid cut from the top scoring models form the last 2ns of MD in explicit solvent.

	PCR (kcal/mol)	Per	ns	CSR (kcal/mol)	Per	ns
	GBTOT	Atom		GBTOT	Atom	
Atcesa1	-3052.25 ± 32.22	-1.491813	21	-2971.81 ± 28.82	-1.828806	21
Atcesa2	-3351.31 ± 33.25	-1.640387	23	-1914.30 ± 28.65	-1.146287	25
Atcesa3	-3204.09 ± 31.89	-1.576816	23	-2028.89 ± 27.66	-1.258617	20
Atcesa4	-3528.53 ± 33.72	-1.703781	18	-2673.66 ± 33.11	-1.294124	9
Atcesa5	-3510.73 ± 32.18	-1.705066	22	-2309.49 ± 27.77	-1.395462	21
Atcesa6	-3255.37 ± 32.68	-1.591868	22	-2645.08 ± 29.28	-1.540524	22
Atcesa7	-2839.67 ± 34.03	-1.390632	14	-1620.83 ± 25.60	-1.244877	24
Atcesa8	-3498.46 ± 30.94	-1.704072	15	-2163.76 ± 29.70	-1.437714	19
Atcesa9	-2857.09 ± 32.43	-1.406741	19	-2188.70 ± 29.41	-1.320881	14
Atcesa10	-3013.16 ± 32.34	-1.47127	12	-2934.51 ± 28.57	-1.777414	22
Ghcesa1	-3458.33 ± 32.23	-1.68043	22	-2151.46 ± 26.69	-1.45369	20
Fra6	-3321.97 ± 34.87	-1.61968	19	N/A	N/A	

**Table 5S.5: TM-Align Scoring of Minimized Structures**

		Similarity Scores of PCR's minimized Reference								
AtCesA	1	2	3	4	5	6	7	8	9	10
1		0.40118	0.5305	0.42486	0.4044	0.45454	0.47687	0.3939	0.47882	0.4569
2	0.40118		0.46661	0.50972	0.50503	0.51196	0.44806	0.4670	0.4959	0.4831
3	0.5305	0.46661		0.45037	0.41976	0.43216	0.41553	0.3476	0.46951	0.4626
4	0.42486	0.50972	0.45037		0.42573	0.47138	0.52682	0.3710	0.42213	0.4966
5	0.4044	0.50503	0.41976	0.42573		0.47011	0.43209	0.4221	0.49637	0.4221
6	0.45454	0.51196	0.43216	0.47138	0.47011		0.38166	0.3635	0.44002	0.4223
7	0.47687	0.44806	0.41553	0.52682	0.43209	0.38166		0.3401	0.41983	0.4920
8	0.39394	0.46704	0.34769	0.36466	0.42216	0.36353	0.34019		0.34441	0.3005
9	0.47882	0.4959	0.46951	0.42213	0.49637	0.44002	0.41983	0.3444		0.4616
10	0.4569	0.48323	0.46264	0.49663	0.42213	0.42232	0.49204	0.3005	0.46167	

		Similarity Scores of CSR's minimized Reference								
AtCesA	1	2	3	4	5	6	7	8	9	10
1		0.32869	0.3378	0.4788	0.35942	0.3665	0.3252	0.3611	0.3176	0.36618
2	0.32434		0.4311	0.3179	0.32934	0.3191	0.3900	0.2946	0.4004	0.36677
3	0.33541	0.43422		0.3144	0.36876	0.3714	0.3549	0.2792	0.4350	0.36825
4	0.39639	0.2766	0.2653		0.32076	0.3477	0.2951	0.3279	0.2321	0.30154
5	0.35699	0.33132	0.3687	0.3790		0.3463	0.3628	0.3376	0.3649	0.33212
6	0.35708	0.31499	0.3641	0.4028	0.33985		0.3118	0.3190	0.3159	0.33665
7	0.36314	0.43984	0.4035	0.3869	0.41762	0.3621		0.4119	0.3857	0.4291
8	0.37686	0.30901	0.2934	0.3959	0.35586	0.3387	0.3837		0.3346	0.32237
9	0.31582	0.40316	0.4350	0.2803	0.36493	0.3222	0.3463	0.3171		0.40994
10	0.36618	0.37196	0.3706	0.3678	0.33457	0.3450	0.3780	0.3079	0.4127	

**Table 5S.6: RMSD Stability of structures in solvent**

	RMSD PCR (Å)	Stdev	Simulation Time (ps)	RMSD CSR (Å)	Stdev	Simulation Time (ps)
Atcesa1	5.925405	1.34464	24456	5.036512	0.926643	20915
Atcesa2	4.300594	0.687619	23654	4.744969	1.024656	25008
Atcesa3	3.960525	0.635046	22933	4.411094	0.75257	20040
Atcesa4	5.623007	1.285725	19798	5.825955	1.046008	12328
Atcesa5	3.878608	0.662219	24867	4.384192	0.567073	23873
Atcesa6	5.477667	0.866864	22624	3.664802	0.645204	22614
Atcesa7	3.411706	0.542951	16931	5.297202	1.262722	27758
Atcesa8	3.033759	0.336699	17588	4.39951	0.760534	23023
Atcesa9	6.114495	0.978301	22292	5.65756	1.213573	16691
Atcesa10	4.44205	0.708471	14057	4.935239	0.909639	24390

# Appendix C

## Supporting Information

For

### Chapter 2 and Chapter 3 Ion Counting Perl Script and Energetics

1. Perl script to obtain ion dynamics over a trajectory from generated PDB snapshots

```
#!/usr/local/bin/perl

use Math::Trig;
#####
#
# Parsing stdout files from the overlap
# routine of Ptraj to get the cylindrical volumetric
# density of ions within a z distance and radius r
# of the precessing midpoint of a kissing loop
# channel
#
# 9 June 2009, added bulk ion statistics
#             highest residency count
#
# 30 May 2009, added selective surface plot
#
# 19 May 2009, added surface plot feature
#             x,y,z -> z,r,ion_density
# May 2009, t_step is a variable
#
# 9 Sept 2009, changed interaction residence to 4 Angstroms
#             from 5 Angstroms
#
# 2 Oct 2009, sourced out ion pocket occupancy
#             to a data file *.pkt
```



```

#          -----
#                                |X2-X1|^2
#
#
#          = |(X2-X1)x(X1-X0)|^2
#          -----
#                                |X2-X1|^2
#
# d_perp      = |(X2-X1)x(X1-X0)|    ; one may reverse X0-X1
to reverse signage
#          -----
#                                |X2-X1|
# Weisstein, Eric W. "Point-Line Distance--3-Dimensional."
#   From Mathworld -- a Wolfram Web Resource.
#   http://mathworld.wolfram.com/Point-LineDistance3-
Dimensional.html
#=====
# Defining the snapshot PDB file format generated from
# ptraj
#
# ATOM ATOM_NUM TYPE RESNAME RESNUM X Y Z OCCUPANCY B-FACTOR
# 0 1 2 3 4 5 6 7 8 9
# when performing a string split on spaces, one should expect
these
# indexation numbers
#
#=====
# Input parameters then the atom numbers for the items of
intererst
# <location of snapshots>
# <phosphate 1 atom no> <phosphate 2 atom no> <phosphate 3>
<phosphate 4>
# <atom number of first ion> <atom number of last ion>
# <output file name>
#
#=====
# Output format
# give the origin x y z
# give the R distance to origin
# give the d_z distance along the cylinder axis
# give the d_perp distance to the cylinder axis
# give the sodium ID

sub trim($);
# x,y,z operations

```

```

sub get_d_perp;
#sub get_r2x0;
sub get_z_dist;
sub get_midpt;
sub get_r_dist;
sub get_u_vec;
sub dot_prod;
sub cross_prod;
sub move_pt;
sub get_z_phi_idx;
sub get_z_phi_raw;
sub report_stats;

sub trim($)
{
    my $string=shift;
    $string =~ s/^\s+//;
    $string =~ s/\s+$//;
    return $string;
}

sub move_pt {
    my ($x0, $y0, $z0, $dx0, $dy0, $dz0, $mag_d) = @_;
    my @new_pt = (0.0,0.0,0.0);
#    print "$x0,$y0,$z0 moving to";
    $new_pt[0]= $x0 + $mag_d*$dx0;
    $new_pt[1]= $y0 + $mag_d*$dy0;
    $new_pt[2]= $z0 + $mag_d*$dz0;
#    print " @new_pt by $mag_d \n";
return @new_pt
}

#####
# Simple vector operations
# using the X1 as the common origin for both rays
#
#-----
sub dot_prod {
    my ($x0,$y0,$z0,$x1,$y1,$z1,$x2,$y2,$z2) = @_; #take in
the list of arguments
    my $d_perp = 0.0; # evaluated values
    my ($a1,$a2,$a3,$b1,$b2,$b3,$c1,$c2,$c3)=
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0);

# construct vector {X2-X1}

```

```

    $a1=($x2-$x1);#a1
    $a2=($y2-$y1);#a2
    $a3=($z2-$z1);#a3
#   $modulus_a =($a1**2+$a2**2+$a3**2)**0.5 ;
#   construct vector {X1-X0}
#   $b1=($x1-$x0);#b1
#   $b2=($y1-$y0);#b2
#   $b3=($z1-$z0);#b3
#   take position 1 as the common origin
    $b1=($x0-$x1);#b1
    $b2=($y0-$y1);#b2
    $b3=($z0-$z1);#b3

#   $modulus_b =($b1**2+$b2**2+$b3**2)**0.5 ;
#   construct the dot product of {X2-X1}x{X1-X0}
    $c1=$a1*$b1;
    $c2=$a2*$b2;
    $c3=$a3*$b3;
#   $modulus_c = ($c1**2+$c2**2+$c3**2)**0.5;
    $dot_value = $c1+$c2+$c3;
return $dot_value
}
#-----
# Generate the
#
#
#-----
# This function returns the perpendicular distance from
# a point to a line defined by X1 and X2 in 3-D space
# Call the subroutine with an & e.g. $val=&myroutine
# The data will be read as strings from the file, and
# must be converted into a floating point/double number
#-----
sub cross_prod  {
#-----
# X0 X1 and X2 are inputs, to generate the length d_perp
# c.f. construction shown in the beginning
#-----
    my ($x0,$y0,$z0,$x1,$y1,$z1,$x2,$y2,$z2) = @_; #take in the
list of arguments
    my $d_perp = 0.0; # evaluated values
    my ($a1,$a2,$a3,$b1,$b2,$b3,$c1,$c2,$c3)=
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0);
    my ($c1_n, $c2_n, $c3_n) = (0.0,0.0,0.0);

```

```

    my @xx_vec = (0.0,0.0,0.0,0.0,0.0,0.0,0.0); # cross vector
product with normalized unit
# construct vector {X2-X1}
    $a1=($x2-$x1);#a1
    $a2=($y2-$y1);#a2
    $a3=($z2-$z1);#a3
    $modulus_a = ($a1**2+$a2**2+$a3**2)**0.5;
# construct vector {X1-X0}
    $b1= ($x0-$x1);#b1
    $b2= ($y0-$y1);#b2
    $b3= ($z0-$z1);#b3
    $modulus_b = ($b1**2+$b2**2+$b3**2)**0.5;
# construct the cross product of {X2-X1}x{X1-X0}
    $c1=$a2*$b3-$a3*$b2;
    $c2=$a3*$b1-$a1*$b3;
    $c3=$a1*$b2-$a2*$b1;
    $modulus_c = ($c1**2+$c2**2+$c3**2)**0.5; # area of
parallelogram formed by rays X10 and X1
    $d_perp = $modulus_c/$modulus_a;
#normalized unit vector
    $c1_n=$c1/$modulus_c;
    $c2_n=$c2/$modulus_c;
    $c3_n=$c3/$modulus_c;
    @xx_vec = ($c1, $c2, $c3, $d_perp, $c1_n, $c2_n, $c3_n);
return @xx_vec;
}

#=====
# Calculates the magnitude of a ray, r, pointing
# from position X1 to position X2
#-----
sub get_r_dist {
    my ($x1, $y1, $z1, $x2, $y2, $z2) = @_;
    my $r_dist = 0;
    my ($a1,$a2,$a3) = (0.0,0.0,0.0);
    my $modulus_a = 0.0;
# construct vector {X2-X1}
    $a1=($x2-$x1);#a1
    $a2=($y2-$y1);#a2
    $a3=($z2-$z1);#a3
    $modulus_a =($a1**2 +$a2**2 + $a3**2)**0.5 ;

# return the x,y,z and modulus
    @r_dist= ($a1,$a2,$a3,$modulus_a);
return @r_dist;

```

```

}

#=====
# This routine returns the mid point between X1 and X2
# position X1 points to position X2
# the magnitude of this ray is also computed
#-----
sub get_midpt {
  my ($x1,$y1,$z1,$x2,$y2,$z2) = @_;
  my ($xm,$ym,$zm) = (0.0,0.0,0.0);
  my ($xp,$yp,$zp) = (0.0,0.0,0.0);
  my @mid_point = (0.0,0.0,0.0); # only
  my $rm = 0.0;

  $xm= ($x2-$x1);
  $ym= ($y2-$y1);
  $zm= ($z2-$z1);
  $rm= ($xm**2 + $ym**2 + $zm**2)**0.5; # radial distance

  $xp= 0.5*$xm + $x1;
  $yp= 0.5*$ym + $y1;
  $zp= 0.5*$zm + $z1;
  @mid_point = ($xp,$yp,$zp);#,$rm);
return @mid_point; # returns the x,y,z,modulus of the
midpoint
}
#=====
# This routine normalizes the ray pointing from x1 to x2
# as a unit vector. The magnitude of the
# actualray is given as the third component
#-----
sub get_u_vec {
  my ($x1,$y1,$z1,$x2,$y2,$z2) = @_;
  my ($xr,$yr,$zr) = (0.0,0.0,0.0);
  my ($xp,$yp,$zp) = (0.0,0.0,0.0);
  my @u_vec= (0.0,0.0,0.0,0.0);
  $xr= ($x2-$x1);
  $yr= ($y2-$y1);
  $zr= ($z2-$z1);
  $rr= ($xr**2 + $yr**2 + $zr**2)**0.5; # radial distance

#normalize the unit vector
  $xr=$xr/$rr;
  $yr=$yr/$rr;
  $zr=$zr/$rr;

```

```

#   $xp= 0.5*$xm + $x1;
#   $yp= 0.5*$ym + $y1;
#   $zp= 0.5*$zm + $z1;
    @u_vec = ($xr,$yr,$zr,$rr);
return @u_vec;
}

#=====
# Returns the geometric difference between two numbers
#
#-----
sub get_z_dist {

    my ($c_z,$b_z) = @_;
    my $d_z = 0.0; # float
    if ($c_z > $b_z) {
        $d_z=($c_z**2-$b_z**2)**0.5;
    } else {
        $d_z=($b_z**2-$c_z**2)**0.5;
    }
}

return $d_z;
}

#=====
# Returns the z index of a given r,z pair value
# r,z are incremented in 0.5 blocks
# z is in +/- the input value
# may make the block size arbitrary later
#-----
sub get_zr_index {
    my ($z_raw, $r_raw, $c_val) = @_;
    my $zr_idx = 0; # integer
    my ($z_idx, $r_idx) = (0, 0);

    if ($c_val > 0) {          # $z_max is defined globally
        $z_idx = int (($z_raw + $z_max)/0.5); # x form value
    } else {
        $z_idx = int (($z_max - $z_raw)/0.5); # x form value
    }

    $r_idx = int ($r_raw/0.5); # y form value

    $zr_idx = ($z_idx)%(($z_max/0.5)*2) +
($r_idx*($z_max/0.5)*2);
}

```

```

return $zr_idx;
}
#####
# Returns the raw z and r grid values in order to
# generate the x,y,z surface map coordinates
# row major order
#-----
sub get_zr_raw {

    my ($zr_idx) = @_;
    my @zr_raw = (0.0,0.0);
    my ($z_idx,$r_idx)= (0.0,0.0);

#    print "raw index is \ $zr_idx \n";

    $z_idx = ($zr_idx)%(($z_max/(0.5))*2);
    $r_idx = $zr_idx - $z_idx;
    $r_idx = (($r_idx/((($z_max/(0.5))*2))*0.5); # converting
back to half units

    $z_idx = ((($z_idx)-(2*$z_max))*0.5);
    @zr_raw = ($z_idx, $r_idx);

return @zr_raw;
}

#####
# Z and Phi tabulation routines
#-----
# phi is fed in radians, but will be stored as degrees
#-----
sub get_z_phi_idx {
    my ($z_raw, $phi_raw,) = @_;
    my $z_phi_idx = 0; # integer
    my ($z_idx, $phi_idx) = (0, 0);
    # z_raw is fed signed
    $z_idx = int (($z_raw + $z_max)/0.5); # x form value
    #phi_raw is fed signed
    $pi = 3.1415926535897932384626433832795;
    if ($phi_raw > 0) {
        $phi_idx = int ( (((phi_raw/$pi)*180))/5); # y form
value 0 to 76
    } else {

```

```

        $phi_idx = int ( ((360 + ($phi_raw/$pi)*180) )/5); # y
form value 0 to 76
    }
    print "phi_idx $phi_idx \n";
    $z_phi_idx = ($z_idx)%(($z_max/0.5)*2) +
($phi_idx*($z_max/0.5)*2));

return $z_phi_idx;
}

sub get_z_phi_raw {

    my ($z_phi_idx) = @_ ;
    my @z_phi_raw = (0.0,0.0);
    my ($z_idx,$phi_idx)= (0.0,0.0);

#   print "raw index is \ $zr_idx \n";

    $z_idx = ($z_phi_idx)%(($z_max/(0.5))*2); # module
remainder
    $phi_idx = $z_phi_idx - $z_idx;#0 to 71 # remainder in
x/phi direction

#   if ($phi_idx < 36) {
#       $phi_idx = ($phi_idx*5); # converting back to half
units
#   } elsif ($phi_idx > 36) {
#       $phi_idx = $phi_idx*5 - 360;
#   } else {
#       $phi_idx = 180; # equivalent to zero
#   }
    $phi_idx = ($phi_idx*5)/(($z_max/0.5)*2);

    $z_idx = ((($z_idx)-(2*$z_max))*0.5);

    @z_phi_raw = ($z_idx, $phi_idx);

return @z_phi_raw;
}

# write statistics and data files
sub report_stats {

    open(MYFILE, '>>' . $base . 'out') || die;

```

```

#=====
# calculate occupancy statistics
#-----
    $mean_occu = 0;
    $std_dev_occu = 0;
    $mean_occu = ($ion_total/$frame_cnt);
    $variance_occu = ($sqr_occu_sum/$frame_cnt) -
$mean_occu*$mean_occu;
    $std_dev_occu = sqrt($variance_occu);

    print MYFILE "Occupancy Statistics -- var= \
$variance_occu \ \ mean= \ $mean_occu \ \ +/- \ \
$std_dev_occu \n";

    #          close (MYFILE);

    print "Occupancy Statistics -- var= \ $variance_occu
\ \ mean= \ $mean_occu \ \ +/- \ \ $std_dev_occu \n";
    print "Totals= \ \ $ion_total \ \n";
#=====
# calculate bulk statistics
#-----
    $mean_bulk = 0;
    $std_dev_bulk = 0;
    $mean_bulk = ($bulk_total/$frame_cnt);
    $variance_bulk = ($sqr_bulk_sum/$frame_cnt) -
$mean_bulk*$mean_bulk;
    $std_dev_bulk = sqrt($variance_bulk);

    print MYFILE "Bulk Occupancy Statistics -- var= \
$variance_bulk \ \ mean= \ $mean_bulk \ \ +/- \ \
$std_dev_bulk \n";

    #          close (MYFILE);
    print "Bulk Occupancy Statistics -- var= \
$variance_bulk \ \ mean= \ $mean_bulk \ \ +/- \ \
$std_dev_bulk \n";
    print "Totals= \ \ $bulk_total \ \n";
    $i = 0;
    $i_dum = 0;
    $highest_time = 0;
    $highest_id = 0;
    while ($i < $ion_high_numid) {
        $i_dum = $i + 1;

```

```

        if ($ion_high_residency_cnt[$i_dum] >
$highest_time) {
            $highest_time =
$ion_high_residency_cnt[$i_dum];
            $highest_id = $i_dum;
        }

        # print "Ion -- $i_dum \ \ No. \ \
$ion_residency_cnt[$i_dum] \ \n";
        $i = $i + 1;
    }
    $highest_time = $t_step*$highest_time;
    print "Highest Occupancy Ion -- \ $highest_id
\ \ time= \ $highest_time \ \n";

    #=====
    # calculate cylinder statistics
    #-----
        $mean_cylinder = 0;
        $std_dev_cylinder = 0;
        $mean_cylinder =
($cylinder_total/$frame_cnt);
        $variance_cylinder =
($sqr_cylinder_sum/$frame_cnt) -
$mean_cylinder*$mean_cylinder;
        $std_dev_cylinder = sqrt($variance_cylinder);

        print MYFILE "Cylinder Occupancy Statistics -- var= \
$variance_cylinder \ \ mean= \ $mean_cylinder \ \ +/- \ \
$std_dev_cylinder \n";

        close (MYFILE);

        print "Cylinder Occupancy Statistics -- var= \
$variance_cylinder \ \ mean= \ $mean_cylinder \ \ +/- \ \
$std_dev_cylinder \n";
        print "Totals= \ \ $cylinder_total \ \n";
        $i = 0;
        $i_dum = 0;
        $highest_time = 0;
        $highest_id = 0;
        while ($i < $ion_high_numid) {
            $i_dum = $i + 1;
            if ($ion_cylinder_high_residency_cnt[$i_dum]
> $highest_time) {

```

```

        $highest_time =
$ion_cylinder_high_residency_cnt[$i_dum];
        $highest_id = $i_dum;
    }

    # print "Ion -- $i_dum \ \ No.    \ \
$ion_cylinder_residency_cnt[$i_dum] \ \n";
    $i = $i + 1;
}
    $highest_time = $t_step*$highest_time;
    print "Highest Cylinder Occupancy Ion -- \
$highest_id \ \ time= \ $highest_time \ \n";

#=====
# dump ion grid count
#-----
    open(MYFILE, '>'.$base.'.'.'grid');
    close(MYFILE);

    open(MYFILE, '>>'.$base.'.'.'grid')||die;
    print "grid size is \ $ion_grid_sz \n";

    $g_idx = 0;

    for ($grid_idx = 0; $grid_idx < $ion_grid_sz;
$grid_idx++) {
        $g_idx = $grid_idx + 1; # taking advantage
        @zr_raw = get_zr_raw($g_idx);
        # print "$zr_raw[0] \ $zr_raw[1] \
$ion_grid_cnt[$g_idx] \n";
        if ($ion_grid_cnt[$g_idx] > 0) {
            print MYFILE "$zr_raw[0] \ $zr_raw[1] \
$ion_grid_cnt[$g_idx] \ \n";
        }
    }

    close(MYFILE);

#=====
# dump residue HIGHEST residence ion times into
tabular format
#-----
    open(MYFILE, '>'.$base.'.'.'hire');
    close(MYFILE);

```

```

    open(MYFILE, '>>'.$base.'.'.'hire')||die;
print "writing ion high occupancies \n";

    $g_idx = 0;

    print MYFILE "ION ByRes ";
    for ($ii_idx = 0; $ii_idx < $res_B;
$ii_idx++) {
        $iii_idx = $ii_idx + 1;
        print MYFILE ", $iii_idx ";
        }
    print MYFILE "\n";

    for ($ion_idx = 0; $ion_idx < $bulk_ion_tot;
$ion_idx++) {
        $sis_ion = $ion_idx + 1;
        #
        print "doing $res_B residues for ion
$sis_ion\n";
        for ($res_idx = 0; $res_idx < $res_B;
$res_idx++) {
            $sis_res = $res_idx + 1;

            $g_idx = ($ion_idx )*$res_B +
$sis_res;

            if ($res_idx == 0) {
                #
                print "dump one at $g_idx \n";
                if ($ion_is_here_high[$g_idx] > 0)
                {
                    print MYFILE " $sis_ion , \
$ion_is_here_high[$g_idx] \ ";
                } else {
                    print MYFILE " $sis_ion, 0 \ ";
                }
            } else {
                if ($ion_is_here_high[$g_idx] > 0)
                {
                    print MYFILE " , \
$ion_is_here_high[$g_idx] \ ";
                } else {
                    print MYFILE " , 0 \ ";
                }
            }
        }
    }
}

```

```

        print MYFILE "\n"; # end of line
#         print "finished ion $is_ion \n";
    }

    close(MYFILE);

    #=====
    # HIGHEST Ion Residence Times Dump in x,y,z
format for tecplot
    #-----
        open(MYFILE, '>'.$base.'.'.'hitc');
        close(MYFILE);

        open(MYFILE, '>>'.$base.'.'.'hitc')||die;
        print "writing ion high occupancies for
tecplot \n";

        $g_idx = 0;

        for ($ion_idx = 0; $ion_idx < $bulk_ion_tot;
$ion_idx++) {
            $is_ion = $ion_idx + 1;
#            print "doing $res_B residues for ion
$is_ion\n";
            for ($res_idx = 0; $res_idx < $res_B;
$res_idx++) {
                $is_res = $res_idx + 1;

                $g_idx = ($ion_idx )*$res_B +
$is_res;

                #                dumping by ion as x, y as
residue, occupancy as z                if ($ion_is_here_high[$g_idx] > 0)
{
                    print MYFILE " $is_ion, $is_res,
$ion_is_here_high[$g_idx] \n";
                    }# else {
                    # print MYFILE " $is_ion,
$is_res, 0 \n";
                    # }
                }
            }
        }
    }

```

```

        close(MYFILE);
#-----

#=====
# DUMPING CUMULATIVE DATA
#=====

#-----
# CUMULATIVE Residence ion times into tabular
format
#-----
        open(MYFILE, '>'.$base.'.'.'tore');
        close(MYFILE);
        open(MYFILE, '>>'.$base.'.'.'tore')||die;
        print "writing cumulative ion occupancies
\n";

        $g_idx = 0;

        print MYFILE "ION ByRes ";
        for ($ii_idx = 0; $ii_idx < $res_B;
$ii_idx++) {
                $iii_idx = $ii_idx + 1;
                print MYFILE ", $iii_idx ";
                }
        print MYFILE "\n";

        for ($ion_idx = 0; $ion_idx < $bulk_ion_tot;
$ion_idx++) {
                $is_ion = $ion_idx + 1;
                #
                print "doing $res_B residues for ion
$is_ion\n";
                for ($res_idx = 0; $res_idx < $res_B;
$res_idx++) {
                        $is_res = $res_idx + 1;

                        $g_idx = ($ion_idx )*$res_B +
$is_res;

                        if ($res_idx == 0) {
                                #
                                print "dump one at $g_idx \n";
                                if ($ion_is_here_tot[$g_idx] > 0)
{

```

```

        print MYFILE " $sis_ion , \
$ion_is_here_tot[$g_idx] \ ";
        } else {
        print MYFILE " $sis_ion, 0 \ ";
        }
    } else {
        if ($ion_is_here_high[$g_idx] > 0)
    {
        print MYFILE " , \
$ion_is_here_tot[$g_idx] \ ";
        } else {
        print MYFILE " , 0 \ ";
        }
    }
}

#
        print MYFILE "\n"; # end of line
        print "finished ion $sis_ion \n";
    }

    close(MYFILE);
#-----

#=====
# HIGHEST Ion Residence Times Dump in x,y,z
format for tecplot
#-----
    open(MYFILE, '>'.$base.'.'.'totc');
    close(MYFILE);

    open(MYFILE, '>>'.$base.'.'.'totc')||die;
    print "writing cumulative ion occupancies for
tecplot \n";

    $g_idx = 0;

    for ($ion_idx = 0; $ion_idx < $bulk_ion_tot;
$ion_idx++) {
        $sis_ion = $ion_idx + 1;
#
        print "doing $res_B residues for ion
$sis_ion\n";
        for ($res_idx = 0; $res_idx < $res_B;
$res_idx++) {
            $sis_res = $res_idx + 1;

```

```

                                $g_idx = ($ion_idx )*$res_B +
$sis_res;

#                                dumping by ion as x, y as
residue, occupancy as z                                if ($ion_is_here_tot[$g_idx] > 0)
{
                                print MYFILE " $sis_ion, $sis_res,
$ion_is_here_tot[$g_idx] \n";
                                }# else {
$sis_res, 0 \n";                                # print MYFILE " $sis_ion,
                                # }
                                }
                                }

                                close(MYFILE);
#-----

#=====
# cumulative frequency by residue
# dump ion tabular format
#=====
                                open(MYFILE, '>'.$base.'.'.'tab');
                                close(MYFILE);

                                open(MYFILE, '>>'.$base.'.'.'tab')||die;
                                $ion_grid_sz = $z_max*4;
                                print "grid size is \ $ion_grid_sz \n";

                                $g_idx = 0;

                                print MYFILE "z/r"; # first entry
                                # generate header
                                for ($grid_idx = 0; $grid_idx < $z_max*4;
$grid_idx++) {

                                @zr_raw = get_zr_raw($grid_idx);
                                print MYFILE "\ @zr_raw[0]";
                                }
                                print MYFILE " \n";

                                $g_idx = 0;

```

```

        for ($grid_idx = 0; $grid_idx < $ion_grid_sz;
$grid_idx++) {
            $g_idx = $grid_idx + 1; # taking advantage
            @zr_raw = get_zr_raw($g_idx);

#            print "$zr_raw[0] \ $zr_raw[1] \
$ion_grid_cnt[$g_idx]";
            if ($ion_grid_cnt[$g_idx] > 0) {

                if ($zr_raw[0] < -5.5) {
                    print MYFILE "$zr_raw[1] \
$ion_grid_cnt[$g_idx]";
                } else {
                    if ($zr_raw[0] < 5.5) {
                        print MYFILE "\
$ion_grid_cnt[$g_idx]";
                    } else {
                        print MYFILE "\
$ion_grid_cnt[$g_idx] \n";
                    }
                }

            } else {

                if ($zr_raw[0] < -5.5) {
                    print MYFILE "$zr_raw[1] \ 0";
                } else {
                    if ($zr_raw[0] < 5.5) {
                        print MYFILE "\ 0";
                    } else {
                        print MYFILE "\ 0 \n";
                    }
                }

            }

        }

        close(MYFILE);

#=====
# dump z phi count
#-----

# dump residue count

```

```

open(MYFILE, '>'. $base.'.'.'zphi');
close(MYFILE);

open(MYFILE, '>>'. $base.'.'.'zphi')||die;
#print "grid size is \ $ion_grid_sz \n";

$g_idx = 0;
$z_phi_grid_sz = ($z_max/0.5)*2*(360/5);
print "grid size is $z_phi_grid_sz \n";
for ($grid_idx = 0; $grid_idx <
$z_phi_grid_sz; $grid_idx++) {
    $g_idx = $grid_idx + 1; # taking advantage
    # $idx_z_phi = get_z_phi_idx($dzc,
$phi_value); # increment phi by 0.0872664625997 radians or 5
degrees
    #
$z_phi_table[$idx_z_phi]=$z_phi_table[$idx_z_phi] + 1;
    #print "made z phi index, $idx_z_phi for
$dzc, $phi_value\n";
    @z_phi_raw = get_z_phi_raw($g_idx);
    if ($z_phi_table[$g_idx] > $thresh ) { #
was 0
        # print "$zr_raw[0] \ $zr_raw[1] \
$ion_grid_cnt[$g_idx] \n";
        print MYFILE "$z_phi_raw[0] \
$z_phi_raw[1] \ $z_phi_table[$g_idx] \ \n";
    }
}

close(MYFILE);

# Dump stemp B z phi table
open(MYFILE, '>'. $base.'.'.'zphiA');
close(MYFILE);

open(MYFILE, '>>'. $base.'.'.'zphiA')||die;
#print "grid size is \ $ion_grid_sz \n";

$g_idx = 0;
$z_phi_grid_sz = ($z_max/0.5)*2*(360/5);
print "grid sizeA is $z_phi_grid_sz \n";
for ($grid_idx = 0; $grid_idx <
$z_phi_grid_sz; $grid_idx++) {
    $g_idx = $grid_idx + 1; # taking advantage

```

```

        #   $idx_z_phi = get_z_phi_idx($dzc,
$phi_value); # increment phi by 0.0872664625997 radians or 5
degrees
        #
$z_phi_table[$idx_z_phi]=$z_phi_table[$idx_z_phi] + 1;
        #print "made z phi index, $idx_z_phi for
$dzc, $phi_value\n";
        @z_phi_raw = get_z_phi_raw($g_idx);
        if ($z_phi_tableA[$g_idx] > $thresh) { #
was zero
        #   print "$zr_raw[0] \ $zr_raw[1] \
$ion_grid_cnt[$g_idx] \n";
        print MYFILE "$z_phi_raw[0] \
$z_phi_raw[1] \ $z_phi_tableA[$g_idx] \ \n";
        }
    }

    close(MYFILE);

# Dump stem B z phi table
    open(MYFILE, '>'.$base.'.'. 'zphiB');
    close(MYFILE);

    open(MYFILE, '>>'.$base.'.'. 'zphiB')||die;
    #print "grid size is \ $ion_grid_sz \n";

    $g_idx = 0;
    $z_phi_grid_sz = ($z_max/0.5)*2*(360/5);
    print "grid sizeB is $z_phi_grid_sz \n";
    for ($grid_idx = 0; $grid_idx <
$z_phi_grid_sz; $grid_idx++) {
        $g_idx = $grid_idx + 1; # taking advantage
        #   $idx_z_phi = get_z_phi_idx($dzc,
$phi_value); # increment phi by 0.0872664625997 radians or 5
degrees
        #
$z_phi_table[$idx_z_phi]=$z_phi_table[$idx_z_phi] + 1;
        #print "made z phi index, $idx_z_phi for
$dzc, $phi_value\n";
        @z_phi_raw = get_z_phi_raw($g_idx);
        if ($z_phi_tableB[$g_idx] > $thresh ) { #
was zero
        #   print "$zr_raw[0] \ $zr_raw[1] \
$ion_grid_cnt[$g_idx] \n";

```

```

        print MYFILE "$z_phi_raw[0] \
$z_phi_raw[1] \ $z_phi_tableB[$g_idx] \ \n";
    }
}

close(MYFILE);

#=====
# Base Frequency tables and close neighbor
statistics
#-----
open(MYFILE, '>'.$base.'.'.'resf');
close(MYFILE);

$g_idx = 0;
$my_type = -1;
$t_idx = 0;
open(MYFILE, '>>'.$base.'.'.'resf')||die;
for ($grid_idx = 0; $grid_idx < $res_max ;
$grid_idx++) {
    $g_idx = $grid_idx + 1;
    $t_idx = 7*$grid_idx; # 0 to 6 for seven
cells

    $my_type = $atom_type_stat[$t_idx + 6];
# get residue type

    if ($residue_box[$g_idx] > 0) {
        print MYFILE "$g_idx
$residue_box[$g_idx] ";}
    else {
        print MYFILE "$g_idx 0 ";}
    print "$my_type";
    if ($my_type > 0 && $my_type < 2) {
print MYFILE " G ";}
    elsif ($my_type > 1 && $my_type < 3) {
print MYFILE " A ";}
    elsif ($my_type > 2 && $my_type < 4) {
print MYFILE " C ";}
    elsif ($my_type > 3 && $my_type < 5) {
print MYFILE " U ";}
    elsif ($my_type =~ m/0/) { print MYFILE
" T ";} else {print MYFILE " | "; print " | \n"; }
}

```

```

                $total_type = $atom_type_stat[$t_idx +
2];
                if ($total_type > 0) {
                    $mean_type = $atom_type_stat[$t_idx
+ 1]/$total_type;
                    $sqr_type_sum =
$atom_type_stat[$t_idx + 0];

                    $std_dev_type = 0;

                    $variance_type =
($sqr_type_sum/$total_type) - $mean_type*$mean_type;
                    $std_dev_type = sqrt($variance_type);

                    if ($my_type > 0 && $my_type < 2) {
# G
                        print MYFILE " $total_type 06 \
$mean_type +/- $std_dev_type ";
                    } elseif ($my_type < 1) { # T
                        print MYFILE " $total_type 04 \
$mean_type +/- $std_dev_type ";
                    } elseif ($my_type > 2) { # C U
                        print MYFILE " $total_type 04 \
$mean_type +/- $std_dev_type ";
                    } else {
                        print MYFILE " 0 NA \ 0 +/- 0 ";
# not applicable for A
                    }

                }

                $total_type = 0;
                $total_type = $atom_type_stat[$t_idx +
5];
                if ($total_type > 0) {
                    $mean_type = $atom_type_stat[$t_idx
+ 4]/$total_type;
                    $sqr_type_sum =
$atom_type_stat[$t_idx + 3];

                    $std_dev_type = 0;

                    $variance_type =
($sqr_type_sum/$total_type) - $mean_type*$mean_type;

```

```

        $std_dev_type = sqrt($variance_type);

        if($my_type > 0 && $my_type < 2) { #G
            print MYFILE " $total_type N7
$mean_type +/- $std_dev_type ";
        } elseif($my_type > 1 && $my_type <
3) { #A
            print MYFILE " $total_type N7
$mean_type +/- $std_dev_type ";
        } else { #T,C,U
            print MYFILE " $total_type O2
$mean_type +/- $std_dev_type ";
        }
    }

    print MYFILE "\n";

}
close(MYFILE);

print " @residue_type \n";

return 0;
}
=====
# END OF SUBROUTINE DECLARATIONS
=====
# BEGIN MAIN PROGRAM
=====
# if the last element is les than zero, no arguments were
# passed
if($#ARGV < 0){
    print " Incorrect usage...\n";
    exit;
}

$j=1;
$atom_cnt=0; # number of atoms per set in proximity
$frame_cnt=0; # frame occurrence
$atom_total=0; # number of all atoms in the area for all
frames
$k=1;

```

```

    $base =      $ARGV[0]; # base name of pdb files
    $phos1=     $ARGV[1]*1.0; # Atom number of the first
phosphorus
    $phos2=     $ARGV[2]*1.0; # Atom number of the second
phosphorus
    $phos3=     $ARGV[3]*1.0; # Atom number of the third
phosphorus
    $phos4=     $ARGV[4]*1.0; # Atom number of the fourth
phosphorus
    $ion_start= $ARGV[5]*1.0; # First Ion's Atom Number
    $ion_stop=  $ARGV[6]*1.0; # Last Ion Atom Number
    $r_max=    $ARGV[7]*1.0; # Max radius of cylinder
    $z_max=    $ARGV[8]*1.0; # Max Length of cylinder from
center
    $t_step=   $ARGV[9]*1.0; # expected time between frames
    $t_max =   $ARGV[10]*1.0; # optional max length of time
    $br_min =  $ARGV[11]*1.0; # radius of RNA plus 5 Angstroms
    $bz_min =  $ARGV[12]*1.0; # 1/2 z length of entire RNA
    $pi_n     = $ARGV[13]*1.0; # pi division by n used to check
helical ion distribution
    $t_start = $ARGV[14]*1.0; # starting time
    $res_A =   $ARGV[15]*1.0; # end of strand A
    $res_B =   $ARGV[16]*1.0; # end of strand B
    $thresh =  $ARGV[17]*1.0; # threshold occupancy for z phi
charts
    $ia_dist = $ARGV[18]*1.0; # minimum interactin distance
btw ion & nearest nuclei
#   $res_T =   $ARGV[18]*1.0; # occupancy of specific residue
id

    print  "p1\ $phos1 \ p2\  $phos2 \ p3\  $phos3 \ p4\
$phos4\n";
    print  "Na start \ $ion_start \ Na stop \ $ion_stop \n";
    print  "Channel radial dist max \ $r_max \ radial axis max \
$z_max \n";
    print  "Bulk radial dist min \ $br_min \ axis min \ $bz_min
\n";
    print  "Residues Number \ $res_B \ min occupancy is \
$thresh \n";
    print  "start time is \ $t_start \ stop time is \ $t_max
\n";
#foreach $i ( 0.. $#ARGV ) {
    open(MYFILE, '>' . $base . '.' . 'out') || die; # overwrite old
files

```

```

close(MYFILE);
$k = 1;
if ($t_start > 0) {
$k = $t_start;
}

# atomic interaction distance
if ($ia_dist < 1.0) {
    $ia_dist = 3.2; # default interaction distance
}

print "Interaction distance is \ $ia_dist \n";

if ($pi_n > 0) {
    $pi_n = 180/$pi_n; # in degrees
}
$frame_cnt = 0;
$ion_total = 0;
$atom_numid = 0;
$ion_numid = 0;
$attempfile = 0;
# do for the first 20 ns
$bulk_ion_tot = $ion_stop - $ion_start + 1.0; # total bulk
ions

print "total ions \ $bulk_ion_tot \ \n";
print "----- Beginning Analysis -----
\n";

# allow for proper counting
$ion_start = $ion_start - 1;
$ion_stop = $ion_stop + 1;

#

    $r_time = 400;
if ($t_step > 1) {
    $r_time = $t_step + 20000/$t_step;
    if ($t_max > 1) {
        $r_time = $t_step + ($t_max - $k)/$t_step;
        #print "$r_time\n";
    }
}

#=====
# Statistical Variables

```

```

#=====
# Circular distribution by phi
#-----
@atom_tab=(0,0,0,0,0); #initial $d_idx
$d_idx = 0;
@residue_box=(0,0,0); # will grow to res_max
@residue_type=(0,0,0,0,0); # only use the last four
#=====
# channel occupancy
#-----
$sqr_occu_sum = 0;
$occu_cnt = 0;
$variance_occu =0;

#=====
# bulk occupancy statistics
#-----
# $bulk_ion_tot = $ion_stop - $ion_start + 1.0; # total bulk
ions:w
$sqr_bulk_sum = 0;
$bulk_cnt = 0; # observed ions in the bulk per frame
$bulk_total = 0; # total ions observed in bulk
$variance_bulk = 0;

$ion_high_numid = 0;
@ion_residency_cnt = (0);
@ion_high_residency_cnt = (0);

$ion_num_idx = 0; # index of ion for the following arrays
@ion_is_here_flag = (0); # flag for ion being present
@ion_is_here_cnt = (0);
@ion_is_here_high = (0); # residency near a residue
@ion_is_here_tot = (0); # total cumulative residency
$ion_is_here_dist = 3.2; # distance an ion is considered
being present
#=====
# cylinder occupancy statistics
#-----
$sqr_cylinder_sum = 0;
$cylinder_cnt = 0; # observed ions in the bulk per frame
$cylinder_total = 0; # total ions observed in bulk
$variance_cylinder = 0;

# $ion_high_numid = 0;
@ion_cylinder_residency_cnt = (0);

```

```

@ion_cylinder_high_residency_cnt = (0);

#=====
# surface topology array counter
#-----
    $zr_index = 0;
    @ion_grid_cnt = (0.0);
    $ion_grid_sz = 0; # initialize the grid size values
    $ion_grid_sz = int (($z_max*4)*($r_max/0.5));
    $ion_grid_cnt[$ion_grid_sz] = 0.0; # size it up to

#=====
# Residue by Atom Array Bins
#-----
    @atom_type_name=(0); # G:1,2-> O6,N7; A:3->N7
    @atom_type_stat=(0.0,0.0,0.0, 0.0,0.0,0.0);
    #purines ->#O6 sum_sqr, sum, n; #N7 sum_sqr, sum, n
    #pyrimidines ->#O4 sum_sqr, sum, n; #O2 sum_sqr, sum, n

#=====
#=====
# BEGIN MAIN PROGRAM
#-----

while ($j < $r_time) { # operations
#while ($j < 10) { # testing
#   $frame_cnt = $frame_cnt + 1;
#=====
# General File Operations
#-----
#   $checkfile = $filein;
#   $filepdb = $base.$k;
#   $k=1;
#   $filepdb=$base.'.'. $k; # open the file

# see if the file exists
    if (-e $filepdb) {
        $frame_cnt = $frame_cnt + 1;
        $attemptfile = 0;
        open(DATFILE, $filepdb) || die "Cannot open file $filepdb
-- $!\n";
#   open(DATFILE, $filepdb) || attemptfile;
# case where the file skips a frame or two
# attempt opening successive k files

```

```

print "Processing output file ($filepdb)... \n";
@lines = <DATFILE>;

# basic statistics on occupancy
$ions_inside = 0;
$occu_cnt = 0; # reset ion count

$atom_cnt = 0; # integer
$set_cnt = 0; # reset the frame count for each file
$atom_total=0; # reset total atoms

#=====
# Phosphorous Atomic X,Y,Z & Holder variables for the ions
# ION X,Y,Z
#-----
    $p1x=0.0;
    $p1y=0.0;
    $p1z=0.0;

    $p2x=0.0;
    $p2y=0.0;
    $p2z=0.0;

    $p3x=0.0;
    $p3y=0.0;
    $p3z=0.0;

    $p4x=0.0;
    $p4y=0.0;
    $p4z=0.0;

#check variables
    $p1c=0.0;
    $p2c=0.0;
    $p3c=0.0;
    $p4c=0.0;

    $m0x=0.0; # midpoint of cylinder
    $m0y=0.0; #
    $m0z=0.0; #

    $t0x=0.0; # end zero of the cylinder
    $t0y=0.0;

```

```

    $t0z=0.0;

    $tlx=0.0; # the other end of the cylinder
    $tly=0.0;
    $tlz=0.0;
#starts from zero indexing
#   open(MYFILE, '>'.$base.'out')|| die; # overwrite old files
#   close(MYFLIE);
#####
# BEGIN LINE PROCESSING OF A FILE #
# Assumes that the main atoms (phosphorous etc) occur before #
# the atom identifications of the ions #
#####
    @mid_cylinder = (0.0,0.0,0.0);
    @midp12 = (0.0,0.0,0.0);
    @midp34 = (0.0,0.0,0.0);
    $mid_made = 0.0; # assigned 1 after the final phosphorous
is found
    foreach $line(@lines) {

        $nlx=0.0; # ion x
        $nly=0.0; # ion y
        $nlz=0.0; # ion z
        @n1= (0.0,0.0,0.0);
        $dlc=0.0; # vertical to point X0 from the ray X21
        @rlc=(0.0,0.0,0.0,0.0); # radial distance from Xm to X0

# Search for the Phosphorus first
# Calculate the Cylinder Parameters
# Search for the Ions Next
# Calculate the line to the mid section of the cylider and
# distance from the mid section of the channel

        # throw out the first data after the "Processing Amber
trajectory"

        @parsed_values = split(/\s+/, $line);
#       $atom_numid = trim($parsed_values[1])*1.0;
        # was set to $parsed_values[3]=~ /atoms/
        if ($parsed_values[0] =~ /TER/) {
            #print "skip on TER \n";
        }
        elsif ($parsed_values[0] =~ m/ATOM/) { # ensure that
we are checking an atomic position

```

```

        $atom_numid = ($parsed_values[1])*1.0;

        if ($phos1 < 0 ) { # entering arbitrary cylinder
mode when phos1 is negative
#           print "Arbitrary Cylinder Mode \ \n"; # define
a midpoint with the remaining three phosphorous id's
                                                    #x, y, z,
the axis is along the z direction
#           print "Cylinder Center at x,y,z ->\ \ $phos2 \
\ , \ \ $phos3 \ \ , \ \ $phos4 \ \n";
        $p1x=$phos2 - $r_max;
        $p1y=$phos3 - $r_max;
        $p1z=$phos4 - $z_max;
        @p1=($p1x,$p1y,$p1z); # array
        $p1c=1.0;
        $p2x=$phos2 + $r_max;
        $p2y=$phos3 + $r_max;
        $p2z=$phos4 - $z_max;
        @p2=($p2x,$p2y,$p2z); # array
        $p2c=1.0;
        $p3x=$phos2 - $r_max;
        $p3y=$phos3 - $r_max;
        $p3z=$phos4 + $z_max;
        @p3=($p3x,$p3y,$p3z); # array
        $p3c=1.0;
        $p4x=$phos2 + $r_max;
        $p4y=$phos3 + $r_max;
        $p4z=$phos4 + $z_max;
        @p4=($p4x,$p4y,$p4z); # array
        $p4c=1.0;
        $mid_made = 1.0;
#           $atom_numid = ($parsed_values[1])*1.0;
        } else {

#=====
#
# NORMAL OPERATIONS
#-----
#
#           $atom_numid = ($parsed_values[1])*1.0;
#           print "read \ $parsed_values[1] \ \ $atom_numid \ \
$phos1 \n";
        if ($parsed_values[2]=~ m/P/ ) {
            if ($atom_numid eq $phos1) {

```

```

        $plx=($parsed_values[5])*1.0;
        $ply=($parsed_values[6])*1.0;
        $plz=($parsed_values[7])*1.0;
        @p1=($plx,$ply,$plz); # array
        $p1c = 1.0;
        print "got 1 @p1 \n";
    } elseif ($atom_numid eq $phos2) {
        $p2x=($parsed_values[5])*1.0;
        $p2y=($parsed_values[6])*1.0;
        $p2z=($parsed_values[7])*1.0;
        @p2=($p2x,$p2y,$p2z); # array
        $p2c = 1.0;
        print "got 2 @p2 \n";
    } elseif ($atom_numid eq $phos3){
        $p3x=($parsed_values[5])*1.0;
        $p3y=($parsed_values[6])*1.0;
        $p3z=($parsed_values[7])*1.0;
        @p3=($p3x,$p3y,$p3z); # array
        $p3c = 1.0;
        print "got 3 @p3 \n";
    } elseif ($atom_numid eq $phos4){
        $p4x=($parsed_values[5])*1.0;
        $p4y=($parsed_values[6])*1.0;
        $p4z=($parsed_values[7])*1.0;
        @p4=($p4x,$p4y,$p4z); # array
        $p4c = 1.0;
        $mid_made = 1.0;
        print "got 4 @p4 \n";
    } # end phos search
    #dump last count to the tally and append the
datafile
} # end of search for phosphorous

} # end encapsulation of arbitrary cylinder mode

# check if we have all the phosphorous coordinates
# to begin calculating the cylinder dimensions

#=====
# Create Midpoint and extrapolate
# axis
#-----
if ($p1c*$p2c*$p3c*$p4c*$mid_made gt 0.0) {
# call up

```

```

        @midp12=&get_midpt(@p1,@p2); # get middle of 1 2
        @midp34=&get_midpt(@p3,@p4); # get middle of 3 4
        # calculate the mid point of the channel
        @mid_cylinder=&get_midpt(@midp12,@midp34); # get
middle of middle
        print "p12 \ @midp12 \ ,p34 \ @midp34 \ ,center->
\ @mid_cylinder \n";
        @axis_u_vec = get_u_vec(@midp12,@midp34); # get
unit vector pointing parallel to the axis
        $mid_made = -1.0; # switch off once the mid point
has been calculated
#         print "made midpoint \ @u_vec \n";
    }

#=====
# Populate full atom position, ID array
# for distance and association matrices
#-----
if ($parsed_values[0] =~ m/TER/) {
    print "skip on TER data cull \n"; } # null
operation
elseif ($atom_numid < $ion_start) {
    # array indices being at zero for perl
    # record atom positions for distance
measurements
        $d_idx = $parsed_values[1]*1.0; #data index
        $d_idx_nn1 = ($d_idx - 1)*5;

        $atom_tab[$d_idx_nn1 + 0]
=( $parsed_values[5])*1.0; # x
        $atom_tab[$d_idx_nn1 + 1]
=( $parsed_values[6])*1.0; # y
        $atom_tab[$d_idx_nn1 + 2]
=( $parsed_values[7])*1.0; # z
        # record residue type, unassigned is T which
is in box zero
        if      ($parsed_values[3]=~ m/G/) {
$atom_tab[$d_idx_nn1 + 3] = 1; }
        elsif ($parsed_values[3]=~ m/A/) {
$atom_tab[$d_idx_nn1 + 3] = 2; }
        elsif ($parsed_values[3]=~ m/C/) {
$atom_tab[$d_idx_nn1 + 3] = 3; }
        elsif ($parsed_values[3]=~ m/U/) {
$atom_tab[$d_idx_nn1 + 3] = 4; }

```

```

        $atom_tab[$d_idx_nn1 + 4] =
$parsed_values[4]*1.0; # RESID Number

        $res_max= $parsed_values[4]*1.0;

        # atom type array for statistics # was
$d_idx_nn1

$atom_type_name[$d_idx]=trim($parsed_values[2]); # store the
atom type

        #      if ($atom_type_name[$d_idx_nn1]=~ m/O1P/){
        #      print " got $atom_type_name[$d_idx_nn1]
$d_idx_nn1 \n";
        #      exit;
        #      }

        # @atom_type_name=(0); # G:1,2-> O6,N7; A:3-
>N7
        # @atom_type_stat=(0.0,0.0,0.0, 0.0,0.0,0.0);
        # #purines ->#06 sum_sqr, sum, n; #N7
sum_sqr, sum, n
        # #pyrmimdines ->#04 sum_sqr, sum, n; #O2
sum_sqr, sum, n

    }

    #=====
    # Perform Analysis
    #-----
    # check if atom id is the ion id
    if ($mid_made lt 0.0) {
        if ($atom_numid > $ion_start) {
            if ($atom_numid < $ion_stop) {
                #=====
                # ions are sequential in their residue order
                #          print "ion \ $atom_numid \n";
                # Determine ION Sequence ID
                $ion_numid = $ion_numid + 1;
                #=====
                # Channel Statistics/Quantitative localizations
                #-----
                $n1x=$parsed_values[5]*1.0; # assign
positional coordinates

```

```

        $nly=$parsed_values[6]*1.0; #
        $nlz=$parsed_values[7]*1.0; #
        @n1=($nlx,$nly,$nlz); # assign
        @x_vec
=cross_prod(@n1,@mid_cylinder,@midp34);
        # get the projection of the ion ray onto
the ray from midpoint cylinder axis to P34 anchor
        $d1c = dot_prod(@n1,@mid_cylinder,@midp34);
# projection along the ray midpt->34

        @r1c = get_r_dist(@n1,@mid_cylinder); # ray
to the midpoint, magnitude is in index 3
        $dzc = get_z_dist($r1c[3],$x_vec[3]);
#
        print "@x_vec \ -> \ $dzc \n";
        if($x_vec[3] < $r_max) {
            if ($dzc < $z_max) {

                # print "ion $atom_numid \ @x_vec[3] \
$r_max -> \ $dzc \ $z_max -> \ $d1c \n";

                $ion_total = $ion_total + 1; # increment
total ion count
                $occu_cnt = $occu_cnt + 1; # increment
ion count for this frame

                # tabulate the grid count
                $zr_index =
get_zr_index($dzc,@x_vec[3],$d1c); # get the zr grid index
                $ion_grid_cnt[$zr_index] =
$ion_grid_cnt[$zr_index] + 1; # increment grid index
                # $ion_frame = $ion_frame + 1; # number
of ions within this frame

                @zr_raw = get_zr_raw($zr_index);
                # print "$zr_index -> x, y \ $zr_raw[0] \
$zr_raw[1] \ $ion_grid_cnt[$zr_index] \n";
                #
                open(MYFILE, '>>'.$base.'.'.'out')|| die;
                print MYFILE "$frame_cnt\ $atom_numid \
$x_vec[3]\ $dzc\ $d1c\ $ion_total\n";
                close (MYFILE);

                #=====
                # Calculate the phi angle relative to the

```

```

# Xprod of midCylinder->P1 and
midCylinder->mid(P3,P4)
# counter clockwise direction
# looking into the p34 direction / end
view
#
#
#
#
#
#
#
#
#(end view)
#
# direction |_| axis going to the right
p1 is zero degrees/ 0 radians
# direction |_| axis toward p2 is
180degrees/pi radians
#-----
# get the unit vector perp to the axis
and anchor point 1
@phi_vec = cross_prod(@p1, @mid_cylinder,
@midp34); #
# get the new point on the axis for the
ion perp
if ( $d1c < 0){ # make z negative
$dzc = $dzc*(-1.0);
}
@new_z_point =
move_pt(@mid_cylinder,$axis_u_vec[0],$axis_u_vec[1],$axis_u_ve
c[2],$dzc);
@new_perp_point =
move_pt(@new_z_point,$phi_vec[4],$phi_vec[5],$phi_vec[6],1.0);
# magnitude should be equal to x_vec[3];
@r_perp_ion = get_r_dist(@n1,
@new_z_point); # ray perp to the axis
#make distant point in space
@up_one_pt =
move_pt(@new_z_point,$axis_u_vec[0],$axis_u_vec[1],$axis_u_vec
[2],1.0);

```

```

        # determin angle sign
        @p1_uperp_vec = cross_prod(@up_one_pt,
@new_z_point, @new_perp_point); #points in direction of p1
anchor

        @p1_phi_point =
move_pt(@new_z_point,$p1_uperp_vec[4],$p1_uperp_vec[5],$p1_uperp_vec[6],1.0);

        $d_phi_plus_minus_val = dot_prod(@n1,
@new_z_point, @new_perp_point);
        # print "p1 phi point @p1_phi_point
\n"; # points toward p1 _|_ to axis
        # print "anchor phi point @new_z_point
\n"; # on the cylinder axis
        # print "perp point @new_perp_point
\n"; # points _|_ to axis toward negative phi
        # print "phi vec @phi_vec \n";

        # determine angle from vector pointing
along the cross of phos1 x midp34
        $d_new_perp = dot_prod(@n1,
@new_z_point, @p1_phi_point);

        $raw_ratio = $d_new_perp/$r_perp_ion[3];
        $phi_value =
acos($d_new_perp/$r_perp_ion[3]); # is an even function

        if ($d_phi_plus_minus_val > 0) {
            # print "is negative
$d_phi_plus_minus_val \n";
            $phi_value = $phi_value*(-1.0);
            # print "pi dot val $d_new_perp ,
radius val $r_perp_ion[3] \n";
        }

        #$m_verify_perp =
dot_prod(@midp34,@new_z_point,@new_perp_point);
        # $m_verify_perp =
dot_prod(@midp34,@new_z_point,@p1_phi_point);

        # print "phi is $phi_value for $raw_ratio
check perp $m_verify_perp \n\n";

```

```

# print " $midp34[0], $midp34[1],
$midp34[2] \n";

# if ($atom_numid > 726) {exit;}

# }

#=====  

# find highest occupancy to a chosen
residue
#-----  

# $ion_num_idx = $atom_numid -
$ion_start; # has already been adjusted by minus 1

for ($e_idx = 0; $e_idx < $res_B;
$e_idx++) { #clear the loop
    $ion_is_here_flag[($ion_numid -
1)*$res_B + $e_idx + 1] = 0;
}
#=====  

# find the closest atom & residue type
#-----  

$go_find_nearest = 1.0;
@this_atom = (0.0,0.0,0.0);
$last_type = 0;
$special_type = 0;
$special_dist = $ia_dist; #3.2; # max
bonding distance, was 5
$special_residue_bin = 0;
$residue_box_bin = 0;
$last_res_dist = $ia_dist; #3.2;
#$r_max; #nominal value is 5 angstroms for any meaningful
electrostatic interactions

# print "entering search for nearest -
> last residue $res_max at $r_max \n";
for ($d_idx= 0; $d_idx < $ion_start;
$d_idx++) { #atom loop
    $this_atom[0] = $atom_tab[0 +
5*$d_idx]; #x
    $this_atom[1] = $atom_tab[1 +
5*$d_idx]; #y

```

```

                    $this_atom[2] = $atom_tab[2 +
5*$d_idx]; #z
                    @nearest_residue =
get_r_dist(@n1,@this_atom); # get the smallest value greater
than zero
                    #$dummy_check = $atom_tab[3 +
5*$d_idx];
                    $dummy_check2 = 1*$atom_tab[4
+ 5*$d_idx]; # actual residue id
                    # =====
                    # occupancy near residue at
least below 4 Angstroms, was hard-coded, now in the
ion_is_here_dist variable
                    # -----
                    if ($nearest_residue[3] > 0 &&
$nearest_residue[3] < $ion_is_here_dist && $dummy_check2 > 0)
{
                    $k_idx = ($ion_numid -
1)*$res_B + $dummy_check2;
                    $ion_is_here_flag[$k_idx]
= 1; # doesn't matter how often
                    #if ($ion_numid < 2 &&
$dummy_check2 < 2 ) {
                    #print "idx $k_idx got it
for $ion_numid $res_B $dummy_check2 \n"; # okay for one loop
in a set
                    #exit;
                    #}
                    # if ($ion_numid < 2) {
                    # print "idx $k_idx for
$ion_numid didn't get it $dummy_check2 \n";
                    # exit;
                    # }
                    }
                    # =====
                    # cylinder occupancy data
                    # -----

```

```

                                #print "$d_idx, nearest is
$nearest_residue[3], $dummy_check, $dummy_check2, @this_atom
\n";
                                if ($nearest_residue[3] > 0 &&
$nearest_residue[3] < $last_res_dist && $dummy_check2 > 0) {
                                    $last_res_dist =
1.0*$nearest_residue[3] ; # take the lowest value
                                    $residue_type_bin =
1*$atom_tab[3 + 5*$d_idx]; # record the type TGACU-> 0,1,2,3,4
into the type box
                                    $residue_box_bin =
1*$atom_tab[4 + 5*$d_idx]; # record residue id number
                                    $go_find_nearest = 0.0; #
found the nearest atom within the magic distance
                                    if ($residue_box_bin < 1)
{
                                        print "read
fault\n";
                                        exit;}

                                #collect atom statistics
on the special atoms O6, N7, O4, O2
                                #print
"$atom_type_name[$d_idx] , $residue_box_bin, $d_idx \n";
                                if ($last_res_dist <
$special_dist) {
                                    # print "$special_dist
\n";
                                    $atom_nomen =
$atom_type_name[$d_idx];
                                    if
($atom_type_name[$d_idx]=~ m/O6/) {
                                        $special_type =
$special_dist =
$last_type = 1;
                                        $last_res_dist;
                                        $special_residue_bin=$residue_box_bin;
                                        print "O6 dist
$special_dist \n";
                                    }elseif
($atom_type_name[$d_idx]=~ m/N7/) {

```

```

$special_type =
$residue_type_bin;
$last_type = 2;
$special_dist =
$last_res_dist;
$special_residue_bin=$residue_box_bin;
print " N7 dist
$special_dist \n";
} elseif
($atom_type_name[$d_idx]=~ m/'/) {
print "before
$atom_type_name[$d_idx] at $last_res_dist \n";
$residue_box_bin = -1
; # don't count back bone interactions
$ion_is_here_flag[$k_idx] = -1; # take back the flag
} else {
if
($atom_type_name[$d_idx]=~ m/O4/) {
$special_type =
$residue_type_bin;
$last_type = 3;
$special_dist =
$last_res_dist;
$special_residue_bin=$residue_box_bin;
print "O4 dist
$special_dist\n";
}
elseif
($atom_type_name[$d_idx]=~ m/O2/) {
$special_type =
$residue_type_bin;
$last_type = 4;
$special_dist =
$last_res_dist;
$special_residue_bin=$residue_box_bin;
print "O2 dist
$special_dist\n";
}
elseif
($atom_type_name[$d_idx]=~ m/P/) { # is a phosphate backbone
screening

```

```

print "phosphate
$atom_type_name[$d_idx] \n";
$residue_box_bin = -
1 ; # don't count phosphate screening ions
$ion_is_here_flag[$k_idx] = -1; # take back the flag
}
else {#$last_type = 0;
print "atomic $d_idx
$atom_type_name[$d_idx] at $last_res_dist \n";
# if
($atom_type_name[$d_idx]=~ m/ /) {
# exit; }
} #don't count toward
being special
}
}
}
} # end of atom loop

#=====  

# Tabulate occupancy cumulative,  

highest, etc.  

#-----  

for ($e_idx = 0; $e_idx < $res_B;  

$e_idx++) { #residue loop
$f_idx = ($ion_numid -  

1)*$res_B + $e_idx + 1; # mapping function
if ($ion_is_here_flag[$f_idx] >
0 ) {
$ion_is_here_cnt[$f_idx] =
$ion_is_here_cnt[$f_idx] + 1;
# print " got it \n";
# exit;
} else {

```

```

                                if ($ion_is_here_cnt[$f_idx]
> $ion_is_here_high[$f_idx]) {
                                $ion_is_here_high[$f_idx]
= $ion_is_here_cnt[$f_idx];
                                #print "$ion_num_idx lost
it $ion_is_here_high[$ion_num_idx] \n";
                                #exit;
                                }

                                $ion_is_here_tot[$f_idx] =
$ion_is_here_tot[$f_idx] + $ion_is_here_cnt[$f_idx]; #
accumulate

                                $ion_is_here_cnt[$f_idx] =
0; # clears the current count
                                }

                                } # end residue loop

                                #=====
                                # atom type statistics
                                #-----

                                if ($go_find_nearest < 1.0 &
$residue_box_bin > 0.0) {
                                $ions_inside = $ions_inside + 1;
                                $residue_box[$residue_box_bin] =
$residue_box[$residue_box_bin] + 1; #increment frequency count
                                $residue_type[$residue_type_bin] =
$residue_type[$residue_type_bin] + 1; # only five types
frequency

                                #T is zero box
                                print "ion $ion_numid at $last_res_dist Ang
$atom_nomen -> res $residue_box_bin type $residue_type_bin |
$k ps \n";
                                $atom_nomen = "";

                                if ($special_residue_bin > 0)
{
                                $idx_type = 0;
                                $idx_type =
($special_residue_bin - 1)*7; # index by residue ID
                                print "$last_type is atom
type at $special_dist of $special_residue_bin of $special_type
\n";

```

```

                                if ($special_residue_bin !=
$residue_box_bin) {print "non special O6,N7,O4,O2 binding!\n";
}

                                if ($last_type < 2.0 &&
$last_type > 0.0) { # O6 G only

#$atom_type_stat=(0.0,0.0,0.0, 0.0,0.0,0.0); }
                                $atom_type_stat[$idx_type +
0] = $atom_type_stat[$idx_type + 0] +
$special_dist*$special_dist;
                                $atom_type_stat[$idx_type +
1] = $atom_type_stat[$idx_type + 1] + $special_dist;
                                $atom_type_stat[$idx_type +
2] = $atom_type_stat[$idx_type + 2] + 1;
                                $atom_type_stat[$idx_type +
6] = $special_type;

                                } elseif ($last_type < 3.0) { #
N7 A,G
                                $atom_type_stat[$idx_type +
3] = $atom_type_stat[$idx_type + 3] +
$special_dist*$special_dist;
                                $atom_type_stat[$idx_type +
4] = $atom_type_stat[$idx_type + 4] + $special_dist;
                                $atom_type_stat[$idx_type +
5] = $atom_type_stat[$idx_type + 5] + 1;
                                $atom_type_stat[$idx_type +
6] = $special_type;

                                print "put into N7 \n";
                                } elseif ($last_type < 4.0){ #
O4 type U,T
                                $atom_type_stat[$idx_type +
0] = $atom_type_stat[$idx_type + 0] +
$special_dist*$special_dist;
                                $atom_type_stat[$idx_type +
1] = $atom_type_stat[$idx_type + 1] + $special_dist;
                                $atom_type_stat[$idx_type +
2] = $atom_type_stat[$idx_type + 2] + 1;
                                $atom_type_stat[$idx_type +
6] = $special_type;

                                } elseif ($last_type < 5.0) { #
O2 type C,U,T

```

```

3] = $atom_type_stat[$idx_type + 3] +
$special_dist*$special_dist;
4] = $atom_type_stat[$idx_type + 4] + $special_dist;
5] = $atom_type_stat[$idx_type + 5] + 1;
6] = $special_type;

} else {
    print "not interested in
this atom type\n";
}
#purines ->#06 sum_sqr, sum, n;
#N7 sum_sqr, sum, n
sum, n; #02 sum_sqr, sum, n
#pyrmimdines ->#04 sum_sqr,
}
}

#=====
# z - phi tables
#-----
---
# $dzc is signed by now
$idx_z_phi = get_z_phi_idx($dzc,
$phi_value); # increment phi by 0.0872664625997 radians or 5
degrees

# composite table

$z_phi_table[$idx_z_phi]=$z_phi_table[$idx_z_phi] + 1;

# stem 1 table z_phi_table of
those near the helices
if ($residue_box_bin > 0 &&
$residue_box_bin < ($res_A + 1) ) {

$z_phi_tableA[$idx_z_phi]=$z_phi_tableA[$idx_z_phi] + 1;
} elsif ($residue_box_bin > $res_A
&& $residue_box_bin < ($res_B + 1) ) {

```

```

                                # stem 2 table z_phi_table

$z_phi_tableB[$idx_z_phi]=$z_phi_tableB[$idx_z_phi] + 1;
                                } else { print "not in the box\n";
}

                                #print "made z phi index,
$idx_z_phi for $dzc, $phi_value\n";
                                #@dummy_z_phi =
get_z_phi_raw($idx_z_phi);

                                #print "raw return z-
>,$dummy_z_phi[0] ,phi-> $dummy_z_phi[1], cnt->
$z_phi_table[$idx_z_phi] \n";
                                #exit;
                                #-----
---
                                # end of z-phi tabulations
                                #-----
---
                                #-----

                                } # greater than z_max
                                } # greater than r_max

#=====
# bulk statistics
#-----
# remains in the annular region
    if ($x_vec[3] > $br_min && $x_vec[3] < 40)
{ # was greater than
    $bulk_cnt = $bulk_cnt + 1;
    $bulk_total = $bulk_total + 1;
    $ion_residency_cnt[$ion_numid] =
$ion_residency_cnt[$ion_numid] + 1;

    } elsif ($dzc > $bz_min && $dzc < 47) {
    $bulk_cnt = $bulk_cnt + 1;
    $bulk_total = $bulk_total + 1;
    $ion_residency_cnt[$ion_numid] =
$ion_residency_cnt[$ion_numid] + 1;

    }
#=====
# case that is in the RNA zone
# both radial distance and axial line are

```

```

# less than the boxed min
# the bulk ion count is updated entering the
cylinder
#-----
if ($x_vec[3] < $br_min || $x_vec[3] > 40)
{ # was less than
    if ($dzc < $bz_min || $dzc > 47 ) {
        if ($ion_residency_cnt[$ion_numid] >
$ion_high_residency_cnt[$ion_numid]) {
            $ion_high_residency_cnt[$ion_numid] =
$ion_residency_cnt[$ion_numid];
        }
        $ion_residency_cnt[$ion_numid] = 0; #
clear the residency count
    }

    }# elsif ($x_vec[3] > $br_min || $dzc >
$bz_min) {#travels farther out from the annular region

        # if ($ion_residency_cnt[$ion_numid] >
$ion_high_residency_cnt[$ion_numid]) {
            # $ion_high_residency_cnt[$ion_numid] =
$ion_residency_cnt[$ion_numid];
            # }
            # $ion_residency_cnt[$ion_numid] = 0; #
clear the residency count
            # }

# }
#=====
#-----
# end bulk statistics
#-----

#=====
# cylinder statistics
#-----
# when leaving the cylinder, the occupancy
# count is moved to the the high counter
array

if ($x_vec[3] > $br_min ) { # was greater
than

```

```

        if
($ion_cylinder_residency_cnt[$ion_numid] >
$ion_cylinder_high_residency_cnt[$ion_numid]) {

$ion_cylinder_high_residency_cnt[$ion_numid] =
$ion_cylinder_residency_cnt[$ion_numid];
        }

$ion_cylinder_residency_cnt[$ion_numid] = 0; # clear the
residency count

        } elsif ($dzc > $bz_min ) { # a larger
cylinder of +/- 47 Angstroms

        if
($ion_cylinder_residency_cnt[$ion_numid] >
$ion_cylinder_high_residency_cnt[$ion_numid]) {

$ion_cylinder_high_residency_cnt[$ion_numid] =
$ion_cylinder_residency_cnt[$ion_numid];
        }

$ion_cylinder_residency_cnt[$ion_numid] = 0; # clear the
residency count

        }
#=====
# case that it is in the RNA zone
# both radial distance and axial line are
# less than the boxed min, occupancy is
updated
#-----
        if ($x_vec[3] < $br_min) { # both cases
must be satisfied for the ion to be in the cylinder

                if ($dzc < $bz_min) {
                $cylinder_cnt = $cylinder_cnt + 1;
                $cylinder_total = $cylinder_total + 1;
                $ion_cylinder_residency_cnt[$ion_numid]
= $ion_cylinder_residency_cnt[$ion_numid] + 1;
                }
        }
}

```

```

#=====
#-----
# end cylinder statistics
#-----
    if ($cylinder_cnt != $occu_cnt) {
        print "$x_vec[3] \ \ $dzc \ \ $z_max \
\ $bz_min \n";

        print "count inequality \n";
        print "bulk cnt \ \ $bulk_cnt \n";
        print "occu cnt \ \ $occu_cnt \n";
        print "cylinder_cnt \ \ $cylinder_cnt
\n";

        exit;
    }

} # atom ion inner functions
} # atom ion functions
} # middle made

} # end of search for atoms in a file
} # end of the foreach lines in a file
    $k=$k + $t_step; # increment time is 50ps, may change
this to a command line variable
    $j=$j + 1;

#clear ion sequence identifier
$ion_high_numid = $ion_numid;
$ion_numid = 0;

#appending the output file with the results
#    open(MYFILE, '>>'.$base.'out')|| die;
#    print MYFILE "$frame_cnt\ $atom_cnt\ $atom_total\
$set_cnt\n";
#    close (MYFILE);

    print "ions in channel \ \ $occu_cnt \n";

# add key statistical data
    $sqr_occu_sum = $sqr_occu_sum + $occu_cnt*$occu_cnt; #
running sum of squares
    $occu_cnt = 0; # clear ion count

# bulk statistics
    print "ions in bulk \ \ $bulk_cnt \n";

```

```

        $sqr_bulk_sum = $sqr_bulk_sum + $bulk_cnt*$bulk_cnt; #
running sum of squares
        $bulk_cnt = 0; # clear bulk ion count

# cylinder statistics
        print "ions in cylinder \ $cylinder_cnt \n";
        $sqr_cylinder_sum = $sqr_cylinder_sum +
$cylinder_cnt*$cylinder_cnt;
        $cylinder_cnt = 0;

# non screening ions
        print "ions inside \ $ions_inside \n";
        $ions_inside = 0;

        close(DATFILE);

    } else {

        if ($attemptfile < 99) {
#            $attemptfile = $attemptfile + 1;
            print "attempting \ $filepdb \ $attemptfile \n";
#
# Attempts are made until the terminal loop is reached
#
            if ($attemptfile < 1) {
                $k=$k - $t_step; # increment by one to see if it
is going up
                $k=$k + 1;
            } else {
                $k=$k + 1;
            }
            $attemptfile = $attemptfile + 1;

        } else {
            report_stats;
            exit;
        }
    }
} # end of file series processing and
# if it hasn't exited will exit now
report_stats;

```

2. ANAL script to calculation ion interaction energies over a trajectory from generated PDB snapshots.

```

#!/bin/csh -f

#BSUB -R em64t
#BSUB -o log.%J
#BSUB -e error.%J
#BSUB -n 2
#BSUB -J i17r3
#BSUB -W 109:00
#BSUB -q yingling

source /usr/local/apps/env/intel.csh
source /home/lsethap/.alias
source /home/lsethap/.cshrc

#rmkdir -rf i17r3/
#mkdir i17r3/

# user values
set myion = "i17r3"
set start_time = 7327
set stop_time = 8512
set step_val = 5
set ion_number = 17
set ion_res_number = 39
set res_number = 3
set mydata = "$myion" # folder with all the inpcrd and prmtop files
set ion_res = "$mydata/$myion"
#===== prmtop coord generator =====
set script_file = "leap_.$myion.scr"
set time_start = $start_time
set time_stop = $stop_time
set qtidx = $time_start
set file_src = "snap_5ps/polyA_snapshot"
set file_dest = $mydata

#output file name
set mydatfile = "$ion_res.out"

rmdir -rf $myion/
mkdir $myion/

#===== run snapshots =====
# optional if snapshots have not been run
#cat <<eof> i2r8_snapshot.in
#trajin /gpfs_yingling/ygyingli/abhishek/poly/polyA/polyA_md1.x 12990
14080 10
#
#center :1-22
#image

##strip :WAT
##strip :Cl-
#solvent byres :WAT
#closestwater 400 :6

```

```

##grid polyA.xplor 100 0.25 100 0.25 100 0.25 :Na+ max 0.9

#trajout $file_src pdb
#go
#
#eof

#ptraj /gpfs_yingling/ygyingli/abhishek/poly/polyA/polyA.prmtop <
i2r8_snapshot.in > i2r8_snapshot.out

#sleep 30

#===== run leap =====
echo "" > $script_file
while ($qtidx <= $time_stop)

echo "a = loadpdb $file_src.$qtidx" >> $script_file
echo "saveamberparm a $file_dest/$qtidx.prmtop $file_dest/$qtidx.inpcrd"
>> $script_file

echo "$qtidx"
@ qtidx += $step_val

end

echo "quit" >> $script_file

tleap -f leaprc.ff99SB -f $script_file

sleep 30

#=====
# begin program
set tidx = $start_time
#=====
# input is as format is as follows:
# c.f. Amber8 manual for description
# line (1) Title |FORMAT(20A4)
# line (2) NTX , NTXO , NRC , NGRPX , KFORM, |FORMAT(6I)
# line (3) NTB, BOX(1) , BOX(2) , BOX(3) , BETA |FORMAT(I,4F)
# line (4) NTF , NTID , NTN , NTN B , NSNB , IDIEL, |FORMAT(6I)
# line (5) CUT , SCNB , SCEE , DIELC, |FORMAT(4F)
# line (6) IMAX , EMAX(I), I = 1..9 |FORMAT(I,9F)
# line (7) IOPT , |FORMAT(A)
# .... groups
# line (penultimate) END
# line (last) STOP
#=====
cat > $ion_res.analin <<EOF
test of anal energies, compare to those from sander
  1  0  0  0  51  1
  0 0.0 0.0 0.0 0.0
  1  0  0  0  80  1
99.0 2.0 1.2 1.0

```

```

    1  2.0  2.0  2.0  2.0  2.0  2.0  2.0  2.0  2.0
ENERGY
ion $ion_number
RES $ion_res_number
END
base $res_number
RES $res_number
END
END
STOP
EOF

#set mydatfile = "$ion_res.out"

echo "ion $ion_number at residue $res_number Edecomp $start_time to
$stop_time" > $mydatfile

while ($tidx <= $stop_time)
# command of execution
#   usage: anal [-O] -i analin -o analout -p prmtop -c inpcrd -ref refc
#           -r rmscrd -s compac -p1 pdb1 -p2 pdb2 -z zmat

set mytopology = "$mydata/$tidx.prmtop"
set mycoord = "$mydata/$tidx.inpcrd"
set myoutfile = "$ion_res.$tidx.out"

anal -O -i $ion_res.analin -p $mytopology -c $mycoord -ref $mycoord -o
$myoutfile || goto error

sleep 30
echo "finished $tidx"
echo "----- $tidx -----" >> $mydatfile
echo "  type  groups      description          dist          vdwnb
eelnb" >> $mydatfile
grep "NBOND  2-  1" $myoutfile >> $mydatfile
# grep -B 1 -A 9... one before 9 after match
grep -A 8 "MATRIX" $myoutfile >> $mydatfile
echo "----- $tidx -----" >> $mydatfile
echo "-----" >> $mydatfile
# remove outfile for this time
rm -f $myoutfile

# snapshots are spaced ten "frames" from each other in this instance

    @ tidx += $step_val

end

#remove input file
rm -f $ion_res.analin
rm -f $file_dest/*.prmtop

```

```

rm -f $file_dest/*.inpcrd

# if snapshots were generated from the trajectories, remove them from the
data folder
rm -f $file_dest/*snapshot.*

grep -A 7 "TOTAL INTERACTION ENERGY" $mydatfile | awk '{print $2}' | grep
 "-" > $ion_res.Total_E.dat
grep -A 7 "ELECTROSTATIC (N-B + 1-4) INTERACTION ENERGY MATRIX" $mydatfile
 | awk '{print $2}' | grep "-" | egrep -v 'B' > $ion_res.EEL.dat
grep -A 7 "VDW (N-B + 1-4) INTERACTION ENERGY MATRIX" $mydatfile | awk
 '{print $2}' | egrep -v "B" | grep ".." | egrep -v "0.000" >
 $ion_res.vdw.dat

echo "total E stats" > $ion_res.stats
calc_stats $ion_res.Total_E.dat >> $ion_res.stats
echo "Electrostatic stats" >> $ion_res.stats
calc_stats $ion_res.EEL.dat >> $ion_res.stats
echo "VDW stats" >> $ion_res.stats
calc_stats $ion_res.vdw.dat >> $ion_res.stats

exit(0)

error:
echo " ${0}: Program error"
rm -f analin
exit(1)
#=====

```