Chapter 1:
Introducing the Importance of RNA: a Reader’s Guide to the Dissertation
1.1 RNA: *Sine qua non*

The participation of RNA as a reagent of molecular biology is ubiquitous, arguably unavoidable. Equally ubiquitous and unavoidable is the subject of RNA as a focus of research; there are over a dozen research programs within the University of Rochester investigating the roles of this molecule. The obsession to understand and characterize RNA is equally prevalent beyond the boundaries of this campus. Shortly after the double helix structure of DNA was proposed, George Gamow formed a club with James Watson, Francis Crick, Leslie Orgel, Sydney Brenner, and fifteen other members (one for each amino acid) with the intent of understanding the role of RNA and its relationship to protein synthesis [1]. This club was coined the "RNA Tie Club", and it fostered the communication that lies central to productive inquiry. Rochester has an RNA club as well, and although amino acids are not assigned to each member, the same fostering occurs. In fact, an entire RNA Society, formed in 1993, is dedicated to RNA research. Needless to say, one doesn’t need a fraternity to make valuable contributions to the RNA field, but such associations facilitate fame and funding.

In the course of this introduction, the reader will be introduced to the historical importance of RNA, as described by the "RNA World" hypothesis, as well as its importance to the complexity of modern life. The mysteries of RNA have yet to be fully revealed, a fact attested to by the sheer volume of researchers active in the field. With each year, new discoveries blur the lines of the central dogma of molecular biology (DNA to RNA to protein) [2]; RNA is so much more than a passive intermediate. This dissertation will highlight two specific research areas that challenge the central dogma, namely the role of RNA as a catalyst, and as an agent subject to “editing”, a process which, among other things, diversifies the proteome. It is the author's biased belief that not only was it an “RNA World”, but it remains very much so, today.
1.2 The RNA World

The RNA World hypothesis states that ribonucleic acid (RNA) begat life. In this primitive world, RNA served as both the genetic material and the principal biocatalyst. The idea that RNA was principal in the genesis of life originated in the late 1960s [3], right around the time R.W. Holley, H.G. Khorana and M.W. Nirenberg shared the Nobel Prize in Medicine for their work relating the genetic code to protein synthesis. That is to say, the description of tRNA, rRNA and mRNA rapidly gave rise to the idea that RNA might have initiated life. RNA’s ability to encode genetic information was obvious, given its similarity to DNA. At the time, RNA’s potential for catalysis seemed the only barrier. That barrier was broken in the early 1980s, when S.Altman [4] and T.Cech [5] identified enzymatic RNA molecules, and was obliterated entirely in 1992, when H.F. Noller demonstrated that the enzymatic component of the ribosome was most probably RNA [6]. In the 20 years since then, the RNA community has learned that the functions of RNA are limited primarily by our ability to imagine them. And thus, the dreamer yet thrives in the scientific realm.

The RNA World hypothesis explains a number of paradoxes in the biological realm. For example, protein is required to replicate DNA and yet DNA is required to generate protein. RNA offers a plausible solution to this ‘chicken and egg’ dilemma, as it can serve in both roles. In fact, Howard Temin and David Baltimore received the Nobel accolade for demonstrating that DNA could be synthesized from RNA via the reverse transcriptase enzyme. The RNA World hypothesis also provides a rationale for why DNA is synthesized by reduction of a ribonucleotide [7], and why DNA replication requires an RNA primer [8].

Research investigating the feasibility of the RNA World hypothesis has generated supportive results, although mysteries endure. Nucleosides have been isolated from meteorites, as well as synthesized from proposed ingredients of the primordial soup (reviewed in [9]), and non-enzymatic, templated replication has also been reported [10].
Ancient RNA catalysts have been identified (RNase P and the ribosome are conserved across all kingdoms of life [11]), and new ribozymes have been created in vitro through directed evolution [12]. Together with nucleotide cofactors such as flavin-adenine dinucleotide (FAD) and nicotinamide adenine dinucleotide (NAD), RNA is believed to be competent to sustain primitive metabolic function (see [13] and references therein).

1.3 RNA: the Beta Enzyme

As with the beta version of many software programs, RNA is imperfect. As a ‘jack of all trades’, it masters little. RNA is neither the most efficient catalyst, as will be described here, nor the most stable genetic storehouse, as will be described below. Although there are several examples where fast ribozymes are on par with slow protein enzymes [14], RNA has never been observed to achieve the $10^6$ sec$^{-1}$ rate of catalysis reported for the $k_{cat}$ of carbonic anhydrase protein enzyme [15]. As a reference for comparison, a $k_{cat}$ of 50 sec$^{-1}$ has been reported for peptide elongation by the E.coli ribosome [16].

The poor catalytic rates of RNA have been attributed primarily to its physical properties, beginning with the limited repertoire of functional groups as compared to proteins [17]. Also, with larger building blocks, RNA’s ability for precise sculpting of an active site is somewhat more restricted. The folding pathways of these two types of macromolecules also differ. The hydrophobic core of globular proteins ([18] and references therein) is not present in RNA to the same extent. In fact, RNA molecules exhibit a reduced tendency toward unfolding, but a greater tendency toward mis-folding, as compared to proteins (see [19] for more information). Additionally, RNA nucleobases are neutrally charged at physiological pH with pK$_A$ values at least 2 pH units above or below neutral [20]. In contrast, the selection of protein side groups offers positively and negatively charged groups, including the histidine residue with its pK$_A$ of 6.0 [15]. The imidazole side group is partially protonated at neutral pH and frequently exploited for the purpose of acid / base catalysis [15].
The physical properties and folding pathways of RNA enzymes may be unique, but the mechanisms of rate-enhancement are probably not. Doudna and Lorsch quipped as much in a 2005 review of RNA enzymes, titled “Ribozyme Catalysis: Not Different, Just Worse” [17]. Fourteen years before this, Jeremy Knowles noted that the same principles of rate acceleration were employed for uncatalyzed organic reactions as well as protein enzymes, which are often many fold better, in a review titled, “Enzymes: Not Different, Just Better” [21]. Similarly, a comparison is warranted of the mechanistic preferences exhibited by protein and RNA enzymes. However, the mechanisms of rate-enhancement that are employed by ribozymes are poorly defined at present. The principles of transition-state stabilization and acid / base catalysis have been frequently invoked, but often from different groups about the same ribozyme (for examples, see ref [22] versus [23]). Hence, a consensus is lacking.

Model ribozymes that are amenable to structural and functional experimentation are, therefore, of obvious value. Several families of ribozymes have been identified that are worthy of consideration in this regard, including the self-splicing Group I and Group II introns, RNase P, the ribosome, and the small ribozymes. (Notably, telomerase and the spliceosome will likely join this list in the very near future). These families each perform unique biological functions, and react to give distinct products. One particularly attractive feature of the small ribozymes is their diminutive size, which makes them well suited to a variety of experimental approaches, including the introduction of non-canonical residues via chemical synthesis of the RNA strand.

1.3.1 Small Ribozymes

As more ribozymes are discovered and added to this class, the similarities between members are reduced. The time may come, and soon, when this family warrants further subdivision. In the meantime, small ribozymes are related by the products that they generate. Site-specific cleavage of the phosphodiester backbone of RNA yields a 2',3'-cyclic phosphodiester and free 5'OH [24]. These products are generally achieved through an $S_{N}2$
mechanism, whereby the O2' of the residue preceding the cleavage site (N-1) makes a nucleophilic attack on the phosphorus center [25]. A trigonal bipyramidal intermediate is resolved when the O5' of the next residue (N+1) leaves. The RNA backbone must deform to achieve the in-line geometry between the O2', P, and O5' atoms that is required for this reaction [26] [27]. As it occurs in nature, the reaction is intramolecular, and thus the small ribozymes are not true enzymes, as they are consumed in the course of the reaction. However, most small ribozymes have been engineered to work in trans for the purposes of kinetic characterization, and this caveat won’t be considered further [28].

Members of this family include the hammerhead ribozyme, the hairpin ribozyme, the genomic and antigenomic hepatitis delta virus ribozymes (HDV), the varkud satellite (VS) ribozyme, and the glmS ribozyme. Each ribozyme exhibits a unique global fold and active site architecture. Crystal structures are available all but the VS and antigenomic HDV ribozymes [29-32]. The size of the small ribozymes ranges from 63 nucleotides for the hammerhead, to 150 nucleotides for the VS ribozyme. Notably, the minimal boundaries required for catalysis are not necessarily the same as the wild-type form in the context of the genome. In a few cases, most notably the hammerhead ribozyme, the inclusion of additional neighboring helices improves activity by imparting tertiary stabilization [29, 33]. As another example, the full glmS ribozyme was reported to be 246 nucleotides [34], although the 125 nucleotide construct used in crystallographic studies demonstrated cleavage activity [32]. The glmS ribozyme is the latest novel structure to join the small ribozyme family; the most recent small ribozyme sequence was reported in 2006 when Szostak’s lab isolated an HDV-like sequence from the human CPEB3 gene [35]. The glmS ribozyme is the only family member that requires a cofactor, glucosamine-6-phosphate, for activity [34]. This sensitivity to the presence of a cofactor earns glmS the distinction of being the first metabolite sensing ribozyme. Non-catalytic metabolite sensing riboswitches have also been reported (reviewed in [36]), and one can’t help but wonder if the metabolite sensing ribozymes will constitute a new ribozyme family once more members of this regulatory element are fully characterized.
Several small ribozymes were first isolated from satellite RNAs found in the presence of a host virus. The ribozymes were important for the replication of these satellite RNAs [28]. The first to be identified were the hammerhead [37] and hairpin ribozymes [38], which were isolated from satellite Tobacco RingSpot Virus (sTRSV) RNA. These prototypes were soon joined by the hepatitis delta virus ribozyme [39] and the varkud satellite ribozyme [40]. We will now consider the biological activity of the founding members of this family.

1.3.2 The satellite Tobacco RingSpot Virus (sTRSV) RNA

sTRSV RNA was first identified by Irving R. Schneider in 1969 [41]. This satellite RNA is composed of ~400 nucleotides [42] and is frequently (but not always) found in plants that are infected with one of several strains of the Tobacco Ringspot Virus (TRSV). It is packaged into TRSV virus particles through association with the capsid, but is not infectious or pathogenic and does not generate a message [43]. This satellite RNA has been described as a parasite of the Tobacco Ringspot Virus because its presence reduces the pathogenicity of TRSV in an infected plant [37]. (This is not always the case for satellite RNAs. The Hepatitis Delta Virus, a satellite RNA associated with Hepatitis B virus, leads to increased pathogenicity of the host virus through expression of a single protein product [44].)

Observations of both monomeric and multimeric forms of sTRSV RNA were consistent with a model of rolling circle replication, in which transcription machinery would run multiple laps around a circular template to generate a concatenated multimeric transcript that was subsequently processed into monomer units [45-47]. Within the cell, these monomeric and multimeric forms were observed for both the (+) and (-) strands of the satellite RNA, suggesting processing of both strands [45, 48, 49]. Processing was not observed to depend on protein [42], but benefited from the presence of Mg$^{2+}$ [37]. This 'autolysis' process has been attributed to a specific region within each strand of RNA; a 97 nucleotide segment isolated from the (+) strand RNA [37] came to be known as the hammerhead ribozyme (it has since been minimized to ~60 nt [33]), whereas a 64 nucleotide segment isolated from the (-) strand is now referred to as the hairpin ribozyme [50]. Notably, the hammerhead motif has
since been identified from other sources, including the Avocado Sun Blotch Viroid (ASBV) [28] and Satellite 2 DNA found in the chromosomes of newts [51].

1.3.3 The Hairpin Ribozyme as a Model for RNA Catalysis

The hairpin ribozyme comprises two helix-loop-helix domains that are connected at one end through a four-way junction, which is not required for activity or included in the minimal 64 nt construct [38, 50]. The helix-loop-helix domains fold together in the shape of an “X” and the active site is located where the two loop regions intersect with each other [27]. The name “hairpin” originates from the observation that one strand is continuous through both domains, to the effect that in the folded conformation, it is bent in half and assumes a ‘hairpin’ structure [52]. Folding, or "docking", and activity are most efficient in the presence of divalent metal ions [53, 54]. However, metals do not contribute to the mechanism of catalysis, as evidenced by equal reactivity in the presence of either cobalt hexamine or magnesium [55]. Metal independent catalysis appears to be the norm for the small ribozyme family [14].

The first crystal structure of the hairpin ribozyme was solved by Ferre D’Amare’s group [30, 56] and included the four-way helical junction. Since then, the Wedekind group has developed a robust minimal construct (no four-way junction) and utilized it for a number of high throughput structure / function studies. One aspect of the development will be detailed in Chapter 2. This construct is amenable to crystallization over a wide pH range (5.5 to 8.8) [57-59], and, as it is produced by solid phase synthesis, non-canonical nucleotides are readily introduced. This minimal ribozyme successfully explained the rate enhancement observed in the gain-of-function U39C mutation [58, 60], and identified the binding sites for cobalt hexaamine. The high resolution of the data derived from this construct enabled the identification of active site waters [61], a finding that stimulated new hypotheses about how solvent might be used by ribozymes to achieve rate enhancement. The minimal ribozyme has also been solved with vanadium bound at the active site [62], as was the four-way junctioned ribozyme [56], in an effort to mimic the transition state.
Lastly, this construct has been used to structurally investigate a series of active site ‘point mutations’ that were kinetically characterized by other groups [22, 61, 63]. One such study will be detailed in Chapter 3, in which a series of Ade38 variants were solved. This residue is located in the active site and plays a critical role in catalysis [22]. The role of Ade38 remains unclear (it has been proposed as a general acid or as a means to stabilize the transition state); however, these structures demonstrate that several of the mutants have a propensity to misfold, a finding that enhances the interpretation of the kinetic results and emphasizes the labiality of the RNA tertiary fold.

RNA catalysis will be an active area of research for years to come. However, RNA experimentation has already benefited greatly from the technological advances made in the ‘protein’ years, particularly in the field of crystallography. The same might be said for the biology of RNA itself. The diversity of RNA function in the modern world is more than partly dependent on the contributions of proteins. Next, we turn our attention to one example of how protein has enabled one aspect of RNA’s diversity.

1.4 RNA: the Beta Genome

There are fewer evolutionary improvements in the advance from RNA to DNA as compared to RNA to protein: there are exactly two. Uridine was replaced by thymidine, which enhanced the stability and integrity of the genome. (The frequent cytidine to uridine conversion events could be readily recognized as mutations and corrected by DNA repair enzymes [15].) Also, the 2’OH group was replaced by 2’H, which does not have the nucleophilic capabilities of the 2’OH, as described above. As a result, hydrolysis of the DNA backbone is a fairly rare event, an obvious advantage for the long-term storage of information. However, the reactivity of RNA offers unique advantages as well. RNA is labile and can be readily altered; each nucleotide in an mRNA even comes with its very own nucleophile! RNA is temporary; it has been reported that the mRNA half-life is on the order of
1-2 minutes in prokaryotes [15]. As such, RNA is an excellent molecule by which to ‘tweak’ cellular function, particularly in a temporally or spatially restricted manner. If DNA is the song, RNA is the singer, modulating the message from lullaby to stadium anthem, from bass to soprano.

The human genome has been sequenced [64]. The annotated genes cannot explain the complexity of the human proteome (i.e., tissue-specific protein expression). RNA will, at least this author proposes as much. When non-coding RNAs are annotated, when all the splicing events and message modulation are understood, the complexity of the living system will be encompassed. The opus will be honored.

1.4.1 RNA Editing

Alternative splicing takes one big step toward explaining the gap between genomic and proteomic diversity [65]. RNA editing takes another step in the same direction. RNA editing can be defined as the co- or post-transcriptional alteration of an RNA sequence from what was originally encoded in the DNA. This can occur through the enzymatic modification of an existing nucleobase, or through the substitution, insertion or deletion of nucleotides. Editing introduces new information into the transcriptome, and is thus distinct from alternative splicing or other diversifying mechanisms, which rearrange existing information [66].

Rob Benne and colleagues first discovered RNA editing in Trypanosome mitochondria in 1986 [67]. In this organism, insertion of uridine nucleotides is necessary to establish the correct reading frame in an mRNA. Over the next 10 years, editing events were identified in several tRNAs of both eukaryotes and prokaryotes [68, 69], a vast number of mRNAs found in the mitochondria and chloroplasts of plants [70-73], as well as mammalian nuclear mRNAs that encoded for the apolipoprotein B lipid transporter [74, 75] and the glutamate gated ion channel [76]. These editing events were found to occur through base modification, rather than whole nucleotide substitution or insertion as was observed by Benne [77-80].
Two mammalian editing events occur through deamination, in which an exocyclic amine is replaced by a keto group (reviewed in [81]). This deamination reaction converts adenosine to inosine and cytidine to uridine. Three families of enzymes perform this reaction in the context of nucleic acid polymers, the ADAT (Adenosine Deaminases that Act on tRNA) or TAD (tRNA Adenosine Deaminases) family, the ADAR (Adenosine Deaminases that Act on RNA) family, and what was originally referred to as the CDAR (Cytidine Deaminases that Act on RNA) family [82]. Editing of the glutamate receptor mRNA was attributed to an ADAR enzyme [83], whereas apoB mRNA editing was found to be the action of the first characterized CDAR enzyme, APOBEC-1 (APOlipoprotein B Editing Catalytic subunit – 1) [84]. The CDAR family was renamed upon the observation that most other family members target single-stranded DNA rather than RNA [85, 86]; this family is now referred to as the APOBEC family [87].

Each editing family recognizes a unique substrate structure; ADATs (or TADs) recognize the anticodon loop of tRNA [88], ADARs recognize double-stranded RNA (reviewed in [89]), and APOBEC family members act on single stranded DNA or RNA [90-94]. The three-dimensional structures of representative family members provide a rationale for each type of recognition; these will be described in greater detail in Chapter 4. Deamination performed by the latter two families, the ADARs and APOBECs, can occur in a controlled site specific manner, as was originally observed on the apoB and GluR mRNAs, or in a promiscuous fashion that results in hypermutation of the substrate (reviewed in [89] and [95]). As one might imagine, these two activities have very different functional implications. A select few of these functions will be described here in order to highlight the utility of RNA and DNA editing.
1.4.1.1 The Site-Specific Activity of an ADAT Enzyme on Adenosine34 of a tRNA Anticodon

Adenine 34 is converted to inosine in a number of tRNAs throughout prokaryotes and eukaryotes [96, 97]. As such, this is the most ancient example of deamination editing of RNA. The ancient origin of this enzyme is further supported by the structural analysis of deaminase enzymes that is described in Chapter 4. Among the editing enzymes, the TadA enzyme structure closely mimics the free cytosine and guanine deaminases; this link suggests a potential pathway for evolution of the editing enzymes.

The inosine modification at position 34 was first identified for tRNA<sub>ALA</sub> in yeast [98], and the responsible enzyme is essential for life. Notably, the form of the enzyme varies depending on species; a homodimer performs the editing reaction in <i>E.coli</i> [97], whereas in <i>Saccharomyces cerevisiae</i>, the enzyme is a heterodimer [99]. Regardless, the inosine modification is located at the wobble position of the anticodon and allows a single tRNA to recognize multiple mRNA codon variations. This is because inosine is able to base pair with A, C, or U, as originally postulated by Crick [100]. Thus, inosine at the wobble position streamlines the translation process, as fewer tRNAs are necessary to decode the codons.

1.4.1.2 APOBEC-1 Site-specifically Edits ApoB mRNA

APOBEC-1 deaminates C<sub>6666</sub> of the serum apolipoprotein B (apoB) mRNA transcript to transform a Gln codon (CAA) into a stop codon (UAA) [74, 101]. The resulting truncated protein, ApoB48, is 48% the length of the unedited 512 kD ApoB100 protein [75]. ApoB48 and ApoB100 proteins associate with triglycerides and cholesterol, and are secreted, respectively, from the small intestine as rapidly cleared chylomicrons, and from the human liver as very low density lipoprotein (VLDL) [15]. VLDL that contains the full length ApoB100 is metabolized to low density lipoprotein (LDL), which exhibits a long serum half-life of ~20 hours and is considered an atherosclerotic risk factor [102].
Site specific editing by APOBEC-1 occurs in the nucleus [103, 104] in the context of a large 27S editosome [105]. Minimally, this reaction requires a homodimer of the APOBEC-1 enzyme [84, 106], an mRNA substrate containing the eleven nucleotide ‘mooring sequence’ recognition site [107-109], and the RNA binding cofactor, ACF (APOBEC-1 Complementation Factor) [110]. The cellular localization of ACF is dependent on phosphorylation [111], and efforts to localize the sites of ACF phosphorylation will be described in Appendix 1.

1.4.1.3 APOBEC3G Promiscuously Edits HIV DNA

APOBEC3G was first characterized in 2002 as a determining factor for the characterization of cells as ‘permissive’ or ‘nonpermissive’ to infection by (vif-) HIV virus [112]. ‘Nonpermissive’ cells, when infected by HIV virus from which the viral infectivity protein has been deleted, fail to produce viral particles in their subsequent round of infection [113-116]. Something happens to the HIV virus in nonpermissive ‘producer’ cells that disrupts the infectivity of the packaged virus. Current understanding suggests that the host factor, APOBEC3G, is carried in the viral capsid where it becomes activated in the next infected host cell. Its ‘spacially privileged’ access to the reverse transcription complex allows the APOBEC3G enzyme to hypermutate cytidines within the first strand of DNA synthesized from the HIV RNA genome [86, 117, 118]. The subsequent (+) strand synthesis uses the hypermutated (-) strand as a template, thus having the appearance of catastrophic G to A hypermutations in the pro-viral DNA. These mutations lead to defective virus incapable of subsequent infections. However, normal HIV produces vif (viral infectivity factor), which targets APOBEC3G for ubiquitination, resulting in degradation of APOBEC3G by the proteasome [119-121]. This explains why ‘permissive’ and ‘nonpermissive’ cell types are both susceptible to HIV that contains the vif protein.

1.4.1.4 ADAR Mediated Site-specific Editing Activity of the Hepatitis Delta Virus

Site-specific ADAR activity has been reported on cellular miRNAs and mRNAs. Editing of miRNAs can redirect miRNAs to specific mRNA targets, or alternatively, result in
degradation of the miRNA (reviewed in [122]), thereby suppressing translational inhibition. A to I editing of mRNAs generates protein isoform diversity vital to nervous system function (reviewed in [89]). Editing of HDV is but one example of how ADAR editing can diversify the proteome. The Hepatitis Delta Virus was previously introduced by way of the ribozyme contained within its sequence. This satellite RNA of the Hepatitis B Virus is also a substrate for ADAR editing (reviewed in [123]). Unlike many other satellite RNAs, HDV contains an open reading frame that generates a protein product, the delta antigen (HDAg). Through the action of an ADAR enzyme, this single product is diversified into two products, which are expressed at alternate times during the viral life cycle. These protein products contribute to Hepatitis B pathogenicity; when the HDV satellite RNA is found in association with the Hepatitis B virus it increases the incidence and severity of liver disease observed during infection (reviewed in [44]).

The 195 residue unedited protein product, HDAg-p24, has been observed to enhance the activity of RNA polymerase II [124]. However, when the open reading frame that encodes the HDAg message undergoes ADAR editing of A\textsubscript{1012}, the stop codon (UAG) is converted into a Trp codon (UGG). This extends the open reading frame by an additional 19 residues [125, 126]. A farnesylation signal sequence located within the extended sequence targets the larger delta antigen, HDAg-p27, to the cell membrane [127], where it contributes to viral packaging [128].

Site-specific substrate recognition is performed by ADAR proteins without the assistance of a cofactor [129]. This feat is accomplished via one or more copies of the double-stranded RNA Binding Motif (dsRBMs) located N-terminal of the catalytic domain [89]. Specificity is believed to depend on the local structure of the substrate, specifically, through deviations to the duplex [130, 131]. The presence of loops, bulges, and mismatches appears to direct the editing enzyme to the correct site through something akin to a process of elimination, as the edited adenosine is generally surrounded by a few base pairs of duplexed
RNA rather than within a loop. The HDV editing substrate will be depicted in Chapter 5 where the pursuit of an x-ray crystal structure of this site is described.

1.4.1.5 Promiscuous ADAR Editing of dsRNA in the Nucleus

The functions of promiscuous ADAR editing are not yet well understood. Functions under investigation include modulation of mRNA stability via editing of non-coding sequences, modulation of RNAi activity, RNA-mediated gene silencing, and quality control of mRNAs (reviewed in [132]). ADAR enzymes have been observed to hyperedit up to ~50% of the adenosine bases within a perfectly duplexed dsRNA \textit{in vitro} [133], and editing is most efficient on longer dsRNA duplexes of greater than thirty nucleotides (reviewed in [89] and [132]). The high occurrence of inosine in these transcripts has been observed to recruit specific subsets of proteins, including a complex of p54\textsuperscript{nrb}, PSF, and matrin 3 [134]. It has been hypothesized that these proteins are responsible for retaining hyperedited transcripts in the nucleus, possibly through association with the nuclear membrane [132]. This retention was originally reported for hyperedited transcripts from a polyoma virus model [135]. As a surveillance function this would benefit the cell by preventing translation of mutated transcripts.

ADAR hyperediting has also been proposed as a protective mechanism against retrotransposons. In a characterization of A to I editing sites, 92% of 12,723 identified sites were found within Alu retrotransposon sequences [136]. This suggests that ADAR recognition of dsRNA may help to control the propagation of repeat elements. APOBEC enzymes have also been reported to inhibit the propagation of repeat elements [137-139]. In short, the functional roles of RNA and DNA editing are still being revealed; the next ten years promise to be an exciting time for research in this field.
1.5 Crystallography: United by Methodology

Each chapter of this dissertation involves the methodology of crystallography, regardless of whether the subject is a ribozyme or an editing enzyme. Each chapter focuses on a different aspect of the discipline of x-ray crystallography. Chapter 2 describes the optimization of a crystallization construct of the hairpin ribozyme, and demonstrates how alterations of the construct can influence crystal packing and the resulting diffraction quality. The biological relevance of crystal structures is a frequent topic of discussion; however, obtaining well-diffracting crystals is an under-appreciated task. Crystallization constructs must be biologically relevant, but they must also be amenable to crystal formation. Crystal contacts, (points where two symmetry-related molecules contact each other within a crystal), are generally regarded in an undesirable light when inspecting the final model. These are areas where the structure is more likely to be perturbed in a way that is not biologically relevant. However, these contacts are essential for formation of the lattice. Lattice stability directly influences diffraction quality; diffraction quality directly influences the level of detail and the types of questions that can be reasonably addressed by a final model. The importance of crystal contacts will be demonstrated in Chapter 2.

In Chapter 3, the engineered hairpin ribozyme construct is utilized for a structure / function study. This chapter highlights the utility of a robust construct that is amenable to high-throughput studies. In this study, point mutants that were previously kinetically characterized are solved crystallographically. These results offer one example of the academic conversation that takes place via scientific literature; i.e., "No man is an island, entire of itself" (John Donne). The structures described in Chapter 3 offer explanations for several kinetic observations that might not otherwise be considered, thereby highlighting the cooperative value of combining structural and functional results; the sum is greater than the parts, and each study builds upon the last.
In Chapter 4, the focus is shifted away from the methodology of structure determination and toward comparative analyses of related protein structures. This chapter highlights the phylogenetic aspects of structural biology and the insight that can be gained by placing a protein in the context of its family. Structures of proteins with the 'cytidine-deaminase-like' protein fold were collected from the Protein Data Bank [140] and assembled in this chapter for analysis from a phylogenetic perspective. Comparisons of proteins that exhibit the same three-dimensional fold are used to separate sequence elements that contribute to the core fold from unique regions that are likely to contribute to the function of an individual protein. Primary sequence alignments were then used to extend this information toward family members for which structural information is lacking. These results are valuable to guide future investigations into the modes of oligomerization and substrate recognition that are used by an enzyme. Such an extension is reminiscent of Woodrow Wilson, “I use not only all the brains I have, but all I can borrow.” After all, “it is difficult to begin without borrowing” [141].

Chapter 5 returns to the methods of RNA crystallography to describe the structural investigation of a double-stranded RNA from the Hepatitis Delta Virus that is a substrate for adenosine to inosine editing. This publication emphasizes the pitfalls of poor crystal growth and merohedral twinning, and details the approaches used to overcome such adversity. Appendix 1 contains a publication to which the author contributed a homology model. This model was constructed to aid in the interpretation of results gained from an investigation of ACF phosphorylation sites. Homology models are another example of how existing structural information can be ‘borrowed’. Finally, Appendix 2 provides a list of structures determined in the course of these studies, and the identification codes required to retrieve the x-ray coordinates and structure factor amplitudes from the Protein Data Bank. These depositions have been made so that others may also benefit from ‘borrowing’.
References


Chapter 2:

A *posteriori* Design of Crystal Contacts to Improve the X-ray Diffraction Properties of a Small RNA Enzyme

*The data presented data in this chapter were published in:*


*Co-Authors who contributed to this research are:*

*Spitale, R.C., Krucinska, J., and Wedekind, J.E.*
2.1 Footnotes

Abbreviations: 4WJ, four-way junction; C3L, 3-carbon propyl linker; CHESS, Cornell High Energy Synchrotron Source; JL, junctionless ribozyme; nt, nucleotide; PEG 2K MME, polyethylene glycol 2000 monomethylether; RMSD, root mean square deviation; S9L, 9 atom triethylene glycol linker; U2N, 2'-deoxy, 2'-amino uridine

Data Deposition: Atomic coordinates and structure factors are available from the Protein Data Bank (www.rcsb.org) under PDB IDs: 2OUE (redeposited junctionless structure), 2NPY (S9 minimally junctioned ribozyme), and 2NPZ (C3 minimally junctioned ribozyme).
2.2 Abstract

The hairpin ribozyme is a small catalytic RNA comprising two helix-loop-helix domains linked by a four-way helical junction (4WJ). In its most basic form, each domain can be formed independently and reconstituted without a 4WJ to yield an active enzyme. Production of such minimal, junctionless hairpin ribozymes is achievable by chemical synthesis, which has allowed structures to be determined for numerous nucleotide variants. However, abasic and other destabilizing core modifications hinder crystallization. This investigation describes the use of a dangling 5’ uracil to form an intermolecular U•U mismatch, as well as the use of synthetic linkers to tether the loop A and B domains, including: (i) a 3-carbon propyl linker (C3L) and (ii) a 9-atom tri-ethylene glycol linker (S9L). Both linker constructs demonstrated similar enzymatic activity, but S9L constructs yielded crystals that diffracted to 2.65 Å resolution or better. In contrast, C3L variants diffracted to 3.35 Å and exhibited a 15 Å expansion of the c*-axis. Crystal packing of the C3L construct showed a paucity of 61 contacts, which comprise numerous backbone to 2'-OH hydrogen bonds in junctionless and S9L complexes. Significantly, the crystal packing in minimal structures mimics stabilizing features observed in the 4WJ hairpin ribozyme structure. The results demonstrate how knowledge-based design can be used to improve diffraction and overcome otherwise destabilizing defects.
2.3 Introduction

The hairpin ribozyme is a small, self-cleaving RNA derived from the negative strand of the 359-nt tobacco ringspot virus satellite RNA. In vivo, this ribozyme generates unit-length circular transcripts from concatenated replication intermediates produced by rolling-circle genome replication [1]. As with other naturally occurring small ribozymes such as the hammerhead, hepatitis delta virus, Varkud satellite and metabolite sensing varieties, hairpin ribozyme cleavage generates a free 5'-hydroxyl and a cyclic-2',3' phosphate as cleavage products. The reaction occurs without need of a metal-hydroxide, and leads to site-specific phosphodiester bond cleavage on the order of 0.1 to 0.3 min⁻¹ using minimal hinged constructs that lack a natural four-way helical (4WJ) junction (reviewed in [2]). The active site of the hairpin ribozyme forms at the confluence of two helix-loop-helix domains, called A and B, that dock together through interactions between the respective internal loops. Docking is stabilized by a cross-strand Watson-Crick base-pair between G+1 and C25, as well as a ‘ribose zipper’ motif [3, 4]. In the wild type sequence, the A and B domains are joined distally through a single adenine residue, and this connection is bolstered further by a 4WJ (reviewed in [5]). Although dispensable for the catalytic activity, the 4WJ enhances folding relative to minimally hinged constructs [1, 6].

The global fold and active site architecture of the hairpin ribozyme were revealed first in a 2.4 Å resolution crystal structure (Fig. 2.1a) [3]. With the exception of the substrate strand, this 113-nucleotide construct was generated by in vitro transcription and included both the 4WJ motif that flanks helices 2 and 3, as well as a stem-loop sequence located at the end of helix 4. The latter stem-loop was enlarged and adapted to bind a well-characterized RNA Recognition Motif (RRM) derived from the U1A splicing factor [7]. This protein oligomerizes while bound to specific RNA sequences, thus creating a self-assembling crystallization platform [8].
In an alternative minimalist approach, an all-RNA construct of the hairpin ribozyme was produced by solid phase chemical synthesis. The ease of manipulation of such minimal constructs offered two advantages. First, this approach enabled well-controlled, high-throughput structural investigations designed to elucidate the roles of specific functional groups located in the enzyme active site [9, 10]. Secondly, a posteriori engineering of constructs enabled incremental improvements in x-ray diffraction resolution [9, 10]. Initial efforts were based on the observation that minimal constructs, which included the A14 hinge residue, but lacked the surrounding 4WJ motif, had a tendency to misfold via coaxial stacking of helix 3 on helix 2 [11]. This problem could be overcome by reconstituting the fold from independent loop A and B domains in trans (i.e. a junctionless hairpin ribozyme) [12], and led to the crystallization of a 64-mer construct that diffracted to 3.17 Å resolution [13]. Empirical observations based on the 64-mer lattice led to the design of a 61-nt junctionless construct that diffracted to 2.05 Å resolution [9, 10]. This variant eliminated the A14 residue at the interdomain junction to yield a 5'-dangling U on the opposite strand. This residue was designed to engage in an intermolecular U•U mismatch and thereby promote pseudo-continuous helical stacking [9, 10].

Junctionless ribozymes demonstrate a 104-fold increase in the apparent $K_M$ compared to hinged variants that include A14. The reduced association of these domains has been attributed to an increase in the entropic penalty of docking as well as an enhanced dependence of domain association on the rate of diffusion [12]. These qualities represent potential caveats in the crystallization of junctionless ribozymes that incorporate transition-state mimics or abasic residues, since these modifications have the potential to further reduce interdomain docking and substrate affinity because they require either: $i$) the introduction of breaks into the RNA backbone [14] or $ii$) the potential to form cavities within the ribozyme core [15].

In this study, we describe strategies to improve the interdomain docking properties of minimal, all-RNA hairpin ribozyme crystallization constructs in order to produce variants that
are more resilient to the destabilizing modifications that are central to our structure/function studies. Strategies include the use of 5'-U•U overhanging mismatches and incorporation of flexible interdomain linkers (Fig. 2.1b), whose effects on crystal packing cannot be readily established with solution activity assays. To assess the outcome of this approach, four hairpin ribozyme crystal structures were compared, including: the 4WJ structure (1M5K), the refined junctionless hairpin ribozyme (2OUE), and two new structures harboring a linker at position 14 that included an S9 linker (Fig. 2.1c, 2NPY) or a shorter C3 linker (2NPZ). The results provide a set of knowledge-based principles that can be applied generally to RNA constructs to optimize x-ray diffraction for structure/function studies.

2.4 Materials and Methods

2.4.1 Minimal junctioned all-RNA hairpin ribozyme construct

The 61-nt minimal, junctionless (JL), all-RNA hairpin ribozyme construct was described previously [9]. The sequence of the ribozyme used in the current study (Fig. 2.1b) included the gain-of-function mutation U39C to prevent conformational heterogeneity in the S-turn. Position A14 was substituted with either of two commercially available synthetic linkers (Fig. 2.1b) that permitted the preservation of the standard phosphodiester bond. In the C3L residue, a 3-carbon propyl linker was included as a flexible mimic of the ribose backbone. The S9L residue contained three ethylene glycol subunits (9 atoms), which added 5 more atoms to the backbone compared to a standard nucleotide. Cleavage activity during crystallization was inhibited by modification of the 2’OH nucleophile of A-1 in the substrate strand: a 2’-deoxy A residue was utilized in the C3L structure, whereas a 2’-deoxy- 2’-amino U (U2N) group was incorporated into the S9L structure. The stem-loop sequence present at the end of helix 4 in the natural sequence was removed to promote blunt-ended crystal packing as in other minimal hairpin ribozyme constructs. Strand sequences outside the conserved core (Fig. 2.1b) were optimized to prevent self-complementarity, and the G-C
content of helical ends was enriched to minimize helical end fraying. All strands were synthesized by Dharmacon Inc. (Lafayette, CO) with subsequent deprotection and HPLC purification performed at home as described [16]. The three strands comprising the substrate and ribozyme were mixed together in 10 mM sodium cacodylate pH 6.0; these included a 13-mer substrate, a 19-mer strand comprising the S-turn, and a 29-mer harboring either the C3L or S9L linker. The final ribozyme concentration was 0.3 mM. Docking was promoted by addition of Co(NH$_3$)$_6$Cl$_3$ from a 100 mM stock solution, which was slowly added to the RNA with mixing until reaching a final metal concentration of 1 mM.

2.4.2 Crystallization and X-ray Diffraction Experiments

Crystals grew as hexagonal rods in hanging-drop vapor-diffusion experiments conducted at 293 K containing the following components for the S9L and C3L constructs, respectively: 20 % (w/w) PEG 2K MME, 0.1 M sodium cacodylate pH 6.4 or 6.2, 0.25 M Li$_2$SO$_4$, 2 mM spermidine-HCl and 1 mM Co(NH$_3$)$_6$Cl$_3$. All crystals were cryoprotected by 3 min serial transfers through 4 synthetic mother liquors containing 5 to 18 % (v/v) glycerol. Crystals were captured in thin nylon loops (Hampton Research) and flash cooled in a 100 K stream of nitrogen gas (X-stream, Rigaku/MSC). Diffraction data for the C3L structure were recorded at home on an R-Axis IV image plate system equipped with confocal optics (Osmic). CuK$_\alpha$ x-rays were generated using a Rigaku RUH2R rotating anode operated at 4.5 kW and equipped with a 0.3 mm focal cup. 220 images were collected at a crystal-to-detector distance of 12.5 cm and an exposure time of 25 min per 0.5° oscillation. Diffraction data for the S9L crystal were collected at 100 K on a Quantum210 CCD (ADSC) at the A1 station of the Cornell High Energy Synchrotron Source (CHESS) using x-rays of wavelength 0.977 Å. A total of 300 images were collected at a crystal-to-detector-distance of 24 cm using an exposure of 30 sec per 0.5°. All data were reduced and scaled using Crystal Clear [17]. Intensity and data reduction statistics are provided in Table 2.1.
2.4.3 Structure Determination and Refinement

Each hairpin ribozyme construct of this study crystallized in space group P6\textsubscript{1}22 with a single 61-mer per asymmetric unit. Phases were derived through difference Fourier analysis starting from the highest resolution JL structure available, i.e., PDB ID 1ZFR. Refinement employed conventional methods for RNA as implemented by CNS and O [18, 19]. Rigid body refinement was performed first on the entire structure and subsequently on the individual loop A and B domains, followed by simulated annealing, positional minimization and individual B-factor refinement. Model building employed reduced bias $\sigma_A$-coefficients throughout the refinement process [20].

Parameter and topology files for the C3 linker were adapted from existing ribonucleotide files available in CNS v.1.0. The S9L files were developed with XPLO2D [21-23] based on a compilation of atomic resolution phosphate and ethylene glycol molecules obtained from the Cambridge Structural Database [21-23]. Waters were assigned to structures based on the following criteria: >1 $\sigma$ electron density in 2$mF_o$-$DF_c$ maps, reasonable geometry, hydrogen bond distances between donor and acceptor groups of 2.6 to 3.7 Å, and refined temperature factors comparable to surrounding RNA atoms. One Co(NH\textsubscript{3})\textsubscript{6} molecule, located near G21 and A40 of the S-turn, has been consistently observed in all minimal hairpin ribozyme crystal structures; a second previously characterized site within the major groove of H2 has been modeled here as a water in the C3L and S9L structures due to low occupancy. Sugar pucker restraints were derived from the 2.05 Å resolution 1ZFR structure where appropriate. Refinement statistics are provided in Table 2.1.

To verify the orientation of U-5 and linker residues, simulated annealing omit maps with coefficients ($mF_o$ - $DF_o$) were generated in CNS (Figs. 2.2c, 2.2d & 2.3). Superpositions were generated from conserved residues using LSQKAB as implemented in CCP4 [24, 25]. Solvent accessible surface areas were calculated using GRASP [26]. Helical parameters were analyzed using CURVES and are based on the best curvilinear fit of the helical axis [27]. All figures were generated in PyMol, supplemented by Nuccyl in Fig. 2.1 [28-30].
Coordinates were deposited in the Protein Data Bank under PDB ID codes: 2NPY (S9L), 2NPZ (C3L) and 2OUE (redeposited 2.05 Å 1ZFR JL structure).

### 2.4.4 Activity Assays

Cleavage reactions for the hairpin ribozyme were conducted at 298 K under single-turnover conditions in order to compare the relative activities of the various minimal crystallization constructs of this study. The respective loop A and B strands of the JL ribozyme, or the three strands of the hinged S9L or C3L constructs (Fig. 2.1b), were combined in a reaction buffer comprising 0.10 M Tris-HCl pH 7.5 and 0.20 M NaCl. To assist with pre-reaction folding, the respective loop A and B domains of the JL construct were heated to 343 K for 3 min, and cooled to 298 K over 15 min. The JL hairpin ribozyme reaction was initiated upon addition of 50 mM MgCl$_2$ [12]; the hinged ribozyme reactions were initiated by addition of 12 mM MgCl$_2$ [1, 4]. The starting concentration of ribozyme strand for each assay was 200 µM for all constructs, whereas the concentration of substrate in each reaction was 2.0 µM.

Aliquots of 10 µL were removed from each reaction at time points spanning 4 hours and quenched by addition of 2 volume equivalents of denaturing sample buffer comprising 7 M urea at pH 8.0, without tracking dye. Time points were measured in duplicate from independent assays, and the variance between duplicates was <10%. The cleavage products were heated to 370 K for 1 min then cooled rapidly on ice before separation on 15% polyacrylamide gels containing 7 M urea [31]. The 13-mer substrate strand was pre-labeled with Cy5 at the 5'-end and handled under light restricted conditions. Substrate and product (8-mer) were detected directly within the gel by use of a Storm™ 860 imaging system (GE Healthcare, Inc) operated in red fluorescent mode (662 nm). The fraction of substrate cleaved relative to input substrate was quantified using ImageQuant™ software (Molecular Dynamics). Experimental data from time-dependent cleavage assays were fit to a double exponential equation as described [11, 32, 33]:
\[ F(t) = A_o + A_1(1-e^{-k_1t}) + A_2(1-e^{-k_2t}) \]

where \( A_1 \) and \( A_2 \) represent amplitudes of the biphasic time course, \( k_1 \) and \( k_2 \) are the corresponding first order rate constants of the fast and slow phases, and \( t \) is time. \( A_o \) represents the initial amount cleaved, which was about 3%. Amplitudes and rate constants were estimated by the Marquardt-Levenberg non-linear least squares regression routine (SigmaPlot 9.0). The standard error for fitted parameters was <15%.

### 2.5 Results and Discussion

#### 2.5.1 Significance of and Modeling a U•U mismatch

A common strategy to improve the x-ray diffraction properties of an RNA crystal is to alter the construct itself, rather than subject it to numerous screening conditions [34]. The initial minimal, JL hairpin ribozyme 64-mer was designed without the benefit of a known structure and included residue A14 at the end of helix H2 (Fig. 2.1b). This seemed sensible even after the 4WJ hairpin ribozyme structure was solved since an A14•U-5 Watson-Crick pair was apparent in this structure (Fig. 2.2a). Nonetheless, the diffraction of 64-mer crystals was limited to a nominal 3.17 Å resolution [9, 10] and, upon solving the structure (PDB ID 1X9K) it was apparent that the lattice packing environment necessitated the extrusion of A14 from helix H2 in deference to a symmetry related U-5 (Fig. 2.2b). As such, it was hypothesized that the two U-5 residues, vertically separated by 3 Å along the c-axis, might base-pair across a strict crystallographic dyad if A14 were removed, thereby promoting pseudo-continuous helical packing.

A thermodynamic consideration of the U•U pairing possibility led us to consider the work of Turner & colleagues. Their results revealed that while a dangling 3'-U in the context of a helix improved stability by ~1 kcal/mol, the 5'-variant that we used (Fig. 2.1b) would contribute a negligible 0.1 kcal/mol [35]. This result suggests that any observed advantages in crystallization would not be the result of increased stability of the folded ribozyme in
solution. In contrast, the 5′ overhang optimized the free energy gained through formation of a pseudo-continuous helix. The base-stacking interaction at the helix-helix interface of a sticky-ended coaxial stack has been reported to contribute ~1 kcal/mol more favorability to the free energy than if all nucleotides were connected by a phosphodiester linkage [36, 37]. The rationale for this observation is that a missing linkage in an overhanging base stack allows the nucleotides more freedom to maximize intermolecular interactions and reduce charge repulsion. This provides one explanation for why such pseudo-continuous helices predominate in the crystal packing of nucleic acids, particularly since such interfaces involve two missing backbone linkages. Lastly, the sequence context of the 5′-U•U mismatch (i.e., flanked by 5′-G and 3′-C, Fig. 2.1b) offered the greatest favorable free energy of internal loop formation among those sequences characterized by Kierzek (1999). The latter study also suggested that the U•U mismatch is more favorable in regions where there are fewer helical constraints on its shape, as demonstrated by observed increases in stability when placed near a helix end [38].

Pairings of U•U mismatches have been reported in several conformations that must be considered for modeling into electron density maps. U•U mismatches most commonly exhibit at least one hydrogen bond between an imino group and a keto oxygen, according to a survey of structures reported in the Noncanonical Base-Pair Database [39]. This configuration could be fit readily into our electron density maps with two stipulations. First, the requirement for strict 2-fold symmetry for this interaction assumed static conformational disorder, as supported by the broadened electron density features (Fig. 2.2c). Secondly, an alternate syn conformation modeled for U-5 was consistent with maximization of hydrogen bonding, although this configuration is observed rarely [40]. Nonetheless, a similar syn-anti orientation was reported in the 30S T. thermophilus ribosomal subunit [41]. Care was taken to avoid an O2 to O5′ clash when modeling the syn conformation and the resulting model exhibited both reasonable stereochemistry and a good fit to the electron density (Fig. 2.2c),
although the sugar-base hydrogen bond distances were quite long (3.4 and 4.5 Å) and hence are not depicted.

The *syn-anti* conformer model for U-5 required that uridine simultaneously occupy the *syn* and *anti* conformations with equal occupancy \( q = 0.5 \) due to its location near a crystallographic two-fold axis of symmetry (Fig. 2.2c). However, NMR studies and molecular dynamics simulations each suggested that uridine seldom assumes the *syn* orientation [42, 43], leading us to re-evaluate our original published model (PDB ID 1ZFR). As such, we explored an alternative model in which both bases adopt the *anti* orientation as proposed in two structures of superior resolution with PDB IDs 1OSU and 413D [44, 45]. This conformation, termed the ‘Calcutta’ base-pair by Sundaralingam, exhibited one hydrogen bond between N3 and O4 and one between O4 and C5 that presumably involved the hydrogen of the latter atom. This model provided a slightly improved fit to the electron density of the current study, although a 4.2 Å distance between O4 and C5 ruled out a second hydrogen bond in the JL structure (Fig. 2.2d). Parallel refinements of both possible U•U pairs were conducted (Fig. 2.2c versus 2.2d) in which \( R \) values were compared after identical rounds of positional and individual \( B \)-factor minimization (holding weighting factors constant in CNS). The Calcutta conformation yielded \( R_{\text{cryst}}/R_{\text{free}} \) values 0.13/0.08% lower than the *syn-anti* combination. This observation and the paucity of known *syn-anti* conformers suggested that the Calcutta conformation was likely to be a more accurate representation of the data. As such, new coordinates of the JL hairpin ribozyme at 2.05 Å resolution (Table 2.1) were redeposited in the PDB with ID 2OUE, and the hinged structures of this study were also refined with the Calcutta conformation. This result emphasizes that crystallographic contacts and alternative conformations must be modeled carefully and provides an important precedent for the use of databases to assist in the assignment of unusual base-pairing conformations during model building.
2.5.2 Design and Crystallization of Synthetically Hinged Hairpin Ribozymes

The limitations of our JL hairpin ribozyme construct became apparent during attempts to crystallize the hairpin ribozyme with transition-state analogues or abasic substitutions in the active site. In attempts to capture vanadium oxide as a transition-state mimic between A-1 and G+1 (Fig. 2.1b), it was necessary to introduce a break into the cleavage site of the substrate strand. As such, the four-stranded JL construct became five strands plus a vanadium ligand. This six component complex failed to crystallize, presumably due to improper folding and interdomain docking. Thus, reconnecting the loop A and B domains was a logical step towards reducing the number of RNA strands. This effort proceeded with the stipulation that constructs incorporating the interdomain connection remained small enough for efficient solid-phase chemical synthesis and were amenable to crystallization (and high-quality x-ray diffraction) in the lattice obtained previously.

The native A14 hinge residue found in the 4WJ structure is base-paired to U-5 and further stabilized by the surrounding 4WJ [3]. However, the intermolecular U•U mismatch formed in the JL crystal packing scheme precludes re-introducing a base at the hinge position. Additionally, natural nucleotide linkers adjoining H3 and H2 were observed to produce coaxial helical stacks in other minimal constructs [11]. Taken together with the observation that lattice formation by RNA favors pseudo-continuous helical packing [46], the presence of a natural nucleotide linker during crystallization was deemed likely to exacerbate improper docking between the loop A and loop B domains. It was therefore hypothesized that tethering the loop A and B domains with flexible synthetic linkers (or spacers) would introduce sufficient conformational freedom at the helical interface to dissuade unfavorable H2-to-H3 end-to-end stacking with minimal disruption to the existing crystal lattice. Similar modifications using non-nucleotide linkers were employed in single-molecule FRET studies. Specifically, when a C3L spacer (Fig. 2.1b) was incorporated between A14 and A15 of a minimal hinged hairpin ribozyme, a 35-fold improvement in the interdomain docking rate constant was observed relative to the wild type sequence [33]. The apparent favorability of
this substitution was tempered by increased docking heterogeneity and a 25-fold slower bond breaking step (i.e., \( k_{\text{cleave}} \)). However, this construct still exhibited overall cleavage rates nearly double that of the A14 linkage alone. In light of these data, we chose to substitute A14 with the S9L spacer (Fig. 2.1b), yielding a ribozyme with the same number of backbone atoms as that used in the FRET investigation, but with no adenine base present to interfere with the U•U mismatch. Similarly, we mimicked the native linkage with the C3L spacer, a substitution that preserved the native backbone in atom number and type but offered increased flexibility since it lacked the furanose ring.

Cleavage assays were conducted to compare relative activities of the S9L, C3L, and JL constructs prior to crystallization. The hinged constructs exhibited similar cleavage profiles (Fig. 2.3) that fit best to a biphasic double exponential equation as described previously [11, 32, 33]. S9 and C3 linked constructs exhibited total amplitudes \( (A_1 + A_2) \) of 81% and 78%, respectively, for cleavage of a 13-mer substrate and showed similar first order rate constants for the fast phase, where \( k_1 = 0.390 \text{ min}^{-1} \pm 0.055 \) and \( k_1 = 0.325 \text{ min}^{-1} \pm 0.080 \). The slow phase, described by \( k_2 \), displayed rates of \( 0.041 \text{ min}^{-1} \pm 0.007 \) and \( 0.024 \text{ min}^{-1} \pm 0.005 \). For the purposes of this study, we believe these values are indiscernible. Likewise, the JL construct showed a fast-phase first order rate constant, \( k_1 \), of \( 0.429 \text{ min}^{-1} \pm 0.150 \), and a slow phase with \( k_2 = 0.014 \text{ min}^{-1} \pm 0.001 \). The \( t_{1/2} \) for the fast phase of each construct was \( \sim 2 \text{ min} \), whereas the slow phases displayed \( t_{1/2} \) values of 17 min (S9L), 29 min (C3L) and 50 min (JL), respectively. Furthermore, the JL construct cleaved only 54% \( (A_1 + A_2) \) of the input substrate. The results suggested that each ribozyme construct could adopt a catalytically competent fold, as demonstrated by the comparable rate constants in the fast phase. However, greater populations of hinged hairpin ribozymes docked productively relative to the JL construct based on the total fraction that it cleaved \( (A_1 + A_2) \). Significantly, the conformation needed for solution activity appeared equally accessible by both the C3L and S9L variants, which has implications for their use in the development of hinged constructs for crystallization.
Despite the kinetic similarities between the hinged ribozymes, constructs harboring the S9L spacer exhibited superior crystal growth relative to C3L constructs. The latter crystals displayed multiple morphologies, with the best samples exhibiting a hexagonal habit that reached a size of 0.2 mm x 0.2 mm x 0.15 mm. These crystals were harvested after 3 weeks and diffracted x-rays to a maximum resolution of 3.35 Å. S9L crystals grew much more consistently in a hexagonal habit that reached dimensions up to 0.4 mm x 0.3 mm x 0.6 mm within 2-3 weeks. The initial S9L crystal used for this investigation diffracted to 2.65 Å resolution, although comparable examples have been refined since to resolutions as high as 2.05 Å. Both S9L and C3L constructs crystallized in space group \( P6_122 \), although the C3L unit cell was longer by 15 Å along the \( c \)-axis, which equates to a difference of 2.5 Å per asymmetric unit. Crystals of the minimal, JL 61-mer hairpin ribozyme have been described elsewhere [9, 10]. In general, JL crystals require higher concentrations of RNA for growth, and reach a maximum size of 0.25 mm x 0.25 mm x 0.35 mm in 1-3 months.

2.5.3 Structural Comparison of C3L and S9L Hinged Hairpin Ribozymes

2.5.3.1 Overall Model Quality

The quality of the models is indicated by the observation that both linker structures fit well to electron density maps and refined with reasonable geometric parameters (Table 2.1 and Figs. 2.4a & b). However, the C3L structure was inferior in several respects. Most notably, the resolution of x-ray diffraction and agreement of reflections measured in multiplicity were poorer than the JL and S9L structures (Table 2.1). The 3.35 Å resolution C3L structure (PDB ID 2NPZ) refined to an \( R_{\text{cryst}} \) of 26% and an \( R_{\text{free}} \) of 27%, as compared to the 2.65 Å resolution S9L structure (PDB ID 2NPY), which refined to an \( R_{\text{cryst}} \) of 20% and an \( R_{\text{free}} \) of 22%. Coordinate errors estimated from cross-validated \( \sigma_A \)-weighting were 0.67 Å and 0.49 Å, respectively. Electron density maps were contiguous for all regions of each structure with the exception of the C3L residue itself, whose central carbon could not be observed above the three sigma contour level in an \( mF_o-DF_c \) omit electron density map (Fig. 2.4a
versus 2.4b). Taking this into account, the conformation of the C3 linker was built according to the positions of the flanking phosphorus atoms and stereochemical constraints.

Helix 1 was the region of each hinged structure with the poorest quality electron density. This trend has been documented in all crystal structures of the minimal JL hairpin ribozyme [9, 10], and thus is not a consequence of the linkers. Increasing the G-C content in this region minimized the helical fraying as compared to the initial 1X9K 64-mer structure, but temperature factors for this region remained high. The most likely reason for this is the scarcity of crystal packing contacts in this region. This is evident by an examination of H1 of the 4WJ structure, which does not exhibit the same difficulties. Although it is engaged in a similar end-to-end base-stacking crystal contact as the minimal construct, H1 of the 4WJ structure is stabilized further by a modest number of packing interactions along the exposed length of the helix [14].

An all-atom superposition of the C3L and S9L hinged structures revealed an overall RMSD of 0.77 Å. The A and B domains exhibited similar molecular dimensions and their active site residues overlaid well. Both structures displayed nearly the same ‘pre-catalytic’ active site conformation described previously for the minimal JL 61-mer [10]. Deviations between the two structures were more pronounced at the helical ends, which are areas where the minimal hairpin ribozyme engages in few tertiary contacts, thus rendering them more susceptible to crystal packing forces. The H2 stem of the S9L structure formed an intermolecular U•U mismatch (Fig. 2.2d) that was stabilized by a 2.7 Å hydrogen bond between N3 and O2 and a 3.3 Å hydrogen bond between O4 and C5, as reported in similar U-U pairs [44, 45]. The equivalent distances for the C3L model were longer (3.3 and 3.7 Å), but were also subject to greater coordinate error (Table 2.1). In light of the 4.2 Å distance exhibited by the JL structure for the second hydrogen bond (discussed above), it seems plausible that steric effects attributable to the linker favorably influenced the spatial proximity of the symmetry-related U-5, leading to the improved hydrogen bonding distances observed here.
2.5.3.2 Contrasting Features of the S9L and C3L Inter-domain Hinges

A comparison of the distance between the O3' atoms of G13 and G15, which flanked the inter-domain junction, revealed significant differences between the two constructs. As anticipated, this distance was longer for the S9 linker (12.7 Å) relative to the C3 linker (10.6 Å) (Figs. 1.4a & b), due to the lengths of the respective linkers. An all-atom superposition of these hinged ribozymes suggested that structural differences arising from variable linker lengths were not distributed equally along the lengths of the newly tethered helices (H2 and H3) (Fig. 2.4c, note ribose superposition of residue 13 versus 16). Rather, structural disparities localized mostly to the top of H3, as observed by the 2.8 Å displacement between the G15-C49 base pairs of each hinged structure (Figs. 2.4d & e). The basis of this localized change can be explained by the observation that the terminal base pairs of H2 and H3 were oriented nearly perpendicular to each other (Fig. 2.1c). G13 was buried further into the core of the loop B domain, whereas G15 was more solvent exposed. Thus, G15 was better suited to adapt to the structural influences of the linker. The 2.8 Å movement of the terminal H3 base pair in the C3L structure toward the center-of-mass of the hairpin ribozyme increased the local twist of H3 to 38˚, as compared to 32˚ for the S9L, effectively narrowing the end of the C3L variant. This twist was alleviated gradually as H3 progressed toward the internal loop of the B domain.

Contraction of the C3L H3 helical end affected H4' of the symmetry related molecule, due to pseudo-continuous helical crystal packing interactions (Fig. 2.4d). The terminal base pair of H4' (A31'-U31') experienced a modest lateral movement similar to, but less than, that observed for H3. This is illustrated in a comparison of the distances between the C3' of C49 and the O4' of U31' (4 Å in S9L versus 2.9 Å in C3L) (Figs. 2.4d versus e). Because the lateral shift exhibited by the A31'-U31' base-pair in the C3L structure was not fully commensurate with that of H3, it was necessarily accompanied by a 1.5 Å upward shift of U31' (Fig. 2.4f). This upward movement circumvented an otherwise inescapable steric clash between the ribose moieties of U31' and C49. The net result was a more flush base stack
between H3 and H4’ in the C3L structure, as compared to the more staggered interaction observed in the S9L (Figs. 2.4d & 2.4e) and JL structure (not shown). From an engineering and design perspective, there are two opposing factors that appeared to influence this end-to-end stacking interaction. The first was the local influence of the shortened C3 linker, which drew the G15-C49 base-pair closer to the centre-of-mass of the RNA, effectively improving the crystallographic base stacking interactions. The second factor originated from crystal contacts observed in both the S9L and JL hairpin ribozyme structures. In these two structures the terminal residues of both H3 and H4’ engaged in backbone and minor groove hydrogen bonds that conferred stability along the 6₁-fold screw axis (Fig. 2.5a). These interactions supported the more staggered helical packing interaction, but were notably absent from the C3L structure, which is relevant to understanding the principles of RNA that influence high-resolution x-ray diffraction.

2.5.3.3 Variations in Crystal Packing Interactions within the Minimal Constructs

All termini of the three minimal hairpin ribozyme structures discussed here (Table 2.1) engaged in end-to-end helical base stacking. These interactions sustained the 2-fold and 2₁ axes present in space group P6₁22. The base-stacks observed for H1, H3 and H4 were blunt-ended, whereas the U•U mismatch of H2 was sticky. Each pseudo-helical packing interaction buried approximately 400 Å² of hydrophobic surface. The terminus of H1 and H2 each stacked self-to-self on dyad axes; the self-to-self (i.e., H1 to H1’) nature of this interaction offered little support to the mostly solvent-exposed length of H1 and may have contributed to the disorder observed in this helix. In contrast, H3 and H4’ stacked onto each other in support of the 2₁ symmetry axis. In this manner, linker influences originating in H3 were propagated into H4’ of the symmetry related molecule.

Minor groove and backbone interactions contributed to the formation of the 6₁ screw axis (Fig. 2.5). The backbone of H4 of a ‘reference’ molecule (Fig. 2.5a, dark molecule)
bridged the blunt stack formed by helices H3’ and H4’, which were related by a crystallographic 2₁-symmetry axis (Figs. 2.5b & c). According to the symmetry operators of space group P6₁22, H4 of the reference molecule formed a blunt-end stack to symmetry related molecule H3”’, which was comparable to the H3’ interaction with H4’ (Fig. 2.5b or c). If all helices were included at this junction, the rudimentary shape would be a plus sign, ‘+’.

[Note; H3”’ was omitted for clarity, thus resulting in the letter ‘T’ in Figs. 2.5b & c]. The cross was formed by two base-stacked smaller helices that formed a single pseudo-continuous helix. These longer helices contributed the respective horizontal and vertical cross components.

The intersection of these components was comprised of minor groove and backbone contributions of varied extent that depended upon the choice of linker (Figs. 2.5b & c). At the intersection of these helices, H4 buried 1200 Å² of surface through interactions with three of the six ribozymes that contacted the asymmetric unit. The minor groove/backbone component of this interaction within the S9L structure buried 800 Å², within which there were 11 hydrogen bonds (Figs. 1.5d-g). In contrast, the comparable interaction of H4 in the C3L structure buried only 700 Å². While the 2₁ base stacks were equivalent between the two structures (400 Å²), the minor groove contacts were largely absent in the C3L structure, giving rise to a large cleft (Fig. 2.5c). Significantly, no stabilizing hydrogen bonds were identified within the 300 Å² of buried surface that constituted the minor groove symmetry contact in the C3L structure (Figs. 2.5h-k). The bridged lattice packing interaction observed in the S9L structure was also lacking in the C3L ribozyme, such that H4 made contact with only H3’, rather than with both H3’ and H4’ as in the S9L structure (Fig. 2.5b versus c). The packing deficits in C3L structure undoubtedly contributed to its poor diffraction properties relative to the S9L variant.

The paucity of crystallographic packing interactions along the 6₁ axis of the C3L structure appeared directly attributable to the choice of inter-domain linker. The shift of H3 toward the RNA centre-of-mass of the asymmetric unit moved it away from the minor groove
interaction with symmetry related helices H3´ and H4´ (Fig. 2.5c versus b). The ensuing loss of hydrogen bond interactions and changes in shape complementarity between intermolecular surfaces appeared to destabilize the 6₁-fold packing scheme (Figs. 2.5h-k), thus contributing to the elongated c-axis (Fig. 2.5a). These observations suggest that base-stacking interactions promote formation of the crystal lattice, but the 2'-OH mediated intermolecular hydrogen bond contacts dictate stability of the lattice. Thus, the C3L lattice exhibited sufficient flexibility to maintain pseudo-continuous helical packing at all 4 termini, but the 6₁-fold minor groove interactions were necessarily sacrificed. Although a number of small molecule compounds from the crystallization medium, such as spermidine or glycerol, could theoretically span the gap between H4 and H4´ of C3L, no such ligands from the crystallization/cryoprotection medium were observed in 2mFo-DFc or mFo-DFc electron density maps, which has implications for the use of such additives to stabilize an RNA lattice.

2.5.4 Comparison of Four Hairpin Ribozyme Structures with Different Interdomain Linkages

2.5.4.1 Local Comparison

Four alternatively hinged ribozyme structures are now available, offering the first opportunity to examine the impact of the hinge region on the overall RNA fold. The S9L and C3L structures of this investigation are the first to link the A and B domains without the addition of the remaining two helices of the four-way helical junction. As such, they provide a means to understand the structural influence of the hairpin ribozyme interdomain linkage (hinge) outside the context of the four-way helical junction. This comparison also has implications for the use of synthetic linkers to produce crystal contacts in RNA constructs, as well as the inclusion of crystal contacts in searches aimed at identifying potentially biologically relevant motifs that promote tertiary or quaternary RNA folding.

A superposition of all four hairpin ribozyme variants demonstrated excellent agreement between residues at the active site. The pairwise RMSD values between the 4WJ
structure and the C3L versus S9L structures were 1.5 Å and 1.6 Å, respectively. A closer inspection of the local differences indicated that the C3 linker most closely resembled the 4WJ A14 linkage (Fig. 2.4c). This result seems reasonable since each of these linkages possessed three carbon atoms between the phosphate groups at positions 14 and 15. The G13 O3’ to G15 O3’ distances clustered accordingly (Figs. 2.4a & b and described above); this distance in the 4WJ structure was measured as 11.1 Å (10.6 Å for C3L) and the JL distance was 12.7 Å (also 12.7 Å for S9L).

Local twist values for G15 at the top of H3 showed the same trend; i.e., the C3L and 4WJ structures exhibited increased twist (38˚ and 48˚), whereas the S9L and JL structures were more relaxed (31.9˚ and 31.6˚). For reference, standard A-form RNA exhibits a twist of 32.7˚ [47]. While the overwinding of H3 in the C3L and 4WJ structures was likely to result from the shortened linker, the mildly under-wound twist values observed for the S9L and JL structures may reflect crystal packing. This is because twist exhibits a linear dependence on minor groove width, as calculated by CURVES [48]. The 6-fold packing scheme of the S9L and JL structures was mediated largely by backbone and ribose contacts along the outside of the minor grooves of H3 and H4, with two genuine minor groove hydrogen bonds to the guanine base of residue 16 (Fig. 2.5d). These contacts may have contributed to the widened minor groove values calculated for the penultimate basepair using CURVES, reported as 10.2 Å (JL) and 10.4 Å (S9L) versus 9.4 Å (C3L) and 9.3 Å (4WJ) for the base-pair between residues 16 and 48 of H3. (Notably, the 15-49 base pair interaction was not amenable to measurement by CURVES due to its terminal location.) Standard A-form RNA exhibits a minor groove width of 11 Å [47], and the discrepancy from this value may be accounted for by the higher than normal local twist.

Variation in the minor groove width for H4, which contacted H3 in the crystal lattice, exhibited a much smaller range of values across the four structures (±0.2 Å), with a mean 9.5 Å. The twist values for H4 also exhibited a smaller range (±0.4˚), with an average of 33˚. Most importantly, all distortions were mitigated approaching the active site of the hairpin
ribozyme. This observation has evolutionary implications for RNA enzymes, and suggests that the folds of globular RNAs may be somewhat self-correcting and tolerant of flanking structural perturbations, even though they lack the true hydrophobic core that confers proteins with the substantial plasticity and stability required to accommodate insertions [49].

2.5.4.2 Implications for Molecular Mimicry through Engineered RNA Crystal Packing

A novel feature of the minimal ribozyme lattice was observed during the comparison of minimally hinged and 4WJ structures. The prevalence of pseudo-continuous helical packing within the lattice of minimal constructs (Fig. 2.6a) enabled the formation of a rudimentary 4WJ motif. H4' of one symmetry mate stacked bluntly onto H3, while H2' of another symmetry mate stacked onto H2 in a staggered fashion through the U•U interaction (e.g., Fig. 2.6b). Additionally, the 2'-OH of a symmetry-related U-5 molecule was positioned appropriately to engage in a ~3 Å hydrogen bond with a non-bridging phosphoryl oxygen of either the C3 or S9 linker (Figs. 1.6b & 1.6c). Overall, the stacking interactions at this helical intersection buried 490 Å² and 540 Å² of surface area within the C3L and S9L structures, respectively, and exhibit an uncanny resemblance to the 4WJ motif (Fig. 2.6d). A similar series of interactions were described above as stabilizing forces for the inter-domain linker residue, A14, present in the 4WJ structure. In the latter molecule, the base-pair between U-5 and A14 at the top of H2 was flanked by a coaxial interaction to form an energetically favorable flush stack; the end of H3 was supported similarly (Fig. 2.6d) [14]. The strand of RNA in the 4WJ structure that comprised these two coaxial stacks was single stranded as it crossed over A14, further contributing to the ~710 Å² of surface area buried in the interface of this biologically relevant motif. Although the natural 4WJ is an intramolecular interaction, as compared to the intermolecular contacts of the minimal constructs, a superposition of these three structures emphasizes the excellent agreement between the modes of stacking at this interface (Fig. 2.7).
As demonstrated here, base-stacking of pseudo-continuous helices is a powerful packing restraint commonly observed in natural RNA structures and should be considered as a key driving force when engineering blunt or overhanging duplex sequences for crystallization. The inclusion of synthetic linkers between helices and the generation of junctions through symmetry contacts represent rational approaches to tailor RNA crystallization constructs to the needs of specific structural studies. This approach requires \textit{a posteriori} structural knowledge, but the outcome demonstrates that significant advances in crystallization and diffraction can be attained by modifying a crystallization construct in a manner that uses variations of the naturally occurring RNA architecture.

### 2.6 Summary and Conclusions

This investigation represents the first crystallographic characterization of the hairpin ribozyme incorporating synthetic linkages between the loop A and B domains. As such, it is relevant on two fronts. From a crystallographic point of view, it represents a novel method for connecting RNA strands. Synthetic linkers appear less prone to non-productive coaxial stacking as compared to natural nucleotide linkages that have limited flexibility, as well as greater hydrophobicity and steric bulk. The structures of this study also highlight the potential for deliberately promoting 4WJ mimicry in RNA crystal-packing interactions, with an obvious rationale for screening multiple linkers. Moreover, the use of crystallographic screening and x-ray diffraction analyses on multiple constructs is still worthwhile, despite apparent similarities of constructs in solution enzymatic assays.

From a functional perspective, comparison of the four structures discussed here (JL, C3L, S9L, and 4WJ) provides the first opportunity to analyze the influences of the hinge region on the global fold of the ribozyme. Significantly, this investigation was conducted in the context of multiple lattice packing schemes. As such, the comparisons have facilitated a dissection of characteristics relevant in a biological setting as compared to those generated
solely from crystal lattice contacts, which can be especially problematic to parse in nucleic acid structures [50, 51]. Overall, these results suggest a rationale to produce minimal RNA constructs based on natural 4WJ motifs that have led to well-diffracting, hinged hairpin ribozyme constructs amenable to incorporation of synthetic abasic residues and transition-state analogues for use in structure/function studies.

2.7 Acknowledgments

The authors thank A. Torelli for critical remarks on the manuscript. We are grateful to the staff of CHESS for assistance with data collection. Support for this project was derived from grant NIH/NIGMS R01 GM63162 and Petroleum Research Fund award 45534-AC 4 to J.E.W. C.M. was supported in part from a pre-doctoral T32 NIH Training award GM068411. CHESS is supported by the NSF under award DMR-0225180 and the NIH through NCRR award RR-01646.
References


Table 2.1 Intensity and Refinement Statistics for Alternatively Junctioned Ribozymes

<table>
<thead>
<tr>
<th>Construct</th>
<th>61-mer Junctionless (JL)</th>
<th>61-mer Position 14 C3 Linker†</th>
<th>61-mer Position 14 S9 Linker††</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range (Å)</td>
<td>2.05 to 30.0 (2.05 to 2.12)</td>
<td>3.35 to 34.7 (3.35 to 3.47)</td>
<td>2.65 to 38.9 (2.65 to 2.74)</td>
</tr>
<tr>
<td>Space group</td>
<td>P6₁22</td>
<td>P6₁22</td>
<td>P6₁22</td>
</tr>
<tr>
<td>Cell dimensions (Å)</td>
<td>a = 93.3, c = 131.3</td>
<td>a = 91.1, c = 145.0</td>
<td>a = 94.3, c = 131.1</td>
</tr>
<tr>
<td>Solvent content (%)</td>
<td>78%</td>
<td>79%</td>
<td>78%</td>
</tr>
<tr>
<td>No. measurements</td>
<td>178418</td>
<td>61410</td>
<td>158609</td>
</tr>
<tr>
<td>No. unique reflections</td>
<td>21669 (5296)</td>
<td>5453 (522)</td>
<td>10389 (1006)</td>
</tr>
<tr>
<td>Average redundancy</td>
<td>8.2 (4.7)</td>
<td>11.3 (12.4)</td>
<td>15.3 (16.1)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.3 (99.1)</td>
<td>98.6 (99.1)</td>
<td>98.6 (99.6)</td>
</tr>
<tr>
<td>R_{sym} (%)</td>
<td>3.0 (43.2)</td>
<td>10.6 (51.0)</td>
<td>5.9 (45.7)</td>
</tr>
<tr>
<td>I/σ (I)</td>
<td>24.0 (3.3)</td>
<td>16.2 (4.7)</td>
<td>25.5 (6.0)</td>
</tr>
</tbody>
</table>

Refinement Statistics

<table>
<thead>
<tr>
<th></th>
<th>1295</th>
<th>1304</th>
<th>1308</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. RNA atoms</td>
<td>87</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>No. water atoms</td>
<td>2 Co(NH₃)₆³⁺</td>
<td>1 Co(NH₃)₆³⁺</td>
<td>1 Co(NH₃)₆³⁺</td>
</tr>
<tr>
<td>1 SO₄²⁻</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R_{cryst}/R_{work}/R_{free} (%)</td>
<td>24.8 / 25.2 / 26.8</td>
<td>26.1 / 26.6 / 27.3</td>
<td>19.9 / 19.5 / 22.2</td>
</tr>
</tbody>
</table>

RMS deviations from ideality

<table>
<thead>
<tr>
<th></th>
<th>0.006</th>
<th>0.008</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bond lengths (Å)</td>
<td>1.4</td>
<td>1.4</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Temperature Factors

<table>
<thead>
<tr>
<th></th>
<th>69.5</th>
<th>95.6</th>
<th>78.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>B factor RNA (Å²)</td>
<td>74.0</td>
<td>78.4</td>
<td>77.0</td>
</tr>
</tbody>
</table>

†From PDB ID 2OUE. †From PDB ID 2NPZ. ††From PDB ID 2NPY. Parenthetical values correspond to the highest resolution shell.

\[ R_{sym} = \frac{\sum |I(h) - \langle I(h)\rangle|}{\sum I(h)} \times 100, \] where \( I(h) \) is the observed intensity of the jth measurement of reflection h and \( \langle I(h)\rangle \) is the mean intensity of reflection h.

\[ R = \frac{\sum |F_o| - k|F_c|}{\sum |F_o|} \times 100, \] where \( F_o \) and \( F_c \) are the observed and calculated structure factor amplitudes and k is a scale factor. \( R_{cryst} \) is calculated using the residual target in CNS with all reflections in the resolution range of refinement. \( R_{work} \) is calculated using all reflections except those randomly removed for the test set (5-7%); \( R_{free} \) is calculated using the test set of reflections.
Figure 2.1  Schematic depictions of the global hairpin ribozyme fold and RNA sequences  (a) Ribbon diagram of the four-way helical junction (4WJ) hairpin ribozyme, PDB ID 1M5K. The junction region, not present in minimal constructs, is shown in pink. RNA and protein residues included for the purposes of forming the U1A crystallization platform are colored gray. Other strands are colored as described in (b). (b) Secondary structure of the minimal hinged hairpin ribozyme adapted from RNAview [30]. The substrate strand is depicted in green, the S-turn strand in red, and the 29-mer strand in blue. The linker position is colored orange. The helix-loop-helix domains are labeled loop A and loop B; helices are labeled H1 - H4. The inset displays two alternate synthetic linkages, the C3L and S9L residues. The site of enzymatic cleavage is highlighted by a star. The A-1 residue is a 2'-deoxy A in the C3L structure, and a 2'-deoxy-2'-amino U in the S9L structure. Conserved residues are boxed; blue backgrounds indicate residues of the ribose zipper; yellow boxes indicate E-loop residues; grey circled residues belong to the S-turn. Hydrogen bond pairings: open-square, Hoogsteen; open triangle, trans-sugar; open circle, Watson-Crick face; closed circle, wobble pair. Double and single lines indicate Watson-Crick pairs; black dashed lines indicate single hydrogen bonds.  (c) Ribbon diagram of the minimal hinged hairpin ribozyme solved in this investigation.
Figure 2.2  Ball-and-stick and electron density maps of the engineered U•U mismatch at the top of helix 2 (H2)  

(a) The U-5•A14 base pair as observed in one of two molecules in the asymmetric unit of the 4WJ structure, 1M5K. Oxygen atoms are colored red, nitrogen atoms are blue. Dashed gray lines indicate putative hydrogen bonds with corresponding distances.  

(b) The minimal junctionless ribozyme 64-mer structure, 1X9K, with an equivalent orientation as in panel a. Two symmetry-related U-5 residues are separated by a 3 Å translation and are not base-paired.  

(c) The original syn-anti base-pairing conformation of U-5 modeled for the 2.05 Å resolution junctionless ribozyme (PDB ID 1ZFR). Alternate conformations were modeled with equal occupancy and are colored green or cyan. A single set of hydrogen bonds is indicated; the symmetry-related hydrogen bonds and labels were omitted for clarity. A $\sigma_A$-weighted simulated annealing omit electron density map calculated for the U-5 residue and nearby solvent molecules is shown contoured at 3.5 $\sigma$ with coefficients $mF_o$-$DF_c$.  

(d) The revised 2.05 Å resolution, minimal hairpin ribozyme structure of this study depicting the new anti-anti or ‘Calcutta’ model for U-5 fit into the omit map of (c).
Figure 2.3  Kinetic analysis of hairpin ribozyme constructs  Substrate cleavage reactions were carried out under single turnover conditions. For each construct (circles = S9 linked, diamonds = C3 linked and triangle = junctionless; JL) time courses were fit to a double-exponential kinetic equation (see Section 2.4.4 Materials and Methods). Amplitudes ($A_1$ and $A_2$) and rates ($k_1$ and $k_2$) for the S9 linked construct were: $A_1 = 0.461$ & $A_2 = 0.346$; $k_1 = 0.390 \text{ min}^{-1} \pm 0.055$, $k_2 = 0.041 \text{ min}^{-1} \pm 0.007$. For C3, $A_1 = 0.393$ & $A_2 = 0.387$; $k_1 = 0.325 \text{ min}^{-1} \pm 0.080$, $k_2 = 0.024 \text{ min}^{-1} \pm 0.005$. For the junctionless construct, $A_1 = 0.114$ & $A_2 = 0.423$; $k_1 = 0.419 \text{ min}^{-1} \pm 0.154$, $k_2 = 0.014 \text{ min}^{-1} \pm 0.002$. Inset: representative gel images of the cleavage time course for the S9 and JL constructs, similar results were observed for the C3 construct (data not shown). Bands represent 5' Cy5-labeled 13-mer substrate and 8-mer product strands.
Figure 2.4  Comparison of the local structural effects of the synthetic C3L and S9L residues observed in the hinge region between the loop A and B domains  
(a) $\sigma_\alpha$-weighted simulated annealing $mF_o-DF_c$ omit electron density map calculated for the C3L residue and flanking G15 phosphate. Blue density is contoured at 3\(\sigma\) and green density at 9\(\sigma\). The distance between atom O3' of residues 13 and 15 is shown as
a dashed line. (b) Omit map for the S9L residue as described in (a); blue density is contoured at 3.5σ and green density at 15σ. (c) Linker and neighboring ribose positions as observed in the superposition of four alternatively hinged hairpin ribozyme structures. Note the trend in the ribose positions of residue 13 versus 16. The C3L structure (blue) most closely mimics the natural A14 linkage of the 4WJ structure (pink), whereas the S9L structure (magenta) most closely mimics the junctionless structure (green). Bases are omitted for clarity with the exception of A14. (d) Structural influence of the C3L structure on the terminus of H3. The C3L structure is depicted by blue sticks, and the S9L structure by magenta lines. The lateral shift of the H3 terminal base-pair is represented by the 2.8 Å distance between O3’ atoms of residue C49 in the respective C3L and S9L structures. The 2₁ pseudo-helical base-stacking interaction with H4’ is shown to demonstrate the more flush base-stacking interaction exhibited by the C3L structure, as compared to S9L. The 2.9 Å distance between the O4’ and O3’ atoms of ribose moieties engaged in crystal packing in the C3L structure is compared to the equivalent distances of 3.5 Å and 4.0 Å displayed for the S9L structure in (e). (e) The equivalent region in the S9L structure, as described in (d). S9L residues are shown as magenta sticks, C3L residues as blue lines. Equivalent O4’ to O3’ distances between H3 and H4’ helices are longer in the S9L structure, representative of a more staggered base-stacking interaction. (f) A superposition of the terminal H4 base-pair of the C3L and S9L structures demonstrates the influence of the shortened linker on pseudo-helical base stacking. The more flush base stack in the C3L structure is necessarily accompanied by a 1.5 Å upward shift of U31’ to avoid a steric clash with the symmetry related C49 residue. Note that even after this movement, U31’ and C49 are still closer in the C3L structure (c, d).
Figure 2.5  Schematic surface and ball-and-stick diagrams illustrating the 6₁ packing interactions of hinged hairpin ribozyme constructs  
(a) Comparison of C3L (blue) and S9L (magenta) unit cells. The perspective represents the interaction of molecules about the 6₁ screw axis. (For 2-fold and 2₁ operations refer to Fig. 2.6a). Superposition of C3L onto S9L initiates at the third molecule from the bottom (i.e., the ‘reference’ molecule), represented in bold magenta and blue overlay. Symmetry operations used to generate the remaining molecules demonstrate the degeneration of the superposition as a result of the 15 Å elongated unit cell of the C3L structure. C3L molecules are shown in blue and green with S9L structures shown in pink. The dashed box denotes the inset for (b & c).  
(b) Expanded view of the 6₁ packing scheme for the S9L structure. Atoms engaged in potential hydrogen bonds are depicted as white spheres. Helices are labeled as in Fig. 2.1c.
and primes (’) denote symmetry-related molecules. An additional H3'' helix that is packed in a blunt-ended base stack with H4 has been omitted for clarity. The asymmetric unit (H4) is colored magenta with symmetry molecules in red (H3’) or light pink (H4’). (c) The C3L 6₁ packing scheme as described in (b), but the asu is colored blue and symmetry mates are colored teal (H3’) or green (H4’). White spheres identify equivalent atoms in the C3L structure that are engaged in hydrogen bonding in the S9L structure. (d - g) Detailed view of hydrogen bond interactions with each of four H4 residues. The A31 and U31 residues are base-paired; A31 ends the linker strand while U31 begins the S-turn strand. (h - k) Equivalent distances in the C3L structure demonstrate the loss of 6₁-fold packing interactions.
Figure 2.6  Schematic surface and ball-and-stick diagrams illustrating a pseudo-four-way helical junction generated by crystal packing interactions of minimal hairpin ribozymes  (a) Unit cell perspective of the 2-fold and 21-fold symmetry contacts within the asymmetric units of both C3L and S9L structures. Superposed C3L (blue) and S9L (magenta) molecules are shown overlaid in the centre of the diagram (i.e., reference molecule). The minimal difference in the lengths of the a- and b-axes of the C3L and S9L unit cells (3.2 Å) allows greater conservation in the superposition as it is extended from the reference molecule’s asymmetric unit; therefore, symmetry-related S9L molecules (pink surface representations) are largely hidden behind the C3L molecules (blue and green
cartoon and stick depiction). Helix 1 is labeled to demonstrate its high degree of solvent exposure. The 2
screw axis depicts the relationship between H3 and H4’. The boxed region is expanded in panels b-d. (b-d) Individual panels of the junction region of the C3L, S9L and 4WJ structures. Residues are colored according to strand conventions in Fig. 2.1b and numbered according to hairpin ribozyme conventions [52]. Surface representations for the C3L molecules are colored as observed in panel a, and rotated 90° anti-clockwise from that perspective. Oxygen atoms engaged in putative hydrogen bonds are designated as red spheres connected by dashed lines.
Figure 2.7  A stereo-view stick diagram of a superposition of the 4WJ, C3L and S9L hairpin ribozyme structures  The orientation of the stereo-view is rotated 180° around the vertical axis compared to Fig. 2.4, and was chosen to emphasize the agreement between residues of the intramolecular four-way junction with those of the intermolecular packing environment observed in the C3L and S9L structures. The 4WJ structure is shown in salmon, C3L in blue, and S9L in magenta. Residues are numbered according to hairpin ribozyme conventions and labeled at U-5, where the sequence is conserved between all three structures.
Chapter 3:

Structural Effects of Nucleobase Variations at Key Active Site Residue Ade38 in the Hairpin Ribozyme

The data presented data in this chapter were submitted for publication.

Co-Authors who contributed to this research are:

Krucinska, J., Salter, J.D., Williams, K., Janda, A., and Wedekind, J.E.
3.1 Footnotes

Abbreviations: *Ade*, adenosine; *2AP*, 2-aminopurine; *IsoG*, isoguanosine; *Pur*, purine, *Cyt*, cytidine; *Gua*, guanosine; *Uri*, uridine; *nt*, nucleotide; *2’-OMe*, 2’-O-methyl modification of ribose; *2’-5’, 3’-deoxy, 2’-5’-phosphodiester linkage; *CHESS*, the Cornell High Energy Synchrotron Source Ithaca, NY; *WT*, wild type with Ade38 intact; *CSD*, Cambridge Crystallographic Database;
3.2 Abstract

The hairpin ribozyme requires functional groups from Ade38 to achieve phosphodiester bond cleavage or ligation. To identify features that contribute to catalysis, structures of Ade38 base variants, cytidine (Cyt), guanine (Gua), 2,6-diaminopurine (DAP) and 2-aminopurine (2AP), were determined in a resolution range between 2.2 and 2.8 Å. Two substrate modifications were then compared for each variant: a 2´-O-methyl-modified nucleophile at A-1 to mimic the pre-catalytic state and a 3´-deoxy-2´,5´-phosphodiester linkage between A-1 and G+1 to mimic the reaction intermediate. The results revealed the importance of the N1 and N6 atoms of Ade38. Lack of N6 in 2AP38 coincided with failure to localize the scissile phosphate. The Cyt38 variant suggested that its poor reactivity was due to a non-productive orientation of its imino nitrogen. Specifically, the Cyt38 base exhibited an anti configuration in which the N3 atom pointed away from the active site. The Gua38 variant is highly deleterious to activity and produced alternate conformations for the S-turn in the 2´-5´ structure, suggesting that the pro-R oxygen of the scissile bond is incompatible with an electro-negative group nearby. In general, Ade38 modifications solved in the presence of 2´-O-Me A-1 caused deviations of the scissile phosphate from ‘in-line’ geometry, whereas those with 2´-5´ modifications exhibited greater S-turn variability. These findings demonstrate the importance of the Ade38 Watson-Crick face in promoting and stabilizing the catalytic geometry, as well as the sensitivity of the local RNA fold to misfolding when electrostatic and shape features fail to complement.
3.3 Introduction

The hairpin ribozyme is a small RNA enzyme whose family members catalyze a reversible, site-specific phosphodiester bond cleavage reaction within a complementary RNA substrate specified by Watson-Crick base pairing [1, 2]. Cleavage occurs when the 2'-hydroxyl group at position A-1 performs a nucleophilic attack on the G+1 phosphorus (Fig. 3.1) [3]. A trigonal-bipyramidal intermediate is resolved through formation of a cyclic 2',3'-phosphodiester and free 5’-OH [4]. This reaction is essential to the replication of several plant viroids [5], the human hepatitis delta virus, the mitochondrial Varkud satellite from *Neurospora* [6], and the *glmS* metabolite-sensing riboswitch encountered among gram positive bacteria [7]. Larger, more complex ribozymes are biologically pervasive including the ribosome [8, 9], self-splicing introns [10, 11], as well as RNase P, which processes the 5’-ends of tRNA precursors [12]. As such, the small ribozymes provide excellent model systems to investigate the general principles of RNA catalysis using structure and function approaches [13, 14].

Previous investigations of the hairpin ribozyme demonstrated the importance of nucleobases Gua8 and Ade38. The latter residues were poised in crystal structures to participate directly in catalysis [15, 16]. At a minimum, these residues contribute to electrostatic stabilization of the transition state [16-20]. However, a role for these nucleobases as general acid/base catalysts has not been dismissed entirely based upon nucleobase substitutions at Gua8 [21, 22], elegant simulations of pH rate profiles that accounted for both the Gua8 and Ade38 bases [23], and alkylation of G8 through contact with a 2’-bromoacetamide group added at A-1 in lieu of the nucleophile [24].

However, the significance of Gua8 has been challenged by investigations of abasic hairpin ribozymes in which various nucleobases are ablated, while leaving a 2'-deoxyribose in its place. Abasic modifications at Ade9 and Ade10 within the compact core reduced hairpin ribozyme activity less than 10-fold [18]. In contrast, an abasic substitution at Gua8 exhibited a
350-fold loss in activity [19], whereas nucleobase removal at Ade38 resulted in a 5,100-fold loss in activity [20]. Rescue experiments in which exogenous nucleobases were added in trans to abasic variants suggested that the key functional group feature of Ade38 was the presence of the amidine group (i.e., an exocyclic amine alpha to a ring nitrogen), as well as the pK\textsubscript{a} of the ring nitrogen.

A series of Ade38 ‘mutants’ was constructed and analyzed for activity to complement the results of the abasic variants. With exception of DAP38, which was 2.7-fold more reactive than the wild type ribozyme at pH 7.5, each variant showed reduced cleavage activity relative to wild type in the order: 2AP38 (~50-fold) > IsoG38 (~170-fold) > Pur38 (4,700-fold) > Cyt38 (8,000-fold) > Gua38 (13,300-fold) and Uri38 (19,000-fold) [20]. The covalent nature of these modifications presented a unique set of considerations when interpreting results, as compared to rescue experiments. Occupancy was assured, but the possibility that global or localized non-productive folding events influenced the kinetic results was difficult to address. Fortunately, such covalent modifications are particularly well suited to crystallographic structural investigations.

The goal of this investigation was to correlate prior functional observations of Ade38 hairpin ribozyme variants with detailed structural information. Four Ade38 modifications spanning a range of activities were examined in the context of a minimal, hinged hairpin ribozyme that was rendered into two distinct active site configurations. A ‘pre-catalytic’ conformation with near in-line geometry was produced by incorporating a 2′-OMe modification at A-1, thereby inactivating the nucleophile as reported for the wild type ribozyme [15, 16], as well as several Gua8 variants [25, 26]. A ‘reaction-intermediate’ conformation was generated by use of an inert 2′-5′ linkage in place of the scissile bond. This ‘reverse’ linkage restrains the non-bridging oxygens of the scissile phosphate at G+1 to interact with the exocyclic amines of Ade38 and Gua8, as observed for the oxo-vanadium transition-state mimic [27]. The localized structural rearrangements reported herein highlight
the flexible nature of the hairpin ribozyme and provide a molecular perspective that accounts for activity losses by means that might not otherwise be considered.

3.4 Experimental Procedures

3.4.1 Hairpin Ribozyme Constructs.

The 61-nt hinged hairpin ribozyme used for this investigation comprised three strands and was described previously [27, 28]. The sequence of the ribozyme used in the current study is shown in Fig. 3.2a. The nucleobase at position 38 was substituted with DAP, 2AP, Cyt or Gua. The Uri39Cyt gain-of-function mutation [29] was included throughout this study to prevent conformational heterogeneity in the S-turn as documented [25]. Position Ade14 was substituted with a commercially available 9-atom synthetic linker that preserved the phosphodiester bond (Fig. 3.2a, S9L). Two of the four helices that compose the four-way junction of the natural hairpin ribozyme [30] were removed, although the junction was closely mimicked through engineered intermolecular contacts provided by the crystallographic packing environment [28].

Mutations to the substrate strand included a 2´-OMe A-1 or a 2´-5´-phosphodiester linkage between A-1 and G+1, both of which prevented cleavage in the crystal. Oligonucleotide strands were obtained from Dharmaco (Lafayette, CO) with the exception of the 2´-5´ linked substrate 13-mer strand. For this modification, the phosphoramidite was purchased from Glen Research (Sterling, VA) and was incorporated into the 13-mer through chemical synthesis at the W.M. Keck Facility (Yale University, New Haven, CT). RNA strands were subjected to in-house deprotection, reverse phase HPLC purification, desalting and domain docking as described [31, 32].

3.4.2 Crystallization and X-ray Diffraction Experiments

Crystals grew as hexagonal rods in hanging-drop vapor diffusion experiments at 20 °C from screens based on previously established conditions [25, 26]. Conditions common to all
constructs included 0.25 M Li$_2$SO$_4$, 1 mM Co(NH$_3)_6$Cl$_3$, and 2 mM spermidine. Poly (ethylene)glycol 2000 monomethyl ether was used as the precipitant in a range of 20% to 24% (w/v). Solutions were buffered using 0.10 M sodium cacodylate for pH ranges from 6.0 to 6.8, 0.10 M HEPES for pH ranges from 7.0 to 7.8, and 0.10 M TRIS for pH ranges from 8.0 to 8.8. Table 3.1 describes the specific pH values of each crystal.

All crystals were cryo-protected by 3 min serial transfers through 4 synthetic mother liquors containing 5% to 18 % (v/v) glycerol. Crystals were captured in thin nylon loops (Hampton Research) and flash cooled in a -178 °C stream of nitrogen gas (X-stream, Rigaku/MSC, TX). Diffraction data for the 2AP38 2´-OMe structure were recorded at home on an R-Axis IV image plate system equipped with confocal optics (Rigaku/MSC). CuKα x-rays were generated using a Rigaku RUH2R rotating anode operated at 4.5 kW. A total of 130 images were collected at a crystal-to-detector distance of 15.0 cm and an exposure time of 25 min per 0.5° oscillation. All other crystals were stored in dewars of liquid nitrogen until they were retrieved for data collection using synchrotron radiation. Data for other 2´-OMe modified ribozymes were collected at -178 °C using a Quantum210 CCD (ADSC) at the A1 station of CHESS using x-rays of wavelength 0.977 Å. For each sample, high resolution data were recorded from approximately 150 to 250 images (dependent on the c* orientation) at a crystal-to-detector-distance of 21.0 cm, and an exposure time of 15 to 20 sec per 0.5° rotation. Low resolution diffraction data were collected similarly, except the exposure time was ~2 sec. Diffraction data for 2´-5´ modified hairpin ribozymes were collected at -178 °C using a Quantum270 CCD (ADSC) at the CHESS F1 station using x-rays of wavelength 0.918 Å. High resolution data were collected from approximately 200 to 300 images at a crystal-to-detector-distance of 27.5 cm with exposure times ranging between 10 to 20 sec per 0.5° rotation. Low resolution data were recorded similarly with a 1 sec exposure. Low resolution data were collected subsequent to high resolution data except DAP38 2´-5´.
3.4.3 X-ray Data Reduction, Structure Determination and Refinement

Diffraction data were reduced to intensities and scaled using Crystal Clear [33]. High and low resolution data sets were merged after integration for all structures, except Gua38 2´-5´ for which only the high-resolution scan was used. Data reduction statistics are provided in Table 3.1. Each construct crystallized in space group \( P_{6_1} \) with a single molecule per asymmetric unit, as observed previously [25, 28, 32]. Preliminary structures were determined by difference Fourier methods using refined models, such as the 2.05 Å resolution 2´-OMe WT structure (PDB ID 2OUE) or the 2.35 Å resolution 2´-5´ modified WT structure (PDB ID 2P7F) [26, 27, 34]. Refinement was performed using cross-validated maximum likelihood targets as implemented in CNS [35, 36]. Rigid body refinement was performed initially on the entire structure (devoid of ions and waters) and subsequently on the individual loop A and B domains. Cartesian simulated annealing, positional minimization and individual B-factor refinement were conducted as necessary. Simulated annealing omit maps were used to guide manual refinement and to reduce model bias [37]. Model building was conducted using the interactive graphics program O [34]. Electron density maps were calculated using reduced-bias \( \sigma_A \)-coefficients throughout the refinement process [38]. Upon convergence, all data were recombined for a final round of positional minimization, during which a residual least squares target was employed, thereby giving a final \( R_{\text{factor}} \) (Table 1).

The default CNS parameter and topology files for RNA (version 1.0) were adapted to accommodate non-standard nucleotides as described [28]. Bond angles, distances and geometry were derived from high-resolution small molecule structures (identification codes: MIUDAP, MTYDAP, WIPCAZ, BUREAP, FUREAP, and YACHAL) obtained through the CSD [39]. A dual conformation was built at U-5 for all structures, as described [28]. Alternate conformations were also modeled for the scissile phosphate of the 2AP38 2´-OMe structure, as well as S-turn residues Uri37, Gua38 and Cyt39 of the Gua38 2´-5´ structure. Occupancies were adjusted to minimize the \( R_{\text{factor}} \) and \( R_{\text{free}} \) statistics, peaks in difference electron density maps, and B-factor discrepancies between models of partial occupancy.
relative to those of uniform occupancy in the local structure. Sugar pucker restraints were derived from higher-resolution hairpin ribozyme structures determined previously [26, 27], and were verified further through simulated annealing omit maps.

Waters were assigned to each structure manually based on the following criteria: peak height $>3.0 \sigma$ in $mF_o - DF_c$ maps; $\geq 0.8 \sigma$ electron density in $2mF_o - DF_c$ maps; reasonable geometry for hydrogen bonding, with distances between donor and acceptor groups between 2.6 to 3.8 Å; and refined temperature factors comparable to surrounding RNA atoms. Two Co(NH$_3$)$_6^{3+}$ ions have been observed consistently in all minimal hairpin ribozyme crystals located ($i$) near G21 and A40 of the S-turn, and ($ii$) within the major groove of H2 [25]. New cobalt sites were observed in this investigation and were treated as described [25] to optimize the fit to electron density maps while maximizing hydrogen bond formation between the hexaamine coordination sphere neighboring RNA ligands.

All-atom superpositions were generated from conserved residues using LSQKAB as implemented in CCP4 [40, 41]. Solvent content was calculated in CNS using a partial specific volume (PSV) of 0.57 mL/g for RNA [42]. All figures were generated in PyMol, supplemented by Nuccyl in Figure 1b [43-45].

3.5 Results

3.5.1 Quality of the Hinged Hairpin Ribozyme Structures

The goal of this investigation was to elucidate how various nucleobase substitutions at Ade38 influence the active site architecture. The Ade38 nucleobase substitutions of this investigation included DAP, 2AP, Gua and Cyt. Each variant was crystallized in the context of two active site configurations representing pre-catalytic (2’-OMe A-1) and reaction intermediate (2’-5’) states. The structures discussed herein were solved and refined to resolutions between 2.20 and 2.75 Å. The quality of each structure at the active site is indicated by the fit of the model to representative simulated-annealing omit electron density
maps calculated independently for the respective position 38 base, and nucleotides A-1 through G+1 of the scissile bond (Fig. 3.3). Each structure exhibited reasonable refinement statistics including a final $R_{\text{factor}} \leq 23.7\%$, with ‘working’ and ‘free’ $R$-values $\leq 23.6$ and $\leq 26.2\%$, respectively; coordinate errors were between 0.46 and 0.73 Å (Table 3.1). In general, the RNA atoms for each structure were well-ordered in electron density maps, although the quality of maps was noticeably poorer for the nucleotides of H1 as observed previously [25]. Maps with coefficients $2mF_o-DF_c$ were contiguous for the RNA of all structures with exception of the S-turn of the Gua38 2´-5´ structure (discussed below). In the latter case, two alternate conformations were modeled for residues 37-39 according to features in $mF_o-DF_c$ and $\sigma_A$-weighted simulated annealing omit electron density maps. Occupancy was modeled at 50 / 50 as described in the Experimental Procedures section.

3.5.2 DAP38 is a Gain of Function Substitution

The DAP38 variant exhibited a 2.7-fold acceleration in the rate of cleavage, but still exhibited a bell-shaped pH rate profile comparable to WT [20]. A superposition of the respective wild type and DAP38 structures revealed only minor differences between either the 2´-5´ or 2´-OMe variants of this study (Fig. 3.4). The distance between the N1 atom of residue 38 and the O2’ atom of A-1 was approximately 3.0 Å in both 2´-5´ structures. The largest deviation between the DAP38 variant and the corresponding reaction-intermediate WT structure was a distance of 0.7 Å located between the N6 atoms of the Ade38 and DAP38 residues (Fig. 3.4A, inset). However, this difference did not influence hydrogen bond formation between the N6 exocyclic amine of position 38 and the $\text{pro-R}$ oxygen of G+1, which maintained a distance of 2.9 Å in both molecules.

Of course the major structural difference between Ade and DAP is that the latter base possesses an additional N2 exocyclic amine in lieu of a hydrogen. As such, the DAP38 variant formed new hydrogen bonds from N2 to two water molecules in the pre-catalytic structure, W3 and W4 (Fig. 3.4B). Both waters were observed previously in the minimal hairpin ribozyme structure in complex with vanadate [27]. In fact W3 is well represented in a
number of minimal hairpin ribozyme structures and was also observed in the 2'-OMe
modified WT, the G8A structure which had a 2'OH at the A-1 position [26], and the 2'-OMe
A38C and A382AP structures described below. The conserved W3 interactions included the
O4' atom of G+1 (Fig. 3.5b) and the 2'OH of U37 (not shown), thus the interaction with the N2
amine of DAP constitutes a new interaction (that was also observed in the A382AP structure).
The role of this water is not presently understood, and it has been previously suggested that
both W3 and W4 could easily exchange with bulk solvent during substrate turnover [27]. W3
has not been modeled in any of the 2'-5' modified ribozymes. A 3σ peak was observed at the
W3 position in mF_o-DF_c electron density maps for the 2'-5' modified WT [27], A38DAP and
A382AP structures; however, the aforementioned criteria for water placement were not met
and W3 was therefore not included in the final model. No novel water sites were identified in
either A38DAP structure that would be in a position to enhance reactivity.

Given the close structural similarity of the DAP and WT structures, it seems reasonable
that the increased activity of the DAP variant resulted from the increased pK_A at the N1
position of this residue (Fig. 3.2b, 5.5 versus 3.9), combined with the increased neutralization
of negative charge provided by the N2 amine.

3.5.3 2AP38 Destabilizes the Scissile Phosphate

The structure of the A382AP ribozyme is an excellent candidate for analyzing the
structural influences of the N6 amine, as adenosine and 2-aminopurine exhibit similar pK_A
values for the N1 position (Fig. 3.2b, 3.9 versus 3.6), and both exhibit a single exocyclic
amine group extending from the purine ring. At the optimal pH for WT cleavage (7.5), a 50-
fold reduction in activity was reported for the A382AP variant [20]. However, this difference in
activity was only 17-fold at pH 6.0. The A382AP substitution was the only variant that
exhibited optimal activity at low pH.

In structures of the A382AP variant, the absence of an N6 amine was manifested
primarily in the conformation of the scissile phosphate. In contrast, the A38G substitution
exhibited a disrupted S-turn, but a well-ordered scissile phosphate. The A382AP 2'-OMe
structure, which was solved at a pH of 7.6, exhibited disorder in the scissile phosphate, as
evidenced by the split electron density observed in a simulated annealing omit map (Fig.
3.3b, left). This was modeled in two distinct conformations, with a relative distribution of 60 / 40. In the dominant population, the pro-R oxygen engaged in a 2.6 Å H-bond with the N2 of Gua8. In the lesser conformation, the pro-S oxygen engaged in a 2.6 Å H-bond with the N2 of 2ap38.

An important active site water molecule (W52) observed in the WT, A38DAP (Fig.
3.4b), and A38C (Fig. 3.6a) structures was not sufficiently resolved in electron density maps of the A382AP 2'-OMe ribozyme to warrant modeling. In the WT structure, this water was tetrahedrally coordinated to the N6 amine of Ade38, the O3' atom of A-1, the N6 amine of Ade9 and, lastly, to another water molecule [26]. W52, as it will be hitherto referred to, has been implicated as a potential specific base catalyst, responsible for abstracting a proton from the 2'OH nucleophile [46] [26]. However, W52 is replaced by the pro-R oxygen in structures representative of a reaction intermediate [16, 27]. As such, it cannot be known if the loss in activity that results from substitution of groups that interacted with W52 [47] was due to loss of the water, destabilization of the transition state, or perhaps both.

Although the A382AP variant lacked an N6 amine at residue 38, the remaining RNA groups previously observed to interact with W52 were conserved. The small elongated 3σ peak in the mF_o-Df_c electron density maps of the A382AP structure suggested that W52 was at least occasionally present, but disordered. This would presumably be due to the loss of the N6 amine and the disorder of the scissile phosphate. When considering the importance of W52, it is significant to note that electron density maps of the 2.35 Å A38G 2'-OMe structure exhibited no peak for W52, and yet the scissile phosphate of this structure was well ordered. This would seem to suggest that the disorder observed for W52 in the A382AP structure was not the cause of the destabilized scissile phosphate.

The 2'-5' modified A382AP ribozyme was solved at a pH of 6.5 and exhibited a much more ordered scissile phosphate than that of the 2'-OMe version. (Note the small 12 σ peak
present around the P atom of G+1 in the A382AP 2′-5′ omit map, Fig. 3.4b rt. side.) This structure was solved to 2.75 Å resolution (compared to 2.35 Å resolution for the A382AP 2′-OMe structure). In this structure, the pro-S oxygen was positioned to engage in a bifurcated H-bond with protons from either the N1 or N2 of G8. The O2′ was positioned 3.4 Å from the N1 of 2ap38, such that an H-bond to an imino proton was possible, but not presumed (Fig. 3.5b). Should N1 be de-protonated, no steric clash would occur between these groups.

A comparison of the 2′-5′ modified WT and A382AP ribozymes revealed a shift of 0.8 Å in the position of the N1 atom of residue 38 (Fig. 3.5b). A similar shift (0.6 Å) observed for the A38DAP ribozyme suggests that the N2 amine was responsible for the alteration. The placement of the N2 amine of 2ap38 was 3.6 Å from the O4′ atom of G+1, and 3.7 Å from its own O1P atom. These long distances do not support stabilization by specific H-bonds, but rather a non-specific attraction of the positive amine toward the negative pocket.

The most significant deviations observed between the 2′-5′ WT and A382AP ribozymes occurred at the scissile phosphate. In fact, the P, O5′ and pro-R oxygen atoms of the A382AP variant exhibited the greatest deviation from WT of any of the mutations addressed. The pro-R oxygen was shifted 1.3 Å with respect to WT, such that, even if an N6 moiety were present at position 38, there would be no interaction between these two atoms. Additionally, the distance between the O5′ leaving group and the N1 atom of residue 38 was elongated from 3 Å in the WT structure to 4 Å in the A382AP structure. Taken together, the 2′-OMe and 2′-5′ structures suggest that the loss of an N6 amine destabilized both the ground state and the transition state.

3.5.4 C38 Exhibits an Anti Conformation

Structures of the A38C variant further emphasized the ribozyme’s penchant for positioning a positive group at the site of the N6 amine of Ade38. Like Ade, Cyt has an exocyclic amine (N4), and is neutral in the deprotonated state. The pKₐ of the N3 imino proton is ~4.5 when free in solution, similar to Ade’s N1 pKₐ of 3.9 [48]. Given these similarities, one might imagine that the 8000-fold reduction in cleavage activity observed for
the A38C ribozyme resulted from the smaller size of the pyrimidine ring [20]. The N3 and N4 atoms of Cyt would be expected to position similarly to the N1 and N6 of Ade, but at a greater distance from the scissile phosphate. However, crystal structures of the Cyt38 residue deviated significantly from this expectation.

Ade38 exhibited a syn conformation in the WT ribozyme [15, 25]. However, the electron density maps for Cyt38 indicated an anti conformation in both the 2´-5´ and 2´-OMe modified crystal structures, which were solved at a pH of 6.5 (Fig. 3a,b). This finding was closely examined upon the observation of broadened electron density in the omit maps calculated for the 2´-OMe structure. This characteristic was suggestive of a mixture of syn and anti conformations, and was not observed in equivalent maps of the 2´-5´ structure (compare the inset of Fig. 3.3a versus 3.3b). mF_o-DF_c and 2mF_o-DF_c electron density maps were generated from a syn model, but these maps strongly indicated the anti conformation. Modeling the syn orientation also created local steric issues; the O2 atom of Cyt38 was less than 3 Å from both the O2' and O4' atoms of the Cyt38 ribose. Adjusting the glycosidic geometry to correct the ribose clash created a steric clash between the C5 atom of Cyt38 and the N6 atom of Ade24, given that both atoms would be associated with protons. Lastly, the syn conformation placed the O2 atom of A38C 2.6 Å from W3. As such, W3 was within H-bonding distance of 3 H-bond acceptors but only 1 donor. Therefore, Cyt38 was modeled as a single anti conformation in the 2´-OMe structure.

The position of the N4 atom of Cyt38 remained unchanged regardless of whether the chi (χ) angle was modeled in an anti or syn conformation. Strikingly, a superposition of WT and A38C models revealed that the N4 amine of Cyt38 occupied the same pocket as the N6 amine of Ade38 (Fig. 3.6a,b). In this position it formed an H-bond with W52 in the 2´-OMe structure (Fig. 3.6a). N4 was, therefore, also positioned to engage in a putative 3.4 Å H-bond with the pro-R oxygen of A-1 in the 2´-5´ structure (Fig. 3.6b).

The most significant effect of the anti conformation was the removal of the N3 atom from the active site. With no imino proton available to fill the role of general acid, and no
evidence in electron density maps of any additional metal or solvent ions that might act as a specific acid, it is not known how the A38C variant achieved catalysis. It could be proposed that the rate reported at pH 6.0 (5 x 10^{-6} \text{ min}^{-1}), reflects catalysis in the absence of an N1 proton. Given that an abasic ribozyme was similarly inhibited (5100-fold), this seems a reasonable explanation.

The \textit{anti} conformation was stabilized by three new interactions. The N6 amine of Ade24 was positioned 3.2 Å from the N3 atom of A38C. N3 would be expected to be deprotonated at pH 6.5 in this orientation (solution pK\textsubscript{A}, 4.5), and therefore able to accept an H-bond from the amine group (Fig. 3.6c). The N6 amine was also positioned to engage in a 2.9 Å H-bond with the O2 atom of A38C. This O2 atom was also positioned to interact with an amine moiety from a novel Co(NH\textsubscript{3})\textsubscript{6} ion that was observed in both the 2'-OME and 2'-5' modified A38C crystal structures (Fig. 3.6c). Notably, these stabilizing interactions would not be expected to titrate under the pH range of 6 to 8.6.

Crystallization solutions included 1 mM Co(NH\textsubscript{3})\textsubscript{6}Cl\textsubscript{3}; two binding sites for Co(NH\textsubscript{3})\textsubscript{6} were previously characterized for the minimal hairpin ribozyme [25]. The Co(NH\textsubscript{3})\textsubscript{6} ion observed here constitutes a new metal site that was novel to the A38C variant. \textit{B}-factors for this ion were higher than those observed for the S-turn Co(NH\textsubscript{3})\textsubscript{6}, but lower than those of the Co(NH\textsubscript{3})\textsubscript{6} ion modeled at the top of H2. Importantly, all six amine groups were positioned to engage in H-bonds with the RNA molecule.

Cleavage rates measured for A38C and A38abasic ribozymes in the presence of Co(NH\textsubscript{3})\textsubscript{6} were 10-fold slower than rates measured in the presence of Mg\textsuperscript{2+} [20]. In order to investigate the influence of Co(NH\textsubscript{3})\textsubscript{6} on the \textit{anti} / \textit{syn} configuration of Cyt38, the A38C variant was crystallized from a solution in which 10 mM MgCl\textsubscript{2} was substituted for 1mM Co(NH\textsubscript{3})\textsubscript{6}. Cyt38 was also observed to be \textit{anti} in the Mg\textsuperscript{2+} containing structure (Fig. 3.6d). \textit{F}_{o} - \textit{F}_{o} maps supported the third Co(NH\textsubscript{3})\textsubscript{6} site, but showed no evidence of a Mg\textsuperscript{2+} ion coordinated to the Cyt38 residue. Of note, a Mg\textsuperscript{2+} ion was observed in inner sphere coordination with the
N7 atom of residue Gua21 in the S-turn. This Mg\(^{2+}\) site was near, but not identical, to a previously characterized Co(NH\(_3\))\(_6\) site [25].

The fact that both abasic and A38C ribozymes were less reactive in the presence of Co(NH\(_3\))\(_6\) could indicate an alternate mechanism of cleavage by these ribozymes. A38C, A38U, A38G and abasic mutants exhibited an altered pH rate profile in which activity was maximized at a high pH (8.6), providing further support of an alternate mechanism [20]. pH studies to investigate the structural basis of the recovered activity observed at high pH are ongoing.

3.5.4.1 The A38C-Mg\(^{2+}\) Structure Exhibits a 3'-endo Sugar Pucker at A-1

The A38C crystal structure solved in the presence of magnesium exhibited the same anti configuration for Cyt38. However, there are two significant features of this structure. First, crystals grown in the presence of Mg\(^{2+}\) exhibited superior diffraction properties; this structure, at 2.2 Å resolution, is the most detailed of the series described here. As such, water molecules were resolved in the active site of this structure (Fig. 3.6d) that were not observed in the other Ade38 mutant structures described here. These waters were, however, observed previously in the WT 2'-OMe and 2'5' modified ribozymes solved to similarly high resolution.

The second, and most exciting, novel feature of the A38C-Mg\(^{2+}\) structure is the C3'-endo sugar pucker observed at A-1. This is the first 2'-OMe modified structure to exhibit a C3'endo pucker at this position. A superposition of the A38C-Mg\(^{2+}\) structure with the A38C-Co(NH\(_3\))\(_6\) structure revealed that the O2' atom of A-1 displaced W52 (data not shown). Notably, the location of W52 in the WT enzyme was 1.3 Å from W52 in the A38C mutant structures. As such, the methyl group of A-1 in the A38C-Mg\(^{2+}\) structure is superimposed on the WT W52 (Fig. 3.6d).

3.5.5 G38 is Not Compatible with Formation of the Transition State

Structures of the A38G variant highlighted the value of the 2'-5' modification. The A38G 2'-OMe structure was very similar to the WT structure (Fig. 3.7a), and offered little
insight into the 13000-fold reduction in cleavage activity observed for the A38G variant at pH 7.5 (~2000-fold at pH 8.6) [20]. The most interesting feature of the A38G 2'-OMe structure, crystallized at pH 8.6, was the position of the O5' atom of G+1. The 3.2 Å distance between the O5' leaving group and the N2 amine of Gua38 was amenable to formation of an H-bond. This interaction was not observed in the A38DAP and A382AP structures, although the N2 amine of all three structures superimposed very well (within 0.5 Å of each other). The O5' atom of G+1 was also 3.2 Å from the N1 of Gua38. Although the geometry was not ideal in this structure, a minimal rearrangement would be required to realize an H-bond to the N1 proton. At a pH of 8.6, the N1 atom Gua38 was likely to be protonated in this structure (solution pK_a: 10.0, Fig. 3.1b). The protonation state of Dap38 and 2ap38 was less certain. It may be of significance that N1 protonation of Gua yields a neutral nucleobase, whereas N1 protonation of Dap or 2ap induces a positive charge.

All in all, the orientation of the leaving group in this 2'-OMe structure was better suited for cleavage chemistry than that observed for any previously solved 2'-OMe hairpin ribozyme structure. The tau (τ) angle (defined as the O2' of A-1, to the P of G+1, to the O5' of G+1) was 166°; the ideal value for in-line geometry is 180°. In comparison, the tau (τ) angle observed in the WT 2'-OMe structure was 158° [26]. The orientation of the leaving group in the A38G structure superimposed well with that of the vanadate structure (PDB ID 2P7E), although the differences in the position of the phosphorus, pro-R and pro-S oxygens was similar to that described previously for other 2'-OMe structures [16, 26, 27]. The significance of this ideal O5' position is not known, but could suggest over-stabilization of the ground state by the A38G modification in the context of a 2'-OMe modification.

In contrast, the 2'-5' structure suggested that the A38G mutant caused destabilization of the transition state, and thus provided strong evidence for the poor activity. The S-turn of this structure was disordered, although the region of the Gua38 base was fairly well defined (Fig. 3.3d). Evidence for 2 distinct conformations of the S-turn was observed and modeled accordingly. The two conformations were modeled with equal occupancy in an effort to
optimize $R_{\text{work}}$ and $R_{\text{free}}$ while minimizing the differences in local $B$-factors. In the conformation that was most similar to the WT ribozyme (Fig. 3.7b), the Gua38 base was repositioned to avoid a steric clash between the O2' of A-1 and the O6 atom of Gua38. The superposition of the WT and A38G structures revealed a 1.6 Å separation between the N6 atom of Ade38 and the O6 atom of Gua38. The N2 amine of Gua38 appeared to stabilize the shifted conformation of Gua38, as it was positioned to engage in H-bonds with the 2'OH from U37 (2.8 Å), the O4' of G+1 (2.7 Å), and the O1P atom of its own phosphate (2.8 Å). These H-bond distances were shorter than those observed in the A382AP 2'-5' structure (3.8, 3.6, and 3.7 Å, respectively), and constitute specific stabilizing interactions. W3 was not observed in the A38G structure. In fact, the shifted position of the N2 atom observed in the A38G structure (1.4 Å closer to the 2'OH of U37 as compared to DAP) precluded the presence of W3. The shifted position of the Gua38 base in this 2'-5' structure was probably caused by repulsion between the O6 atom of Gua38 and the pro-R oxygen of G+1. This is supported by consideration of the 2'-OMe structure, where the N2 atom of Gua38 superimposed well on the N2 atom of Dap38 and 2ap38. The pro-R oxygen in the 2'-OMe structure was not positioned near Gua38's O6 atom. Without a repulsion to drive the repositioning of the Gua38 base, the distances observed between the N2 atom of Gua38 and the O4', O1P, and O2' atoms mentioned above were less conducive to H-bond formation (3.1, 4.5 and 4.4 Å, respectively) in the 2'-OMe structure. In fact, a small $mF_o$-$DF_c$ peak suggested that W3 was occasionally present in the A38G 2'-OMe structure, albeit at a much lower occupancy than that of other 2'-OMe structures.

The second conformation of the S-turn in the A38G 2'-5' structure avoided a steric clash between the O6 atom of G38 and the pro-R oxygen through a refolding event. In this conformation, the N2 amine of Gua38 was positioned into a pocket occupied by O6 in the above alternate conformation (Fig. 3.7c). As such, the N2 atom of Gua38 was superimposed 0.9 Å from the position of the N6 atom of Ade38 in the WT ribozyme. This required a significant refolding of the S-turn residues, such that if the two Gua38 models were
superimposed, only the bases would overlay (Fig. 3.7d). In this manner, the base-stacking interaction with G+1 was conserved in both alternate conformations. However, the placement of the ribose and phosphate moieties was unique and non-overlapping. In the ‘mis-folded’ conformation (Fig. 3.7c), the base of Gua38 exhibited an anti orientation that directed the N2 atom toward the O2’ and pro-R atoms of the active site. This unique orientation allowed a 3.3 Å interaction between the N2 atom of Gua38 and the O2’ atom of A-1. The anti configuration was further stabilized by a 3 Å H-bond between the O6 atom of Gua38 and the N6 atom of Ade24.

3.6 Discussion

Experiments designed to investigate the chemistry of the phosphodiester cleavage reaction inevitably influence the local structure as well, thus complicating the interpretation of results. The hairpin ribozyme presents an advantage in this regard, as it is well-suited to both structural and functional investigations. This combination makes it an ideal model system for the study of RNA catalysis. The crystallographic studies presented here offer a window into the perturbations caused by mutating the Ade38 base and lend insight into the existing kinetic characterizations [20].

The structural results presented here emphasize the sensitivity of the hairpin ribozyme active site structure. Superpositions of similarly modified ribozymes offered insight into the respective stabilities of the ribozyme active site in both a ground state and reaction intermediate conformation. Specifically, in the context of the 2´-OMe A-1 modification, Ade38 mutations exhibited their greatest influence on the orientation of the scissile phosphate. Ade38 mutations made in the context of the 2´-5´ modification, however, primarily influenced the residues of the S-turn. This observation suggests that the local environment of the active site stabilizes the transition state of the scissile phosphate more so than the ground state, in following with the general principles of catalysis.
A comparison of the 2´-OMe and 2´-5´ modifications also enabled some parsing of the influence of each Ade38 mutant. For example, the A38G mutation destabilized the S-turn only in the context of the 2´-5´ modification, a finding that suggested this mutation destabilized the transition state, rather than the ground state. In contrast, the A382AP mutation altered the scissile phosphate conformation in the context of either active site modification.

The structures described here further emphasize the importance of the N1 atom of Ade38. The A38C structures were particularly beneficial in this regard. The 8000-fold loss in activity appears attributable to the loss of the imino nitrogen, caused by the anti configuration of the Cyt38 residue. Significantly, the placement of the exocyclic amine and associated water molecule (W52, Fig. 3.6), was conserved between the A38C and WT structures. This contrasts with the A382AP structure, which exhibited a conserved placement of the N1 atom, but lacked an N6 amine. The superior reactivity of the A382AP ribozyme (160-fold more reactive than A38C) suggests that N1 is more crucial to catalysis than the N6 amine.

Removal of the N6 amine was, however, disruptive to the structure of both the scissile phosphate and a key water molecule. This water (W52, WT and A38DAP Fig. 3.4b; A38C Fig. 3.6b) has been proposed as a specific base catalyst [25, 46], although it was replaced by the pro-R oxygen of G+1 in all ribozyme structures representative of a reaction intermediate [27], and by the 2′O-Me moiety in the A38C-Mg$^{2+}$ structure. Based on these observations one might suggest that W52 is an artifact of the 2´-endo sugar pucker observed at A-1 in most 2´-OMe modified ribozymes and does not participate in catalysis. The strong preference for an electronegative atom at this position might be satisfied in vivo first by the O2´ atom of A-1, and then by the pro-R oxygen of G+1. If W52 serves as a specific base catalyst for the 2′OH, there must be some minor rearrangement in the ground state that avoids a clash with the O2´ atom. Additionally, the 2′5′ structures suggest that W52 must vacate this pocket during formation of the transition state.
In short, the role of W52 cannot be determined from these studies. The A38C structure contained W52, but suffered from an 8000-fold loss in activity. Transition state formation would not be expected to be impaired in this structure. The 2′-OMe A382AP structure showed weak evidence of a disordered W52, and suffered a 50-fold loss in activity. The active site of the 2′-5′ modified A382AP structure suggested that loss of the N6 amine destabilized the transition state. The 2′-OMe A38G structure was well ordered but did not contain any evidence of W52. This mutant exhibited a 13,000-fold loss in activity, and the A38G 2′-5′ structure demonstrated that formation of the transition state could not occur without significant refolding of the S-turn. Taken together, these results might suggest that the observation of W52 in 2′-OMe modified structures is indicative of transition state stability.

These data support previous observations that there is a preference for an electronegative atom at this position, regardless of whether that atom is a water or an RNA atom. In this regard, it is interesting that W52 was not present in the 2.35 Å A38G 2′-OMe structure. Water could donate an H-bond to the O6 atom just as easily as it could accept an H-bond from an amine. Thus, the absence of W52 in this structure highlights the intricate nature of the ribozyme’s electronic environment, and would seem to suggest that the solvent network observed in the ribozyme active site is not amenable to rearrangement. The conserved placement of the pro-R oxygen across the suite of 2′5′ modified Ade38 mutant structures (Fig. 3.8b) further emphasizes the robust stabilization that is provided to this position by the ribozyme and its associated solvent network.

In investigating the role of waters, it is significant to note that none were observed to mediate an interaction between the N1 atom and the scissile phosphate, regardless of the orientation or type of residue at position 38. The catalytic contributions of the N1 atom are likely to be direct. The only other water that mediated an interaction between residue 38 and the scissile phosphate was W3. W3 was positioned to accept an H-bond from the 2′OH of Urr37, as well as the N2 amine of Dap38 or 2ap38. It was positioned to donate an H-bond to the O4′ of G+1, and in the context of the A38C 2′-OMe structure, also to the O5′ atom of G+1.
Notably, a similar interaction was previously observed in the G8A structure [26]. This interaction might seem to implicate W3 as a specific acid in the cleavage reaction, where it could donate a proton to the leaving group. However, in the vanadate structure (which represents the closest known mimic of the transition state), and the product mimic structure (which exhibited a free 5'OH atom at G+1), W3 was ~5 Å from the O5' leaving group. Lastly, the large swing in the scissile phosphate that initiated the W3-O5' interaction in the A38C 2'-OMe structure was not recapitulated in A38C 2'-5' structure.

It therefore seems likely, given the prevalence of electronegative groups in the pocket surrounding W3, that this molecule neutralizes the local build-up of negative charge. A number of experiments that disrupted the H-bond between W3 and the 2'OH of Uri37 exhibited fairly small effects on cleavage rates [25, 49].

In summary, these structures provide a rationale for the kinetic parameters of four Ade38 mutants. Each mutant was analyzed in the context of modifications representative of both the ground state and a reaction intermediate. This approach was validated by the unique structural perturbations observed in each context. The altered folding events observed for the A38C and A38G mutants highlight the underappreciated role of the Ade38 residue in achieving a catalytically competent fold at the active site of the hairpin ribozyme.

3.7 Acknowledgements

C.M. was supported in part by an Elon Huntington Hooker Fellowship. Additional funding provided by grants to J.E.W. from the NIH/NIGMS R01 GM63162 and the Petroleum Research Fund 45534-AC 4. Gratitude is expressed to the staff of the Cornell High Energy Synchrotron Source (CHESS, Ithaca NY) for assistance with X-ray data collection. CHESS is supported by the NSF under award DMR-0225180 and the NIH through NCRR award RR-01646.
References


### Table 3.1. Intensity and Refinement Statistics

<table>
<thead>
<tr>
<th>Construct</th>
<th>A38DAP 2’-5’ pH6.8 PDBID:</th>
<th>A38DAP 2’Ome pH7.0 PDBID:</th>
<th>A382AP 2’-5’ pH6.5 PDBID:</th>
<th>A382AP 2’Ome pH7.6 PDBID:</th>
<th>A38G 2’-5’ pH 7.8 PDBID:</th>
<th>A38G 2’Ome pH8.6 PDBID:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range (Å)</td>
<td>2.70 to 40.48 (2.70 to 2.80)</td>
<td>2.25 to 40.49 (2.25 to 2.35)</td>
<td>2.75 to 31.04 (2.75 to 2.85)</td>
<td>2.35 to 23.31 (2.35 to 2.43)</td>
<td>2.65 to 29.8 (2.65 to 2.74)</td>
<td>2.35 to 44.08 (2.35 to 2.43)</td>
</tr>
<tr>
<td>Space group</td>
<td>P6₁22</td>
<td>P6₁22</td>
<td>P6₁22</td>
<td>P6₁22</td>
<td>P6₁22</td>
<td>P6₁22</td>
</tr>
<tr>
<td>Cell dimensions (Å)</td>
<td>a = 93.5, c = 133.8</td>
<td>a = 93.5, c = 133.4</td>
<td>a = 94.0, c = 134.3</td>
<td>a = 93.2, c = 132.8</td>
<td>a = 93.4, c = 132.3</td>
<td>a = 93.6, c = 131.7</td>
</tr>
<tr>
<td>Solvent content (%)</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td>No. unique reflections</td>
<td>9960 (962)</td>
<td>16724 (1615)</td>
<td>9589 (935)</td>
<td>14740 (1433)</td>
<td>10328 (1020)</td>
<td>14645 (1416)</td>
</tr>
<tr>
<td>Average redundancy</td>
<td>7.8 (6.1)</td>
<td>10.4 (8.8)</td>
<td>9.8 (6.5)</td>
<td>6.5 (6.4)</td>
<td>10.2 (11.7)</td>
<td>11.7 (9.0)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.4 (99.9)</td>
<td>98.5 (98.1)</td>
<td>99.5 (99.4)</td>
<td>99.6 (100.0)</td>
<td>99.1 (99.9)</td>
<td>99.1 (99.4)</td>
</tr>
<tr>
<td>R&lt;sub&gt;sym&lt;/sub&gt; (%)&lt;sup&gt;²&lt;/sup&gt;</td>
<td>5.0 (45.9)</td>
<td>3.7 (47.6)</td>
<td>6.1 (46.6)</td>
<td>3.1 (44.9)</td>
<td>5.7 (47.1)</td>
<td>5.3 (47.9)</td>
</tr>
<tr>
<td>I/σ (I)</td>
<td>23.8 (3.0)</td>
<td>25.5 (3.5)</td>
<td>23.2 (3.0)</td>
<td>23.4 (3.3)</td>
<td>20.7 (3.8)</td>
<td>25.6 (3.9)</td>
</tr>
<tr>
<td>No. RNA atoms</td>
<td>1327</td>
<td>1329</td>
<td>1326</td>
<td>1333</td>
<td>1390</td>
<td>1329</td>
</tr>
<tr>
<td>No. ions</td>
<td>2 Co(NH&lt;sub&gt;3&lt;/sub&gt;)&lt;sup&gt;3+&lt;/sup&gt; 1 SO₄&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>2 Co(NH&lt;sub&gt;3&lt;/sub&gt;)&lt;sup&gt;3+&lt;/sup&gt; 1 SO₄&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>2 Co(NH&lt;sub&gt;3&lt;/sub&gt;)&lt;sup&gt;3+&lt;/sup&gt; 1 SO₄&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>2 Co(NH&lt;sub&gt;3&lt;/sub&gt;)&lt;sup&gt;3+&lt;/sup&gt; 1 SO₄&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>2 Co(NH&lt;sub&gt;3&lt;/sub&gt;)&lt;sup&gt;3+&lt;/sup&gt; 1 SO₄&lt;sup&gt;2-&lt;/sup&gt;</td>
<td>2 Co(NH&lt;sub&gt;3&lt;/sub&gt;)&lt;sup&gt;3+&lt;/sup&gt; 1 SO₄&lt;sup&gt;2-&lt;/sup&gt;</td>
</tr>
<tr>
<td>R&lt;sub&gt;factor&lt;/sub&gt;/ R&lt;sub&gt;work&lt;/sub&gt;/ R&lt;sub&gt;free&lt;/sub&gt; (%)&lt;sup&gt;²&lt;/sup&gt;</td>
<td>22.4 / 22.3 / 25.0</td>
<td>22.3 / 22.3 / 24.1</td>
<td>21.8 / 21.6 / 24.3</td>
<td>22.8 / 22.8 / 23.8</td>
<td>23.7 / 23.6 / 26.2</td>
<td>21.8 / 21.6 / 23.8</td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.007</td>
<td>0.009</td>
<td>0.008</td>
<td>0.007</td>
<td>0.006</td>
<td>0.008</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.5</td>
<td>1.6</td>
<td>1.4</td>
<td>1.5</td>
<td>1.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Coordinate Error*</td>
<td>0.54</td>
<td>0.46</td>
<td>0.73</td>
<td>0.53</td>
<td>0.51</td>
<td>0.60</td>
</tr>
<tr>
<td>B factor RNA (Å&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>80.0</td>
<td>70.1</td>
<td>93.8</td>
<td>66.5</td>
<td>83.2</td>
<td>74.2</td>
</tr>
<tr>
<td>B factor Water (Å&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>60.6</td>
<td>70.1</td>
<td>80.3</td>
<td>68.3</td>
<td>82.5</td>
<td>71.3</td>
</tr>
</tbody>
</table>
### Table 1

Intensity and Refinement Statistics for Hairpin Ribozyme Constructs

<table>
<thead>
<tr>
<th>Construct</th>
<th>A38C 2’5’ pH6.5 PDB ID:</th>
<th>A38C 2’Ome pH6.5 PDB ID:</th>
<th>A38CMg 2’Ome pH 6.0 PDB ID:</th>
<th>A38G 2’5’ pH 7.8 PDB ID:</th>
<th>A38G 2’Ome pH6.8 PDB ID:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range (Å)</td>
<td>2.75 to 30.95 (2.75 to 2.85)</td>
<td>2.65 to 30.86 (2.65 to 2.74)</td>
<td>2.25 to 30.89 (2.25 to 2.33)</td>
<td>2.65 to 29.8 (2.65 to 2.74)</td>
<td>2.35 to 44.08 (2.35 to 2.43)</td>
</tr>
<tr>
<td>Space group</td>
<td>P6;22</td>
<td>P6;22</td>
<td>P6;22</td>
<td>P6;22</td>
<td>P6;22</td>
</tr>
<tr>
<td>Cell dimensions (Å)</td>
<td>a = 93.5, c = 134.0</td>
<td>a = 93.6, c = 133.5</td>
<td>a = 93.6, c = 133.6</td>
<td>a = 93.4, c = 132.3</td>
<td>a = 93.6, c = 131.7</td>
</tr>
<tr>
<td>Solvent content (%)</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td>No. unique reflections</td>
<td>9471 (920)</td>
<td>10427 (1023)</td>
<td>16729</td>
<td>10328 (1020)</td>
<td>14645 (1416)</td>
</tr>
<tr>
<td>Average redundancy</td>
<td>13.8 (10.5)</td>
<td>5.9 (4.6)</td>
<td>11.05 (10.97)</td>
<td>10.2 (11.7)</td>
<td>11.7 (9.0)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.6 (99.9)</td>
<td>98.4 (99.7)</td>
<td>98.2 (99.9)</td>
<td>99.1 (99.9)</td>
<td>99.1 (99.4)</td>
</tr>
<tr>
<td>$R_{sym} (%)$</td>
<td>5.1 (49.1)</td>
<td>4.9 (39.9)</td>
<td>4.0 (54.1)</td>
<td>5.7 (47.1)</td>
<td>5.3 (47.9)</td>
</tr>
<tr>
<td>$I/\sigma (I)$</td>
<td>34.4 (4.0)</td>
<td>20.5 (3.6)</td>
<td>25.8 (4.4)</td>
<td>20.7 (3.8)</td>
<td>25.6 (3.9)</td>
</tr>
<tr>
<td>No. RNA atoms</td>
<td>1324</td>
<td>1326</td>
<td>1326</td>
<td>1390</td>
<td>1329</td>
</tr>
<tr>
<td>No. ions</td>
<td>3 Co(NH$_3$)$_6$^{3+} 1 SO$_4^{2-}$</td>
<td>3 Co(NH$_3$)$_6$^{3+} 1 SO$_4^{2-}$</td>
<td>1 Mg$^{2+}$ 1 SO$_4^{2-}$</td>
<td>2 Co(NH$_3$)$_6$^{3+} 1 SO$_4^{2-}$</td>
<td>2 Co(NH$_3$)$_6$^{3+} 1 SO$_4^{2-}$</td>
</tr>
<tr>
<td>$R_{factor}/R_{work}/R_{free} (%)$</td>
<td>19.7 / 19.6 / 22.1</td>
<td>20.6 / 20.6 / 21.8</td>
<td>--/ 21.9 / 24.6</td>
<td>23.7 / 23.6 / 26.2</td>
<td>21.8 / 21.6 / 23.8</td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.006</td>
<td>0.008</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.5</td>
<td>1.8</td>
<td>1.7</td>
<td>1.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Coordinate Error*</td>
<td>0.51</td>
<td>0.60</td>
<td>0.61</td>
<td>0.51</td>
<td>0.60</td>
</tr>
<tr>
<td>B factor RNA (Å$^2$)</td>
<td>88.3</td>
<td>83.7</td>
<td>67.5</td>
<td>83.2</td>
<td>74.2</td>
</tr>
<tr>
<td>B factor Water (Å$^2$)</td>
<td>84.4</td>
<td>75.9</td>
<td>70.1</td>
<td>82.5</td>
<td>71.3</td>
</tr>
</tbody>
</table>

* Parenthetical values correspond to the highest resolution shell.

$R_{sym} = \frac{\sum |I(h)| - <I(h)>}{\sum |I(h)|} \times 100$, where $I(h)$ is the observed intensity of the jth measurement of reflection $h$ and $<I(h)>$ is the mean intensity of reflection $h$.

$R_{factor} = \frac{\sum |F_o| - |F_c|}{\sum |F_o|} \times 100$, where $F_o$ and $F_c$ are the observed and calculated structure factor amplitudes. $R_{factor}$ is calculated using the residual target in CNS with all reflections in the resolution range of refinement. $R_{work}$ is calculated using all reflections except those randomly removed for the test set (9%). $R_{free}$ is calculated using the test set of reflections.

* Based upon cross-validated plots of sigma A coefficients vs. resolution for full resolution range.
Figure 3.1 Reaction and nucleotide schematics  (A) Schematic of the transesterification reaction of the hairpin ribozyme. (B) Structure and pKₐ values of nucleobases investigated in this study. Solution pKₐ values are for a 5' monophosphorylated nucleotide, as reported in refs [50] and [48]. Bases are presented in decreasing order of reactivity as reported at pH 7.5 in [20].
Figure 3.2 Depictions of the global hairpin ribozyme fold and RNA sequences used in this investigation. (A) Secondary structure of the minimal hinged hairpin ribozyme adapted from RNAview [45]. The substrate strand is depicted in green, the S-turn strand in red, and the 29mer strand in beige. The linker position is colored orange. The helix-loop-helix domains are labeled loop A and loop B; helices are labeled H1-H4. The star between A-1 and G+1 indicates the site that was modified by either a 2′-OMe or 2′-5′ phosphodiester linkage. Conserved residues are boxed; blue backgrounds indicate residues of the ribose zipper; yellow boxes indicate E-loop residues; gray circled residues belong to the S-turn. Hydrogen bond pairings: open-square, Hoogsteen; open triangle, trans-sugar; open circle, Watson-Crick face; closed circle, wobble pair. Double and single lines indicate Watson-Crick pairs; black dashed lines indicate single hydrogen bonds. (B) A ribbon diagram of the minimal hinged hairpin ribozyme (prepared with Nuccyl [44]) is shown in an orientation equivalent to that of the sequence schematic in panel A, then rotated 90° into the conformation represented in subsequent figures. The boxed area is enlarged in the inset and highlights the position of A-1, G+1, A38 and C39.
Figure 3.3  Representative $\sigma_A$-weighted simulated annealing omit electron density maps for the nucleobase at position 38 and the A-1 and G+1 residues that border the scissile bond The blue mesh represents $mF_o-DF_c$ maps contoured at 4$\sigma$; the orange mesh is contoured at 12$\sigma$. The model is shown in stick representation; C1' of the ribose at 38 is shown for clarity but was not omitted during the map calculation. Inset panels display the electron density observed for the nucleobase at residue 38 and were generated by a 90° rotation, toward the reader, of the larger panel. Nitrogen atoms are colored blue, oxygen atoms are red, and phosphorus atoms are orange. Stick representations are colored as in Fig. 3.2. Left hand panels depict the 2´-OMe modification; right hand panels depict the 2´-5´ modification. (A-D) The Ade38 variants as they are presented in the text: (A) A38DAP, (B) A382AP, (C) A38C, (D) A38G. The middle panel depicted in (C) represents the 2´-OMe modified A38C structure that was solved in the presence of Mg$^{2+}$. 
**Figure 3.4 Superpositions of the A38DAP mutants on the WT ribozymes**

Stick representation of the A38DAP structure is shown in green; WT is shown in grey. Perspective for the full panel and inset are as shown in Fig. 3.3, except that the G+1 phosphate is included in the inset to detail the scissile phosphate conformation in addition to the Dap38 nucleobase. Active site waters are represented as spheres in the same color as the RNA. WT waters are shown only when the equivalent water was observed in the mutant structure. Nitrogen atoms are shown in blue, oxygen in red, phosphorus in orange. Putative H-bond interactions are indicated in dashed grey lines, distances are reported in angstroms. Pro-R oxygens are indicated by arrows. (A) A38DAP variant with a 2'-5' modification at the active site. (B) Stereo diagram of the A38DAP variant with a 2'-OME modified active site.
Figure 3.5  Superpositions of A382AP structures on WT ribozymes

The A382AP structure is shown in orange. Details are as described in Fig. 3.4. (A) The 2’-OMe modified A382AP variant exhibited 2 alternate conformations, depicted in orange (60% occupancy) and yellow (40% occupancy). The WT 2’-OMe structure is shown in grey. (B) The 2’-5’ modified A382AP variant shown superimposed on the WT 2’-5’ structure.
Figure 3.6  Depictions of the A38C mutant structures  The A38C structures are shown in red, WT in grey.  (A) A superposition of the active sites of the 2'5' modified A38C and WT structures is shown.  (B) A superposition of the 2'-OMe modified A38C and WT active sites is depicted. The equivalent W52 molecules are separated by 1.3 Å in the A38C-Co(NH₃)₆ and WT structures.  (C) Stabilizing interactions for the Cyt38 anti configuration and the novel Co(NH₃)₆ ion observed in the A38C-Co(NH₃)₆ variant are depicted. The substrate strand is green; the 29-mer strand is brown; the S-turn strand is red. Co(NH₃)₆ is yellow. The nucleobase of Ade24 is engaged in two H-bonds with the nucleobase of Cyt38. Ade23 stacks on top of Ade24; the O2P atom of Ade23 coordinates to an amine group of Co(NH₃)₆.  (D) A stereo figure of the A38C-Mg²⁺ active site superimposed on the 2'-OMe modified WT ribozyme.
Figure 3.7  Superpositions of A38G structures onto WT ribozymes  The A38G structures are shown in blue, WT in grey. (A) The A38G 2´-OMe modified structure is shown superimposed on the WT ribozyme. (B) The A38G 2´-5´ structure is shown superimposed on the WT ribozyme. The dark blue stick representation of the Gua38 base depicts the conformation most similar to the WT ribozyme. (C) A38G 2´-5´ structure depicting the ‘mis-folded’ conformation of Gua38 is shown superimposed on the WT ribozyme. (D) A superposition of the base and ribose moieties of Gua38 and the base moiety of the WT ribozyme illustrates the large change in conformation. The standard S-turn fold is depicted in dark blue for the A38G structure; the mis-folded conformation is in light blue. The riboses are well separated; however, the base-stacking between Gua38 and G+1 is well conserved.
Figure 3.8  The suite of Ade38 mutant structures  (A) A superposition of the 2’-OMe modified Ade38 mutant structures illustrates the variation in position of the scissile phosphate. Color scheme is conserved from Figs. 3.4-3.7; A38DAP is green; A382AP is orange; A38C is red; A38G is blue. The inset panel displays the base at position 38 and the G+1 phosphate, with oxygen atoms colored red and nitrogen atoms blue.  (B) A superposition of the 2’5’ modified Ade38 mutant structures illustrates the conservation in the orientation of the scissile phosphate. Colors are as described for (A), with the addition that the ‘folded’ Gua38 nucleobase is shown in dark blue, whereas the ‘misfolded’ Gua38 base is shown in light blue.
Chapter 4:

Chemistry, Phylogeny and Three-Dimensional Structure of the APOBEC Protein Family


Wedekind, J.E. contributed to this work.
4.1 Footnotes

Abbreviations: ACF, APOBEC Complementation Factor; ADAR, Adenosine Deaminase that Acts on RNA; ADAT, Adenosine Deaminase that Acts on tRNA; AID, Activation Induced Deaminase; APOBEC, APOlipoprotein B Editing Catalytic subunit; CD, Catalytic Domain (Section 3.6.1); CD, Cytosine Deaminase (Section 3.7); CDA, Cytidine Deaminase; Cdd1, Cytidine Deaminase 1 (S. cerevisiae); dCD (Deoxycytidylate Deaminase); dCMP, 2’-deoxycytidine 5’-monophosphate; dCTP, 2’-deoxycytidine 5’-triphosphate; dsRBD, double stranded RNA Binding Domain; dTTP, 2’-deoxythymidine 5’-triphosphate; EST, Expressed Sequence Tag; fnCDA, free nucleotide Cytidine Deaminase; FRET, Fluorescence Resonance Energy Transfer; GD, Guanine Deaminase; HIV, Human Immunodeficiency Virus; HMM, High Molecular Mass; HSP70, Heat Shock Protein 70 (kD); LINE, Long Interspersed Element; LMM, Low Molecular Mass; NCD, NonCatalytic Deaminase domain; ORF, Open Reading Frame; PCD, Pseudo-Catalytic Domain; RibG, Riboflavin biosynthesis protein G; SAXS, Small Angle X-ray Scattering; SIV, Simian Immunodeficiency Virus; TadA, tRNA Adenosine Deaminase A; UTR, UnTranslated Region; Vif, viral infectivity factor; ZDD, Zinc Dependent Deaminase
4.2 Abstract:

The ability of the mammalian enzyme APOBEC1 to catalyze C-to-U deamination or ‘editing’ of the apolipoprotein B mRNA, thereby generating a truncated protein of altered function, has been known for nearly two decades. The recent discovery of a family of human APOBEC1-related proteins has stirred excitement in the field, especially since several family members edit single-strand (ss)DNA, and have proven essential for immunoglobulin gene diversification, as well as innate antiretrotransposon and antiretroviral activities. Although crystal structures of APOBEC family members would do much to elucidate structure/function relationships at the molecular level, such analyses are in a nascent stage at present. Nonetheless, in this chapter we provide a demonstration of the available tools and approaches that can be employed to glean useful three-dimensional information about APOBEC1 related proteins, which should prove useful to gain molecular insight and guide experimentation until high-resolution structures become available.
“If you want to understand function, study structure”.
-- Francis H.C. Crick (in *What Mad Pursuit*)

4.3 Introduction to Nucleic Acid Deamination with Implications for Biological Activity

Nucleobase oxidation in the context of DNA or RNA is generally considered an undesirable biological phenomenon due to its potential to threaten genomic integrity. However, there are a growing number of circumstances in higher organisms in which enzymes of the APOBEC family perform site-specific deamination of DNA or RNA substrates. APOBEC enzymes are highly specialized and are required for such activities as proteomic diversification, codon modification, adaptation of the humoral immune response, and innate cellular immunity against invading retroviruses or retrotransposable elements. As one can imagine, these 'editing' activities must be highly regulated and compartmentalized since indiscriminate DNA deamination can lead to genome instability. Indeed, the overexpression or misregulation of APOBEC deaminases has been associated with the development of cancers, lymphomas and other neoplastic diseases [1-10]. In this chapter, we will be concerned especially with the chemistry of deamination mediated by the action of APOBEC family members. Using a hierarchical approach, it will become apparent (i) how a defining amino acid signature sequence (the Zn\(^{2+}\) dependent deaminase motif) folds into a conserved three-dimensional structure, (ii) how this fold is part of a conserved subunit that is shared by a diverse group of deaminase enzymes, and (iii) how variations to the core fold impart specialization in terms of biological function. With regard to the latter point, structural precedents drawn from well characterized enzymes will be used to gain insight into the APOBEC family, for which there is limited structural data. A review of current three-dimensional data for the APOBEC family will be followed by a phylogenetic, sequence-driven analysis with the aim of relating family members to each other. Finally, a discussion of the
influences of selective pressure and chromosome remodeling highlights the evolutionary forces that have sculpted the form and function of the APOBEC family.

4.4 The Chemistry of the Zinc Dependent Deaminase Amino Acid Signature Motif

The enzymes that catalyze the deamination of cytidine to uridine are known as cytidine deaminases or CDAs (Enzyme Classification or EC 3.5.4.5). Much of the chemistry and structural details of the CDA reaction were established by Wolfenden & Carter, who demonstrated that the Zn$^{2+}$ ion is bound tightly by the His and Cys residues of the bacterial enzyme from *E. coli* that acts on the free nucleosides [11]. The Zn$^{2+}$ ion is an essential catalytic component that serves as a Lewis Acid to activate a water molecule for nucleophilic aromatic substitution at the C4 position of the cytosine ring [12-14] (Fig. 4.1). As such, maintenance of the Zn$^{2+}$ ion coordination and stereochemical positioning of water are essential aspects of the CDA mechanism of action. In function studies, site-directed mutations of the respective Cys or His residues to Ala resulted in activity losses on the order of $10^3$-fold in $k_{cat}$ and 3- to 10-fold in $K_M$ relative to wild type [11]. These changes, as well as metal composition analysis of the mutant enzymes, suggested that loss of function is mostly due to the absence of Zn$^{2+}$ participation in catalysis, rather than its participation in substrate binding. A conserved Glu serves as a proton shuttle that facilitates transfer of hydrogen from the activated water to the imino N3 position of the pyrimidine ring [14]. Site directed mutagenesis of this Glu to Ala resulted in a $10^6$-fold loss in $k_{cat}$, but a 30-fold gain in affinity for the cytidine substrate as reflected in $K_M$. A comparable gain in affinity was attained for the product uridine, as measured by the equilibrium $K_D$ [15]. These results suggest that Glu contributes significantly to the chemistry of catalysis, but not to substrate binding. Collectively, these amino acids form part of a conserved amino acid signature sequence known as the Zinc Dependent Deaminase (ZDD) motif, which implies a common chemistry and three-dimensional fold [16]. Detection of this active site constellation within the amino
acid sequence of a protein can be distilled into the pattern (H/C)xEx_{25-30}PCxxC, where x is any residue [17]. As such, the ZDD motif is a defining characteristic of the CDA superfamily that includes a number of pyrimidine metabolism deaminases, as well as APOBEC family members. Moreover, the ZDD sequence is also a common feature of Adenosine Deaminases that Act on tRNA (ADATs) and Adenosine Deaminases that Act on RNA (ADARs). Although the latter enzymes deaminate adenosine to inosine, conservation of the ZDD motif is evidence of a common ancestor, which supports classification of these molecules into distinct clades based on amino acid sequence identity, but ultimately establishes their grouping into a single CDA superfamily.

4.5 The ZDD Signature Motif Implies a Specific Three-Dimensional Arrangement of Amino Acids

To the structural biologist, the ZDD motif invokes not only an image of the catalytic residues within the linear amino acid sequence, but also their arrangement in three-dimensional space leading to stereo-specific substrate binding and catalysis [16]. Residues of the ZDD interact in the context of a helix-strand-helix super-secondary structure element (Fig. 4.2a). This motif is integral to the formation of the CDA active site and is representative of all known structures of this superfamily. In this motif, the three Zn$^{2+}$ coordinating residues reside at the N-terminal ends of two adjacent α-helices that are connected topologically by a single, intervening β-strand which imparts a right-handed crossover element (Fig. 4.2a). The coordination geometry of the Zn$^{2+}$ ion is tetrahedral and is fixed by the three ZDD ligands such that a water molecule completes the coordination sphere. This activated water is located ~2.8 Å from atom C4 on the si (i.e., left handed) face of the cytosine ring in the crystal structure, as represented by the CDA enzyme structure from mouse (Fig. 4.2a). In contrast, the leaving group, NH$_3$, departs from the cytosine ring from the opposite re or right-handed face. The departure of ammonia is guided via formation of a hydrogen bond between NH$_3$
and the carbonyl oxygen of the residue preceding a conserved Pro of the ZDD. The role of this Pro is likely to be conformational restriction of the backbone, while simultaneously assuring the presence of a pocket to receive the leaving group [14, 18]. In the crystal structure of the *E. coli* CDA in complex with the transition-state mimic 5-fluorozebularine, the attacking water hydrogen bonds to the amide backbone of Cys immediately C-terminal to Pro in the PCxxC motif [14]. Therefore, the presence of Pro has a direct influence on the orientation of catalytic groups through positioning of the local backbone. Moreover, the evolutionary selection of an imino acid at this position (i.e., one that is missing a backbone amide to donate a hydrogen bond) may serve to eliminate non-productive hydrogen bond interactions between the attacking water and the main-chain on both si and re faces of the cytosine ring. Finally, the conserved Glu residue of the ZDD motif resides at the N-terminal end of the first α-helix where it is poised to serve as a proton shuttle as depicted in Fig. 4.1. In the crystal structure, this activity appears plausible due to the apparent hydrogen bonds to the Watson-Crick face of cytidine at the N3 (imino) position (Fig. 4.2a, broken pink lines), as well as the leaving ammonia (not shown).

Despite the high level of detail known about the mode of CDA interactions with individual nucleoside substrates, no structural information exists for any member of the APOBEC family on the mode of single-stranded DNA or RNA recognition. As such, crystal structures of individual APOBEC family members will provide a valuable resource for understanding substrate affinity and specificity. One aspect that is certain about all CDA family members is that the ZDD motif is part of a larger core structure (Fig. 4.2b). The need to embed the ZDD within this large, folded structure is sensible since the hydrophobic core of the protein imparts a great number of peripheral interactions that stabilize the helix-beta-helix motif, while simultaneously providing elements that bind and orient the substrate, such as Loop 3 (Fig. 4.2b). In this chapter, the larger α/β-fold is referred to as the ‘CDA fold’ or ‘CDA domain’, which implies independent folding of the polypeptide chain. Based upon current
knowledge, is it very likely that APOBEC family members will conform to the core CDA fold as well.

4.6 Rationale for a Combined Structural and Phylogenetic Approach to Understand APOBEC Evolution

Although not implied in Crick’s adage, it has been proven that three-dimensional structure is a better prognosticator of function than amino acid sequence identity. This is because tertiary (three-dimensional) structure is more evolutionarily conserved than primary structure [19]. In this section, we examine the relationship between comparative primary sequence analysis versus three-dimensional structure analysis. In doing so, we wish to convey that combining both methods is a powerful tool to infer function and evolutionary relatedness, even in cases of low amino acid homology.

In Section 4.5, the ZDD sequence motif was revealed as a useful consensus predictor of deaminase function in amino acid sequence searches seeking to unveil candidate members of the APOBEC family. (Note; for those in doubt try the string [HC]xEx(25,30)PCxxC in psi-blast (http://www.ncbi.nlm.nih.gov/blast/ in combination with one known CDA sequence). However, beyond the apparent amino acid sequence homology within the ZDD, overall sequence similarities among members of the CDA superfamily are often quite poor [17, 20]. Nonetheless, the secondary structure elements of the core CDA fold are spatially well conserved (Fig. 4.2b), which provides a powerful restraint in our efforts to relate APOBEC structure to function. The relationship between structural conservation and amino acid sequence identity was formalized more than 2 decades ago by Chothia and Lesk who showed that sequence similarity between structures provides a predictable metric for the variation in overall core folding [21]. This relationship established a useful limit regarding what three-dimensional information one may accurately infer about an unknown structure based solely on protein sequence. As a heuristic, the ability to confidently model a
core three-dimensional fold requires sequence identity between the empirically defined ‘template’ structure and unknown target sequence of \( \geq 35\% \) [22]; of course, the greater the identity the more reliable the result of modeling. Under circumstances of low identity (e.g., \( \leq 30\% \)), but conserved function, the core architecture and functional residues can be expected to remain largely unaltered. Changes to the peripheral loops and elements flanking the core fold are likely to be divergent, thus giving rise to specialized functions such as substrate binding, subunit oligomerization or allosteric regulation (see Sections 4.7 & 4.8). The latter two points provide some of the most powerful tenets for structure-based alignments and will be the premise of our later analysis. Yet, despite the utility of structure as a tool to relate the amino acids of proteins in a pairwise manner, phylogeny is often inferred solely through sequence alignments unguided by structural restraints. This realization is sobering, especially in light of the tens of thousands of experimentally derived, high-resolution three-dimensional structures available in the Protein Data Bank [23] (www.pdb.org).

To illustrate the power of structure-based alignment, let us consider a comparison of the amino acid sequences from the \( \alpha \)-chain of human hemoglobin, human myoglobin and leghemoglobin from lupine, a flowering plant. (Note; see ref. [19] for a more complete discussion). Despite the fact that the primary sequences of these molecules do not exhibit more than 16\% identity, the pairwise root mean square (r.m.s.) deviations between the conserved atoms of lupine hemoglobin or human myoglobin and human hemoglobin are 1.9 Å and 1.4 Å, respectively. Although not necessarily conveyed in the r.m.s numbers to the average reader, this degree of structural similarity is considered high and is interpreted by the structural biologist to be the result of selective pressure that favored the evolution of respective proteins to coordinate heme for \( \text{O}_2 \) binding and transport [24]. Significantly, the relatedness of these proteins would likely be missed with a simple primary sequence comparison [25], whereas knowledge of their common function implies structural similarity that favors the use of pairwise approaches.
Of course, there are also circumstances in which proteins possess the same fold even though they exhibit diverse biological activities. One example is the remarkable structural similarity between the ATP binding domain of the mammalian heat shock protein 70 (HSP70), solved by McKay & colleagues, and the globular actin subunit determined by Holmes [26]. Again, these proteins exhibit ~16% sequence identity, but their pairwise r.m.s. deviation is 1.6 Å between 193 shared atoms. The high degree of structural similarity suggests that they evolved from a common ancestor and subsequently diverged to fulfill special biological functions (i.e., protein folding and formation of the cytoskeleton). Such proteins are designated paralogs to indicate that they arose from a gene duplication event; this term also describes the evolutionary relationship between APOBEC1 and AID of the APOBEC family [20]. The take home message here is that structural conservation does not automatically equal functional equivalence, but it is an excellent means to establish evolutionary commonality. [Note; the reader is referred to the DALI Server as a resource to search for structural homologs based on an input three-dimensional structure [27]. See http://www.ebi.ac.uk/dali/].

As pointed out by Chothia and Lesk, the use of simple sequence alignments can be especially informative under circumstances with moderate to high identity (>40%) [21]. Under such conditions, these comparisons may establish a conserved core fold, as well as evolutionary ‘embellishments’ (deletions and insertions) that diversify an ancient fold for a variety of uses (e.g., the ATP binding domain of actin versus HSP70). Such a scenario appears to be the case for APOBEC family members. However, inferring evolutionary relatedness to more distant ZDD-containing proteins quickly reveals the tenuousness of such relationships, especially when one considers that three-dimensional folding topology, subunit oligomerization, and substrate binding may have been altered significantly through evolution to fulfill specialized biological niches. This knowledge can be crucial in the choice of a three-dimensional structural template for homology modeling [20]. In the end, there is a balancing act between inferring common structure based solely on sequence relatedness, and
assigning common function based on known three-dimensional structure (e.g., the case of paralogs). In this review, we will walk this ‘tightrope’ to relate sequence and structure, in so far as the experimental details are available. This strategy allows us to relate members of the APOBEC family to more ancient CDA folds to identify a general set of principles that provide insight into the maintenance of conserved chemical function versus core diversification elements among CDA family members.

4.6.1 The Starting Point for Structural and Phylogenetic Analyses of APOBEC Family Members.

When considering how to proceed in terms of a phylogenetic comparison between APOBEC family members and enzymes of related function, it is appropriate to ask what these proteins have in common. The clear answer is that both the conserved ZDD amino acid signature motif \((\text{H/C})_x\text{Ex}^{25-30}\text{PCxxC},\) as well as surrounding residues [17] form the basis for much of the phylogenetic classification of individual APOBEC family members (Fig. 4.3). Because it resides within the core fold and imparts deaminase activity, the ZDD signature sequence is slow to change, making it the most reliable region for sequence comparisons among family members [28-30]. To illustrate this point, we cite a comprehensive phylogenetic analysis that was generated from protein alignments of a ~150 residue segment encompassing the ZDD [31]. As seen in Section 4.6 (above) regarding the relationship between form and function, the observation that the ZDD motif is conserved across the ADAR, ADAT and AID/APOBEC1/APOBEC3 genes implies that these families share a common ancestor with the CDA, dCMP DA and Riboflavin DA families (Fig. 4.3). The ZDD commonality exists despite the diverse array of substrates targeted by each branch of the superfamily, and is significant because the ZDD motif itself is not shared by all deaminases known in biology. As such, deaminase enzymes as a whole are considered an example of convergent evolution. That is, the adenosine deaminases of purine metabolism and bacterial cytosine deaminases developed through selective pressure to achieve the same chemical
reaction as members of the CDA superfamily. However, each enzyme class achieves catalysis by its own unique three-dimensional fold, which exhibits no sequence or spatial similarity to the ZDD motif employed by CDA proteins [32, 33]. Such themes are common in biology. For example, carbonic anhydrases come in multiple forms. Whereas the α-form of the human enzyme adopts an α/β-fold, the γ-enzyme from the archaeon *Methanosarcina thermophila* is a left-handed β-helix [34]. The common feature of both enzymes is that they each coordinate to Zn$^{2+}$ to catalyze conversion of CO$_2$ into carbonic acid or HCO$_3^-$. However, these enzymes have such disparate three-dimensional folds they could not possibly have arisen from a common ancestor. Such is the case for ADAR enzymes relative to free nucleoside adenosine deaminases, suggesting it is highly improbable that the former evolved from the latter or *vice versa*.

The starting point for exploring the evolutionary relationship between APOBEC and CDA family members was the structure determination of the CDA from *E. coli* [14]. This crystal structure was published nearly a decade before any three-dimensional information became available for the RNA (or DNA) editing enzymes listed in Fig. 4.3. As such, the *E. coli* enzyme was employed as a template in early homology models for members of the APOBEC family [20, 35, 36]. In retrospect, one fascinating but cautionary aspect of the *E. coli* CDA structure was the observation that the enzyme folded as a dimer comprised of two identical polypeptide subunits related by two-fold symmetry [14]. Moreover, each polypeptide exhibited two highly conserved CDA folds itself, despite the fact that only a single ZDD motif was detectable per amino acid chain (Fig. 4). This observation led to the suggestion that the bacterial enzyme evolved by gene duplication of a single CDA domain that gave rise to a second CDA-like domain that lost its ability to bind Zn$^{2+}$ through divergent evolution [36]. Such divergence is reminiscent of the relationship between actin and HSP70 except that it occurred within the same polypeptide chain. Hence, although N- and C-terminal halves of the amino acid sequence exhibit ~16% identity, each polypeptide segment adopts a highly similar CDA fold (i.e., their Ca backbones superpose with an r.m.s. deviation of 1.5 Å). The
inability of the C-terminal CDA to bind Zn\textsuperscript{2+} or substrate led to its classification as a catalytically inert ‘pseudo-catalytic’ domain or PCD, as compared to the \textit{bona fide} catalytic domain (CD) at the N-terminus [14, 17]. Retention of the PCD as part of the \textit{E. coli} subunit fold appears to stabilize the global structure by burying a large hydrophobic area in the dimer interface. Maintenance of this interface is necessary to form the active site. The internal symmetry of the CDA subunit from \textit{E. coli} coupled with the intermolecular two-fold relationship between subunits gives rise to pseudo-222 or D2 symmetry [14].

The example of symmetry in the \textit{E. coli} CDA is important to illustrate a possible mode of folding and subunit \textit{inter}molecular organization by other CDA superfamily members. Specifically, several members of the \textit{APOBEC3} family clearly contain tandem ZDD motifs, implying two CDA domains within a single polypeptide [37]. These double-deaminase enzymes appear to coordinate zinc in both active sites, although frequently only a single CDA domain demonstrates deaminase activity [38, 39]. In this case, the inactive ZDD has been implicated in alternative functions including RNA binding [40].

Following the structure determination of the cytidine deaminase from \textit{E. coli}, the enzymes from other species were solved, including: \textit{B. subtilis}, \textit{S. cerevisiae}, human and mouse. These enzymes also exhibited an overall globular structure, but without the CD–PCD domain architecture [20, 41-43]. Instead, the equivalent of a single CD domain was the fundamental catalytic subunit. Rather than dimerization with pseudo-222 symmetry, these enzymes oligomerized as tetramers and exhibited proper 222 symmetry [20, 41]. As such, each tetrameric enzyme coordinated four Zn\textsuperscript{2+} ions rather than two as observed for the \textit{E. coli} CDA. The yeast enzyme (Cdd1) was solved in the Wedekind lab and was among a handful of CDA structures used as a template for APOBEC1 and AID comparative modeling [20]. Although classified as a pyrimidine metabolism enzyme [44], Cdd1 is distinguished by its ability to edit apolipoprotein B mRNA in reporter assays [45]. This discovery provided a tenable connection between the structure and function of metabolic nucleoside deaminases and those that act on DNA or RNA [17, 20]. The resulting comparative model of APOBEC1
was constrained to be a dimer, as documented experimentally. With evidence that both proteins were dimeric and of comparable size (APOBEC1: 27 kD vs. EcCDA: 31.5 kD), the N- and C- terminal domains were assumed to adopt similar CD–PCD architectures, whereby formation of the active site would depend on dimerization requiring loops that originate from each of the two identical polypeptide chains [20]. Overall, the APOBEC1 model satisfactorily explained dominant negative effects of the dimer [46] and how its active site accommodated single-stranded RNA [17] or DNA substrates [47, 48]. However, the model lacked sufficient molecular detail to account for substrate specificity [20, 49], thereby underscoring a need for experimental structural approaches. As tRNA editing TadA enzymes were solved, it was proposed that editing enzymes more closely mimicked the deoxycytidylate and cytosine deaminases, which typically exist as dimers rather than tetramers [50, 51]. This negated the requirement for a PCD to exist to complete the anticipated dimeric architecture of APOBEC1- or AID-like family members. Most recently, the crystal structure of APOBEC2 solidified the finding that APOBEC family members do not exhibit the CD-PCD architecture within a single ZDD, but instead fold as a single domain [52].

4.6.2 The CDA Superfamily: Overview of Conserved Fold Topology in the Core and Common Variations.

In Section 4.5, we alluded to the fact that the ZDD signature motif is embedded in a larger fold that maintains the integrity of the active site and Zn$^{2+}$ coordination by contributing core structural elements that flank the ZDD βαβ supersecondary structure (Fig. 4.2b). A comparison of available structures within the CDA superfamily demonstrates that the core of the CDA domain is a conserved fold that extends beyond the boundaries of the embedded ZDD signature motif (Fig. 4.2). This fold is characterized by a five-stranded mixed (parallel and antiparallel) β-sheet flanked by α-helices on either broad face of the pleated surface. Typically, the CDA sheet is triangular in shape when looking at its broadest face with the C-terminal strands being shorter than those of the N-terminus. Two helices contribute the
necessary residues for both Zn\(^{2+}\) coordination and catalysis (Fig. 4.2). All of the conserved ZDD residues reside on a single face of the β-sheet (Fig. 4.2b); a long N-terminal helix usually juxtaposes this motif on the opposing side of the triangular surface (Fig. 4.2b, helix α1).

The arrangement of these secondary structure elements, as one progresses along the polypeptide chain from N- to C-terminus, is α1β1β2α2β3α3β4β5 with the Zn\(^{2+}\) coordinating helices assigned as α2 and α3. However, the β-strand order in the three-dimensional structure is 2-1-3-4-5 when viewing the top of the sheet from the N- to C-terminal direction. This ‘topological’ description is useful since it defines the unique arrangement of β-strands as they are arranged spatially, which is often sufficient to define the uniqueness of a particular three-dimensional fold in relation to other known structures. In this chapter, we define a ‘standard’ orientation of the CDA that places the zinc ligands at the top of the diagrams, but behind the β-sheet, which gives an apparent topology of 5-4-3-1-2 from left to right (Fig. 4.5, top left panels). Strands β2, β3 and β4 are directed towards the active site whereas β1 is directed downwards away from the active site (Figs. 4.2b and 4.5). Variability is observed in the topological orientation of strand β5, providing the basis for dividing this superfamily into two smaller groups [53].

The free nucleotide cytidine deaminases (fnCDAs), including the *E. coli*, *B. subtilis* and mammalian CDAs, are distinguished by the downward directionality of strand β5, which points away from the catalytic zinc ion and is antiparallel to strand β4 (Fig. 4.5) [14, 41-43]. This orientation allows for a rather short loop between the fourth and fifth strands. In all other CDA superfamily members, including the editing enzymes ADAR2, TadA and APOBEC2, the orientation of strand β5 is upward towards the active site, parallel to strand β4 (Figs. 4.6-8) [52, 54, 55]. In the latter case, the loop from strand β4 to β5 is often longer because the change in direction of β5 necessitates the presence of an extended topological cross-over element. This connector region commonly exhibits helical character (Figs. 4.6 & 4.7). The connectivity of these family members is therefore α1β1β2α2β3α3β4α4β5. Further
complicating matters, even in those deaminases where \( \beta 5 \) runs antiparallel to strand \( \beta 4 \), the \( \alpha 4 \) helix is spatially, although not topologically, conserved to yield \( \alpha 1\beta 1\beta 2\alpha 2\beta 3\alpha 3\beta 4\beta 5\alpha 4 \). This is illustrated by comparing the \( B. \ subtilis \) CDA (Fig. 4.5) to the cytosine deaminase of \( S. \ cerevisiae \) (Fig. 4.6). In the bacterial enzyme, the smaller C-terminal helix follows rather than precedes \( \beta 5 \) in the polypeptide sequence. However, structural superpositions place both helices, labeled \( \alpha 4 \), in the same location. Defining the orientation of such structural elements among APOBEC3 family members is especially important since this region has been shown to bind the HIV virion infectivity factor (Vif) [56], which leads to APOBEC3G and APOBEC3F degradation through polyubiquitination and subsequent targeting to the proteasome [57, 58].

C-terminal helices are also observed in those enzymes with strand \( \beta 5 \) oriented upward, i.e., parallel to strand \( \beta 4 \). In these cases (exemplified in Fig. 4.7), the C-terminal sequences are directed upward toward the active site, and the significant diversification observed in this region is situated to provide substrate specificity with minimal disturbance to the core fold. From a practical perspective, the significance of the alternate orientation of strand \( \beta 5 \) becomes clear when designing experiments for family members that have not yet been structurally characterized. For example, the application of domain boundaries derived from a bacterial fnCDA onto an APOBEC protein would undoubtedly result in truncation to the core fold of the expressed sequence.

As observed for actin and HSP70, selective pressures yielding specialized function manifest as embellishments at the periphery of the core CDA fold. For example, the N- and C-termini are diversified commonly in the CDA superfamily, as well as the loops that connect the core \( \beta \)-strand and \( \alpha \)-helical secondary elements. Loop 3 (L3) is an especially common site of diversification, as it occurs between strand \( \beta 2 \) and helix \( \alpha 2 \) in the ZDD motif and is therefore situated near the active site (Fig. 4.2b); L3 commonly exhibits helical character. However in T4 dCMP deaminase, a sixty residue insertion in this region coordinates a second, non-catalytic, \( \text{Zn}^{2+} \) ion [59] (Fig. 4.7, middle panel). In comparison, APOBEC2 exhibits a rather shortened L3 (Fig. 4.7, lower panel) as a result of an elongated strand \( \beta 2 \).
This compensatory alteration leaves the putative active site more exposed, as might be expected for a protein fold designed to edit ssDNA or ssRNA (Fig. 3.3).

4.6.3 Comparison of the Common CDA Superfamily Core Reveals Broad Peripheral Diversification.

Members of the CDA superfamily share a core fold comprising a series of topologically conserved secondary structure elements that encompass the ZDD motif and enable deamination chemistry via Zn\(^{2+}\) coordination and cytidine binding. Diversifications to the core fold have imparted differences in substrate specificity and functional regulation among family members [49]. As such, members of the CDA superfamily target a remarkably diverse array of substrates including: RNA, DNA, deoxycytidylate, cytidine, cytosine, guanine and blasticidin S, as well as intermediates of riboflavin and purine biosynthesis. Most surprisingly, the adaptations that have enabled such functional diversity have evolved in a manner that does not alter the core CDA-like polypeptide fold. Consequently, the locations in primary sequence that demonstrate divergence are restricted to insertions or deletions on the periphery of the core fold, and are identifiable by structural comparisons among family members. For example, the gray elements shown in Figs. 4.6-4.8 demonstrate how non-core elements fold in a manner that flanks the conserved Zn\(^{2+}\) ion held in place by the ZDD motif. Differentiating between functional adaptations and the conserved core fold continues to be of special interest for researchers aiming to target regions that impart specialized biological activity. For example, CDA superfamily members have long been targets in the development of chemotherapeutic agents. However, the cross-reactivity and ensuing toxicity resulting from such therapies [60-63] underscores the importance of discerning one family member from the next at the molecular level to fine tune drug specificity. Greater discussion of the modes of substrate interaction will be presented in Section 4.8.
4.7 Modes of Oligomerization

4.7.1 Free Nucleotide Cytidine Deaminases (CDA): Strand $\beta_5$ Antiparallel to Strand $\beta_4$

A dominant theme for the diversification of the CDA superfamily is variation in the mode of subunit oligomerization. The structurally well-characterized fnCDAs, in which strand $\beta_5$ runs anti-parallel to strand $\beta_4$, exist as tetramers (Fig. 4.5), with the exception of the enzyme from *E. coli*. Although even the latter case, with its CD-PCD configuration (Fig. 4.4), generates a comparable pseudo-tetrameric assembly comprised of four CDA-like domains (Fig. 4.5, *third panel*). The subunits of the tetramer are related by 222, or D2 symmetry, as described in *Section 4.6.1*. The oligomerization interface of each tetramer is extensive, and in the case of the yeast enzyme Cdd1, is maintained by contributions from core helices 2, 3 & 4, a C-terminal 3$_{10}$ helix (5), loops 1, 3 & 4, and a few residues from strand $\beta_5$ (Fig. 4.5 and ref. [20]). The take home message is that the Cdd1 subunit interface is representative of enzymes of this class and that multiple polypeptide chains (3 for the strict tetramers) are essential to complete the formation of a single active site (i.e., *trans* complementation).

4.7.2 Cytosine Deaminase, Guanine Deaminase and TadA: Strand $\beta_5$ Parallel to $\beta_4$

Cytosine deaminase (CD), guanine deaminase (GD) and TadA exist as dimers, and share a common mode of oligomerization (Fig. 4.6) [64-66]. Subunits are related by a proper two-fold rotation axis in which each active site forms near the interface of the two subunits. The interface is maintained by contributions from the core helices 2, 3 and 4 with additional contributions from residues in loops 3 and 4. Again, the formation of a single active site appears dependent upon contributions from two polypeptide chains. Interestingly, a putative case of 3-D domain swapping was identified in the guanine deaminase structure, whereby
helix α4 and strand β5 are exchanged between the two subunits, thus increasing the buried surface area while maintaining all the elements of the core fold; for more information on 3-D domain swapping see refs. [67, 68]. In yeast cytosine deaminase the C-terminal helices fold over the active site and contact, among other things, a non-conserved helical element present in L4’ of the opposing subunit. This bulky configuration has been noted previously to restrict access of the substrate to the active site, possibly discriminating free cytosine from larger nucleosides [49]. In contrast, the C-terminal helix of TadA (a tRNA editing deaminase that converts adenosine to inosine) extends straight upward and has little opportunity to engage in oligomerization. The same would be true for guanine deaminase, but for the observed 3-D domain swapping effect (Fig. 4.6, two bottom-most panels).

The TadA structures are more similar to the GD enzyme as compared to the CD enzyme, as might be expected considering their preference for purines. TadA does not engage in domain swapping. However, two unique features of GD are present in TadA that are not present in CD. Helix α2 is elongated on the lower (C-terminal) end, such that when inspecting the dimer, this helix protrudes from the opposing subunit and is available to aid in substrate binding. The dimerization interface of both TadA and GD exhibit a more staggered packing of helices 2 and 3 as compared to CD, which further enhances the protrusion of helix α2.

4.7.3 Deoxcytidylate Deaminases (T4, N. e.) and APOBEC2: Strand β5 Parallel to β4

The deoxcytidylate deaminases (dCDs) exhibit a comparable helical oligomerization interface when compared to the CD and GD enzymes (Fig. 4.6 versus 4.7). A dimeric form of this enzyme derived from the nitrogen fixing bacteria N. europaea is diversified to include an additional N-terminal helix and β-strand (Fig. 4.7, upper panel). The extra β-strand interdigitates with the equivalent region from the opposing monomer and contributes to subunit dimerization. This structure was solved by the Midwest Structural Genomics
Consortium, and consequently the endogenous substrate for this protein has not been identified at present. As such, the classification of this enzyme is tenuous at present.

The T4 bacteriophage dCMP deaminase structure tells a unique story. The active form of the enzyme is a hexamer and therefore each subunit has two oligomeric interfaces [59]. This molecule is allosterically regulated by dCTP and dTTP in a region of the protein at the C-terminal (bottom) end of α-helix 2 [69] (Fig. 4.7, middle panel). A point mutation at this site, R115E, resulted in a dimeric enzyme (as characterized by gel filtration) that was largely devoid of allosteric regulation [70]. Nonetheless, this mutant form of the enzyme crystallized as a hexamer, perhaps as a result of ionic shielding due to the high salt conditions of crystallization [59]. In the crystal structure the R115E mutation localized to the helical oligomerization interface, thereby providing a rationale for its contribution to oligomerization. Notably, an R115Q mutation was dimeric in the absence of dCTP, but hexameric in the presence of dCTP, whereas the wild type enzyme is hexameric regardless of dCTP concentration [70]. Together, these findings implicate allosteric regulation through modulation of an oligomerization interface, which is a common mode of regulation [71, 72].

The elements maintaining this interface are extensive and familiar with respect to other CDA superfamily members, including core helices 2, 3 and 4, and loops 3 and 4. This information is particularly relevant in light of current hypotheses that suggest that both AID and APOBEC3 family members, which act on DNA, may be allosterically inhibited through RNA binding [73, 74]. We will see in Section 4.10 that RNA has a definite influence on the molecular mass of APOBEC3G.

Interestingly, the dimeric interface of T4 dCD is created through the antiparallel association of two β2 strands that form a continuous β-sheet surface (Fig. 4.7, middle panel, yellow inset). This subunit interaction is stabilized further through interactions between the N- and C-terminal helices, the former of which pack side by side lengthwise with additional contributions from helix α2 and loop L1 [59]. Although α-helical packing is the most common
mode of subunit oligomerization for members of the CDA superfamily, edge-to-edge β-sheet-mediated oligomerization is an emergent theme, as observed recently for APOBEC2.

APOBEC2 has the most extensive edge-to-edge β-sheet subunit interface of any structurally characterized deaminase (Fig. 4.7, lower panel). This interface is maintained through the canonical carbonyl and amide backbone interactions in a largely sequence-independent interaction, with additional contacts made by two residues at the bottom of helix 2 [52]. The N-terminal 40 residues of APOBEC2 were omitted from this construct, presumably to obtain crystals. Hence, the possibility exists that additional supportive contacts might originate from this unique N-terminal region and further stabilize the dimeric interface, akin to the interdigitating strands of N. europaea dCD or the extra 4 helix bundle of EcCDA, although there is no apparent sequence homology among the respective N-termini of these structures to support this assertion.

APOBEC2 includes all the canonical elements of the CDA superfamily topology, although these elements have been rearranged somewhat. Helix 1 is shortened and juts out from the β-sheet surface rather than laying across it. Thus, tertiary contacts from α-helix 1 are made primarily to the C-terminal features of APOBEC2, rather than to the β-sheet surface typically contacted by the α-helix 1 of other deaminases (Fig. 4.7, compare upper and lower panels). APOBEC2 encodes two C-terminal helices beyond strand β5; the first of these (α5) traverses the broad face of the β-sheet, akin to α-helix 1 of other deaminases [52]. A long intervening loop (L10) enables the second C-terminal helix (α6) to cap the side of the β-sheet, contacting β-strand 5 and α-helix 4. In this manner, the C-terminus is directed away from the active site, which is in contrast to the comparable, but upwardly directed helix observed in TadA [55] (Fig. 4.6, lower panel).

Like the T4 dCMP deaminase enzyme, APOBEC2 migrated as a dimer in gel filtration studies, but crystallized as a higher order tetramer [52]. The amount of buried surface area in the tetramer interface was 1745 Å², which is near the average value observed in a survey of biologically relevant protein-protein interactions [75]. The tetramer interface exhibits
contributions from loops L1 & L7, helix α4 and the last C-terminal helix, α6. The rearrangement of the C-terminal helices places the tetramer interface on the opposite end from the dimer interface, such that the tetramer observed in the crystal lattice is formed through a tail-to-tail interaction. (Fig. 4.7, lower panel – right). Gel filtration results not withstanding, if the tetrameric form is relevant, there is nothing to prevent APOBEC2 from further oligomerizing to form an extended polymer. In contrast, the helical interface of the T4 dCD enzyme is located on the back surface of the protein rather than its edge, thus resulting in formation of a toroidal hexamer (Fig. 4.7, middle panel). Further investigation of APOBEC2 will be necessary to establish how its oligomeric state confers biological function.

4.7.4 Multi-domain Enzymes RibG and ADAR2 of the CDA Superfamily: Strand β5 Parallel to β4

The CDA superfamily members RibG and ADAR2 demonstrate that the CDA-like fold can be fused to additional, discrete domains within a single polypeptide chain (Fig. 4.8) [51, 76, 77]. In RibG, the additional domain performs a discrete chemical reaction and is classified as a Rossmann fold, which is a classical nucleotide binding structure [78-80]. RibG is a bifunctional deaminase and reductase involved in riboflavin biosynthesis [51, 81]. Tetramer formation is mediated by one surface of each domain. The deaminase domain interface is mediated by an antiparallel side-by-side arrangement of the β2 strands (Fig. 4.8, upper panel inset) similar to that of T4 dCD and APOBEC2. In RibG, the β2 strands are twisted and not positioned to interact as a sheet, but engage in a few side-chain-mediated contacts further supported by contributions from α-helices 1 and 2 and Loops 1, 2 and 3.

ADAR proteins incorporate both a deaminase domain and one or more copies of a double stranded RNA Binding Domain (dsRBD). The latter domain(s) contributes to both RNA-independent dimer formation and substrate recognition [82-84]. Significantly, the crystallization construct of Bass and colleagues contained only the deaminase domain, which was demonstrated to be a moderately active monomer [54]. The monomeric state of this
construct has been disputed in gel filtration studies but validated by FRET analyses [84, 85]. Assuming the A-to-I editing activity of the monomeric form is genuine, this property represents a unique characteristic of this CDA family member. The active site of most other family members arises through *trans* complementation of subunits engaged in oligomerization.

ADAR2 is the most highly diversified fold yet observed for members within the CDA superfamily, and the self-contained fold supports the case for monomeric activity. The interfaces observed to engage in oligomerization in other family members are unavailable in the ADAR2 structure, such that if this construct is dimeric, the mode of oligomerization is likely to be novel among family members. Both ends of the central ADAR2 β-sheet exhibit auxiliary β-strands, beyond the canonical five strands of the CDA fold, that would interfere with a strand β2-mediated interface, such as that observed in APOBEC2. Additionally, the 'back side' of helices 2 and 3 are covered by a number of additional helices, thereby preventing formation of a dimer like that observed in TadA enzymes. These additional helices in ADAR2 coordinate an inositol hexakisphosphate (IP6) molecule that is essential for activity [54]. Significantly, ADAT1 proteins have been shown to exhibit a similar IP6 dependence, but ADAT2/3 proteins, the human homologs of the more primitive TadA proteins, do not [54]. Phylogenetic analyses further support this division of relatedness between tRNA editing enzymes (Fig. 4.3) [31].

### 4.8 Modes of Substrate Interaction

Although there is currently no structural information that describes how members of the immediate APOBEC family bind and deaminate ssDNA or RNA substrates, a comparison of members of the CDA superfamily reveals a central theme in substrate recognition. Namely, these deaminase enzymes share a common active site fold that originates from restraints imposed by the ZDD motif (Fig. 4.9). Like many enzymes that utilize β-scaffolds,
the substrate binds at the periphery of the sheet, most likely because this region is easiest to diversify through alteration of loops. Similar observations have been made for triose phosphate isomerase (TIM) α/β barrels, which typically bind substrate via loops connecting α-to-β elements at the carboxylic end of the barrel [86, 87]. Likewise, the ZDD motif places the essential catalytic Zn$^{2+}$ ion at the periphery of the beta sheet while using loops, particularly L3, to bind and orient the substrate for nucleophilic attack by activated water (Fig. 4.2).

Substrate specificity is achieved through variations on this theme, including alterations in the size of the binding pocket, the extent of solvent exposure, contributions from neighboring subunits, and supplemental contributions from additional elements within the structure, such as C-terminal helices; for a review see ref. [49]. APOBEC2 is a prime example of the variance that can be achieved with regard to solvent exposure (Fig. 4.7, lower panel). Here, the putative active site is strikingly open relative to all other family members, including those editing enzymes that also bind polynucleotide substrate such as ADARs and TADs [52]. The fnCDAs offer another 'variation on a theme'. As described in Section 4.7.1, all of the fnCDAs require multiple subunits to form their active sites, yet the size of the active site varies (Fig. 4.5). For example, the active site of blasticidin S deaminase is significantly larger than that of mouse cytidine deaminase (Figs. 4.9a versus b). Several factors, including a rearrangement of loops 1 and 3, contribute to the more open active site of the former enzyme. These observations demonstrate the remarkable variability of CDA superfamily members despite a conserved core fold and the presence of the ZDD motif.

### 4.8.1 Tetrameric fnCDAs Favor Flexible Flaps: RNA Editing and the Case of Cdd1 from Yeast

Difficulties in preparing crystal structures of bona fide APOBEC family members led investigators to seek three-dimensional structural information from more operable sources, such as Cdd1 from yeast. This fnCDA site-specifically edited reporter apolipoprotein B mRNA constructs during late log phase growth in yeast [45], albeit at levels considerably
lower than APOBEC1. Furthermore, both APOBEC1 and Cdd1 have been shown to target ssDNA in a bacterial mutator assay, although again, Cdd1 activity was inferior to that of APOBEC1 [47, 88]. These similarities in RNA and DNA editing function suggested that Cdd1 might provide structural insights into APOBEC1. In light of the recent APOBEC2 structure, this now seems less likely [20, 52]. Nonetheless, the Cdd1 structure suggests that the auxiliary nucleic acid editing activity of Cdd1 may be attributable to the accessible nature of fnCDA active site, as well as the consequence of protein overexpression.

In addition to Cdd1, several ZDD containing yeast proteins were examined for their ability to target the reporter apolipoprotein B substrate, but only Cdd1 exhibited significant activity [45]. Notably, Cdd1 was the only fnCDA family member tested; other family members included a deoxycytidylate deaminase, a cytosine deaminase, and a deaminase active in riboflavin biosynthesis. Compared to these other enzymes, the Cdd1 active site appears to be the most accessible in terms of the degree to which multiple structural elements from neighboring subunits fold over the active site (e.g., compare the Zn$^{2+}$ accessibility of Cdd1 in Fig. 4.5, versus cytosine deaminase in Fig. 4.6, the deoxycytidylate deaminases of Fig. 4.7, and RibG in Fig. 4.8). Upon analyzing the x-ray structure, Wedekind and colleagues tested the requirement for C-terminal flexibility by making a series of chimeric constructs predicted to possess flexible C-termini versus highly ordered termini, such as the E. coli CDA. Editing analysis revealed that constructs with flexible C-termini were capable of editing reporter apolipoprotein B RNA, whereas constructs with rigid C-termini or the intact E. coli CDA, whose active site is occluded by a rigid linker connecting the CD to the PCD, were incapable of C-to-U deaminase activity [20]. As such, the working hypothesis for the Cdd1 editing of RNA (or DNA) is that the C-terminus is flexible and can open to accommodate a polynucleotide substrate (Fig. 4.9b, flap harboring Phe137`).

Notably, Cdd1 required late log phase growth in the presence of galactose to elicit RNA editing [45, 89]. This observation suggests that the synthesis of a new protein may be required for Cdd1’s RNA editing activity, which would make such activity more of an
exception than a rule. Over-expression of APOBEC enzymes also resulted in general promiscuous editing activity on non-canonical substrates [6, 90-93]. Thus, both of these enzymes exhibit auxiliary activities in the appropriate context. It is not known if the explanation behind these activities is the same. Given the precedent of the APOBEC2 structure, it seems unlikely. Although Cdd1 editing of apoB mRNA may be the result of structural flexibility in the C-terminus, ‘promiscuous’ APOBEC1 activity may rely more upon cellular localization and the availability of cofactor necessary for specific substrate recognition.

4.8.2 A Topological Transformation Obstructs Active Site Accessibility in Dimeric Deaminases that Bind Bases

A parallel orientation of strands $\beta 4$ and $\beta 5$ facilitates the engagement of C-terminal helices in substrate binding. This is accomplished to an extreme extent in yeast CD, where two well developed C-terminal helices fold over the active site and sequester substrate to such an extent that only the monomer contacts cytosine (Figs. 4.6 & 4.9d) [64]. Oligomerization remains important for sequestering substrate from solvent and aids in defining the binding pocket in each of the examples presented here. In guanine deaminase, the dimer contributes to pocket formation and substrate binding to a larger extent than yeast CD (Fig. 4.6). Substrate modeling suggested that two aromatic residues from the opposing subunits are likely to interact directly with substrate, one from the opposing helix $\alpha 3'$, and another from the domain swapped C-terminal helix [65]. The T4 dCD enzyme has a relatively under-developed C-terminal helix, and contacts substrate directly with only one subunit (Fig. 4.7, middle panel). Nucleotide recognition is dominated by a sixty residue insertion in Loop 3 that contains two helices, two small $\beta$-strands and twenty residues of unknown structure that were not visible in the crystal structure [59]. Although absent in the human ortholog, this region coordinates a second zinc ion that acts to stabilize the binding pocket and ameliorate the localized negative charge of the substrate’s phosphate group (Fig. 4.9c).
The take home message is that nature evolves enzymes to accommodate a specific substrate, but there appear to be a finite number of approaches employed to achieve this goal to preserve the fundamental CDA architecture. Knowledge of the common variations is significant as a means to direct experiments on structurally uncharacterized enzymes. As such, the suggested *modus operandi* is to align sequences of interest to known structures to identify segments that have been evolutionarily embellished (i.e., C-terminal helices or L3 expansions). These areas can then be targeted experimentally to define functional relevance. We propose that it may be necessary to compare the sequence in question to a representative from each of the major CDA classes presented (Figs. 4.5-8).

### 4.8.3 Substrate Selection by Polynucleotide Editing Enzymes Remains Elusive

Family members that edit polynucleotide substrates must accommodate relatively large dsRNA, ssRNA or ssDNA substrates. Structures representative of each editing family (TADs, ADARs, and APOBECs) suggest that they achieve such targeting uniquely, although only the TadA enzyme from bacteria has been solved in the presence of a substrate analog (Fig 4.6 & 4.9e) [94]. The TadA enzyme looks strikingly similar to GD, but binds a much larger substrate (Fig. 4.6, *middle* versus *lower panels*). The absence of 3-D domain swapping in TadA allows for a more open active site. In particular, a single, long C-terminal helix extends upward from strand $\beta_5$ and engages in several contacts to the tRNA substrate. Contacts on the opposite side of the active site are provided by the protrusion of the opposing subunit’s long helix $\alpha_{2'}$. Together, these elements extend the binding pocket up from the active site, such that no fewer than five nucleotides are accommodated (Fig 4.9e) [94]. This arrangement is in stark contrast to yeast CD, which cannot accommodate a single nucleoside (Fig. 4.9b).

The active site of ADAR2 represents the opposite scenario in terms of accessibility. Each subunit harboring the fundamental CDA domain appears to fold autonomously as a monomer, with helices $\alpha_2$ and $\alpha_3$ of the ZDD poised such that the active site zinc resides slightly above the plane of the $\beta$-sheet and is thus far more solvent exposed by comparison to
the TadA enzymes (Fig. 4.6, lower panel versus Fig. 4.8, lower panel). This shallowness seems sensible in light of the requirement for ADAR2 to bind an adenosine in the context of dsRNA. To gain access to the edited position, the base is hypothesized to flip out of the duplex into the superficial opening revealed in ADAR2. This flipping mechanism is supported by fluorescence studies from Beal’s lab [95]. Consequently, the substrate would not be able to extend as far into the core of ADAR2 as the anti-codon loop of tRNA does with TadA without significant melting of the duplex structure.

The APOBEC family recognizes single-stranded nucleic acid substrates, and as with TadA and ADAR2, the crystal structure is reflective of the functional data. The putative active site of the APOBEC2 structure is strikingly open [52] (Fig. 4.7, bottom panel). This effect is achieved by i) the placement of the C-terminal helices on the opposite face of the β-sheet rather than folded over the active site, and ii) the unique head-to-head and tail-to-tail modes of oligomerization that avoid the more common interaction between helices α2, α3 and α4 (Fig. 4.7, bottom panel). Notably, these helices would be exposed in ADAR2 as well, if not for the considerable adaptations necessary to bind the IP6 molecule. The β-strands in APOBEC2 are considerably elongated relative to other CDA domains, and Loop 1 is also expanded. Collectively, these elements limit substrate entry from the ‘top’ as compared to ADAR2 and TadA.

4.9 The APOBEC Family: Insights into a Structurally Under-represented Family

Phylogenetic analysis places a protein in the context of evolution, between its ancestors and descendents. The preceding sections used structure to provide a phylogenetic perspective of the APOBEC family, thereby placing this family in the context of the CDA superfamily. This served to highlight common themes and unique features for the investigator interested in pinpointing regions of functional significance. However, the
absence of structural information for the bulk of APOBEC enzymes prevents such lucid comparisons between APOBEC family members. Consequently, alternative methods must be considered in the analysis of these enzymes. Comparative sequence analysis, the tool most commonly used to predict evolutionary relationships, can be combined with information about gene structure to glean insight into how other APOBEC enzymes might compare to the structure of APOBEC2. For example, sequence comparisons of APOBEC family members demonstrate a number of conserved regions (Fig 4.10). Based on the APOBEC2 structure, some of these sequences appear to stabilize the core CDA fold ($^{70}$LCY$^{72}$, $^{122}$YVSS$^{125}$), or C-terminal helical structures ($^{187}$VWxxFV$^{192}$, $^{219}$LxxIL$^{223}$), whereas others are positioned to contribute to either the tetramer interface or substrate recognition ($^{153}$RLF$^{155}$). Clearly, there is an obvious benefit in distinguishing between such regions when designing experiments.

Sequence based analyses can also suggest evolutionary timelines including the order and pace with which family members evolved. The pace at which a protein evolves is representative of the selective pressure it has experienced. It has been asserted by Gillespie and others that the evolution of proteins is consistent with the principles of natural selection to some extent [96, 97]. As such, external influences can cause a protein to remain relatively the same over time (purifying selection), or select for a new functionality (positive selection). The hallmark of positive selection is an excess of non-synonymous substitutions (i.e., those that alter the amino acid being encoded) relative to synonymous substitutions (that retain the encoded amino acid). These changes can be benchmarked by pairwise sequence comparisons of orthologous proteins from different species. The ratio between these two values (non-synonymous versus synonymous) is referred to as the $d_N/d_S$ ratio. Values greater than one are suggestive of positive selection [98]. For example, proteins involved in viral defense commonly exhibit rapid rates of evolution, as indicated by high $d_N/d_S$ ratios, as a means to counteract rapidly evolving viruses [99].

After considerable discussion of the similarities of the APOBEC family members to other deaminases in which strands $\beta 4$ and $\beta 5$ are parallel (Sections 4.7.2 to 4.7.4), it is
interesting to observe that the APOBEC proteins form a distinct phylogenetic cluster (Fig. 4.3). This branching is indicative of the age of the APOBEC family and supports the rationale for initial modeling efforts derived from the CDA enzyme structures rather than the CD enzymes; i.e., APOBEC proteins appeared equally related to either class, but the CDA structures were elucidated first [20, 35]. This branching also supports the prediction that the structures of other APOBEC family members will more closely resemble the APOBEC2 structure rather than other members of the CDA superfamily.

Estimation of the evolutionary relationship of APOBEC family members requires identification of a progenitor member. The APOBEC proteins are divided into five groups according to function and order of identification. These groups include AID, APOBEC1, APOBEC2, APOBEC3 and APOBEC4. Each class is represented by a single member with the exception of the APOBEC3 clade, which has undergone rapid expansion in primates [31]. Members of the AID, APOBEC1, and APOBEC3 clades exhibit fairly non-specific deamination activity on ssDNA without any cofactor or ATP requirement [100-102]. No editing activity has yet been demonstrated for the APOBEC2 or APOBEC4 proteins [28]. APOBEC1 is unique among family members in that its endogenous substrate is RNA rather than DNA, albeit an RNA binding cofactor is required to achieve specificity. This precedent suggests that other family members may have the capacity to target specific RNA, although the substrate sequences and co-factors have not been forthcoming. In the next sections, we review available phylogenetic and structural data for the APOBEC family to highlight the relationships between individual enzymes and thereby guide future research efforts.

4.9.1 Activation Induced Deaminase (AID): An Ancient Enzyme with Essential Roles in Adaptive Immunity

AID was identified in 1999 independent of its relationship to other APOBEC family members, thus giving rise to its unique nomenclature [103]. The gene encoding AID is located on human chromosome 12 and contains 5 exons [104]. The protein is expressed in
germinal center B cells, and is essential for class switch recombination and somatic hypermutation during antibody gene diversification, which is found in all extant jawed vertebrates [105] [106-108]. As expected for a protein performing a highly-conserved function, the $d_{ni}/d_s$ ratio generated by comparing human and chimp orthologs indicates strong purifying selection [109]. Two variants of AID were identified in lamprey (P. marinus), a jawless vertebrate that emerged ~480 million years ago [110, 111]. One variant, PmCDA1, represents a 208 residue protein from a single exon, whereas PmCDA2 is a 331 residue protein from four exons. Each of these genes exhibits a single characteristic ZDD signature motif. The longer PmCDA2 also contains an AT-hook motif in the C-terminus, which is a small DNA-binding motif comprised of Gly, Arg and Pro residues that recognizes the minor groove of AT-rich DNA sequences [112, 113]. Expression of these ancient proteins was detectable in lymphocytes from blood and hematopoietic tissues [110]. While AID has now been identified in early vertebrates, it is absent in the protochordate, Ciona intestinalis [114]. Thus, AID is currently believed to be the oldest APOBEC family member and emerged in early vertebrate evolution. As such, it is the most likely progenitor for the more recently evolved APOBEC1 and APOBEC3 families, as evidenced by the conservation of multiple exon boundaries within each of these genes (Fig. 4.10) [31, 37].

The APOBEC2 structure represents the strongest model available to provide insight into AID. Prochnow, C. et al. functionally examined a number of AID mutations using this assertion [52]. AID is hypothesized to dimerize utilizing the essential region $G_{47}$ to $G_{54}$ [115]. This corresponds with the β-sheet interface of APOBEC2 (Fig. 10). Interestingly, mutants R112C and Y114A/F115A (analogous to the conserved $^{153}$RLF$^{155}$ patch described above) exhibited no deaminase activity [52]. This patch has been hypothesized as important for tetramer formation in APOBEC2, but may also play a role in substrate recognition or regulation. Both AID and APOBEC3G have exhibited unique properties in the presence of RNA. AID has been shown to require RNAse treatment before deamination can occur on DNA [73]. Based on the precedent of allosteric regulation in T4 dCD via nucleotide binding
within the hexameric interface (Section 4.7.3), and given that both APOBEC2 and T4 dCD enzymes were characterized as dimers in gel filtration but crystallized as higher order oligomers [52, 70], investigation of this interface as a site of regulation is warranted. If the tetramer interface of AID were a site of allosteric regulation, analogous to that of T4 dCD, the regulatory molecule would be in a prime location to influence active site architecture and substrate recognition.

R24E is another inactive AID mutant formulated from the APOBEC2 model [52]. In APOBEC2, the equivalent residue (R65) forms a cation-π interaction with Y61 to stabilize the β-hairpin conformation observed in Loop1 of the outer monomers (described in Section 4.9.2). It is hypothesized that the R24E mutation inactivates the AID enzyme by collapse of Loop1 and subsequent inactivation of the active site, although further investigation will be required to definitively prove this assertion.

### 4.9.2 APOBEC2: A Divergent Ancestral Protein of Unknown Function

If AID is the oldest family member, then APOBEC2 is the penultimate member [116, 117]. Discovered in 1999 during a search for APOBEC1 homologs, this protein has been identified in the teleostei, including pufferfish and zebrafish, but not in the chondrichthyes (cartilaginous fish) [31]. Interestingly, the bony fish contain two paralogous copies of the APOBEC2 gene, presumably a product of the genome duplication that took place in the ray-finned fish [118, 119]. It is not understood why there is only one copy of AID in these animals, but the AID and APOBEC2 genes are located on different chromosomes and gene deletion events are common in evolution. The APOBEC2 gene is located on human chromosome 6 and contains three exons [116]. The core APOBEC fold, including all C-terminal helices, is encoded within the second exon; the first exon contributes the N-terminal 40 amino acid residues that are absent from all other APOBEC family members (and were omitted from the crystallization construct). Together, the unique exon boundaries and sequence-driven phylogenetic analyses suggest that APOBEC2 and AID diverged prior to the evolution of APOBEC1 and APOBEC3 family members (Figs. 4.3 & 4.10) [31].
Akin to AID, the \( d_{s} / d_{s} \) ratio calculated between human and orangutan orthologs of APOBEC2 indicates strong purifying selection [109]. Robust protein expression was observed in cardiac and skeletal muscle tissues, and to a lesser extent in a number of other tissues including kidney, liver and small intestine [116, 117, 120]. Expression in human liver cells is regulated by pro-inflammatory cytokines via NFκB activation, and two NFκB Regulatory Elements (NREs) were identified in the 5’ flanking region of the APOBEC2 gene [120]. These results suggest APOBEC2 has a function, although it has not been identified.

Davidson and colleagues demonstrated that APOBEC2 could inhibit the RNA editing activity of APOBEC1 [117]. However, deletion of the mouse ortholog exhibited no identifiable phenotype [121]. Lastly, although multiple labs have reported low levels of deamination activity on free nucleotides in vitro, it has been suggested that this activity may be due to assay contamination. [116, 117, 121].

Importantly, analysis of the APOBEC2 structure provides some insight into the apparent absence of detectable activity. Within the tetrameric form of APOBEC2, two distinct loop conformations were observed at the active site [52]. The internal subunits at the ends of the tetramer are characterized by a β-hairpin turn between helix α1 and strand β1 that allows for a more open active site. The Zn\(^{2+}\) ion of each outer subunit is coordinated to the His, Cys and Cys residues of the ZDD motif, although the distance between Zn\(^{2+}\) and the second Cys residue is slightly longer than expected for an inner sphere coordinated Zn\(^{2+}\). Additionally, the fourth ligand for an active deaminase is water, which is present in only one of the two outer monomers and is also characterized by an unusually long Zn\(^{2+}\) coordination distance of 3.5 Å. These discrepancies are suggestive of local disorder. The arrangement of these outer subunits contrasts with the two internal subunits, in which Glu60 serves as the fourth Zn\(^{2+}\) ligand, thereby displacing the nucleophilic water and ‘inactivating’ this subunit. Glu60 localizes to L1, which is more curved in the two internal subunits as a result of the encroaching tetramer interface [52]. To clarify, it is Glu100, not Glu60, that is believed to be the conserved proton shuttle depicted in Fig. 4.1. Interestingly, these inactivated subunits
exhibit the expected Zn$^{2+}$ coordination distances, suggesting greater definition of the atoms within the experimental electron density maps.

One possible reason for the suboptimal conformation of the outer, “active” subunits of APOBEC2 involves crystal packing contacts. The formation of the APOBEC2 crystal lattice involves a series of four hydrogen bonds to residues in the L1 regions of the outer subunits. The possibility that these interactions are the driving force behind the more structured versions of L1 cannot be dismissed, and warrants further investigation using experimental approaches. Hence, it is plausible that there are only two active sites in the APOBEC2 tetramer, but it is also conceivable that neither is active in the solution state, which would account for the preponderance of biochemical evidence to date.

Predictions of the mode of substrate binding for APOBEC2 are purely speculative at present since the protein has no known substrate or biological editing activity [116, 117, 121]. However, comparative structural analysis supports the hypothesis that C-terminal helices will contribute to substrate binding (Section 3.8.2). Additionally, the presence of a phosphate backbone in the substrate suggests that whatever the trajectory of the RNA, it is likely to be complemented by a trail of positively charged residues on APOBEC2. Indeed, such a trail is observed in maps of electrostatic surface potentials (Fig 4.7, bottom panel) prompting the hypothesis that an RNA strand might wrap around the outermost end of APOBEC2, nestled between helices α1 and α6 to contact basic residues in L1 and L7, then extending down the backside of helices α2 and α3 toward a patch of basic residues contributed by loops 2, 4, and 6.

Basic residues will occur on the surface of a protein for a number of reasons, not the least of which is to create the hydrophilic exterior necessary to maintain solubility. Likewise, aromatic residues readily engage in base-stacking with single stranded nucleic acid (Fig. 4.9e) and can be indicators of substrate binding when located on a protein surface; however, such residues may also contribute to protein-protein interactions. Such disparate possibilities demand a synergistic approach that implements structural and functional methods.
Moreover, without a known function or assayable substrate, there is no way to directly test the molecular details of substrate recognition or to define them crystallographically. However, all is not lost. The shrewd investigator recognizes that amino acid sequence similarity equals structural similarity. As such, the next rational step is to extend knowledge of the APOBEC2 coordinates onto APOBEC family members better suited for experimental validation.

4.9.3 APOBEC1: The Historical Archetype of C-to-U Editing Enzymes

APOBEC1 was the first C-to-U editing enzyme to be characterized and therefore is the namesake of the family, i.e., the APOlipoprotein B Editing Catalytic subunit-1 [122-125]. In isolation, this enzyme exhibits strong mutator activity on ssDNA in a bacterial assay, as observed for several other family members [100]. However, the endogenous substrate of APOBEC1 is a single cytidine within the ~14,000 nucleotide mRNA transcript that encodes the lipid transporter, apolipoprotein B. Deamination activity influences lipid transport, and is achieved in the context of a 27S editosome ribonucleoprotein complex comprised minimally of APOBEC1, an obligate RNA binding auxiliary factor named APOBEC Complementation Factor (ACF), and the apolipoprotein B mRNA (see Appendix for further details on ACF) [126]. The specificity of APOBEC1 for substrate is accomplished through ACF, which contains 3 RRMs (RNA Recognition Motifs) [127]. Several groups have suggested that the apolipoprotein B mRNA substrate exhibits secondary structure that enhances recognition [128, 129]; however, such structures appear to be melted away by the ACF cofactor such that APOBEC1 acts on a ssRNA substrate [129]. Thus far, APOBEC1 is the only family member with an identified essential cofactor.

APOBEC1 editing is limited to the small intestine in humans [130]. In mice and rats (and to a lesser extent, dogs and horses) editing also occurs in the liver due to the presence of three additional 5'-exon regions that encode a hepatic-specific promoter [131]. Notably, the APOBEC1 variant present in these species is 7 residues shorter on the C-terminus and the mRNA is more widely expressed, having been identified in spleen, kidney, lung, muscle and
heart. The APOBEC1 gene is located ~1Mb from AID on chromosome 12 in humans [132, 133]. The location of the rodent versions are syntenic with human chromosome 12, but are located only 30 Kb apart and the transcriptional orientation of APOBEC1 is reversed. Genome comparisons suggest that a 1 Mb inversion containing the APOBEC1 locus occurred after the rodent/primate divergence [31], although it has not been indicated whether this inversion could be the underlying cause for loss of APOBEC1 expression in the human liver.

Current estimates of the age of APOBEC1 hinge on the observation of apoB mRNA editing in the American opossum [134], but not in amphibians or birds [130]. This indicates that APOBEC1 is younger than either AID or APOBEC2, since marsupials diverged ~170 million years ago [135]. Together, the conservation of sequence, exon junction boundaries, and chromosome assignment support the hypothesis that APOBEC1 evolved from AID (Fig. 4.10). The more recent evolution of the APOBEC1 protein presents the possibility that its activity on RNA may be a uniquely evolved adaptation. In contrast, if the ancient AID were found to deaminate RNA it would be a strong argument that RNA deamination is a conserved trait of the APOBEC family, which appears to be the case for ADARs. The \( d_n / d_s \) ratios for APOBEC1 vary considerably according to species; however, it is generally agreed that this protein has experienced a moderate level of positive selection, similar to that observed for the apoB protein itself [29, 131], and greater than that of either AID or APOBEC2 [109]. Moreover, a comparison of human and orangutan orthologs suggests that current pressures are focused on the N-terminal 100 residues [109].

Investigations into the cellular trafficking patterns demonstrated that nuclear localization requires residues within the segments 1-56 and 97-172 and may be ACF dependent, whereas a C-terminal leucine rich region (173L-187L) is sufficient for cytoplasmic retention [136]. This enzyme is believed to function as a dimer and truncation studies suggest that the C-terminus is a key component of this interaction [137], although the N-terminus has also been implicated [138]. If the C-terminal data are accurate, the tetrameric,
helical interface of the APOBEC2 structure may be more relevant to APOBEC1 than the \( \beta \)-sheet interface. This highlights how variations on the themes of oligomerization and substrate recognition may contribute to the diverse activities of APOBEC enzymes.

A number of APOBEC1 truncation and point mutations generated by Teng and Navaratnam can now be mapped onto the APOBEC2 model for further consideration [137, 138]. R16A and R17A mutants retain dimerization capability but are unable to crosslink or edit RNA [138]. In APOBEC2, the equivalent residues localize before the start of helix 1 on the front face of the \( \beta \)-sheet. This patch is not likely to contribute to either oligomerization or catalysis, but may be involved in substrate recognition or stabilization of the long flexible Loop 10. The single point mutations R33A and K34A do not significantly reduce editing, however the double mutant edits RNA at levels only 3% of wild-type [137]. Interestingly, R33 in APOBEC1 is analogous to R24 in AID (Section 4.9.1). It is not understood why R24E yields an inactive AID enzyme, whereas R33A APOBEC1 is fully functional. Regardless, this mutation is likely to contribute to substrate recognition or active site stabilization.

Other residues implicated in substrate recognition include F66 and F87, which abolish editing when mutated to Leu [139]. Both of these residues are conserved across the APOBEC family; position 66 is always Phe, whereas position 87 is observed to be Phe or Tyr. In APOBEC2, the residue equivalent to F66 (F103) extends from the bottom of helix \( \alpha_2 \) into the cleft between helix 2 and helix 3. In this manner it could contribute to either substrate recognition or the stability of the core fold, although the inability of F66L APOBEC1 to edit free nucleotide cytidine supports the latter of these two possibilities [139]. The residue equivalent with F87 (Y122) extends from the front face of the \( \beta \)-sheet just below helix 5 in APOBEC2, where it most likely stabilizes the protein fold. I185 is another highly conserved residue that yields an inactive enzyme upon mutation to Ala in APOBEC1 [137]. In APOBEC2, this residue appears to stabilize the orientation of C-terminal helix \( \alpha_6 \) with respect to helix \( \alpha_4 \). A number of additional mutations have been made to the C-terminus of APOBEC1; however, poor sequence conservation between family members in this region
precludes rational prediction. The observation of numerous Pro and Leu residues in the C-terminus of APOBEC1 has generated hypotheses about protein-protein interactions, but it remains to be seen if these residues are important for recognition of ACF. As stated for APOBEC2, the identification of hydrophobic residues on a protein surface may be suggestive of protein-protein interactions. As such, the APOBEC2 coordinates provide a tenable starting point to initiate such an analysis.

4.9.4 APOBEC4: Pushing the Envelope of APOBEC Boundaries

APOBEC4 is the least well studied member of the APOBEC family and was identified through an iterative database search [28]. The protein has no known function, although EST analysis indicated expression in testes. As reported in the Ensembl database, this protein is found on chromosome 1 and contains two exons. The ORF is confined to exon 2, and is predicted to encode a protein of 367 amino acids. The length of this sequence is suggestive of a double-deaminase, but there is a single copy of the ZDD consensus motif; interestingly, the HxE sequence exists as HPE in both APOBEC4 and the non-catalytic N-terminal deaminase domain of APOBEC3G. Additionally, the PCxxC motif involved in Zn\(^{2+}\) coordination has been altered to PCx\(_6\)C in APOBEC4, which calls into question the assignment of this protein as a ‘canonical’ APOBEC family member. Sequence motifs conserved within the APOBEC family are also present in APOBEC4, although most are distinguished by conservative mutations. For example, the \(^{153}\text{RLF}^{155}\) motif found in APOBEC2 is present in human APOBEC4 as \(^{156}\text{QLY}^{158}\). Analysis of the \(d_i/d_s\) ratio between mouse and human orthologs indicate modest selective pressure, similar to that of APOBEC1 [28]. The gene for APOBEC4 has been identified in mammals, frogs and chickens, but not in fish, suggesting that it emerged after AID and APOBEC2, but prior to APOBEC1 and APOBEC3. Further research will be necessary to clarify the role of this protein and its relationship to other family members.
4.10 APOBEC3: Radiative Expansion of Proteins Involved in Viral Defense

The APOBEC3 genes have received extensive investigative interest due to the exciting discovery that many proteins function as defenders against retroviruses and retrotransposons. In light of this significance, and the multiple members that comprise the APOBEC3 family, the remainder of this chapter will focus on these proteins. The emphasis will be upon understanding the origin and sequence relationships between family members, which is paramount to undertaking a comparative modeling analysis. Although their exact mechanisms of action are unknown, it appears that the efficacy of some APOBEC3 family members in restricting viral infectivity is due to C-to-U deaminase activity that targets viral genome intermediates resulting from reverse transcription [140-142]. A deaminase independent activity has been invoked as well [143-145]. The expression profile further supports a protective role. APOBEC3 proteins are most commonly and abundantly expressed in lymphoid tissues, which are relevant to retroviral replication, and in the germline where genome restricted mobile genetic elements need to transpose to ensure generational survival [37, 102, 146-150].

Among the APOBEC family members, the APOBEC3 genes are the most recently evolved and appear to be confined to mammals. To date, a single APOBEC3 gene has been identified in rodents, cats, pigs and sheep, two in cows, three in dogs and horses, and eight in humans [149]. There is, as yet, no standard nomenclature for the APOBEC3 genes identified in lower mammals. However, in humans the genes are labeled alphabetically in order of occurrence. A-G (including D/E, which are now considered one gene) are arranged tandemly on chromosome 22, followed 14 kb downstream by APOBEC3H [17, 31, 37]. Lastly, a pseudo-gene devoid of introns (the likely product of a retrotransposition event [151]) is hypothesized to be the most recent duplication of APOBEC3G, and was identified on
chromosome 12 [31]. Given current estimates of mammalian evolution [152] and the absence of APOBEC3 in avian lineages [153], the window of emergence for the APOBEC3 gene is between 100 - 300 million years ago. Based on sequence homology, exon boundaries and chromosomal location, AID is the suggested parent of this family [31]. Further, the respective sites on human chromosomes 12 and 22 where the AID and APOBEC3 genes reside have been hypothesized to originate from the same ancestral chromosome [154].

Sequence-based phylogenetic trees reveal three distinct groupings of the APOBEC3 CDA domain, referred to as Z1a, Z1b and Z2. The amino acid sequence characteristics driving this classification have not been systematically documented, although one example is the four residue insertion exclusive to Loop1 of Z2 domains (Fig. 4.10, C-term of mouse A3 and macaque A3H located between α1 and β1). Notably, all three APOBEC3 domain forms are present in dogs, horses, cows and humans [149] indicating that three forms of APOBEC3 emerged early on in the evolution of this gene. Further, this division would seem to support unique functionalities for each sub-type, such as modes of oligomerization or substrate recognition, based on the general principles observed for the CDA superfamily. Of unknown significance is the observation that in all species to date, the Z2 domain is present in a single copy, either as the C-terminal portion of a fusion gene (rodents, cows and pigs), or as a single domain (dogs, horses and humans) (Fig. 4.11) [149]. In hominids, including the gorilla, chimpanzee and human, the sole Z2 gene representative (i.e., APOBEC3H) exhibits poor ssDNA deaminase activity, resulting from a truncation mutation that occurred sometime after the separation of gibbons from other hominids ~18 million years ago [149, 155, 156]. Notably, a fully active -3H that efficiently hypermutates retroviral genomes is present in Old World monkeys including the macaque and sooty mangabey [149].

The APOBEC3 proteins are the only members of the APOBEC family that are found in both single CDA and double-CDA forms. Akin to the *E. coli* CDA enzymes, a gene duplication event within the APOBEC3 cluster is hypothesized to have produced a single
polypeptide chain that encodes the equivalent of an AID dimer within a single subunit (Fig. 4.11). Such double-CDA domain APOBEC3 variants are present in rodents, sheep, cows and pigs, as well as primate -B, -D/E, -F, and -G forms [37, 147, 157, 158]. Single domain forms of APOBEC3 genes are also present, as in the examples of cat, dog and horse, as well as the primate forms -A, -C and -H [37, 149, 159]. There is no correlation between the type (Z1a, Z1b, Z2) of APOBEC3 CDA domain and its occurrence as either a single domain CDA protein, or as the N-terminal or C-terminal portion of a double-CDA protein. Several of the double-CDA forms of APOBEC3 exhibit deaminase activity in only one CDA domain. In mice (Z1a, Z2), the N-terminal domain is catalytically active [160], whereas in human APOBEC3G (Z1a, Z1b) the C-terminal domain is catalytically active [38-40]. Human APOBEC3B (Z1a, Z1b) is the only double-CDA domain protein that traffics to the nucleus, and both the Z1a and Z1b domains exhibit deaminase activity [145, 158, 161, 162]. Thus, there is no apparent correlation between domain type and catalytic activity. Oligomerization may contribute to the differential activity of these domains, given the apparent inactivation of the APOBEC2 active site at the tetramer interface [52]. Notably, lack of deaminase activity is not necessarily an indicator of functional relevance, as deaminase independent viral inhibition has been reported [143, 163].

4.10.1 Mechanisms of Primate-Specific Expansion of the APOBEC3 Proteins

The CDA structural domains of the tandemly arranged APOBEC3A-G are either Z1a or Z1b [31, 149] and occur in the same transcriptional orientation. Based on conservation of both non-coding nucleotide sequence as well as protein sequence, these genes are paralogs, i.e., the result of local duplication events [31, 37]. For example, APOBEC3C is highly related to the C-terminal domain of APOBEC3F. Similarly, the single CDA domain APOBEC3A is the result of a partial gene duplication of the double CDA protein APOBEC3B. APOBEC3A has a unique exon 1 derived from the adjacent chromosome 22 sequence, but the next four exons are derived from the last three exons of APOBEC3B [37, 164]. In fact, a 29.5 kB gene deletion polymorphism described by Eichler and colleagues [165] eliminates the C-terminal
portion of APOBEC3A and most of APOBEC3B. However, due to the homologous nature of these regions, the resultant mRNA encodes an intact APOBEC3A followed by the 3’-UTR of APOBEC3B. This deletion is rare in Africans and Europeans, more common in Asians, and nearly fixed in the Oceanic populations of the Far East. The implications of this polymorphism with respect to human disease are unknown at present.

Based on the observation of unique Z-type architecture, it seems likely that the initial duplication event that gave rise to the double-domain proteins in primates is independent of that which occurred in lower mammals. All double-CDA APOBEC3 proteins in lower mammals exhibit a Z1a, Z2 architecture; the only Z2 present in primates is APOBEC3H. In contrast, it has been hypothesized that the first double-CDA APOBEC3 protein generated in the primate locus was duplicated in its entirety to generate additional double-CDA APOBEC3 genes [37]. This is supported by the conservation of several repeated sequences present in the 5’ region as well as within intron 4 (i.e., the intron that separates the two CDA domains) [37]. Repeats and duplications of any kind, once established, promote further rearrangement through their own misalignment and subsequent recombination [166-168]. These misalignments can result in further duplications or deletions. Thus, a chromosome with two or three duplicated APOBEC3 genes is more likely to experience further remodeling events.

The extensive expansion of APOBEC3 genes is believed to be exclusive to primates, but is not necessarily identical in all primates [149]. One indicator of how this process may have conferred a selective advantage is the dramatic decline in retrotransposon activity observed in primates over the past 35-50 million years [169]. This trend is not seen in the mouse genome, presently believed to encode only one APOBEC3 enzyme, where a number of transposable elements are still active. In fact, it is estimated that about 1 in 600 mutations in the human genome are caused by retrotransposons, whereas in mice the odds are 1 in 10 [170]. The discussion of retrotransposons brings up a rather ironic scenario. APOBEC3 proteins may have evolved to curb the very mechanism that facilitated their initial expansion. Primate genomes, as a whole, experience a greater incidence of segmental duplications
compared to other mammals [171], a finding which has been partially attributed to Alu repeat elements [166].

Significantly, both Alu and LINE repeat elements were identified in the BAC from which APOBEC3 genes were originally identified [37]. In fact, chromosome 22 exhibits the highest incidence of segmental duplications among the autosomal chromosomes [171], as well as the third highest percentage of sequence encoding Alu elements (18%) [172]. Not surprisingly, this chromosome contains a number of gene families, in addition to APOBEC3, that have diversified via duplication events [164, 173]. Lastly, the current address of the APOBEC3 locus at 22q13.1 is believed to have resulted from a double translocation event that occurred between human chromosomes 12 and 22 subsequent to the divergence of anthropoids from prosimians (~50 million years ago) [174]. Consideration of such relationships is common among structural biologists since they are useful in the generation of homology models. As such, it is apparent that the duplicated nature of APOBEC3 family members precludes a facile description of their three-dimensional structures without a more sophisticated experimentally defined starting model than APOBEC2. This observation is especially true for a tandem CDA protein such as APOBEC3G.

4.10.2 Alternative Methods to Obtain Structure: The Molecular Envelope of APOBEC3G by Small Angle X-ray Scattering

Although the use of x-ray crystallography to define three-dimensional structures of APOBEC3 family members represents the 'gold standard' in the field, other biophysical approaches can provide low resolution structural information that can contribute to understanding the shape and oligomeric state of a protein in solution. As such, Wedekind & Smith and colleagues restored the molecular envelope of human APOBEC3G by use of small angle x-ray scattering (SAXS) methods [175]. This technique provided the global shape of APOBEC3G in solution, at ~3 nanometer resolution, thus making it possible to
detect where specific CDA structural domains are located based on the volumetric
distribution of the particle relative to known structures (e.g., Figs. 4.5-8).

Because RNA binding is an inherent property of APOBEC3G, it was necessary to
digest samples with ribonuclease A during the purification process (similar to findings with
AID). The SAXS results from RNA depleted samples revealed that the molecule adopts a
140 Å-long extended shape in solution [175] (Fig. 4.12a). The SAXS data, as well as gel
filtration and dynamic light scattering suggested that APOBEC3G forms a homodimer in
solution of ~95 kDa [175]. These observations implied that previously determined fnCDA
structures (Fig. 3.5) were unsuitable as templates for APOBEC3G comparative modeling
since their oligomeric subunit arrangements are distinctly ‘square’ in shape (Fig. 4.12a
versus b) with dimensions of 55 Å by 55 Å by 40 Å. By comparison, the elongated large-
small-large-small volume distribution of APOBEC3G is evident based upon visual
inspection of the SAXS envelope (Fig. 4.12a). Therefore, although a fnCDA tetramer did
not fit into the electron density for the SAXS dimer, simple volumetric arguments suggested
an isolated CDA domain (e.g., Fig. 4.12b, green subunit) could be fit into the SAXS
envelope at either of two distinct locations within a single APOBEC3G subunit (Fig. 4.12a,
ribbon diagrams at CD1 and CD2). Of course, the exact orientation of the CDA domains is
unknown due to the resolution limits of the SAXS method. However, this model is
appealing because it agrees with the tandem CDA domain composition apparent in the
linear amino acid sequence (Fig. 4.10). As such, the current working model for
APOBEC3G is that the individual APOBEC3G subunits engage in a tail-to-tail interface with
a domain/subdomain distribution comprising: (i) an N-terminal, CDA-like catalytic domain
(CD1), (ii) an N-terminal non-catalytic domain (NCD1) with no known structural homology,
(iii) a second, C-terminal CDA (CD2), and (iv) a C-terminal non-catalytic domain (NCD2)
comparable to NCD1 (Fig. 4.12a).

Following the SAXS analysis of APOBEC3G, Goodman & Chen reported the crystal
structure of APOBEC2 [52]. Importantly, this molecule forms an elongated 127 Å tetramer
featuring a tail-to-tail subunit interaction analogous to that proposed for the APOBEC3G SAXS envelope (Fig. 4.12c and Section 4.9.2). At 224 amino acids APOBEC2 is <50% of the length of APOBEC3G (Fig. 4.10). However, the fact that the tandem ZDD motifs of APOBEC3G arose by gene duplication of an ancestral AID-like gene makes it possible to align each of its N- and C-terminal CDA domains with an APOBEC2 subunit (Fig. 4.10). Despite this primary structure similarity, the APOBEC2 tetramer, which appears cylindrical, does not fit readily into the boundaries of the dimeric APOBEC3G molecular envelope (Fig. 4.12a versus c). This difference may suggest a fundamental difference in the orientation of hA3G domains and subunits compared to APOBEC2, or could be a consequence of the SAXS method itself, which is representative of a solution structural ensemble [175]. Hence, although the level of sequence identity suggests the core ZDD fold of each APOBEC3G domain is represented by the APOBEC2 subunit, modes of substrate binding, dimerization and Vif interactions cannot be deduced for APOBEC3G at present, and will require high resolution crystallographic studies.

4.10.3 Characterization of Structural Changes in APOBEC3G Morphology in the Presence of RNA

A somewhat controversial aspect of the APOBEC3G story is the observation that the enzyme has been isolated from activated CD4+ T cells as an HMM (High Molecular Mass) ribonucleoprotein complex [176-178]. The dC-to-dU deaminase activity of such species is nominal, but can be restored in vitro by digestion with ribonuclease, which coincides with formation of enzymatically active, LMM (Low Molecular Mass) variants whose mass is consistent with a monomer or dimer (~47 kDa to 95 kDa) [176]. As such, it has been hypothesized that APOBEC3G binding to RNA, and subsequent formation of HMM complexes from LMM variants, serves to switch off its deaminase-dependent antiviral activity [177, 179]. Although this phenomenon appears to contradict the antiretroviral activity of this enzyme, it has been demonstrated that HMM complexes incorporate HIV
RNA, as well as non-autologous Alu and retrotransposable hY RNAs [177]. This RNA-binding activity may have evolved to inhibit retro-element reverse transcription and retrotransposition into the primate genome [180], rather than as a means to combat lentiviruses. The observation of an HMM complex also supports evidence for the observed deaminase-independent antiviral activity of APOBEC3G [143], which may simply be a consequence of the enzyme’s innate ability to bind non-specifically to ssRNA [181].

Genuine HMM complexes are localized within cellular P-bodies or stress granules that comprise multiple proteins, many of which are associated casually with APOBEC3G via RNA binding as ribonucleoprotein complexes [177, 181]. Nonetheless, HMM variants of reduced complexity, ranging in size from 5-15 MDa, have been isolated from CD4+ T cells using gel filtration chromatography [176]. HMM-like complexes from recombinant sources display similar size distributions and deaminase activity in vitro, and these complexes can be reduced to deaminase-active LMM-like species via RNase A treatment [175, 176, 182, 183] making them tractable targets for structural and functional investigations.

To address the question of how RNA affects the morphology and deaminase activity of APOBEC3G, Wedekind and Smith collaborated to examine a minimal, HMM-like form by SAXS [175]. The results revealed a significantly larger particle than the LMM-like sample with a maximum dimension of 210 Å (Fig. 4.12a versus d). The results of a circular dichroism analysis of LMM- versus HMM-like variants indicated no significant changes in the overall protein secondary structure composition in the presence of RNA, implying that RNA did not significantly alter the global protein fold [175]. Efforts to fit envelopes of LMM-like dimers into the HMM-like electron density utilized six-dimensional computational searches. This effort yielded two possible packing configurations, one of which is depicted in Fig. 4.12d. In both instances, optimal packing solutions suggested the possibility that APOBEC3G associates symmetrically with itself as a dimer of dimers (Fig. 4.12d, inset). Work by Landau and colleagues identified previously that the N-terminal CDA domain
(CD1) of APOBEC3G binds RNA, whereas the C-terminal CDA (CD2) deaminates ssDNA
[40]. If the proposed mode of packing in the minimal HMM particle is correct, it implies that
the N-terminal CDA domains of closely packed LMM-like particles (i.e., the dimer of
dimers) could be bridged via intermolecular RNA interactions. Coalescence of multiple
HMM-like particles would result in sequestration of CD2 domains responsible for dC-to-dU
deaminase activity. Thus, inactivation of APOBEC3G deaminase function by RNA and
formation of HMM complexes may simply be a property of locking the enzyme into an
inactive conformation through molecular crowding and steric overlap at the active site.
RNase degradation would necessarily break the RNA cross links and free each active site
to engage in substrate deamination. Importantly, the HMM-like model also predicts the
presence of discrete protein-protein interaction sites between APOBEC3G molecules (Fig.
4.12d, inset), but these interactions alone may be insufficient to form a core (stable)
structure without RNA binding.

Whether the HMM-like molecular envelope from SAXS is representative of a bona
fide cellular complex has yet to be determined. However, recent studies provide evidence
for the existence of minimal HMM-like ribonucleoprotein complexes inside the viral capsid.
Pulse labeling suggested that newly translated APOBEC3G exists in the cell in the LMM
form for only a short period ($t_{1/2} = 20$ min) before interacting with cellular RNA to become
HMM [74]. This investigation also implied that only the LMM form of APOBEC3G interacts
with viral components that become integrated into viral particles. Interestingly, LMM
APOBEC3G in viral particles is HMM-like in that it is associated with the RNA genome and
exhibits reduced deaminase activity. Moreover, the deaminase activity of this Intra-Virion
APOBEC3G Complex (IVAC) is recovered only after post-entry viral replication. The
authors suggested that restoration of deaminase activity is the result of the RNase H
activity of HIV reverse transcriptase, which removes viral genomic RNA from RNA-DNA
duplexes. This investigation further contends that prior work in which APOBEC3G was
isolated with viral particles, but did exhibit deaminase activity on ssDNA in vitro, was
caused by enzyme trapped within the viral particle, but outside the core assembly (i.e., not bound to viral RNA). This condition was attributed to overexpression of APOBEC3G in producer cells. Collectively these data suggest that conversion of minimal HMM to LMM variants \textit{in vivo} would promote APOBEC3G antiviral deaminase activity and deaminase-independent activity through binding to reverse transcription initiation and elongation complexes [184, 185]. Again, the relevance of the IVAC with respect to the HMM-like particles from SAXS (Fig. 4.12d) is uncertain, but the results of the biophysical characterization are consistent with a protein that undergoes morphological changes that are modulated by RNA, but do not impact the global fold of the fundamental protein structure [175].

It is not yet known if this RNA mediated HMM / LMM modulation is common to other APOBEC3 family members. Nor is it known if APOBEC3 family members will share a single mode of oligomerization. APOBEC3G is the first APOBEC3 enzyme to be analyzed from a biophysical perspective. Prolific duplication events, such as that observed for APOBEC3 enzymes, have been hypothesized to play an important role in primate adaptive evolution, particularly when a diversity of responses was advantageous [171]. Thus, we might expect each enzyme to exhibit unique features with regards to both its structure and function. With so many versions present, the forces of selection can tailor each APOBEC3 enzyme to fulfill a specific niche.

4.10.4 Positive Selection Exerted on the APOBEC3 Family

Studies of the selective pressures on the APOBEC3 family have also focused on APOBEC3G, as this was the enzyme first shown to diminish HIV infectivity [186]. Protein alignments of the APOBEC3G sequence among primate species demonstrated high sequence variation, thereby prompting an analysis of the evolutionary forces that influenced this protein. Positive selection is often observed for host proteins that are directly involved in pathogen defense [187], as suggested for a number of the APOBEC3 family members. This type of genetic conflict between host and pathogen drives rapid change in interacting host
and pathogen proteins as they attempt to increase or decrease interactions with one another in the battle for host cell dominance. As assessed across several primate species by the $d_{\text{N}}/d_{\text{S}}$ ratio, the APOBEC3G enzyme has been under strong positive selection for at least 33 million years [109, 149, 169]. Actually, with the sole exception of APOBEC3A, all human APOBEC3 enzymes exhibit some degree of positive selection as compared to chimpanzee. APOBEC3A, akin to AID, is under purifying selection [109]. The $d_{\text{N}}/d_{\text{S}}$ ratio for APOBEC3H is on par with that of A3G over a wide range of primate species, implying that these proteins have experienced similar evolutionary influences [149]. APOBEC3D/E, as compared between humans and chimpanzees, is the only example that appears to experience greater selective pressure than APOBEC3G. This pressure is specific to the “E” domain. When considered independently, the “E” domain exhibits the highest $d_{\text{N}}/d_{\text{S}}$ ratio of any APOBEC3 protein, whereas the “D” domain experiences mild positive selection comparable to that of APOBEC1 [109].

The take home message of these results is that most APOBEC3 family members show evidence of positive selection, with the possible exception of APOBEC3A. This would suggest that each protein has evolved on a species specific basis, and moreover, that the needs of each species are unique. In other words, there are two aspects to the selection observed in APOBEC3 family members. Upon a gene duplication event, there is the evolution of a specific activity for each family member. That is not observed in an analysis of $d_{\text{N}}/d_{\text{S}}$ ratios, as it likely occurred prior to the speciation events compared here. Rather, the positive selection described here is more supportive of the species specific adaptations (e.g., human APOBEC3B versus chimpanzee APOBEC3B).

Analysis of $d_{\text{N}}/d_{\text{S}}$ ratios are often more informative when generated for small regions of a sequence because, as established previously, not all regions of the CDA fold are equally amenable to mutation. Such sliding-window analyses of APOBEC3G suggest that pressures on APOBEC3G are widely dispersed throughout its sequence [109, 169]. Specifically, it has been estimated that 30% of the APOBEC3G protein residues have undergone strong positive
selection [109]. The proportion of nonsynonymous substitutions exceeds that of synonymous substitutions in all but one region of the sequence [169]. Reports of the strongest site of selection differ according to methods of analysis and species selected for comparison [109, 169, 188]. Together, they indicate that episodes of genetic conflict have applied pressure throughout the APOBEC3G sequence. This pressure is observed in a variety of primate species and is not exclusive to those susceptible to HIV/SIV infection [169].

Such dispersed pressure supports Zhang's hypothesis that APOBEC3G is experiencing diversifying positive selection rather than directional selection [169]. Directional selection occurs when natural selection favors a single allele, whereas diversifying selection may favor multiple alleles. This hypothesis was based on the observation that the HIV Vif infectivity factor that targets APOBEC3G for proteasomal degradation must evolve specific structures and sequences to bind APOBEC3G (directional selection), but APOBEC3G need only acquire changes that abolish viral recognition while maintaining deaminase activity (diversifying selection). Such escape behavior by APOBEC3G is evidenced by the preference for charge reversal observed in nonsynonymous substitutions [169]. A specific example is the observation that Asp128 in human APOBEC3G co-varies with the first basic amino acid in the HIV Vif sequence $^{14}$D$^{15}$R$^{16}$MR$^{17}$ [189-193]. Asp128 maps to Loop7 of the APOBEC2 structure (Fig. 10); the equivalent APOBEC2 residue (E159) engages in a salt bridge across the tetramer interface [52]. The comparable position in APOBEC3G from several non-human primates such as African green monkey is Lys(128), which interacts with SIV Vif through the equivalent sequence $^{14}$S$^{15}$E$^{16}$RQ$^{17}$ [194]. As an analogy for escape behavior, when a fox (vif) chases a rabbit (APOBEC3G), the rabbit can choose from a number of escape routes; the fox has only one route, that of the rabbit. This observation is a reversal of the typical roles played by pathogen and host, and perhaps is the reason why APOBEC3G continues to maintain a tandem CDA domain structure (i.e., one is a decoy for Vif at the N-terminus, whereas the other C-terminal CDA deaminates ssDNA).
In light of these correlations between HIV and APOBEC3G, it is important to note that the evidence of APOBEC3G selection far precedes the emergence of modern lentiviruses in the primate population [109], which can be traced back only as far as ∼7 million years [195]. Thus, although APOBEC3G may currently be running from the HIV Vif protein, this is only a recent phenomenon on the evolutionary timescale. Indeed, APOBEC3G exhibits antiviral activity against other, non-HIV viruses, such as the hepatitis B virus [140]. The very nature of APOBEC3G’s antiviral dC-to-dU deaminase activity supports the hypothesis of a broad target range. While each APOBEC3 enzyme exhibits a preference for the nucleotide neighbors nearest the editing site (e.g., APOBEC3G prefers the context 5'-C<sup>C</sup>A-3') [196, 197], no specific consensus sequence has been identified that is equivalent to the APOBEC1 mooring sequence (Section 4.9.3).

4.11 Conclusions and Future Prospects

It is appropriate to reflect on the generalities that are now apparent regarding the chemistry, phylogeny and three-dimensional structure of the APOBEC protein family. A hierarchical examination revealed that the ZDD signature motif, (H/C)xEx<sub>25-30</sub>PCxC, is highly conserved throughout evolution in terms of its primary and tertiary structures. Phylogenetic analysis revealed that the perpetuation of the ZDD motif in numerous species is attributable to the successful spatial presentation of key amino acids, as well as a Lewis acid Zn<sup>2+</sup> ion, that form the basis of a robust hydrolytic deamination module. The ZDD motif itself is incapable of adopting the requisite three-dimensional structure for function, and therefore must be supported by a larger self-organizing CDA domain of the form α1β1β2α2β3α3β4β5 with the conserved β-strand topology 2-1-3-4-5, in which helices α2 and α3 coordinate the essential Zn<sup>2+</sup>. Changes in the direction of the C-terminal β-strand, as well as deletions and insertions to loop elements that join core secondary structure elements appear to influence active site accessibility and substrate binding. Adaptations have culminated in specialized
enzymes that act on free nucleobases, nucleosides, as well as nucleotides embedded within single-stranded DNA or RNA substrates. The mode of subunit oligomerization plays a major role in substrate binding, activity and regulation. However, despite a rather divergent evolution, the core CDA domain remains intact, thus providing consistency with respect to the stereochemistry and regio-specificity of the overall reaction. One exciting result from the human genome project was the discovery of the APOBEC3 sub-family of CDAs, which exhibit antiretroviral activity that most likely evolved to quell the explosion in retrotransposable elements in primates ~40 million years ago. At present, some family members, such as APOBEC3F and APOBEC3G, exhibit an innate anti-HIV activity that manifests as a dC-to-dU hypermutation in viral DNA. The challenge of the next generation of scientists will be to develop ‘shields’ for these natural defense proteins that protect them from proteasomal degradation by masking their binding to viral proteins, such as the HIV Vif factor. To accomplish this task it will be necessary to understand the molecular basis of host-viral protein interactions in order to target the viral binding site, while leaving the host protein function intact. With this challenge, we return to Crick’s adage, “If you want to understand function, study structure”.

References


Figure 4.1  Schematic diagram for the cytidine deaminase mechanism of action  (Left) The cytidine-containing substrate binds in the CDA active site and proceeds to undergo nucleophilic attack by an activated water coordinated to Zn$^{2+}$. A conserved Glu donates a proton to the N3 position.  (Middle) The tetrahedral intermediate at C4 breaks down through loss of ammonia, which gains a proton from the conserved Glu.  (Right) The uridine product is free to diffuse away, leaving a protonated Glu that is poised for the next cycle of catalysis. A new water molecule enters the Zn$^{2+}$ coordination sphere.
Figure 4.2 Ribbon diagrams of the representative zinc dependent deaminase motif bound to cytidine  The coordinates were derived from the mouse cytidine deaminase crystal structure [43]. (A) The helix-strand-helix structure of the ZDD motif indicating the spatial orientation of key residues. Dashed lines (blue) indicate ionic coordination to Zn$^{2+}$; pink lines indicate hydrogen bonds. The activated water makes a close contact to atom C4 of the cytidine ring (gray line). Black arrows indicate the progression of the polypeptide chain from N- to C-terminus. (B) The ZDD signature motif in the context of the CDA domain (gray, semitransparent). The cytidine substrate is depicted as a space-filling model.
Figure 4.3  Relationships of the deaminase sequences in the vicinity of their ZDD motif

The neighbor-joining phylogenetic tree was generated from a protein alignment of the region encompassing the ZDD motif. Stars indicate families for which a representative crystal structure is discussed. Cytosine deaminases are labeled CD, whereas cytidine deaminases are identified as CDA. DA is an abbreviation of 'deaminase'. The scale bar represents the number of substitutions per amino acid. Clusters highlighted in dark gray are supported by a bootstrap value greater than 50, those grouped in a medium shading of gray
are clustered by function, and the largest groupings in the lightest shading reflect gene families based on sequence similarity. Bootstrap values for selected nodes are indicated. The labels at the end of each branch indicate the organism from which the sequence is derived. (Anopheles gambiae, Ag; Ascaris suum, As; Aspergillus terreus, Atr; Arabidopsis thaliana, At; Brugia malayi, Bm; Candida albicans, Ca; Caenorhabditis elegans, Ce; Ciona intestinalis, Ci; Dictyostelium discoideum, Dd; Dirofilaria immitis, Di; Drosophila melanogaster, Dm; Danio rerio, Dr; Encephalitozoon cuniculi, Enc; Echinococcus multilocularis, Em; Gallus gallus, Gg; Homo sapiens, Hs; Monodelphis domestica, Md; Macaca fascicularis, Mf; Mus musculus, Mm; Neurospora crassa, Nc; Oryctolagus cuniculus, Oc; Oryzias latipes, Ol; Oryza sativa, Os; Plasmodium falciparum, Pf; Plasmodium yoelii yoelii, Py; Rattus norvegicus, Rn; Saccharomyces cerevisiae, Sc; Schizosaccharomyces pombe, Sp; Trypanosoma cruzi, Tc; Tetraodon fluviatilis, Tf; Takifugu rubripes, Tr; and Xenopus laevis, Xl). Within each darker shaded area, multiple sequences from the same organism are distinguished by use of asterisks (where their distinct functions are unknown) or by use of digits, where they can be assigned to separate groupings (e.g. 1, 2, 3 for ADAR1, 2, or 3 family members; 1, 2, 3 for APOBEC1, 2, 3 or AID family members; and 1, 2, 3 [...] 9 for hypothetical plant CDAs). In the case of double-domain-containing APOBEC3 proteins, an [N] or [C] indicates whether the sequence is derived from the N-terminal or C-terminal ZDD motif. CeC33G8 is a gene of unknown function. This figure was adapted and reproduced with permission from Oxford University Press and originally appeared in [31].
Figure 4.4  Superposition of N- and C-terminal domains from the *E. coli* CDA

(A) The CDA subunit in monomeric form. The respective catalytic domain (CD) harboring Zn$^{2+}$ is colored dark. The C-terminal "pseudo" catalytic domain (PCD) is colored lightly.  (B) Separation of the domains by removal of the long linker observed in (A) at the far left. The PCD can be oriented similarly to the CD by $180^\circ$ rotation about the $x$-axis, which is indicative of the internal symmetry of CDA subunits.  (C) Translation of the CD (downwards) and the PCD (upwards) results in the superposition of these core domains.
Figure 4.5  Free nucleotide cytidine deaminases (fnCDAs): β-strand 5 antiparallel to β-strand 4  Representative members of the fnCDA family are labeled with species, enzyme name and PDB ID, and are shown in three forms. Left, a simplified topology diagram displays the connectivity of secondary structure elements. Secondary structure elements that comprise the conserved core fold characteristic of the CDA
superfamily are shown in red and blue and numbered sequentially. Embellishments to the core fold are colored gray and are excluded from the numbering scheme. All loops are labeled in the first example, and only as pertinent to the text throughout the rest of the figures. The zinc ion is shown in green. **Center,** ribbon diagram displays the three-dimensional structure of a subunit. Secondary structure elements are colored as in the topology diagram. Numbering is limited to α-helices and selected loops in the interest of clarity. N- and C-termini are indicated. The second *E. coli* CDA domain, believed to have resulted from a domain duplication event is colored in darker shades of red and blue. **Right,** ribbon diagram displays the three-dimensional structure of the oligomer. Scale bars indicate the length of the protein. A single subunit is colored as in subunit panel, but rotated 90° along the axis of the arrow. Remaining oligomers are colored green (and dark green for the duplicated domain of *E. coli* CDA), orange, and yellow, with non-conserved embellishments shown in tan. The final perspective looks down from the top to view the substrate binding region and catalytic zinc ion, shown in green. Substrates, if available, are depicted in gray stick format.
Figure 4.6  Cytosine Deaminase, Guanine Deaminase, and TadA: β-strand 5 parallel to β-strand 4  Figure layout is as described for Fig. 4.5. 3-D domain swapping in guanine deaminase is indicated in the following manner: the subunit is colored in red, blue, and gray as described previously. However, α-helix 4 and β-strand 5 do not contribute to the core fold of the subunit, but rather to the core of the opposing monomer. α-helix 4' and β-strand 5' of the opposing monomer, which contribute to the core fold of the displayed subunit, are labeled with primes (') and colored in green; non-conserved embellishments of the opposing subunit are shown in tan.
Figure 4.7  Deoxycytidylate Deaminases (dCDs) and APOBEC2: β-strand 5 parallel to β-strand 4  Figure layout is as described for Fig. 4.5. (Upper) For *N. europaea* multiple rotation axes indicate the relationship of the subunit to its position in the oligomer. The view of oligomer is still from the top down, displaying the substrate binding pocket. (Middle) The ~20 residues missing from L3 of the T4 dCD crystal structure are represented by a dashed line. T4 dCD exists as a hexamer shown above rotation axis. For clarity, the two subunit interfaces are boxed and expanded; the green box denotes the helical interface, whereas the yellow box denotes the β-sheet interface. The β-sheet interface boxed in yellow
is orientated similarly to the subunit depiction. (Lower) In APOBEC2, the subunit represented in the topology diagram and in the subunit ribbon diagram is that of the outer-most subunit of the tetramer. A depiction of the electrostatic potentials is mapped onto the APOBEC2 surface to demonstrate the relative charge of the protein surface. Red indicates a negative charge; blue indicates a positive charge, as predicted at pH 7.0. The orientation of the tetramer is as shown in the ribbon diagram. Arrows indicate a basic cleft with the potential to bind nucleic acid substrate. Figure generated by Swiss-PDBViewer [198].
Figure 4.8  Fusion protein Deaminases - RibG and ADAR2: β-strand 5 parallel to β-strand 4  The figure layout is as described for Fig. 4.5.  (Upper) For RibG, the monomer shown in the ribbon diagram is rotated 180° from the topology diagram, such that the ‘backside’ helices face forward. The fusion domain, comprised of a Rossmann fold, is shown in gray. The oligomeric region of interest is the β-sheet subunit interface; this is expanded and boxed in red. The full tetramer is shown at far right for perspective.  (Lower) ADAR2 is depicted as a single subunit and does not appear to oligomerize in the crystal structure. The topological orientation is similar to that of the three-dimensional subunit orientation. A second subunit orientation rotated 90° is depicted at the far right.
Figure 4.9  Schematic diagrams of substrate binding in the active sites of representative CDA superfamily members  (A) blasticidin S deaminase in complex with blasticidin S.  (B) Mouse cytidine deaminase in complex with activated cytidine.  (C) T4 deoxycytidylate deaminase in complex with deoxycytidine monophosphate.  (D) Yeast cytosine deaminase in complex with the substrate analog 4-(R)-hydroxyl-3,4-dihydropyrimidine.  (E) TadA deaminase from S. aureus in complex with the anti-codon stem loop RNA substrate (orange). Subunits are colored as in Figs. 3.5-8, substrates are colored white. Oxygen atoms are red, nitrogen atoms are blue. Residues omitted for clarity are represented as shaded ovals.  Those residues implicated in hydrogen bonds to substrate are represented with three letter codes; hydrogen bonds are indicated with dashed gray lines.  Ionic interactions with Zn$^{2+}$ are indicated in dashed green lines. Selected waters that mediate substrate-protein interactions are shown as red spheres. Residues are represented by side-
chain only, except for those where the backbone atoms are implicated in hydrogen bonding. Secondary structure features are selectively displayed and labeled according to Figs. 4.5-8. A cartoon secondary structure is provided for the RNA backbone. The perspective for each panel was chosen in the interest of clarity and key residues, and is not identical throughout the panels.
Figure 4.10 Structure-based sequence alignment of representative CDA superfamily members and a subset of APOBEC sequences. Four known structures corresponding to T4 dCD, ScCD, SaTadA and APOBEC2 are aligned with APOBEC family members whose structures are unknown. Red letters denote identical residues; green letters indicate similarity. Residues that coordinate Zn$^{2+}$ are indicated by green spheres. Residues comprising the ZDD signature motif are shown as white letters boxed in black. Secondary structure elements are indicated approximately by cartoons above the alignment; specific boundaries for known structures are indicated by shaded backgrounds such that blue represents β-strand, orange indicates α-helix and gray indicates non-conserved C-terminal helical ‘embellishments’. Residues that contact substrate, if known, are boxed in pink. Residues that contribute to oligomerization via the helical interface are boxed in blue; those that contribute to a β-sheet interface are boxed in black. Exon junction boundaries are indicated by a vertical line (golden). Representative members of the various types of APOBEC3 sequence are included: human APOBEC3G is a Z1a-Z1b combination, mouse APOBEC3 is comprised of a Z1a-Z2 combination and macaque APOBEC3H is an active example of the only primate Z2.
Figure 4.11  Phylogenetic tree of APOBEC3 family members illustrating the division into three main ‘Z’ groups  

A neighbor-joining tree was constructed in CLUSTAL_X based on a protein alignment of APOBEC3 sequences from humans and non-primate eutherian (placental) mammals. The N- and C-terminal domains of double-domain APOBEC3 proteins have been split and are designated -N and -C, respectively. APOBEC3 proteins from mammals in which more than one APOBEC3 was identified are numbered. An APOBEC3H-like domain has been conserved in a number of mammalian species, including mice, rats, dogs, horses, pigs, cows and primates. This phylogeny is rooted with the most closely related enzyme, AID, as an outgroup lineage; placement of this root is shown by an arrow. The scale bar represents a mutational frequency of 0.1 substitutions per residue.
Bootstrap support for the groupings is indicated by numbers next to the relevant branches; those nodes supported by a Bayesian maximum-likelihood analysis are indicated with an asterisk. This figure was reproduced with permission from the American Society for Microbiology and originally appeared in [149].
Figure 4.12  The global molecular envelope for APOBEC3G based on shape restorations from SAXS, and subunit interactions of crystallographically defined cytidine deaminases  

(A) APOBEC3G RNase treated at 3.4 nm resolution. The putative subunit boundaries are indicated in green and magenta. Possible domains and subdomains are labeled CD1, NCD1, CD2 and NCD2. Individual CDA domains from (B) are docked into the envelope at CD1 and CD2.  
(B) Transparent surface model and underlying ribbon representation of fnCDA Cdd1 from yeast defined crystallographically (PDB ID 1R5T). Individual CDA domains are colored green, red, blue and purple.  
(C). Transparent surface model and underlying ribbon diagram for the APOBEC2 crystal structure (PDB ID 2NYT). The scale bar in (A) applies to (A-C).  
(D) The molecular envelope restoration for a minimal, HMM-like variant of APOBEC3G. Computational searches led to a ‘dimer-of-dimers’ configuration of LMM-like subunits within the HMM molecular boundary (cyan). The remaining molecular envelope may comprise RNA, which is hypothesized to bridge neighboring subunits via an intermolecular CD1 (red) to CD1 (purple) interaction (inset), thereby sequestering the DNA deaminase sites at CD2.
Chapter 5:

Crystallization and X-ray Diffraction Analysis of the Trp/Amber Editing Site of Hepatitis Delta Virus (+)RNA: A case of rational design

The data presented in this chapter were published in:


Co-Authors who contributed to this research are:

Wedekind, J.E.
5.1 Footnotes

Abbreviations: *ADAR*, Adenosine Deaminase that Acts on RNA; *HDV*, Hepatitis Delta Virus; *HPLC*, High Pressure Liquid Chromatography; *I*, Inosine; *PEG MME*, Poly Ethylene Glycol Mono-Methyl Ether;
5.2 Abstract

RNA editing by mammalian ADAR1 (Adenosine Deaminase Acting on RNA) is required for the life cycle of the hepatitis delta virus (HDV). Editing extends the single viral open reading frame to yield two protein products of alternate length. ADARs are believed to recognize double-stranded RNA substrates via a "structure-based" readout mechanism. Crystals of 10-mer duplexes representing the HDV RNA editing site diffracted to 1.35 Å resolution, but suffered from merohedral twinning and averaging of the base registry. Expansion of the construct to include two flanking 3 x 1 internal loops yielded crystals in the primitive tetragonal space group $P4_12_12_1$ or $P4_32_12$. X-ray diffraction data were collected to 2.8 Å resolution, revealing a unit cell with dimensions $a = 62.5$ Å and $c = 63.5$ Å. We report the crystallization and x-ray analysis of multiple forms of the HDV RNA editing substrate, our encounters with common RNA crystal growth defects, and our approach to overcome these problems.
5.3 Introduction

RNA editing by members of the ADAR family converts adenosine to inosine through hydrolytic base deamination of the C6 exocyclic amine [1]. This activity fills two roles. Site-specific nuclear editing of endogenous mRNA transcripts diversifies the proteome [2], while promiscuous editing of duplex RNA has implications in RNAi and viral defense [3, 4]. The key to these dual functions resides, to some extent, in the RNA structure itself. Duplex mismatches contribute to bulging of targeted adenosines [5], although base-paired substrates also exist. Internal loops may act to either prevent binding of ADARs, thereby directing the enzyme to neighboring base-paired regions, or to create recognition sites through localized bending or widening of RNA grooves [6].

ADAR1 edits within the only open reading frame of the hepatitis delta virus, a subviral pathogen associated with chronic Hepatitis B infection [7, 8]. ADAR1 targets A_{1012} of an A•C mismatch within a double stranded region of the ~1700 nt circular, single-stranded HDV RNA anti-genome. Although there are multiple A•C mismatches in the HDV genome, editing is exclusive to A_{1012}. The UAG amber ‘stop’ codon is edited to UIG, which is read by the ribosome as the tryptophan codon UGG. This modified transcript generates a new delta antigen protein (HDAg-p27), 19 amino acids longer than the unedited precursor (HDAg-p24). These two forms are reciprocally required in the viral life cycle; p24 enhances transcription, whereas p27 allows packaging [9, 10].

To provide insight into the special structural features of RNA that promote ADAR1 binding and editing, we undertook a structural analysis of several minimal substrates. This manuscript describes our progress toward a structure of the HDV editing site. Specifically, we document encounters with crystal twinning and base-averaging, as well as strategies to overcome this problem that should be generally applicable to the crystallization of RNA helices.
5.4 Materials and Methods

5.4.1 Materials

RNA oligonucleotides were synthesized by Dharmacon Inc. (Lafayette, CO). Poly(ethylene) glycol mono methyl ether (PEG MME) and triethylamine (TEA) were from Fluka (Buchs, Switzerland). Spermidine was from Sigma (Milwaukee, WI). HPLC grade acetonitrile was from JT Baker (Phillipsburg, NJ) and glycerol was from MP Biomedicals (Irvine, CA). All other chemicals were from Fisher Scientific (Fairlawn, NJ). Crystal Tools were from Nextal (Montreal, QC); VDX plates and silanized glass cover slips were from Hampton Research (Laguna Niguel, CA). Cover slips were autoclaved prior to use. All water was autoclaved and originated from a UV/UF Nanopure Infinity™ water polishing system (Barnstead, IA).

5.4.2 RNA Purification

Oligonucleotides were deprotected according to manufacturer’s instructions except that heating was for 1 hour at 338 K. Individual RNA strands were purified on a 1.9 x 30 cm μBondapack C18 column (Waters, Milford, MA) on a Waters HPLC equipped with Model 510 pumps, an automated gradient controller and a Waters 484 MS Absorbance Detector interfaced with the Maclntegrator software (Varian, Palo Alto, CA). HPLC separations employed the TEA/TEAA (triethyl ammonium acetate) ion-pairing system at pH 7.0. Buffer A was 0.1 M TEA/TEAA; Buffer B comprised Buffer A plus 50% (v/v) acetonitrile. Preparative-scale RNA separations were eluted at 7 mL min\(^{-1}\) with linear gradients using 15-30% (v/v) Buffer B over 60 min intervals. RNA elution was monitored at 260 nm; fractions were pooled and lyophilized to ~1 mL volume. Concentrated RNA was desalted using SepPak Vac C18 cartridges (Waters), which entailed loading, water washing, and batch elution with 40% (v/v) acetonitrile. Variations of HPLC purification and desalting have been described previously [11]. Individual strands of desalted RNA were lyophilized to dryness and resuspended in 10 mM sodium cacodylate buffer pH 6.0 to a final working concentration of 1-2 mM. Equimolar
concentrations were mixed to a final concentration of 0.5 mM for the 10 x 10-mer and 0.9 mM for all other constructs. Annealing was performed on longer constructs at 348 K for 5 min, then slow-cooled to room temperature. Samples were spun for 10 min at 12045g at 277 K prior to crystal set-ups.

5.4.3 Crystallization

Eight unique HDV RNA constructs were subjected to crystallization by hanging-drop vapor diffusion at 20°C, based on screens previously established for small ribozymes [11, 12]. RNAstructure version 4.2 was used as a design resource for thermodynamic folding calculations [13] (Fig. 5.1). A volume of 1 µL of precipitating agent from a 1 mL reservoir was added to 1 µL of RNA. Precipitants included combinations of Li⁺, Na⁺, K⁺, NH₄⁺ or Mg²⁺ with counter ions Cl⁻, SO₄²⁻ or acetate, as well as high molecular weight polymers PEG and PEG MME. All trials included 2 mM spermidine, 10 mM MgCl₂, and 0.1 M Na-cacodylate buffer pH 6.0. Droplets were examined 3 to 6 days after set-ups and weekly thereafter by use of an 80x MZ9.5 dissecting microscope (Leica, Wetzler, Germany) equipped with a polarizer and analyzer.

5.4.3.1 Streak-Seeding, Microseeding and Macroseeding

To improve crystal quality and size, three seeding methods were employed in conjunction with other optimization methods. Streak-seeding was performed on equilibrated drops containing 1:1 mixtures of RNA and reservoir solution. A natural fiber was dragged once through an over-nucleated “source” drop then once through a clear “receiving” drop at lower precipitant concentration. For microseeding, over-nucleated crystals were crushed, serially diluted (10¹ to 10⁶-fold) with mother liquor, and 0.5 µL was added to the receiving drop. For macroseeding, small crystals were transferred, using either a 20 µm rayon loop (Hampton Research) or a P10 Gilson pipette (Middleton, WI), into a receiving drop with 0-5% less precipitant. Receiving drops were prepared 24 hr prior to seeding with 2 µL each of precipitant and RNA, at concentrations of 0.5 or 0.9 mM.
5.4.4 X-ray Diffraction Experiments

Single crystals were cryoprotected by 3 min serial transfers through synthetic mother-liquors containing 5, 10, 15 and 20% (v/v) glycerol or MPD. Crystals were captured by surface tension in 20 µm rayon loops and flash-cooled to 100 K by a stream of cold nitrogen gas (X-stream, Rigaku/MSC, TX). X-rays were generated by an RU-H2R rotating-anode generator (Rigaku/MSC) operated at 4.5kW and equipped with a 0.3 mm focal cup. X-rays were filtered and focused by confocal optics (Rigaku/MSC) and collimated through a 0.5 mm front-end pinhole. Diffraction data were recorded on an R-AXIS IV image-plate detector equipped with a vertical φ-rotation axis at a crystal-to-detector distance of 15 cm. A total of 78˚ was collected for the 23 x 19-mer as 0.5˚ oscillations per 15 min exposure.

Additional data were collected at station A1 of the Cornell High Energy Synchrotron Source (CHESS, Ithaca, NY) on a Quantum 210 CCD detector (ADSC, Poway, CA) at crystal-to-detector distances of 13.5 and 15.0 cm for 10 x 10 edited and unedited samples, respectively. Data were collected with a ∆φ of 0.5˚ per image and exposure times of 20 to 30 sec totaling 140˚ and 120˚. An additional low resolution pass was collected for the edited crystal at a distance of 30 cm with an exposure time of 3 sec per 0.5˚. Intensities were reduced and merged with CrystalClear [14]. X-ray diffraction and data reduction statistics are reported in Table 5.1.

5.4.5 Structure Determination, Refinement and Twinning Analysis

Structures of the RNA duplexes were solved by molecular replacement using AMoRe [15]. Amplitudes were generated by TRUNCATE, a component of CCP4 [16]. Structures were refined in CNS [17] and electron density maps were examined in O [18]. Twinning was diagnosed on the Merohedral Crystal Twinning Server (http://nihserver.mbi.ucla.edu/Twinning/) courtesy of the Todd Yeates laboratory using 95.5% of reflections from 3 to 6 Å resolution [19]. Detwinning was performed in CNS. Refinement statistics for the edited 10 x 10-mer are reported in Table 5.1.
5.5 Results and Discussion

5.5.1 10 x 10-mer Results

5.5.1.1 Crystallization Screening

The initial target of these studies focused on duplex 10-mers encompassing the minimal substrate capable of supporting ADAR1 activity [20]. A•C mismatch and I•C base-pair targets were tested in both blunt and sticky-ended configurations (Fig. 5.1b). Crystals grew quickly for all four constructs, producing showers of needles in 0.5 M Li⁺, Na⁺ and K⁺ acetates with 15-30% (w/v) PEG 2000 MME; additional crystals were observed in 0.5 – 3.8 M solutions of Li⁺, Mg²⁺ and NH₄⁺ sulfates with and without PEG 2000 MME. Streak seeding was performed on clear drops, yielding single crystals in the form of long rods ~1.0 mm x 0.1 mm x 0.1 mm for the blunt-ended, unedited construct and in a trigonal habit, of size 0.5 mm x 0.15 mm x 0.15 mm for the blunt-ended, edited crystals (Figs. 5.2a & b).

5.5.1.2 Diffraction Experiments

Edited crystals (Fig. 2a) were serially cryoprotected into a synthetic mother liquor containing 20% (v/v) glycerol, 30% (w/v) PEG 2000 MME, 0.10 M Na-cacodylate, 2 mM spermidine, 10 mM magnesium chloride and 0.50 M lithium acetate. Crystals diffracted to 1.6 Å resolution in-house and better than 1.4 Å resolution using synchrotron radiation. The space group and unit cell dimensions were determined to be rhombohedral and were treated in the hexagonal setting, H3, with a = 39.4 Å and c = 248.2 Å. Low and high resolution synchrotron-derived data sets were integrated together during processing with CrystalClear. An analysis of the Matthew’s coefficient by CCP4 suggested 2 or 3 10-mer duplexes (20 or 30-nt) per asymmetric unit (Vₘ = 2.9 Å³Da⁻¹ with 58 % solvent or Vₘ = 2.0 Å³Da⁻¹ with 36.6 % solvent).

Unedited crystals (Fig. 5.2b) were cryoprotected similarly with a synthetic mother liquor containing 20% (v/v) glycerol, 0.10 mM Na-cacodylate, 2 mM spermidine, 10 mM
magnesium chloride and 2.3 M ammonium sulfate. These crystals diffracted anisotropically to ~3 Å resolution at home. Crystals also indexed as H3 with unit cell dimensions \( a = 49.1 \text{ Å} \) and \( c = 239.5 \text{ Å} \). The difference between edited and unedited unit cells (10 Å along the \( a \)-and \( b \)-axes and 9 Å along the \( c \)-axis) was considerable, given the single atom change between constructs. Data collection at CHESS extended the resolution of the unedited construct to 2.8 Å. An analysis of \( V_M \) gave \( 2.9 \text{ Å}^3 \text{Da}^{-1} \) (56.8% solvent), consistent with 3 molecules per asymmetric unit.

5.5.1.3 Structure Determination and Twinning

Molecular replacement was performed using a 10-mer duplex search model derived from PDB ID 405D [21]. Trials to independently place one model in three positions using data from 2.0 to 15 Å resolution yielded a correlation coefficient of 80.1% with an R-factor of 44%. Refinement proceeded in CNS with 7% of the data reserved for cross-validation (\( R_{\text{free}} \)). After rigid body refinement, positional and individual B-factor refinement, the \( R_{\text{work}} \) failed to drop below 35%. Inspection of electron density maps prompted submission of data to the Merohedral Crystal Twinning Server, which identified a positive result for a 2-fold operator in the [110] plane. Considerable twinning fractions were obtained for all 10-mer duplex data sets, ranging from 0.35 to 0.45. The CNS routine ‘detwin partials’ was implemented with a twinning fraction of 0.38 and refinement repeated. Electron density maps were much improved. Specifically, the electron density for ribose and phosphates was well resolved, however the base density was incomplete and the \( R_{\text{work}} \) could not be brought below 30% at 1.6 Å resolution. This was likely a result of base-averaging due to the high symmetry of the construct, as previously observed [22]. Analysis of the existing model using CURVES suggested this duplex may exhibit elements of A′-form character, including a widened major groove and decreased inclination angle [23, 24]. Poly(I)-poly(C) RNA is known to form A′-form RNA, however this has not been documented resulting from a single inosine [25]. A fully refined structure will determine if this deviation is a propagated effect of the single IC pair, or more likely, a result of the afore-mentioned base averaging.
The unedited construct was also solved by molecular replacement using the edited model of this study, which was divided into two search models, a 20-mer duplex and a 10-mer duplex. Utilizing data between 2.9 and 15 Å yielded a correlation coefficient of 57.4% with an \( R\)-factor of 52.6%. Packing suggested a potential solution, although a 6.9 Å gap existed between end-to-end packing molecules. Detwinning (twin fraction: 0.41) and subsequent refinement yielded traceable phosphate backbones in electron density maps, although these were of much poorer quality due to the lower resolution and anisotropy of the diffraction data. These data were set aside when the \( R_{work} \) could not be brought below 49%.

**5.5.2 23 x 19-mer Results**

### 5.5.2.1 Crystallization

Following the experience of twinning and base-averaging, RNA crystallization constructs were expanded to include two flanking 3 x 1 internal loops within the endogenous HDV sequence (Fig. 5.1a), thereby eliminating the high symmetry of the original 10-mers. Two modest modifications were made to the native sequence to promote folding and stability: (i) a closing G-C pair was reversed to C-G and (ii) a stem U•G wobble pair was converted to a G-C pair (Fig. 5.1c). The final blunt-ended construct incorporated a 23-nucleotide edited strand and a complementary 19-mer.

Screening yielded small crystals 7-14 days after set-ups in conditions of 0.5 \( M \) Na\(^+\), Li\(^+\) and K\(^+\) acetates with ~25% (w/v) PEG 2000 MME for both edited and unedited constructs. Additional edited crystals were observed in 2.5 \( M \) ammonium sulfate. Crystals from acetate salts diffracted to ~10 Å resolution after cryoprotection in 10% (v/v) glycerol. Further cryoprotection induced cracking. Increasing acetate and PEG concentrations yielded edited crystals with a cubic habit and dimensions of ~150 µm. These crystals diffracted to 5 Å resolution after serial cryoprotection to 20% (v/v) glycerol. A series of micro- and macro-seeding experiments were performed to increase crystal size. Macroseeding proved successful for the I•C crystals, and although deformations could be visualized, these crystals
diffracted to ~4 Å resolution when capillary mounted and exposed on the home x-ray source. Beginning as a 100 µm$^3$ seed, the best macroseeded crystal grew to 0.25 mm x 0.2 mm x 0.2 mm (Fig. 5.2c) after 4 weeks in conditions of 0.75 M potassium acetate and 25% PEG 2000 MME. This crystal diffracted to 2.8 Å resolution at home after cryoprotection in MPD; MPD was used as a cryoprotectant after trials in PEG 550, sucrose, ethylene glycol and glycerol were observed to damage crystals. Attempts to improve crystal growth by removing the last A-U base pair (making a 22 x 18-mer, Fig. 5.1c) and by converting the first 3 x 1 internal loop into a 3-nt bulge loop (22 x 17-mer, Fig. 5.1c) yielded no crystals.

**5.5.2.2 Diffraction Experiments**

CrystalClear indexed the 23 x 19-mer crystals as cubic. However, crystals appeared birefringent and reduction in $P23$ produced an $R_{merge}$ of 61%; data were therefore processed as $P422$. Analysis of $V_M$ suggested one molecule per asymmetric unit with a value of 2.3 Å$^3$ Da$^{-1}$ and solvent content of 47%. Zones displaying systematically absent reflections were measured and $I/\sigma(I)$ values for the $h00$ reflection class were consistent with screw axes along $a$ and $b$ ($2n$ present, where $n$ is an integer). Similarly, the $00l$ reflections supported a $4_1$ or $4_3$ screw axis along the $c$-axis ($4n$ present). This suggested a space group of $P4_12_12_1$ or its enantiomorph.

**5.5.2.3 Structure Determination**

Multiple molecular replacement attempts with AMoRe and Phaser 3.1 [26] failed to produce a satisfactory solution using models derived from PDB IDs: 1SDR, 405D or 1NUJ [21, 27, 28]. Experimental MAD phasing methods using halogenated bases, 2'-deoxy-2'-seleno ribose or 5'-seleno are currently being pursued [29].

**5.6 Acknowledgements**

This work was funded in part by NIH grant GM63162 (J.E.W.) and T32 training award GM068411 (C.M.). The authors thank Jolanta Krucinska for assistance with crystal growth,
and the staff of CHESS for assistance with data collection at the A1 station. CHESS is supported by the NSF under award DMR-0225180 and NIH through NCRR award RR-01646.
References


Table 5.1 X-ray diffraction data and refinement statistics

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>Edited 10 x 10</th>
<th>Unedited 10 x 10</th>
<th>Edited 23 x 19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space Group</td>
<td>H3</td>
<td>H3</td>
<td>P4₁2₁2 or P4₃2₁2</td>
</tr>
<tr>
<td>Cell Dimensions (Å)</td>
<td>a = 39.5</td>
<td>a = 49.0</td>
<td>a = 62.5</td>
</tr>
<tr>
<td></td>
<td>c = 246.4</td>
<td>c = 235.8</td>
<td>c = 63.6</td>
</tr>
<tr>
<td>Resolution</td>
<td>29.92-1.35</td>
<td>23.38 – 2.85</td>
<td>36.29 – 2.80</td>
</tr>
<tr>
<td></td>
<td>(1.40-1.35)</td>
<td>(2.95 – 2.85)</td>
<td>(2.90 – 2.80)</td>
</tr>
<tr>
<td>Total Reflections</td>
<td>122008</td>
<td>16052</td>
<td>16572</td>
</tr>
<tr>
<td>Unique Reflections</td>
<td>30558</td>
<td>4798</td>
<td>3307</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.99 (1.78)</td>
<td>3.35 (3.67)</td>
<td>5.01 (4.82)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>96.8 (77.4)</td>
<td>97.2 (99.8)</td>
<td>97.0 (92.7)</td>
</tr>
<tr>
<td>&lt; I / σ (I) &gt;</td>
<td>13.1 (2.2)</td>
<td>10.3 (1.9)</td>
<td>11.4 (2.2)</td>
</tr>
<tr>
<td>R(sym) (%)</td>
<td>6.3 (35.4)</td>
<td>6.1 (37.9)</td>
<td>9.9 (33.4)</td>
</tr>
<tr>
<td>Twinning fraction</td>
<td>0.38</td>
<td>0.41</td>
<td>None</td>
</tr>
</tbody>
</table>

Refinement Statistics

| Resolution       | 30.0 – 1.5 |
| Number of atoms (RNA/water) | 1257 / 26 |
| R(work) / R(free) (%) c,d | 31.9 / 31.8 |
| r.m.s.d. Bonds (Å) | 0.0038 Å |
| r.m.s.d. Angles (%) | 0.94° |

a Highest resolution shell statistics are noted parenthetically.

b \( R_{sym} = \Sigma | I | - < I > | / \Sigma | I | \times 100 \)

c \( R_{work} = \Sigma | F_o - k F_c | / \Sigma | F_o | \times 100 \). \( R_{work} \) was calculated using 93% of the data.

d \( R_{free} \) is defined as the \( R_{work} \) calculated using 7% of the data selected randomly and excluded from the refinement.
Figure 5.1  Schematic diagram of HDV RNA secondary structures harboring the Trp/Amber editing site at an A•C mismatch  Mutations from the native sequence are in green. Boxed or circled bases were removed during construct development to optimize crystallization. (a) A double-stranded region of the native antigenomic RNA. The edited position is shown in red. (b) The 10 x 10-mer minimal substrate constructs, including sticky- and blunt-ended, edited and unedited versions (4 total). Sites where inosine was substituted for adenosine in the ‘edited’ constructs are marked by a red ‘I’. (c) The asymmetric 23 x 19-mer crystallization constructs (edited and unedited). The 23 x 19-mer was truncated to an edited 22 x 18-mer, then to 22 x 17-mer, which were used in crystallization trials without success.
Figure 5.2  Representative crystals of the HDV Trp/Amber RNA editing site

(a) Unedited 10 x 10-mer RNA constructs grew as 1 mm rods with a diameter of 0.1 mm. (b) Edited 10 x 10-mer crystals displaying a trigonal habit are 0.5 mm in length (c). Edited 23 x 19-mer crystals grown from macroseeding experiments reached 0.25 mm in each dimension. Crystals were photographed under polarized light.
Appendix 1:

Functional Characterization of APOBEC-1

Complementation Factor Phosphorylation Sites

The data presented in this chapter were published in:

*Biochim Biophys Acta.- Molecular Cell Research*


The authorship order of this publication was as follows: David M. Lehmann, Chad A. Galloway, Celeste MacElrevey, Mark P. Sowden, Joseph E. Wedekind, and Harold C. Smith.
A1.1 Footnotes

**Abbreviations:** *APOBEC-1*, apolipoprotein B editing catalytic subunit 1; *ACF*, *APOBEC-1* Complementation Factor; *apoB*, apolipoprotein B; *CREB*, cAMP-response element-binding protein; *RRM*, RNA Recognition Motif; *PP1*, Protein Phosphatase 1; *PKC*, Protein Kinase C; *PKA*, Protein Kinase A; *NT*, N-terminal; *CT*, C-terminal.

**Author Contribution:** The author’s contributions to this work are encompassed by Figs. A.6 and A.7. The author generated a sequence alignment between two RNA binding proteins HuD and ACF. This alignment was used to generate a homology model of RRM1 and 2 of ACF.
A1.2 Abstract

ApoB mRNA editing involves site-specific deamination of cytidine producing an in-frame translation stop codon. Editing minimally requires APOBEC-1 and APOBEC-1 Complementation Factor (ACF). Metabolic stimulation of apoB mRNA editing in hepatocytes is associated with serine phosphorylation of ACF localized to editing competent, nuclear 27S editosomes. We demonstrate that activation of protein kinase C (PKC) stimulated editing and enhanced ACF phosphorylation in rat primary hepatocytes. Conversely, activation of protein kinase A (PKA) had no effect on editing. Recombinant PKC efficiently phosphorylated purified ACF64 protein in vitro, whereas PKA did not. Mutagenesis of predicted PKC phosphorylation sites S154 and S368 to alanine inhibited ethanol-stimulated induction of editing suggesting that these sites function in the metabolic regulation of editing. Consistent with this interpretation, substitution of S154 and S368 with aspartic acid stimulated editing to levels comparable to ethanol treatment in control McArdle RH7777 cells. These data suggest that phosphorylation of ACF by PKC may be a key regulatory mechanism of apoB mRNA editing in rat hepatocytes.
A1.3 Introduction

ApoB mRNA editing occurs post-transcriptionally and causes the site-specific deamination of cytidine 6666 to uridine thereby creating an in-frame translation stop codon [1, 2]. Consequently, two different apoB proteins are expressed, full-length apoB100 and the truncated protein apoB48, each with distinct physiological functions [3]. Minimally, apoB mRNA editing requires the cytidine deaminase APOBEC-1 as a homodimer [4, 5], APOBEC-1 complementation factor (ACF) [6], and the RNA substrate. Limited tissue expression of APOBEC-1 restricts editing in humans to the small intestine (> 85% editing), whereas in some rodents apoB mRNA editing also occurs in the liver where it is subject to metabolic regulation [2, 3, 7, 8].

Under normal, physiological conditions, apoB mRNA editing is a nuclear event occurring on spliced and polyadenylated RNA [9, 10]. However, in vitro editing activity could be detected in both cytoplasmic and nuclear S100 extracts [11, 12]. In vitro and in vivo data demonstrate that the proteins involved in editing exist in two distinct complexes; active nuclear 27S editosomes and inactive, 60S cytoplasmic complexes that can be re-organized into active 27S complexes in vitro [11, 13, 14]. Proteins that form these complexes traffick between the cytoplasm and nucleus and their subcellular compartmentalization is regulated to control editing activity [14-17].

ApoB mRNA editing is regulated by developmental, hormonal, and dietary factors as well as being under tissue-specific control [7, 18-23]. Editing can be stimulated primarily through two mechanisms: increased APOBEC-1 expression [24] and/or through modulation of pre-existing editing factors [25]. The fasting/refeeding and insulin induced models of editing are associated with increased APOBEC-1 abundance [24]. Conversely, induction of editing with ethanol was not dependent on de novo protein or RNA synthesis [25], suggesting modification or redistribution of pre-existing auxiliary factors were sufficient to induce editing.
In both cases, the mechanism of induction involved the redistribution of ACF from the cytoplasm to the nucleus [14].

Recently, ACF was shown to be a phosphoprotein in rat primary hepatocytes [26]. The phosphorylated form was strictly localized to nuclear 27S editing complexes, the biological site of editing [26]. Dephosphorylation of ACF diminished in vitro editing activity and the interaction between ACF and APOBEC-1, but did not significantly affect RNA-binding activity. Induction of editing by ethanol and insulin resulted in nuclear accumulation of phosphorylated ACF [14, 26] suggesting that phosphorylation is a common mechanism to modulate editing in response to metabolic stimuli. Inhibition of protein phosphatase 1 (PP1) activity was associated with nuclear accumulation of ACF, increased recovery of phosphorylated ACF and robust editing levels [26]. However, the protein kinase(s) responsible for phosphorylating ACF remain an outstanding question.

We identified serine 154 (S154) and serine 368 (S368) as two candidate phosphorylation sites. McArdle RH7777 cell lines expressing S154 and S368 mutated to alanine maintained normal basal editing levels, but these levels were refractory to ethanol induction. Conversely, substitution of aspartic acid at these sites significantly increased editing in the absence of ethanol to levels comparable with those achieved by ethanol treated cells expressing wild type ACF. The data support the possibility that these sites are necessary for ethanol-induced editing. Residues S154 and S368, are predicted PKC phosphorylation sites, and are conserved between rat and human ACF. Studies with PKC and PKA activators demonstrated that modulation of PKC activity enhanced editing levels in rat primary hepatocytes, whereas PKA did not. Similarly, purified ACF protein was efficiently labeled by PKC isozymes, but not by PKA. These findings are discussed in terms of structural and functional predictions of ACF phosphorylation.
A1.4 Materials and Methods

A1.4.1 Isolation of Rat Primary Hepatocytes and Protein Kinase Activator Studies

Rat primary hepatocytes were isolated from normal, fed male Sprague-Dawley rats (Charles River Laboratories, Wilmington, MA) (250-275 g BW) and cultured on BIOCOAT type I collagen coated, 60 mm plastic dishes (Becton Dickinson Labware, Franklin Lakes, NJ) in Waymouth’s media (Sigma Chemical Co., St. Louis, MO) containing 0.1 nM porcine insulin (Sigma) as described previously [14]. Primary hepatocytes were treated for 6 hours with protein kinase activators indolactam V, 8-cpt cAMP, and forskolin at the concentrations encompassing their respective in vivo EC\text{50} values, as described by the manufacturer (Calbiochem, La Jolla, CA). Total cellular RNA was harvested in TriReagent (MRC, Cincinnati, OH) RNA was harvested and processed through the poisoned-primer extension assay to quantify the effect on apoB mRNA editing [14].

To evaluate the effect of protein kinase activation on ACF phosphorylation, cultures were pre-incubated for 2 hours with 12 or 120 µM indolactam V in phosphate-free Waymouth’s media and subsequently supplemented with 0.5 mCi \textsuperscript{32}PO\textsubscript{4} and incubated for an additional 4 hours. Cultures were harvested and subcellular extracts prepared. ACF was immunoprecipitated, resolved by SDS-PAGE, and specific activity was quantified as described previously [26].

A1.4.2 Recombinant ACF Purification and \textit{in vitro} ACF Phosphorylation

Recombinant ACF64 was expressed in \textit{E. coli} and purified as described [27]. \textit{In vitro} ACF labeling reactions were carried out using 2 µg of purified recombinant ACF64 in 5 mM MgCl\textsubscript{2}, 100 µM CaCl\textsubscript{2}, 1 mM DTT, 5 µCi \textsuperscript{32}P-\gammaATP, 25 mM Tris-HCl pH 7.5 and 1X Lipid Activator (Invitrogen, Carlsbad, CA) with 25 mU of each PKC isozyme (Calbiochem) or 40 mM Tris-HCl pH 7.4, 20 mM Mg Acetate, 5 µCi \textsuperscript{32}P-\gammaATP and 25 mU of bovine PKA catalytic subunit (Promega, Madison, WI). Reactions were incubated at 30 °C for 10 minutes,
terminated by acetone precipitation, resolved by 10.5% SDS-PAGE, and transferred to nitrocellulose (Bio-Rad, Hercules, CA). Western transfers were probed with anti-HA monoclonal antibody (Covance Research Products, Berkeley, CA) and developed using Western Lightning reagent (NEN, Boston, MA). $^{32}$P incorporation was monitored by phosphorimager scanning densitometry. PKA enzymatic activity was confirmed via labeling of Kemptide peptide (Promega).

**A1.4.3 ACF Site-directed Mutagenesis and Functional Analysis**

Rat ACF64 with a C-terminal V5 epitope tag was cloned into a modified pcDNAIII vector (Invitrogen) [14]. ACF64 cDNA was mutated at specific bases to convert specific serine or threonine codons to alanine or aspartic acid codons using the QuikChange® Multi System (Stratagene, La Jolla, CA). Mutagenic primers (T49AT50A; CCAGGCTGGGATGCTGACCTCCTGAAAGGGGCTGC, T49DT50D; CCAGGCTGGGATGACGATCCTCCTGAAAGGGGCTGC, S132A; GGGCGTCTGTGCTGTGGACAACGTGCGG, S132D; GGGCGTCTGTGCTGTGGACAACGTGCGG, S154A; GAGAGAAGAAATCTTGGCAGAGATGAAAAAGGTC, S154D; GAGAGAAGAAATCTTGGACGAGATGAAAAAGGTC, S171A; GTCATTGTCTTACCAGCCGCTGCGGATAAAACC, S171D; GTCATTGTCTTACCAGCCGCTGCGGATAAAACC, S176A; GTCTACCACGCGCTGCTGATAAAACCGGGG, S176D; GTCTACCACGCGCTGCTGATAAAACCGGGG, S188A; GCCTTTGTGAATATGAAGCTCACCAGCGGACGCG, S188D; GCCTTTGTGAATATGAAGCTCACCAGCGGACGCG, S368; CTACCAAGGACATCTCGCCCACAGAGCTCCTCAGCCG, S368D; CTACCAAGGACATCTCGCCCACAGAGCTCCTCAGCCG) were extended using PfuTurbo® DNA polymerase according to the manufacture’s recommendations. The accuracy of the mutations was verified by sequencing the entire cDNA using T7 (Promega, Madison, WI), V5
(Sigma Genosys), and Seq481 (Sigma Genosys, The Woodlands, TX) sequencing primers (T7, TAATACGACTCTATAGGG, V5; CTAGAAGGCACGTCGAGGC, Seq481; GAACGAGTTGTGATGTCATTG) and using Big Dye sequencing system (Applied Biosystems, Foster City, CA).

McArdle RH 7777 (McArdle) cells maintained in DMEM (Gibco-BRL, Carlsbad, CA; supplemented with 10% FBS, 10% horse serum, 1% Penicillin, Streptomycin, Fungizone (Invitrogen)) were transfected with 2 µg plasmid containing each ACF64 mutant using Fugene 6 (Roche, Mannheim, Germany). Stable cell lines were created by G418 selection. When applicable, stable cell lines were treated with 0.9% ethanol for 4 hours. RNA editing of endogenous apoB mRNA was assessed by poisoned-primer extension analysis as previously described [14].

Expression of ACF mutants was assessed by western blotting of whole-cell extracts prepared by lysing cells for 30 minutes in RIPA buffer (50 mM Tris pH 7.4, 1% NP-40, 0.25% sodium deoxycholate, 150 mM NaCl, 1 mM EDTA, 0.2 mM DTT) supplemented with 50 mM NaF and protease inhibitors. Equal amounts of cell extract protein were resolved by 10.5% SDS-PAGE, transferred to nitrocellulose, and probed with anti-V5 antibody (KPL; Gaithersburg, MD).

McArdle RH7777 cells were transfected with pcDNA3 (Invitrogen) containing rat ACF cDNA mutated at suspected phosphorylation sites using FuGene6 (Roche; Basel, Switzerland). Cells were harvested 48 hours post-transfection and fractionated into nuclear and cytoplasmic fractions using the NE-PER kit (Pierce; Rockford, IL). Lysates, 40 µg of nuclear and 60 µg of cytoplasmic, were resolved using 8% or 10% SDS-PAGE and transferred to nitrocellulose (BioRad, Hercules, CA). Blots were probed with anti-histone H1 (Santa Cruz Biotech; Santa Cruz, CA) and anti-actin (Sigma) to ensure proper fractionation. Exogenous ACF was detected using anti-V5 (Invitrogen) antibody, while both ACF populations were observed when probing with an anti C-terminal (CT) ACF antibody [26]. To
evaluate total cellular ACF expression 60 g of total cell lysate was resolved and probed with anti-ACF CT antibody or actin to ensure equal loading.

A1.4.4 Immunological Techniques

Affinity purified, peptide-specific polyclonal antibodies were raised against ACF N-terminal (NT) sequence (NHKSGDGLSGTQKE) and C-terminal (CT) sequence (HTLQTLGIPTEGGD) in rabbits ([14]); Bethyl Laboratories, Inc., Montgomery, TX). ACF CT antibody was conjugated to Sepharose (Bethyl Laboratories, Inc.) and ACF immunoprecipitated as follows. Whole cell extracts were incubated with the antibody-Sepharose bead column overnight at 4 °C. The resin was washed 3X with PBS, 1X with 10 mM phosphate buffer pH 6.8, and eluted using 100 mM glycine pH 2.5. The eluates were neutralized with 1/10th volume 1 M Tris pH 7.5 and precipitated with 5 volumes acetone at –80 °C. Samples were resolved by 10.5% SDS-PAGE, probed with anti-phosphoserine, threonine, or tyrosine antibodies (Zymed Laboratories, San Francisco, CA), stripped and re-probed with anti-ACF NT antibody.

A1.4.5 Comparative Modeling of ACF

RNA Recognition Motifs (RRMs) 1 and 2 of rat ACF64 (residues 54-223) were aligned with RRM 1 and 2 (residues 37-203) of human HuD protein using ClustalX version 1.83 [28]. ACF RRM boundaries were predicted according to alignments with existing RRM crystal structures from HuD [29], polyA binding protein [31], and sex lethal [32]. ACF RRMs 1 and 2 were then comparatively modeled onto the 1.8 Å resolution crystal structure of RRM 1 and 2 of HuD [29] using MODELLER 6v2 [30]. Details of template assembly and utilization were as described in the MODELLER manual. HuD and ACF ribbon diagrams were constructed with PyMol [33]. Model quality was assessed by Ramachandran plots using RAMPAGE [34].
A1.5 Results

A1.5.1 Evaluation of ACF Phosphorylation

Previous studies of in vivo $^{32}$P labeling of rat primary hepatocytes demonstrated that ACF possesses metabolically regulated, high turnover serine phosphorylation sites [26]. However, the possibility of threonine and tyrosine phosphorylation at lower turnover or constitutive sites was not excluded and thus remains a formal possibility. Rat ACF64 protein sequence was analyzed by NetPhosK (http://www.cbs.dtu.dk/services/NetPhosK/), ProSite (http://expasy.org/tools/scanprosite), ProteinScan (http://nci.nih.gov/MPR/ScanProteinForPKCSitesPageaspx), and MotifScan (http://scansite.mit.edu/ motifscan_seq.phtml) algorithms for prediction of potential sites of phosphorylation. The algorithms detected several potential serine, threonine, and tyrosine residues with varying probabilities of being phosphorylated. To examine the types of residues phosphorylated in ACF we initially used an immunological approach.

ACF was immunoprecipitated from extracts of primary hepatocyte cultured in basal insulin using immobilized ACF CT antibody, resolved by SDS-PAGE and probed with anti-phosphoserine, phosphothreonine, and phosphotyrosine antibodies. ACF reacted with anti-phosphoserine and anti-phosphothreonine, but not anti-phosphotyrosine antibodies (Fig. A1.1, right hand column). To confirm that phospho-ACF was responsible for the observed immunoreactivity, western transfers were stripped and re-probed with anti-ACF NT antibody and the images overlaid (Fig. A1.1, left hand column). ACF and anti-phosphoserine/threonine immunoreactivities were superimposed demonstrating that ACF was responsible for the observed anti-phosphoserine/threonine reactivity.
A1.5.2 Protein Kinase Activator Studies

The majority of serine and threonine residues predicted to be phosphorylated fit consensus PKC or PKA target sequences (Table A1.1). Studies were therefore initiated to evaluate the effect of PKC and PKA activation on apoB mRNA editing in rat primary hepatocytes. Since inhibition of PP1 stimulated editing activity [26], we reasoned that activation of appropriate protein kinases would also induce editing. Commercially available activators of PKC (indolactam V, [28]) and PKA (8-cpt-cAMP [29] and forskolin [30]) were selected for evaluation.

As anticipated, control primary hepatocytes edited 60% of apoB mRNA and DMSO treatment did not significantly increase editing (Table A.2). Activation of PKC with 12 µM and 120 µM indolactam V significantly stimulated editing to 92% (P < 0.01). Activation of PKA with 2 µM, 20 µM, or 200 µM 8-cpt-cAMP (EC_{50} 2 µM) had no effect on editing (P > 0.05, Table A.2) suggesting that PKA activation does not alter editing. To exclude the possibility that 8-cpt-cAMP was not active under our experimental conditions, we tested a second PKA activator, forskolin. Treatment with 4 µM, 40 µM, or 120 µM forskolin had also no effect on editing (Table A1.2). Furthermore, additional adenylate cyclase activators, glucagon and isoproterenol also failed to induce editing (data not shown). Taken together, these data suggest that PKC plays a role in editing regulation, while PKA does not.

In order to correlate the observed effects on editing with changes in ACF phosphorylation, ACF was immunoprecipitated from extracts isolated from hepatocytes treated with increasing concentrations of indolactam V (Fig. A.2) in the presence of ^{32}P. ACF phosphorylation was increased in hepatocytes treated with 12 µM and 120 µM indolactam V. Taken together, these data demonstrate that the cellular effects of indolactam V that lead to enhanced editing activity are associated with increased ACF phosphorylation.
A1.5.3 PKC Phosphorylates ACF64

To determine if PKC can phosphorylate ACF, recombinant rat ACF64 was expressed and purified [27] and then incubated with equal units of various PKC isozymes or PKA catalytic subunit. Reactions were resolved by SDS-PAGE, transferred to nitrocellulose, and evaluated by autoradiography. Liver-expressed PKC isozymes (including \( \alpha \), \( \beta_{II} \), \( \delta \), \( \varepsilon \), and \( \zeta \); reviewed in [31]) phosphorylated ACF64 in vitro (Fig. A.3, upper panel). To verify that the observed \(^{32}\)P incorporation was specific to ACF, western transfers were probed with anti-HA and the images overlaid (Fig. A1.3, lower panel). PKC isozymes \( \alpha \), \( \beta_{II} \) and \( \zeta \) demonstrated the highest specific-activity with ACF64. No \(^{32}\)P labeling was detected in the absence of protein kinase demonstrating that the signal is PKC-dependent. These data are consistent with the PKC activator studies (Table A1.2) and suggest that liver-specific PKC isozymes are capable of phosphorylating ACF64.

Given that ACF was predicted to contain potential PKA phosphorylation sites, the ability of PKA catalytic subunit to phosphorylate ACF was also evaluated. PKA catalytic subunit was unable to phosphorylate ACF64 in vitro using the same number of enzyme units as PKC. PKA-dependent phosphorylation of Kemptide peptide (Promega) demonstrated that the enzyme was active under our assay conditions (data not shown). However, 10-times greater number of units of PKA resulted in \(^{32}\)P incorporation into ACF (data not shown) and may represent non-selective phosphorylation due to excess enzyme. Although our data are consistent with phosphorylation of ACF by PKC, we cannot rule out PKA phosphorylation since in vivo phosphorylation of ACF may require other regulatory factors (e.g., PKA may require prior posttranslational modification of ACF by PKC).

A1.5.4 Site-directed Mutagenesis and ApoB mRNA Editing

To identify candidate ACF phosphorylation sites, ACF cDNA was selectively mutated to generate alanine or aspartic acid at predicted PKC phosphorylation sites (Table A1.1). Sites were selected for mutagenesis based on the following three criteria: (1) ACF contains
serine and threonine phosphorylation sites (Fig. A1.1 and [26]) predicted to be phosphorylated by PKC (Table A1.1), (2) the amino terminal 380 amino acids of ACF contain all domains necessary to complement editing including; APOBEC-1 interaction [32], RNA-binding, [16], and a nuclear localization sequence [33], and (3) only sites conserved in both rat and human ACF64 were considered for our studies (see sequence alignment, Fig. A1.6). We also selected sites for mutagenesis with low and medium probabilities of being phosphorylated (Table A1.1) as there are examples of empirically determined and validated sites of phosphorylation that were not initially predicted by consensus sequence algorithms alignments [34].

Eight of the predicted sites of ACF phosphorylation were separately mutated to either alanine or aspartic acid (Figs. A1.4 & A1.5). ApoB mRNA editing activity was evaluated in stable McArdle cell lines that expressed similar levels of each phosphorylation site mutants to ensure that modulation of editing activity was not due to large differences in ACF expression. Expression of each mutant was assessed by western blot of whole cell extracts using an anti-V5 epitope antibody (Fig. A1.4a). Cell lines were treated without or with 0.9% ethanol for 4 hours to test the ability of each mutant in ACF to support editing activity under basal and stimulated conditions.

Editing in vector only transfected cells was indistinguishable from control, parental McArdle cells without ethanol treatment (data not shown). ApoB mRNA editing efficiency in wild type McArdle cells was 14% (+/- 4.4) (Fig. A1.5, shown as a vertical line for comparison to other treatments). As anticipated [25], the addition of ethanol significantly stimulated editing by 50% (P < 0.05) in wild type McArdle cells (Fig. A1.5). A 13% (+/- 4.4%) increase in editing was observed in McArdle cells stably transfected with wild type ACF64. Editing activity in ACF64 transfected McArdle cells retained the ability to respond to ethanol stimulation (45% increase +/- 3.5%, P<0.05). These data demonstrated that over expression of ACF does not adversely affect editing activity and that ACF64 transfected McArdle cells responded robustly to ethanol stimulation.
There is a vast literature on protein phosphorylation supporting the use of mutagenesis to identify sites of phosphorylation. In this approach, mimicking a phosphorylation site by mutation to aspartic acid should induce editing in control cells, whereas ablation of a phosphorylation site by mutation to alanine should inhibit ethanol-induced editing. Cells transfected with an ACF64 S154 mutated to alanine (S154A) had basal editing values comparable to control untransfected cells. McArdle cells expressing ACF64 S368 mutated to alanine (S368A) demonstrated editing values only 8% (+/- 5.0) greater than control. In contrast, editing activity in cells transfected with S154A or S368A had a severely blunted the response to ethanol treatment with editing values only 10% (+/- 7.1) and 22% (+/- 4.4) greater than control untreated McArdle cells, respectively (Fig. A1.5). These data suggested these mutants had a dominant negative effect on the ability of the endogenous McArdle cell machinery to respond regulated activity in response to ethanol treatment.

Mutation of ACF64 S154 to aspartic acid (S154D) stimulated editing 48% greater than control in the absence of ethanol (P ≤ 0.05). Mutation of ACF64 S368 to aspartic acid (S368D) also enhanced editing by 48% greater than control (P ≤ 0.05). Treatment of McArdle cells expressing either S154D or S368D with ethanol resulted in editing 43% (+/- 11.0) and 33% (+/- 4.8) greater than control, respectively (Fig. A1.5). Taken together, the data suggested that the phenotype of Ala and Asp mutation at S154 and S368 was consistent with regulation of editing through these sites by phosphorylation.

Several other serine and threonine residues were mutated to alanine or aspartic acid, but failed to behave in a manner consistent with regulated sites of phosphorylation. To facilitate discussion, the remaining mutants have been categorized as: (i) those for which mutation to alanine was inhibitory (e.g., S171), (ii) those for which mutation to aspartic acid inhibited editing (e.g., S132), (iii) those for which ethanol treatment rescued editing of alanine and aspartic acid mutations (e.g., S188 and T49/T50), and (iv) those whose editing activity never reached wild type McArdle cell editing levels (e.g., T176).
In the first case, cells transfected with S171 mutated to alanine (S171A) could not maintain editing levels comparable to cells transfected with wtACF. While these data suggest that this site is not essential to ACF64 function and S171 phosphorylation is not critical for ethanol stimulation of editing activity mutation of S171 to aspartic acid (S171D) stimulated editing compared to control cells and cells transfected with wtACF. The addition of ethanol further stimulated editing in cells expressing S171D to 40% greater than control (P < 0.05). The data suggested that the ability of S171D to stimulate editing activity is likely due to the change in local charge that lies close to S154 (see structural model, Fig. A.7).

S132 was also in its own category where mutation to alanine (S132A) displayed a phenotype similar to transfection with wild type ACF (stimulated editing by 17% (+/- 6.4) and 29% (+/- 4.3) in control and ethanol treated McArdle cells, respectively (Fig. A1.5). However, mutation to aspartic acid (S132D) did not stimulate editing activity, on its own, but cells retained the ability to respond to ethanol stimulation. The data suggest a preference for nonpolar or uncharged amino acids at position 132 and that ethanol treatment can activate ADF (presumably through phosphorylation at S154 and S368 regardless of the mutations at S132 studied here).

Mutation of S188 or T49/T50 to either alanine (S188A and T49A/T50A) or aspartic acid (S188D and T49D/T50D) markedly inhibited editing activity (Fig. A1.5). Treatment with ethanol increased editing activity to varying levels greater than control cell levels, but not to the extent seen in ethanol-treated parental McArdle cells. The data suggest that S188 and T49/T50 are inherently important for ACF64 structure and function and that it is unlikely that they are phosphorylation sites. However, consistent with phosphorylation of S154 and S368, cell lines expressing S188D or T49/T50D mutants resulted in increased editing. Similar results were obtained when ACF64 alanine and aspartic acid substitution mutagenesis was performed at S253, S274, T160 and T316 (data not shown).

Mutation of T176 to either alanine (T176A) or aspartic acid (T176D) resulted in the most marked dominant negative editing phenotype (Fig. A1.5). With this mutation ethanol
treatment was only able to marginally restore editing to the levels, suggesting that this site is essential as a threonine. Similar results were obtained when ACF64 alanine and aspartic acid substitutions were evaluated at S241, S242, S243, S377, S308, T14 and T146 (data not shown).

Our findings suggest that mutagenesis of phosphorylation sites had an effect on ACF’s ability to complement APOBEC-1 and form functional editosomes. To rule out the possibility that ectopically expressed ACF mutants altered editing by changing the expression of the endogenous ACF we evaluated the abundance of endogenous ACF in the presence of our mutants. Irrespective of ectopic ACF64 expression, endogenous ACF abundance (normalized to cellular actin) was not changed (Fig. A1.4). The subcellular distribution of endogenous ACF also was not significantly changed through ectopic expression of wtACF or mutant ACF (data not shown). These findings demonstrate that the observed changes in editing activity are due to the properties of individual ACF mutants.

While not exhaustive, the mutagenesis data suggested that phosphorylation of 154, 171 and 368 would stimulate editing activity, whereas phosphorylation at other serines or threonines would markedly inhibited ACF function and editing activity.

**A1.5.5 ACF Structure-based Prediction of Phosphorylation Site**

**Accessibility**

The positions of ACF phosphorylation-site mutants were evaluated based on the comparative model to HuD. This model served to demonstrate that the predicted sites of ACF phosphorylation might be in accessible regions of the protein. Alignment of rat [14] and human [3] ACF64 protein sequences demonstrated 93.5% identity (Fig. A1.6). All predicted sites of ACF phosphorylation selected for mutagenesis were conserved between both species (Fig. A1.6). HuD and ACF both contain three RRMs and regulate mRNA stability through binding AU-rich 3’-UTRs [42]. Alignment of the first two RRMs of ACF (residues 55 – 223) and HuD (37 – 203) maintained RNP consensus alignment and yielded a primary
sequence identity of 23%, with 57% homology. A structural model of ACF was generated from this alignment using the 1.8 Å structure of HuD RRMs 1 and 2 bound to a short AU-rich RNA oligomer (PDB ID 1FXL) [29]. These RRMs exhibited the canonical four-stranded antiparallel β-sheet flanked by two α-helices (Figs. A1.6 & A1.7). Ninety-one percent of the modeled ACF residues fell into favored regions of the Ramachandran plot, with an additional 5.4% in allowed regions. Compared to the equivalent residues in HuD, S132 of ACF is conserved while S154, T176, and S188 are substituted with glutamine, serine and lysine, respectively. ACF residues T49/T50 and S368 fall outside the boundaries of the model, and therefore could not be evaluated in this context.

In HuD, single-stranded RNA contacts the β-strands and loops of RRMs 1 and 2 (Fig. A1.7), whereas the flanking α-helices are more solvent exposed and therefore accessible to protein-protein interactions (Fig. A1.7 and [35, 36]). It is important to note that while the position and orientation of each residue within the modeled ACF RRMs can be evaluated, the relative spatial orientations of each RRM cannot be deduced without empirical structural information for ACF. Hence, although multiple tandem RRM structures exist and the individual RRM fold is highly conserved, the conformation of the inter-RRM linker varies greatly among structures [37, 38]. Nonetheless, since comparative modeling of a single RRM does not rely upon knowledge of the inter-domain linker, there is confidence that S154 resides in a solvent exposed region of the second helix of RRM 2 (Fig. A1.7). As such, the hydroxyl side chain of this residue should be accessible to protein kinases and phosphatases. Likewise, T176 and S171 are located within the loop joining β-strands 2 and 3 of RRM 1 (Fig. A1.7). In contrast, S132 is located in the linker between RRMs 1 and 2, which suggests solvent accessibility, but precludes rigorous modeling (Fig. A1.7).

A1.6 Discussion
ApoB mRNA editing requires the coordinated assembly of a multi-protein editosome [14] that governs substrate and site specificity [39], the proportion of substrate edited [10] and possibly export of the edited product from the nucleus. Thus, regulation of editosome assembly and trafficking of editing complexes is necessary for efficient editing. ACF is an obligate component of the editing machinery and is the site-specific RNA-binding protein [39, 40] that docks APOBEC-1 to cytidine 6666. Recently, phosphorylation of ACF [26] was identified as a mechanism whereby apoB mRNA editing can be regulated. Hyper-phosphorylation of serine residues, in active nuclear 27S editosomes, was implicated in ethanol-stimulated editing [26].

To better understand the role of ACF phosphorylation in basal and metabolically stimulated editing, the sites of phosphorylation and the enzymes responsible were characterized. The expression of all mutants was validated to be at similar levels through the selection of McArdle cell lines stably expressing each variant. While it is likely that overexpressed wtACF or mutant ACF competed with endogenous ACF leading to the observed changes in editing activity, the overall abundance and subcellular distribution of endogenous ACF was unchanged in response to ectopic expression of phosphorylation site mutants. We have demonstrated in this report that mutation of S154 and S368 to alanine inhibited the ability of McArdle cells to regulate (increase) editing activity in response to ethanol. Mutation of these residues to aspartic acid fully stimulated editing activity even in the absence of ethanol treatment.

S154 and S368 were predicted as phosphorylation sites, potentially targeted by either PKA or PKC. Protein kinase activator studies demonstrated that only PKC activators stimulated editing activity and enhanced ACF phosphorylation in hepatocytes. The role of PKC in ACF phosphorylation was further supported by in vitro protein kinase assays in which liver expressed PKC isoforms phosphorylated recombinant ACF, whereas PKA had to be used at 10-fold higher concentrations to affect ACF phosphorylation. Thus our data support
the role of PKC phosphorylation of ACF64, particularly PKC isoforms α, βII and ζ, in ethanol regulated hepatic editing activity.

Immunological data indicated that rat liver ACF possesses both serine and threonine phosphorylation sites although only serine phosphorylation has been observed by radiolabeling during metabolic stimulation [26]. We believe this discrepancy to be due to inherent differences in the experimental systems used. Detection of phosphorylated residues by antibodies does not require phosphate turnover (unlike metabolic labeling with $^{32}$P). These data suggest that threonine phosphorylation is subject to slower turnover than serine phosphorylation in rat primary hepatocytes and/or that phosphothreonine is significantly less abundant. Our mutagenesis studies were designed to identify sites of phosphorylation that affect ACF64 function during metabolically stimulated editing. Given that threonine residues were not identified in this mutagenic screen, our collective data only support the role of regulated ACF64 serine phosphorylation in the metabolically modulation of apoB mRNA editing. Phosphorylated threonines are likely constitutive or subject to slow phosphate turnover. ACF threonine phosphorylation however could play an inhibitory role, by maintaining editing at low levels under basal metabolic conditions. Consistent with this possibility are data showing that inhibition of protein synthesis stimulated editing in primary hepatocytes [25]. This could be due to reduced protein synthesis of a negative regulator of editing, such as a protein kinase active on ACF threonines. If constitutively expressed, a protein kinase with activity on ACF T49/T50 and T176 for example could inhibit editing activity. The phenotype of S145D and S368D suggested that ethanol stimulated phosphorylation of S154 and/or S368 will override the inhibitory effects of threonine phosphorylation and activate editing activity without necessarily requiring dephosphorylation of phosphothreonine sites. We therefore predict that at steady state hepatic ACF64 will be phosphorylated at different sites and to varying degrees depending on the metabolic demand. ACF containing multiple phosphorylations was predicted by two-dimensional gel analysis [26].
Hepatic ethanol-stimulated editing [41, 42] coincides with hyper-phosphorylation of ACF [26]. Our results demonstrate that liver expressed PKC isozymes α, βII and ζ have the greatest ability to phosphorylate recombinant ACF. PKC θ over expression has been reported to modulate editing in rat hepatoma cells, whereas PKC isozymes α, βII, ε and ζ were unable to modulate editing [43]. However, the authors did not provide direct evidence that individual isoforms phosphorylated proteins involved in editing. Moreover, PKC θ is not a physiologically relevant isozyme since it is not expressed in liver [44]. Further, the relative expression level of each PKC isoform was not indicated and therefore, non-selective phosphorylation of proteins may have been induced experimentally by overexpression of protein kinase. In this regard, although PKA is predicted to phosphorylate ACF and is regulated by ethanol [45], PKA had very low activity in vitro on recombinant ACF64 and PKA-specific activators did not modulate editing in primary hepatocytes (Table A1.2). Neither direct activation of PKA with cAMP analogs nor indirect activation via adenylate cyclase stimulated editing. Conversely, activation of PKC with indolactam V significantly enhanced apoB editing (Table A1.2) and PKC isozymes phosphorylated purified ACF in vitro. Our data is in agreement with results previously published demonstrating that activation of PKC induces editing in intestinally derived Caco-2 cells [43].

Although phosphorylated ACF has only been recovered in the nucleus of hepatocytes, the cellular domain in which ACF is phosphorylated remains to be determined. In vivo, PKC substrate specificity is believed to be imparted by the subcellular distribution of the enzyme and substrate [46]. Therefore, PKC interaction with and phosphorylation of ACF in the perinuclear environment, at some point during nuclear import or at some point in time subsequent to nuclear import remain formal possibilities.

ACF is a single polypeptide that comprises three tandem RRMs [6]. The protein itself has been implicated in apoB mRNA-binding, APOBEC-1 binding and protection of edited apoB mRNA from NMD [6, 17]. In general, each RRM is composed of four antiparallel β-
strands flanked by two α-helices (reviewed in [47]). RNA recognition occurs through the β-strands and their connecting loops. Other complex embellishments to the RRM fold have been documented that allow protein-protein interactions at the α-helices [36, 48, 49], or β-strands [50, 51], although the nature of the ACF-APOBEC-1 interaction is unknown at present. In HuD, ssRNA makes contact primarily to the β--strands of RRMs 1 and 2 (Fig. A1.7, red residues), whereas the outward facing α-helices are exposed to solvent [35]. Based on ACF sequence homology to HuD and the conserved structure of the RRM fold, S154 of ACF is predicted to localize within the second helix of RRM 2, where it would be accessible to protein kinases, but would not be expected to contact RNA or affect RNA-binding when phosphorylated. As such, it is plausible that this residue could be involved in protein-protein interactions (Figs. A1.6 & A1.7). The three-dimensional model for the position and orientation of S154 within RRM 2 is consistent with previous studies demonstrating that treatment of extracts with calf-intestinal alkaline phosphatase reduced co-immunopurification of APOBEC-1 with ACF, but did not affect ACF’s ability to bind to apoB RNA [26]. Regulation of protein-protein interactions by changes in the phosphorylation state of one or more proteins has been well documented [52, 53]. Additionally, phosphorylation may alter tertiary fold promoting protein-protein interactions as is the case with CREB and CREB binding protein [54].

Another residue of interest in ACF is S171, which is a glutamine in HuD. This residue is located in the loop between β–strands 2 and 3 of RRM 2. Loop 3 is the most genetically divergent region among RNP proteins and is important for substrate specificity [47, 55]. Q171 of HuD is 4 Å from the non-bridging phosphate oxygen of U5, suggesting that the introduction of negative charge in this area could affect substrate specificity irrespective of the phosphorylation state. Editing activities for the S171D mutant were 15% and 28% greater than untreated McArdle cells in the absence and presence of ethanol, respectively. The introduction of negative charge by aspartic acid mutation in S171D may have contributed to the slight elevation in editing activity, the data suggest that it is unlikely that S171 is an ethanol regulated phosphorylation site.
Our mutagenesis analyses predicted that some serine/threonine residues in and of themselves are important for ACF structure and also editing. The introduction or ablation of negative charge at these residues could affect protein folding, protein-protein interactions, protein-RNA interaction and thus function independent of protein phosphorylation. In this case, mutagenesis of these sites to alanine and aspartic acid would disrupt editing. Case in point, S132 is conserved in rat and human ACF as well as in HuD and is located in between RRM1s 1 and 2. The data suggested that mutation to alanine at this site was tolerated since editing was comparable to wild type ACF64. However, mutation to aspartic acid was detrimental. Although serine residues are polar, the R group is uncharged at pH 7.0. Mutation to alanine, a nonpolar amino acid, with only a methyl moiety as the R group would be less likely to interfere with protein folding. To the contrary, substitution with negatively charged aspartic acid could alter protein folding. Therefore, S132 is unlikely to be a phosphorylation site.

Limited structural information precludes modeling outside of ACF RRM1s 1 and 2. Consequently, mutations amino terminal to amino acid 55 (T49/T50) and those carboxy terminal to amino acid 223 (i.e., S368) cannot be evaluated in the context of this model. However, S368 is located within ACF’s nuclear localization sequence (amino acids 360-401) [33]. ACF is a nucleocytoplasmic shuttling protein related to hnRNP proteins [33] and its subcellular distribution is sensitive to metabolic perturbations [33]. Significantly, phospho-ACF accumulates in the nucleus of hepatocytes treated with ethanol or insulin [14]. The S368 to alanine mutation was refractory to ethanol, whereas the aspartic acid mutation stimulated editing in untreated McArdle cells to levels comparable to cells treated with ethanol. These data suggest that phosphorylation of S368 may modulate ACF nucleocytoplasmic shuttling. In fact, several examples exist in the literature where phosphorylation mediated nuclear localization ([53, 56, 57] reviewed in [58]). The coordinated action of PP1 [26] and PKC could regulate that subcellular distribution of ACF, and therefore the proportion of apob mRNA that is edited.
In conclusion, the data suggest a model of editing regulation in McArdle cells whereby ACF is phosphorylated minimally on serine and threonine residues in basal McArdle cells. Metabolic stimulation of editing by ethanol may be accompanied by phosphorylation of S154 and S368 by PKC. Protein phosphatase 1 activity has been implicated in removing phosphate from metabolically regulated sites of ACF phosphorylation, enhancing export of ACF from the nucleus [26] and reducing editing activity. Our findings suggest that phosphorylation of ACF by PKC in response to ethanol is part of the mechanism for nuclear import of ACF and activation of editing activity. In this regard, PKC and PP1 activities on ACF are predicted to act in concert to modulate the overall phosphorylation status of ACF, regulating its protein interactions and subcellular distribution.

A1.7 Acknowledgements

We thank Dr. N. Ballatori for assistance in the preparation of rat primary hepatocytes and Jenny M.L. Smith for the preparation of Tables and Figures. This work was supported in part by a Public Health Services Grant DK43739 awarded to HCS and RR15934 and GM63162 awarded to JEW. DML was supported through a Toxicology Training Grant 5T32 ES07026, a Grant in Aid from the American Liver Foundation and Sigma Xi. CM was supported by a Training in Cellular, Biochemical and Molecular Sciences Training Grant 5T32 GM068411.
References


Table A1.1 Sites selected for site-directed mutagenesis

<table>
<thead>
<tr>
<th>Site</th>
<th>Sequence</th>
<th>Average Prediction Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>T49</td>
<td>GGPPPGWD-T-TPPERGCE</td>
<td>92</td>
</tr>
<tr>
<td>T50</td>
<td>GPPPGWDT-T-TPPERGCE</td>
<td>97</td>
</tr>
<tr>
<td>S132</td>
<td>GRLLGVCA-S-VDNCRLFV</td>
<td>36</td>
</tr>
<tr>
<td>S154</td>
<td>TKKREEIL-S-EMKKVTEG</td>
<td>52</td>
</tr>
<tr>
<td>S171</td>
<td>VVDVIVYP-S-AADKTKNR</td>
<td>66</td>
</tr>
<tr>
<td>T176</td>
<td>VVYSAADK-T-KNRGFAFV</td>
<td>7</td>
</tr>
<tr>
<td>S188</td>
<td>GFAFVETE-S-HRAAAMAR</td>
<td>39</td>
</tr>
<tr>
<td>S368</td>
<td>FPATKGHLS-NRALIRTP</td>
<td>7</td>
</tr>
</tbody>
</table>

The sequence was analyzed using ProteinScan (http://nci.nih.gov/MPR/ScanProteinForPKCSitesPageaspx) algorithm to predict potential PKC phosphorylation sites. As an example of the range of site scoring, the average of PKC \( \delta \) and \( \zeta \) percentiles are shown. The specific serine (S) or threonine (T) predicted to be phosphorylated are shown in bold.
### Table 2

<table>
<thead>
<tr>
<th>Indolactam V</th>
<th>Mean</th>
<th>STD</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary hepatocytes</td>
<td>63.3</td>
<td>3.9</td>
<td>3</td>
</tr>
<tr>
<td>DMSO</td>
<td>62.2</td>
<td>4.8</td>
<td>3</td>
</tr>
<tr>
<td>12 µM</td>
<td>*92.5</td>
<td>0.8</td>
<td>3</td>
</tr>
<tr>
<td>120 µM</td>
<td>*91.6</td>
<td>1.1</td>
<td>3</td>
</tr>
</tbody>
</table>

PKC EC$_{50}$ = µM

<table>
<thead>
<tr>
<th>8-cpt cAMP</th>
<th>Mean</th>
<th>STD</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary hepatocytes</td>
<td>58.8</td>
<td>4.8</td>
<td>4</td>
</tr>
<tr>
<td>DMSO</td>
<td>58.2</td>
<td>4.6</td>
<td>4</td>
</tr>
<tr>
<td>2 µM</td>
<td>61.2</td>
<td>4.3</td>
<td>4</td>
</tr>
<tr>
<td>20 µM</td>
<td>59.3</td>
<td>(-)</td>
<td>2</td>
</tr>
<tr>
<td>200 µM</td>
<td>59.6</td>
<td>3.4</td>
<td>4</td>
</tr>
</tbody>
</table>

PKA EC$_{50}$ = 2 µM

<table>
<thead>
<tr>
<th>Forskolin</th>
<th>Mean</th>
<th>STD</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary hepatocytes</td>
<td>57.6</td>
<td>3.4</td>
<td>7</td>
</tr>
<tr>
<td>DMSO</td>
<td>61.8</td>
<td>4.1</td>
<td>6</td>
</tr>
<tr>
<td>4 µM</td>
<td>70.6</td>
<td>5.2</td>
<td>6</td>
</tr>
<tr>
<td>40 µM</td>
<td>68.5</td>
<td>7.5</td>
<td>4</td>
</tr>
<tr>
<td>120 µM</td>
<td>69.0</td>
<td>0.8</td>
<td>4</td>
</tr>
</tbody>
</table>

AC EC$_{50}$ = 4 µM

---

**Table A.2**  **Activation of PKC modulates apoB mRNA editing**

Hepatocytes were treated with the indicated concentrations of protein kinase activators for 6 hours. ApoB mRNA editing activity was determined using apoB-specific RT-PCR and poisoned-primer extension analysis [14]. The mean, standard deviation (STD) and number of replicates (n) are listed below for each concentration tested. *Statistical significance was determined to be P < 0.05 by unpaired t-test relative to DMSO control when n ≥ 3.
Figure A1.1  Rat hepatic ACF immunoreactivity with anti-phosphoserine and threonine antibodies  ACF was immunoprecipitated from liver nuclear extracts with anti-ACF CT antibody conjugated to Sepharose and western blotted with anti-phosphoserine, threonine, or tyrosine antibodies (right column). The blots were stripped and re-probed with anti-ACF NT antibody. ACF, position of ACF; anti-pSer, anti-pThr, anti-pTyr denotes position of phosphoserine, threonine, or tyrosine immunoreactivity.
Figure A1.2  ACF phosphorylation is modulated by protein kinase C activation

*In vivo* ACF $^{32}$P incorporation was determined by Phosphorimager scanning of ACF immunoprecipitates prepared from rat hepatocytes treated with 12 or 120 μM indolactam v (Indo V). ACF specific activity (relative to control hepatocytes) was calculated as the ACF $^{32}$P density (Phosphorimager) divided by the recovery of ACF determined from densitometric scanning of ACF western blots (Image J). ACF immunopurified from control hepatocytes was arbitrarily assigned a value of 1 ($n = 3$). Exp. 1 and Exp. 2 denote independent experiments 1 and 2.
Figure A1.3  PKC phosphorylates ACF64 in vitro

Recombinant ACF64 was reacted with PKC isozymes (α, βII, δ, ε and ζ) or PKA, transferred to nitrocellulose and analyzed by autoradiography. Western transfers were subsequently probed with anti-HA antibody. (-) denotes kinase-free control reaction only containing buffer and ACF.
Rat ACF64 was previously cloned [14] into modified pcDNAIII vector containing 3' V5 epitope tag. Equivalent number of cells and amount of protein from McArdle cell lines overexpressing V5-tagged ACF64 constructs (as labeled) were resolved by 10.5% SDS-PAGE, transferred to nitrocellulose and probed with anti-V5 antibody. (B) Abundance endogenous ACF is unaffected by ectopic expression of ACF64 site-directed mutants. Whole cell extracts were generated from McArdle cells expressing ACF mutants. Equal micrograms of protein were resolved by SDS-PAGE, transferred to nitrocellulose, and blotted with anti-ACF CT antibody (upper panel), anti-actin (lower panel). Data shown is representative of 3 independent experiments.
Figure A1.5 Mutagenesis of predicted ACF64 phosphorylation sites modulates apoB mRNA editing in McArdle cells McArdle cells stably transfected with various ACF site-directed mutants were treated with 0.9% EtOH or as control for 4 hours. RNA was isolated, apoB RNA was amplified and evaluated for editing using the poisoned-primer extension assay. The central vertical line represents the level of wild type McArdle editing. Percent change from wild type (wt) editing was calculated by \[ \frac{(\text{ACF mutant} \% \text{ editing} - \text{wt ACF} \% \text{ editing})}{(\text{wt ACF} \% \text{ editing})} \times 100 \] and is shown as histograms to either side of the control value vertical line. *Statistical significance was defined as \( P < 0.05 \) compared to control wtACF expressing cells when \( n \geq 3 \). Error bars represent the calculated standard deviation (\( n \geq 3 \)).
Figure A1.6  Comparative modeling of ACF  Rat and human ACF sequence alignment and secondary structure prediction. The human ACF amino acid sequence is shown above that of rat. Black residue pairs are conserved; unique regions between ACF variants are blue and unique HuD residues are green. Red letters indicate contacts to HuD RNA. Arrows mark mutagenized residues.
Figure A1.7  Predicted tertiary structure of ACF and potential phosphorylation sites  A ribbon diagram of the HuD structure that served as a template for the ACF homology model is shown at left bound to substrate RNA. The resulting ACF model is shown at right. Labeled residues were targeted as potential sites of phosphorylation and are shown in ball and stick representation. Figure was made using PyMol [59].
Appendix 2:

Identification Codes for Ribozyme Coordinates and Structure Factors Available from the Protein Data Bank (PDB) or Nucleic Acid Data Bank (NDB)
<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2NPZ</td>
<td>Crystal structure of hinged hairpin ribozyme incorporating 3-atom synthetic propyl linker.</td>
</tr>
<tr>
<td>2NPY</td>
<td>Crystal structure of a hinged hairpin ribozyme incorporating 9-atom linker and 2'-deoxy 2'-amino U at A-1.</td>
</tr>
<tr>
<td>2OUE</td>
<td>Redeposited crystal structure of a junctionless all-RNA hairpin ribozyme at 2.05 angstroms resolution with anti-anti configuration for the U-U mismatch. (old PDB ID: 1ZFR)</td>
</tr>
<tr>
<td>3B5A</td>
<td>Crystal structure of a minimal, hinged hairpin ribozyme incorporating A38G mutation with a 2'OMe modification at the active site.</td>
</tr>
<tr>
<td>3B58</td>
<td>Minimal, hinged hairpin ribozyme incorporating A38G mutation and a 2',5'-phosphodiester linkage at the active site.</td>
</tr>
<tr>
<td>3B5S</td>
<td>Minimal, hinged hairpin ribozyme incorporating A38DAP mutation and 2'-O-methyl modification at the active site.</td>
</tr>
<tr>
<td>3B5F</td>
<td>Crystal structure of a minimal, hinged hairpin ribozyme incorporating the Ade38DAP mutation and a 2',5'-phosphodiester linkage at the active site.</td>
</tr>
<tr>
<td>3BBI</td>
<td>Minimal, hinged hairpin ribozyme incorporating A382AP and A-1 2'O-Me modifications near active site.</td>
</tr>
<tr>
<td>3B91</td>
<td>Minimal, hinged hairpin ribozyme incorporating Ade382AP and 2',5'-phosphodiester linkage mutations at the active site.</td>
</tr>
<tr>
<td>3BBM</td>
<td>Minimal, hinged hairpin ribozyme incorporating A38C and 2'O-Me modification at active site.</td>
</tr>
<tr>
<td>3BBK</td>
<td>Minimal, hinged hairpin ribozyme incorporating A38C and 2'5'-phosphodiester linkage within active site.</td>
</tr>
</tbody>
</table>

**Table A2.1 Structures Deposited to the Protein Data Bank**  Coordinates and structure factors are available for downloading at http://www.rcsb.org/pdb/home/home.do.