DNA Substrate Length and Surrounding Sequence Affect the Activation Induced Deaminase Activity at Cytidine

Kefei Yu, Feng-Ting Huang, and Michael R. Lieber*

USC Norris Comprehensive Cancer Ctr., Rm. 5428
Departments of Pathology, Biochemistry & Molecular Biology, Molecular Microbiology & Immunology, and Biological Sciences
1441 Eastlake Ave., MC9176
Los Angeles, CA 90089

Key Words: class switch recombination, somatic hypermutation, immunoglobulin gene recombination, cytidine deaminase, cytosine

Running Title: AID substrate length and sequence effects

*Corresponding author. Ph. 323-865-0568. lieber@usc.edu
The first two authors contributed equally.

Copyright 2003 by The American Society for Biochemistry and Molecular Biology, Inc.
Abstract

Activation-induced deaminase (AID) is required for both immunoglobulin class switch recombination and somatic hypermutation. AID is known to deaminate cytidines in single-stranded DNA, but the relationship of this step to the class switch or somatic hypermutation processes is not entirely clear. We have studied the activity of a recombinant form of the mouse AID protein that was purified from a baculovirus expression system. We find that the length of the single-stranded DNA target is critical to the action of AID at the C’s positioned anywhere along the length of the DNA. The DNA sequence surrounding a given C influences AID deamination efficiency. AID preferentially deaminates C’s in the WRC motif and additionally has a small but consistent preference for purine at the position after the WRC, thereby favoring WRCr (the lower case r corresponds to the smaller impact on activity).
Introduction

Three genetic modifications occur at immunoglobulin loci. The first is V(D)J recombination, which involves the assembly of the variable domain exon of the heavy and light chains from V, D, and J subexons. The second is class switch recombination (CSR), which involves changing the heavy chain from Igµ to Igα, Igγ, or Igε (1). This results in a change in the Ig protein from IgM to IgG, IgA or IgE, respectively. The third genetic modification is the introduction of point mutations in the Ig variable domain exons of the heavy and light chains. This is called somatic hypermutation (SHM) and is responsible for the affinity maturation of antibodies (2). The latter two processes are not dependent on each other, although they both occur in the germinal centers of peripheral lymphoid tissues and share some common features. Specifically, they both require transcription through the target region (2) and require a 24 kDa protein called activation-induced cytidine deaminase (AID) (3-5).

AID was initially identified from a cDNA subtraction between B cells capable of spontaneous CSR from cells that were not (6). AID null mice are completely deficient in both CSR and SHM (3). A group of human patients with mutations in the AID gene have hyper-IgM syndrome because of blocked CSR, and they are also deficient for SHM (4). AID is normally only expressed in stimulated B cells. Ectopic expression of AID in non-lymphoid cell types can confer the ability to carry out CSR and SHM (7-9), indicating that AID is the only lymphoid-specific factor required.

AID shares sequence homology to a cytidine deaminase, Apobec-1 (6), the catalytic subunit of a large mRNA editing enzyme complex. Because of this sequence homology, AID was initially thought to represent an mRNA editing activity critical for changing the property of the CSR recombinase and/or SHM mutagenesis machinery (10). However, key observations from the Neuberger group indicated that, instead of being an RNA deaminase, AID actually deaminates cytosines in DNA (11,12). Therefore, instead of being an RNA modifier, AID actually appeared more likely to be a key component of the CSR recombinase machinery. This hypothesis was further strengthened by
biochemical evidence demonstrating that purified recombinant AID deaminates C residues in single-stranded DNA (ssDNA) (13-16).

Given that AID deaminates cytosines in single-stranded DNA, how does this permit it to function in CSR and SHM? Previous work has shown that RNA:DNA hybrids form upon transcription through class switch regions in vitro (17-19). Recently, we showed that the conformation of these in vitro RNA:DNA hybrids is that of an R-loop (20). More importantly, we have shown that chromosomal DNA has an R-loop conformation at the in vivo switch regions of stimulated primary B cells (20). Interestingly, single-stranded DNA is an intrinsic property of the R-loop structure found in the switch regions of stimulated B cell chromosomes (20). The R-loop structure is thought to be a key target element in CSR. It has also been suggested that a transcription bubble provides enough single-strandedness in the VDJ region to allow AID-mediated SHM (21). All of these findings favor a DNA deaminase model for both CSR and SHM (22), and an R-loop structure for the CSR substrate.

Despite a recent series of reports on AID biochemistry, much remains to be learned concerning AID function. Different forms of recombinant AID from various laboratories have shown some contradictory properties. One group has provided data that the AID deamination target is a WRC (W = A or T; R = A or G) hotspot motif (23), which is similar to the RGYW/WRCY hotspot identified from in vivo studies (24,25). However, two other groups did not observe such a hotspot predilection in biochemical studies of AID action (15,16). One group has shown that AID is activated by removal of RNA by RNase A (23), but two other groups reported that RNase A had no effect (15,16), leaving the issue of RNase A-activation unclear.

Here, we examined the activity of a recombinant form of the mouse AID protein purified using a baculovirus expression system. We find that the length of the single-stranded DNA target is critical to the action of AID at the C's positioned anywhere along the length of the DNA. Having established the minimum optimal length, we did further analyses on how the DNA sequence context influences AID activity. We have kept the sequence surrounding each C identical so that the
sites are fully comparable. We find that AID preferentially deaminates C’s in the WRC motif and additionally has a small but consistent preference for purine at the position after the WRC, thereby favoring WRCr, where the lower case r designates the smaller impact.

**Experimental Procedures**

**Oligonucleotide substrate**

All oligonucleotides were synthesized by Qiagen/Operon Technologies (Richmond, CA). Oligonucleotides used to determine the suitable AID substrate length are as follows: KY333 (5'TTTTTTTTTACGATTTTTTTT3', 20mer); KY334 (5'TTTTTTTTTTTACGATTTTTTTT3', 24mer); KY335 (5'TTTTTTTTTTTTTACGATTTTTTTTTT3', 30mer). UDG substrates are: KY414 (5'TTTTTTTAUGATTTTT3', 16mer); KY415 (5'TTTTTTTTTAUGATTTTTTT3', 20mer); KY416 (5'TTTTTTTTTTTAUGATTTTTTTTT3', 24mer). All other oligonucleotides used are indicated in the specified figures.

**Recombination protein**

Mouse AID cDNA was kindly provided by Dr. Tasuku Honjo. The glutathione-S-transferase coding sequence was obtained from pAcG2T (Pharmingen, San Diego, CA). A PCR reaction was used to construct the GST-mAID DNA fragment with an enterokinase recognition site between the GST and mAID. The GST-mAID coding region was fully sequenced. Recombinant baculovirus (vKY9) was made using the Bac-to-Bac baculovirus expression system from Invitrogen/Life Technologies (Carlsbad, CA) according to the manufacturer's instructions. SF21 cells (Invitrogen) were infected with vKY9 at a multiplicity of infection (MOI) of 10 and harvested 3 days after infection. Cells were resuspended in buffer A (50mM Tris-Cl, pH8.0; 250mM NaCl; 10mM NaF; 5mM 2-mercaptoethanol) that contains a cocktail of proteinase inhibitors and sonicated. The cell lysate was ultracentrifuged at 100,000g for 1 hr, and the clear cell extract was mixed with glutathione agarose (Sigma, St. Louis, MO). The glutathione agarose was washed 5 times with 10 volumes (relative to bed
volume) of buffer A and GST-mAID recombinant protein was eluted with an equal volume (relative to bed volume) of buffer B (50mM Tris-Cl, pH 8.0; 100mM NaCl; 10mM reduced glutathione; 1mM dithiothreitol). The AID protein on a Coomassie stained SDS polyacrylamide gel is shown in Figure 1A.

Deamination assay

The AID activity assay is shown schematically in Fig. 1B. Uracil generated by AID can be removed from DNA by UDG, resulting in an abasic (AP) site. DNA cleavage at AP sites can be introduced by alkali treatment at elevated temperatures (95°C) (26). Therefore, DNA cleavage can serve as a readout for AID-mediated cytidine deamination. The reaction mixture (10 µl) containing 200 fmol of radioisotope-labeled oligonucleotide substrate, 50 ng (1 pmol) of GSTmAID, 100 ng of RNase A and 0.1 unit (1 unit catalyzes the release of 1 nmol of uracil in one hr at 37°C) of uracil-DNA-glycosylase (UDG) (Invitrogen) was incubated for 10 min at 37°C in a buffer containing 25 mM Tris-Cl, pH 8.0; 50 mM NaCl and 5 mM EDTA. The reaction was stopped by the addition of 1 µl of 2 M NaOH and heated for 5 min at 95°C. Eleven microliters of formamide were then added and samples were heated at 100°C for 5 min and plunged in an ice water bath. Reaction products were separated on 10% denaturing polyacrylamide gels running in 1X Tris-borate-EDTA (TBE) buffer. Gels were visualized by autoradiography using a phosphor-imager FX (BioRad Laboratories, Hercules, CA) and quantified with Quantity One software (version 4.2). A time course of a typical reaction for two different DNA substrates is shown in Fig. 2.

Results

AID activity assay

Recombinant mouse AID with an N-terminal GST fusion was expressed in baculovirus-infected insect cells and affinity-purified to near homogeneity (Fig. 1A). Purified GST-mAID protein was tested for its ability to deaminate C residues in DNA. Excess uracil glycosylase was added (see Experimental Procedures) in the assay to ensure that AID-mediated deamination is the rate-limiting step in the complete reaction. The difference in AID-mediated
deamination efficiency is directly reflected in the percentage of cleaved product in each reaction. Consistent with earlier studies (13), our GST-mAID protein is active only on C residues located in the ssDNA region (Fig. 3), but not on dsDNA or on an RNA:DNA hybrid (data not shown). GST-mAID activity requires RNase A treatment (Fig. 3A), consistent with one earlier study (13). A time course showed that the deamination reaction plateaus at approximately 15 to 20 minutes (Fig. 2). Therefore, all deamination assays were carried out for 10 min at which time the reaction is still in the linear range.

**DNA length-dependence of the action of AID at cytidines**

In the course of optimizing the assay, we wondered whether the length of the ssDNA substrate would influence the activity of AID at a single C imbedded in the middle of the DNA. Oligonucleotides with identical sequence surrounding the C residue, but differing only in length, show markedly different deamination efficiencies (Fig. 3). A 20-mer gave a minimal 5.3% conversion; while under the same reaction conditions, a 24-mer resulted in 49% conversion. Increasing substrate length to 30 nt further enhanced the efficiency to 62%. Additional studies showed that this length effect plateaus at 26 nt (Fig. 3B). Repositioning of the C residue (AGCA) closer to the 5' or 3' end of the single-stranded DNA substrate, without changing the oligonucleotide length, also showed little effect on deamination efficiency (data not shown). The observed substrate length effect was not conferred by UDG, because it cleaves the synthetic uracil-containing oligonucleotides (mimicking the AID deamination products) with the same efficiency, independent of the length of the substrate (Fig. 3C). It is also known that UDG can act on oligonucleotides as short as three nucleotides (27). The molecular basis for the AID substrate length effect is not known. However, it is an important factor to be taken into consideration when oligonucleotide substrates are designed and as the mechanism of AID is considered at endogenous substrate sites.

**Quantitation of the sequence preferences of AID on single-stranded DNA**

The simple substrate configuration chosen for our studies permits a characterization of how the immediate sequence surrounding the C influences
the activity of AID. We synthesized 27 oligonucleotides to determine AID preferred deamination sites (Fig. 4A). The length of each substrate was 26 nt, sufficient for maximal AID-mediated cytidine deamination. Each substrate had a string of 11 Ts on both sides of the 5'NNCN3' motif. N can be any of the 3 nucleotides, A, G or T. This creates a total of 27 possible combinations. C is excluded in creating the combination to avoid the complication of having two AID targets in the same substrate. Within the NNCN sequence, we named the first position –2, the second position –1, the C, 0, and 3'-most position +1.

These substrates were tested for AID deamination efficiency. We found that all combinations that fit the WRCN motif (W=A or T, R=A or G) gave more than 50% conversion in our assays, in contrast to the non-WRCN substrates (Fig. 4A). The +1 position of N_2N_1C_0N_+1 has a minor but consistent impact on AID deamination activity. We found that a T residue at this position confers a consistently lower activity compared to A or G. This is true for all 9 groups of substrates, whose members differ only at that position. Hence, the optimal substrate is WRCr, where the lower case r indicates that this position has a smaller impact than the WR positions at –2 and –1, respectively. It is important to note that the full range of sequences can be deaminated, despite the preferences noted.

Next, we tested substrates with more than one C. Because the +1 position has a small impact on AID activity, we chose not to vary the nucleotide at this position; hence, seven multiple C substrates were tested with an A maintained constant at the +1 position. A caveat concerning the analysis of oligonucleotides with more than one C is that deamination (and therefore, cleavage) of the C closest to the 5' label will mask the deamination of further downstream C's. Nevertheless, the WRC rule observed above is largely preserved with these substrates (Fig. 4B). In most of the seven double-C substrates, the C residues do not conform to a WRC motif, and these substrates generally have low deamination efficiency. The first C in substrate #30 (TGCC) fits the WRC motif and deaminates very efficiently. However, the second C in substrate #28 (TACC), which also fits the WRC motif, deaminates only
moderately. The other interesting exception is that the second C in substrate #33 (CGCT) gives moderate deamination efficiency, even though it does not conform to the WRC motif (Fig. 4B).

Overall, our deamination comparisons clearly show that AID prefers C residues that are in the WRCr context.

Discussion

We determined the AID site preference using a comprehensive set of fully comparable oligonucleotide substrates. Our data and that of others (23) suggest that the 5'RGYW3'/5'WRCY3' hotspot motif found in somatic hypermutation may be largely due to AID enzyme site preference. Our consensus is WRCr, given a small but consistent preference at the +1 position for purines.

RGYW is generally regarded as the hotspot for somatic hypermutation based on the analysis of endogenous genomic immunoglobulin genes (24). G is the most heavily mutated residue within that motif. The complementary strand of the 5'RGYW3' hotspot is 5'WRCY3'. It is easy to understand the presence of G in the RGYW hotspot because AID deaminates the complementary C in WRCY. However, we did not observe the preference for Y in the WRCY (corresponding to the R in RGYW) motif in our deamination assays. Instead, we found that it is slightly favored to have R at that position. Considering the smaller effect at this position, the hotspot preference determined by our assay is best described as a 5' WRCr 3' motif (and the complementary sequence would be 5' yGYW 3'). We also observed some deviations from the WRC preference. For example, the second C in CGCT, which does not conform to a WRC motif, shows a moderate to high deamination efficiency. This illustrates that while AID activity is influenced by the surrounding sequence, local structural factors probably underlie this in a manner that is not fully described by primary sequence.

Unlike class switch regions, V(D)J regions are unlikely to form RNA:DNA hybrids upon transcription, and hence, the required single-strandedness in the V(D)J region could not be generated by an extensively displaced non-template strand. It has been hypothesized that the non-template strand is single-stranded
transiently at the transcription bubble, which would then serve as a target for AID. Indeed, recent studies have shown that AID preferentially acts on the non-template strand when transcription is carried out in vitro or in E.coli (16,21,23). It is interesting to note, however, that the RGYW motif found at hotspots for SHM is not strand-specific (28,29). Therefore, both strands of the V(D)J region should be equally accessible to AID. Reconciliation of the substrate specificity of AID with the precise structural features of its endogenous substrates awaits further investigation.

A recent study used AID deamination to introduce mutations into a single-stranded LacZα genetic reporter located on an otherwise double-stranded circular substrate (23). That study reached a similar conclusion regarding the sequence preference, namely that AID prefers the WRC motif. The agreement on the sequence preference is stronger given that both a biological readout and a purely biochemical one have yielded similar results.

We observed a DNA substrate length dependence for AID action on single-stranded DNA. Given the small size of AID, 24 kDa, we anticipated that 16 to 20 nt would be sufficient. In fact, > 24 nts is necessary to reach full or nearly-full activity on single-stranded DNA. This surprising length dependence may reflect a larger than expected footprint size of AID, but multimerization of AID is another possibility (14,30).

Acknowledgements. This research was supported by NIH grants to MRL.

References
Figure Legends

**Figure 1.** AID protein and deamination assay.

A. **Purified AID protein.** Recombinant AID protein with an N-terminal GST fusion was purified, resolved on a 10% SDS-polyacrylamide gel, and stained with Coomassie brilliant blue. The identity of each band is indicated.

B. **Scheme of the AID activity assay.** A single-stranded oligonucleotide substrate is labeled at the 5'-end (asterisk). Uracil generated by AID is removed from DNA by UDG, resulting in an abasic (AP) site. DNA cleavage at AP sites is introduced by alkali treatment at elevated temperatures. The cleaved product is separated from the substrate on a denaturing polyacrylamide gel.

**Figure 2.** Time-dependent AID deamination. AID-mediated cytidine deamination was measured every 5 min for the first 30 min. Both a high efficiency (T₁₁AGCAT₁₁, open circles) and a low efficiency (T₁₁GTCTT₁₁, open triangles) substrate were tested. The sequences surrounding the target C residues are listed.

**Figure 3.** Substrate length effect on AID activity

A. **AID deamination assay.** Radioisotope-labeled oligonucleotide substrate (single-stranded) was incubated with recombinant AID protein. Oligonucleotides were heat-cleaved in alkali conditions at the deaminated position. The numbers listed on the right of the gel indicate the length of each oligonucleotide.

B. **Variation of AID activity as a function of substrate length.** The x-axis indicates the ssDNA substrate length. The y-axis indicates the percentage of product resulting from AID deamination (see Experimental Procedures).

C. **UDG activity is independent of substrate length.** UDG removes uracil from synthetic uracil-containing oligonucleotide substrates (equivalent to an AID deamination product). The resulting AP site was then cleaved with alkali treatment. The cleaved product was resolved on a denaturing polyacrylamide gel. The amount of UDG used and the length of the substrates are indicated. Lane 1, 4 and 7, KY414; Lane 2, 5 and 8, KY415; Lane 3, 6 and 9, KY416.

**Figure 4.** Site preference in the AID deamination.
A. Deamination assay with 27 C-containing substrates. Oligonucleotide substrates with sequence variation around the C (N₂N₁C₀N₁⁺) were assayed for deamination efficiency. The exact sequence of each substrate around the C is listed on the top of the gel. A unique number was assigned to each substrate (bottom of the gel). S, substrate; P, product. The histogram shows the average conversion efficiency for each substrate based on three independent experiments. The four groups that fit the WRC motif are emboldened above the histogram.

B. Deamination assay with substrates containing two or more C’s. The sequences surrounding each C are listed on the top of the gel as described in panel A. The deamination efficiencies are listed in the histogram in the order (product 1, 2 and 3) from the longest (C that is furthest away from the radioisotope label) to the shortest product.
Figure 1

A

B

MW

GSTmAID

GST

*  C  AID

*  U  UDG

*  OH⁻
Figure 2
Figure 3
Figure 3C

UDG

0.01U  0.0033U  0.001U

Substrate

Product
Figure 4A
DNA substrate length and surrounding sequence affect the activation induced
deaminase activity at cytidine
Kefei Yu, Feng-Ting Huang and Michael R. Lieber

J. Biol. Chem. published online November 25, 2003

Access the most updated version of this article at doi: 10.1074/jbc.M311616200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 0 references, 0 of which can be accessed free at
http://www.jbc.org/content/early/2003/11/25/jbc.M311616200.citation.full.html#ref-list-1