Structural Studies of the Deaminase Domain of the Human HIV-1 Restriction Protein APOBEC3G

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Kuan-Ming Chen

IN PARTIAL FULFILLMENT OF THE REQUIRMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Hiroshi Matsuo, Adviser

August 2010

ACKNOWLEDGEMENTS

When I remind my seven years life at the University of Minnesota, a lot of people come to my mind. First, I would like to thank my thesis adviser, Dr. Hiroshi Matsuo, for his support, encouragement and patience during my graduate studies. I always remember when we got the first HSQC spectrum of A3G-198-384-2K, he was so happy to pat on my head. From him, I not only learned how to design experiments, analyze the data but also the attitude to difficulties with the experiments; when facing a problem, instead of giving up easily, he always tries thousand of methods to overcome the problems. I am so glad to have such a great adviser.

I would also like to thank members in my lab. Dr. Matthew Devany taught me cell culture and how to purify proteins; Dr. Yongjian Lu provided mutant plasmids; Dr. Elena Harjes worked with me on the APOBEC3G project and helped with structural calculations. Dr. Aya Kitamura helped to study the APOBEC3G protein at the beginning. I would also like to thank Dr. Takahide Kono and Jingying Zhang for providing joy and maintaining the clean lab. Without them, the APOBEC3G project wouldn't have progressed so soon.

I would like to extend my gratitude to our collaborator, Dr. Reuben Harris. He provided many helpful ideas for our project and taught me how to make beautiful figures for publication. His lab members, Dr. Keisuke Shindo helped us with protein and cell assays; Natalia Martemyanova and Phillip J Gross provided mutation data.

During my studies here, I am very lucky to have some great scientists around me who like to discuss my project and provide suggestions, they are Dr. Naixia Zhang, Dr. Xiang Chen, Dr. Yang Kang, Sarmistha Chakrabarty and Dr. Issam El Ghazi. Especially, I would like to thank Dr. Melanie Rogers, she gave me feedbacks on my thesis and helped me during my preparation for my defense.

Finally, I am so grateful for the support and love from my father, my mother, my brother and my sister during my 7 years in the US. I also like to thank the encouragement and morning calls from my girlfriend Min Lu during my thesis writing and preparation for my defense. And most of all, I would like to thank God for his guidance for every step in my life.

ABSTRACT

According to a report from the UNAIDS on the global AIDS epidemic in 2009, about 33.4 million people were living with the HIV virus, with a 2.7 million increase each year. HIV can lead to AIDS, which is a set of symptoms and infections resulting from the damage to the human immune system. However, due to the rapid mutability and productivity of the HIV, identifying treatments and therapeutic intervention remains challenging and has seen limited progress.

In the past year, one novel innate defense against HIV infection was discovered, in which the human protein APOBEC3G (A3G) plays an important role. A3G was identified as a single-strand DNA deaminase that potently inhibited the replication of the HIV-1ΔVif virus. It produces the nonfunctional provirus by deaminating the cytosines to uracils on the minus-strand viral cDNA. Consequently, A3G can genetically inactivate HIV and recent studies have demonstrated that this activity is as potent as any current anti-retroviral drug. However, the exact model of this mechanism at atomic level has not yet been elucidated due to the low solubility of A3G which presents an obstacle for biochemical and structural studies. The research in this thesis took a structural approach to screen for a catalytically active and more soluble C-terminal deaminase of APOBEC3G (A3G-ctd) derivative, determine the NMR solution structure of this variant (A3G-2K3A) and characterize its interaction with single-strand DNA.

Due to intrinsically low solubility of A3G-ctd, a strategy to design more soluble derivatives of the catalytic domain was performed prior to NMR structure determination. Two key methods: a solubility test and a *E. coli*-based mutation assay were used to test the solubility and catalytic activity of APOBEC3G variants. Deletion mutant analyses of

APOBEC3G found the minimal catalytic region consisted of amino acids 198-384 (A3G198-384). Various alanine and lysine substitution variants based on this fragment were constructed and examined to screen for improved solubility and enhanced activity. One variant A3G198-384-2K3A (L234K-C243A-F310K-C321A-C356A) showed a significant improvement in both assays, and was purified as a monomer.

The three-dimensional structure of A3G198-384-2K3A was then determined by triple resonance NMR spectroscopy. It consists of five β-strands that form a hydrophobic platform surrounded by five α-helices. Summarizing the DNA titration data, *E. coli*-based catalytic activity, conserved residues and computational modeling, the DNA binding mechanism of A3G was proposed in which a canyon formed by positively charged residues guides single-strand DNA binding and positions the target cytidine for deamination. Subsequently, a longer catalytic domain, A3G191-384-2K3A, was found to have higher activity than that of the A3G198-384-2K3A derivative. The longer domain has an additional α1-helix (residues 201–206) that was not observed in the shorter variant and part of the last α-helix (residues 191-194) of the N-terminal domain. The truncated model of the N-terminal domain was generated from the C-terminal NMR structures based on the sequence homology. Finally, a novel full-length A3G model was constructed by physically overlapping the α-helix (residues 191-194) of the N-terminal domain model and the C-terminal domain structure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

The human APOBEC3G Protein

1.1 APOBEC family

The human APOBEC3G protein belongs to a family of polynucleotide cytidine deaminases which are involved in diverse enzymatic reactions [1]. This family was named after the first family member, identified APOBEC1 (apolipoprotein B mRNA-editing enzyme catalytic polypeptide 1). In humans, the APOBEC family currently consists of 11 members, APOBEC1, AID, APOBEC2, APOBEC3A-3H and APOBEC4 (Figure 1.1) [2]. Originally, APOBEC3D and APOBEC3E were thought to be two different proteins, however, it was recently discovered that these two proteins form the N- and C- terminal deaminase domain of a single protein (APOBEC3DE) [3]. All members of this protein family are defined by at least one highly conserved zinc-binding catalytic domain with the consensus amino-acid sequence His-X-Glu-X_{23-28}-Pro-Cys-X_{2-4}-Cys (where X stands for any amino acid) (Figure 1.1). APOBEC3B, APOBEC3DE, APOBEC3F and APOBEC3G all have two deaminase domains, therefore, their molecular weights are roughly twice as large as the rest of family members. Interestingly, these four members can all block the replication of HIV-1 viruses in the absence of viral Vif protein. In addition, the zinc domain can be further classified into three groups, Z1, Z2 and Z3 based on their sequence differences (Figure 1.2B) [4].

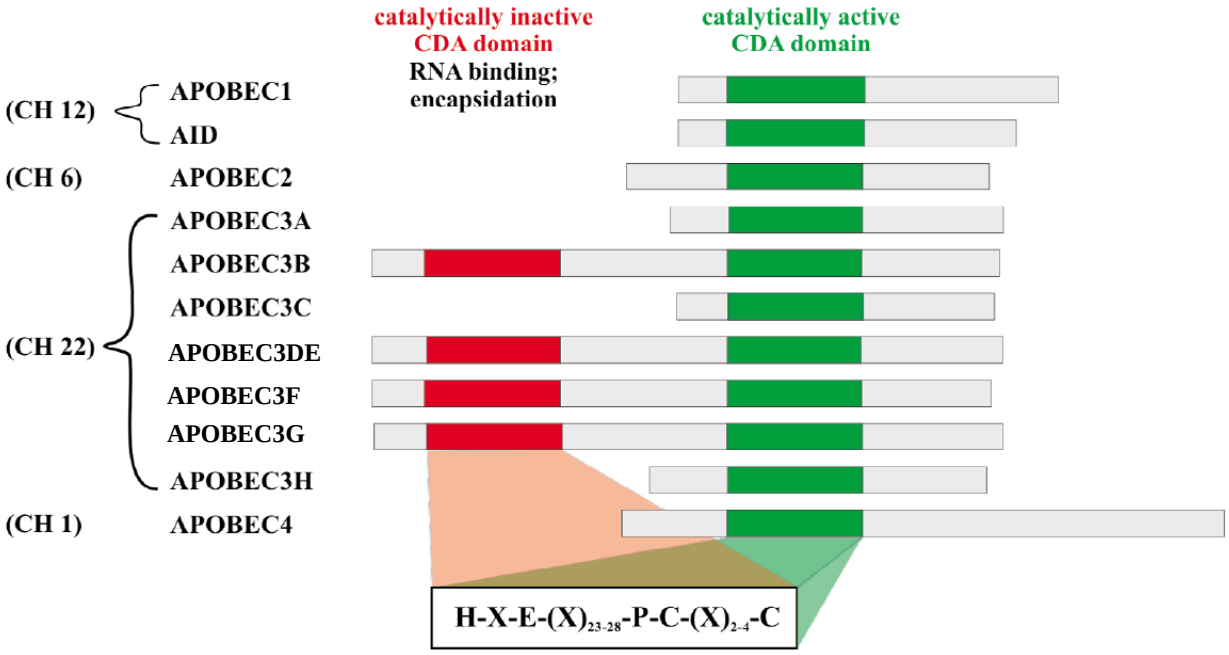


Figure 1.1. The human APOBEC family. Domain organization of human APOBEC proteins. The catalytically active cytidine deaminase domain (CDA) is colored in green, while the catalytically inactive cytidine deaminase domain is colored in red. The chromosomal location of each gene is indicated on the left. (Adapted from Holmes, R.K., Malim, M.H. and Bishop, K.N., (2007) *Trends Biochem Sci*, **32**(3), 118-28)

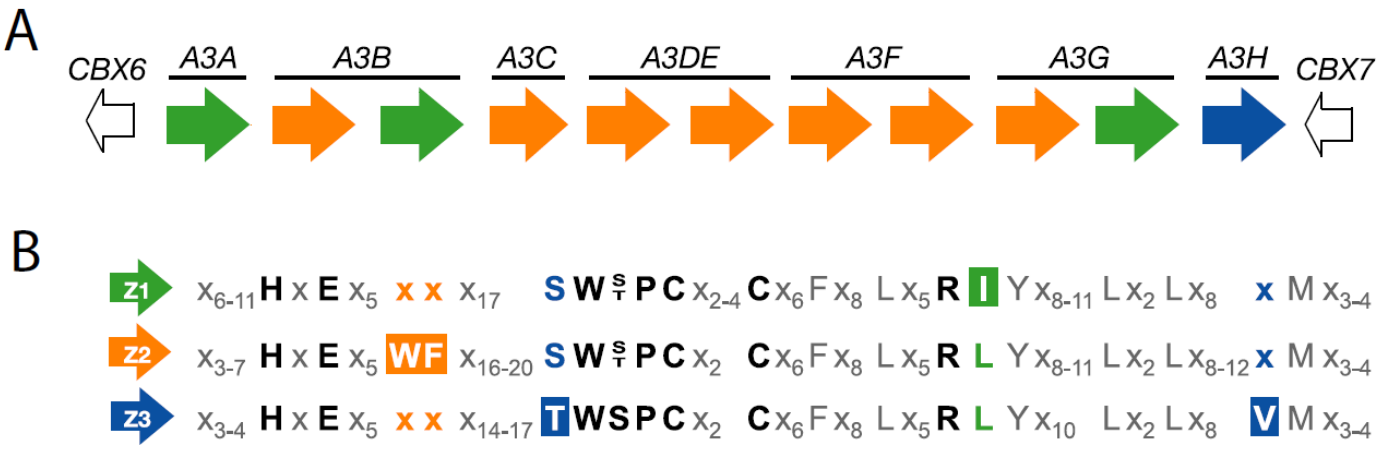


Figure 1.2. Classification of deaminase domains of APOBEC3 proteins. (A) Schematics of the human *APOBEC3 family members.* Z1, Z2, and Z3 domains are shown in green, orange and blue, respectively. (B) The sequence alignments of the three distinct Z-domain groups, the difference among sequence are in colored, and "x" represents any amino acid. (Adapted from LaRue, R.S., *et al.*, (2009) *J Virol*, **83**(2), 494-7)

The zinc domains of each A3 protein are colored based on the groups they belong to (Fig. 1.2A). It clearly demonstrates each Z1, Z2 and Z3 all have the ability to exert deaminase activity in at least one family member.

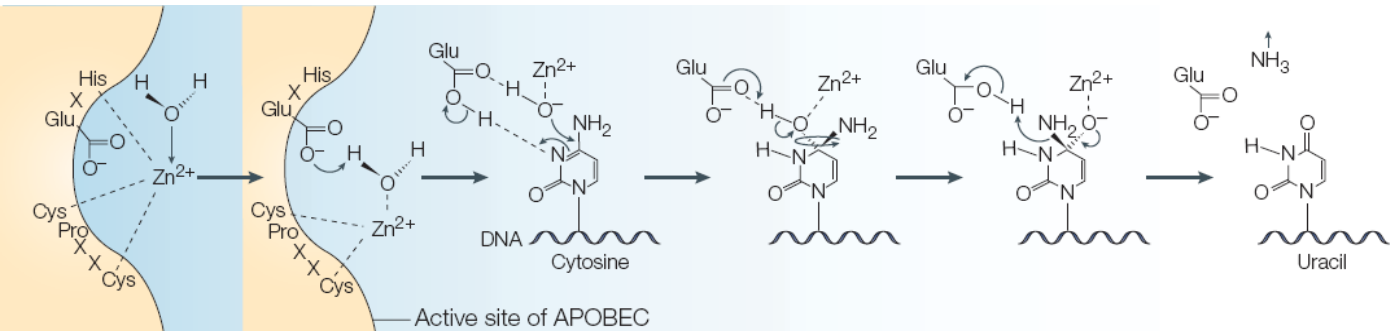


Figure 1.3. A possible mechanism for APOBEC-dependent cytosine deamination. (1) Two cysteine (Cys) and histidine (His) residues of the active site coordinate a zinc ion (Zn^{2+}). (2) A hydroxide ion (OH^-) is produced when water reacts with glutamate (Glu) and the zinc ion. (3) The glutamate protonates the pyrimidine ring of cytosine at the N3 position. (4) This results in the formation of a tetrahedral intermediate. (5) The amino group (NH_2) receives this proton in a reaction that causes cleavage of the C–N bond and formation of a C4–O double bond. (6) Uracil and ammonia (NH_3) are released from the active site. (Adapted from Harris, R.S. and Liddament, M.T. (2004) *Nat. Rev. Immunol.* **4**(11), 868-77)

Based on the crystal structures of cytidine deaminases in bacteria [5, 6] and yeast [7, 8], a possible mechanism for APOBEC protein deamination of cytosine was proposed (Figure 1.3). These proteins are defined by a highly conserved zinc-coordinating motif, His-X-Glu-X_{23-28}-Pro-Cys-X_{2-4}-Cys, where the histidine and the two cysteines position the zinc ion while the glutamate generates a hydroxide ion (OH^-) to remove the amino-group from the cytosine [9]. The consequence of this catalysis is to transform cytidine to uridine, a process named “cytidine deamination”.

The function of the first family member, identified APOBEC1, was shown to catalyze deamination of the cytidine of a CAA codon for Gln to a UAA codon for stop translation on the apolipoprotein B mRNA [10, 11]. This catalysis results in translating a truncated apolipoprotein, apoB-48 (with only 48% mRNA translated), instead of the full length protein, apoB-100, in the intestine [12]. ApoB-100 containing VLDL (very low density lipoprotein) is important in the formation of the LDL (low density lipoprotein, bad cholesterol) in the liver where no APOBEC1 is expressed [13]; while ApoB-48 containing VLDL, which lacks the component to bind to LDL, are cleared from the blood stream too soon to be metabolized to become LDL. APOBEC1 protein itself binds to apoB mRNA weakly and nonspecifically, it requires the auxiliary proteins p44, p55, p66, p100 and p240 for apoB mRNA specific binding and editing [14].

Both AID (Activation-induced cytidine deaminase) and APOBEC2 are the most ancient members in APOBEC family, the sequence homology of these two proteins is 44% [9]. AID plays important physiological roles in both adaptive and innate immunity by deaminating the cytidines in the DNA. It initiates immunoglobulin gene diversification reactions, somatic hypermutation and class switch recombination by triggering the deamination of cytosine bases within immunoglobulin genes variable and switch region [15, 16]. The resulting uracil residues promote additional DNA processing events that ultimately enable the formation of high-affinity, isotype-switched antibodies. AID and antibody-mediated immunity are features of all vertebrates.

APOBEC2 contains one zinc-binding deaminase domain, and its function and physiological roles are still unclear. Recent studies indicated that mouse APOBEC2 does

not have any cytidine deaminase activity and is not required for mouse development and survival [17]. The crystal structure of APOBEC2 is the first structure of the human APOBEC family. The protein was in a tetrameric form [18].

The APOBEC3 subgroup later evolved within this family, and, unlike other subgroup that contains only one member, currently has seven members. The genomic locus of A3 genes of all animals featured flanking of two conserved genes CBX6 and CBX7 [4, 19]. Unlike other single member subgroups, why the APOBEC3 gene expands in human is still unknown. Recently, studies on the physiological functions of this subgroup have been rigorously explored, approximately, their functions can be classified as retroelement restriction, such as retroviruses, parvoviruses and retrotransposon, antibody gene diversification and newly discovered foreign DNA restriction (Table 1.1) [20-28]. Specifically, the family members A3B, A3DE, A3F and A3G have potent antiviral activity against HIV-1 virus in the absence of viral Vif protein [9, 23, 29-32]. Interestingly, all of them contain two deaminase domains, where only one deaminase domain has catalytically activity while the other deaminase domain is responsible for nucleotide binding and virus encapsidation [33-37].

APOBEC4 was identified in 2005 by searching for AID sequence homology in the PSI-BLAST database [38]. Its homology in mammals, chicken and frog are identified, but the function and catalytic activity of this protein are unclear, more biochemical experiments are needed to obtain more specific information about this protein.

Supplementary Table 1. Physiological functions of DNA cytidine-to-uridine deamination.

	Foreign DNA restriction	Antibody gene diversification	Retroelement restriction		
			Retroviruses (*e.g.*, HIV-1)	Parvoviruses (*e.g.*, rAAV)	Retro-transposons (*e.g.*, L1)
Enzyme(s)	A3A & other A3s (not A3G)	AID	A3G & A3F	A3A	A3A, A3B, A3F & A3G
Substrate	Naked double-stranded DNA (exogenous DNA; *e.g.*, bacterial)	Double-stranded chromosomal DNA (nuclear)	Single-stranded viral cDNA replication intermediates	Undefined DNA replication intermediates	Undefined transposon replication intermediates
Deaminase activity requirement	Yes	Yes	Yes	No	No
Hypermutation of substrate	Yes (C/G-to-T/A transitions on both strands)	Yes (all 6 types of base substitutions)	Yes (cDNA strand-specific C/G-to-T/A transitions)	No	No
UNG2 (uracil excision) requirement	Yes	Yes (and MSH2)	No	No	No
Interferon-inducible	Strong	None	Weak	Strong	Strong/Weak
Biological function	Innate immunity to foreign DNA	Adaptive immunity through antibody diversification	Innate immunity to retrovirus infection	Unknown	Innate immunity to endogenous retro-transposition
Relevant Cell Type	Phagocytes (specialized) & other cells (housekeeping)	B-cells	T-cells & possibly other cell types	Unknown	Germ & somatic cells
Fate of substrate	Degradation & hypermutation (C/G-to-T/A transitions)	Somatic hypermutation (6 types of base substitutions) & DNA breaks (in class switch recombination)	Degradation / hypermutation	Degradation	Failed integration

Table 1.1. Physiological functions of APOBEC family. (Adapted from Stenglein, M.D. *et al.*, (2010) *Nat Struct Mol Biol.*, 17(2), 222-9)

1.2 The antiviral activity of APOBEC3G

It has been observed that HIV-1ΔVif virus could replicate in some cell lines (called permissive cells, such as 293T and SupT1) but not in other cell lines (called non-permissive cells, such as H9 T cell and macrophages) [39]. In the presence of Vif protein, the ΔVif virus can overcome the intrinsic immune defense mechanism and replicate in the non-permissive cells [40]. By comparing the results of subtractive hybridization studies between permissive and non-permissive cells, APOBEC3G (formerly named CEM15) was observed to express only in the non-permissive cells [41].

A3G is a 384 amino acids protein, it contains two zinc-binding motifs, one is in N-terminal pseudo-deaminase domain (A3G-ntd) which counteracts with HIV Vif protein and the other is in C-terminal deaminase domain (A3G-ctd), but only the C-terminal deaminase domain catalyzes the conversion of DNA cytosines to uracils [42]. The solubility of A3G is very low and aggregation as well as precipitation was observed during the purification process. So far, the maximum *in vitro* concentration A3G achieved is merely 350 nM, in high concentration of detergent and glycerol buffer (1% Triton X-100 and 10% glycerol) [33, 43]. An *in vitro* deaminase assay demonstrates that A3G works only on single-stranded DNA from the 3' to 5' direction, with a dissociation constant (with 10nt or longer ssDNA) of about 50 nM [43]. A possible model of how APOBEC3G exerts its antiviral activity against HIV-1 Δ Vif virus is shown in Figure 1.4. In the figure, the cellular APOBEC3G is incorporated into the HIV Δ Vif budding virions in the producer cell. When virions arrive and fuse into the target cell, A3G deaminates the cytosine in the minus-strand (first strand) viral cDNA, and this may produce a non-functional provirus [44-47].

The second antiviral activity of APOBEC3G was elucidated and named "deaminase-independent inhibition" (Figure 1.4). The studies show the deaminase activity is not a requirement for antiviral activity against HIV-1 or HBV viruses [34, 35, 48-50]. The chimera mutants of A3G that abolish enzymatic functions of C-terminal deaminase were still able to exert antiviral activities [51]. The phenotype of this deaminase-independent inhibition suppresses the synthesis of viral reverse transcripts and prevents the viral cDNA integration [51, 52]. Moreover, the physical interactions between APOBEC3G with HIV-1 RNA, Nucleocapsid and Integrase are implied to correlate to the phenotype [46].

APOBEC3G can form low-molecular-mass (LMM) and high-molecular-mass (HMM) complexes in the cytoplasm, HMM A3G (>700 kDa) is inactive and can be converted to active LMM A3G by treating RNase. In the peripheral blood, APOBEC3G is in LMM complex in the resting $CD4^+$ T cells that are resistant to HIV infection. Once the $CD4^+$ T cells are activated, the APOBEC3G forms HMM complex through RNA and, thereby, abolishes its antiviral activity [53].

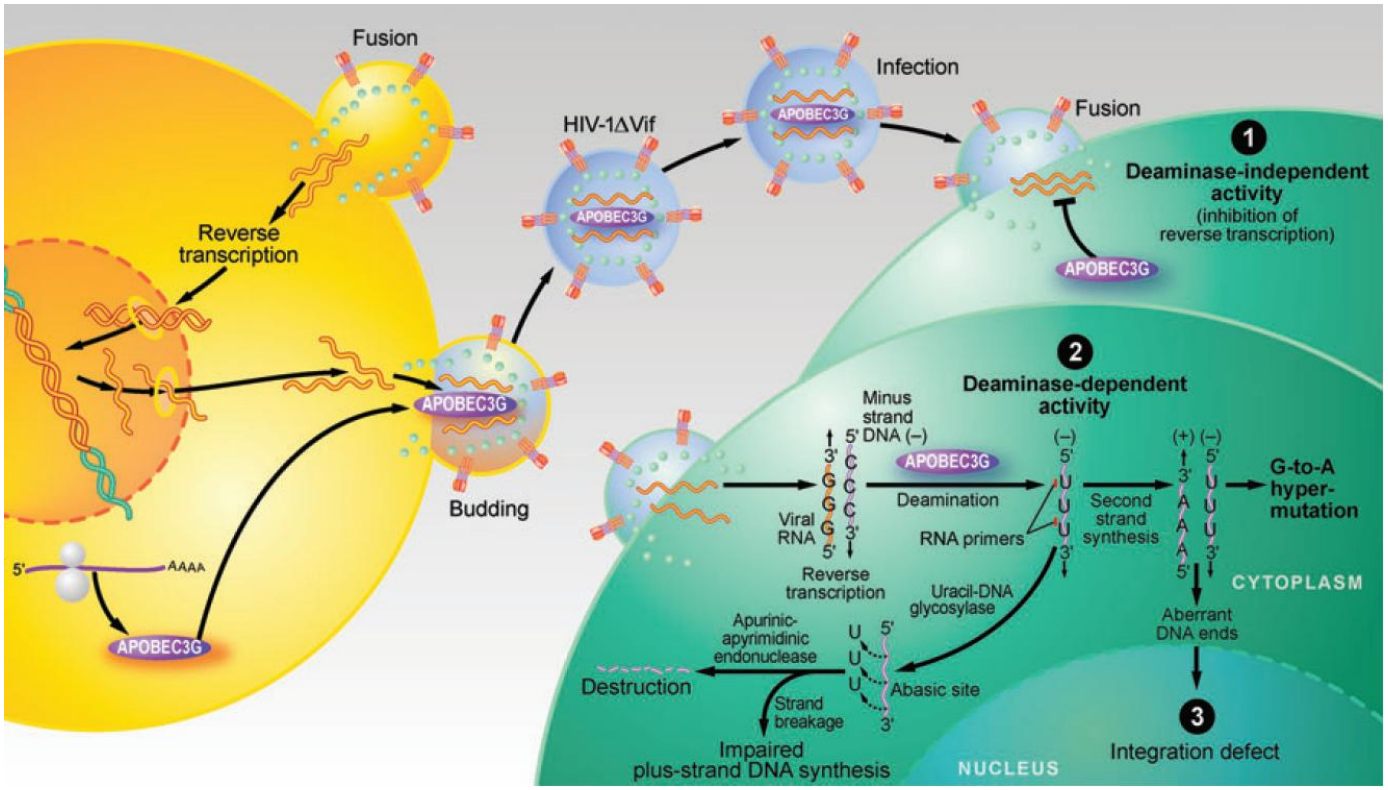


Figure 1.4. The APOBEC3G exerts its antiviral activity through two mechanisms. In the absence of vif, APOBEC3G is encapsidated into budding HIV-1 virions in the producer cells. It then stays associated with the reverse transcription complex in the target cells, where it inhibits viral replication by both deaminase-independent and deaminase-dependent pathways. (1) Human A3G bound to HIV-1 RNA may physically prevent reverse transcriptase accumulation, resulting in a deaminase-independent decrease in the production of early reverse transcripts. (2) Human A3G mediates extensive cytidine to uridine deamination in the minus-stranded cDNA. This effectively inhibits viral replication due either to the accumulation of dG-to-dA hypermutations in the subsequently synthesized plus-strand DNA or because the uracil-containing minus-strand DNA is degraded by the uracil DNA-glycosylase and apurinic-apyrimidinic endonuclease. (3) Additionally, decreased chromosomal integration of the double-stranded viral DNA required for virion formation may occur owing to defects in tRNALys3 primer cleavage leading to the formation of viral DNA with aberrant ends. (Adapted from Chiu, Y.L., and Greene, W.C., (2008) *Annu. Rev. Immunol.*, 26, 317–53)

1.3 Degradation of APOBEC3G by Vif protein

It has been over 20 years since the nine genes of the HIV-1 (human immunodeficiency virus) genome were identified, however, unlike the other eight genes, the function of the Vif (viral infectivity factor) protein was elucidated only in recent years [41, 54]. HIV-1 Vif is a 23 kDa protein consisting of 192 amino acids, and is required for the HIV-1 virus to replicate in the nonpermissive cells [55, 56]. So far, due to the native partially folded property of the Vif family, the structures of viral Vif proteins have yet to be determined. A mechanism of how Vif protein inhibits the antiviral activity of APOBEC3G has been proposed [57-61]. The Vif protein blocks APOBEC3G packaging into the budding virions by triggering the ubiquitin-proteasome degradation pathway in the producer cell. The model is depicted in Figure 1.5, in which Vif recruits the cellular ubiquitin-ligase complex, which includes Cullin5 (Cul5), ring-box-2 (RBX2), Elongin B and Elongin C, to ubiquitinate APOBEC3G for proteasomal degradation [62]. Through mutagenesis studies, four solvent-exposed lysine residues (297, 301, 303, and 334) on the surface of A3G were shown to participate in ubiquitination [63]. The Cul5 complex together with Vif thereby prevent the virion packaging of the ubiquitinated A3G [61]. In addition, the residue D128 of A3G has been shown to be crucial for binding to Vif [64]. Also, the region, $Y^{40}RHHY^{44}$, on the N-terminus of Vif is an important motif interacting with APOBEC3G [65-67]. The studies have shown HIV-1 Vif restrain human A3G but not African green monkey (AGM) A3G, demonstrating species difference. Interestingly, the species specificity can be switched by substituting aspartic acid 128 of human A3G for lysine as found in AGM. The introduction of a D128K mutation in human A3G confers resistance to HIV-1 Vif with sensitivity to SIV_{AGM} Vif; while AGM A3G with a

K128D substitution abolishes sensitivity to SIV_{AGM} Vif protein [68-70]. Based on a full-length model of APOBEC3G built by the Matsuo group, the location of D128 is on the loop between β4 and α5 and has a solvent-exposed position on the N-terminal deaminase domain [71].

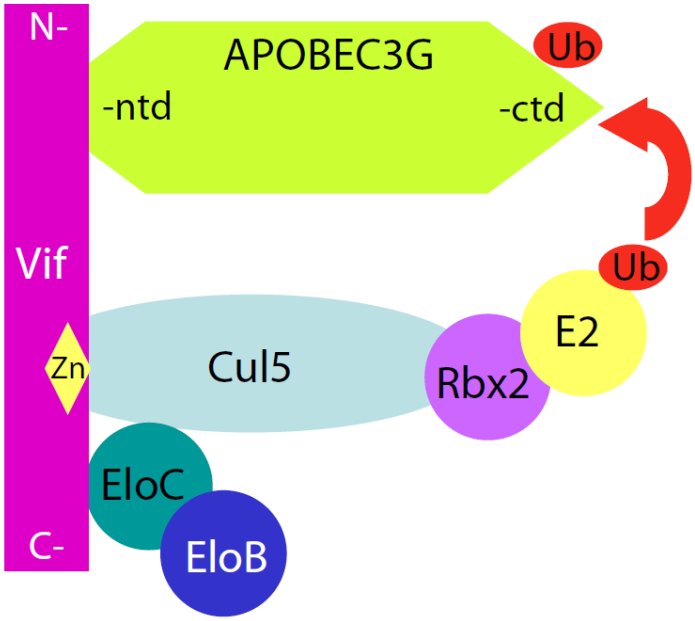


Figure 1.5. APOBEC3G degradation by the Vif-Cul5-EloB-EloC E3 ubiquitin ligase. The N-terminal part (N-) of Vif binds to the N-terminal domain (ntd) of APOBEC3G and forms a complex that recruits the Elongin B (EloB) and Elongin C (EloC) proteins in its C-terminal part (C-). The zinc-binding motif of Vif then mediates a second interaction with the Cullin-5 (CUL5)–ring-box-2 (RBX2) complex of the E3 ligase. Finally, the E2 ubiquitin-conjugating enzyme transfers ubiquitin to the C-terminal domain of A3G for proteasome degradation.

1.4 Rational and Overview of the thesis work

This thesis provides the solution NMR structures of the catalytic deaminase domain of APOBEC3G. For NMR studies, soluble and stable proteins are required. However, wild-type APOBEC3G has a low solubility and is prone to rapid precipitation; therefore a

new strategy to design a more soluble and catalytic active APOBEC3G has been taken. In Chapter 2, the minimal catalytic region of APOBEC3G protein was determined, and residues within this region that could be mutated to alanine without losing its catalytic activity were identified. Next, the lysine substitutions on these mutable positions were generated to screen variants with enhanced solubility. By combining promising lysine variants, the double lysine variant with catalytic activity and enhanced solubility was selected. The variant was then subjected to non-detrimental cysteine to alanine mutations and buffer optimization for long-term storage. Chapter 3 demonstrates how the soluble and active mutants were generated for NMR structural studies based on the information in Chapter 2. It also details how the structures of the deaminase domain of APOBEC3G were determined using NMR spectroscopy. This structure was then used to construct a model of a single strand DNA binding complex. Finally, a full-length model of APOBEC3G was generated based on the structure of the longer deaminase domain of APOBEC3G which contains the linker region connecting N-terminal and C-terminal deaminases.

Chapter 2

Screening for more soluble and active variants of the catalytic domain of APOBEC3G

2.1 Introduction

APOBEC3G has two structural domains, the N-terminal domain (A3G-ntd) which interacts with the HIV Vif protein and the C-terminal domain (A3G-ctd) which catalyzes the deamination that converts cytidine to uridine on viral DNA. To date, the mechanism of how A3G-ctd binds to polynucleotide DNA and exerts its enzymatic activity has not yet been fully elucidated. Many biochemical and biophysical studies have been hindered due to the low solubility and instability of A3G-ctd. The wild-type A3G-ctd-198-384 protein degrades and precipitates rapidly even at 15 ℃ within 3 days of purification. NMR spectra show aggregation of this protein at low concentration (20-80 uM).

Solubility of a protein is a physical property describing its ability to be dissolved in solution. It is affected by temperature, pressure, salt concentration, detergent concentration, pH and buffer type (solvent). The term "intrinsically low solubility" here is defined as a protein that cannot be dissolved well in solution, and this is often due to the distribution of hydrophobic and hydrophilic residues on the surface of proteins. The

charged or polar side chains of hydrophilic residues will interact with water and thereby increase the protein's solubility, while the non-polar side chains of hydrophobic residues do not favor interaction with water and often resulting in self association and precipitation. Of the twenty amino acids found in proteins, nine are hydrophobic and eleven are hydrophilic, each with different mass and charge. Most hydrophobic residues are likely structurally important (perhaps forming part of the enzyme core), whereas, others would be positioned on the surface of the protein (perhaps defining interaction sites). Therefore, the hypothesis proposed here is that the substitution of hydrophilic residues for hydrophobic ones on the protein surface will improve the intrinsic solubility. In this chapter, a systematic mutagenesis (Table 2.1) on the C-terminal catalytic domain of APOBEC3G was performed to screen for soluble and active A3G-ctd variants. Firstly, the minimal catalytic region was determined by an *E. coli*-based mutation assay. Secondly, single alanine substitutions of the hydrophobic residues (except proline but including cysteine) were constructed individually and were tested for their catalytic activity and solubility. At this step, the positions that could be mutated to alanine without compromising activity were identified. Thirdly, a single lysine was introduced into the position of the active alanine variants individually, and these variants were tested for their activity and solubility. At this step, the residues that could be mutated to single lysine without losing activity but with enhanced solubility were identified. Fourthly, the double lysine derivatives which were constructed based on the information obtained from single lysine variants were subjected to the activity assay and the solubility test. Thus, we were able to select active A3G-ctd mutants with enhanced solubility. Then, the buffer conditions were optimized for NMR study of A3G-ctd. Finally, to avoid intermolecular

disulfide bonds, three cysteine to alanine replacements, which did not hinder the catalytic activity, were introduced into the double lysine variants.

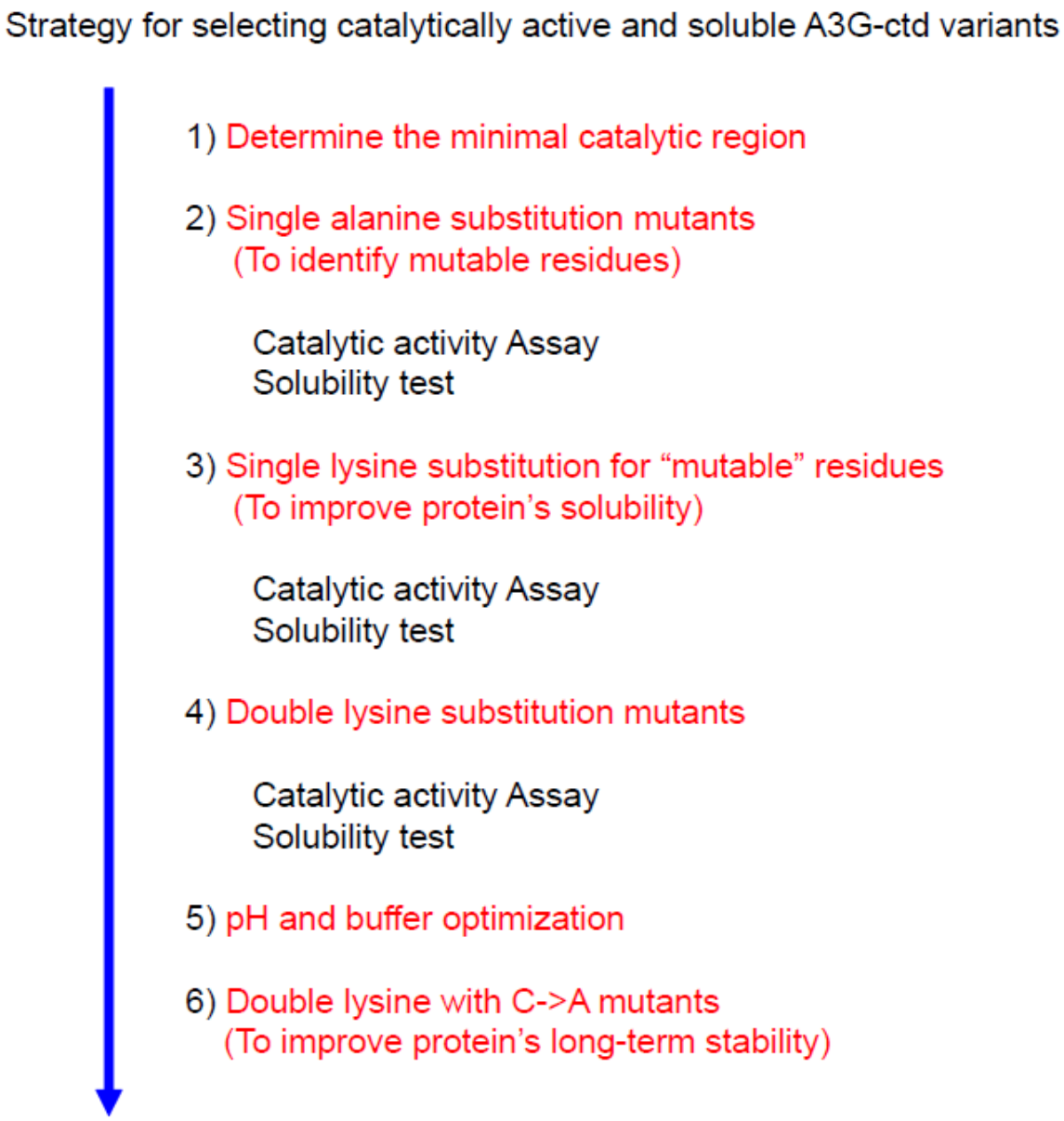


Table 2.1. Flow chart for screening soluble and active A3G-ctd derivatives.

2.2 Materials and Methods

2.2.1 Plasmid constructs

The A3G cDNA used here matches NM_021822. A3G and mutant derivatives were expressed as GST fusion proteins using pGEX6P1 or pGEX6P2 (GE Health). An EcoRI-SalI DNA fragment from pTrc99A-A3G encoding full-length A3G [42] was sub-cloned directly into pGEX6P1. A3G deletion mutants were constructed by amplifying the relevant A3G coding regions, digesting the resulting PCR products with SmaI and SalI and ligating them into the SmaI and XhoI sites of pGEX6P2. All amino acid substitution derivatives were constructed using the QuikChange protocol (Stratagene) and verified by DNA sequencing.

2.2.2 *E. coli*-based Rif^R mutation assays

The *E. coli*-based, rifampicin-resistance (Rif^R) mutation assay has been used extensively to monitor the intrinsic DNA cytosine deaminase activity of several A3 proteins including A3G (*e.g.*, [42, 49, 72, 73]. Each construct was expressed constitutively in *E. coli* strain BW310 (uracil excision-defective). Aliquots of saturated overnight cultures (LB plus 200 µg/mL ampicillin) were spread onto LB plates containing 100 µg/mL rifampicin to select for Rif^R mutants and diluted aliquots were spread onto LB plates containing 200 µg/mL ampicillin to determine the number of viable cells. Mutation frequencies were calculated by dividing the number of Rif^R mutants by the number of viable cells in each culture. For the deletion experiments, the mutation frequencies of 8 individual cultures were plotted and the median indicated. For the alanine and lysine mutant experiments, the Rif^R mutation frequency for each

construct was determined by assaying the median Rif^R mutation frequency for 8-12 independent cultures, calculating the fold-difference relative to the median value of the vector control cultures and averaging at least two (and up to 12) independent experiments. Representative raw frequencies were shown (Figures 2.7A& 3.3D) to facilitate comparisons, the median values were normalized and data from a minimum of 2 independent experiments were averaged (Figures 2.4& 2.8).

2.2.3 Solubility test

The plasmids of GST-A3G-ctd derivatives were transformed into *E. coli* strain BL21 DE3 RIL (Stratagene). The cells were cultured in 20 ml LB medium and induced at OD_{600} of 0.6 with IPTG 0.4mM at 17 ℃ for 16 hrs. Then, the cells were collected by centrifugation (2000 g for 15 minutes) and resuspended in 4 ml buffer [50 mM Tris, 200 mM NaCl, pH 8.0, 50 uM Zinc, 0.5% Tween 20, 7 mM β-mercaptoethanol, and protease inhibitor (Roche)], followed by sonication and centrifugation. The supernatant and pellet (resuspended with fresh 4ml buffer) were collected individually. The ratio of S/P was calculated by integrating the protein bands of supernatant and pellet on the SDS-PAGE gel. Immunoblots were done with antibodies from GE Healthcare (anti-GST) and from J. Lingappa (anti-A3G; [49]), and Image J software was used to quantify protein levels (http://rsb.info.nih.gov/ij/).

2.2.4 Protein expression and purification

GST-based constructs were expressed in *E. coli* strain BL21 DE3 RIL (Stratagene). Unlabeled proteins were produced by overnight expression at 17 ℃ in LB medium

containing 0.4 mM IPTG and 100 µg/mL ampicillin. Proteins were purified by sonicating cell pellets in lysis buffer [50 mM Na_2HPO_4/NaH_2PO_4 , pH 7.4, 50 uM Zinc, protease inhibitor (Roche)], separating the soluble (supernatant) and insoluble (pellet) fractions by centrifugation (12,110 g, 20 min, 4 ℃), binding to glutathione sepharose (GE healthcare), washing with lysis buffer and eluting with PreScission protease (GE Healthcare) in buffer [1 mM DTT, 50 mM Na_2HPO_4/NaH_2PO_4 (pH 7.4), 0.005% Tween 20] and, finally, concentrating with Centricon filters (Millipore) to desired volume.

2.2.5 Size exclusion column

For size exclusion experiments, GST-A3G198-384 was bound to glutathione sepharose, washed several times with lysis buffer, eluted with PreScission protease (GE Healthcare) in 1 mM DTT, 50 mM Na_2HPO_4/NaH_2PO_4 [pH 7.4] buffer, quantified and immediately 1 ml of 1mg protein was passed through a Superdex 75 size exclusion column (GE Healthcare) and detected in fractions by UV (absorbance 280). GST and lysozyme (both were purchased from Sigma) were gone through the column individually under the same condition.

2.2.6 CD spectroscopy and sedimentation velocity analytical ultracentrifugation experiments

Circular Dichroism (CD) Spectroscopy - Proteins were purified as described above and diluted to 6 µM in 50 mM Na_2HPO_4/NaH_2PO_4 (pH 7.4), 50 µM $ZnCl_2$. CD spectra were collected on a Jasco 710 dichrograph using 10 mm thick quartz cells at 10 ℃. Data were acquired between 190 and 250 nm at 50 nm/min with a bandwidth of 1 nm.

Sedimentation Velocity Analytical Ultracentrifugation Experiments - A3G-2K3A was diluted to 0.15, 0.4, 0.8 or 1.2 mg/ml in a buffer containing 50 mM Na_2HPO_4/NaH_2PO_4 [pH7.4], 0.005% Tween 20, 5 mM dithiothreitol and 50 µM $ZnCl_2$. Samples were then sedimented using a 4-hole rotor at 20 ℃ and 55,000 rpm. The rotor and cells were pre-equilibrated at 20 ℃ and the samples were kept on ice during protein preparation and dilution. Synthetic boundary cells were loaded with 430 µl of buffer and 420 µl of the appropriate sample solution. The cells were placed in the rotor and accelerated to 24,000 rpm while monitoring the transfer of the excess buffer in each cell. Subsequently, the rotor was stopped and then gently inverted to thoroughly mix the contents of the cells. The rotor was then equilibrated under vacuum at 20 ℃ and after a period of ~1 hour at 20 ℃ the rotor was accelerated to 55,000 rpm. Interference scans were acquired at 1 minute intervals for 6 hours. The data for each loading concentration was analyzed using the program DcDt+ (version 2.0.7) [74, 75]. The normalized g(s*) plots of all four concentrations of A3G-2K3A are shown in Figure 2.7D. The complete data set for A3G-2K3A was analyzed using Sedphat v4.4b using the model of a hybrid local continuous distribution and global discrete species [76]. The fit yielded a value of 22.15 kDa [21.85, 22.45; 95% confidence limits] for the molecular weight, and a corrected sedimentation coefficient, s20, w of 2.42S, with an r.m.s. error of 0.0034 mg/ml. A similar analysis using Sedanal v4.37 gave a value of 22.3 kDa [21.9, 22.8; 95% confidence limits] for the molecular weight, and a corrected sedimentation coefficient, s20, w of 2.39S, with an r.m.s. error of 0.0025 mg/ml [77].

2.2.7 Dialysis buttons

The 30 µl of 5mg/ml A3G-ctd protein was put in a dialysis button (Hampton Research) covered with a 10 kDa semi-permeable membrane (Spectrum Laboratories, Inc.) which allows buffers, ions and salt to pass. Next, the button was placed in the beaker filled with 30 ml tested buffer for buffer exchange. After two days, the solution was transferred out of the button in an eppendorf tube and was spun down at 4000 rpm for 10 minutes. The protein amount retained in the supernatant is analyzed by the SDS-PAGE gel.

2.3 Results and Discussion

2.3.1 The minimal region required for deaminase activity

Chimeric APOBEC3 proteins and site-directed mutants have been used to demonstrate that the intrinsic DNA cytosine deaminase activity of human APOBEC3G resides in the conserved C-terminal zinc-coordinating domain [33, 35, 42, 49]. To further delineate the minimal domain required for catalysis, nine A3G deletion constructs were tested for mutability in the *E. coli*-based Rif^R mutation assay (Figure 2.1). Bacteria expressing GST showed a median of 2.5 Rif^R mutants per 10^7 viable cells. Expression of full-length A3G fused to GST caused a 4.4-fold increase in mutation frequency. Apart from two notable exceptions, all of the A3G deletion constructs were inactive. The inactivity of constructs lacking 22 or 40 C-terminal residues was consistent with prior work showing that A3G lacking 38 or 89 C-terminal amino acids was unable to inhibit HIV-1Δvif [78, 79]. A more interesting result emerged from analyses of N-terminal deletion set. A3G variants encoding residues 175-384 or 198-384, but not 215-384, were

considerably more active and mutable than full-length A3G. These data demonstrated that the entire N-terminal, zinc-binding domain is dispensable for activity, and only A3G residues 198-384 are required.

2.3.2 A3G198-384 wild-type protein is monomer

Previous studies indicated that an A3G C97A mutant, which fails to form multimerization and contains inactive N-terminal catalytic domain, was incapable of co-immunoprecipitating wild-type A3G, but was still capable of DNA deamination [34, 35]. Consistent with these studies, A3G198-384 profiled as a 22 kDa monomer in size exclusion experiments, migrating clearly between the elution positions of lysozyme (14 kDa) and GST (25 kDa) (Figure 2.1C). It is not likely that stable oligomeric forms of A3G198-384 exist, because fractions 1-27 did not contain protein peaks.

2.3.3 The *E. coli*-based mutation assay result of alanine substitution variants defines essential and non-essential residues of A3G198-384

To more precisely delineate the residues and domains required for DNA deamination, the A3G198-384 fragment was used to construct a series of 69 alanine mutants (Figure 2.2). Mutagenesis was concentrated to hydrophobic residues and cysteines. This strategy was motivated partly by the likelihood that some of the hydrophobic amino acids would likely be important structurally (perhaps forming part of the enzyme core) whereas, more intriguingly, others would be positioned on the surface of the protein (perhaps defining interaction sites).

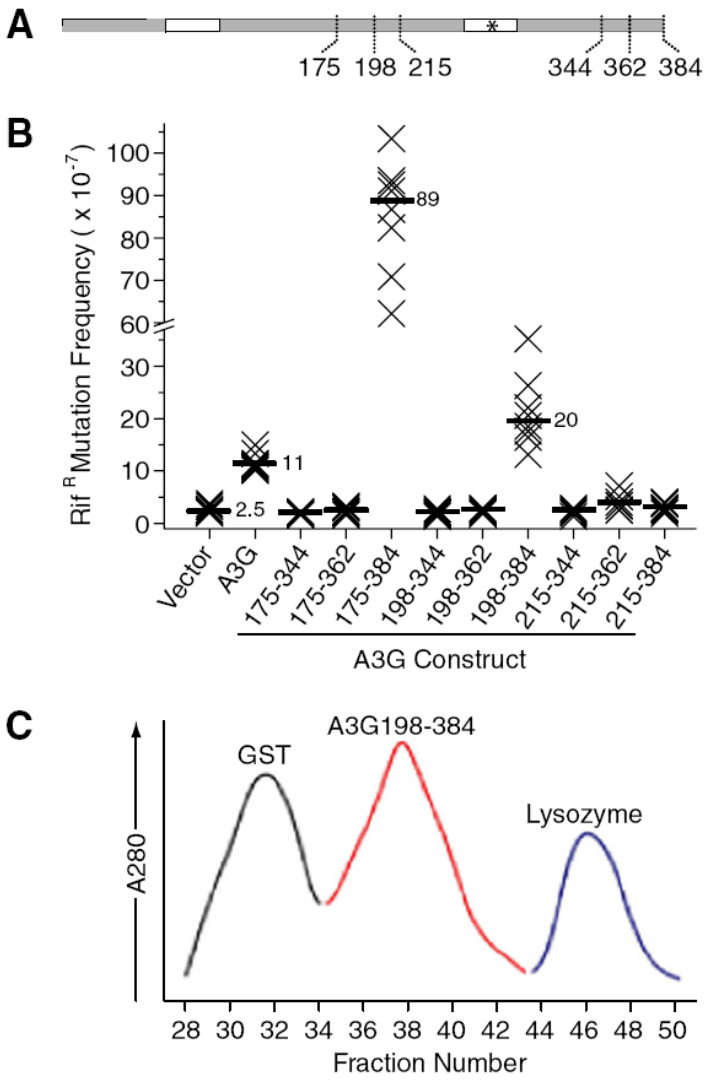


Figure 2.1. APOBEC3G deletion mutants delineate a minimal active domain. (A) An illustration showing the amino acid boundaries used for deletion constructs. The deaminase motifs are depicted by open boxes. (B) Rif^R mutation frequencies of the indicated GST-A3G constructs. The median mutation frequency for cells expressing the vector control, GST-A3G, GST-A3G175–384 and GST-A3G198–384 is indicated. (C) Size exclusion profiles of GST (25 kDa), A3G198–384 (22 kDa) and lysozyme (14 kDa) indicate that A3G198–384 is monomeric.

We also envisaged that mutating select hydrophobic residues might help overcome the solubility issues of A3G and other family members [33, 34, 43]. All 69 mutants were analyzed using the *E. coli*-based Rif^R mutation assay, because *in vitro* experiments were hindered by the fact that wild-type A3G198-384 frequently precipitated during biochemical purification and during long-term storage.

Twelve independent Rif^R experiments, each with at least 10 constructs (and 8-12 independent cultures per mutant), were required to analyze 69 alanine mutant derivatives of A3G198-384 (Figure 2.2). It was not feasible to simultaneously examine all 69 mutants, and minimizing inter-experiment variability was therefore important. This was done by normalizing the median Rif^R mutation frequencies of cells expressing the control vector (*i.e.*, the spontaneous mutation frequency was set to 1 and used as a baseline). The actual vector control values ranged from 0.8 to 3.9 Rif^R mutants per 10^7 viable cells (n=12 experiments). This approach readily enabled the mutagenic effects of A3G198-384 and derivatives to be compared between multiple experiments. For instance, the first two columns of each row of Figure 2.2 report the relative mutation frequency of the vector control and A3G198-384, which increased the Rif^R mutation frequency 14.9-fold (SEM: 1.5-fold; n=12 experiments; the actual values ranged from 15.1 to 54.5 Rif^R mutants per 10^7 viable cells). Although the raw experimental values fluctuated modestly between experiments (attributable to factors such as incubation temperatures, freshness of the rifampicin-containing plates, colony sizes upon counting, time in stationary phase, *etc*.), the small SEMs indicated that the relative inter-experiment values were remarkably constant and therefore readily comparable.

We were first struck by the surprising number of alanine mutants that elicited levels of DNA deaminase activity that were at least 3-fold greater than those of the vector control (Figure 2.2). In total, 31/69 mutants met this threshold. Several of the other mutants also triggered mutation frequencies significantly above those of the vector control cells and above those of cells expressing a catalytically dead A3G variant, E259A [49, 80].

The second notable feature of this dataset was an obvious clustering of non-essential and essential residues, defined by alanine substitution mutants that retained or lost significant activity, respectively. Approximately two-thirds of the non-essential amino acids were found in the A3G region spanning 198-275, with 224-253 particularly striking. Conversely, the majority of the essential residues were found in the C-terminal interval, 276-384, suggesting that the C-terminal end of A3G is required for the integrity of the enzyme.

Third, the alanine mutant data agreed with those from the deletion studies. Toward the N-terminal end, F202, F206 and W211 were required for Rif^R mutagenesis, explaining why all of the A3G variants starting at position 215 were inactive. Similarly, L364, L375, I378 and L379 were clearly required, offering a reasonable explanation for why all three C-terminal deletion constructs were inactive.

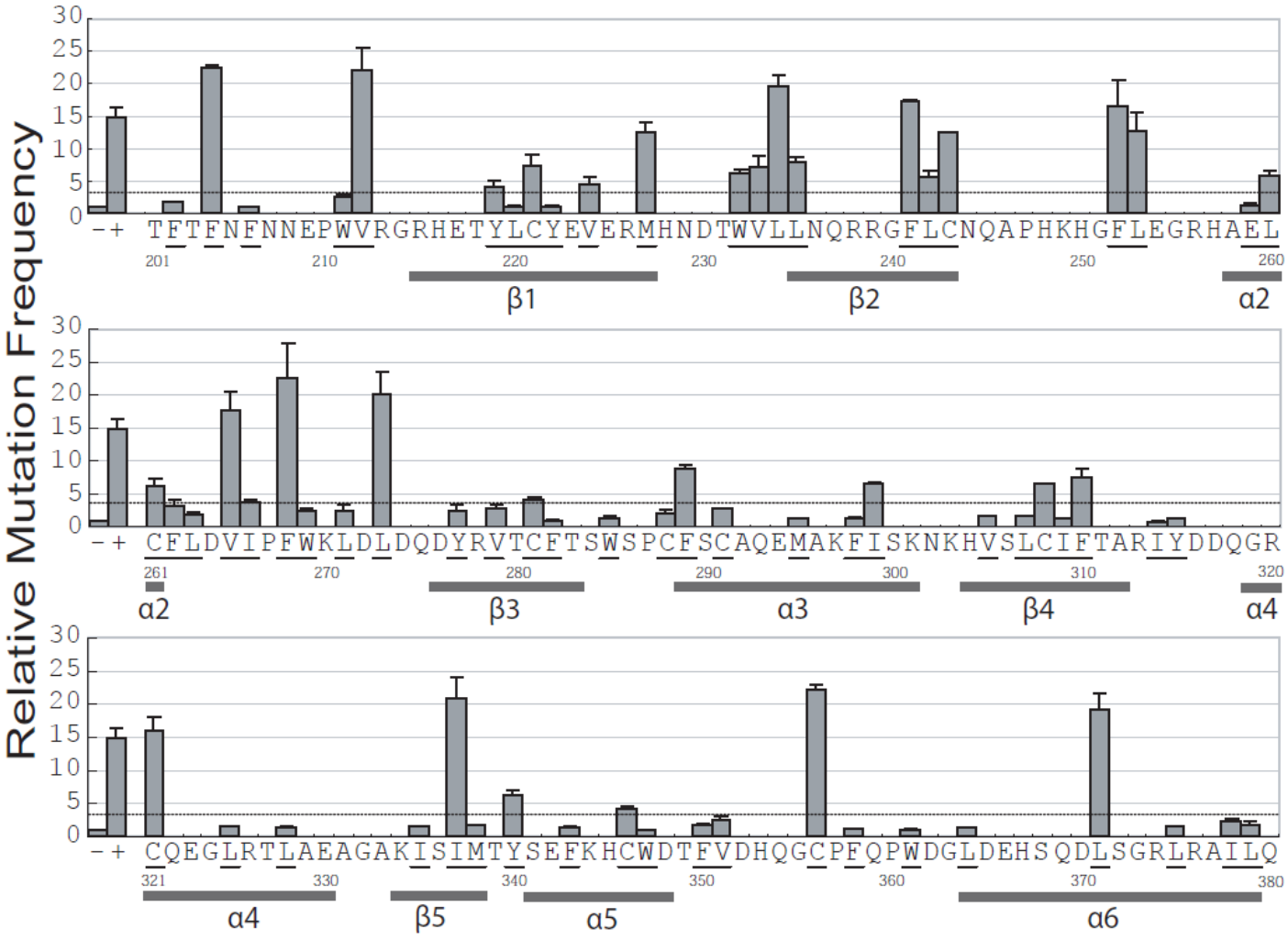
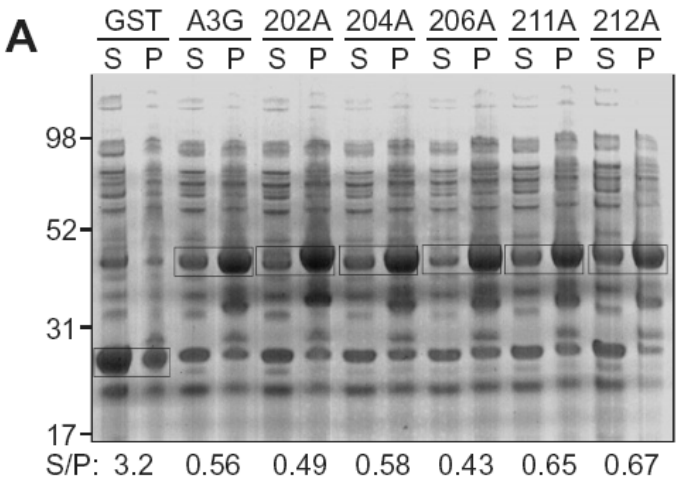
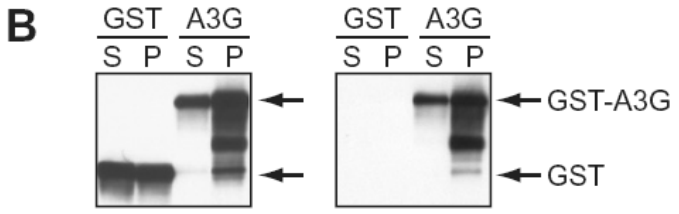


Figure 2.2. Mutator phenotype of 69 APOBEC3G alanine substitution mutants. Histograms showing the relative RifR mutation frequencies of cells expressing the vector control (-), A3G198–384 (+) or derivatives with alanine substitutions at the underlined amino acid positions. Each histogram bar reports data from two to five independent experiments, except the bars for vector and A3G198–384, which reflect data from 12 (±S.E.M.). The dotted line represents the 3-fold significance threshold that was used to distinguish active and inactive mutant constructs. A3G amino acids 198–200 and 381–384 were not shown.

2.3.4 The expression level of most alanine variants is similar to that of the wild-type

To help exclude the possibility that the lack of deaminase activity might simply be due to reduced solubility or expression, the relative abundance of each protein was analyzed via a solubility test. Rather than examine whole cell extracts, the soluble (supernatant) and insoluble (pellet) fractions were considered separately (the sum of course reflecting whole cell levels). A representative Coomassie-stained gel is shown for the first five GST-A3G198-384 derivatives, F202A, F204A, F206A, W211A and V212A (Figure 2.3A). The supernatant/pellet ratio of all five mutants was similar to that of the parent construct, indicating that the low activity levels for F202A, F206A and V212A were not simply due to solubility or expression deficiencies. Anti-GST and anti-A3G immunoblots confirmed the identity of the Coomassie-stained bands (Figure 2.3B). The majority of the remaining mutants were as soluble and some were even more soluble than GST-A3G198-348 (Figure 2.3C). However, six mutants were less soluble. Four of these mutants, L260A, C261A, C281A and C308A, caused significant Rif^R mutation frequency increases indicating that these variants are catalytically active and that the corresponding mutation frequencies may be underestimates. Correction factors were not calculated because these mutants did not impact major conclusions. However, two of the six less soluble mutants, W269A and C288A (one of the conserved zinc-coordinating positions), showed no activity. We were therefore unable to determine whether this was due to gross insolubility, to a loss of enzyme activity or to both of these reasons. Nevertheless, the expression data indicated that the vast majority of the mutants are well expressed and modestly soluble and, therefore, that the corresponding *E. coli*-based activity data are informative.





C S/P relative to A3G198-384

> 2 SD higher	234A, 235A, 241A, 253A, 371A
≤ 2 SD	202A, 204A, 206A, 211A, 212A, 221A, 227A, 232A, 233A, 243A, 252A, 259A, 265A, 268A, 273A, 289A, 291A, 299A, 310A, 321A, 337A, 340A, 346A, 356A
> 2 SD lower	260A, 261A, 269A, 281A, 288A, 308A

Figure 2.3. GST-A3G198-384 expression data. (A) A representative SDS-PAGE gel showing the soluble (supernatant, S) and insoluble (pellet, P) amounts of GST, GST-A3G198-384 (wild-type) and 5 mutant derivatives. The S/P ratio of the boxed bands is shown below each lane. The *E. coli* protein(s) that migrates indistinguishably from GST-A3G198-384 was present in every sample and it provided a constant (but lower) background signal. (B) Anti-GST (left panel) and anti-A3G (right panel) immunoblots confirm the identities of the bands boxed in A. (C) Mutants with higher, near equivalent or lower S/P values than those of wild type GST-A3G198–384 are listed. Total protein amounts (S + P) were similar for every construct ((A) and data not shown). S/P is protein ratio in supernatant and pellet, and SD is standard deviation of wild-type.

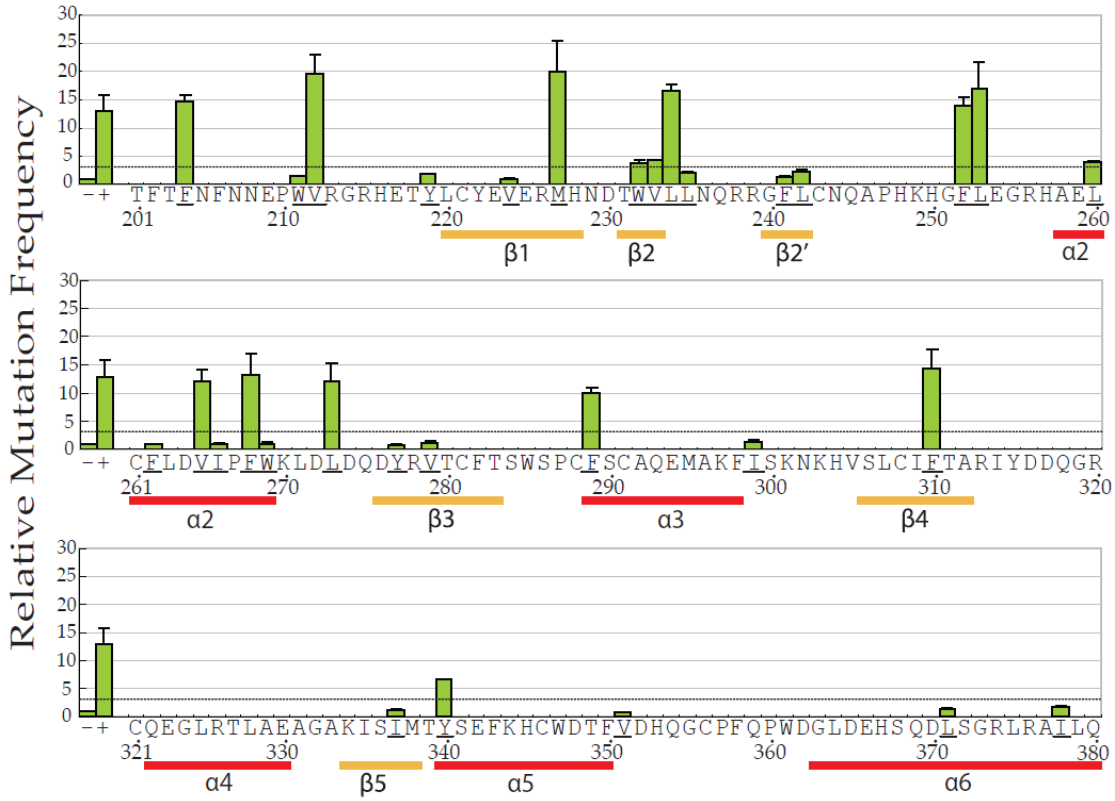


Figure 2.4. Mutator phenotype of 31 A3G198-384 lysine substitution mutants. The histogram bars show the relative Rif^R mutation frequencies of cells expressing the vector control (-), A3G198-384 (+) or derivatives with lysine substitutions at the underlined amino acid positions. A3G amino acids 198-200 and 381-384 were not mutated or illustrated.

Solubility test result of active single Lys mutants

S/Pratio relative to A3G 198-384 wt	
>2 SD higher	L234K, L273K, F310K
≤2 SD	V212K, M227K, F252K, L253K, V265K, F268K
< 2SD lower	F204K

S/P: protein ratio in Supernatant and Pellet
SD: Standard Deviation of wt

Table 2.2. Solubility test result of single lysine mutants

2.3.5 Replacement of mutable residues by single lysine

From the result described above, 31 mutable positions that can be substituted for alanine without losing activity were identified. According to our hypothesis, the substitution of hydrophilic residues for hydrophobic ones should improve solubility. There are 11 hydrophilic residues found in the primary structure of proteins, and it was not clear which one should be targeted for substitution in this study. Considering APOBEC3G is a DNA deaminase and A3G-ctd is acidic protein (PI is 5.8), therefore, replacing a hydrophilic residue with a positively charged group should increase the intrinsic solubility without reducing the binding affinity of DNA. Therefore, lysine which has a positive charge at pH 7.4 and has a minimal side chain length was selected for further mutation study. Based on the positions of these 31 active mutants, we substituted lysine for alanine and tested the variants produced for their activity and solubility. Finally, the assays demonstrated 11 out of 31 variants were active (Figure 2.4), and these 11 variants were then tested for their solubility. Table 2.2 shows that three variants (L234K, L273K and F310K) have higher solubility than wild-type A3G198-384 while six variants have solubility similar to wild-type.

2.3.6 Screening for a catalytically active, soluble derivative of APOBEC3G: A3G-198-384-2K

Subsequently, we tested the activity of the double lysine mutants by combining these three residues (L234K, L273K and F310K). Surprisingly, one variant, L234K and F310K were combined to yield a protein that was 2.4-fold more active and 4-fold more soluble (Figures. 2.5A and 2.7A), and it was named A3G-198-384-L234K-F310K or

simply A3G-198-384-2K. The 2K variant is soluble even after removing the GST tag (Figure 2.5B), with a maximum concentration of 300 uM. However, precipitation of this protein was observed at 25 ℃ after three days and this variant (data not shown), therefore, it will not be a suitable protein for NMR studies since one three-dimensional NMR experiment usually requires 5 days to collect data.

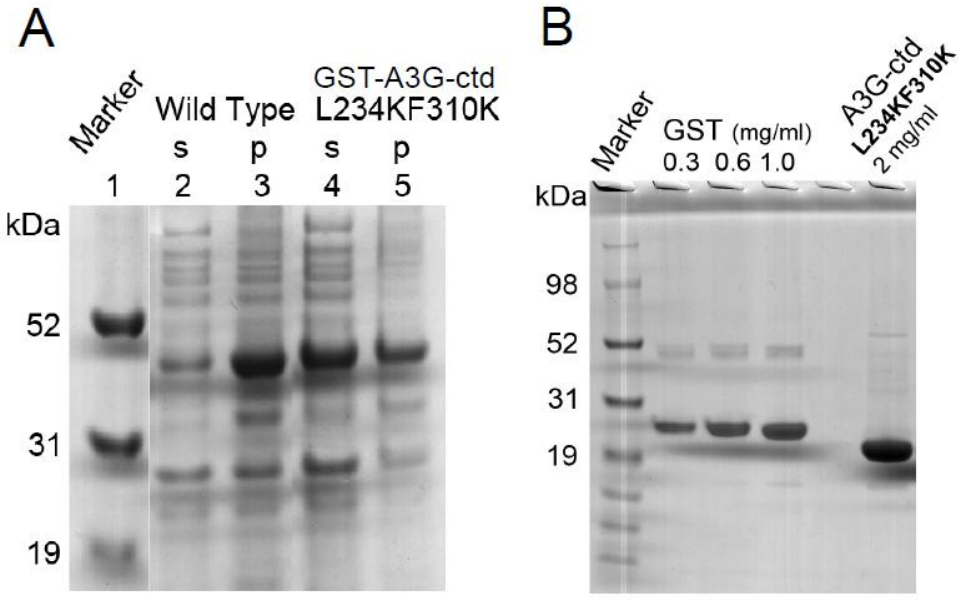


Figure 2.5. The biophysical properties of A3G-198-384-2K. (A) Solubility test of wild-type GST-A3G-198-384 (48 kDa) and GST-A3G-198-384-2K. (B) The SDS-PAGE gel of A3G-198-384-2K (22 kDa) after removing the GST tag.

2.3.7 Buffer optimization

Initially, the GST buffer (Tris-Cl 50 mM, NaCl 200 mM, pH 8.0, Zn 50 μM, Tween 20 0.5% and 5 mM DTT)) was used for above protein purification. However, to better stabilize the 2K mutant protein in solution, a buffer optimization was performed. Here, dialysis buttons (Hampton Research) were employed for buffer optimization.

Briefly, this method can determine the amount of protein remaining in the test buffer and thereby reveals buffer's capacity and protein's stability.

We first tested the stability of the 2K mutant in Tris-Cl, sodium phosphate and HEPES buffers at different concentration and pH (Figure 2.6 and data not shown). Our data clearly shows 50 mM sodium phosphate at pH 7.4 retains most of the protein. The term "salting in" is well-known; it describes how the solubility of a protein at low ionic strength (low salt concentration) generally increases with the increasing salt concentration. However, salt also decreases the signal intensity (or S/N ratio) in the NMR spectra. Surprisingly, the data shows the 2K variant is still stable without any added sodium chloride in the 50 mM sodium phosphate buffer at pH 7.4 (data not shown). Detergents are molecules that have the ability to cover the hydrophobic regions on the surface of the protein, and thereby increases the protein solubility. At detergent concentrations higher than the CMC (critical micelle concentration) the micelle will form in the solution, which again will result in a reduced NMR signal intensity in the case where the protein interacts with a large detergent micelle. Tween-20 is a non-ionic detergent with a molecular weight of 1228 and a CMC of 0.007% in water at room temperature. To reduce NMR signal line broadening, the concentration of Tween 20 can be decreased to 0.005% even though lowering the detergent concentration can reduce A3G-ctd-2K solubility. Therefore, the optimized buffer for the following NMR experiments consisted of Sodium phosphate 50 mM, Zinc 50 µM, pH 7.4, Tween 20 0.005% and 1 mM DTT.

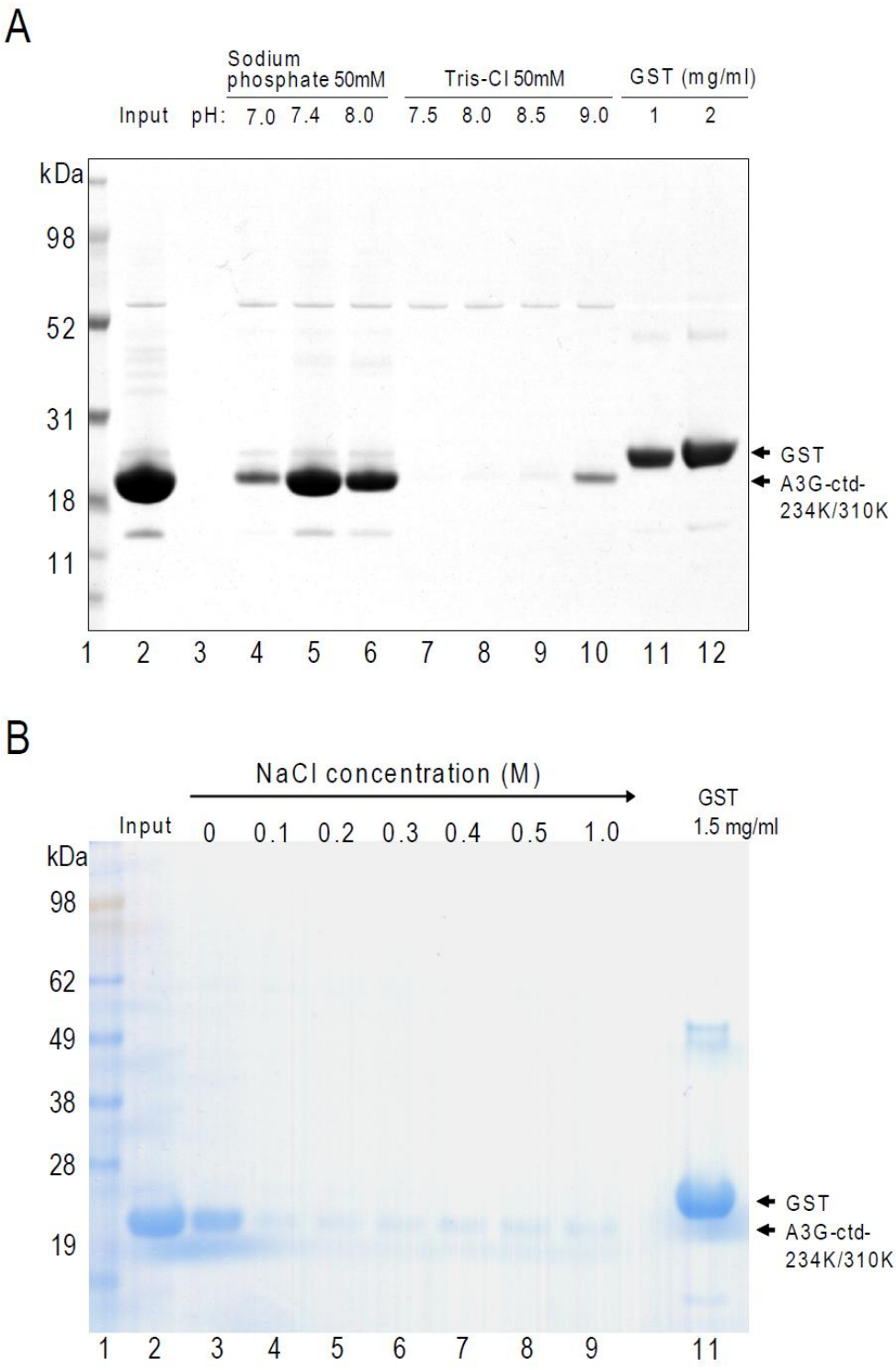


Figure 2.6. Optimization of buffer condition supports increased A3G-ctd-2K solubility. (A) The solubility of A3G-ctd-2K in sodium phosphate 50 mM and Tris-Cl 50 mM buffers at different pH values (lane 4-10) after two days incubation at 25 ℃. The input which is the protein at the start of the experiment (lane 2) and GST (lane 11 &12) are indicated. (B) The solubility of A3G-ctd-2K in sodium phosphate 50 mM, pH 7.4 with different NaCl concentrations (lane 3-9). The input which is the protein at time zero (lane2) and GST (lane11) are indicated.

2.3.8 The stable, soluble and enzymatically active APOBEC3G derivative: A3G-198-384-2K3A

Finally, based on the data in Figure 2.2, three additional non-detrimental substitutions [73], C243A, C321A and C356A, were added to A3G-198-384-2K construct to minimize the possibility of inter-molecular disulfide bond formation and maximize long-term stability (Figure 2.8). The resulting variant, A3G-2K3A, was 2.7-fold more active and 4-fold more soluble than the parental protein (Figure 2.7A, B) and could be storage at 25 ℃ for 2 weeks. Importantly, the DNA cytosine deamination activity of A3G-2K3A was fully dependent on the catalytic glutamate E259 (Figure 2.7A) [33, 35, 49]. Moreover, the NMR HSQC spectra of 2K and 2K3A (Figure 2.9), clearly shows the signal linewidth of 2K3A is sharper than that of 2K, with further analysis indicating that 2K3A is in a monomer state while the 2K protein is in a dimeric state.

2.3.9 A3G-2K3A is structurally intact and monomeric

Gel filtration assays have previously shown that the wild-type A3G198-384 is monomeric [73]. To confirm this finding and to assess the oligomeric state of A3G-2K3A, the parental protein and the five-substitution derivative were compared using circular dichroism (CD) spectroscopy. The CD spectra of A3G198-384 and A3G-2K3A virtually superimposed indicating that the five-substitution derivative had secondary structures like the wild-type protein (Figure 2.7C). Moreover, A3G-2K3A sedimentation velocity analytical ultracentrifugation profiles were nearly identical over a range of

concentrations, providing strong evidence that a monomer-dimer or higher order equilibrium is not occurring (Figure 2.7D). These sedimentation velocity data were also used to calculate an A3G-2K3A molecular weight of 22.3 kDa (considered to be within error of the theoretical value of 22.6 kDa).

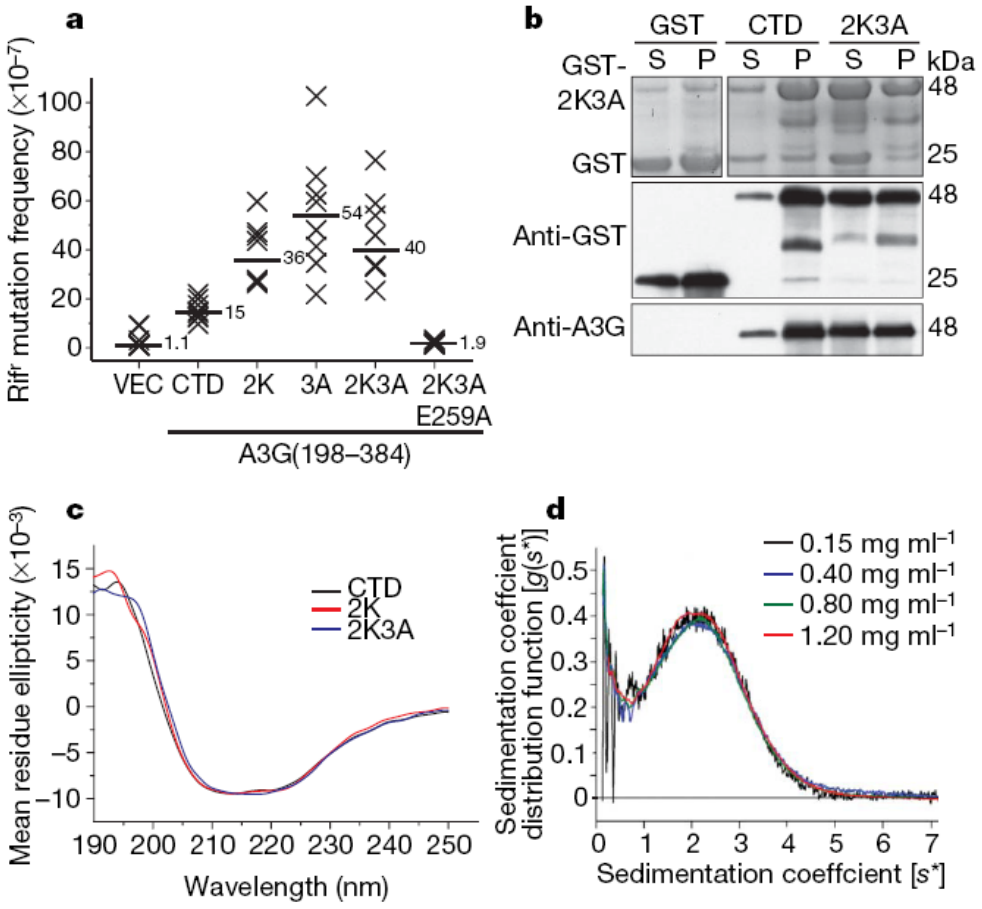


Figure 2.7. Functional and biophysical properties of A3G-2K3A. (A) Capacity of GST–A3G(198–384) (CTD), the indicated mutant derivatives or an empty vector control (VEC) to trigger Rif^R mutations in E. coli. (B) Solubility of GST, GST–A3G(198–384) (CTD) and GST–A3G-2K3A, as monitored by SDS–PAGE and coomassie blue staining (top panels) or immunoblotting (anti-GST middle panel and anti-A3G bottom panel). (C) Circular dichroism spectra of A3G(198–384) (CTD), 2K and 2K3A derivatives. (D) Sedimentation velocity analytical ultracentrifugation profiles for A3G-2K3A. The sedimentation coefficient distribution function g(s*) is shown for various concentrations of A3G-2K3A. The single peak of the g(s*) distribution indicates that A3G-2K3A is homogenous and monomeric.

2.4 Conclusion

Our experimental strategy provides a method to select for active and soluble variants of the catalytic domain of APOBEC3G. The deletion variants showed the minimal catalytic region required for deaminase activity is contained in residues 198-384. Size exclusion assays indicated this fragment to be monomeric. From the activity results of the alanine substitutions, we identified functional non-essential residues which, in most case, are located on the surface. Thirty-one out of 69 alanine derivatives retained significant to full levels of activity. By mutating these "non-essential" positions to lysine, three potential variants with enhanced solubility were identified. This systematic analysis of mutants yielded a highly soluble (400% more soluble than wild-type), highly active (3 times more active than wild-type) and highly stable (stable for 2 weeks at 25 ℃) variant of A3G-ctd. This mutant A3G198-384-2K3A is then being subjected to NMR experiment for determining the three dimensional structure of A3G-ctd. The rational for generating soluble and catalytically active variants of insoluble enzymes described in this thesis could be applied to other APOBEC family members that have deaminase activity.

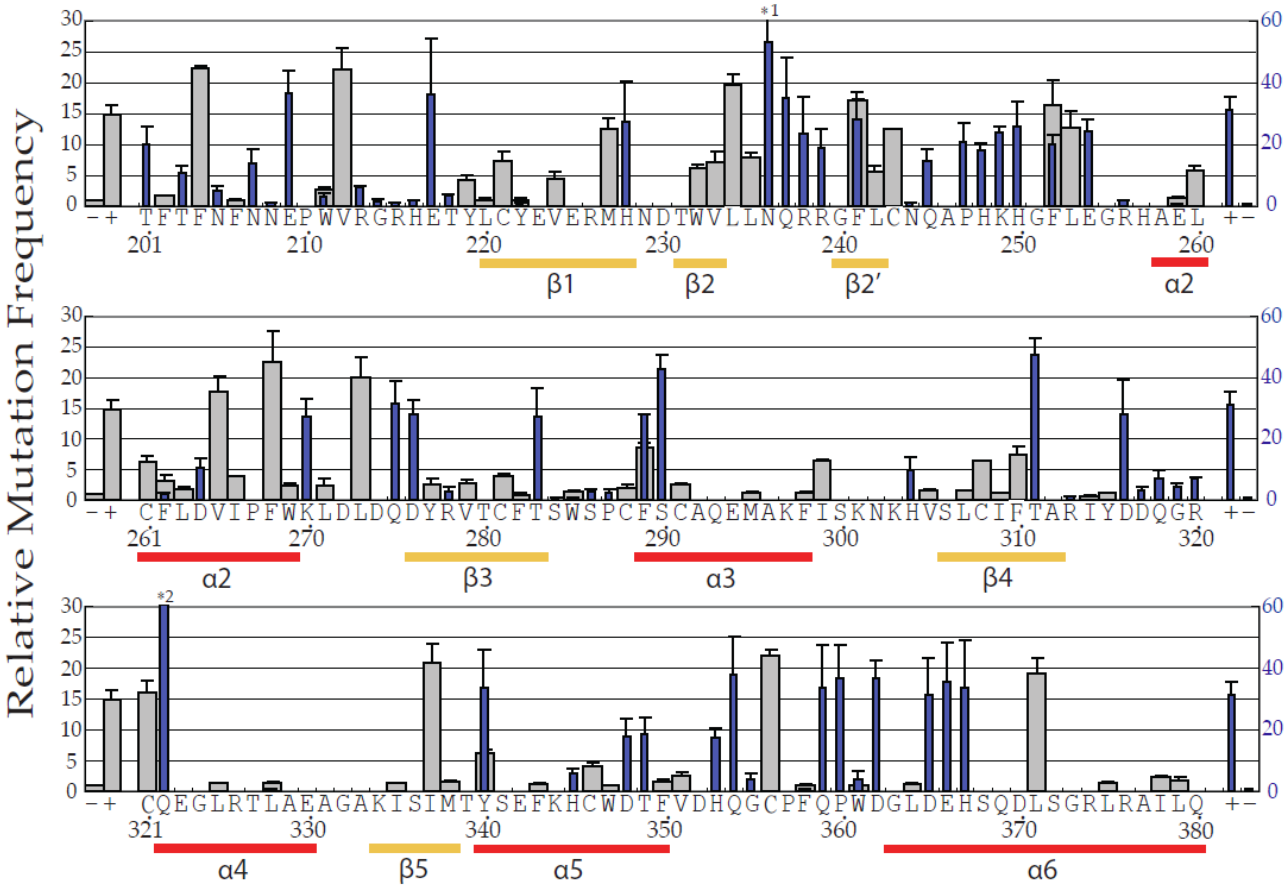


Figure 2.8. Rif^{R} mutation phenotypes of A3G198-384 and A3G-2K3A alanine substitution derivatives. Data for A3G198-384 and derivative constructs are represented by left Y-axis values and grey histogram bars. These 69 mutants were reported previously, but they are shown again here to facilitate comparisons with the A3G-2K3A alanine substitution data (represented by right Y axis values and blue histogram bars). Each histogram bar reports the average median mutation frequency and SEM from 2-5 independent experiments, except the bars for the controls (left Y axis: vector and A3G198-384, n=12; right Y axis: vector and A3G-2K3A, n=13). Histogram bars for two A3G-2K3A derivatives exceeded the right Y axis [N236A (53 + 19) and Q322A (61+12); indicated by asterisks]. The primary amino acid sequence matches wildtype A3G and the secondary structure is derived from the observed A3G-2K3A solution structure.

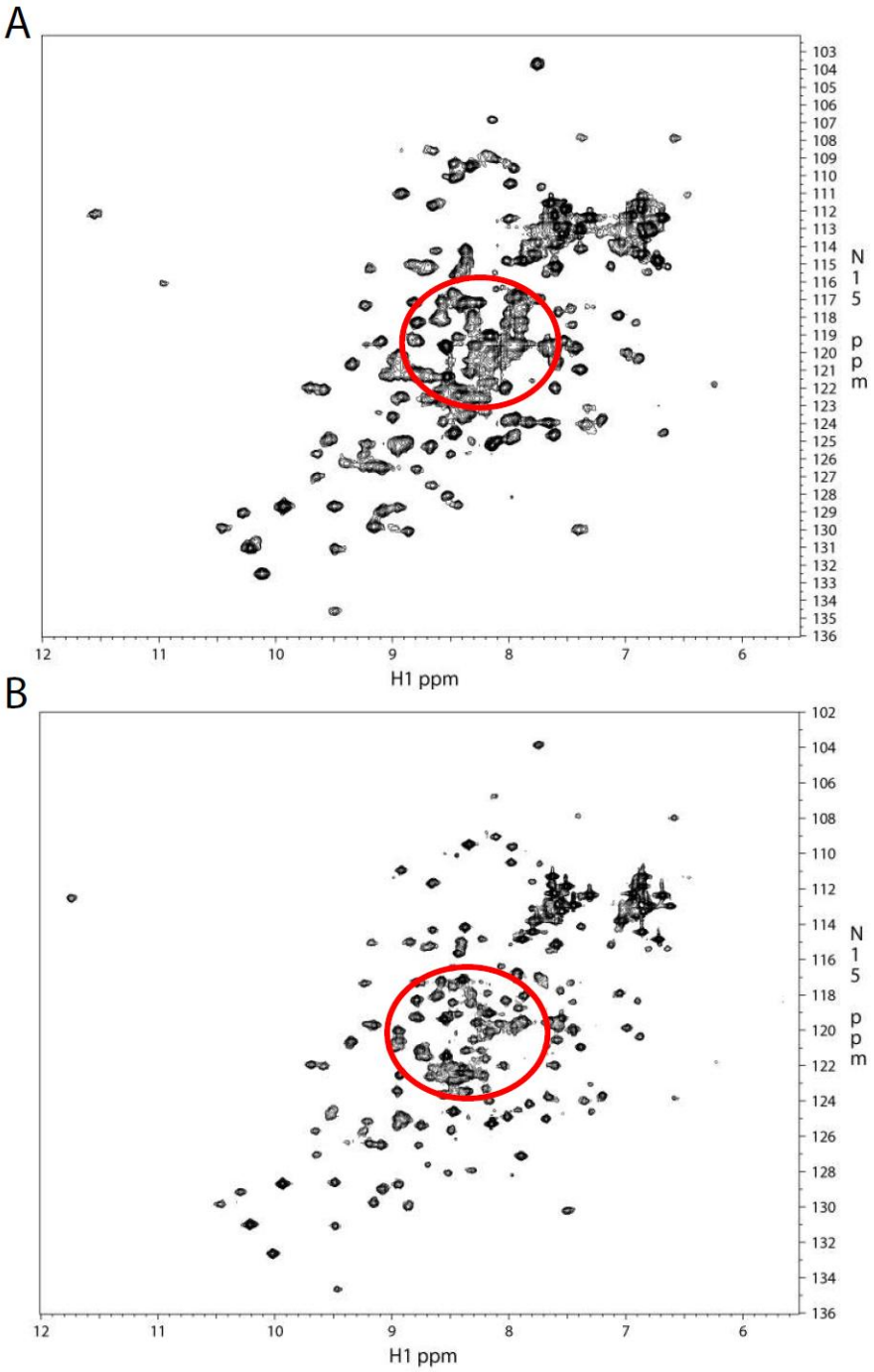


Figure 2.9. The (300 μM) HSQC spectra of 2K (A) and 2K3A (B) at 600 MHz at 15 ℃ show the linewidth of 2K3A is sharper than 2K. The linewidth analysis indicates the 2K protein is oligomerization while 2K3A protein is monomeric. The buffer of both samples is the optimized buffer (50 mM Sodium phosphate, 50 μM Zinc, PH 7.4, 0.005% Tween 20 and 1 mM DTT).

Chapter 3

Solution structure of the DNA deaminase Domain of APOBEC3G

3.1 Introduction

APOBEC3G is a member of a family of polynucleotide cytosine deaminases, several of which also target distinct physiological substrates, such as APOBEC1 editing APOB mRNA and AID deaminating antibody gene DNA. Although related structures of other family members exist, none of these proteins have demonstrated polynucleotide cytosine deaminase or anti-viral activity. Despite remarkable roles for DNA deaminases in adaptive and innate immune responses, none of these proteins has yielded a high-resolution structure. Current knowledge is derived mostly from the crystal structures of cytosine (free-base) [7] or cytidine (nucleoside) deaminases [5, 6, 8]. Some additional and important insights have been provided by the recent structures of APOBEC2 [18] and TadA (tRNA adenosine deaminase) [81]. However, APOBEC2 is a conserved vertebrate protein with no obvious biochemical or genetic function [17], and TadA is a transfer RNA adenosine deaminase which deaminates adenosine to inosine [82]. Therefore, the understanding of the deamination mechanism at an atomic level has been impeded as none of these structurally characterized enzymes deaminate polynucleotide DNA

cytosines.

In this study, we have determined the solution structure of the catalytic domain of human A3G with the active and improved solubility variant A3G198-384-2K3A. A complex model of the deaminase variant complexed with DNA was also generated to provide details of the deamination mechanism. We also describe the structure of the longer and more active construct A3G191-384-2K3A, and the differences in structure and the enzymatic activity of these two variants are discussed. Finally, the novel full-length A3G model was constructed based on the new elongated structure.

3.2 Materials and Methods

3.2.1 DNA constructs

The construction of GST expression vectors, pGEX6P1-A3G1-384 (full length), pGEX6P2-A3G198-384 and pGEX6P2-A3G191-384, was described in chapter 2 [73]. All amino acid substitution derivatives were constructed using the QuikChange protocol (Stratagene) and verified by DNA sequencing.

3.2.2 *E. coli*-based Rif^R mutation experiments

The detailed protocol is illustrated in chapter 2. Briefly, the *E. coli*-based Rif^r mutation assay has been used extensively to report the intrinsic DNA cytosine deaminase activity of A3G [21, 49, 73]. Each construct was expressed constitutively in *E. coli* strain BW310 (uracil-excision-defective), and at least eight independent cultures were used to determine the median Rif^r mutation frequency. Representative raw frequencies and median frequencies were shown (Figures 2.7A and 3.3D) or, to facilitate comparisons, the

median values were normalized and data from a minimum of two independent experiments were averaged (Figures 2.4& 2.8).

3.2.3 Protein purification and sample preparation

GST-based constructs were expressed in *E. coli* strain BL21 DE3 RIL (Stratagene). Unlabeled proteins were produced by overnight expression at 17 ℃ in LB medium containing 0.4 mM IPTG and 100 µg/mL ampicillin. Isotope-labeled proteins were produced by overnight expression at 25 ℃ in M9 supplemented with $^{15}NH_4Cl$, ^{13}C-labeled D-glucose and ^{2}H water as previously described [83]. Proteins were purified by sonicating cell pellets in lysis buffer [50 mM Na_2HPO_4/NaH_2PO_4 (pH7.4), 50 µM Zinc and protease inhibitor (Roche)], separating the soluble (supernatant) and insoluble (pellet) fractions by centrifugation (12,110g, 20 min, 4 ℃), binding to glutathione sepharose (GE Healthcare), washing with lysis buffer and eluting with PreScission protease (GE Healthcare) in 1mM DTT, 50 mM Na_2HPO_4/NaH_2PO_4 [pH 7.4] and, finally, concentrating with Centricon filters (Millipore) to desired volume and concentration.

3.2.4 NMR spectroscopy and structure determination

For signal assignments, NMR spectra of ^{13}C- and ^{15}N- labeled A3G-2K3A as well as amino acid specific ^{15}N-labeled Ala, Val, Leu, Phe, Tyr, Trp, Met, Cys, Thr, His, Lys or Arg A3G-ctd were used. The backbone ^{1}H, ^{13}C and ^{15}N resonances of the uniformly ^{13}C and ^{15}N labeled and 90% perdeuterated protein were assigned using triple resonance HNCA [84-86], HNCO [86], HNCACB [87], HNCOCACB [88], HNCOCA [84] and

HNCACO [89-91] experiments. The side chain assignments were completed using 3D CCONH [92], HCCH-TOCSY [93] and ^{15}N-, ^{13}C- edited NOESY-HSQC [94] with 80 ms mixing time. NOE-derived distance restraints were obtained from ^{15}N- or ^{13}C- edited NOESY-HSQC and 2D NOESY spectra acquired with the 200 ms (for ^{15}N-edited NOESY of perdeuterated protein), 150 ms (for ^{15}N-edited NOESY of non-deuterated protein) or 100 ms mixing time (for ^{13}C-edited NOESY and 2D NOESY). To collect NOEs between amide proton and methyl proton, aromatic proton and methyl proton, or aromatic proton and aromatic proton, these protons were selectively protonated in an otherwise fully deuterated protein.

NMR spectra were processed with NMRPipe [95] and analyzed with CARA [96]. For A3G-198-384-2K3A structure, 200 torsion angles restraints were taken from TALOS prediction [97]. 142 hydrogen bonds were set for residues consistent with the chemical shifts deviations and NOE pattern defined secondary structure. 1004 NOE distance restraints were picked manually from NOESY data, and 1004 additional NOEs were assigned using Atnos [98] /Candid [99] structure dependent cycles. The final calculation employed 242 intra-residue, 604 sequential, 506 medium-range and 656 long-range NOEs. 100 structures were calculated with CNS [100] torsion angle molecular dynamics. Ten of the calculated structures were chosen based on energy and Ramachandran plot to represent in the Figure 3.1 ensemble. NMR calculation statistics for A3G-198-384-2K3A and A3G-191-384-2K3A are summarized in Tables 3.1 and 3.2.

The position of Zn^{2+} in the A3G-2K3A structure was estimated based on how it is coordinated in existing deaminase superfamily member crystal structures [5-8, 81]. Table

3.3 lists the covalent bond lengths and angles that were used to link Zn^{2+} to H257, C288 and C291 in the A3G-2K3A structure.

Table 3.1. Detailed NMR statistics for A3G-198-384-2K3A (2jyw)

Distance Restraints		**Average RMSD**	
NOEs		from distance restraints	0.0191±0.0005
All	2008	from dihedral restraints	0.3626±0.0335
Intra-residue	242	from idealized geometry:	
Inter-residue	1772	bonds (Å)	0.0026±0.00004
Sequential (|i-j|=1)	604	angles (°)	0.4622±0.0124
Medium	506	impropers (°)	0.3531±0.0217
Long	656	average pairwise between 10 structures:	
Hydrogen bonds	142	backbone (Å) *	0.81±0.15
Dihedral angle restraints:		all heavy atoms (Å) *	1.6±0.17
ϕ	115	backbone (Å) **	0.55±0.14
ψ	115	all heavy atoms (Å)**	1.26±0.21
		Ramachandran plot appearance	
		Most favored regions (%)	66
		Additional allowed (%)	29
		Generously allowed (%)	3.4
		Disallowed regions (%)	1.6

Table 3.2. NMR refinement statistics of A3G-191-384-2K3A

Distance restraints	
NOEs	
All	2224
Intraresidue	259
Interresidue	1772
Sequential ($\|i-j\|=1$)	640
Medium	613
Long	756
Hydrogen bonds	145
Dihedral angle restraints	
ϕ	150
ψ	150
Average RMSD	
From distance restraints	0.0321±0.0010
From dihedral restraints	0.5424±0.0371
From idealized geometry	
Bonds (Å)	0.0030±0.0001
Angles (°)	0.4912±0.0096
Impropers (°)	0.3933±0.0091
Average pairwise between 10 structures	
Backbone (Å)[a]	0.86±0.18
All heavy atoms (Å)[a]	1.84±0.26
Backbone (Å)[b]	0.54±0.13
All heavy atoms (Å)[b]	1.26±0.15
Ramachandran plot appearance	
Most favored regions (%)	64
Additionally allowed regions (%)	27

[a] All residues, except loops between β2 and β2′ and loops between β2′ and α2, are included.
[b] All secondary structure elements are included.

Table 3.3. Constraints for Zn^{2+} positioning*

Covalent bonds			
Atom 1	Atom 2	Length (Å)	
H257 Nδ1	Zn^{2+}	2.0	
C288 Sγ	Zn^{2+}	2.3	
C291 Sγ	Zn^{2+}	2.3	
Bond angles			
Atom 1	Atom 2	Atom 3	Angle (degree)
H257 Cγ	H257 Nδ	Zn^{2+}	126
C288 Cβ	C288 S	Zn^{2+}	108
C291 Cβ	C291 S	Zn^{2+}	108

* Zn^{2+} molecule was constrained in the structure calculations of A3G-2K3A. Covalent bonds were created between Zn^{2+} and His 257, Cys288 or Cys291. The bond lengths and angles used for the calculation are listed.

3.2.5 NMR titration experiments

Non-labeled ssDNA was titrated into ^{15}N-labeled A3G-2K3A at protein:DNA molar ratios of 1:0, 1:1, 1:2, 1:4, 1:8 and 1:16. A heteronuclear single quantum coherence (HSQC) spectrum was recorded at each molar ratio, which enabled specific amino acid chemical shift perturbations to be detected. Chemical shift perturbation was calculated using the following equation $\Delta\delta = [(\Delta^1H)^2 + (\Delta^{15}N/5)^2]^{1/2}$ where Δ^1H and $\Delta^{15}N$ are differences in ^{1}H chemical shifts and ^{15}N chemical shifts, respectively.

3.2.6 Single-strand DNA binding model

The target cytosine was positioned in the catalytic active site (under H257 and adjacent to E259) based on existing crystal structures of active cytidine deaminases [101, 102]. This positioning fixed the target cytosine base and enabled calculations of all possible rotamer configurations of the 5'-C (C_1), the 3'-T (T_3) and all 25 of the amino acid side chains that showed significant NMR chemical shift perturbations (R215, T218,

Y219, C221, H228, L242, A246, E254, R256, A258, E259, V265, C281, N302, C308, R313, I314, Y315, D316, G332, I337, H345, W347, G355, and H367; *i.e.*, signals that were 1 SD above mean in Figure 3.7). All rotable bonds were varied (6 bonds per nucleotide) to obtain lowest energy configurations for both the DNA and the affected A3G-2K3A residues. Although these calculations resulted in a very large number of possible configurations, the number was efficiently reduced using the dead-end elimination method, which eliminates configurations of side-chains or nucleotides that are unlikely to be part of the global minimum structure [103, 104]. The reduced number of configurations was enumerated systematically to arrive at a minimum energy model.

3.2.7 A3G1-194 and full-length structural models

The full-length A3G model was constructed in three steps. In the first step, a model for A3G pseudo-catalytic domain residues 1–194 was generated based on amino acid sequence homology between the A3G pseudo-catalytic domain, the A3G catalytic domain, and A2 (Figure 3.6A) [105]. Sequence alignment was generated by pairwising A3G1-196 with A3G197-384, then with A2, using the National Center for Biotechnology Information algorithm BLAST 2 SEQUENCES set to the similarity matrix BLOSUM62. This similarity matrix uses amino acid side-chain and genetic code similarities, and it calculated 35% identity and 52% similarity for A3G1-196 and A3G197-384 alignment, and 29% identity and 46% similarity for A3G1-196 and A2 alignment. The actual A3G191-384 secondary structural elements from this study were superimposed on the alignment to delineate core residues (Figure 3.6A). Considering only these residues within the secondary structural elements, the percentages of identity and similarity were

recalculated to yield 49% identity and 62% similarity for A3G1-196 and A3G197-384 alignment, and 39% identity and 49% similarity for A3G1-196 and A2 alignment. These high levels of identity strongly indicated that the A3G1-196 model structure would closely approximate the pseudo-catalytic domain of A3G because homology models with 30–50% identity typically position greater than 90% of the backbone atoms within 1.5 Å RMSD of the real structure [106]. The A3G1-194 model was generated using a program called YASARA [105]. This program used multiple parameters to systematically generate an A3G1-194 model structure. First, each of the three most related structures—A3G191-384-2K3A [Protein Data Bank (PDB) ID code 2kem], A3G197-380 (PDB ID code 3e1u [107]), and A2 (PDB ID code 2nyt [18])—was aligned with the A3G1-194 primary and predicted secondary amino acid structures (Figure 3.6A). Second, YASARA accommodated insertions by searching the PDB for loops with start and end points that superimposed well with anchor points in the model, and then loops were further optimized to adopt the lowest-energy conformation. Third, amino acid side-chain geometry was refined using electrostatic and knowledge-based interactions. Fourth, unrestrained high-resolution refinement was performed, and one model was built for each related structure. Individual models were ranked based on a residue-for-residue score. Finally, the top model was taken, and lower-scoring regions were replaced iteratively with higher-scoring regions of the other two models. Hence, the resulting model resembled both the A3G catalytic domain (most regions) and A2 (β2 region). The pseudo-catalytic domain β2 region modeled as a continuous short strand because it shared little homology with the catalytic domain β2 region (Figure 3.6B and C). In the second step, residues 191–194 occurred in both the A3G1-194 model and the A3G191-384-

2K3A NMR structure; importantly, these residues form one helical turn in both structures. Therefore, these helical turns were simply superimposed to connect the N- and C-terminal domains and to render a full-length A3G model. In the third step, three loops were located close to the opposing domains, including residues 140–149 of the A3G1-194 model and residues 210–217 and 245–256 of the A3G191-384-2K3A NMR structure. Since these loops were not well determined in NMR solution structures or the A3G1-194 model, they were regenerated in the full-length A3G model using YASARA [105] to find the lowest-energy conformation. All structural schematics were made with PyMOL (DeLano Scientific) and labeled manually.

3.3 Results

3.3.1 NMR structure of the DNA deaminase domain of APOBEC3G

A3G198-384-2K3A was used for NMR spectroscopy experiments. A total of 2,004 distance constraints were obtained and used to calculate a solution structure (Figure 3.1, Tables 3.1 & 3.2). The superimposition of the ten lowest energy structures demonstrated that this enzyme has a well-defined core structure comprised of five β-strands and five α-helices, arranged from N- to C-terminus as β1-β2/2'-α2-β3-α3-β4-α4-β5-α5-α6 (Figure 3.1 A, B, C). The zinc-coordinating active site, α2-β3-α3, is anchored within the platform of β strands.

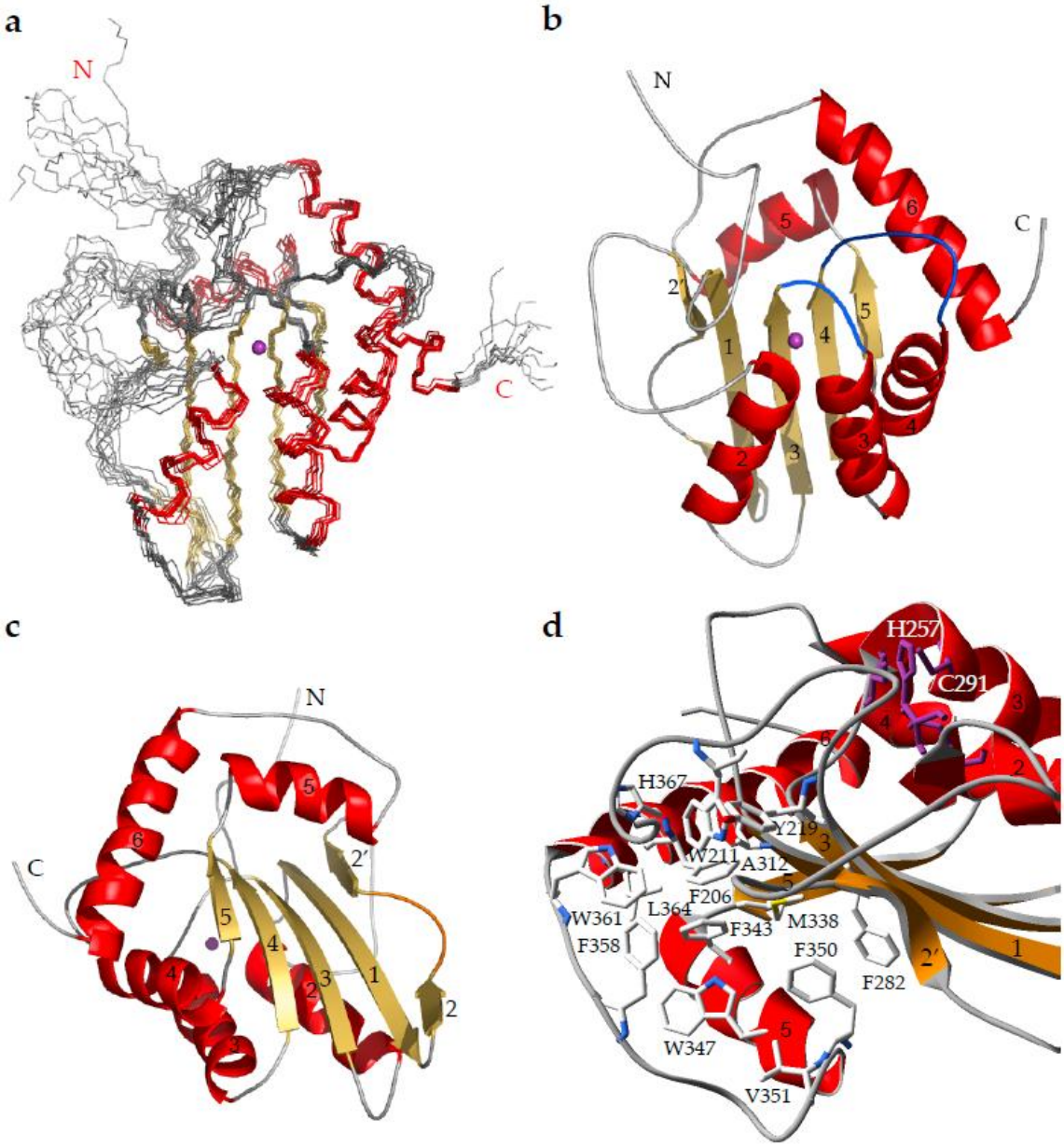


Figure 3.1. NMR structure of A3G-2K3A. A, Superimposition of ten NMR structures showing α-helices in red, β-sheets in yellow and Zn^{2+} in purple. B, C, Ribbon diagrams of the NMR structure shown in the same (B) and 180° (C) angle rotation, respectively. The β3-to-α3 and β4-to-α4 loops are colored blue in B, and the β2-bulge-β2' is colored orange in C. (D) Hydrophobic contacts between α5 and the β-strands and loops of the indicated regions (β1, β3, β4, N-terminal-loop, β3-α3-loop and β4-α4-loop). Amino acid side chain atoms are colored yellow (sulfur), red (oxygen), blue (nitrogen) and white (carbon). Zn^{2+}-binding side chains are colored purple.

The catalytic site is further supported by the α5 and α6 helices, which make extensive stabilizing hydrophobic contacts with the β strand platform (Figure 3.1D). The secondary structural elements are connected by loops of varying lengths, with the β3-to-α3 loop being remarkably well-defined (blue in Figure 3.1B). This loop consists of S284, W285, S286 and P287, residues that are conserved among DNA deaminases and likely important for active site integrity (Figure 3.9 & below).

The A3G catalytic domain shares some features with prior structures. First, the α-β-α Zn^{2+}-binding motif, α2-β3-α3 in A3G-2K3A, is the clearest structural feature of this deaminase superfamily (Figure 3.2, top) [5-8, 18, 81]. Second, a subset of the superfamily members, including human APOBEC3G, *S. aureus* tRNA adenosine editing protein TadA [81] and human APOBEC2 [18] (and likely all of the other APOBEC family members), has the β-strand of the zinc-coordinating motif and the two subsequent β strands arranged in parallel (Figure 3.2, bottom). As hypothesized previously, this organization is likely a key determinant of substrate specificity, enabling a loop and additional structural elements to be accommodated between the latter two β-strands [18, 73, 108, 109]. In contrast, cytidine deaminases of *E. coli*, *B. subtilis*, *S. cerevisiae* and humans have an anti-parallel β4-β5 organization separated by a small loop (Figure 3.2, bottom right) [5, 6, 8, 110]. Finally, closer family members, such as APOBEC2 [18], have a common overall fold and similar secondary structures (Figures. 3.2 & 3.3). Several prior reports have discussed and modeled this similarity [8, 18, 73, 108, 109, 111].

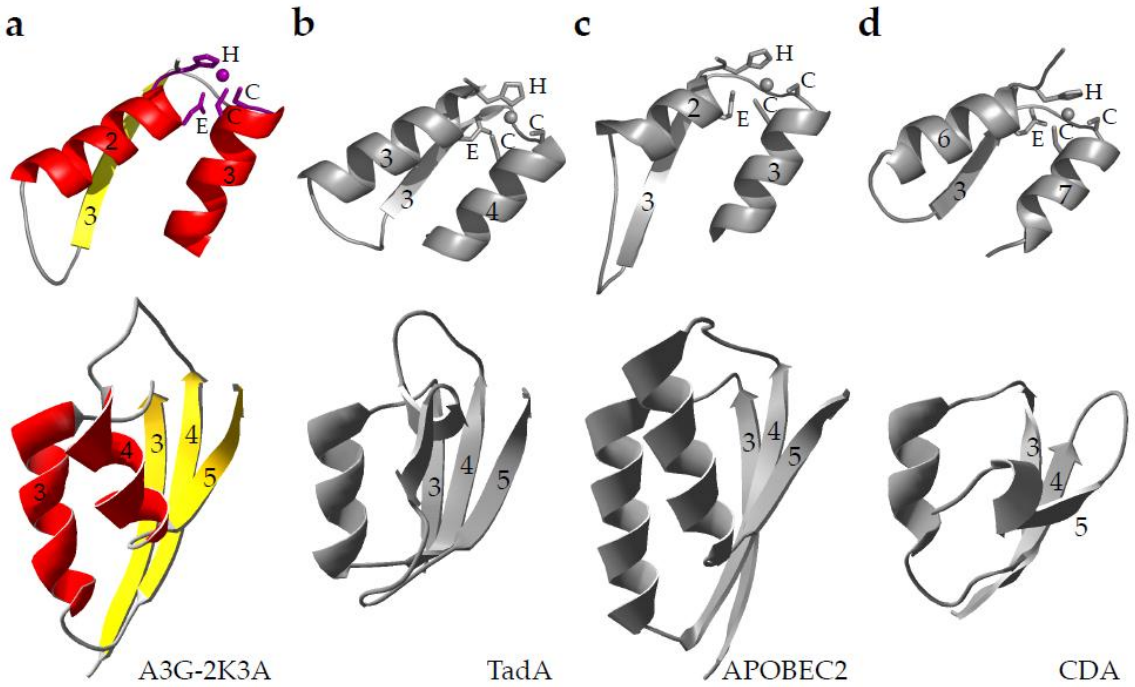


Figure 3.2. The relationship of the catalytic domain of A3G to selected family members. A–D, Human A3G (2jyw), S. aureus TadA (2b3j), human APOBEC2 (2nyt) and E. coli cytidine deaminase (CDA, 1ctu), Zn^{2+}-binding motifs (top row) and β-strand organization (bottom row). The amino acid side chains of the catalytic glutamic acid (E) as well as the Zn^{2+}-binding histidine (H) and cysteines (C) are indicated.

However, A3G-2K3A differs significantly from all prior structures. For instance, the closest family member with structural information [18], APOBEC2, shares 31% identity overall (Figure 3.4). As inferred previously [8, 18, 73, 108, 109, 111], the majority of these residues are located within the protein core (35/86 residues), consistent with the likelihood that these amino acids are critical for forming the overall scaffold (Figure 3.3). In contrast, much less identity (11/68 residues) occurs among solvent-accessible residues, which mediate substrate recognition, catalysis and interactions with other macromolecules (Figure 3.3). This makes sense in light of evolution, as more than 400

million years have passed since these two proteins were encoded by a single gene (prior to vertebrate radiation) [108]. Thus, as described below, the A3G-2K3A structure will help us understand why APOBEC3G and other family members (but apparently not APOBEC2) [18, 21, 40, 108, 112] are endowed with DNA cytosine deaminase and retrovirus restriction activities.

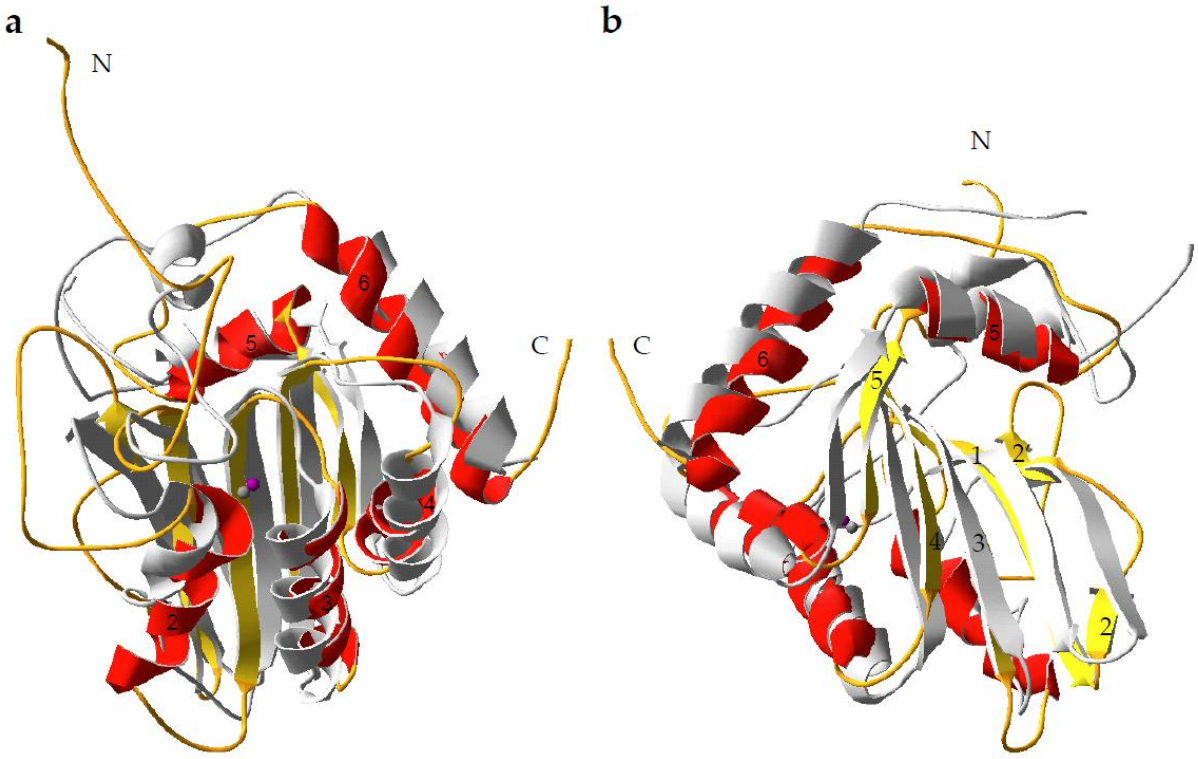


Figure 3.3. A comparison of the ribbon structures of A3G-2K3A and APOBEC2 (PDB 2NYT1). (A) view of the zinc-coordinating faces. (B) view of the β-sheet cores [rotated approximately 180 degrees around the central axis relative to (A)]. A3G-2K3A secondary structures and termini are colored and numbered. Cα atoms were superimposed to find the minimum r.m.s.d. using the Swiss-Pdb viewer (expasy.org/spdbv/).

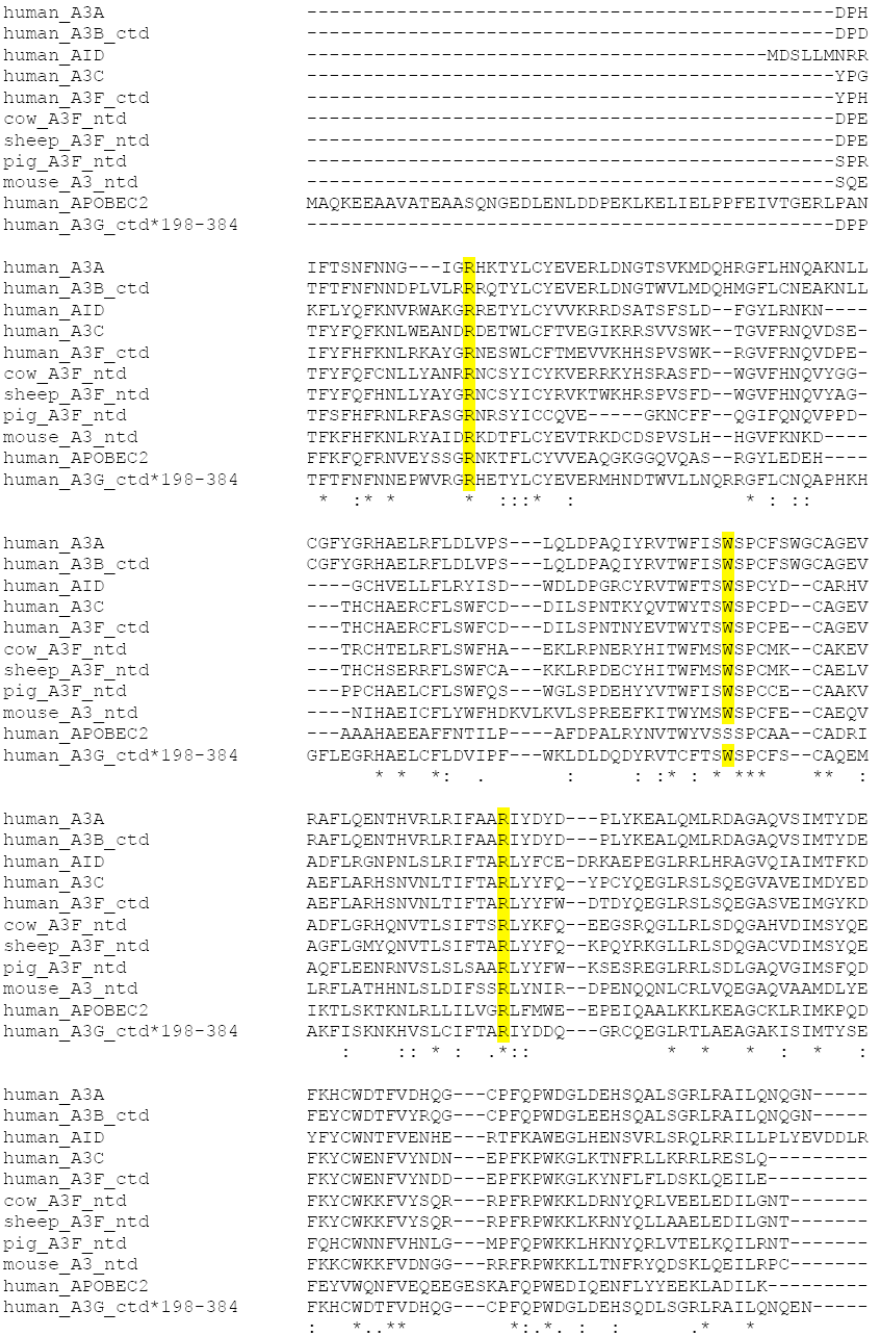


Figure 3.4. Alignment of A3G198-384 with selected mammalian APOBEC family members (Clustal W). Residues corresponding to A3G R215, W285 and R313 are highlighted together with homologous residues of family members. Accession numbers: human A3A (NP_663745.1), human A3B (NP_004891.3), human AID (NP_065712.1), human A3C (NP_055323.2), human A3F (AAH38808.1), cow A3F (NM_001077845.1), sheep A3F (NM_001093784.1), pig A3F (NM_001097446.1), mouse A3 (NP_084531.1), human APOBEC2 (NP_006780.1), and human A3G (NP_068594.1). Eight amino acids from the C-terminus of human AID are not shown because they go beyond the aligned region.

In addition to surface residue differences, A3G-2K3A has several remarkable structural features. First, A3G-2K3A (or a derivative with L234 restored) has a unique β2 strand, which is interrupted with a bulge of 6 residues (Figures 3.1C, 3.3 & 3.5). In contrast, APOBEC2 has a continuous 11 residue β2 strand, which mediates dimerization through a β2 strand of another molecule [18]. The structural constraints imposed by the β2-bulge-β2' suggest that different contacts will connect N- and C-terminal domains of APOBEC3G. Alternatively, the β2-bulge-β2' may mediate interactions with RNA and/or other proteins (of cellular and/or viral origin), because it appears largely dispensable for DNA deamination activity (Figures 2.4 & 2.8) [73]. However, additional data will be needed to fully discount the possibility that the β2 bulge has a different conformation in the context of the full-length protein. Second, A3G-2K3A begins with β1, whereas APOBEC2 has a small α-helix preceding its first β-strand [18]. Amino acid alignments suggest that residues 198-202 of APOBEC3G may form an analogous α helix (ExPASy proteomix tools, http://ca.expasy.org/), but this awaits experimental confirmation. Finally, there are many less obvious differences between A3G-2K3A and APOBEC2 (Figures 3.3 & 3.4). For instance, the zinc-coordinating α2 helix in A3G-2K3A is considerably longer than the corresponding helix in APOBEC2, and the conserved S-W285-S motif in APOBEC3G and other DNA deaminases is an S-S-S motif in all known APOBEC2 proteins. Given the prominence of W285 within the APOBEC3G catalytic site (discussed further below), it is likely that the S-S-S motif of APOBEC2 contributes to this protein's substrate specificity.

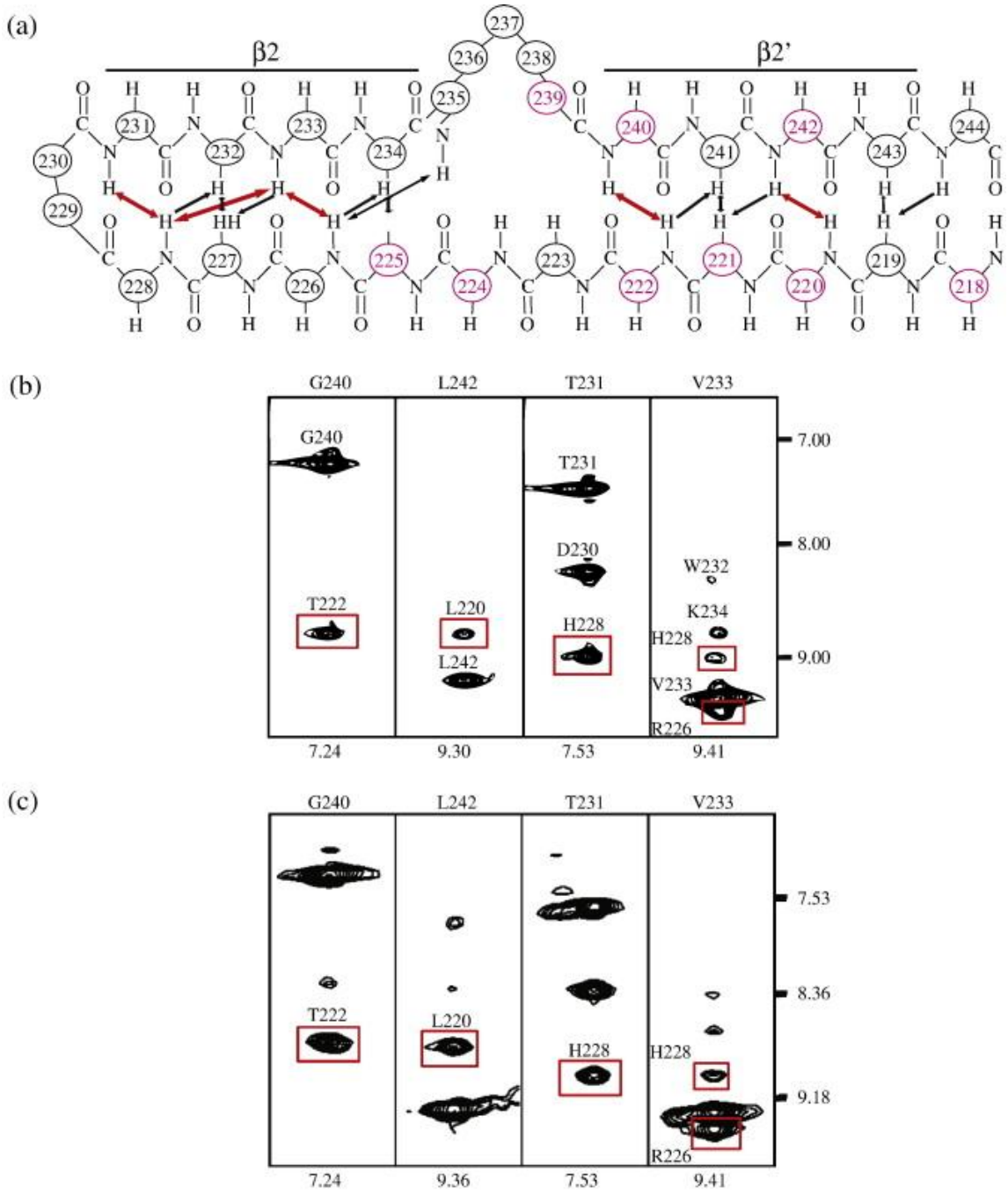


Figure 3.5. NOESY spectra showing that neither K234 nor A243 affects the β2–loop-like-bulge–β2′ secondary structure. (A) A schematic diagram of the interactions detected between β1 and β2/β2′. Red arrows represent NOE interactions observed in NMR spectra and correspond to the boxed signals in (B) and (C). Black broken arrows represent observed NOE interactions not shown in (B) and (C). (B) Representative strips of the ^{15}N-edited three-dimensional NOESY spectrum of A3G-2K3A showing NOE signals between β1 and β2/β2′. (C) Representative strips of the ^{15}N-edited three-dimensional NOESY spectrum of A3G-1K2A (both K234 and A243 were reverse engineered to L234 and C243, respectively) showing NOE signals nearly identical with those in (B).

3.3.2 Interaction Model of A3G-2K3A and DNA complex

A fundamental question is how APOBEC3G and related family members recognize single-strand DNA (ssDNA). Like many other nucleic acid-interacting proteins, we imagined that A3G-2K3A would have a prominent positively charged surface that would define the DNA interacting region. However, the electrostatic potential of the active site face of A3G-2K3A was largely negative apart from a few positively charged residues arranged on an apparent brim surrounding the concave active site region (Figure 3.6A). To directly test whether any of these residues interacted with DNA, NMR chemical shift perturbation experiments were conducted with ^{15}N-labeled A3G-2K3A and varying concentrations of a 21 base ssDNA oligonucleotide, which contained an APOBEC3G 5'-CC deamination hotspot. As expected, significant chemical shift perturbations occurred predominantly on the active site face of A3G-2K3A (Figures 3.6B & 3.7). Notable perturbations were detected for conserved arginines R215 and R313 and for the catalytic glutamate E259. Residues adjacent to R313 (within the β4-to-α4 loop) and E259 also showed strong chemical shift perturbations. The two other brim domain arginines, R213 and R320, could not be detected with this technique.

The NMR titration data were used to build a model for ssDNA binding (Figures 3.6B, C & 3.7). First, we selected an APOBEC3G hotspot containing trinucleotide 5'-C_1-C_2-T_3-3' to model the DNA interaction. This short sequence was selected because ssDNA interactions were detected predominantly around the active site and this sequence spans that region. Second, the target cytosine (C_2) was positioned under H257, analogous to how it orients in cytidine deaminase crystal structures [101, 102].

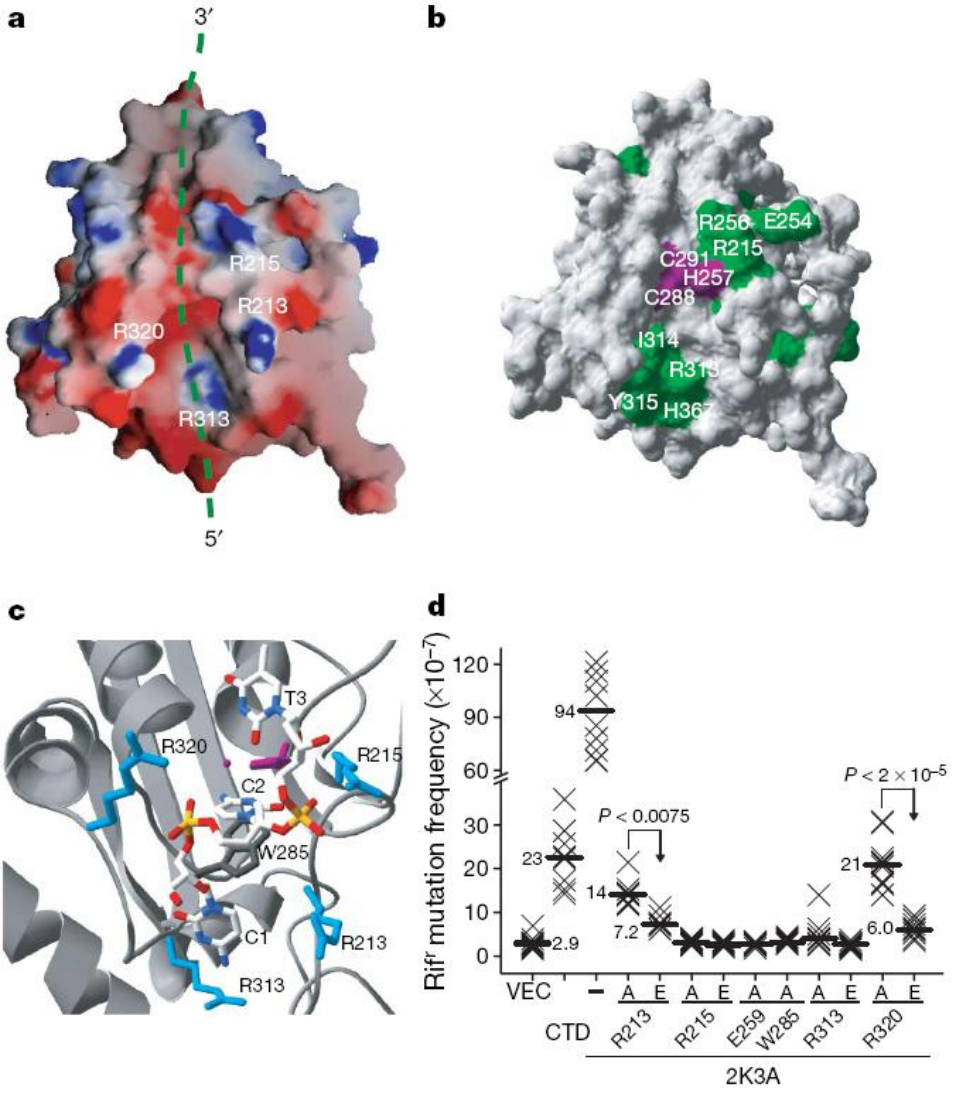


Figure 3.6. A3G catalytic domain DNA interaction model. (A) Surface representation of A3G-2K3A, highlighting positions of positive (blue), negative (red) or neutral (white) charge. The hypothesized position and polarity of ssDNA is indicated (green dashed line). (B) NMR ssDNA-titration data summary. Residues with chemical shift perturbations more than 1 s.d. above average are colored green (E259 is perturbed but hidden by H257). H257, C288 and C291 are shaded purple. (C) Model depicting the interaction between A3G-2K3A and ssDNA (5'-C1-C2-T3-3'). H257 (purple) is shown partially stacked with the ring of the flipped-out target cytosine (C2). W285 (grey) helps to form a hydrophobic catalytic cavity. Arginines surrounding the positively charged brim of the active site are indicated. The ssDNA is colored white (carbon), blue (nitrogen), red (oxygen) and yellow (phosphate). (D) DNA deaminase activity of A3G-2K3A derivatives. The significance of the A versus E substitution at R213 or R320 is indicated.

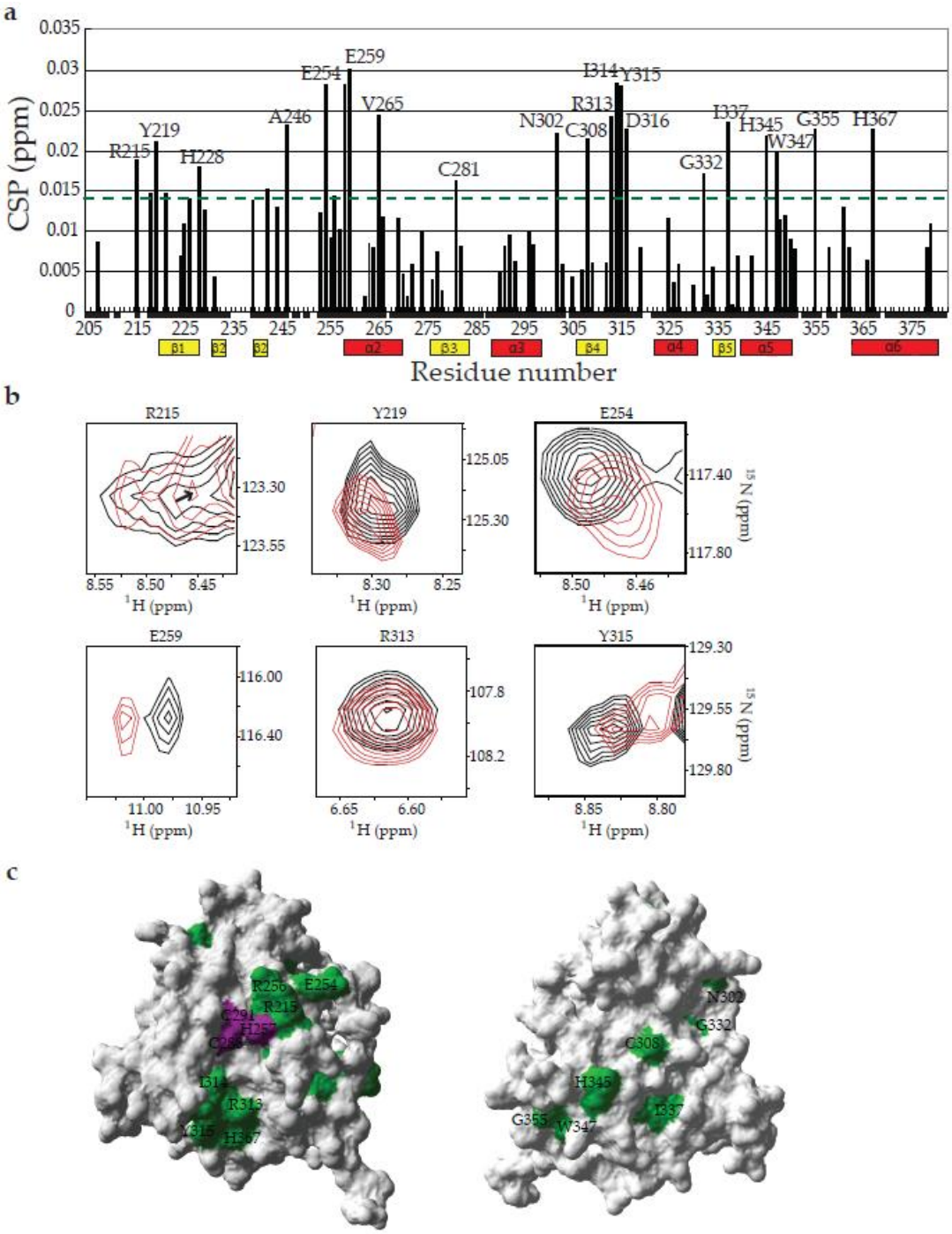


Figure 3.7. A3G-2K3A NMR chemical shift perturbations caused by ssDNA. (A) NMR chemical shift perturbation data for A3G-2K3A incubated with a ssDNA (5'-GCT TCT TCT ACC TTC TCT TGA-3') at a molar ratio of 1 protein to 4 ssDNA. The chemical shift perturbations (parts per million) are shown for the underlined residues of A3G-2K3A. A schematic of the observed secondary structure is shown for reference. (B) representative A3G-2K3A NMR signal shifts caused by ssDNA. Red and black plots are with and without ssDNA, respectively. (C) ssDNA induced A3G-2K3A NMR chemical shift perturbations occurred predominantly on the active site side of the protein (left panel, identical to Figure 3.6B) and not on the opposite side of the protein (right panel).

Finally, we used all residues that showed significant chemical shift perturbations to calculate the lowest energy structure of an A3G-2K3A-trinucleotide 5'-C_1-C_2-T_3-3' complex (Figure 3.6C).

One notable feature of the DNA binding model is that the target cytosine is predicted to be flipped-out from the phosphodiester backbone (*i.e.*, without flipping it cannot access the catalytic glutamate E259). A similar substrate contortion was described previously for TadA [81] and a number of other DNA metabolism proteins. The model further predicted that the 5' nucleotide C_1 would be sufficiently close to interact with conserved R313. C_1 has a large interaction surface that contributes significantly to the overall trinucleotide binding energy (-44.7 kcal mol^{-1}). This strong interaction may help explain the observed specificity of APOBEC3G for 5'-C<u>C</u> dinucleotides, which underlies the retroviral genomic strand 5'-<u>G</u>G to -<u>A</u>G hypermutation bias [113]. We hypothesize that DNA deaminases with different dinucleotide preferences such as AID (5'-R<u>C</u>) or APOBEC3F (5'-T<u>C</u>) will make similarly robust contacts with the 5' nucleotide. The model also predicted that the phosphate of the 3' nucleotide T_3 would contact both R215 and R213 and that the C_2 phosphate would interact with R320.

To test this brim domain model for DNA binding, we first asked whether conserved residues would be required for activity. The model predicted that R215 and R313 would promote DNA binding, W285 would help form the hydrophobic active site and E259, as shown previously, would mediate catalysis. As expected, all of these residues proved essential for activity (Figures 3.6D & 2.8). Second, since R213 and R320 were predicted to interact with the phosphate backbone of ssDNA, we hypothesized that they would be influential but non-essential for activity. Accordingly, a non-invasive

substitution at these positions might be tolerated, but a negatively charged substitution might render the protein inactive by repelling the phosphate backbone. Indeed, R213A and R320A derivatives still retained 20% of wild-type activity, whereas R213E and R320E derivatives were nearly dead (Figures 3.6D & 2.8). Thus, the A3G-2K3A solution structure, NMR DNA titration data, computational modeling, phylogenetic conservation and DNA cytosine deaminase activity data combined to support the brim domain model for ssDNA binding.

The catalytic domain of the HIV-1 restriction factor APOBEC3G represents the first high-resolution ssDNA deaminase structure. This structure will facilitate studies on related proteins such as the mRNA editor APOBEC1, the antibody gene deaminase AID and other family members that elicit retroelement restriction activity. Although our data strongly support a novel model for DNA interaction, future studies are necessary to understand all of the molecular contacts and to explain major substrate differences such as the specificity for ssDNA over RNA (and *vice versa* for other family members such as APOBEC1). As a practical consideration, we anticipate that similar mutagenesis strategies may be used to improve the solubility of other family members. Moreover, the A3G-2K3A structure may be used to build accurate models of the N-terminal, Vif-interacting domain of APOBEC3G and therefore also models of the full-length protein. The structure presented here therefore provides a crucial step toward a molecular definition of the A3G-Vif interaction, which will benefit the development of AIDS therapeutics that function by modulating this battle between host and pathogen.

3.3.3 A longer and more active A3G catalytic domain variant: A3G191-384-2K3A

To better understand the A3G catalytic domain and how it is oriented with respect to the N-terminal pseudo-catalytic domain, we revisited prior data showing that A3G175-384 was approximately threefold more active than A3G198-384 in an *Escherichia-coli*-based rifampicin-resistant (Rif^R) mutation assay [73]. To better define the residues responsible for this elevated mutation frequency, we examined a series of single-alanine mutants and observed that residues 175–190 were dispensable (data not shown).
We next generated untagged and glutathione *S*-transferase (GST)-tagged A3G191-384 and A3G198-384 constructs for extensive head-to-head genetic, biochemical, and structural characterizations. Derivatives with five amino acid substitutions, L234K, C243A, F310K, C321A, and C356A (2K3A), were also made to enable concentrated protein preparations for NMR studies [114]. Similar to our prior observations with A3G175-384 [73], A3G191-384 was threefold more active than A3G198-384 under noninduced basal expression conditions (Figure 3.8). 2K3A derivatives of these constructs had even higher Rif^R mutation frequency increases, but the relative activity difference between longer and shorter constructs was still threefold. Immunoblots showed that the A3G191-384 and A3G198-384 constructs were expressed similarly at noninduced levels, indicating that the improved activity of the longer protein is not simply due to higher expression levels or improved solubility (Figure 3.8, lower panel).

3.3.4 The α1-helix in A3G191-384-2K3A contributes to active-site stability

Our original NMR studies of A3G198-384-2K3A did not resolve residues 198–210 [114], but the subsequent A3G197-380 crystal structure showed that residues 201–206

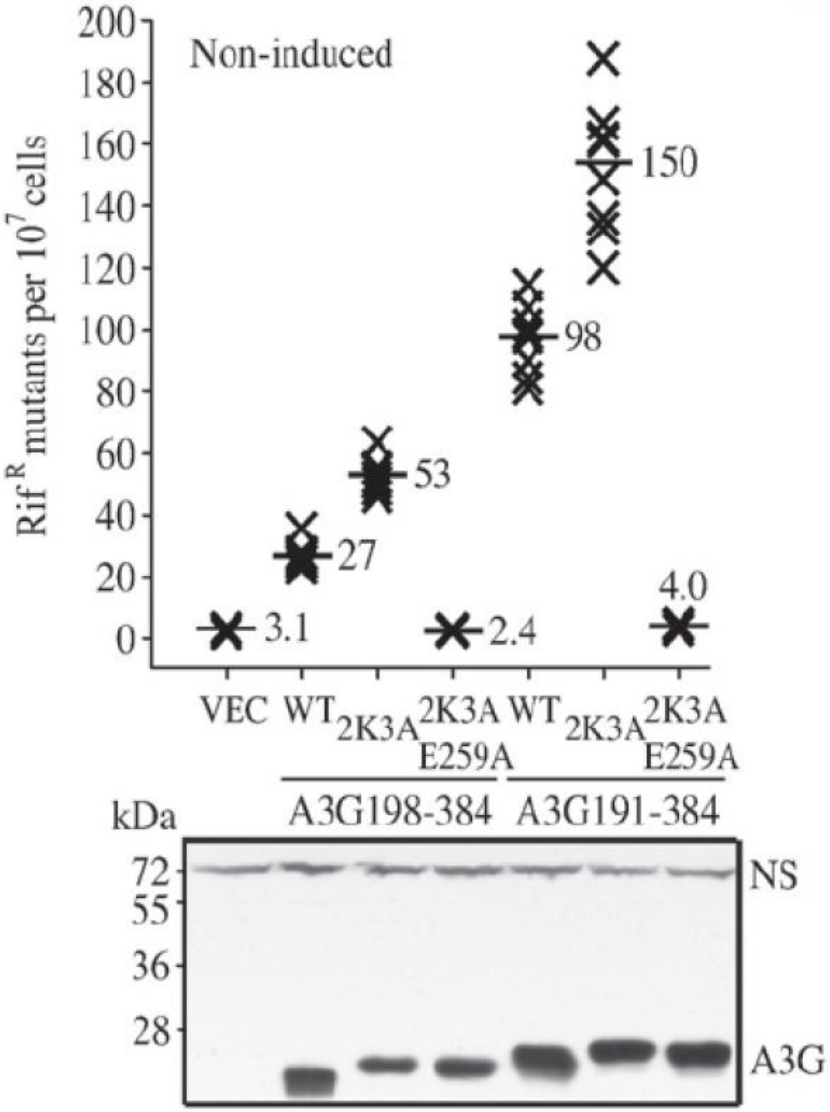


Figure 3.8. *E. coli*-based DNA cytidine deaminase activity of A3G catalytic domain constructs. The capacity of untagged A3G198-384, A3G191-384, and indicated mutant derivatives to trigger RifR mutations in E. coli growing under noninducing conditions. Each X represents the mutation frequency of an independent culture, and median values are indicated. VEC and WT represent empty vector control and the indicated construct with no additional mutations, respectively. A representative anti-A3G immunoblot is shown, and nonspecific (NS) bands provide loading controls. 2K3A derivates migrate slightly slower.

were capable of forming an α-helix [107]. Since the inability to resolve this region in our original NMR experiments correlated with diminished A3G198-384-2K3A activity in *E. coli* and especially *in vitro*, we hypothesized that some additional structure in A3G191-384-2K3A (such as the α1-helix) would be directly responsible for the elevated DNA cytidine deaminase activity. We also considered an alternative hypothesis that one or more of the 2K3A mutations destabilize the α1 region.

The solution structure of A3G191-384-2K3A showed that residues 201–206 form the α1-helix of the C-terminal domain (Figures 3.9A and B). Since 2K3A mutations are present in both A3G191-384 and A3G198-384 constructs, it was clear that the 2K3A substitutions are not responsible for the lack of α1-helix in the 198–384 NMR structure. The α1-helix is positioned anti-parallel with α5, in full agreement with the A3G197-380 crystal structure [107]. This positioning was supported by strong nuclear Overhauser enhancement (NOE) signals between the amide protons of T201/F202/T203 at the beginning of α1 and the methyl protons of V351 at the end of α5. The NMR data also indicated that F202 has hydrophobic contacts with V351. NOEs were also detected between the amide protons of N208/E209 and $H\varepsilon^1$ of W361 immediately proximal to α6. These data were in good agreement with the A3G197-380 crystal structure, which showed hydrophobic contacts between N208 and W361 and a hydrogen bond between $O^{\delta 1}$ of N208 and $H\varepsilon^1$ of W361. These α1 interactions are crucial for DNA deamination because alanine substitutions at these positions were shown either to ablate (F202A, N208A, V351A, and W361A) or to compromise (T201A and T203A) catalytic activity in *E. coli.* [73, 115].

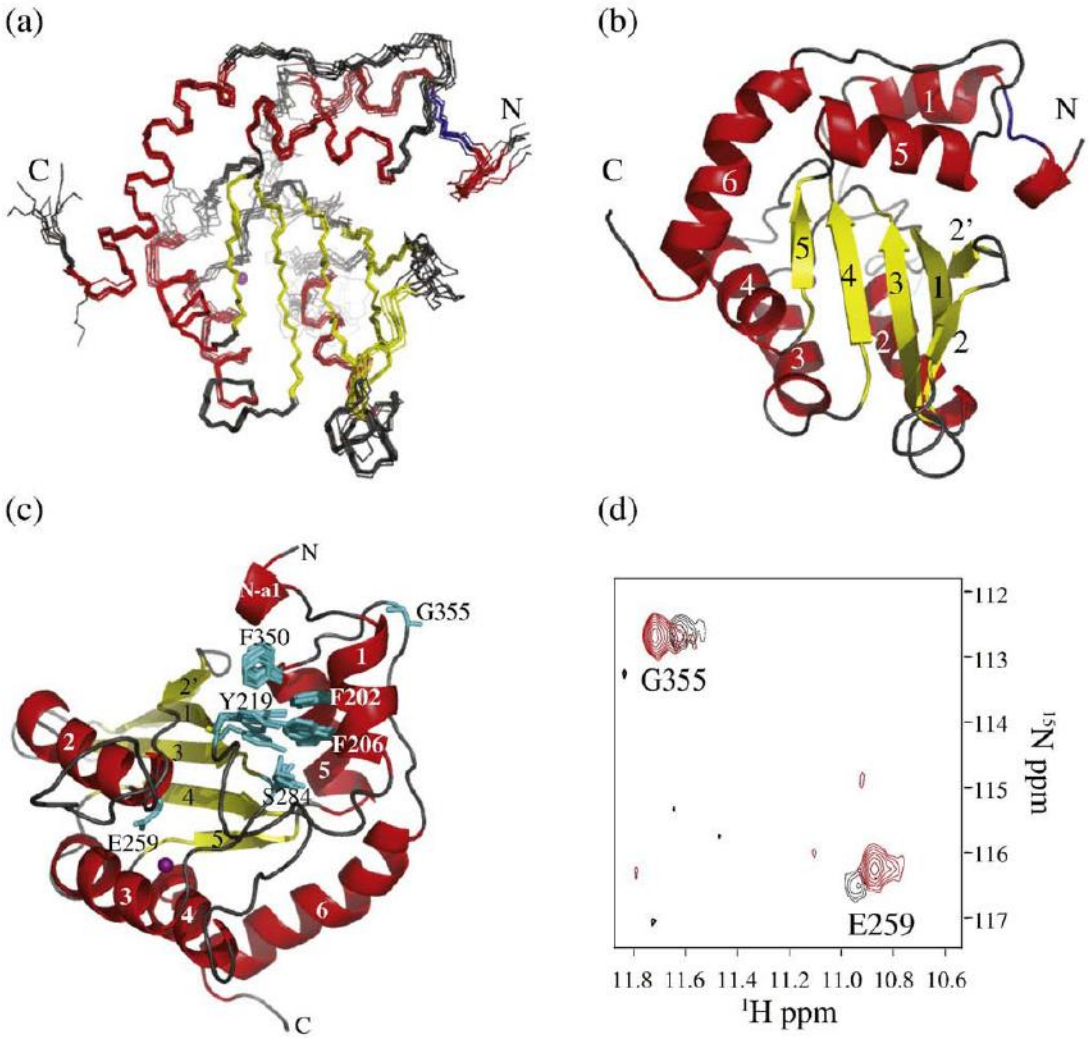


Figure 3.9. NMR structure of A3G191-384-2K3A (PDB ID code 2kem). (A) A superimposition of 10 NMR structures showing α-helices in red, β-sheets in yellow, and Zn^{2+} in purple. Residues 195–200 are shown in blue to highlight their well-defined structure. (B) A ribbon diagram of the NMR structure shown in (A) from the same angle. Since our prior work had detected only five of six α-helices, the helices were renumbered α1 (201–206), α2 (258–269), α3 (289–301), α4 (321–330), α5 (340–351), and α6 (363–380). (C) A schematic of the hydrophobic contacts involving F202, F206, and F350. The side chains of E259 and G355 are also labeled. (D) 1H–15N correlation NMR signals of G355 and E259 of A3G191-384-2K3A (in red) and A3G198-384-2K3A (in black).

Further direct comparison of the A3G191-384-2K3A and A3G198-384-2K3A solution structures indicated that several residues within α1 help stabilize the catalytic core and thereby enhance DNA deaminase activity. For instance, F202 and F206 make hydrophobic contacts with Y219 located at the beginning of β1 (Figure 3.9C). Additionally, F202 helps stabilize the conformation of the active site through hydrophobic contacts with F350 of α5, and the side chain of F206 contacts S284 within the conserved S284-W-S-P-C288 motif (Figure 3.9C). All of these residues are conserved and required for full levels of catalytic activity [73, 114]. Moreover, as anticipated, we have also found that F202 and F206 are essential for the activity of the full-length A3G holoenzyme (see the text below).

The important contributions that α1 makes to catalytic site stability are evidenced by the intensity of the NMR signal of the catalytic glutamate E259, which is located within the substrate-binding pocket. The E259 signal is weak in the A3G198-384-2K3A NMR spectrum, indicating that this residue is undergoing chemical exchange (Figure 3.9D, black). In contrast, the E259 signal is much stronger in the NMR spectrum of A3G191-384-2K3A, indicating increased stability (Figure 3.9D, red). G355 yielded similarly intense NMR signals in each protein, indicating that G355 is unaffected by α1 stability and that similar amounts of labeled protein were used in each NMR experiment.

3.3.5 A model for full-length A3G

Residues 191–194 resolved as a single helical turn in the A3G191-384-2K3A NMR structure (Figure 3.9A and B). Amino acid sequence alignments with the C-terminal domain of A3G and A2 predicted that these residues constitute part of the last helix of the N-terminal pseudo-catalytic domain N-α6, which is predicted to span residues 177–194 (Figure 3.10A; also see Stenglein *et al.* [24]. The A3G191-384-2K3A NMR data also indicated that residues 195–200 form a stable loop (bridge-like structure) that connects the final helix of the pseudo-catalytic domain to the first helix of the catalytic domain (i.e., N-α6 to C-α1). The stability of the bridge may be mediated, in part, by an interaction between L193 and G240/F241, as NOEs were detected between the H^{δ} of L193 and the amide protons of the latter residues.

The well-defined atomic coordinates of the N-terminal residues (residues 191–196) of the A3G191-384-2K3A NMR structure enabled us to assemble a novel full-length A3G structural model consisting of the predicted pseudo-catalytic domain structure overlapped with the actual A3G191-384 catalytic domain structure (Figure 3.10B and C). Interestingly, the resulting full-length A3G model suggested that the N-terminal pseudo-active site and the C-terminal active site are situated on different faces of the holoenzyme (Figure 3.10B and C). Moreover, in contrast to full-length A3G models based on the crystal structure of the β2–β2 homodimeric A2 protein [18, 111], the short N-α6-to-C-α1 connection placed physical constraints upon our model such that it became impossible for the β2 region of the pseudo-catalytic domain (N-β2) to mediate interdomain interactions with the β2–loop–β2′ region of the catalytic domain (in blue and yellow, respectively, in Figure 3.10C).

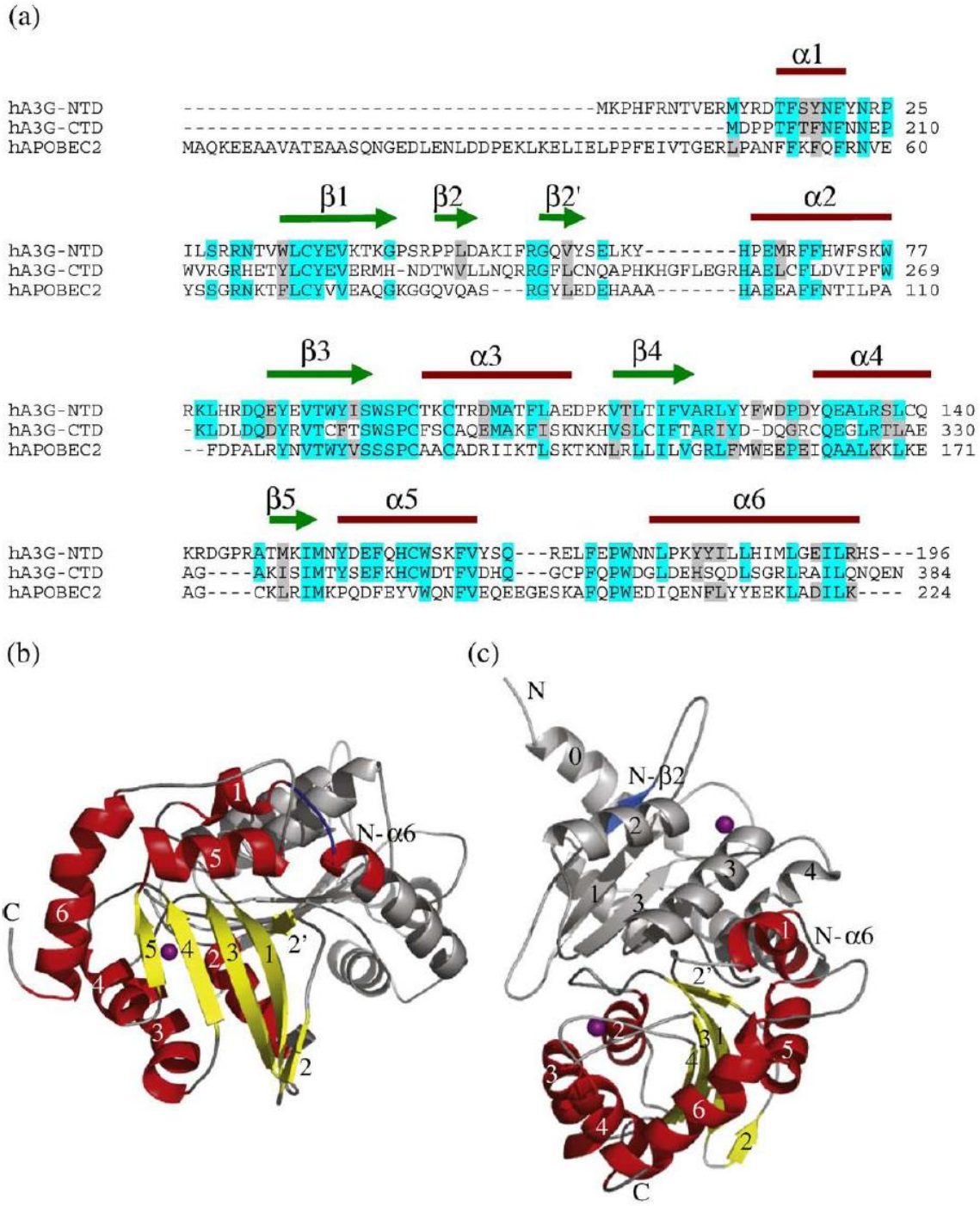


Figure 3.10. A model for full-length A3G. (A) Amino acid alignments of A3G1-196, A3G197-384, and A2. The National Center for Biotechnology Information accession numbers for these proteins are NP_068594 and NP_006780. Identity is indicated in light blue, similarity is indicated by gray shading, and actual secondary structure elements for A3G191-384-2K3A are illustrated above the alignments. (B and C) Two different angles of a full-length A3G structural model. In (B), the catalytic domain is viewed with the same angle as Figure 3.5B, which clearly shows the superposition N-α6 region. The model structure of the N-terminal pseudo-catalytic domain is shown in gray. Helices and β-strands within the C-terminal catalytic domain are shown in red and yellow, respectively. N-β2 and N-α6 are labeled and highlighted in blue and red, respectively. The β2 region of the pseudo-catalytic domain modeled as a continuous strand and, unlike prior A3G catalytic domain structures, the pseudo-catalytic domain has an N-terminal extension that modeled as an α-helix.

It should also be noted that three loops appear close to the domain interface [N-α4–loop–β5 (residues 140–149), C-α1–loop–β1 (residues 210–217), and C–β2′–loop–α2 (residues 245–256)] and that the structures of these loops have yet to be well determined ([114] and this study). For instance, the loop between N-α4 and N-β5 has a four-residue insertion relative to the A3G catalytic domain, making it difficult to model precisely (Figure 3.10A). Moreover, the large loops between α1 and β1 and between β2′ and α2 of the catalytic domain are flexible in solution, and the structures were not resolved by NMR ([114, 115] and this study). It is likely that multiple contact points mediate interdomain interactions and that further structural and functional studies will be needed to precisely delineate all of the residues involved.

3.4 Discussion

3.4.1 A3G-2K3A β2-bulge-β2'

Our original A3G198-384-2K3A solution structure revealed a loop-like bulge within the β2 region [114]. This unique structure is also apparent in our new A3G191-384-2K3A NMR data sets (Figures 3.5 & 3.11) and in a recently published A3G193-384 NMR structure (Figure 3.11B) [115]. In contrast, a recent article describing an A3G197-380 crystal structure reported that the β2 region was continuous and possibly functionally equivalent to the long continuous β2 strand of A2 that mediates homodimerization [107]. However, a detailed atomic comparison of the β2 regions of the A3G191-384-2K3A NMR structure (Figure 3.11A and E), the A3G193-384 NMR structure (Figure 3.11B and F), the A3G197-380 crystal structure (Figure 3.11C and G), and the A2 crystal structure

(Figure 3.11D and H) revealed that the A3G catalytic domain crystal structure also has a discontinuous β2 region. First, for purposes of orientation, the hydrogen bonds between A3G β2′ residues 240–243 and β1 residues 219–222 are identical in both NMR and crystal structures (Figure 3.11E-G). Second, the structures differ at R239, which contacts β1 in the crystal but forms part of the loop in the NMR structures. Third and importantly, in the crystal structure, the next N-terminal residue R238 is bulged out such that the amide nitrogen of Q237 takes its place and bonds with the carbonyl oxygen of V224. In the NMR structure, R238 is part of the loop, and it does not appear to make any chemical contacts. Fourth, in both structures, N236 is located in the central part of the loop, and it does not appear to contact any β1 residues. Finally, in the crystal structure, β2–β1 hydrogen bonding resumes, with L235 interacting with R226. This event is associated with an excess of β2–β1 residues, with three residues in β2 (N236, Q237, and R238) juxtaposing only two residues in β1 (V224 and E225) (Figure 3.11E). The net result of this mismatch is a bulged-out β2 strand. Altogether, the β2 strand in the A3G197-380 crystal structure is discontinuous at three positions: residues N236, Q237, and R238. It is important to emphasize that these three residues constitute a major part of the loop-like bulge in the β2–bulge–β2′ region in the NMR structures ([114, 115] and this study), in reverse-engineered protein preparations with L234 and/or C243 restored ([114] and Figure 3.5), and in wild-type A3G193-384 [115]. We conclude that a loop-like bulge structure is a fundamental part of the β2 region of the A3G catalytic domain. Nevertheless, the significant differences between NMR spectroscopic studies and crystallographic studies are the length and the organization of the β2 portion of the β2–loop-like-bulge–β2′ region (compare Figure 3.11E-G).

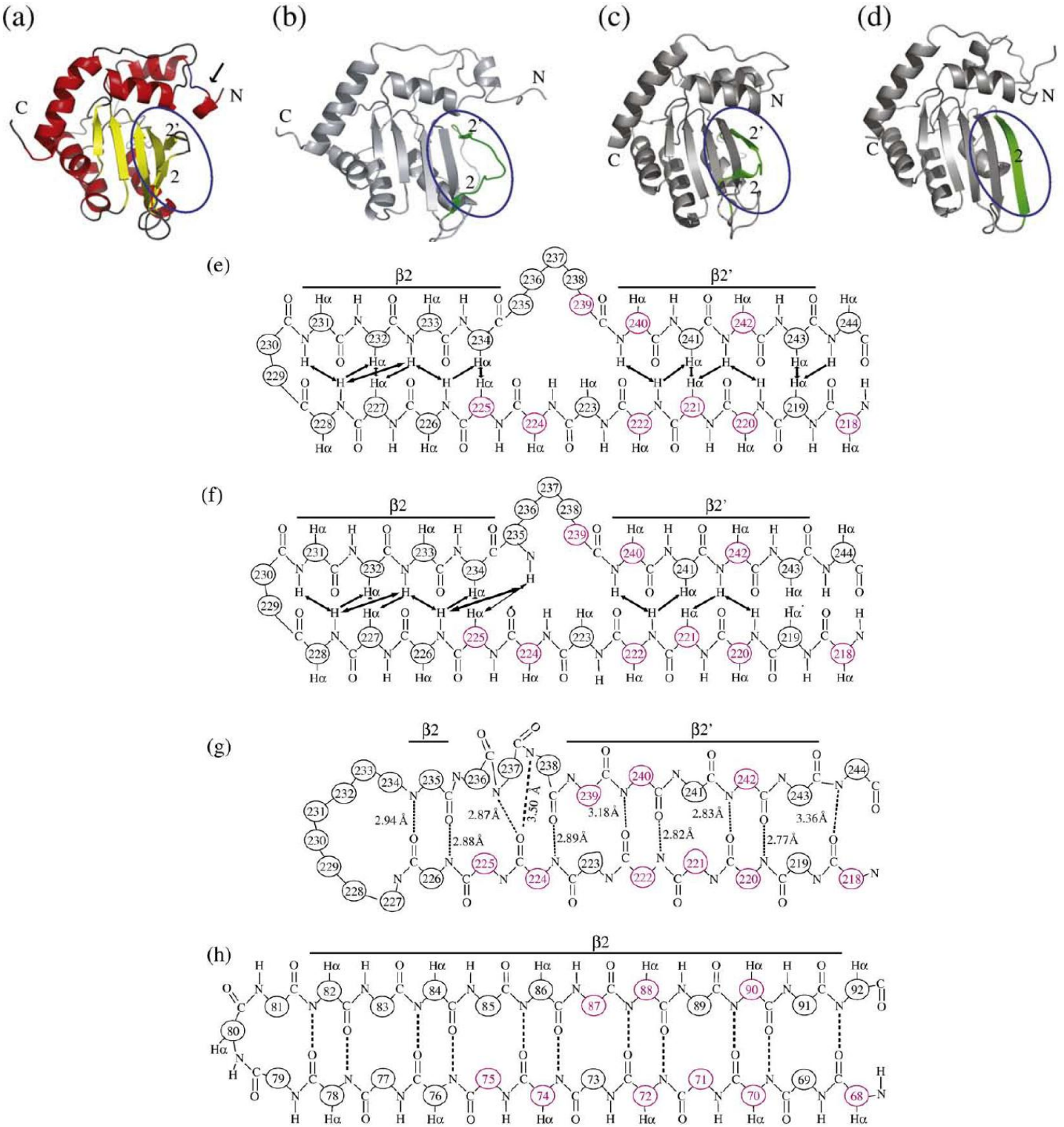


Figure 3.11. A structural comparison A3G191-384-2K3A (PDB ID code 2kem; this study), A3G193-384 (PDB ID code 2kbo), A3G197-380 (PDB ID code 3e1u), and A2 (PDB ID code 2nyt). (A-D) α-carbon-traced ribbon schematics for the aforementioned structures. The β2 regions are highlighted by blue ovals and shown in green in (B)–(D). The interdomain linker region in (A) is indicated by an arrow. (E-H) Polypeptide backbone and chemical contact schematics for the β2 andβ1 regions of A3G191-384-2K3A, A3G193-384, A3G197-380, and A2, respectively. Amino acid positions are numbered and, if the residue is conserved, shown in red. In (E) and (F), observed NOEs are indicated by arrows. In (G) and (H), predicted hydrogen bonds are indicated by broken lines, and labels correspond to distances between the amide nitrogen and the carbonyl oxygen atoms.

The NMR structures showed that residues 229 and 230 form the β1-to-β2 loop and that β1 residues 225–228 form a β-sheet with β2 residues 231–234 (regardless of whether residue 234 was L or K; Figure 3.11E and F; Figure 3.5 [114, 115]). Similar conceptually but contrasting in some details, the crystal structure has a much larger loop from residues 227–234 and only a single β1-to-β2 interaction between R226 and L235. In total contrast, the corresponding region of A2 is clearly part of a continuous β2 strand (Figure 3.11H). At present, we cannot explain this difference between the A3G catalytic domain structures, but it may be due to differences in buffers and/or techniques. However, the difference may suggest that the β2 region has an intrinsic structural flexibility, which enables it to adopt multiple conformations (perhaps also influenced by nucleic acid and/or protein interactions). Indeed, it is striking that most of the A3G β2 region residues are dispensable for deaminase activity in *E. coli* (most single-alanine substitutions within residues 232–243 were at least 50% active, with the exception of G240, which was not tested [73, 114]).

3.4.2 A3G-2K3A/ssDNA binding model

The NMR chemical shift perturbation experiments discussed in the main text and shown in Figures 3.6 and 3.7 used a 21 base ssDNA 5'-GCT-TCT-TCT-ACC-TTC-TCT-TGA-3' because it had been characterized previously as an HIV-1 derived sequence that binds full-length APOBEC3G *in vitro* (P. Henry, A. Stephen and V. Pathak, NCI-Frederick, personal communication), it contained a predicted APOBEC3G deamination hotspot (5'-CC with the target cytosine underlined) and it was among the shortest sequences to trigger significant chemical shift perturbations. Consistent with published

studies from the Goodman and Levin laboratories [33, 43], shorter deoxy-oligonucleotides did not appear to bind A3G-2K3A in gel shift assays and/or cause strong chemical shift perturbations. For instance, at ssDNA to A3G-2K3A molar ratios of 8:1, we failed to detect significant chemical shifts for 5'-ACC-CA, 5'-AAA-CCC-AAA-A and 5'-GGG-AGA-CCC-AAA-GAG [43] or for a transition state analogue 2' deoxy-tetra-hydroxy-uridine (dTHU).

However, at molar ratios of 4 ssDNA to 1 A3G-2K3A, we observed that a 20 base ssDNA 5'-AAA-GGG-AGA-CCC-AAA-GAG-GA [43], a 21 base ssDNA (above) and longer deoxy-oligonucleotides could trigger significant chemical shift perturbations (Figures 3.6 and 3.7). The 21-mer triggered greater NMR chemical shift perturbations, and therefore these datasets were expanded upon for detailed presentation (Figures 3.6 and 3.7). We further noted that the longer deoxy-oligonucleotides (*e.g.*, a 69mer [43]) were not particularly useful for NMR experimentation, because they also tended to cause line-broadening of A3G-2K3A NMR signals and more non-specific chemical shifts perturbations. Overall, these data favor a model in which the C-terminal catalytic domain of APOBEC3G binds ssDNA ($\geq$20 bases) weakly using the catalytic pocket but, consistent with prior reports [33, 35], the N-terminal pseudo-catalytic domain likely functions to promote stronger nucleic acid interactions. Indeed, in comparison to published dissociation constants for full-length APOBEC3G (76 $\pm$ 21 nM for a 20 nucleotide ssDNA [33] and 50 $\pm$ 7 nM for a 69 nucleotide ssDNA [43]), data from NMR chemical shift perturbation experiments were used to estimate that the ssDNA dissociation constant of A3G-2K3A is minimally 450uM (determined using published methods [116]. ^{15}N-^{1}H HSQC NMR signals of R215, E254, R256, I314, Y315, D316 and

H367 were used to estimate dissociation constants (these residues showed significant ssDNA-induced chemical shift perturbations and located near the catalytic site; Figures 3.1, 3.6 and 3.7). Future experiments with full-length APOBEC3G and/or variants containing more of the N-terminal pseudo-catalytic domain are anticipated to be particularly interesting.

3.5 Conclusions

In this work, we determine the structure of C-terminal catalytic domain of APOBEC3G (A3G-198-384-2K3A & A3G-191-384-2K3A). This domain is structurally related to another family member APOBEC2, with the exception of the discontinuous β2 strand. In addition, the structural model of the catalytic domain and DNA complex proposed here is slightly different from the model published by the crystal group [107]. The DNA is placed parallel to catalytic residues containing α2 and α3 helices in this study while it is placed perpendicularly to α2 and α3 helices by the crystal group. The longer fragment (A3G-191-384-2K3A) shows an α1-helix (residues 201–206) that was not observed in the shorter construct (A3G-198-384-2K3A), and it provides 10-fold higher enzymatic activity through interactions with the hydrophobic core region. In addition, the novel full-length A3G model demonstrates the N-terminal deaminase is connected to C-terminal deaminase in a back-to-face orientation (Figure 3.10), which is different from the model published by the crystal group who proposes the parallel side-by-side orientation [117].

3.6 Future consideration

The protein's intrinsic solubility can be improved by introducing hydrophilic residues to "key positions" on the surface. We have developed a strategy to enhance the solubility of the catalytic domain of APOBEC3G. A systematic mutagenesis was performed to construct a series of alanine variants in which hydrophobic residues are individually substituted for alanine residues. The *E. coli* mutation assay, used to test the function of the variants, confirmed the protein structure was not significantly altered by testing for activity. By using mutagenesis, *E. coli* mutation assay and solubility test iteratively, one soluble and catalytically active A3G-ctd variant A3G-2K3A was identified. So far, several high resolution structures of the human APOBEC family are available in the Protein Data Bank, but they are limited to only APOBEC2 and the catalytic domain of APOBEC3G. The strategy described here has the potential to be applied to other family members that have deaminase activity, such as AID, APOBEC3B, and APOBEC3F**.**

While studies of APOBEC3G have increased greatly in recent years, understanding of the central problem about how Vif interacts with APOBEC3G remains elusive. Vif is a viral protein against the antiviral factor APOBEC3G, however, the study of this protein is hindered due to the fact that it is an unfolded or partial folded protein. One speculation is that Vif may form an intact structure and recruit the cellular ubiquitin-ligase complex once bound to it substrate, APOBEC3G. Therefore, obtaining the soluble full-length APOBEC3G would be of great interest in studying the Vif-APOBEC3G interaction. The soluble full-length APOBEC3G would be able to detect the dynamic of Vif's structure formation as well as the folded structure of Vif in the complex using NMR.

In addition, how C-terminal deaminase and N-terminal deaminase of APOBEC3G align to each other in the full-length structure as well as how monomers dimerize to exert their deaminase activity would also been answered. Development of small inhibitors to disrupt the Vif-APOBEC3G interaction would also be a good target in the field against HIV virus infection.

References

1. Jarmuz, A., et al., *An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22.* Genomics, 2002. **79**(3): p. 285-96.
2. Holmes, R.K., M.H. Malim, and K.N. Bishop, *APOBEC-mediated viral restriction: not simply editing?* Trends Biochem Sci, 2007. **32**(3): p. 118-28.
3. Sawyer, S.L., M. Emerman, and H.S. Malik, *Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G.* PLoS Biol, 2004. **2**(9): p. E275.
4. LaRue, R.S., et al., *Guidelines for naming nonprimate APOBEC3 genes and proteins.* J Virol, 2009. **83**(2): p. 494-7.
5. Betts, L., et al., *Cytidine deaminase. The 2.3 A crystal structure of an enzyme: transition-state analog complex.* J Mol Biol, 1994. **235**(2): p. 635-56.
6. Johansson, E., et al., *Crystal structure of the tetrameric cytidine deaminase from Bacillus subtilis at 2.0 A resolution.* Biochemistry, 2002. **41**(8): p. 2563-70.
7. Ko, T.P., et al., *Crystal structure of yeast cytosine deaminase. Insights into enzyme mechanism and evolution.* J Biol Chem, 2003. **278**(21): p. 19111-7.
8. Xie, K., et al., *The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1.* Proc Natl Acad Sci U S A, 2004. **101**(21): p. 8114-9.
9. Harris, R.S. and M.T. Liddament, *Retroviral restriction by APOBEC proteins.* Nat Rev Immunol, 2004. **4**(11): p. 868-77.
10. Anant, S., A.J. MacGinnitie, and N.O. Davidson, *apobec-1, the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme, is a novel RNA-binding protein.* J Biol Chem, 1995. **270**(24): p. 14762-7.
11. Yang, Y., M.P. Sowden, and H.C. Smith, *Induction of cytidine to uridine editing on cytoplasmic apolipoprotein B mRNA by overexpressing APOBEC-1.* J Biol Chem, 2000. **275**(30): p. 22663-9.
12. Chen, S.H., et al., *Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon.* Science, 1987. **238**(4825): p. 363-6.
13. Chan, L., *Apolipoprotein B, the major protein component of triglyceride-rich and low density lipoproteins.* J Biol Chem, 1992. **267**(36): p. 25621-4.
14. Yang, Y., K. Kovalski, and H.C. Smith, *Partial characterization of the auxiliary factors involved in apolipoprotein B mRNA editing through APOBEC-1 affinity chromatography.* J Biol Chem, 1997. **272**(44): p. 27700-6.
15. Di Noia, J.M. and M.S. Neuberger, *Molecular mechanisms of antibody somatic hypermutation.* Annu Rev Biochem, 2007. **76**: p. 1-22.
16. Longerich, S., et al., *AID in somatic hypermutation and class switch recombination.* Curr Opin Immunol, 2006. **18**(2): p. 164-74.
17. Mikl, M.C., et al., *Mice deficient in APOBEC2 and APOBEC3.* Mol Cell Biol, 2005. **25**(16): p. 7270-7.
18. Prochnow, C., et al., *The APOBEC-2 crystal structure and functional implications for the deaminase AID.* Nature, 2007. **445**(7126): p. 447-51.

19. LaRue, R.S., et al., *The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals.* BMC Mol Biol, 2008. **9**: p. 104.
20. Chen, H., et al., *APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons.* Curr Biol, 2006. **16**(5): p. 480-5.
21. Harris, R.S., S.K. Petersen-Mahrt, and M.S. Neuberger, *RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators.* Mol Cell, 2002. **10**(5): p. 1247-53.
22. Franca, R., S. Spadari, and G. Maga, *APOBEC deaminases as cellular antiviral factors: a novel natural host defense mechanism.* Med Sci Monit, 2006. **12**(5): p. RA92-8.
23. Goila-Gaur, R. and K. Strebel, *HIV-1 Vif, APOBEC, and intrinsic immunity.* Retrovirology, 2008. **5**: p. 51.
24. Stenglein, M.D., et al., *APOBEC3 proteins mediate the clearance of foreign DNA from human cells.* Nat Struct Mol Biol, 2010. **17**(2): p. 222-9.
25. Di Noia, J. and M.S. Neuberger, *Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase.* Nature, 2002. **419**(6902): p. 43-8.
26. Bogerd, H.P., et al., *Cellular inhibitors of long interspersed element 1 and Alu retrotransposition.* Proc Natl Acad Sci U S A, 2006. **103**(23): p. 8780-5.
27. Narvaiza, I., et al., *Deaminase-independent inhibition of parvoviruses by the APOBEC3A cytidine deaminase.* PLoS Pathog, 2009. **5**(5): p. e1000439.
28. Kinomoto, M., et al., *All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition.* Nucleic Acids Res, 2007. **35**(9): p. 2955-64.
29. Turelli, P. and D. Trono, *Editing at the crossroad of innate and adaptive immunity.* Science, 2005. **307**(5712): p. 1061-5.
30. Dang, Y., et al., *Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family.* J Virol, 2006. **80**(21): p. 10522-33.
31. Bogerd, H.P., et al., *APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells.* Nucleic Acids Res, 2006. **34**(1): p. 89-95.
32. Liddament, M.T., et al., *APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo.* Curr Biol, 2004. **14**(15): p. 1385-91.
33. Iwatani, Y., et al., *Biochemical activities of highly purified, catalytically active human APOBEC3G: correlation with antiviral effect.* J Virol, 2006. **80**(12): p. 5992-6002.
34. Opi, S., et al., *Monomeric APOBEC3G is catalytically active and has antiviral activity.* J Virol, 2006. **80**(10): p. 4673-82.
35. Navarro, F., et al., *Complementary function of the two catalytic domains of APOBEC3G.* Virology, 2005. **333**(2): p. 374-86.
36. Hache, G., M.T. Liddament, and R.S. Harris, *The retroviral hypermutation specificity of APOBEC3F and APOBEC3G is governed by the C-terminal DNA cytosine deaminase domain.* J Biol Chem, 2005. **280**(12): p. 10920-4.

37. Bogerd, H.P., et al., *The intrinsic antiretroviral factor APOBEC3B contains two enzymatically active cytidine deaminase domains.* Virology, 2007. **364**(2): p. 486-93.
38. Rogozin, I.B., et al., *APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis.* Cell Cycle, 2005. **4**(9): p. 1281-5.
39. Lecossier, D., et al., *Hypermutation of HIV-1 DNA in the absence of the Vif protein.* Science, 2003. **300**(5622): p. 1112.
40. Mariani, R., et al., *Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif.* Cell, 2003. **114**(1): p. 21-31.
41. Sheehy, A.M., et al., *Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein.* Nature, 2002. **418**(6898): p. 646-50.
42. Haché, G., M.T. Liddament, and R.S. Harris, *The retroviral hypermutation specificity of APOBEC3F and APOBEC3G is governed by the C-terminal DNA cytosine deaminase domain.* J Biol Chem, 2005. **280**(12): p. 10920-4.
43. Chelico, L., et al., *APOBEC3G DNA deaminase acts processively 3' --> 5' on single-stranded DNA.* Nat Struct Mol Biol, 2006. **13**(5): p. 392-9.
44. Guo, F., et al., *Inhibition of formula-primed reverse transcription by human APOBEC3G during human immunodeficiency virus type 1 replication.* J Virol, 2006. **80**(23): p. 11710-22.
45. Mbisa, J.L., et al., *Human immunodeficiency virus type 1 cDNAs produced in the presence of APOBEC3G exhibit defects in plus-strand DNA transfer and integration.* J Virol, 2007. **81**(13): p. 7099-110.
46. Luo, K., et al., *Cytidine deaminases APOBEC3G and APOBEC3F interact with human immunodeficiency virus type 1 integrase and inhibit proviral DNA formation.* J Virol, 2007. **81**(13): p. 7238-48.
47. Yang, B., et al., *Virion-associated uracil DNA glycosylase-2 and apurinic/apyrimidinic endonuclease are involved in the degradation of APOBEC3G-edited nascent HIV-1 DNA.* J Biol Chem, 2007. **282**(16): p. 11667-75.
48. Shindo, M., et al., *Synthesis and phorbol ester binding of the cysteine-rich domains of diacylglycerol kinase (DGK) isozymes. DGKgamma and DGKbeta are new targets of tumor-promoting phorbol esters.* J Biol Chem, 2003. **278**(20): p. 18448-54.
49. Newman, E.N., et al., *Antiviral function of APOBEC3G can be dissociated from cytidine deaminase activity.* Curr Biol, 2005. **15**(2): p. 166-70.
50. Nguyen, D.H., S. Gummuluru, and J. Hu, *Deamination-independent inhibition of hepatitis B virus reverse transcription by APOBEC3G.* J Virol, 2007. **81**(9): p. 4465-72.
51. Bishop, K.N., R.K. Holmes, and M.H. Malim, *Antiviral potency of APOBEC proteins does not correlate with cytidine deamination.* J Virol, 2006. **80**(17): p. 8450-8.
52. Iwatani, Y., et al., *Deaminase-independent inhibition of HIV-1 reverse transcription by APOBEC3G.* Nucleic Acids Res, 2007. **35**(21): p. 7096-108.

53. Chiu, Y.L., et al., *Cellular APOBEC3G restricts HIV-1 infection in resting CD4+ T cells.* Nature, 2005. **435**(7038): p. 108-14.
54. Mangeat, B., et al., *Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts.* Nature, 2003. **424**(6944): p. 99-103.
55. Gabuzda, D.H., et al., *Role of vif in replication of human immunodeficiency virus type 1 in CD4+ T lymphocytes.* J Virol, 1992. **66**(11): p. 6489-95.
56. von Schwedler, U., et al., *Vif is crucial for human immunodeficiency virus type 1 proviral DNA synthesis in infected cells.* J Virol, 1993. **67**(8): p. 4945-55.
57. Stopak, K., et al., *HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability.* Mol Cell, 2003. **12**(3): p. 591-601.
58. Marin, M., et al., *HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation.* Nat Med, 2003. **9**(11): p. 1398-403.
59. Conticello, S.G., R.S. Harris, and M.S. Neuberger, *The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G.* Curr Biol, 2003. **13**(22): p. 2009-13.
60. Sheehy, A.M., N.C. Gaddis, and M.H. Malim, *The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif.* Nat Med, 2003. **9**(11): p. 1404-7.
61. Yu, X., et al., *Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex.* Science, 2003. **302**(5647): p. 1056-60.
62. Xiao, Z., et al., *Characterization of a novel Cullin5 binding domain in HIV-1 Vif.* J Mol Biol, 2007. **373**(3): p. 541-50.
63. Iwatani, Y., et al., *HIV-1 Vif-mediated ubiquitination/degradation of APOBEC3G involves four critical lysine residues in its C-terminal domain.* Proc Natl Acad Sci U S A, 2009. **106**(46): p. 19539-44.
64. Huthoff, H. and M.H. Malim, *Identification of amino acid residues in APOBEC3G required for regulation by human immunodeficiency virus type 1 Vif and Virion encapsidation.* J Virol, 2007. **81**(8): p. 3807-15.
65. Schrofelbauer, B., et al., *Mutational alteration of human immunodeficiency virus type 1 Vif allows for functional interaction with nonhuman primate APOBEC3G.* J Virol, 2006. **80**(12): p. 5984-91.
66. Russell, R.A. and V.K. Pathak, *Identification of two distinct human immunodeficiency virus type 1 Vif determinants critical for interactions with human APOBEC3G and APOBEC3F.* J Virol, 2007. **81**(15): p. 8201-10.
67. Pery, E., et al., *Regulation of APOBEC3 proteins by a novel YXXL motif in human immunodeficiency virus type 1 Vif and simian immunodeficiency virus SIVagm Vif.* J Virol, 2009. **83**(5): p. 2374-81.
68. Mangeat, B., et al., *A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action.* J Biol Chem, 2004. **279**(15): p. 14481-3.
69. Schrofelbauer, B., D. Chen, and N.R. Landau, *A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif).* Proc Natl Acad Sci U S A, 2004. **101**(11): p. 3927-32.

70. Bogerd, H.P., et al., *A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor.* Proc Natl Acad Sci U S A, 2004. **101**(11): p. 3770-4.
71. Harjes, E., et al., *An extended structure of the APOBEC3G catalytic domain suggests a unique holoenzyme model.* J Mol Biol, 2009. **389**(5): p. 819-32.
72. Harris, R.S., S.K. Petersen-Mahrt, and M.S. Neuberger, *RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators.* Molecular Cell, 2002. **10**(5): p. 1247-53.
73. Chen, K.M., et al., *Extensive mutagenesis experiments corroborate a structural model for the DNA deaminase domain of APOBEC3G.* FEBS Lett, 2007. **581**(24): p. 4761-6.
74. Philo, J.S., *A method for directly fitting the time derivative of sedimentation velocity data and an alternative algorithm for calculating sedimentation coefficient distribution functions.* Anal Biochem, 2000. **279**(2): p. 151-63.
75. Philo, J.S., *Improved methods for fitting sedimentation coefficient distributions derived by time-derivative techniques.* Anal Biochem, 2006. **354**(2): p. 238-46.
76. Schuck, P., *On the analysis of protein self-association by sedimentation velocity analytical ultracentrifugation.* Anal Biochem, 2003. **320**(1): p. 104-24.
77. Stafford, W.F. and P.J. Sherwood, *Analysis of heterologous interacting systems by sedimentation velocity: curve fitting algorithms for estimation of sedimentation coefficients, equilibrium and kinetic constants.* Biophys Chem, 2004. **108**(1-3): p. 231-43.
78. Li, J., M.J. Potash, and D.J. Volsky, *Functional domains of APOBEC3G required for antiviral activity.* J Cell Biochem, 2004. **92**(3): p. 560-72.
79. Shindo, K., et al., *The enzymatic activity of CEM15/Apobec-3G is essential for the regulation of the infectivity of HIV-1 virion but not a sole determinant of its antiviral activity.* J Biol Chem, 2003. **278**(45): p. 44412-6.
80. Jonsson, S.R., et al., *Evolutionarily conserved and non-conserved retrovirus restriction activities of artiodactyl APOBEC3F proteins.* Nucleic Acids Res, 2006. **34**(19): p. 5683-94.
81. Losey, H.C., A.J. Ruthenburg, and G.L. Verdine, *Crystal structure of Staphylococcus aureus tRNA adenosine deaminase TadA in complex with RNA.* Nat Struct Mol Biol, 2006. **13**(2): p. 153-9.
82. Wolf, J., A.P. Gerber, and W. Keller, *tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli.* EMBO J, 2002. **21**(14): p. 3841-51.
83. Devany, M., N.P. Kotharu, and H. Matsuo, *Solution NMR structure of the C-terminal domain of the human protein DEK.* Protein Sci, 2004. **13**(8): p. 2252-9.
84. Matsuo, H., et al., *Increased sensitivity in HNCA and HN(CO)CA experiments by selective C beta decoupling.* J Magn Reson B, 1996. **113**(1): p. 91-6.
85. Ikura, M., L.E. Kay, and A. Bax, *A novel approach for sequential assignment of 1H, 13C, and 15N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin.* Biochemistry, 1990. **29**(19): p. 4659-67.

86. Lewis E. Kay, M.I., Rolf Tschudin and Ad Bax, *Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins* J. Magn. Reson. , 1990 **89**(3): p. 496-514

87. Wittekind, M.M., *HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha- and beta- carbon resonances in proteins.* J. Magn. Reson., 1993. **101**: p. 201-205.
88. Toshio Yamazaki, W.L., Cheryl H. Arrowsmith, D. R. Muhandiram, Lewis E. Kay, *A suite of triple resonance NMR experiments for the backbone assignment of 15N, 13C, 2H labeled proteins with high sensitivity.* J. Am. Chem. Soc., 1994. **116**(26): p. 11655-11666.
89. Matsuo, H., H. Li, and G. Wagner, *A sensitive HN(CA)CO experiment for deuterated proteins.* J Magn Reson B, 1996. **110**(1): p. 112-5.
90. Matsuo, H., et al., *Use of selective C alpha pulses for improvement of HN(CA)CO-D and HN(COCA)NH-D experiments.* J Magn Reson B, 1996. **111**(2): p. 194-8.
91. Clubb, R.T., Thanabal, V., and Wagner, G., *A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue 1HN, 15N, and 13C' chemical shifts in 15N---13C-labelled proteins.* J. Magn. Reson. , 1992. **97**: p. 213-217.
92. Grzesiek, S.e.a., *Correlation of backbone amide and aliphatic side-chain resonances in 13C/15N-enriched proteins by isotropic mixing of 13C magnetization.* J. Magn. Reson., 1993. **101**: p. 114-119.
93. Clore, G.M., et al., *Assignment of the side-chain 1H and 13C resonances of interleukin-1 beta using double- and triple-resonance heteronuclear three-dimensional NMR spectroscopy.* Biochemistry, 1990. **29**(35): p. 8172-84.
94. Zhang, O., et al., *Backbone 1H and 15N resonance assignments of the N-terminal SH3 domain of drk in folded and unfolded states using enhanced-sensitivity pulsed field gradient NMR techniques.* J Biomol NMR, 1994. **4**(6): p. 845-58.
95. Delaglio, F., et al., *NMRPipe: a multidimensional spectral processing system based on UNIX pipes.* J Biomol NMR, 1995. **6**(3): p. 277-93.
96. Keller, R., *Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignment.* . PhD thesis, Swiss Fed. Inst. Tech. Zurich, 2004.
97. Cornilescu, G., F. Delaglio, and A. Bax, *Protein backbone angle restraints from searching a database for chemical shift and sequence homology.* J Biomol NMR, 1999. **13**(3): p. 289-302.
98. Herrmann, T., P. Guntert, and K. Wuthrich, *Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS.* J Biomol NMR, 2002. **24**(3): p. 171-89.
99. Herrmann, T., P. Guntert, and K. Wuthrich, *Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA.* J Mol Biol, 2002. **319**(1): p. 209-27.

100. Brunger, A.T., et al., *Crystallography & NMR system: A new software suite for macromolecular structure determination.* Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 5): p. 905-21.
101. Teh, A.H., et al., *The 1.48 A resolution crystal structure of the homotetrameric cytidine deaminase from mouse.* Biochemistry, 2006. **45**(25): p. 7825-33.
102. Xiang, S., et al., *The structure of the cytidine deaminase-product complex provides evidence for efficient proton transfer and ground-state destabilization.* Biochemistry, 1997. **36**(16): p. 4768-74.
103. Desmet, J., Maeyer, M.D., Hazes, B., Lasters, I.,, *The dead-end elimination theorem and its use in protein side-chain positioning.* Nature, 1992(356): p. 539 - 542
104. Goldstein, R.F., *Efficient rotamer elimination applied to protein side-chains and related spin glasses.* Biophys J, 1994. **66**(5): p. 1335-40.
105. Krieger, E., G. Koraimann, and G. Vriend, *Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field.* Proteins, 2002. **47**(3): p. 393-402.
106. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
107. Holden, L.G., et al., *Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications.* Nature, 2008. **456**(7218): p. 121-4.
108. Conticello, S.G., et al., *DNA deamination in immunity: AID in the context of its APOBEC relatives.* Adv Immunol, 2007. **94**: p. 37-73.
109. Huthoff, H. and M.H. Malim, *Cytidine deamination and resistance to retroviral infection: towards a structural understanding of the APOBEC proteins.* Virology, 2005. **334**(2): p. 147-53.
110. Chung, S.J., J.C. Fromme, and G.L. Verdine, *Structure of human cytidine deaminase bound to a potent inhibitor.* J Med Chem, 2005. **48**(3): p. 658-60.
111. Zhang, K.L., et al., *Model structure of human APOBEC3G.* PLoS One, 2007. **2**(4): p. e378.
112. Esnault, C., et al., *APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses.* Nature, 2005. **433**(7024): p. 430-3.
113. Harris, R.S., et al., *DNA deamination mediates innate immunity to retroviral infection.* Cell, 2003. **113**(6): p. 803-9.
114. Chen, K.M., et al., *Structure of the DNA deaminase domain of the HIV-1 restriction factor APOBEC3G.* Nature, 2008. **452**(7183): p. 116-9.
115. Furukawa, A., et al., *Structure, interaction and real-time monitoring of the enzymatic reaction of wild-type APOBEC3G.* EMBO J, 2009. **28**(4): p. 440-51.
116. Kim, S., et al., *Solution structure of the Reps1 EH domain and characterization of its binding to NPF target sequences.* Biochemistry, 2001. **40**(23): p. 6776-85.
117. Chelico, L., et al., *A structural model for deoxycytidine deamination mechanisms of the HIV-1 inactivation enzyme APOBEC3G.* J Biol Chem, 2010.