NUCLEIC ACID STRUCTURAL RECOGNITION BY TRANSCRIPTION FACTOR AMRZ AND RIBOENDONUCLEASE RNASE HI

By

Ryan R. Davis

A Thesis Submitted to the Graduate Faculty of

WAKE FOREST UNIVERSITY GRADUATE SCHOOL OF ARTS AND SCIENCES

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Department of Biochemistry and Molecular Biology

May 2015

Winston-Salem, North Carolina

Approved By:

Thomas Hollis, Ph.D., Advisor

Rajendar Deora, Ph.D., Chair

Leslie B. Poole, Ph.D.

Acknowledgements

I am exceptionally grateful for Dr. Tom Hollis allowing me to join his lab as well as providing me with the knowledge, skillset, and support to be successful in graduate school. His door was always open and he was always willing to give advice and push me in the right direction when I needed it.

I would also like to acknowledge and thank Dr. Leslie Poole and Dr. Rajendar Deora for being a part of my committee and providing input on my research and projects to help me become a better scientist.

Chris Mauney, you helped me learn and understand various techniques, instruments, and procedures needed to become a biochemist and I am very thankful for all that you have done in training me during my time in the Master's program. Brittney Dodge, thank you for all your advice and assistance during our time together. I always knew when I had to go into the lab late at night you would be there. You made it much easier for my wife and me to move to North Carolina and I am very fortunate that we both chose the same lab.

Going to graduate school was a huge commitment and I am very appreciative of my family, especially my parents, for everything they have done to get me this far. They have always been there and given me the support I needed to overcome any obstacle. They may not have understood what I was talking about but they were always willing to listen and provide help in any way they could.

Finally, I cannot fully express my appreciation for my wife, Shayna. We often had conflicting schedules and my time in the lab could be long and sporadic but your understanding and patience made it much easier. You have always supported my decisions and you were willing to leave your friends and family so that I could pursue my education. You believed in me, even when I did not, and encouraged me to keep going; that is something I will never forget.

Table of Contents

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations

AGS	Aicardi-Goutières syndrome
AmrZ	Alginate and motility regulator Z
c-di-GMP	Cyclic-di-guanosine monophosphate
CAP	Catabolite activator protein
CF	Cystic fibrosis
CFTR	Cystic fibrosis transmembrane conductance regulator
DNA	Deoxyribonucleic acid
dsDNA	Double stranded deoxyribonucleic acid
dsRNA	Double stranded ribonucleic acid
GDP	Guanosine diphosphate
HBD	Hybrid binding domain
HIV	Human immunodeficiency virus
HTH	Helix-turn-helix
IHF	Integration host factor
NMR	Nuclear magnetic resonance
PCNA	Proliferating cell nuclear antigen
PDB	Protein Data Bank
PPT	Polypurine tract
RHH	Ribbon-helix-helix
RNA	Ribonucleic acid
RNAP	Ribonucleic acid polymerase
RNase	Ribonuclease
RIP	Regulated intramembrane proteolysis

Abstract

DNA regulation is a fundamental part of every cell's life cycle. The expression of DNA allows cells to respond, adapt, and live in the ever changing environment and even overcome host cell defenses. This regulation is controlled through DNA-protein binding. Proteins bind to an innumerable amount of sequences, with different mechanisms and have various effects on the DNA.

Sometimes ribonucleotides are found in a duplex with deoxyribonucleotides and need to be removed in order to maintain DNA stability. The insertion of ribonucleotides can be incidental, such as misincorporation by DNA polymerase, or intentional, from Okazaki fragments, depending on the role they served. RNase H enzymes ensure that ribonucleotides are cleaved from DNA. Before cleavage can occur, the RNase H first has to locate where the ribonucleotides are integrated within the DNA strand. RNases H seem to recognize the presence of ribonucleotides by the DNA adopting the A-form conformation, a greater bend, and the presence of a 2'-OH group on ribose while the other strand contains deoxyribose to sustain flexibility in one of the strands. Upon binding, the DNA near the scissile bond conforms to B-DNA characteristics and the major groove widens while the minor groove is narrowed and the DNA bend is decreased to a more linear angle. This evidence demonstrates that local DNA shape is crucial for RNase H recognition.

Transcription factors are the proteins responsible for activating or repressing gene expression. How transcription factors find their high affinity binding sequence and their mechanism of action are still under investigation. Energetically, it is favorable for

transcription factors to bind to DNA and slide until they find their consensus sequence. One transcription factor, AmrZ, functions as both an activator and repressor and studying its binding mechanism could provide more information on the roles of transcription factor regulation. When bound as a repressor to its own gene, *amrZ*, AmrZ binds as a dimer-of-dimers and may interact distally with itself while the DNA is bent to sterically hinder RNA polymerase from binding. As an activator of *algD*, AmrZ seems to bind as a dimer and may favorably interact with RNA polymerase to ensure transcription initiation is completed. The crystal structure of AmrZ bound as a repressor has already been solved and data has been collected from crystals of AmrZ bound as an activator; however, the data with AmrZ bound to *algD* was twinned and provided difficulty in determining the structure. Untwinned crystals of AmrZ bound as an activator are needed in order to definitively determine the binding mechanism of AmrZ as an activator.

Chapter 1:

Introduction to nucleic acid structure recognition

1.1 Characteristics of Nucleic Acids

1.1.1 Deoxyribonucleic Acid

Deoxyribonucleic acid (DNA) is found inside every single cell and controls all aspects of cell viability and survivability. DNA binding proteins allow the DNA to be managed and maintained to ensure the cell responds quickly, but appropriately, to the environment. The structure of unbound DNA also has huge roles in determining which proteins bind and when. DNA provides all the information needed to create a fully developed organism. The fundamental components of DNA allow it to take a shape that is generally consistent across all species. These characteristics are needed for various roles of DNA through maintenance, expression, storage, and replication. DNA is composed of four bases (adenine, guanine, cytosine, and thymine), a deoxyribose sugar, and a phosphate backbone [1]. The bases do not always bond in a planar fashion and can have different orientations and conformations, especially upon protein binding. Many of the conformational forms are shown in Figure 1-1 and will be referred to throughout this thesis. The changes in conformation can also be due to the specific conformation of DNA. There are three main forms that DNA can take: A-, B-, and Z-form DNA (Figure 1-2). However, two more types of DNA have been found but they are not as common as the three aforementioned: W-DNA and TA-DNA [3].

B-DNA is found most commonly in aqueous environments. It is claimed that most double helices exist in the B-form, but there is not currently any strong evidence to support this statement [4]. B-DNA has the following typical characteristics: C2’-endo

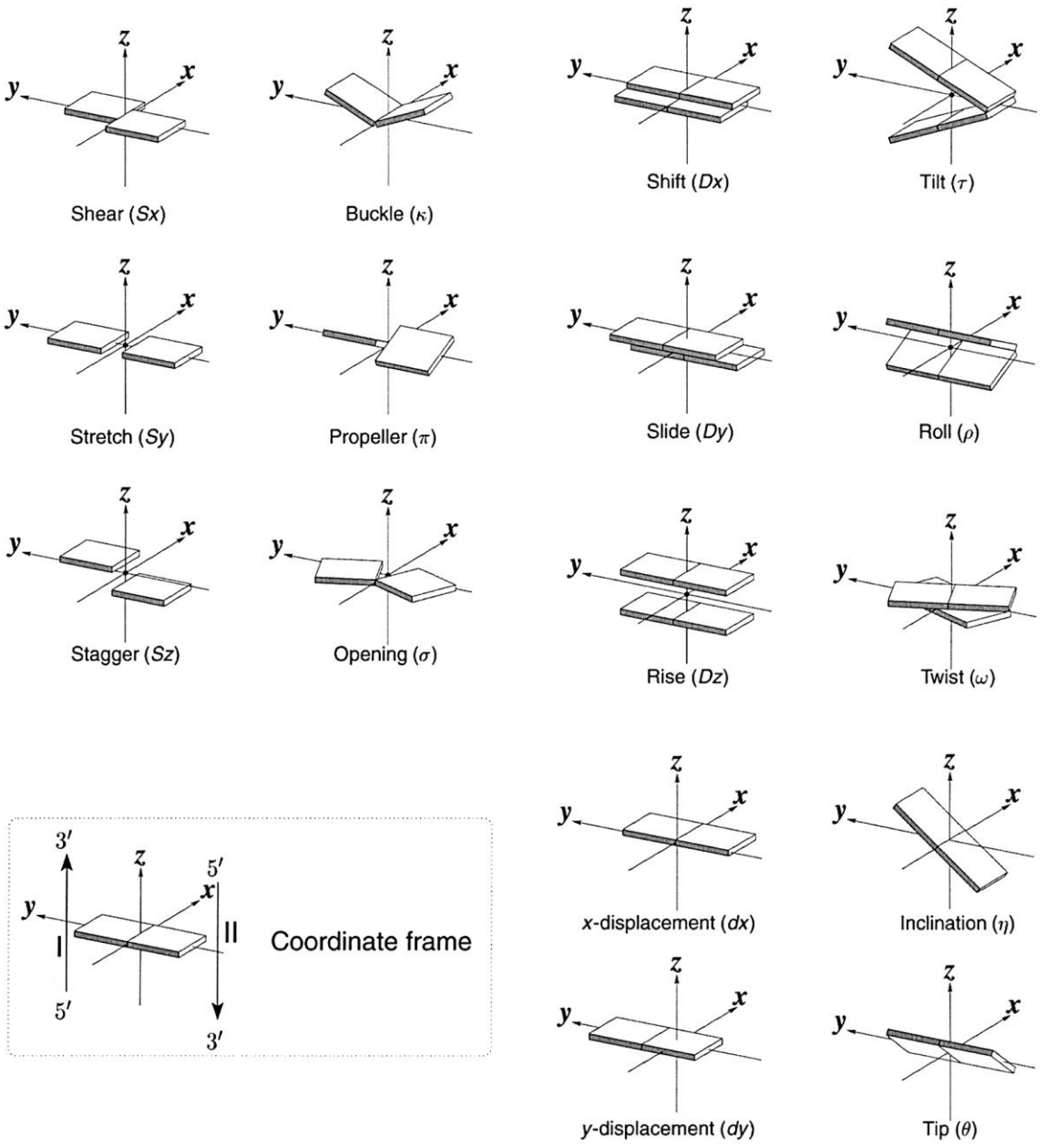

Figure 1-1: **DNA base pair geometry.** DNA base geometry and conformations that the base pairs can adopt. The coordinate frame shows the 5' to 3' direction for the base as well as the x,y,z directions. (Picture used from Lu and Olson [2] with permission).

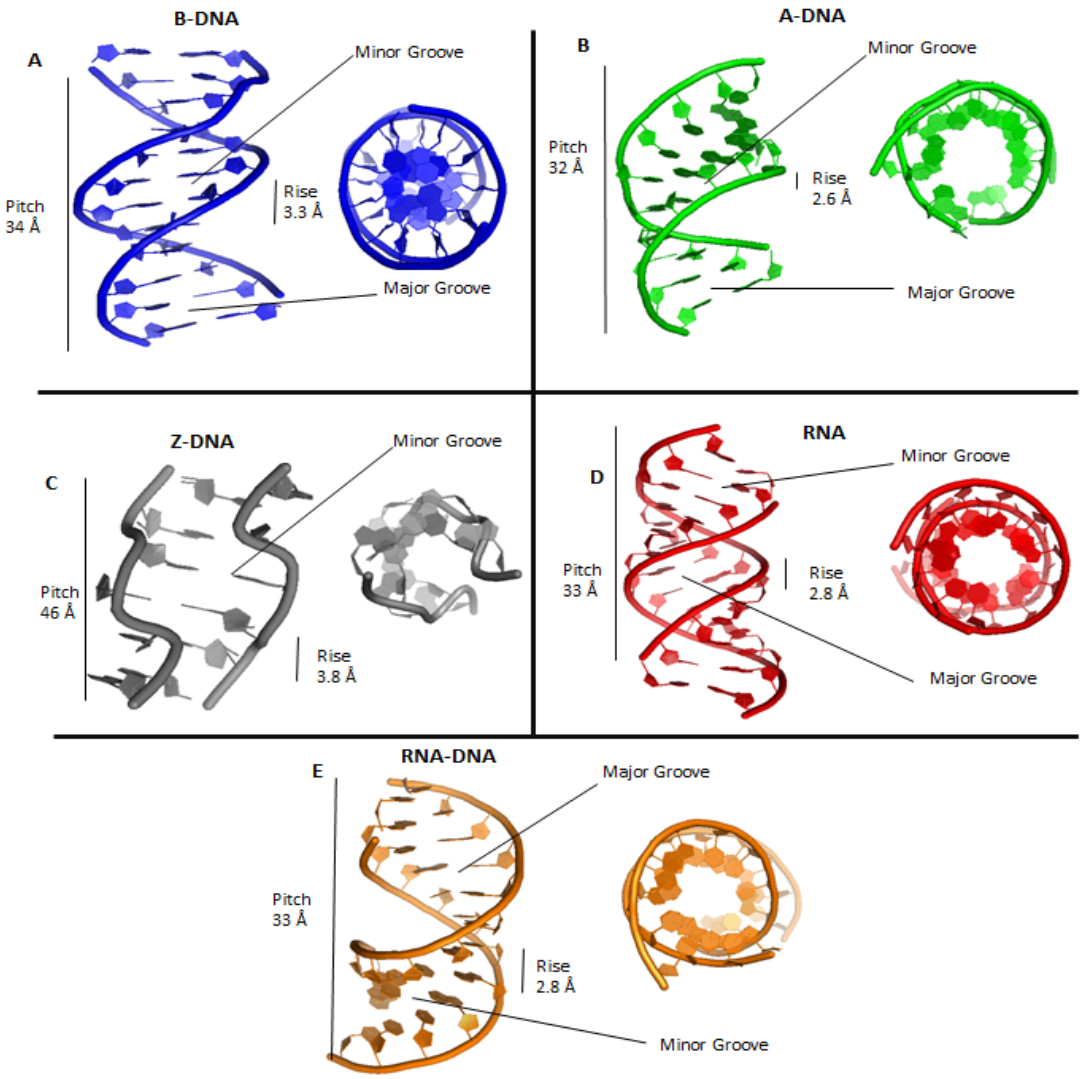

Figure 1-2: **Nucleic acid comparisons.** A) B-DNA (1BNA). B) A-DNA (137D). C) Part of Z-DNA (4OCB). D) RNA (1RNA). E) RNA-DNA hybrid (4WKJ).

---

sugar conformation, bases almost perpendicular to the axis (-7° tilt), 3.34 Å rise per residue, 34 Å pitch, right-handed helix, ten base pairs per turn, wide major groove, narrow minor groove, and no hole down the axis of the helix (Figure 1-2A) [4-7]. Crystal structures of DNA packing show minor groove interactions as well as end-to-end stacking, which has the appearance that the DNA duplexes are continuous [5,6].

A-DNA is formed in dehydrated environments and can also take this conformation when bound to various proteins [7]. DNA that is rich in GC sequences can take the A-form due to the extra hydrogen bond present in GC base pairing, which is lacking in AT base pairing, that allows for a more planar base pair thus promoting base pair sliding, common in A-DNA [6]. The following characteristics are common in A-DNA: C3'-endo sugar pucker, base pairs removed from axis (tilt 19°) leading to a hole in the center when looking down the axis, 2.56 Å rise per residue, wide minor groove, narrow major groove, 11 base pairs per turn, 32 Å pitch, and end-to-end crystal packing (Figure 1-2B) [4,5,7].

Z-DNA is the most uncommon form of DNA out of the three major conformations and exists under high salt conditions. Z-DNA appears to have a zigzagging backbone because of the alternating syn-anti glycosyl bonds with oscillating purines and pyrimidines (Figure 1-2C) [3]. Z-DNA, interestingly, forms a left-handed helix and does not have a pronounced major groove but rather has a negative electrostatic potential that is uniform down the helix through the deep, narrow minor groove [7].

W-DNA is the least common form of DNA out of all the DNA conformations mentioned. The backbone is reversed, compared to Z-DNA, which forces guanine to take a syn conformation and deoxyribose to have a C3'-endo sugar pucker at cytosines and C2'-endo at guanines [3]. Like Z-DNA, the minor groove is deep but narrow and the major groove is very shallow [3].

Finally, TA-DNA is primarily present in TATA boxes. The main difference between TA-DNA and A-DNA is that TA-DNA has an increased inclination of the bases

near 50° [7]. The positive roll of the bases allows the DNA to open easier when proteins bind making DNA melting more energetically favorable to sigma factors (σ).

1.1.2 Ribonucleic Acid

Ribonucleic acid (RNA) is the product transcribed from DNA and contains all the information necessary to code for the translation of proteins. Like DNA, RNA contains four bases (uracil instead of thymine), a phosphate backbone, and a ribose sugar instead of a deoxyribose sugar. The presence of the extra hydroxyl group on the C2' position on ribose increases the rigidity of the helix and forces RNA to solely take the A-form [5,8]. RNA has a pitch of 33.5 Å, which is similar to A-DNA, 2.8 Å rise per residue, right-handed, wide minor groove, narrow major groove, C3'-endo sugar puckers, and an incline of 16.1° (Figure 1-2D) [9].

1.1.3 RNA-DNA Hybrids

RNA-DNA hybrids are fairly common in biology. They can exist as a single RNA nucleotide inserted into a DNA double helix due to misincorporation by a DNA polymerase or extended strands from Okazaki fragments formed during DNA replication and transcription [5,8-11]. Hybrids are also formed in viral reverse transcription, utilized by HIV [8,11]. Typically, RNA is removed from the DNA duplex by ribonucleases, such as the RNase H family. As expected, due to the rigidity of the ribose sugar and the flexibility of the DNA strand, the RNA-DNA duplex primarily takes A-form

characteristics. These include: wide minor groove (although it is slightly narrower than RNA), narrow major groove, C3'-endo on the RNA strand with intermediate forms of C3'-endo and C2'-endo on the DNA strand, 2.8 Å rise per residue, 7.9° incline of the bases, 32.5 Å pitch, 11.6 bases per turn, and end-to-end crystal packing as well as abutting/minor groove interactions (Figure 1-2E) [5,8,9,12-14].

1.2 DNA Recognition and Protein Interactions

Protein-DNA interactions are not fully understood but they are a fundamental part of molecular biology. Proteins bind to DNA for activation, repression, enhancement, repair, replication, expression, and for assisting other molecular processes. The binding sites for transcription factors are called transcription factor binding sites, DNA binding motifs, or *cis*-regulatory sequences and are normally very short and highly degenerative. Because of how short the sequences typically are, it is expected that these *cis*-regulatory sequences will randomly occur approximately every few hundred base pairs, making transcription factor binding site predictions very difficult. In the cell, the concentration of DNA is typically very high which leads to transcription factors and DNA binding proteins almost always bound to DNA at any given time, regardless if there are any high-affinity sites present [15]. But the question arises, how do transcription factors find their specific, high-affinity sites when the DNA can be at a concentration of 100 g/L and consists of mostly non-specific interactions for that particular protein? The answer seems to be that proteins attach to the DNA in a non-specific manner and then slide along the

DNA until they reach their target sequence [16]. This process is driven by electrostatic interactions.

There are two types of DNA-protein interactions: direct (base) readout and indirect (shape) readout and both are required for transcription factor binding [16-18]. Base readout is the specific interaction between amino acid and nucleobase or even interactions with the DNA sequence through water-mediated contacts [19,20]. Shape readout is the recognition of DNA by the protein through various contours of the helix [20]. Shape readout can be classified into two groups, local and global. Local shape readout is the deviation of the helix from B-form such as bending in the major groove or minor groove and global shape readout is the change in the shape of a large part of the helix like backbone bending, or curvature/deformity of the helix [7]. Electrostatic interactions are part of the shape readout category and change based on properties of the major and minor grooves. A narrow minor groove has a larger negative electrostatic potential because the distance between the negatively charged phosphate backbones are closer together [7,16,17]. The minor groove distance is sequence dependent; a higher GC content widens the minor groove and A-tracts narrow the minor groove [16,21]. A-tracts are found in 82% of minor grooves, excluding TpA steps, and cause slight bending in the DNA which promotes twisting and interbase pair hydrogen bond formation in the major groove [21]. The minor grooves in B-DNA are narrow enough that they possess nearly uniform negative electrostatic potential which attracts positively charged amino acids, arginine and lysine. It was found that 28% of the residues inserted into a non-narrow minor groove are arginine but arginine makes up 60% of the residues inserted into a narrow minor groove [21]. Arginine is more prevalent than lysine because the process is

energetically more expensive to remove the amino group of lysine from water than it is to remove arginine's guanidinium group [21].

Because the minor groove is so narrow that it appears to be uniform negative charge, only the major groove of B-DNA can have specific interactions because hydrogen donors and acceptors can contribute individually to the specific recognition for that base [22]. The edges of the bases are exposed in the major groove and gives a specific pattern that a protein can recognize, known as the chemical signature, and bind to as it is sliding along the DNA looking for a high-affinity site [23]. Some amino acids have specific bases with which they preferentially interact. Lysine and arginine have a stronger interaction with guanine compared to adenine. Glutamine and asparagine almost solely interact with adenine, and aspartic acid and glutamic acid interact primarily with cytosine [22,24]. Hydrophobic interactions can occur with the aromatic rings of thymine and cytosine as well as the methyl group on thymine. Alanine and isoleucine are the main residues that have hydrophobic interactions and they only occur with thymine but tyrosine and methionine also interact with thymine occasionally [22]. Hoogsteen base pairing with these amino acids is also possible and may contribute to additional binding specificities [7].

There are common structural motifs found on DNA binding proteins that serve to interact with the nucleobases. Typically, the protein domains are globular and they contain an α-helix or β-sheet that is inserted into the major or minor groove [7,23]. Each binding domain interacts with an average of four to ten base pairs using approximately 24 amino acids [7]. Many interactions occur in the major groove but by kinking the DNA, the bases can unstack and this allows secondary structures to be inserted into the minor

groove [23]. One hydrogen bond between a base and a single residue is not enough to create specificity. Bidentate interactions can happen when two hydrogen bonds are formed from one amino acid with different donors and bifurcate interactions occur when two hydrogen bonds share the same donor [7,23].

These characteristics of protein-DNA interactions are just an introduction to how proteins can bind, change, and find specific sites on the DNA helix. By using cooperativity, transcription factors and other DNA binding proteins, such as RNase H and AmrZ, can assemble to create a higher order complex that can regulate DNA replication, repair and gene expression [7].

1.2.1 Riboendonucleases, RNase H1 and H2

Nucleases are enzymes that have the role of cleaving the phosphodiester bond formed by nucleic acids. Endonucleases cut internal phosphodiester bonds and are commonly used in repair mechanisms to replace damaged nucleotides. Ribonucleases are used to cleave ribonucleotides at the phosphodiester backbone in RNA or RNA-DNA hybrids. Therefore, riboendonucleases cleave ribonucleotides on the internal phosphodiester bond, which is commonly performed during replication. Ribonuclease H (RNase H) is a class of ribonucleases that are able to remove RNA from DNA strands by hydrolysis [25,26]. RNase H is commonly found in most species, both prokaryotic and eukaryotic, and was determined to be crucial in maintenance and development of the organism. Ribonucleotides are detrimental and mutagenic if they are not removed from DNA [27]. There are three main types of RNases H, but they are given different

nomenclature based on the species. RNase H1 in eukaryotes and RNase HI in prokaryotes require at least four ribonucleotides in a row for cleavage the occur [25,28,29]. RNase H2 in eukaryotes and RNase HII in prokaryotes is able to cleave individual ribonucleotides from DNA [25,29]. RNase H3 or RNase HIII is able to cleave double stranded RNA ribonucleotides and is structurally similar to RNase H2 but functions like RNase H1 [30,31].

RNase H1

RNase H1 is able to bind to dsDNA, dsRNA, and RNA-DNA hybrids but has the greatest affinity for hybrid duplexes by over 25-fold, followed by dsRNA because of the presence of the extra 2'-hydroxyl group found on ribose sugars [25,26,32]. Binding of the duplex occurs because RNase H1 recognizes the A-form that is present due to the ribonucleotides forcing the DNA to adopt the A-form conformation. The ability for the DNA to take on B-form characteristics after binding to the enzyme maximizes the number of interactions with the RNase H active site to the backbone and the minor groove of the hybrid which aids in cleavage specificity and efficiency [9]. The hybrid duplex is positioned between two grooves in RNase H1 and the minor groove spans the ridge between these grooves [28]. Surface complementarity and electrostatic interactions seem to be the primary forces driving RNase H1 binding to RNA-DNA [28,32].

RNase H2

RNase H2 commonly interacts with proliferating cell nuclear antigen (PCNA), which provides evidence to its role in DNA replication [25,33]. They are able to remove a single ribonucleotide, which is likely to be accidentally incorporated by a polymerase. Studies have shown that 10,000 misincorporations can occur in yeast in only a single cycle of replication [34,35]. RNase H2 preferentially cleaves ribonucleotides on the 5' end. The nuclease contains three domains: A, B, and C. RNase H2A is the catalytic domain and the B and C domains are likely auxillary components and are thought to assist the catalytic subunit [29,33]. Aicardi-Goutières syndrome (AGS) occurs when there are mutations in any of the three subunits of RNase H2 and consequently the nervous system is affected [33]. These mutations lead to an increase in the quantity of ribonucleotides present in genomic DNA which can cause DNA damage and strand breaks [25,27]. The A domain is positively charged in the active site which complements the overall negatively charged phosphate backbone on the double stranded duplex [35]. Like RNase H1, RNase H2 recognizes hybrid duplexes by interacting with both of the RNA and DNA chain.

1.2.2 Transcription Factor, AmrZ

Transcription factors are DNA binding proteins that bind to regulatory parts of the genomic DNA in order to activate or repress gene expression. These proteins allow for controlled transcription under specific environmental cues and allow a cell to respond. One example of activation can occur by the following: activation can be caused by

recruiting other factors or polymerases to the promoter region upon binding of the DNA binding protein. Repression, however, can occur by transcription factor binding at the same location that a sigma factor would bind, blocking polymerase binding and repressing gene expression. There are many types of motifs transcription factors commonly contain that aid in DNA binding and recognition. A common motif is the ribbon-helix-helix (RHH). RHH is a superfamily that has a β-sheet followed by two α-helices. RHH members typically exist in a homodimerized form [36]. The β-sheet is inserted into the major groove of the DNA in an anti-parallel fashion and one α-helix is normally responsible for dimerization through hydrophobic residues that interact with the α-helix of another RHH protein [36]. In many RHH proteins, an arginine is present in the β-sheet and the guanidinium side group interacts with the nucleotides.

Alginate and motility regulator Z (AmrZ) is an RHH transcription factor found in *Pseudomonas aeruginosa* that activates *algD* expression and autorepresses its own transcription. AmrZ binds as a dimer-of-dimers at its own gene regulatory sites (*amrZ1* and *amrZ2*) and it is believed that a DNA bending protein, integration host factor (IHF), binds and bends the DNA so that the C-terminal ends of AmrZ interact and cause the DNA to sterically hinder the polymerase from binding, resulting in its own repression [37,38]. It is also believed that AmrZ binds as a dimer to the *algD* promoter but acts to recruit other transcription factors and favorably interact with RNA polymerase to promote expression. The *algD* operon hosts 12 genes responsible for alginate production in *P. aeruginosa* which contributes to biofilm formation; this additional mucus build-up in the lungs of cystic fibrosis patients is the leading cause of death [39, 40]. This dual role of AmrZ makes it a very unique and interesting transcription factor to investigate.

1.3 Statement of Purpose

In the past decade, the scientific community has learned a great deal about the structure and function of RNases H and the transcriptional regulation of AmrZ. However, there is still much left to be discovered before a complete understanding is obtained. This thesis aims to answer the following questions in a structural manner:

1) What conformational changes does the RNA-DNA hybrid undergo upon binding to RNase HI? It has been reported that RNase HI undergoes only minor conformational changes after binding to a double helix, but the corresponding free DNA has not been crystallized to directly observe the differences and changes in DNA conformation upon binding.
2) How can AmrZ specifically recognize two DNA binding sites that do not have the same sequence? AmrZ binds as a dimer-of-dimers to its own gene, *amrZ*, which has a palindromic sequence for each homodimer. But the *algD* site to which AmrZ binds as an activator does not have the same sequence or even a palindromic sequence for the dimer-of-dimers to bind, suggesting that the binding mechanism between these promoters is likely to be different.

In order to answer the first question, the crystal structure of an RNA-DNA hybrid was determined and compared to an existing duplex with the same sequence that was bound to RNase HI. The results are presented in Chapter 4. The hybrid adopts a generalized A-form conformation but the DNA strand near the scissile bond has B-form which gives a unique major and minor groove and is probably crucial for RNA-DNA

hybrid recognition by RNase HI. These results gave a clearer understanding of how RNase HI finds ribonucleotides in DNA and the conformational changes the duplex undergoes upon binding in order for RNase HI to cleave the ribonucleotides.

The second question can be answered by solving the crystal structure of AmrZ bound to *algD*. Crystals have been produced but contain a twinning pathology which complicated structural determination. Preliminary data indicate that AmrZ probably only binds as a dimer to the *algD* sequence instead of the dimer-of-dimers observed when bound as a repressor.

The experiments, results, and information contained within this thesis allow us to ascertain a greater understanding about the molecular actions and changes that DNA binding proteins confer upon binding and may prove to be useful in answering questions dealing with substrate identification or conformational changes/deformations of proteins. Hopefully this material will help further research to better explain how DNA binding proteins are able to activate and/or repress genes, identifying their binding sequence in a sea of DNA, and how proteins selectively and specifically bind to their substrate in a timely manner.

References

[1] Watson, J., Crick, F. 1953. Molecular Structure of Nucleic Acids. *Nature*. 4356:737-738.

[2] Lu, X.J., Olson, W.K. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*. 31(17):5108-21.

[3] Dickerson, R. 1992. DNA Structure from A to Z. *Methods in Enzymology*. 211:67-111.

[4] Wang, A., Fujii, S., van Boom, J., Rich, A. 1982. Molecular Structure of the Octamer d(G-G-C-C-G-G-C-C): Modified A-DNA. *Proc. Natl. Acad. Sci. USA.* 79:3968-3972.

[5] Horton, N., Finzel, B. 1996. The Structure of an RNA/DNA Hybrid: A Substrate of the Ribonuclease Activity of HIV-1 Reverse Transcriptase. *J. Mol. Biol.* 264:521-533.

[6] Huange, D. Phelps, C., Fusco, A., Ghosh, G. 2004. Crystal Structure of a Free κB DNA: Insights into DNA Recognition by Transcription Factor NF-κB. *J. Mol. Biol.* 346:147-160.

[7] Rohs, R., Jin, X., West, S., Joshi, R., Honig, B., Mann, R. 2010. Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem.* 79:233-269.

[8] Gyi, J., Lane, A., Conn, G., Brown, T. 1998. Solution Structures of DNA-RNA Hybrids with Purine-Rich and Pyrimidine-Rich Strands: Comparison with the Homologous DNA and RNA Duplexes. *Biochemistry*. 37:73-80.

[9] Davis, R.R., Shaban, N.M., Perrino, F.W., Hollis, T. 2015. Crystal structure of RNA-DNA duplex provides insight into conformational changes induced by RNase H. *Cell Cycle*. 14(4):668-673

[10] Fedoroff, O., Salazar, M., Reid, B. 1993. Structure of a DNA:RNA Hybrid Duplex: Why RNase H Does Not Cleave Pure RNA. *J. Mol. Biol.* 233:509-523.

[11] Hantz, E., Larue, V., Ladam, P., Le Moyec, L., Gouyette, C., Huynh Dinh, T. 2000. Solution Conformation of an RNA-DNA Hybrid Duplex Containing a Pyrimidine RNA Strand and a Purine DNA Strand. *Journal of Biological Macromolecules*. 28:273-284.

[12] Conn,G., Brown, T., Leonard, G. 1999. The Crystal Structure of the RNA/DNA Hybrid r(GAAGAGAAGC)-d(GCTTCTCTTC) Shows Significant Differences to that Found in Solution. *Nucleic Acids Research*. 27(2):555-561.

[13] Xiong, Y., Sundaralingam, M. 2000. Crystal Structure of a DNA-RNA Hybrid Duplex with a Polypurine RNA r(gaagaagag) and a Complementary Polypyrimidine DNA d(CTCTTCTTC). *Nucleic Acids Research*. 28(10):2171-2176.

[14] Han, G., Kopka, M., Langs, D., Sawaya, M., Dickerson, R. 2003. Crystal Structure of an RNA-DNA Hybrid Reveals Intermolecular Intercalation: Dimer Formation by Base-Pair Swapping. *Proc. Natl. Acad. Sci. USA.* 100(16):9214-9219.

[15] Stormo, G., Zhao, Y. 2010. Determining the Specificity of Protein-DNA Interactions. *Nature Reviews*. 11:751-760.

[16] Blanco, F., Montoya, G. 2011. Transient DNA/RNA-Protein Interactions. *FEBS J*. 278:1643-1650.

[17] Dror, I., Zhou, T., Mandel-Gutfreund, Y., Rohs, R. 2013. Covariation Between Homeodomain Transcription Factors and the Shape of their DNA Binding Sites. *Nucleic Acids Research*. 41:1-12.

[18] Nahalka, J. 2013. Protein-RNA Recognition: Cracking the Code. *Journal of Theoretical Biology*. 343:9-15.

[19] Zhao, Y., Ruan, S., Pandey, M., Stormo, G. 2012. Improved Models for Transcription Factor Binding Site Identification Using Nonindependent Interactions. *Genetics*. 191:781-790.

[20] Steffen, N., Murphy, S., Tolleri, L., Hatfield, G., Lathrop, R. 2002. DNA Sequence and Structure: Direct and Indirect Recognition in Protein-DNA Binding. *Bioinformatics*. 18(1):S22-S30.

[21] Rohs, R., West, S., Sosinsky, A., Liu, P., Mann, R., Honig, B. 2009. The Role of DNA Shape in Protein-DNA Recognition. *Nature*. 461(7268):1248-1253.

[22] Mandel-Gutfreund, Y., Margalit, H. 1998. Quantitative Parameters for Amino Acid-Base Interaction: Implications for Prediction of Protein-DNA Binding Sites. *Nucleic Acids Research*. 26(10): 2306-2312.

[23] Garvie, C., Wolberger, C. 2001. Recognition of Specific DNA Sequences. *Molecular Cell*. 8:937-946.

[24] Seeman, N., Rosenberg, J., Rich, A. 1975. Sequence-Specific Recognition of Double Helical Nucleic Acids by Protein. *Proc. Natl. Acad. Sci. USA.* 73(3):804-808.

[25] Cerritelli, S., Crouch, R. 2009. Ribonuclease H: the enzymes in eukaryotes. *FEBS Journal*. 276:1494-1505.

[26] Nowotny, M., Cerritelli, S., Ghirlando, R., Gaidamakov, S., Crouch, R., Yang, W. 2008. Specific recognition of RNA/DNA hybrid and enhancement of human RNase H1 activity by HBD. *EMBO Journal*. 27(7):1172-1181.

[27] Hiller, B., Achleitner, M., Glage, S., Naumann, R., Behrendt, R., Roers, A. 2012. Mammalian RNase H2 removes ribonucleotides from DNA to maintain genomic integrity. *JEM*. 290(8):1419-1426.

[28] Nowotny, M., Gaidamakov, S., Ghirlando, R., Cerritelli, S., Crouch, R., Yang, W. 2007. Structure of human RNase H1 complexed with an RNA/DNA hybrid: Insight into HIV reverse transcription. *Molecular Cell*. 28:264-276.

[29] Shaban, N., Harvey, S., Perrino, F., Hollis, T. 2009. The structure of mammalian RNase H2 complex provides insight into RNA-DNA hybrid processing to prevent immune dysfunction. *J. Biol. Chem.* 285(6):3617-3624.

[30] Chon, H., Vassilev, A., DePamphilis, M.L., Zhao, Y., Zhang, J., Burgers, P.M., Crouch, R.J., Cerritelli, S.M. 2009. Contributions of the two accessory subunits, RNASEH2B and RNASEH2C, to the activity and properties of the human RNase H2 complex. *Nucleic Acids Res.* 37(1):96-110.

[31] Tadokoro, T., Kanaya, S. 2009. Ribonuclease H: molecular diversities, substrate binding domains, and catalytic mechanism of the prokaryotic enzymes. *FEBS J.* 276:1482-93.

[32] Nowotny, M., Gaidamakov, S., Crouch, R., Yang, W. 2005. Crystal structures of RNase H bound to an RNA/DNA hybrid: Substrate specificity and metal-dependent catalysis. *Cell*. 121:1005-1016.

[33] Figiel, M., Chon, H., Cerritelli, S., Cybulska, M., Crouch, R., Nowotny, M. 2010. The structural and biochemical characterization of human RNase H2 complex reveals the molecular basis for substrate recognition and Aicardi-Goutières Syndrome defects. *J. Biol. Chem.* 286(12):10540-10550.

[34] Nick McElhinny, S., Watts, B., Kumar, D., Watt, D., Lundstrom, E., Burgers, P., Johansson, E., Chabes, A., Kunkel, T. 2010. Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases. *Proc. Natl. Acad. Sci. USA.* 107:4949-4954.

[35] Rychlik, M., Chon, H., Cerritelli, S., Klimek, P., Crouch, R., Nowotny, M. 2010. Crystal structures of RNase H2 in complex with nucleic acid reveal the mechanism of RNA-DNA junction recognition and cleavage. *Molecular Cell*. 40:658-670.

[36] Schreiter, E., Drennan, C. 2007. Ribbon-helix-helix transcription factors: variations on a theme. *Nature*. 5:710-719.

[37] Tart, A., Blanks, M., Wozniak, D. 2006. The AlgT-dependent transcriptional regulator AmrZ (AlgZ) inhibits flagellum biosynthesis in mucoid, nonmotile *Pseudomonas aeruginosa* cystic fibrosis isolates. *J. Bact.* 188(18):6483-6489.

[38] Ramsey, D., Baynham, P., Wozniak, D. 2005. Binding of *Pseudomonas aeruginosa* AlgZ to sites upsteam of the *algZ* promoter leads to repression of transcription. *J. Bact.* 187(13):4430-4443.

[39] Pryor, E., Waligora, E., Xu, B., Dellos-Nolan, S., Wozniak, D., Hollis, T. 2012. The transcription factor AmrZ utilizes multiple DNA binding modes to recognize activator and repressor sequences of *Pseudomonas aeruginosa* virulence genes. *PLoS Pathogens*. 8(4):e1002648.

[40] Waligora, E., Ramsey, D., Pryor, E., Lu, H., Hollis, T., Sloan, G., Deora, R., Wozniak, D. 2010. AmrZ beta-sheet residues are essential for DNA binding and transcriptional control of *Pseudomonas aeruginosa* virulence genes. *J. Bact.* 192(20):5390-5401.

Chapter 2:

The role of AmrZ in alginate production in *Pseudomonas aeruginosa*

2.1 *Pseudomonas aeruginosa*

Pseudomonas aeruginosa is a rod-shaped, Gram negative soil bacterium that can cause infections in humans as well as plants but is also able to break down toxic waste and organic compounds that are leached into the ground [1]. *P. aeruginosa* commonly causes burn wound infections, dermatitis, and eye infections. The bacteria can be found in the lungs of cystic fibrosis patients and because of the production of a biofilm they can be extremely difficult to eradicate. Besides forming the protective biofilm, *Pseudomonas* produces an exotoxin that causes tissue damage which only exacerbates the complications in cystic fibrosis [1].

2.2 Cystic Fibrosis Implications with *Pseudomonas aeruginosa*

Cystic fibrosis (CF), also known as mucoviscidosis, is an autosomal recessive disorder that is caused by a mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) gene and affects the lungs, reproductive system, liver, pancreas, and intestines [2]; the lungs are the primary organ impacted and this leads to frequent and chronic lung infections. The CFTR transports chloride anions from the intracellular fluid to the extracelluar matrix in epithelial cells responsible for making mucus. With a defective CFTR, chloride anions do not cross into the extracellular fluid and because of this water is absorbed into the cell causing the sputum to become thick and sticky [2]. The depletion of surface liquids in the lungs causes the area to become dehydrated and sputum retention leads to many symptoms displayed by CF patients, including chronic

broncopulmonary infections which 80% of all CF pateints develop [3]. When *Pseudomonas aeruginosa* inoculates the thick sputum, it converts into a mucoid-producing strain and generates a biofilm to combat the lack of water in its surroundings which allows it to become the predominant pathogen in the lungs [4]. When *Pseudomonas* switches into the mucoid producing strain, it loses flagella- and cilia-based motility and therefore becomes planktonic. Studies have shown that 95% of CF patient deaths are due to respiratory failure caused by the biofilm accumulation [4]. While in the biofilm, *Pseudomonas* can be 1,000 times more resistant to antibiotics than its swimming, mobile counterpart [5]. It has been demonstrated that *P. aeruginosa* secretes three exopolysaccharides: Psl, Pel, and alginate [6]. Psl is composed of mannose and galactose and is responsible for initial attachment to the sputum or epithelial cells. Pel is glucose rich and resembles cellulose to mediate cell-cell interactions and pellicle formation at the surface air-liquid interface. Alginate is a linear polymer composed of mannuronate and guluronate and ensures the microenvironment and lipid membranes are hydrated to protect against osmotic stress [7]. Alginate has been shown to be a hygroscopic polymer but is not required for biofilm formation. If both Psl and Pel are knocked out, then *P. aeruginosa* does not produce any biofilm, but if alginate is knocked out, only slight changes in the architecture of the biofilm are observed.

2.3 Alginate Biosynthesis

Alginate biosynthesis is a well-studied pathway in *Pseudomonas aeruginosa* and has shown regulatory complexity comparable to that of eukaryotes. The alginate trait has

been repeatedly demonstrated to be unstable and the mucoid strains isolated from CF patient's lungs often resort back to the nonmucoid phenotype [8-10]. Alginate is a negatively charged polysaccharide polymer composed of mostly D-mannuronate with randomly interspersed L-guluronate sugars linked by β-1,4 glycosidic bonds. The guluronate has been shown to affect the alginate structure because it exists in a different chair conformation than mannuronate in order to maintain the carboxyl group in the equatorial position which results in a change in viscosity, altering the ability for alginate to form a gel [8]. Mannuronate rich gels are weak and flexible where-as guluronate rich gels are strong but brittle. *Pseudomonas* does not utilize guluronate rich gels but epimerizes mannuronate into guluronate to aid in strengthening the alginate structure and increasing divalent cation attraction, especially calcium, thus bringing more water into the biofilm. Alginate can also have some beneficial roles, especially in the medical field; one example is that it is used to release drugs to patients in surgery in a controlled manner [11].

Alginate is produced in four steps: synthesis of GDP-mannuronic acid, polymerization of GDP-mannuronic acid into β-1,4 polymannuronate, O-acetylation or epimerization of mannuronate, and secretion of the polymer [12]. *Pseudomonas* contains an entire operon of proteins with various roles and functions in the production of alginate and a summary of the pathway is shown in Figure 2-1. The first step is to convert fructose-6-phosphate into mannose-6-phosphate which is performed by the phosphomannose isomerase activity of AlgA. Then AlgC, phosphomannomutase, converts mannose-6-phosphate into mannose-1-phosphate. AlgA then uses its pyrophosphorylase activity to convert mannose-1-phosphate into GDP-mannose [13].

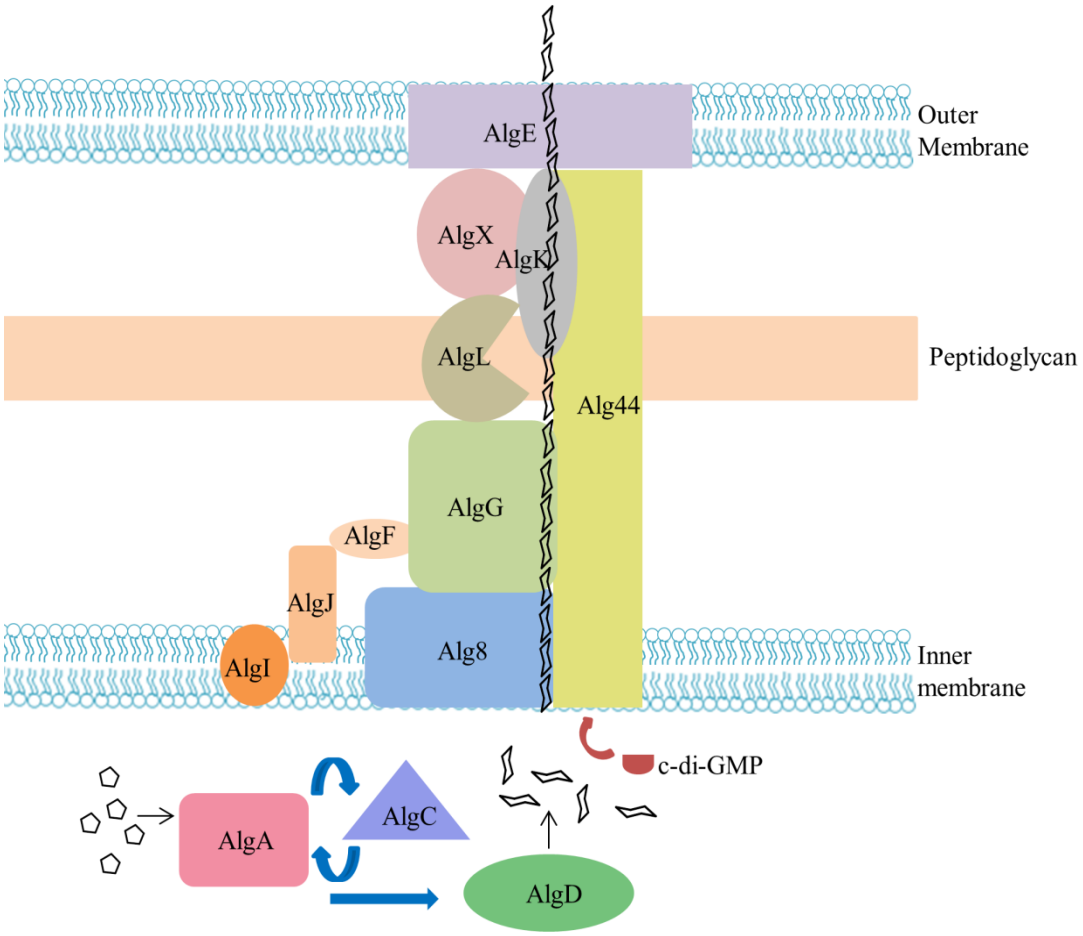

Figure 2-1: **Alginate biosynthesis Pathway.** The alginate pathway is involved in the production of alginate from fructose-6-phosphate to the secretion of the polymer through the outer membrane. AlgA, AlgC, and AlgD convert fructose-6-phosphate into GDP-mannuronate. Alg8 and Alg44 polymerize GDP-mannuronate into polymannuronate which is acetylated by AlgI, AlgJ, and AlgF. AlgG epimerizes some mannuronates into guluronate and is part of the scaffold with AlgL, AlgX, and AlgK to protect the polymer from AlgL until it is secreted through the AlgE porin.

AlgD, GDP-mannose dehydrogenase, converts GDP-mannose into GDP-mannuronic acid and this is the first committed step of producing alginate [14]. The conversion of fructose-6-phosphate to GDP-mannuronate happens in the cytoplasm of the bacteria.

The second step of alginate biosynthesis, after GDP-mannuronic acid has been produced, is to polymerize the monosaccharides into a polymer. Alg44, a membrane

bound signaling protein, likely aids in the formation of alginate by interacting with the polymerase, Alg8 [13]. Alg44 has a PilZ domain which binds to cyclic-di-GMP and induces a conformational change needed for Alg8 to become activated [15]. Alg8, structurally, resembles glycosyltransferases and links the GDP-mannuronates together into a polymer [13]; removing the Alg8 polymerase results in the complete absence of alginate, indicating its importance for saccharide polymerization [6]. In order for Alg8 to add the mannuronates, it must either add two molecules together or rotate the chain as it passes through; if one monomer is added at a time the chain is forced to rotate due to a twofold screw axis that is formed with the addition of one monomer [11]. This process occurs across the inner membrane of *Pseudomonas*.

The third step is responsible for modifying the polymer chain as it passes through the periplasm. The modifications are often limited to O-acetylation and epimerization. O-acetylation assists in the formation of microcolonies in the biofilm and also increases antibiotic resistance and avoids opsonic phagocytosis by blocking the C3b and C4b complement proteins [12,16]. There are three proteins dedicated to O-acetylation; AlgI, AlgJ, and AlgF but their specific roles are currently unknown. AlgI is a membrane protein which interacts with AlgJ. AlgJ, a membrane bound protein, interacts with AlgF, which is located in the periplasm [17]. The two acetylation locations on mannuronate are O-2 and O-3 [12]. Epimerization is carried out by the C-5 epimerase AlgG in which mannuronate is converted into its C-5 epimer guluronate. AlgG, however, cannot place two guluronate sugars beside each other, which prevents the formation of guluronate rich regions in the alginate in *Pseudomonas* [12]. Interestingly, if a mannuronate is transacetylated, then epimerization is prevented on that residue [11].

The final step of alginate production is secreting the polymer through the outer membrane. Moving alginate from the inner membrane to the outer membrane requires cooperation among various proteins that form a scaffold to ensure that all proteins are present and in their required locations. The current hypothesis is that AlgG, AlgK, AlgX, and AlgL form part of the scaffold with Alg44 contributing to remainder of the scaffold [18,19]. AlgK is an outer membrane-bound periplasmic protein that assists the translocation of AlgE, a porin protein, to the outer membrane [20]. AlgE interacts with AlgK and has a positively charged pore that is selective for alginate [21]. AlgX is an acetyltransferase that aids in the O-acetylation process but is also a key component of the scaffold complex [21,22]. AlgL is an alginate lyase responsible for cleaving the polymer, using a β-elimination reaction, if any of the scaffold proteins are not present. Another role of AlgL may also be to control the length of alginate as it passes through the periplasm; deletion of AlgL causes *Pseudomonas* to rupture by separating the inner and outer membranes from the accumulation of alginate [19].

2.4 Regulation of the *algD* Operon

2.4.1 Protein Regulation

The alginate operon is one of the most regulated regions in the entire bacterium. Two main types of regulation occur in the alginate pathway: protein regulation and DNA regulation. The protein regulation happens on the inner membrane near the proteins that produce alginate. These regulatory proteins are called MucA-D and most are negative

effectors of the sigma factor (σ) AlgU (also known as AlgT or σ^{22}) (Figure 2-2). AlgU is the same as σ^{E} in *E. coli* which is a stress response sigma factor required to express the *algD* operon [14]. Any mutations in the Muc proteins results in *P. aeruginosa* switching to the mucoid phenotype because most mutations inactivate the anti-σ^{22}, MucA [10]. MucA binds to AlgU to prevent it from binding to RNAP and initiating transcription; at the same time AlgQ binds to the general sigma factor, σ^{70}, to ensure AlgU can activate the *algD* operon when it is released from MucA [23]. AlgW and MucP are both proteases that can cleave MucA and release AlgU into the cytosol. AlgW is activated by MucE. MucB binds to MucA which protects it from cleavage, thus ensuring AlgU stays sequestered to the membrane [24]. MucD functions to degrade proteins that are responsible for activating AlgW. MucB and MucD both function to ensure that alginate is not produced until the proper signals have been received. However, a mutation in MucA is all that is needed to overcome the inhibition and release AlgU without going through the MucB/D regulation. The entire activation process has been deemed the regulated intramembrane proteolysis (RIP) pathway and three steps were discovered for releasing AlgU from MucA. First, AlgW cleaves the periplasmic region of MucA and then MucP cleaves a second site on MucA which allows MucA bound to AlgU to enter the cytoplasm before it is finally degraded by other proteases so that AlgU is released to bind to the *algD* promoter [23]. MucR produces c-di-GMP which is used to activate Alg44 and signal Alg8 to start polymerization, as discussed previously [25]. The last regulatory protein is AlgO and is another protease, but unlike AlgW and MucP, when AlgO cleaves MucA *P. aeruginosa* retains a permanent mucoid phenotype [9].

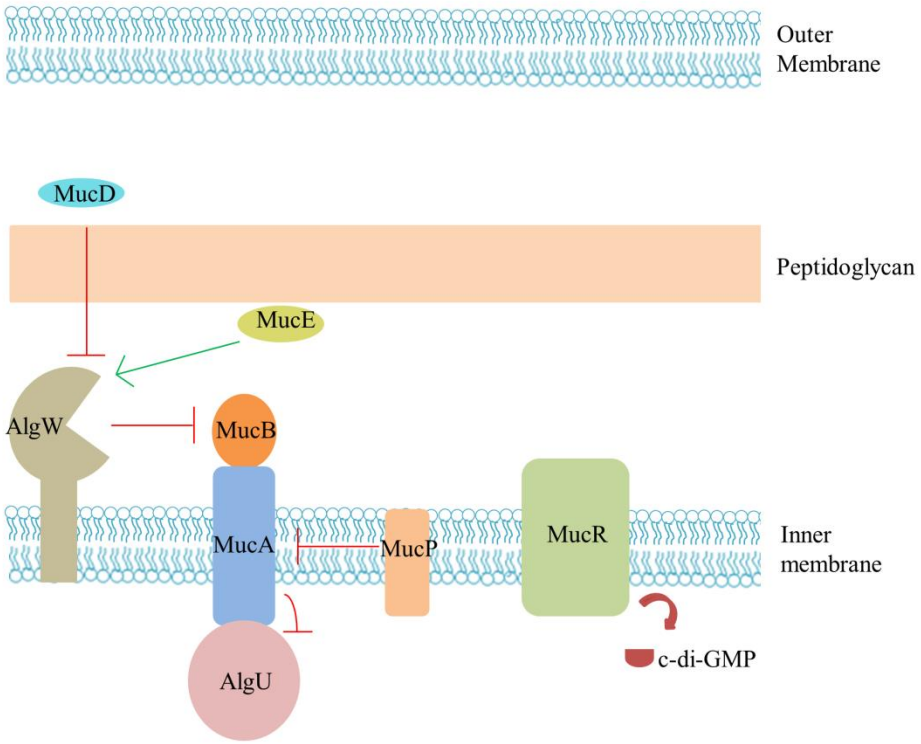

Figure 2-2: **Regulation of AlgU.** Regulated intramembrane proteolysis (RIP) pathway for sequestering AlgU to the membrane to prevent activation of the *algD* operon. MucA binds to AlgU to prevent it from activating *algD* transcription. MucB protects MucA from cleavage by AlgW or MucP. MucE activates AlgW and MucD inhibits AlgW. MucR produces c-di-GMP needed by Alg44 to start alginate polymerization.

2.4.2 DNA Regulation

Transcription factors regulate gene expression by binding to DNA and activating or repressing transcription. Many transcription factors belong to superfamilies with conserved structural motifs that bind to DNA using common features and mechanisms. Common modes of regulation are for DNA to interact with the RNA polymerase (RNAP) to increase affinity to the DNA (positive regulation), in terms of activation, or prevent RNAP from binding or transcribing (negative regulation), in repression [26]. Transcription factors are often regulated by environmental signals and small molecules

which allow them to respond to environmental changes at a faster rate than other proteins.

Most of the transcription factors that regulate the *algD* operon bind distally upstream from the start site. All of the transcription factors mentioned below act as activators and *algD* expression will occur even if some are missing, as long as AlgU and integration host factor (IHF) are present. The main activator proteins are: AlgU, CysB, AlgB, AlgR, CAP, IHF, and AmrZ but unfortunately the exact mechanism as to how these proteins each play a role in alginate activation is unknown (Figure 2-3).

AlgU, as discussed previously, is the stress response σ that is required for alginate expression. CysB is postulated to maintain the 3' region of the *algD* promoter in a supercoiled state to promote activation [4]. AlgB is a transcription factor with a helix-turn-helix motif, discussed below, and binds -270 base pairs upstream from the start site [27]. AlgR was the first activator discovered for the *algD* operon and has three binding sites at -37, -394 and -470. The current hypothesis is that AlgR interacts with AlgU to promote quicker binding of RNA polymerase (RNAP) to the TATA box. The upstream AlgR proteins may either aid in recruitment of other transcription factors or favorably interact with RNAP as transcription is initiated. Catabolite activator protein, CAP, binds -368 base pairs upstream and interacts with the C-terminal domain of the α subunit of RNAP [4]. IHF is a DNA bending protein and bends the DNA 160° by intercalating proline residues into the minor groove of the DNA and causing kinks due to hydrophobic stacking [28]. Most IHF binding sites have A-tracts on either side of the kink, which are characterized by having straight, narrow minor groove and high propeller twists that aid in DNA deformation. IHF scans the DNA and finds binding sites based on the structure

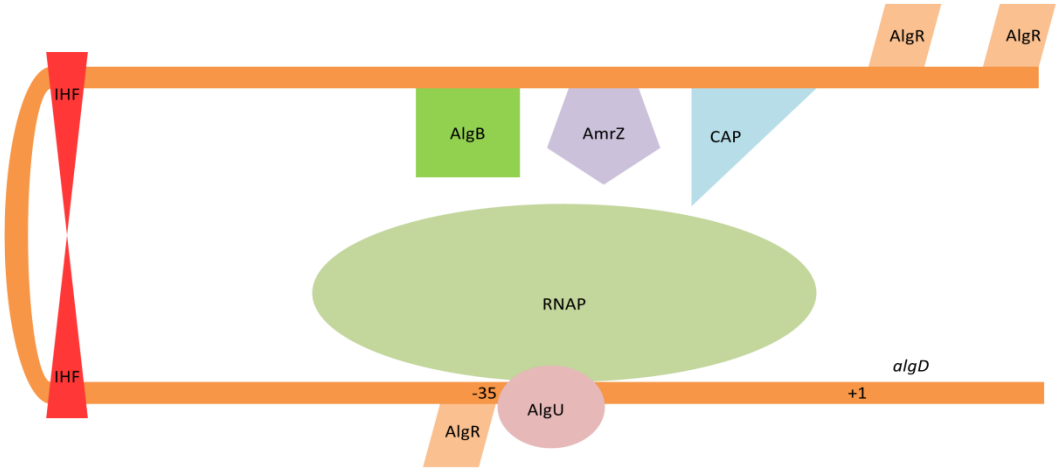

Figure 2-3: **Regulation of the *algD* operon.** Model for regulation of the expression of alginate in the *algD* promoter region by AlgB, AlgR, AlgU, AmrZ, and CAP.

of the DNA (indirect readout) rather than specific base pairs [28]. The summary of the entire alginate biosynthesis from activation of transcription through expression of alginate is shown in Figure 2-4.

2.4.3 AmrZ Repression

One of the most common superfamilies of transcription factors, especially in prokaryotes, is the helix-turn-helix (HTH) [26]. They are so named because of the conserved motif of an α-helix followed by a loop or turn and then another α-helix, where one α-helix is inserted into the major groove of the DNA. MetJ was the first transcription factor determined that fell into a new superfamily; ribbon-helix-helix (RHH) [29]. Arc and Mnt were added to the superfamily shortly after the discovery of MetJ [30,31]. The RHH motif was different because it had a β-sheet followed by two α-helices with the

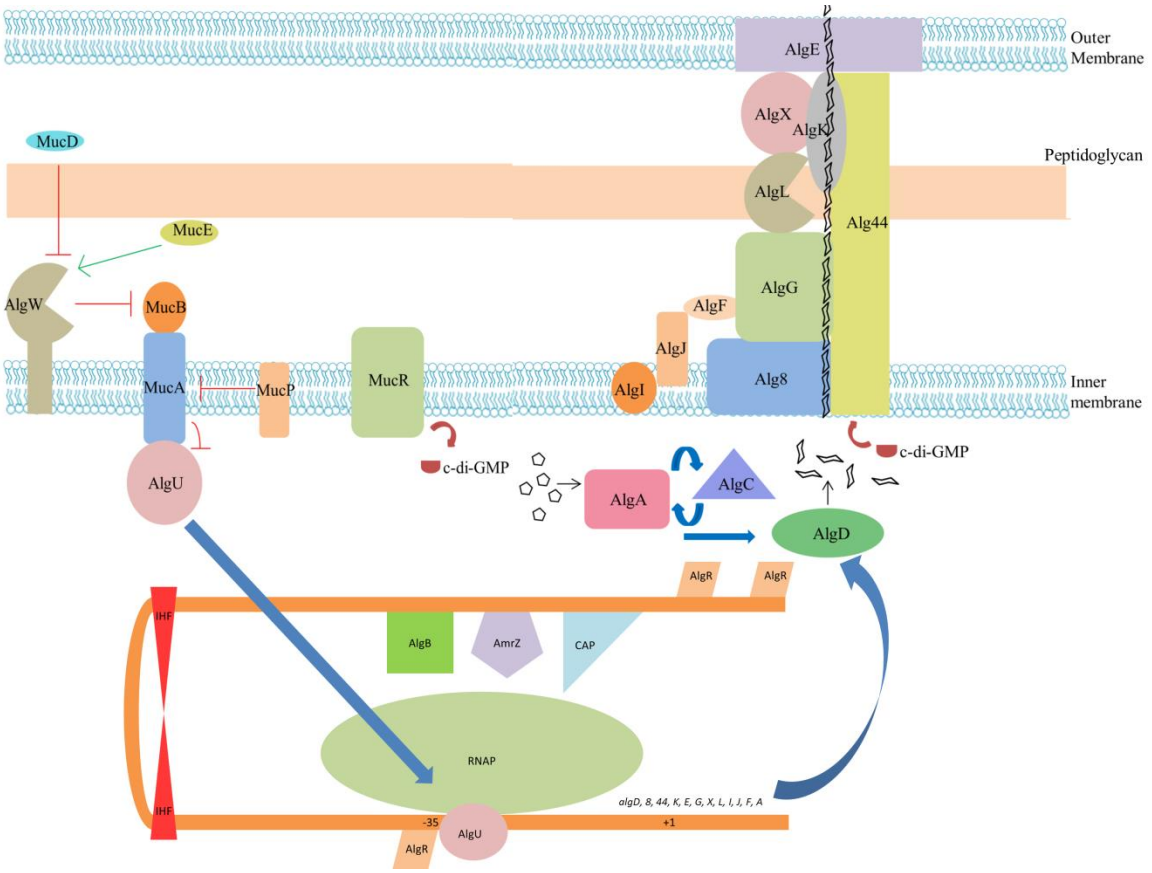

Figure 2-4: **Alginate transcription and biosynthesis regulation.** Regulation of AlgU by the Muc regulatory proteins, transcription initiation, formation, and secretion of alginate through the inner and outer membranes. After MucA is cleaved and AlgU goes to its binding site -35 base pairs upstream, RNAP can bind and start transcribing the genes needed to produce the proteins required to make and secrete alginate.

β-sheet being inserted into the major groove of the operator. Many RHH containing transcription factors exist as homodimers and are tightly bound together [26]. When RHH homodimers bind to DNA the β-sheets are inserted in an antiparallel fashion and approximately three to four residues make direct, sequence-specific contact with the nucleotides [32]. Non-specific interactions are made between amide nitrogens and the phosphate backbone, which are energetically favorable due to the negatively charged phosphate backbone interacting with the positively charged amide groups. Because of

this non-specific binding to the phosphate backbone, it is believed that transcription factors bind to the DNA and slide until specific contacts are made between the residues that are inserted into the major groove and the nucleobases that satisfy hydrogen bonding potentials [33]. RHH proteins typically bind to operator sequences of approximately six bases that are palindromic, or are at least repeated since they commonly exist as homodimers [26,34]; these sequences are normally separated by a linker region of five to ten nucleotides [34,35]. Upon binding to DNA, almost no changes occur in the RHH transcription factor which implies that conformational changes are induced in the DNA strand [26]. Most of the known RHH transcription factors are repressors and have a protein-DNA interaction area of almost 1,500 $Å^2$ [35].

Genomic wide studies have provided a consensus sequence for AmrZ based on the 398 binding sites found in *P. aeruginosa*. The recognition sequence is CAAATTGCCATCA [36]. Only one site was found not to be between -100 and +15 base pairs from the start site which leads to a proposal that AmrZ interacts with RNAP. The crystal structure of AmrZ bound to its repressor, *amrZ1*, has already been determined (Figure 2-5) [35]. The structure revealed that Lys18, Val20, and Arg22 are the essential recognition amino acids and interact directly with the base pairs, while Ser13, Arg28, Ser41, Met42, Asn43, and Ser44 are used to interact with the backbone [35]. When AmrZ is bound to the DNA, the minor groove is significantly narrowed, but how much of the narrowing is an artifact of the DNA sequence is not known. The AT rich region is important for the binding of AmrZ because ApA and TpT steps have narrow minor grooves and less flexibility in the DNA strand [35].

The binding sequence to *amrZ1* is 5′-GGCAAAACGCC-3′ / 3′-CCGTTTTGCGG-5′ (shown with the palindromic sequences in red [35]) which is similar to the consensus sequence found by Jones, et al. (2013) [36]. The base sequence RGCY, where R is a purine and Y is a pyrimidine, forms base pair stacking between the two purines and the two pyrimidines which prevents a propeller twist of the bases and base pair rolling, causing a bend between the G and C due to the increased flexibility [37]. Accompanied by the narrowing of the A-tract minor groove, this could be the reason behind AmrZ's specificity to that consensus sequence. The palindromic sequence allows two dimers, referred to as a dimer-of-dimers, to bind to the sequence at both the *amrZ1* and *amrZ2* sites on the *amrZ* promoter. When IHF binds and bends the DNA, the C-terminal ends, shown to be involved in oligomerization, may interact and tetramerize with the dimer-of-dimers and form a basis for repression.

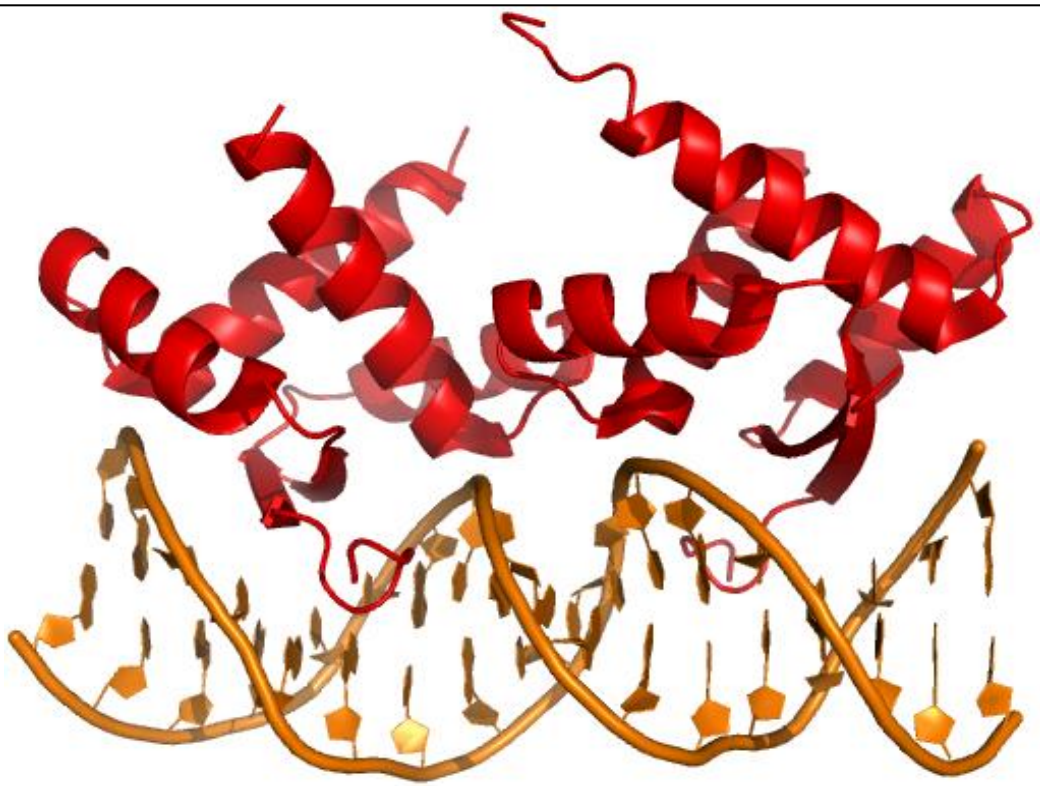

Figure 2-5: **AmrZ bound as a repressor to *amrZ1*.** Two AmrZ dimers bound in the major grooves of *amrZ1* while narrowing the minor groove (3QOQ) [35]. The β-sheets are inserted antiparallel to each other into the major groove. The A-tract between the two major grooves may assist in narrowing the minor groove.

2.4.4 AmrZ Activation

AmrZ has been shown to repress the polymer formation of Psl while activating expression of alginate which shows that Psl and alginate formation are inversely regulated [25]. AmrZ has a similar, but not identical, sequence on the *algD* promoter. The sequence, 5′-GGCCATTACCA-3′ / 3′-CCGGTAATGGT-5′ is not palindromic but still contains the AT rich region. Based on DNA binding analysis and mutational studies, it was believed that AmrZ had a different mode of binding to the *algD* promoter. Given the lack of two consensus sequences on the *algD* promoter, it was hypothesized that AmrZ would only bind as a dimer, which may distinguish when it needs to act as a repressor or activator. When K18, Val20, or Arg22 are mutated, DNA binding to the *algD* promoter is abolished but interestingly, when K18 is mutated and tested with *amrZ1*, there is no observable decrease in binding which implies that K18 may play a role in differentiating between AmrZ's role in activation or repression [35, 38]. The best method to discover the binding interactions of AmrZ as an activator is to determine the crystal structure of AmrZ bound to *algD*, which is discussed in the next chapter.

2.4.5 Mechanism of Activation and Repression

The effect of activators is to decrease the activation energy needed to start transcription initiation, whereas repressors serve to increase the activation energy [39]. When proteins bind as repressors they have multiple methods of repressing transcription: i) block the RNAP α subunits from binding; ii) bind where RNAP needs to bind; iii) act

as anti-σ to prevent binding of the RNAP to σ; iv) bind to the TATA box or -10/-35 regions to increase promoter strength and prevent DNA melting; v) bind the general sigma factors, such as σ^{70} or σ^{54}, where the stress response σ^{22} needs to bind (a process called sigma factor antagonism) [39,40]. Typically, when IHF is involved in binding to the promoter region of a gene, the bending of the DNA plays a huge role in the activation or repression of that gene or operon. DNA looping can serve several roles and can cause either activation or repression depending on the function of the proteins binding to the DNA and the orientation of the -35 region. DNA looping can cause four major modes of repression: i) steric hindrance of the RNAP if the -35 region is on the inside of the loop, ii) distortion in the DNA which causes RNAP not to recognize the binding site, iii) tightening of the DNA so that it requires a much higher activation energy to open the helix, iv) prevention of promoter escape or transcription elongation by increasing the tightness of the DNA helix [41]. Activation can function by similar mechanisms if the RNAP recognition sequence is on the outside of the loop, if the proteins bound near RNAP favorably interact to recruit the RNAP to the promoter quicker, or if the proteins can aid in the start of initiation to get RNAP to dissociate from the σ quicker to move into the elongation phase. AmrZ could act like a class I activator, similar to CAP, to interact with the C-terminal domains of the α subunit to assist in RNAP recruiting to the promoter [42].

The current hypothesis is that AmrZ uses DNA bending in both activation and repression but utilizes different mechanisms of actions (Figure 2-6). Due to the presence of two binding sites on the *amrZ* promoter, AmrZ likely binds to both sites and, once the DNA is bent, tetramerizes, locking the DNA in place and sterically hindering RNAP

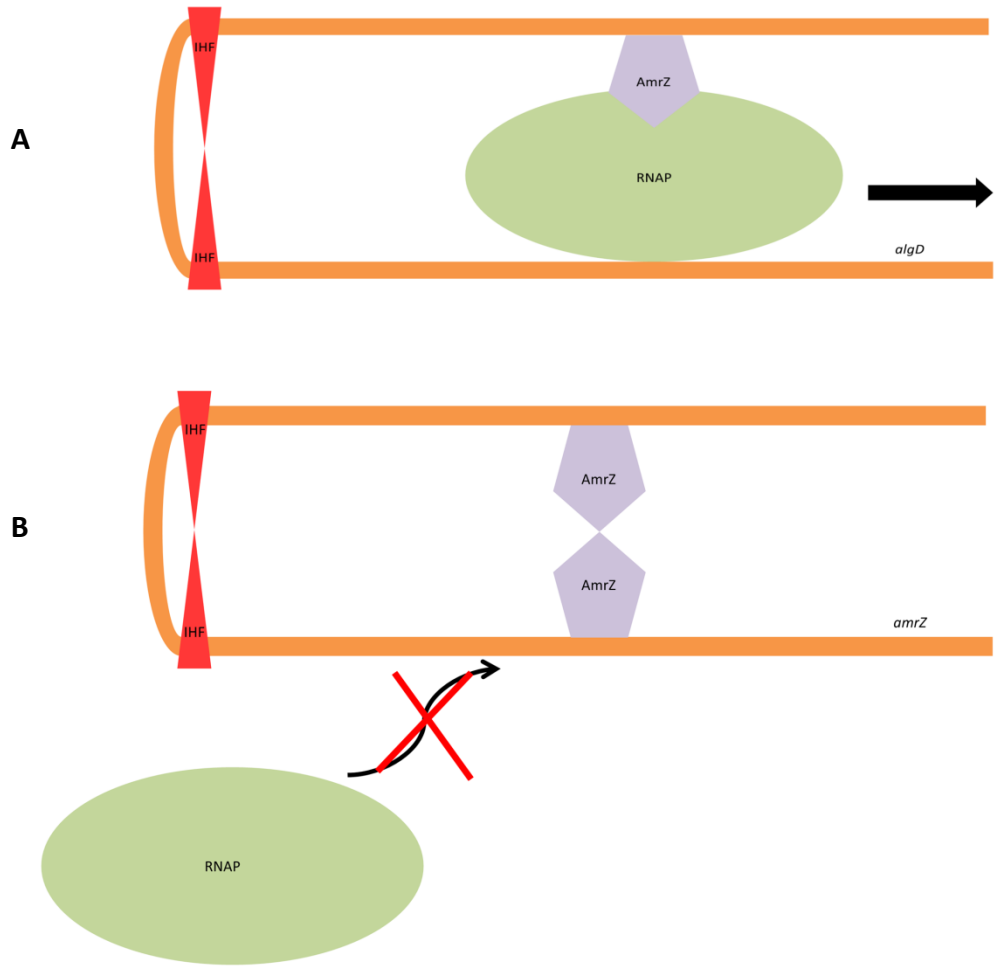

Figure 2-6: **Models of AmrZ activation and repression.** A) Proposed model of the mechanism of action for AmrZ activating *algD* by interacting with RNAP. B) Model showing how AmrZ could tetramerize with its distal dimer to bring the DNA strands close enough to sterically hinder RNAP access on *amrZ* promoter.

from accessing its binding site. For activation of *algD*, AmrZ is positioned in the proximity of the AlgU/RNAP binding site; this mechanism involves either interacting with RNAP to recruit the polymerase quicker or to recruit other transcription factors to the promoter that will directly interact with RNAP.

References

[1] Salyers, A.A., Whitt, D.D. 2002. Bacterial pathogenesis: A molecular approach (2nd edition). Washington, DC. ASM Press.

[2] Sousa, A., Pereira, M. 2014. *Pseudomonas aeruginosa* diversification during infection development in cystic fibrosis lungs-a review. *Pathogens*. 3:680-703.

[3] Moreau-Marquis, S., Stanton, B., O'Toole, G. 2008. *Pseudomonas aeruginosa* biofilm formation in the cystic fibrosis airway. *Pulm Pharm & Ther*. 21(4):595-9.

[4] Baynham, P.J., Brown, A.L., Hall, L.L, Wozniak, D.J. 1999. *Pseudomonas aeruginosa* AlgZ, a ribbon-helix-helix DNA-binding protein, is essential for alginate synthesis and *algD* transcriptional activation. *Mol. Micro.* 33(5):1069-80.

[5] Hentzer, M., Teitzel, G.M., Balzer, G.J., Molin, S., Givskov, M., Matthew, R., Heydorn, A., Parsek, M.R. 2001. Alginate overproduction affects *Pseudomonas aeruginosa* biofilm structure and function. *J. Bact.* 183(18):5395-5401.

[6] Gahfoor, A., Hay, I.D., Rehm, B. 2011. Role of exopolysaccharides in *Pseudomonas aeruginosa* biofilm formation and architecture. *App Env Micro.* 77(15):5238-46.

[7] Chang, W.S., van de Mortel, M., Nielsen, L., Nino de Guzman, G., Li, X., Halverson, L.J. 2007. Alginate production by *Pseudomonas putida* creates a hydrated microenvironment and contributes to biofilm architecture and stress tolerance under water-limiting conditions. *J. Bact.* 189(22):8290-9.

[8] Chitnis, C.E., Ohman, D.E. 1990. Cloning of *Pseudomonas aeruginosa algG* which controls alginate structure. *J. Bact.* 172(6):2894-900.

[9] Sautter, R., Ramos, D., Schneper, L., Ciofu, O., Wassermann, T., Koh, C.L., Heydorn, A., Hentzer, M., Hoiby, N., Kharazmi, A., Molin, S., Devries, C.A., Ohman, D.E., Kalai, M. 2012. A complix multilevel attack on *Pseudomonas aeruginosa algT/U*

expression and *algT/U* activity results in the loss of alginate production. *Gene*. 498(2):242-53.

[10] Devries, C.A., Ohman, D.E. 1994. Mucoid-to-nonmucoid conversion in alginate-producing *Pseudomonas aeruginosa* often results from spontaneous mutations in *algT*, encoding a putative alternate sigma factor, and shows evidence for autoregulation. *J. Bact.* 176(21):6677-87.

[11] Remminghorst, U., Rehm, B.H.A. 2006. Bacterial alginates: from biosynthesis to applications. *Biotech. Letters*. 28(21):1701-12.

[12] Douthit, S.A., Dkakic, M., Ohman, D.E., Franklin, M.J. 2005. Epimerase active domain of *Pseudomonas aeruginosa* AlgG, a protein that contains a right handed β-helix.

[13] Ogleby, L.L., Jain, S., Ohman, D.E. 2008. Membrane topology and roles of *Pseudomonas aeruginosa* Alg8 and Alg44 in alginate polymerization. *Micro*. 154:1605-15.

[14] Baynham, P.J., Wozniak, D.J. 1996. Identification and characterization of AlgZ, an AlgT-dependent DNA-binding protein required for *Pseudomonas aeruginosa algD* transcription. *Mol. Micro.* 22(1):97-108.

[15] Hay, I.D., Remminghorst, U., Rehm, B.H.A. 2009. MucR, a novel membrane-associated regulator of alginate biosynthesis in *Pseudomonas aeruginosa. App. Env. Micro.* 75(4):1110-20.

[16] Paletto, J.L., Ohman, D.E. 2012. Evidence for two promoters internal to the alginate biosynthesis operon in *Pseudomonas aeruginosa. Curr. Micro.* 65(6):770-5.

[17] Franklin, M.J., Ohman, D.E. 2002. Mutant analysis and cellular localization of the AlgI, AlgJ, and AlgF proteins required for O-acetylation of alginate in *Pseudomonas aeruginosa. J. Bact.* 184(11):3000-7.

[18] Jain, S., Franklin, M.J., Ertesvag, H., Valla, S., Ohman, D.E. 2003. The dual roles of AlgG in C-5 epimerization and secretion of alginate polymers in *Pseudomonas aeruginosa*. *Mol. Micro.* 47(4):1123-33.

[19] Jain, S., Ohman, D.E. 2005. Role of an alginate lyase for alginate transport in mucoid *Pseudomonas aeruginosa*. *Infection and Immunity.* 73(10): 6429-6436.

[20] Keiski, C.L., Harwich, M., Jain, S., Neculai, A.M., Yip, P., Robinson, H., Whitney, J.C., Riley, L.M., Burrows, L.L., Dennis, E., Howell, P.L. 2011. AlgK is a TPR-containing protein and the periplasmic component of a novel exopolysaccharide secretion. *Structure*. 18(2):265-73.

[21] Rehman, Z.U., Rehm, B.H.A. 2013. Dual roles of *Pseudomonas aeruginosa* AlgE in secretion of the virulence factor alginate and formation of the secretion complex. *App. Env. Micro.* 79(6):2002-11.

[22] Riley, L.M., Weadge, J.T., Baker, P., Robinson, H., Codee, J.D.C., Tipton, P.A., Ohman, D.E., Howell, P.L. 2013. Structural and functional characterization of *Pseudomonas aeruginosa* AlgX: role of AlgX in alginate acetylation. *J. Biol. Chem.* 288(31):22299-314.

[23] Hay, I.D., Wang, Y., Moradali, M.F., Rehman, Z.U., Rehm, B.H.A. 2014. Genetics and regulation of bacterial alginate production. *Env. Micro.* 1-15.

[24] Hay, I.D., Schmidt, O., Filitcheva, J., Rehm, B.H.A. 2012. Identification of periplasmi AlgK-AlgX-MucD multiprotein complex in *Pseudomonas aeruginosa* involved in biosynthesis and regulation of alginate. *App. Micro. And Biotech.* 93(1):215-27.

[25] Jones, C.J., Ryder,C.R., Mann, E.E., Wozniak, D.J. 2013. AmrZ modulates *Pseudomonas aeruginosa* biofilm architecture by directly repressing transcription of psl operon. *J. Bact.* 195(8):1637-44.

[26] Schreiter, E.R. and Drenan, C.L. 2007. Ribbon-helix-helix transcription factors: variations on a theme. *Nature*. 5:710-719.

[27] Wozniak, D.J., Ohman, D.E. 1991. *Pseudomonas aeruginosa* AlgB, a two-component response regulator of the NtrC family, is required for *algD* transcription. *J. Bact.* 173(4):1406-13.

[28] Rice, P.A., Yang, S., Mizuuchi, K., Nash, H.A. 1996. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell*. 87(7):1295-1306.

[29] Rafferty, J.B., Somers, W.S., Saint-Girons, I., and Phillips, S.E. 1989. Three-dimensional crystal structures of *Escherichia coli* met repressor with and without co-repressor. *Nature*. 341:705-710.

[30] Breg, J.N., van Opheusden, J.H., Burgering, M.J., Boelens, R., and Kaptein, R. 1990 Structure of Arc repressor in solution: evidence for a family of β-sheet DNA–binding proteins. *Nature*. 346:586-589.

[31] Knight, K.L., Bowie, J.U., Vershon, A.K., Kelley, R.D., and Sauer, R.T. 1989. The Arc and Mnt repressors. A new class of sequence-specific DNA-binding protein. *J. Biol. Chem.* 264:3639-3642.

[32] Raumann, B.E., Rould, M.A., Pabo, C.O., and Sauer, R.T. 1994. DNA recognition by β-sheets in the Arc repressor-operator crystal structure. *Nature*. 367:754-757.

[33] Blanco, F. and Montoya, G. 2011. Transient DNA/RNA-Protein Interactions. *FEBS Journal*. 278:1643-1650.

[34] Youderian, P., Bouvier, S., and Susskind, M.M. 1982. Sequence determinants of promoter activity. *Cell*. 30:843-853.

[35] Pryor, E., Waligora, E., Xu, B., Dellos-Nolan, S., Wozniak, D., and Hollis, T. 2012. The transcription factor AmrZ utilizes multiple DNA binding modes to recognize activator and repressor sequences of *Pseudomonas aeruginosa* virulence genes. *PLoS Pathogens*. 8(4):e1002648.

[36] Jones, C.J., Newsom, D., Kelly, B., Irie, Y., Jennings, L.K., Xu, B., Limoli, D.H., Harrison, J.J., Parsek, M.R., White, P., Wozniak, D.J. 2014. ChIP-Seq and RNA-Seq

reveal an AmrZ-mediated mechanism for cyclic-di-GMP synthesis and biofilm development by *Pseudomonas aeruginosa.PLoS Pathogens*. 10(3):e1003984.

[37] Alleman, R.K., Egli, M. 1997. DNA recognition and bending. *Chem and Biol.* 4(9):643-50.

[38] Waligora, E.A., Ramsey, D.M., Pryor, E.E., Lu, H., Hollis, T., Sloan, G.P., Deora, R., Wozniak, D.J. 2010. AmrZ beta-sheet residues are essential for DNA binding and transcriptional control of *Pseudomonas aeruginosa* virulence genes. *J. Bact.* 192(20):5390-401.

[39] Rojo, F. 2001. Mechanisms of transcriptional repression. *Curr. Op. Micro.* 4(2):145-51.

[40] Boucher, J.C., Schurr, M.J., Deretic, V. 2000. Dual regulation of mucoidy in *Pseudomonas aeruginosa* and sigma factor antagonism. *Mol. Micro*. 36:341-51.

[41] Becker, N.A., Greiner, A.M., Peters, J.P., Maher, L.J. 2014. Bacterial promoter repression by DNA looping without protein-protein binding competition. *Nucleic Acids Res.*42(9):5495-504.

[42] Barnard, A., Wolfe, A., Busby, S. 2004. Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Curr. Op. Micro*. 7:102-108.

Chapter 3:

Expression, crystallization and preliminary crystallographic analysis of pseudomerohedrally twinned AmrZ-*algD* crystals from *Pseudomonas aeruginosa*

Ryan R. Davis and Thomas Hollis[+]

Department of Biochemistry, Center for Structural Biology, Wake Forest School of Medicine, Winston-Salem, NC 27157

[+]To whom correspondence should be addressed: Tel: 336-716-0768; Fax: 337-713-1283; E-mail: thollis@wakehealth.edu

In preparation for submission to *Acta Crystallographica Section F*.

3.1 Introduction

Pseudomonas aeruginosa is a Gram negative bacterium commonly found in various infections but is the leading cause of death in cystic fibrosis (CF) patients with 95% of deaths due to respiratory failure from biofilm production [1]. Alginate is a major component of the biofilms produced by *P. aeruginosa* [2-4]. The alginate and motility regulator Z (AmrZ) is a transcription factor that is highly expressed in the lungs of patients with CF and has a dual role as a transcriptional activator and repressor which has been found necessary for the activation of the operon responsible for alginate production [2]. AmrZ is a 108 residue member of the ribbon-helix-helix superfamily and exists in solution as a dimer due to the interaction of two α-helices that are composed of hydrophobic residues [5]. AmrZ is composed of three regions: an N-terminal region with suspected roles in DNA recognition, a DNA binding region, and a C-terminal region responsible for higher order oligomerization [5]. AmrZ has been structurally determined as a repressor bound to its own promoter region *amrZ1* but a structure of AmrZ bound as an activator has not yet been determined. AmrZ binds by inserting two antiparallel β-strands, one from each monomer, into the major groove of the DNA. As a repressor, AmrZ forms a tetramer (dimer-of-dimers) with two palindromic recognition sequences on *amrZ1* to which each dimer bind [5,6]. However, the recognition sequence of *algD* only contains one binding site, indicating that a different mechanism may exist between the activation and repression of genes under the control of AmrZ.

3.2 Results and Discussion

The data was initially processed in the primitive tetragonal point group due to the symmetry in the precession images as well as a reasonable R_{merge}. Molecular replacement in P422 yielded protein-DNA complexes that were overlapping and failed to yield an R_{free} <40% after rigid body refinement, an indicator of twinning [7]. p*henix.xtriage* indicated that P422 did not contain a twin operator but the L tests suggested twinning [8]. The data were reindexed in the lower symmetry primitive orthorhombic space group and molecular replacement gave a model that was much more realistic. The model had a translational Z score of 11.2. *phenix.xtriage* determined the twinning operator was –h,l,k and the twin fraction was 0.36. The native Patterson map did not indicate any off origin peaks greater than 6% indicating that pseudotranslation was not an issue [9,10]. The self-rotation function, calculated by POLARRFN, showed the expected peaks at κ=180° for twofold rotation but also contained two peaks believed to be the twin operators in $P22_12_1$ (Figure 3-1) [11]. A rotation function was applied in the apparent P422 space group with a translation function in the real space group, but this was unsuccessful in producing a valid molecular replacement model [11,12]. Trials in space group P1 have also been unsuccessful [13]. Crystals with varying *algD* sequences and conditions will be used to

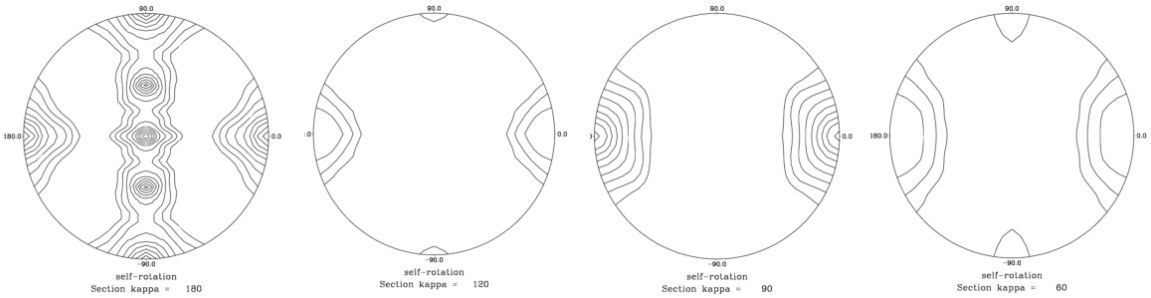

Figure 3-1: **Self-rotation maps**. (a) κ=180°, (b) κ=120°, (c) κ=90°, and (d) κ=60° for AmrZ-*algD* data from 9-4 Å. The plots were generated by POLARRFN [11].

try to grow AmrZ-*algD* crystals without the presence of twinning in order to determine the structure of AmrZ bound as an activator to the *algD* promoter site. The current $2mF_o$-DF_c maps provide evidence that AmrZ only binds as a dimer to the *algD* sequence (Figure 3-2). Although the twinning complications make the statistical refinement difficult to trust, all molecular replacement solutions only yield results with one dimer bound to the 16mer *algD*.

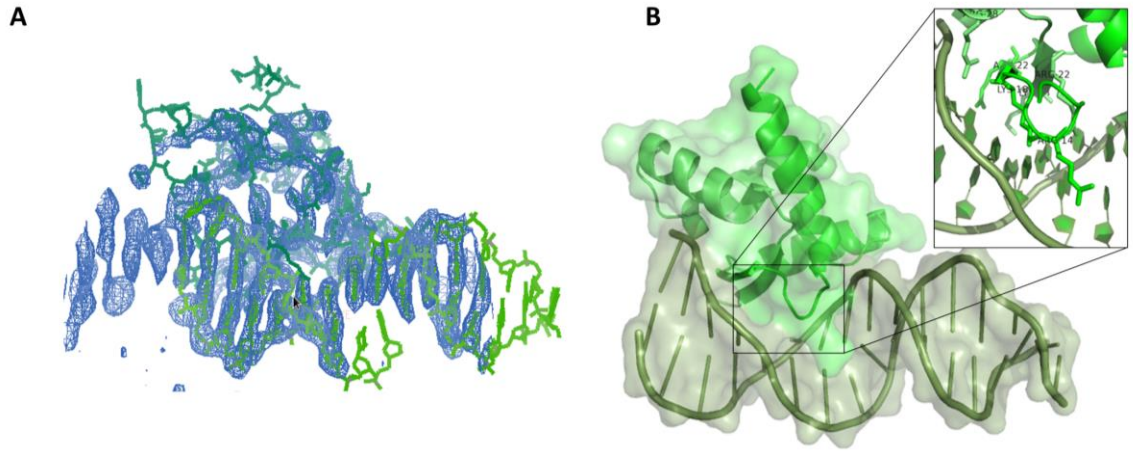

Figure 3-2: **AmrZ bound as a dimer to *algD*.** A) Electron density map ($2mF_o$-DF_c) showing positive electron density for only one AmrZ dimer. B) AmrZ dimer bound to the major groove of *algD* showing the DNA binding residues Arg14, Lys18, and Arg22.

3.3 Materials and Methods

3.3.1 Expression and Purification of AmrZ

A pET-19b expression vector (Novagen), previously described [5], contained an N-terminal polyhistidine tag on the Δ42 truncated AmrZ which only consisted of the first 66 amino acids because the protein failed to crystallize with the C-terminal region present.

The vector was transformed into C41DE3 *Escherichia coli* cells at 310 K overnight in 1L LB broth with 1 mL 100 mg/mL ampicillin until OD_{600} was 0.5. The cells were cooled to 293 K and induced with 1 mM isopropyl β-D-thiogalactopyranoside overnight at 289 K. The cells were harvested by centrifugation at 4,000 rpm for 20 min at 277 K and the pellets were resuspended in lysis buffer (100 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 4 M urea) and lysed with an Emulsiflex-C5 cell homogenizer (Avestin). The solution was centrifuged at 18,000 rpm for 30 min at 277 K to remove cell debris. The supernatant was loaded onto a cobalt metal affinity column (Clontech) which was equilibrated with lysis buffer. AmrZ was recovered with an elution buffer (100 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 500 mM imidazole, 1 M urea). Urea is needed to partially denature AmrZ to increase solubility. PreScission Protease (GE Healthcare) was added and dialyzed at 277 K overnight to remove the polyhistidine tag in a dialyzing buffer (100 mM Bis-Tris pH 5.5, 100 mM NaCl, 5% glycerol, 2 mM DTT, 0.5 mM EDTA). The cleaved AmrZ was applied to a Heparin cation exchange column (GE Healthcare) and was eluted with a 100 mM–1 M NaCl gradient. The fractions containing pure AmrZ were concentrated to 20 mg/mL with a Vivaspin 20 3,000 Da cutoff concentrator (GE Healthcare), flash frozen with liquid nitrogen and stored at 193 K.

3.3.2 Crystallization

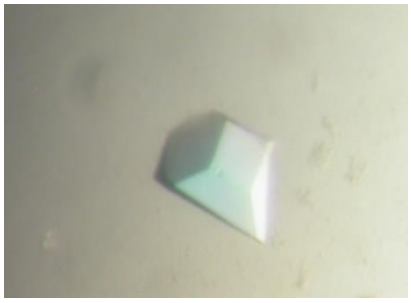

Figure 3-3: **AmrZ crystal.** Crystal of *Pseudomonas aeruginosa* transcription factor AmrZ bound to promoter site *algD*.

AmrZ, at 20 mg/mL, was combined with 50 mM $MgCl_2$ and 5 mM 16mer *algD* DNA in a 1.7:1 protein:DNA ratio and added to a PEG ION screen using the sitting-drop vapour diffusion method with a 60 μL reservoir and a well containing 0.2 μL reservoir

solution and 0.2 µL protein and kept at 289 K. Crystals were optimized in 0.1 M sodium acetate pH 4, 0.1 M NaCl, 3 mM TCEP, 7% PEG 4K in a sitting-drop of 0.6 µL mother liquor and 0.6 µL protein-DNA solution. Crystals grew to about 100 µm in three days and were dipped in a 20% glycerol solution to help with cryoprotection and were vitrified in liquid nitrogen before being shipped to the National Synchrotron Light Source in Brookhaven for data collection (Figure 3-3).

3.3.3 Data Collection and Processing

Four data sets were collected at 1.1 Å on the X25 beamline. The crystal diffracted at a resolution of 3.2 Å and was processed with D*TREK (Figure 3-4) [14]. The data were indexed in the primitive orthorhombic space group with unit cell dimensions of a=88.3, b=106.4, c=106.5. CCP4 and Phenix packages were used for processing [8,11]. *phenix.xtriage* indicated the presence of pseudomerohedral twinning with a twin operator of –h,l,k (α=0.36). Using an AmrZ dimer that has already been structurally determined (3QOQ) as the molecular replacement model, a solution was found with two AmrZ-*algD* complexes per asymmetric unit cell in space group $P22_12_1$ but R_{free} failed to go below 40% upon numerous rounds of refinement.

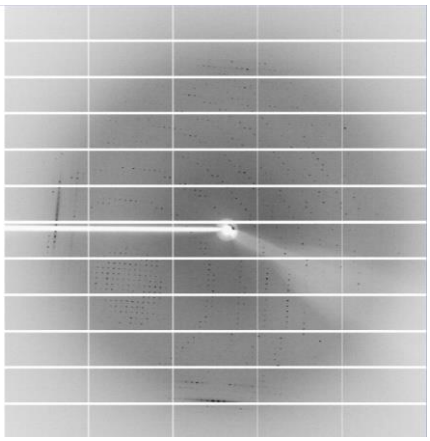

Figure 3-4: **Diffraction pattern.** X-ray diffraction image from AmrZ-*algD* crystal. The crystal diffracted to 3.2 Å resolution and the data was collected on the X25 beamline at NSLS, Brookhaven, NY.

Table 3-1: Data collection and processing
Values for the highest resolution shell are given in parentheses.

Wavelength (Å)	1.1
Space group	$P22_12_1$
a, *b*, *c* (Å)	a=88.3, b=106.4, c=106.5
Mosaicity (°)	1.05
Resolution range (Å)	30-3.2 (3.3-3.2)
Total No. of reflections	108,359
No. of unique reflections	17,082
Completeness (%)	99.7 (99.2)
Redundancy	6.34 (6.34)
$\langle I/\sigma(I)\rangle$	10.69 (3.54)
R_{merge}† (%)	13.6 (24.2)

† $R_{merge}=\sum_{hkl}\sum_i|I_i(hkl)-\langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th observation of reflection *hkl*, and $\langle I(hkl)\rangle$ is the weighted mean of reflection *hkl*.

References

[1] Baynham, P. J., Wozniak, D. J. 1996. Identification and characterization of AlgZ, an AlgT-dependent DNA-binding protein required for *Pseudomonas aeruginosa algD* transcription. *Mol. Microbiol.* 22:97–108.

[2] Baynham, P. J., Brown, A. L., Hall, L. L., Wozniak, D. J. 1999. *Pseudomonas aeruginosa* AlgZ, a ribbon-helix-helix DNA-binding protein, is essential for alginate synthesis and *algD* transcriptional activation. *Mol. Microbiol.* 33:1069–1080.

[3] Baynham, P. J., Ramsey, D. M., Borys, V., Cordonnier, E. M., Wozniak, D. J., Gvozdyev, B. V. 2006. The *Pseudomonas aeruginosa* ribbon-helix-helix DNA-binding protein AlgZ (AmrZ) controls twitching motility and biogenesis of Type IV pili. *J. Bacteriol*. 188:134–140.

[4] Ramsey, D. M., Baynham, P. J., Daniel, J., Wozniak, D. J. 2005. Binding of *Pseudomonas aeruginosa* AlgZ to sites upstream of the *algZ* promoter leads to repression of transcription. *J. Bacteriol*. 197:4430–4443.

[5] Pryor, E. E., Waligora, E. A., Xu, B., Dellos-Nolan, S., Wozniak, D. J., Hollis, T. 2012. Crystallization of *Pseudomonas aeruginosa* AmrZ protein: development of a comprehensive method for obtaining and optimization of protein-DNA cyrstals. *PLoS Pathog.* 8:e1002648.

[6] Waligora, E. A., Ramsey, D. M., Pryor, E. E., Lu, H., Hollis, T., Sloan, G. P., Deora, R., Wozniak, D. J. 2010. AmrZ beta-sheet residues are essential for DNA binding and transcriptional control of *Pseudomonas aeruginosa* virulence genes. *J. Bacteriol*. 192:5390–5401.

[7] Dauter, Z. 2003. Twinned crystals and anamalous phasing. *Acta Crystallogr. Sect. D Biol. Crystallogr*. 59:2004–2016.

[8] Adams, P. D., Afonine, P. V, Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.W., Kapral, G. J., Grosse-Kunstleve, R. W., et al. 2010. *PHENIX:* a

comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.* 66:213–221.

[9] Radaev, S., Agniswamy, J., Sun, P. D. 2009. A case of structure determination using pseudosymmetry. *Acta Crystallogr. D. Biol. Crystallogr*. 65:1334–1340.

[10] Ora, A., Oksanen, E., Kajander, T., Goldman, A., Butcher, S. J. 2010. Crystallization and preliminary crystallographic analysis of mouse peroxiredoxin II with significant pseudosymmetry. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun*. 66:357–360.

[11] Collaborative Computational Project, N. 4 1994. The *CCP4* suite: programs for protein crystallography. *Acta Crystallogr. D. Biol. Crystallogr*. 50:760–763.

[12] Lee, S., Sawaya, M. R., Eisenberg, D. 2003. Structure of superoxide dismutase from *Pyrobaculum aerophilum* presents a challenging case in molecular replacement with multiple molecules, pseudo-symmetry and twinning. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 59:2191–2199.

[13] Zwart, P. H., Grosse-Kunstleve, R. W., Lebedev, A. A, Murshudov, G. N., Adams, P. D. 2008. Surprises and pitfalls arising from (pseudo)symmetry. *Acta Crystallogr. D. Biol. Crystallogr.* 64:99–107.

[14] Pflugrath, J. W. 1999. The finer things in x-ray diffraction data collection. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 55:1718–1725.

Chapter 4:

Crystal structure of RNA-DNA duplex provides insight into conformational changes induced by RNase H

Ryan R. Davis[*], Nadine M. Shaban[*], Fred W. Perrino, Thomas Hollis[+]

Department of Biochemistry, Center for Structural Biology, Wake Forest School of Medicine, Winston-Salem, NC 27157

[*]These authors contributed equally to this work and therefore should be regarded as joint first authors

[+]To whom correspondence should be addressed: Tel: 336-716-0768; Fax: 337-713-1283; E-mail: thollis@wakehealth.edu

Adaptation of:

Davis, R.R., Shaban, N.M., Perrino, F.W., Hollis, T. 2015. Crystal structure of RNA-DNA duplex provides insight into conformation changes induced by RNase H. *Cell Cycle*. 14(4):668-673.

4.1 Introduction

RNA-DNA hybrids exist in all cells and play crucial roles in genomic and mitochondrial DNA replication as well as in transcription. The RNA primers in Okazaki fragments are required for lagging strand synthesis during genomic DNA replication and R-loops are generated in mitochondrial DNA replication [1–5]. RNA-DNA intermediates are formed during transcription and are part of class I retrotransposon element amplifications. They are also a necessary intermediate in retrovirus replication [6,7]. Recently it has been found that ribonucleotides are inserted into DNA during normal replication at a rate of approximately one ribonucleotide for every 7,600 nucleotides added [8,9]. The presence of these single ribonucleotides within DNA act as a strand discrimination signal for mismatch repair of leading strand replication errors [10,11]. Interestingly, removal of these ribonucleotides from DNA is essential for long-term maintenance of genome integrity and development [9,12].

Ribonucleotides within RNA-DNA hybrids are recognized and hydrolyzed by the RNase H enzymes. Three classes of RNase H enzymes exist (for reviews see references 13,14). RNase H1 is a single domain enzyme that cleaves ribonucleotides from DNA but must have at least four ribonucleotides present for activity. RNase H2 recognizes and cleaves a single ribonucleotide within a dsDNA duplex. RNase H3 is structurally similar to RNase H2 but catalytically functions like RNase H1. In humans, mutations in the RNase H2 enzyme lead to the autosomal recessive neurological disorder, Aicardi-Goutières syndrome (AGS), which is a severe autoimmune disorder characterized by the loss of white matter in the brain and increased white cells and cytokine interferon α [15–

19], presumably due to the loss of ability to process ribonucleotides within DNA. All of the ribonucleases have similar catalytic activity and substrate specificity. A nucleophilic attack by water hydrolyzes the phosphate backbone through the assistance of a divalent metal ion and forms a product with a 5'-phosphate and 3'-hydroxyl group [20]. Mg^{2+} and Mn^{2+} are the two most prevalent divalent cations used in catalysis.

Although RNase HI, simply referred to as RNase H, can bind both dsDNA and dsRNA, neither are hydrolyzed by the enzyme. Understanding how RNA-DNA hybrids are recognized and processed by RNase H enzymes is an important step in determining how enzyme dysfunction leads to disease. While structures of RNase H enzyme in complex with RNA-DNA and dsDNA duplexes have been determined [21–26], the recognition requirements for double-stranded RNA-DNA hybrids are still unclear. Relatively few structures of RNA-DNA hybrids alone have been determined, most of which are ten base pairs or less or contain polypurine or polypyrimidine tracts (PPT).

In order to understand structural aspects of the double stranded duplex that are important for RNase H recognition and to determine the structural changes to the nucleic acid structure that are induced upon binding, we have determined the crystal structure of a dodecameric RNA-DNA duplex (Table 4-1). The sequence of this RNA-DNA hybrid is identical to one determined in complex with RNase H enzyme [20] and contains 50% G/C content. Our results show that, like many other crystallographic hybrid structures, the RNA-DNA duplex takes almost solely the A-form parameters but upon binding to RNase H, widens the major groove to widths much closer to B-DNA and decreases the

Table 4-1: X-ray data collection and refinement

Data Collection	
Wavelength (λ) (Å)	1.54
Space group	$P2_1$
Unit cell (a,b,c) (Å)	46.6, 44.5, 83.3
Unit cell (αβγ) (°)	90 105.3 90
Resolution range (Å)	26.3-2.8 (2.9-2.8)
Total Reflections	15,767
Unique Reflections	8,314
Completeness (%)	99.9 (99.9)
R_{merge}	0.086 (0.492)
Mean I/σ(I)	13.4 (2.7)
Average Redundancy	7.3 (7.3)
Refinement	
Resolution Range (Å)	26.3-2.8 (3.0-2.8)
Working set reflections	15,032 (2,989)
Test set reflections	735 (149)
Number of nucleic acid atoms	1,980
R_{cryst}	0.221
R_{free}	0.249
r.m.s.d. bond lengths (Å)	0.008
r.m.s.d. bond angles (Å)	1.4
Average B factors (Å^2)	47.6

Values in parenthesis are for the outermost resolution shell

curvature of the duplex, which may be key steps in the recognition and cleavage of ribonucleotides from DNA [22–24,26]. The RNase H protein contacts the RNA and DNA strands in the minor groove and can differentiate between the A- and B-forms of the duplex as well as the sugar geometry/moiety (ribose or deoxyribose) to ensure specificity for an RNA-DNA hybrid.

4.2 Results

This RNA-DNA hybrid will be referred to as the GAA duplex and Figure 4-1D depicts the numbering system used. The GAA duplex crystallized with four RNA-DNA duplexes in the asymmetric unit. Superimposing the four hybrids indicate that they are structurally similar with a fairly small measured r.m.s.d. between the duplexes ranging from 0.42 Å to 0.75 Å. The differences in structures are primarily due to slight variations in the incline of the bases. The dodecamer RNA-DNA hybrid structure displays a conformation most closely related to A-form RNA and DNA (Figure 4-1). Unlike many

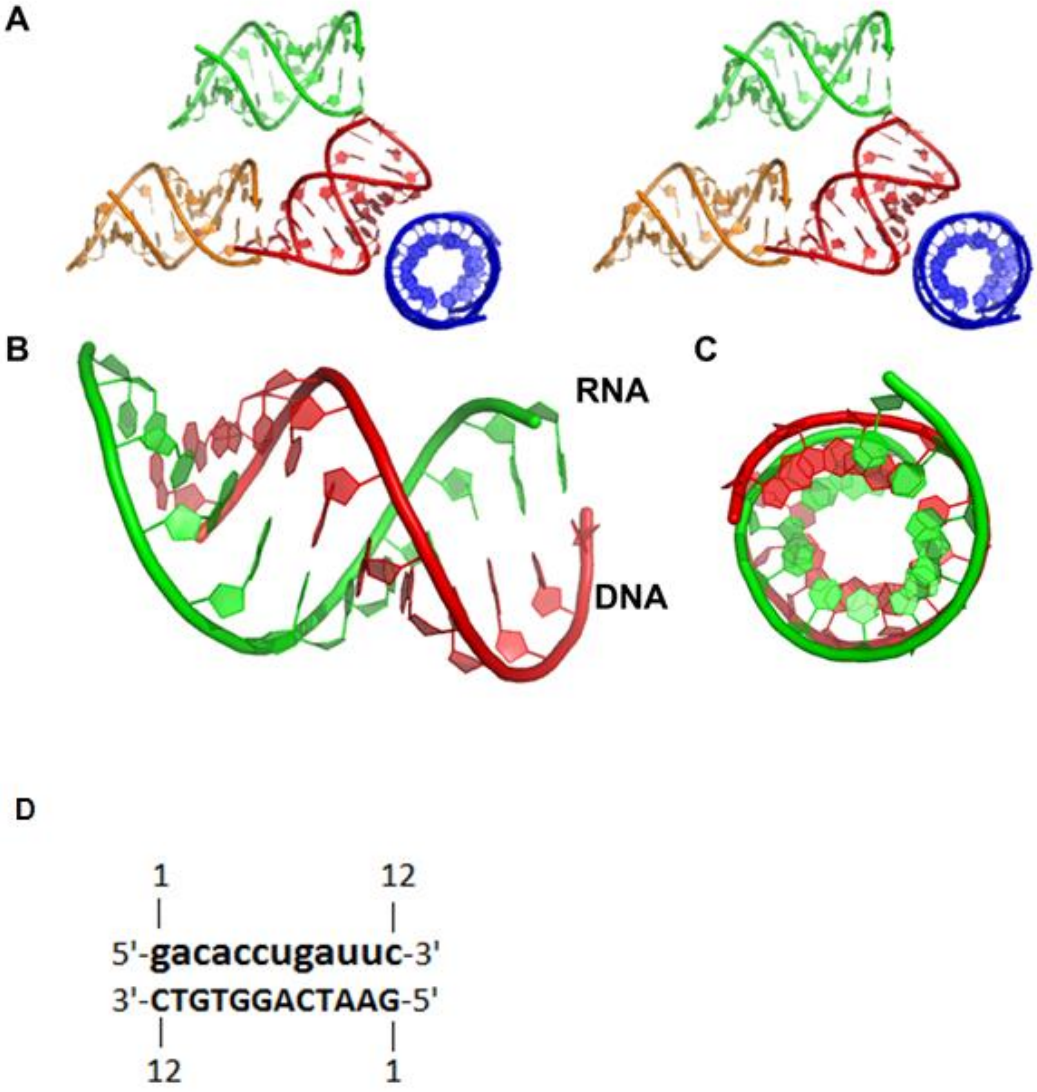

Figure 4-1: **Structure of RNA-DNA hybrid**. A) The stereoimage of the contents in the asymmetric unit shows the packing of four RNA-DNA hybrids with each duplex shown in a different color. B) The RNA-DNA duplex, with RNA strand in green and the DNA strand in red, has a structure most consistent with A-form helices (see Table 4-2). C) Hybrid helix viewed along helical axis. D) The sequence of the RNA-DNA hybrid with the RNA strand in lower case letters and the DNA strand in upper case letters and numbering is shown above the end bases.

structures that have been determined in the absence of an enzyme, this hybrid has a more uniform pyrimidine/purine sequence of 5'-gacaccugauuc-3' and a complementary DNA sequence of 5'-GAATCAGGTGTC-3' compared to PPT hybrids, which are common in

HIV primers and may also contribute to atypical structures of the duplex due to adenine stacking, a-g-a deformations, and "unzipping" caused by unpairing of the bases [21,22,24,25].

4.2.1 RNA-DNA Helical Parameters

The helical parameters of the GAA duplex are almost solely consistent with the A-form helix, which has been also noted in other hybrid duplex structures [27–29]. We compared the parameters of the GAA duplex with typical A-DNA, B-DNA, A-RNA, and RNA-DNA hybrids calculated from structures in the Protein Data Bank (Table 4-2). The data indicate that typical A-DNA and A-RNA have very similar characteristics and that the GAA hybrid tends to share these A-form traits. The structure of GAA duplex reveals an average minor groove width of 9.8 Å and depth of 1.2 Å, comparable with the A-form family. This leads to a narrower and deeper major groove of 4.2 Å and 9.5 Å, respectively. Figure 4-2 shows the changes in the minor groove widths and depths between GAA and the bound duplexes. The x-displacement measures the distance of the

Table 4-2: Average helical parameters of nucleic acid duplexes.

	GAA Duplex[1]	B-DNA[2]	A-DNA[3]	A-RNA[4]	RNase bound to hybrid[5]	RNA-DNA hybrids[6]	PPT hybrids[7]
X-displacement (Å)	-3.6±1.2	0.3±1.0	-4.4±0.7	-3.9±0.8	-3.9±0.8	-4.0±0.6	-3.2±1.8
Bend (°)	17.7	5.7	22.6	13.9	8.7	22.2	17.9
Minor Groove Width (Å)	9.8±0.7	5.4±0.5	10.1±0.8	9.9±0.7	8.6±1.2	9.8±1.2	9.9±0.9
Minor Groove Depth (Å)	1.2±1.3	5.4±0.5	0.7±0.7	0.7±0.8	2.4±1.0	1.0±1.3	1.1±1.4
Major Groove Width (Å)	4.2±2.6	11.3±1.2	4.9±0.5	4.6±2.2	12.5±3.0	3.9±0.6	3.9±0.3
Major Groove Depth (Å)	9.5±1.5	4.7±1.4	11.2±0.4	9.0±0.6	9.4±2.8	10.5±0.8	9.9±0.7

1Average of the four RNA-DNA duplexes in the asymmetric unit

2Average of PDBs: 3U2N, 1BNA

3Average of PDBs: 137D, 240D

4Average of PDBs: 1RNA, 4KYY

5Average of PDBs: 4H8K, 1ZBI

6Average of PDBs: 1D87, 1FIX

7Average of PDBs: 1G4Q, 3SSF

Bases from the center of the helix and the GAA duplex (-3.6 Å) closely resembles that of A-RNA (-3.9 Å) which forms a hole down the axis of the helix (Figure 4-1C). B-DNA has a very small displacement and thus no radial shift of the bases. The curvature of bound RNA-DNA hybrid is lower than the free GAA hybrid (8.7° vs 17.7°) which is likely a result of the binding of the RNase H domain, which is discussed below.

The structure also reveals the sugar geometries have C3'-endo puckers in an anti-conformation for almost all of the nucleotides but a few C1'-exo and C4'-exo sugar puckers are present primarily in the DNA strand which may be due to the rigidity of the RNA backbone forcing the DNA backbone to adopt those conformations. B-DNA typically takes an anti C2'-endo sugar pucker and A-RNA and A-DNA have an anti C3'-endo conformation but at the current resolution subtle nuances are difficult to differentiate. Unlike hybrid structures determined by NMR or duplexes bound to RNase H, the GAA duplex is primarily C3'-endo in both the RNA and DNA strands whereas bound hybrids and those determined by NMR have an RNA strand with a C3'-endo and a DNA strand with C2'-endo, or related, sugar conformation [22–24,30–32].

Figure 4-2: **Comparison of the minor groove widths and depths.** The GAA hybrid (solid lines) has wider minor grooves but they are more shallow than 1ZBI (dashed lines).

To examine how the sequence of the GAA duplex affected the overall structure of the hybrid as compared to previously determined hybrid structures composed of polypurine and polypyrimidine tracks (PPT), two PPT RNA-DNA hybrid structures (PDB: 1G4Q and 3SSF) were used as comparative models with the GAA structure. As shown in Table 4-3, the helical parameters calculated for the other RNA-DNA hybrids are generally similar to the GAA structure with the notable difference that the width of the major groove for GAA is smaller by 0.3 Å than the RNA-DNA and PPT hybrids.

Table 4-3: Helical parameters of RNA-DNA hybrids.

	1G4Q PPT	3SSF PPT	1FIX	1D87	1ZBI Bound	4H8K Bound
X-displacement (Å)	-3.4±1.2	-2.9±1.3	-3.8±0.5	-4.2±0.4	-4.0±0.6	-3.8±0.5
Bend (°)	26.0	9.8	17.0	27.4	7.4	10.0
Minor Groove Width (Å)	9.7±0.7	10.1±0.6	9.7±0.7	9.9±0.9	8.4±0.5	8.8±1.1
Minor Groove Depth (Å)	0.8±1.1	1.4±0.9	1.3±0.8	0.8±1.0	2.6±0.5	2.3±0.9
Major Groove Width (Å)	4.07±0.06	3.6±0.3	4.3	3.4	13.2±1.6	11.8±2.5
Major Groove Depth (Å)	11.13±0.12	8.7±0.7	9.9	11.0	9.6±2.6	9.3±1.2

4.3 Discussion

The structural differences between the bound and unbound hybrid reveal that RNase H interactions unwind the helix to create a local distortion in nucleic acid structure and facilitate protein binding. The free and bound RNase H structures have a difference of 2.0-2.2 Å r.m.s.d. along the Cα backbone, indicating that structural changes will be conferred onto the bound duplex upon binding [20,28,29]. Until now, it has been difficult to determine if RNase H changes the conformation of the RNA-DNA hybrid upon binding, because of the lack of the same nucleic acid structure in both the free and bound state. The free GAA duplex and the hybrid bound to the *Bacillus halodurans* RNase H [20] (PDB 1ZBI), have the same sequence and allow us to make a direct comparison to

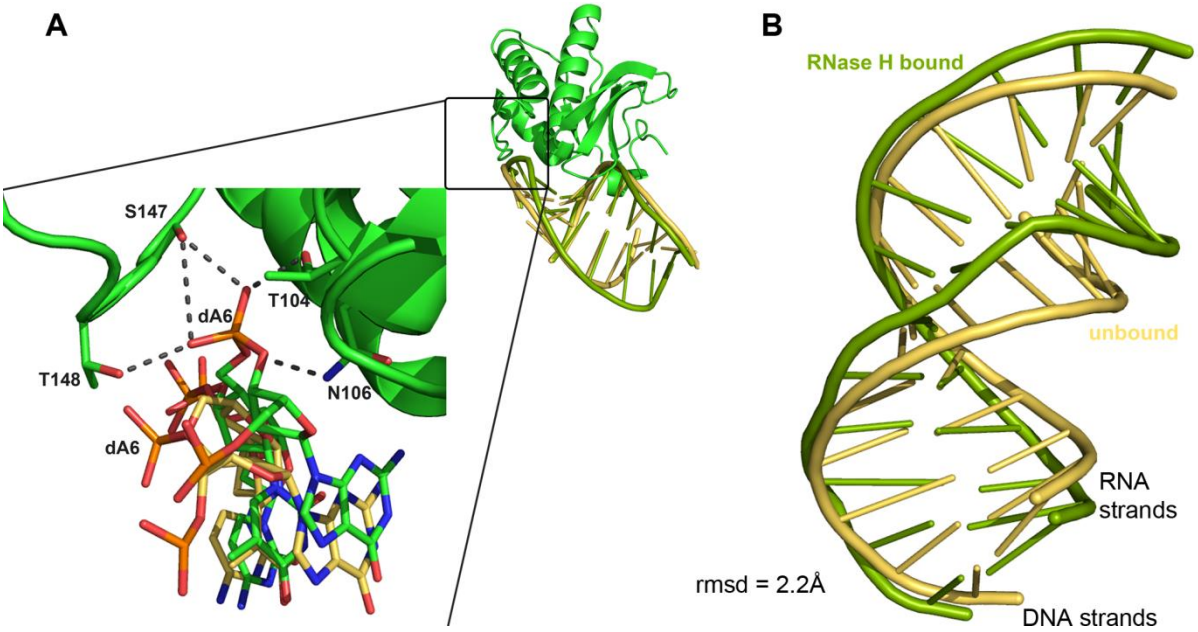

Figure 4-3: **RNase H binding induces conformational changes in RNA-DNA hybrid**. A) The structure of the GAA hybrid (yellow) superimposed onto the hybrid of the same sequence bound to RNase H (green) (PDB: 1ZBI), reveals RNase H distorts the DNA strand of the hybrid by rotating the phosphodiester backbone around nucleotide dA6 about 5 Å into the nucleotide binding pocket. B) The superimposition of the two RNA-DNA hybrids shows the RNA strand has a relatively similar conformation, but the DNA strand has been distorted to significantly widen the major groove of the helix (see Tables 4-2 and 4-3). This flexibility of the DNA strand likely plays a role in substrate recognition and discrimination

address this issue. Superimposing the structure of the GAA hybrid onto the RNase H bound duplex structure reveals dramatic differences between the two (Figure 4-3). The most distinct difference between the RNase H bound duplex and the GAA hybrid is that the protein bound duplex has a smaller minor groove width (8.4 Å vs 9.8 Å) and a significantly larger major groove width (13.2 Å vs 4.2 Å). The GAA structure also shows a significantly larger bend (17.7° vs 7.4°) (Table 4-3). By having a greater bend, the duplex can interact more favorably with RNase H due to an increased exposure of the surface area of the minor groove and preventing water molecules from diminishing hydrogen bonding interactions with the binding pockets [26]. These structural changes are induced by interactions with the protein, most notably the phosphate binding pocket

of RNase H [20], created by residues T104, N106, S147, and T148 (Figure 4-3). RNase H contributes a considerable helical distortion by moving the bound phosphate about 5 Å away from the unbound helical position, resulting in a pinching of the minor groove and widening of the major groove. This leads to a kink in the sugar-phosphate backbone at residue dA6 but the same kink is not seen in the free GAA duplex (Figure 4-3B).

The structure of the hybrid binding domain (HBD) from RNase H in complex with an RNA-DNA hybrid also has the same sequence as the GAA hybrid [20]. The HBD is a small protein domain that preferentially binds RNA-DNA hybrids and functions in protein dimerization and processivity for RNase H [29,33]. The structure shows HBD interacts with the RNA-DNA hybrid solely along the minor groove backbone of the duplex (3BSU), in a manner that is very distinct from RNase H catalytic domains (1ZBI). Interestingly, the duplex bound to the HBD has structural parameters that are similar to the unbound GAA duplex, supporting the idea that binding in the active site of RNase H enzymes induces conformational changes in the RNA-DNA duplex.

It has been proposed that the RNase H enzymes recognize ribonucleotides within dsDNA by the local change in conformation to the A-form having a specific minor groove width of approximately 10 Å, which is slightly less than A-DNA and A-RNA, a pitch and rise similar to A-DNA, curvature different from B-DNA, as well as the presence of a 2'-hydroxyl group on the ribose sugar [20,26,27]. The structure of the GAA duplex indicates that this conformation is likely induced by protein binding. RNases H are not able to cleave dsRNA, suggesting that the added flexibility of a DNA strand is

required for conformational changes to the duplex upon protein binding [34]. This is supported by data showing the flexibility of the DNA strand correlates with the rate of cleavage by RNase H enzymes [13,35]. But the width of the minor groove alone is not enough for ribonucleotide cleavage. It has been demonstrated that the binding of B-DNA to RNase H forces the DNA to have a narrow minor groove but cleavage still does not occur [26]. Another requirement for catalysis to occur could be based on the degree of curvature of the helix. B-DNA is fairly linear with a bend less than 6° but RNA-DNA hybrids have bends greater than 16° but are straightened to about 8° upon binding to RNase H.

In conclusion, distortions created by RNase H binding induce changes in the minor groove width and curvature of the duplex, and can put strain on specific backbone atoms, likely functioning to fine tune binding affinity and substrate discrimination. Similar mechanisms of helical distortion for nucleic acid binding have been seen with a number of other proteins, most notably prokaryotic transcription factors. DNA binding by the bacteriophage λ Cro protein and repressor, as well as the AmrZ protein from *Pseudomonas aeruginosa* facilitates helical distortions in DNA in order to discriminate binding sites [36–39]. RNase H binds nucleic acids independent of sequence but needs to distinguish RNA and DNA polynucleotides inferring that substrate flexibility is likely an important factor. The GAA structure finally allows a direct comparison to be made between a bound and unbound RNA-DNA hybrid to determine the structural aspects of the hybrid that provide specificity for RNase H binding. More hybrids with the same sequence that are bound and free are needed to ensure that the minor groove width,

degree of the duplex bending, and the presence of the rigid 2'-hydroxyl group are the main requirements for substrate recognition and cleavage.

4.4 Materials and Methods

4.4.1 Crystallization and X-ray Data Collection

The RNA and DNA dodecanucleotides (5'-gacaccugauuc-3' and 5'-GAATCAGGTGTC-3') were ordered from Integrated DNA Technologies with desalting purification. Oligonucleotides were annealed in RNase free water with 10 mM MES pH 6.5, 5 mM $MgCl_2$, 20 mM NaCl. Crystals were grown by sitting drop vapour diffusion method at room temperature in conditions containing 20% PEG 3350, 0.2 M magnesium formate, 10% 2-Methyl-2,4-pentanediol. Crystals were frozen in liquid nitrogen prior to data collection. X-ray diffraction data were collected using Cu-Kα radiation on a MicroMax 007 generator and a Saturn 92 CCD detector (Rigaku). Intensity data were processed with the program D*TREK and scaled to 2.8 Å resolution (Table 4-1) [40].

4.4.2 Structure Solution and Refinement

The crystals of the RNA-DNA hybrid grew in space group $P2_1$ with four RNA-DNA hybrids in the asymmetric unit. Phases for the data were obtained by molecular replacement using PHASER [41] and the structure of the RNA-DNA component of the RNase H complex (PDB: 1ZBI) as the search model [20]. The structure was refined with multiple rounds of simulated annealing and composite omit maps (no crystallographic

symmetry averaging) using the program Phenix [42] followed by rounds of model building using Coot [44]. PDB_REDO [43] was used for refinement validation. The model refinement converged at an R_{work}=22.1% and R_{free}=24.9%. The structure has been deposited in the Protein Data Bank with the PDB accession code 4WKJ.

4.4.3 Helical Analysis

Analysis of the nucleic acid structure was performed using the program Curves+ [45,46] as well as on structures from the following PDB files: 3U2N, 1BNA, 137D, 240D, 1RNA, 4KYY, 4H8K, 1ZBI, 1G4Q, 1FIX, 1D87, and 3SSF. The parameters used for Curves+ were wback=2.9 Å and wbase=3.5 Å.

References

[1] Brown T.A., Tkachuk A.N., Clayton D.A. 2008. Native R-loops persist throughout the mouse mitochondrial DNA genome. *J Biol Chem.* 283:36743–51.

[2] Okawa T., Okazaki T. 1980. Discontinuous DNA replication. *Annu Rev Biochem.* 49:421–57.

[3] Rossi M.L., Bambara R.A. 2006. Reconstituted Okazaki fragment processing indicates two pathways of primer removal. *J Biol Chem.* 281:26051–61.

[4] Deininger P., Batzer M. 2002. Mammalian retroelements. *Genome Res.* 12:1455–65.

[5] Clayton D.A., Hughes H., Chase C. 2000. Transcription and replication of mitochondrial DNA. *Hum Reprod.* 15:11–7.

[6] Jern P., Coffin J.M. 2008. Effects of retroviruses on host genome function. *Annu Rev Genet.* 42:709–32.

[7] Temin H.M. 1985. Review reverse transcription in the eukaryotic genome: *Mol Biol Evol.* 2:455–68.

[8] Nick McElhinny S.A., Watts B.E., Kumar D., Watt D.L., Lundström E.B., Burgers P.M., Johansson E., Chabes A., Kunkel T.A. 2010. Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases. *Proc Natl Acad Sci USA.* 107:4949–54.

[9] Nick McElhinny S.A., Kumar D., Clark A.B., Watt D.L., Watts B.E., Lundström E.B., Johansson E., Chabes A., Kunkel T.A. 2010. Genome instability due to ribonucleotide incorporation into DNA. *Nat Chem Biol*. 6:774–81.

[10] Ghodgaonkar M.M., Lazzaro F., Olivera-Pimentel M., Artola-Borán M., Cejka P., Reijns M.A.M., Jackson A.P., Plevani P., Muzi-Falconi M., Jiricny J. 2013. Ribonucleotides misincorporated into DNA act as strand-discrimination signals in eukaryotic mismatch repair. *Mol Cell.* 50:323–32.

[11] Lujan S.A., Williams J.S., Clausen A.R., Clark A.B., Thomas A. 2014. Evidence that ribonucleotides are signals for mismatch repair of leading strand replication errors. *Mol Cell.* 50:437–43.

[12] Reijns M.A.M., Rabe B., Rigby R.E., Mill P., Astell K.R., Lettice L.A., Boyle S., Leitch A., Keighren M., Kilanowski F., et al. 2012. Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. *Cell*. 149:1008–22.

[13] Cerritelli S.M., Crouch R.J. 2009. Ribonuclease H: the enzymes in eukaryotes. *FEBS J*. 276:1494–505.

[14] Tadokoro T., Kanaya S. 2009. Ribonuclease H: molecular diversities, substrate binding domains, and catalytic mechanism of the prokaryotic enzymes. *FEBS J.* 276:1482–93.

[15] Crow Y.J., Black D.N., Ali M., Bond J., Jackson A.P., Lefson M., Michaud J., Roberts E., Stephenson J.B.P., Woods C.G., et al. 2003. Cree encephalitis is allelic with Aicardi-Goutiéres syndrome: implications for the pathogenesis of disorders of interferon alpha metabolism. *J Med Genet*. 40:183–7.

[16] Crow Y.J., Leitch A., Hayward B.E., Garner A., Parmar R., Griffith E., Ali M., Semple C., Aicardi J., Babul-Hirji R., et al. 2006. Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutières syndrome and mimic congenital viral brain infection. *Nat Genet*. 38:910–6.

[17] Crow Y.J., Livingston J.H. 2008. Aicardi-Goutières syndrome: an important Mendelian mimic of congenital infection. *Dev Med Child Neurol.* 50:410–6.

[18] Stephenson J.B.P. 2008. Aicardi-Goutières syndrome (AGS). *Eur J Paediatr Neurol*. 12:355–8.

[19] Rigby R.E., Leitch A., Jackson A.P. 2008. Nucleic acid-mediated inflammatory diseases. *Bioessays*. 30:833–42.

[20] Nowotny M., Gaidamakov S.A., Crouch R.J., Yang W. 2005. Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell*. 121:1005–16.

[21] Gyi J.I., Lane A.N., Conn G.L., Brown T. 1998. Solution structures of DNA:RNA hybrids with purine-rich and pyrimidine-rich strands: comparison with the homologous DNA and RNA duplexes. *Biochemistry*. 37:73–80.

[22] Han G.W., Kopka M.L., Langs D., Sawaya M.R., Dickerson R.E. 2003. Crystal structure of an RNA:DNA hybrid reveals intermolecular intercalation: dimer formation by base-pair swapping. *Proc Natl Acad Sci USA.* 100:9214–9.

[23] Horton N.C., Finzel B.C. 1996. The structure of an RNA/DNA hybrid: a substrate of the ribonuclease activity of HIV-1 reverse transcriptase. *J Mol Biol.* 264:521–33.

[24] Kopka M.L., Lavelle L., Han G.W., Ng H.L., Dickerson R.E. 2003. An unusual sugar conformation in the structure of an RNA/DNA decamer of the polypurine tract may affect recognition by RNase H. *J Mol Biol.* 334:653–65.

[25] Xiong Y., Sundaralingam M. 2000. Crystal structure of a DNA.RNA hybrid duplex with a polypurine RNA r(gaagaagag) and a complementary polypyrimidine DNA d(CTCTTCTTC). *Nucleic Acids Res.* 28:2171–6.

[26] Pallan P.S., Egli M. 2008. Insights into RNA/DNA hybrid recognition and processing by RNase H from the crystal structure of a non-specific enzyme-dsDNA complex. *Cell Cycle*. 7:2562–9.

[27] Rychlik M.P., Chon H., Cerritelli S.M., Klimek P., Crouch R.J., Nowotny M. 2010. Crystal structures of RNase H2 in complex with nucleic acid reveal the mechanism of RNA-DNA junction recognition and cleavage. *Mol Cell.* 40:658–70.

[28] Nowotny M., Gaidamakov S.A., Ghirlando R., Cerritelli S.M., Crouch R.J., Yang W. 2007. Structure of human RNase H1 complexed with an RNA/DNA hybrid: insight into HIV reverse transcription. *Mol Cell.* 28:264–76.

[29] Nowotny M., Cerritelli S.M., Ghirlando R., Gaidamakov S.A., Crouch R.J., Yang W. 2008. Specific recognition of RNA/DNA hybrid and enhancement of human RNase H1 activity by HBD. *EMBO J.* 27:1172–81.

[30] Chou S.H., Flynn P., Reid B. 1989. High-resolution NMR study of a synthetic DNA-RNA hybrid dodecamer containing the consensus pribnow promoter sequence: d(CGTTATAATGCG).r(CGCAUUAUAACG). *Biochemistry*. 28:2435–43.

[31] Hantz E., Larue V., Ladam P., Le Moyec L., Gouyette C., Huynh Dinh T. 2001. Solution conformation of an RNA--DNA hybrid duplex containing a pyrimidine RNA strand and a purine DNA strand. *Int J Biol Macromol*. 28:273–84.

[32] Salazar M., Fedoroff O., Miller J.M., Ribeiro N.S., Reid B. 1993. The DNA strand in DNA.RNA hybrid duplexes is neither B-form nor A-form in solution. *Biochemistry*. 32:4207–15.

[33] Gaidamakov S.A., Gorshkova I.I., Schuck P., Steinbach P.J., Yamada H., Crouch R.J., Cerritelli S.M. 2005. Eukaryotic RNases H1 act processively by interactions through the duplex RNA-binding domain. *Nucleic Acids Res.* 33:2166–75.

[34] Fedoroff O., Salazar M., Reid B. 1993. Structure of a DNA:RNA hybrid duplex why RNase H does not cleave pure RNA. *J Mol Biol.* 233:509–23.

[35] Shiels J.C., Jerkovic B., Baranger A.M., Bolton P.H. 2001. RNA-DNA hybrids containing damaged DNA are substrates for RNase H. *Bioorg Med Chem Let.* 11:2623–6.

[36] Harrison S.C., Anderson J.E., Koudelka G.B., Mondragon A., Subbiah S., Wharton R.P., Wolberger C., Ptashne M. 1988. Recognition of DNA sequences by the repressor of bacteriophage 434. *Biophys Chem.* 29:31–7.

[37] Koudelka G.B., Harbury P., Harrison S.C., Ptashne M. 1988. DNA twisting and the affinity of bacteriophage 434 operator for bacteriophage 434 repressor. *Proc Natl Acad Sci USA.* 85:4633–7.

[38] Pryor E.E., Waligora E.A., Xu B., Dellos-Nolan S., Wozniak D.J., Hollis T. 2012. The transcription factor AmrZ utilizes multiple DNA binding modes to recognize activator and repressor sequences of Pseudomonas aeruginosa virulence genes. *PLoS Pathog.* 8:e1002648.

[39] Koudelka G.B., Harrison S.C., Ptashne M. 1987. Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Lett to Nat.* 326:886–8.

[40] Pflugrath J.W. 1999. The finer things in X-ray diffraction data collection. *Acta Crystallogr Sect D Biol Crystallogr*. 55:1718–25.

[41] McCoy A.J., Grosse-Kunstleve R.W., Adams P.D., Winn M.D., Storoni L.C., Read R.J. 2007. Phaser crystallographic software. *J Appl Crystallogr.* 40:658–74.

[42] Adams P.D., Afonine P.V., Bunkóczi G., Chen V.B., Davis I.W., Echols N., Headd J.J., Hung L.W., Kapral G.J., Grosse-Kunstleve R.W., et al. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*. 66:213–21.

[43] Joosten R.P., Long F., Murshudov G.N., Perrakis A. 2014. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ*. 1:213–20.

[44] Emsley P., Cowtan K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr Sect D Biol Crystallogr.* 60:2126–32.

[45] Lavery R., Moakher M., Maddocks J.H., Petkeviciute D., Zakrzewska K. 2009. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* 37:5917–29.

[46] Blanchet C, Pasi M, Zakrzewska K, Lavery R. 2011. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. Nucleic Acids Res. 39:W68–73.

Chapter 5:

Conclusions

5.1 DNA Structure Recognition

DNA regulation is a crucial component of cellular survivability. The specific characteristics that allow various DNA conformations to exist each serve unique and specialized roles. B-DNA is most commonly associated with aqueous environments with nearly perpendicular bases, ten base pairs per turn, wide major grooves, and narrow minor grooves [1-3]. The separation of the major and minor grooves allows proteins to bind to sequence-specific sites due to the increased exposure from the wide major groove. Minor groove binding proteins will contain positively charged binding residues due to the nearly uniform negative charge from the closely associated phosphate backbones. Unlike B-DNA, A-DNA is mostly formed in dehydrated environments because of the condensed nature of A-form helices [3]. A-form DNA has 11 base pairs per turn and they are tilted to 19° which forms a hole down the center of the helix, not seen in B-form [1,3]. Finally, Z-DNA almost solely exists in high salt conditions which cause ionic interactions with the backbone allowing the DNA to alternate between syn and anti glycosyl bonds and take a zigzag characteristic to allow for a uniform negative electrostatic charge that runs through the narrow minor groove [3,4].

RNA can also hybridize to DNA and serves various roles in DNA maintenance and replication. Failure to remove the ribonucleotides often leads to DNA strand breakage or replication error signals [5-8]. Because the 2'-hydroxyl on the ribonucleotide causes decreased flexibility, RNA-DNA hybrids are forced to take RNA characteristics such as wide minor grooves, narrow major grooves and are even more condensed at 11.6 bases per turn [7].

Proteins are able to recognize the local and global differences among the various forms of DNA whether it is from the shape of the grooves, the charge distribution, or the degree of condensed base pairs. The current hypothesis is that proteins are constantly bound to DNA, in an association-dissociation equilibrium, and scan or slide along the helix until they find a high affinity site for specific binding [9]. This high affinity site can either be from the bases (base readout) or from the helical shape (shape readout). Often times both base and shape readout are needed in high affinity sites. The electrostatic interactions that drive protein sliding are driven by shape readout. Specific amino acids in the protein's binding domain drive which nucleobase sequence the protein will bind to as well as if the protein will bind to the major or minor groove. Arginine, for example, will prevalently bind in narrow minor grooves and/or to guanine [10,11]. A protein's ability to recognize its consensus sequence from the shape of the DNA allows it to bind quicker and more efficiently to the DNA and can expedite a cell's response to its environment.

5.2 RNase HI

Ribonucleases cleave ribonucleotides from DNA. RNase HI can only cleave RNA out of DNA when there are at least four consecutive ribonucleotides. This allows for two ribonucleotides to be on either side of the scissile bond. The ribonucleotides are cleaved on the 5' end to form a 5'-phosphate/3'-hydroxyl product [12]. This cleavage occurs by a phosphate nucleophilic attack from a water molecule with the aid of two divalent cations [12]. RNase HI can bind to RNA-DNA hybrids, dsRNA, and dsDNA but can only cleave hybrids [13-17]. The recognition requirements for RNase H to bind and cleave RNA-

DNA hybrids was unclear which is why an RNA-DNA hybrid with the same sequence to one bound to RNase HI was crystallized and characterized to allow insight into how the ribonuclease can recognize the presence of ribonucleotides within DNA.

Structurally, RNA-DNA hybrids take the A-form conformation, as expected, due to the presence of ribonucleotides and have a wide minor groove, narrow major groove, and tilted base pairs (12°). The RNA-DNA hybrid is bent approximately 17.7° which is linearized upon binding to RNase HI (8.7°). The data indicate that when RNase HI binds, the duplex is unwound slightly causing the major groove to open and the DNA to take a B-form conformation in the area of the binding to put strain on the ribonucleotides, causing a kink in the phosphate backbone where cleavage is to occur. From these data, RNase HI recognizes ribonucleotides in DNA by the adoption of A-form DNA and as well as having a significant bend. Upon binding to this region, the 2'-hydroxyl present on the RNA strand fits into the phosphate binding pocket on the enzyme and the flexible DNA strand is pushed into having B-DNA characteristics, thus putting strain on the ribonucleotides so that a nucleophilic attack can occur (see Chapter 4). Solving the crystal structure of the free RNA-DNA duplex allowed a greater understanding of how ribonucleases, and potentially other nucleases, recognize and bind to their substrate and what conformational changes may be conferred upon the nucleic acids.

5.3 AmrZ

Transcription factors regulate gene expression by either activating or repressing key components required for transcription. After they bind to DNA the transcription factors slide using non-specific interactions with the phosphate backbone until they reach

a high affinity site, at which time specific residues hydrogen bond with specific bases in the major groove [18]. Alginate and motility regulator Z (AmrZ) is a ribbon-helix-helix transcription factor that inserts antiparallel β-sheets into the major groove of the DNA [19]. When bound as a repressor two AmrZ dimers bind to a palindromic sequence with an A-tract linker region in the middle. The minor groove is narrowed and the major grooves are widened when the β-sheets are inserted. On the promoter of *amrZ* there are two AmrZ binding sites distally upstream. This leads to a proposed mechanism that binding of AmrZ to each site allows IHF to bend the DNA so that the C-terminal ends are able to interact and maintain the DNA helices close enough together to sterically hinder RNA polymerase from binding.

AmrZ only has one binding site, upstream of *algD*, to which it binds as an activator. This leads to the hypothesis that only one dimer binds to the promoter; the absence of a palindromic sequence is therefore not surprising, however, the A-tract region is still present indicating that even when only one dimer is bound, the structural characteristics of the A-tract are important for recognition and binding. This leads to a different mechanistic model where AmrZ interacts with RNAP to lower activation energies to increase the rate of transcription of the alginate operon. Empirical evidence will be needed to either confirm or alter these hypotheses.

There are still many unanswered questions that need to be addressed with AmrZ. First, what role does the N-terminal region have? It has been postulated that the N-terminal region has a role in DNA binding but DNA binding analyses have shown that the N-terminal region only has minor contributions to DNA binding. Second, further research is needed to confirm that the C-terminal ends are indeed responsible for higher

order oligomerization. Third, are there additional binding sites on *algD* for AmrZ? Finally, the crystal structure of AmrZ bound as an activator needs to be characterized to determine if the proposed hypothesis of AmrZ only binding as a dimer is correct. If crystals of AmrZ-*algD* cannot be obtained, additional activation promoter sites can be tested.

References

[1] Lu, X.J., Olson, W.K. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*. 31(17):5108-21.

[2] Horton, N., Finzel, B. 1996. The Structure of an RNA/DNA Hybrid: A Substrate of the Ribonuclease Activity of HIV-1 Reverse Transcriptase. *J. Mol. Biol.* 264:521-533.

[3] Huange, D. Phelps, C., Fusco, A., Ghosh, G. 2004. Crystal Structure of a Free κB DNA: Insights into DNA Recognition by Transcription Factor NF-κB. *J. Mol. Biol.* 346:147-160.

[4] Rohs, R., Jin, X., West, S., Joshi, R., Honig, B., Mann, R. 2010. Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem.* 79:233-269.

[5] Wang, A., Fujii, S., van Boom, J., Rich, A. 1982. Molecular Structure of the Octamer d(G-G-C-C-G-G-C-C): Modified A-DNA. *Proc. Natl. Acad. Sci. USA.* 79:3968-3972.

[6] Dickerson, R. 1992. DNA Structure from A to Z. *Methods in Enzymology*. 211:67-111.

[7] Gyi, J., Lane, A., Conn, G., Brown, T. 1998. Solution Structures of DNA-RNA Hybrids with Purine-Rich and Pyrimidine-Rich Strands: Comparison with the Homologous DNA and RNA Duplexes. *Biochemistry*. 37:73-80.

[8] Davis, R.R., Shaban, N.M., Perrino, F.W., Hollis, T. 2015. Crystal structure of RNA-DNA duplex provides insight into conformational changes induced by RNase H. *Cell Cycle*. 14(4):668-673.

[9] Stormo, G., Zhao, Y. 2010. Determining the Specificity of Protein-DNA Interactions. *Nature Reviews*. 11:751-760.

[10] Rohs, R., West, S., Sosinsky, A., Liu, P., Mann, R., Honig, B. 2009. The Role of DNA Shape in Protein-DNA Recognition. *Nature*. 461(7268):1248-1253.

[11] Garvie, C., Wolberger, C. 2001. Recognition of Specific DNA Sequences. *Molecular Cell*. 8:937-946.

[12] Nowotny M., Gaidamakov S.A., Crouch R.J., Yang W. 2005. Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell*. 121:1005–16.

[13] Han G.W., Kopka M.L., Langs D., Sawaya M.R., Dickerson R.E. 2003. Crystal structure of an RNA:DNA hybrid reveals intermolecular intercalation: dimer formation by base-pair swapping. *Proc Natl Acad Sci USA.* 100:9214–9.

[14] Horton N.C., Finzel B.C. 1996. The structure of an RNA/DNA hybrid: a substrate of the ribonuclease activity of HIV-1 reverse transcriptase. *J Mol Biol.* 264:521–33.

[15] Kopka M.L., Lavelle L., Han G.W., Ng H.L., Dickerson R.E. 2003. An unusual sugar conformation in the structure of an RNA/DNA decamer of the polypurine tract may affect recognition by RNase H. *J Mol Biol.* 334:653–65.

[16] Xiong Y., Sundaralingam M. 2000. Crystal structure of a DNA.RNA hybrid duplex with a polypurine RNA r(gaagaagag) and a complementary polypyrimidine DNA d(CTCTTCTTC). *Nucleic Acids Res.* 28:2171–6.

[17] Pallan P.S., Egli M. 2008. Insights into RNA/DNA hybrid recognition and processing by RNase H from the crystal structure of a non-specific enzyme-dsDNA complex. *Cell Cycle*. 7:2562–9.

[18] Blanco, F. and Montoya, G. 2011. Transient DNA/RNA-Protein Interactions. *FEBS Journal*. 278:1643-1650.

[19] Pryor, E., Waligora, E., Xu, B., Dellos-Nolan, S., Wozniak, D., and Hollis, T. 2012. The transcription factor AmrZ utilizes multiple DNA binding modes to recognize activator and repressor sequences of *Pseudomonas aeruginosa* virulence genes. *PLoS Pathogens*. 8(4):e1002648.

Scholastic Vitae

Ryan R. Davis

Personal Information

Birthdate: September 11, 1990

Hometown: Amanda, Ohio

Education

Undergraduate Study: Wittenberg University
Springfield, Ohio
B.S. Biochemistry/Molecular Biology, 2013

Graduate Study: Wake Forest University
Winston-Salem, North Carolina
M.S. Molecular/Cellular Biosciences, 2015

Research Experience

2013-Present: Master's Research: Wake Forest University; Winston-Salem, N.C. Dr. Thomas Hollis, advisor

2010-2012: Undergraduate Research: Wittenberg University; Springfield, OH. Dr. Amil Anderson, advisor

Publications

Davis, R.R.[*], Shaban, N.M.[*], Perrino, F.W., Hollis, T. 2015. Crystal structure of RNA-DNA duplex provides insight into conformational changes induced by RNase H binding. *Cell Cycle.* 14(4):668-673.

Davis, R.R., Hollis, T. Expression, crystallization and preliminary crystallographic analysis of pseudomerohedrally twinned AmrZ-*algD* crystals from *Pseudomonas aeruginosa*. In preparation to be sent to *Acta Crystallographica Section F*.

Presentations

Davis, R.R., Anderson, A.G. "Molecular Dynamics Simulations for Determining the Free Energy Difference in Sugar Binding of the Glucose/Galactose Binding Protein". National Conference of Undergraduate Research (NCUR). Weber State University, Ogden, UT. 2012. *(Poster Presentation)*

Davis, R.R., Muir, A., Mackojc, J., Anderson, A.G. "Free Energy Calculations for the Difference in Sugar Binding of the Glucose/Galactose Binding Protein". Molecular Education and Research Consortium in Undergraduate Computational Chemistry (MERCURY). Bucknell University, Lewisburg, PA. 2011. *(Poster Presentation)*

Davis, R.R., Engdahl, A., Anderson, A.G. "Thermodynamic Free Energy Calculations for the Difference in Sugar Binding of the Glucose/Galactose Binding Protein". Molecular Education and Research Consortium in Undergraduate Computational Chemistry (MERCURY). Hamilton College, Clinton, NY. 2010. *(Poster Presentation)*

Awards

2014	Cystic Fibrosis Foundation Traineeship Grant
2012	Phi Eta Sigma National Honor Society
2011	Wittenberg Board of Director's Grant