The Nuclease Domain of Adeno-Associated Virus Rep Coordinates Replication Initiation Using Two Distinct DNA Recognition Interfaces

Alison Burgess Hickman, Donald R. Ronning, Zhanita N. Perez, Robert M. Kotin, and Fred Dyda

Summary

Integration into a particular location in human chromosomes is a unique property of the adeno-associated virus (AAV). This reaction requires the viral Rep protein and AAV origin sequences. To understand how Rep recognizes DNA, we have determined the structures of the Rep endonuclease domain separately complexed with two DNA substrates: the Rep binding site within the viral inverted terminal repeat and one of the terminal hairpin arms. At the Rep binding site, five Rep monomers bind five tetranucleotide direct repeats; each repeat is recognized by two Rep monomers from opposing faces of the DNA. Stem-loop binding involves a protein interface on the opposite side of the molecule from the active site where ssDNA is cleaved. Rep therefore has three distinct binding sites within its endonuclease domain for its different DNA substrates. Use of these different interfaces generates the structural asymmetry necessary to regulate later events in viral replication and integration.

Introduction

Gene therapy is an approach to treating disease in which an exogenous gene is introduced to correct for a defective or missing protein or to affect a biochemical pathway. Few successes have been reported in humans (Hacein-Bey-Abina et al., 2002), as several technical issues limit its broader application. For example, one question is how to deliver DNA to the appropriate cells. Nature provides one solution in the form of viruses, which are in essence protected gene delivery packages with native ability to introduce their genomes into cells. Once the desired gene is delivered to target cells, another issue that arises is the fate of the DNA. Some strategies rely on long-term expression from extrachromosomal DNA, but there are cases, such as dividing cells, where it would be highly beneficial to permanently insert the gene into chromosomes.

Certain viral genomes can be integrated into host DNA by nonhomologous recombination or, in the case of retroviruses, by virally encoded integrases. While integration is not dependent on target sequence, in vivo, retroviruses, such as HIV and murine leukemia virus, integrate preferentially into active genes (Schröder et al., 2002; Wu et al., 2003), introducing the possibility of insertional mutagenesis. The theoretical danger inherent in retrovirus-based gene therapy has been concretely demonstrated in a recent clinical trial in which the modified retrovirus integrated into the LMO2 locus, causing leukemia in two of the patients (Hacein-Bey-Abina et al., 2003).

To date, only one animal virus—the adeno-associated virus (AAV)—has been identified that integrates its genome into a particular location in human chromosomal DNA. When cells are infected in the absence of helper virus, AAV establishes a latent infection in which the AAV genome integrates into a locus known as AAVS1 on the q arm of chromosome 19 (Kotin et al., 1990, 1992; Samulski et al., 1991; reviewed in Smith and Kotin, 2002). This targeted integration by AAV requires the viral rep gene that encodes the nonstructural Rep protein. Rep is an essential viral protein and catalyzes several reactions during viral replication. The N-terminal domain of Rep, consisting of the first ~200 amino acids, has site-specific endonuclease and sequence-specific DNA binding activities (Im and Muzyczka, 1989, 1990; Owens et al., 1993; Chiorini et al., 1994a, 1994b; Davis et al., 2000). The C-terminal domain of Rep possesses ATPase and 3′→5′ helicase activities (Im and Muzyczka, 1990; Chiorini et al., 1994a; Wonderling et al., 1995; Smith and Kotin, 1998) and is a member of the SF3 helicase superfamily (Gorbalyena et al., 1990). The structures of the isolated domains have recently been described (Hickman et al., 2002; James et al., 2003).

The varied activities of Rep are needed for reactions on unique DNA structures known as inverted terminal repeats (ITRs) found at the ends of the single-stranded viral genome that serve as the viral origins of replication (Figure 1). Each ITR contains interrupted palindromic sequences that allow the formation of a three-way DNA junction with two short (~9 bp) hairpin arms. Within the ITRs are two sequences required for replication: a Rep binding site (RBS) consisting of several direct repeats of a 5′-GCTC-3′ motif and a terminal resolution site, or trs. Viral replication requires Repbinding at the RBS and subsequent cleavage of the top strand at the trs to generate the 3′-OH group so that the viral ends can be converted into linear duplex DNA. Cleavage at the trs is strongly stimulated by the presence in cis of one of the ITR hairpin arms (Chiorini et al., 1994a, 1994b; McCarty et al., 1994a; Ryan et al., 1996; Wu et al., 1999; Brister and Muzyczka, 2000; Wu et al., 2001), an effect attributed to a five base sequence known as the RBE, a Rep binding element, at the tip of the hairpin. The endonuclease domain recognizes its trs substrate in the context of ssDNA or a stem loop generated by the Rep helicase activity (Im and Muzyczka, 1990; Snyder et al., 1993; Brister and Muzyczka, 1999; Smith and Kotin, 2000). Two of the important DNA sequences within the ITR, the RBS and the trs, also occur at AAVS1, suggesting that the same Rep-mediated DNA recognition
and cleavage occur during integration. The 33 bp at AAVS1 that encompass these two sequences are necessary and sufficient for targeted integration (Linden et al., 1996).

In current AAV-based transgene systems, rep is deleted since the small viral capsid does not have space for an exogenous gene; retaining rep would result in a genome too large to be packaged (Flotte and Carter, 1995). Thus, rep AAV-based vectors, which retain other inherent advantages of AAV, either do not integrate or integrate inefficiently in a nontargeted manner, a process that has been reported to occur preferentially into active genes (Nakai et al., 2003). Promising alternatives are hybrid vectors that combine AAV ITRs and rep with a second virus with larger transgene capacity, such as adenovirus (Recchia et al., 1999), herpes simplex virus (Costantini et al., 1999), or baculovirus (Palombo et al., 1998).

Integration into a specific chromosomal location would be an extremely valuable asset to gene therapy, as it bypasses targeting problems associated with other delivery systems. Continued improvement of gene delivery systems will clearly benefit from a detailed understanding of how AAV Rep recognizes its DNA substrates: its own genome and the site of integration. To this end, we have determined two crystal structures of the N-terminal domain of AAV5 Rep complexed with specific regions of viral ITRs. These structures provide a foundation for a model for Rep assembly on the ITRs, a necessary first step in both AAV replication and integration.

Results and Discussion

Complex Formation, Crystallization, and Structure Determination

Encouraged by observations that the N-terminal domain of Rep contains the determinants for ITR binding (Owens et al., 1993) and specifically binds ITR sequences (Yoon et al., 2001), we have obtained crystals of AAV5 Rep residues 1-197 separately complexed with a double-stranded 26-mer containing the AAV5 Rep binding site (“RBS26”) and a 15-mer (denoted stem2) consisting of a stem-loop sequence of one of the ITR hairpin arms. This hairpin arm corresponds to the BB palindrome. Although different lengths of DNA incorporating the AAV5 RBS were tried, only the 26-mer yielded crystals.

Stable complex formation between Rep197 and RBS26 was demonstrated by comigration through a Superdex 200 column (data not shown). We similarly attempted to detect complexes of Rep197 with stem-loop sequences (shaded boxes in Figure 1A) representing the two ITR hairpin arms and the proposed hairpins.
Figure 2. The Rep N-Terminal Domain Specifically Binds One Hairpin Arm of the ITR. Shown are gel filtration elution profiles of AAV5 Rep197 with four possible stem loops of the ITR. DNA and protein were mixed in a 0.5:1 molar ratio, so that binding would result in incorporation of all the DNA into a higher molecular weight complex. Only stem2 (middle, right) comigrates with Rep197. The slight broadening of the trsHP peaks may indicate weak binding accompanied by dissociation during gel filtration.

(Bristor and Muzyczka, 1999; Smith and Kotin, 2000) around the trs site. Only the stem loop containing the RBE' (stem2) formed a stable complex (Figure 2).

The structures of both complexes, Rep197/RBS26 and Rep197/stem2, were solved by molecular replacement using uncomplexed AAV5 Rep197 (Hickman et al., 2002) as the search model (see Experimental Procedures and Table 1).

The Structure of Rep197 Bound to RBS26
The asymmetric unit of the Rep197/RBS26 crystals contains one molecule of RBS26 and five Rep197 monomers. The DNA possesses an overall B-DNA structure and is packed blunt end to blunt end in the crystal lattice. The five Rep197 molecules (designated AA–AE in Figure 3A) are bound to RBS26 and spiral around the DNA axis, offset from one another by four base pairs, or ~138°. As a result, each Rep197 molecule binds RBS26 independently. This binding mode reveals a surprisingly intricate structural basis of the observation that a repeated tetranucleotide sequence is necessary for recognition by Rep (Chiorini et al., 1994a; McCarty et al., 1994a). For purposes of discussion, we refer to the GCTC sequence as a “perfect” tetranucleotide repeat.

Recognition of the GCTC Tetranucleotide Repeat
The two structural elements of Rep197 that are important for RBS binding are the surface loop between strands β4 and β5 (residues 135–144) and helix αC (residues 101–118). These regions of Rep197 are located along one edge of the central β sheet, and the bound monomers are oriented such that all of the active sites face in the direction of the trs (Figure 3A).

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Rep197/RBS26</th>
<th>Rep197/stem2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelength (Å)</td>
<td>1.000</td>
<td>1.5418</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>3.1</td>
<td>2.0</td>
</tr>
<tr>
<td>Total reflections (N)</td>
<td>72,828</td>
<td>158,265</td>
</tr>
<tr>
<td>Unique reflections (N)</td>
<td>23,005</td>
<td>38,009</td>
</tr>
<tr>
<td>Completeness (%) (for I/σI &gt; 3)</td>
<td>87.8</td>
<td>89.7</td>
</tr>
<tr>
<td>R/Iσ1</td>
<td>6.2</td>
<td>19.7</td>
</tr>
<tr>
<td>Rsym</td>
<td>0.10</td>
<td>0.063</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Refinement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
</tr>
<tr>
<td>Atoms (N)</td>
</tr>
<tr>
<td>Reflections F &gt; 0 σ(F)</td>
</tr>
<tr>
<td>R factor (%)</td>
</tr>
<tr>
<td>Rfree (%)</td>
</tr>
<tr>
<td>Average B factor (Å²)</td>
</tr>
<tr>
<td>Rms ϕ bonded atoms (Å²)</td>
</tr>
<tr>
<td>Rms bond lengths (Å)</td>
</tr>
<tr>
<td>Rms bond angles (°)</td>
</tr>
<tr>
<td># water molecules</td>
</tr>
<tr>
<td># Mg²⁺ ions</td>
</tr>
</tbody>
</table>

Rey = \sum |I - <I>|/Σ|<I>\}; R factor = \sum |FP_s - FP|/Σ|FP_s|. 
Figure 3. The Rep N-Terminal Domain Binds the RBS with a Four Base Pair Repeat

(A) RBS26 bound by five Rep197 monomers. In each monomer, the active site tyrosine, Tyr-153, is shown in gold and residues that bind the catalytically essential Mg2+ ion in dark blue. Two helices (αB and αC) that form the hairpin binding site (indicated by arrow) are labeled in monomer C.

(B) Major and minor groove interactions involving monomer C. Note that the two phosphate groups forming the ridge between the two grooves are contacted by residues from both sides.

(C) (Top) Schematic representation showing base and phosphate contacts between Rep197 and RBS26. The interactions of each Rep monomer are shown in a different color. Straight lines indicate main chain hydrogen bonds, and arrows indicate side chain-mediated interactions. The
AAV Rep Nuclease Domain Binding to ITR Sequences

Repeat Recognition
The chromosomal AAVS1 and RBS sequences of various AAV serotypes are related but not identical (Figure 3C), and the Rep197/RBS26 structure explains how Rep accommodates variations at certain positions within the tetranucleotide repeat. Two dinucleotide sequences within each GCTC repeat provide the foundation of sequence-specific DNA recognition. The first of these is recognized by Rep from the major groove where main chain carbonyls at the tip of the \( \text{[4]/[5]} \) loop bind a diagonal C-A, i.e., the second base (C) on the top strand and the third base (A) on the bottom strand. The second important dinucleotide consists of the third and fourth bases on the top strand (TC), which are recognized by Arg-106 in the minor groove. As each monomer recognizes two adjacent repeats, we propose that for binding to occur, at least one of these repeats must contain one of these two dinucleotide sequences, a pattern observed in all RBS sequences and at AAVS1.

Two variations on a perfect repeat are contained within RBS26. In repeat “A,” the C within the diagonal C-A dinucleotide is replaced with T. Although this results in the loss of a hydrogen bond, the overall mode of Rep binding is the same as for perfect repeats. In repeat “D,” the C of the -TC- dinucleotide is replaced by G. The structural consequence of this substitution is a change in the side chain orientation of Arg-106; to conserve the total number of protein/DNA bonds, Arg-106 shifts by one base pair to now bridge the fourth base of repeat “D” and the first of repeat “E.” Thus, the ability of Rep monomers to bind imperfect perfect repeats reflects not only the fact that two adjacent monomers participate in repeat recognition but also the redundancy in the binding interactions (since interactions can be lost, as in repeat A) and their flexibility (seen in the movement of Arg-106 in repeat D).

Interestingly, the distribution of perfect repeats within RBS sequences is not random: repeat “C” is always perfect and is flanked on at least one side by another perfect repeat. This pattern may be a tactic to dictate tight binding to the center of the RBS and weaker binding to the repeats on the flanks, which are generally imperfect. It is worth noting that AAVS1 contains three perfect repeats followed by the GCTG variation seen in repeat D of the AAV5 RBS; thus, our structure provides the structural information necessary to model Rep binding to the chromosomal integration site. To date, only AAV2 has been shown to integrate, but the high sequence similarity between Rep proteins from different serotypes indicates that the mode of DNA recognition will likely be the same.

Conservation of Interacting Residues of Rep197
Of the eight Rep197 residues that form direct side chain interactions with the RBS, seven are strictly conserved among serotypes AAV2-6. Lys-137 of AAV5 is substi-

---

**Repeat Recognition**

The chromosomal AAVS1 and RBS sequences of various AAV serotypes are related but not identical (Figure 3C), and the Rep197/RBS26 structure explains how Rep accommodates variations at certain positions within the tetranucleotide repeat. Two dinucleotide sequences within each GCTC repeat provide the foundation of sequence-specific DNA recognition. The first of these is recognized by Rep from the major groove where main chain carbonyls at the tip of the \([4]/[5]\) loop bind a diagonal C-A, i.e., the second base (C) on the top strand and the third base (A) on the bottom strand. The second important dinucleotide consists of the third and fourth bases on the top strand (TC), which are recognized by Arg-106 in the minor groove. As each monomer recognizes two adjacent repeats, we propose that for binding to occur, at least one of these repeats must contain one of these two dinucleotide sequences, a pattern observed in all RBS sequences and at AAVS1.

Two variations on a perfect repeat are contained within RBS26. In repeat “A,” the C within the diagonal C-A dinucleotide is replaced with T. Although this results in the loss of a hydrogen bond, the overall mode of Rep binding is the same as for perfect repeats. In repeat “D,” the C of the -TC- dinucleotide is replaced by G. The structural consequence of this substitution is a change in the side chain orientation of Arg-106; to conserve the total number of protein/DNA bonds, Arg-106 shifts by one base pair to now bridge the fourth base of repeat “D” and the first of repeat “E.” Thus, the ability of Rep monomers to bind imperfect perfect repeats reflects not only the fact that two adjacent monomers participate in repeat recognition but also the redundancy in the binding interactions (since interactions can be lost, as in repeat A) and their flexibility (seen in the movement of Arg-106 in repeat D).

Interestingly, the distribution of perfect repeats within RBS sequences is not random: repeat “C” is always perfect and is flanked on at least one side by another perfect repeat. This pattern may be a tactic to dictate tight binding to the center of the RBS and weaker binding to the repeats on the flanks, which are generally imperfect. It is worth noting that AAVS1 contains three perfect repeats followed by the GCTG variation seen in repeat D of the AAV5 RBS; thus, our structure provides the structural information necessary to model Rep binding to the chromosomal integration site. To date, only AAV2 has been shown to integrate, but the high sequence similarity between Rep proteins from different serotypes indicates that the mode of DNA recognition will likely be the same.

**Conservation of Interacting Residues of Rep197**

Of the eight Rep197 residues that form direct side chain interactions with the RBS, seven are strictly conserved among serotypes AAV2-6. Lys-137 of AAV5 is substi-
Figure 4. Binding of AAV2 Rep68 and AAV2 Rep197 to AAVS1 and RBS Sequences
(A) Gel filtration elution profiles demonstrate the formation of a single, stable Rep68/AAVS1 complex at protein:DNA ratios ranging from 6:0.25 to 6:3. Using a fixed amount of protein, increasing amounts of DNA were added resulting in increasing amounts of a single complex and no free DNA, until saturation was reached at ~6:1 Rep197:DNA and ~7.5:1 Rep68:DNA. Thereafter, the addition of more DNA did not result in the formation of higher-order complexes. Excess unbound protein precipitates under these experimental conditions and, therefore, is not seen in the chromatograms.

(B) Data obtained from gel filtrations as described in (A) was plotted as the amount of complex formed (filled squares represent integrated peak areas in arbitrary units) as a function of the DNA:protein ratio. The A260:A280 nm ratio for the complex formed at each protein-to-DNA ratio was also calculated (blue open circles) based on integrated peak areas. Top and middle plots correspond to AAV2 Rep68 binding to 32-mers containing the AAVS1 Rep binding site and the RBS, respectively; the bottom plot represents binding of AAV2 Rep197 to the RBS 32-mer. Binding of AAV2 Rep197 to the AAVS1 32-mer is essentially identical to the bottom plot (data not shown).

Assembly of a Multimeric Complex on Viral RBS Sequences
In the Rep197/RBS26 complex, five Rep197 monomers are observed bound to the five direct repeats of RBS26. Although several lines of biochemical evidence indicate that two consecutive tetranucleotide repeats are sufficient for Rep binding as long as flanking DNA is present (Chiorini et al., 1995; Wonderling and Owens, 1997), the presence of at least four repeats within the RBS sequences of serotypes AAV2-6 (Figure 3C) suggests that two repeats are not enough to assemble a functional complex. Up to six different bound species can be detected by gel shift (McCarty et al., 1994b) and chemical crosslinking (Smith et al., 1997). Furthermore, dinucleotide transversion mutations anywhere within the five repeats in the AAV2 RBS result in an increase in $K_d$ for Rep68 binding; while the most marked effects are centered over the middle three perfect repeats, the limits correspond to the 20 bp of the five repeats (Ryan et al., 1998). Our structural result that each Rep monomer binds to two sequential repeats and each repeat is recognized by elements from two monomers suggests that, at the edges of the RBS, one recognizable repeat next to a random tetranucleotide sequence might be bound by two Rep monomers.

To address the issue of stoichiometry of Rep binding at RBS sequences and to confirm that our structure of AAV5 Rep197/RBS26 is relevant to the more thoroughly biochemically characterized AAV2 Rep, we performed a series of binding studies using purified AAV2 Rep68 (Chiorini et al., 1994b) and AAV2 Rep197 on 32-mers containing AAVS1 and AAV2 RBS sequences. A single, stable complex was detected when a fixed amount of protein and increasing amounts of dsDNA were mixed and subsequently analyzed by gel filtration (Figure 4).
As the amount of DNA was increased, the amount of complex increased until a saturating level was reached at a protein:DNA molar ratio of ~1:0.17 (or 6:1) for Rep197 and ~1:0.13 (7:5:1) for Rep68; this difference is likely within the error of this nonequilibrium method, but may also reflect slight degradation of Rep68 (data not shown). The A260:A280 nm of each complex was constant at all protein-to-DNA ratios examined, suggesting that under these experimental conditions a single species was present throughout, a result also suggested by the constant elution position of the complexes. These results are therefore consistent with the notion that both the isolated N-terminal domain and full-length Rep assemble as multimeric complexes, most likely hexamers, on viral and AAVS1 sequences. The A260:A280 nm at all protein-to-DNA ratios examined, suggesting that under these experimental conditions a single species was present throughout, a result also suggested by the constant elution position of the complexes. These results are therefore consistent with the notion that both the isolated N-terminal domain and full-length Rep assemble as multimeric complexes, most likely hexamers, on viral and AAVS1 sequences.

The Structure of Rep197 Bound to stem2
Rep197 and the DNA stem loop corresponding to the RBE-containing hairpin arm of the AAV5 ITR (stem2) form a 1:1 complex in which stem2 is bound to an extensive positively charged surface on the opposite side of Rep197 from the active site (Figure 5A). Only three consecutive bases, T7–T9, interact directly with the protein. The base-paired stem of the hairpin extends away from Rep197, orthogonal to the protein surface. Apart from some side chain rotamer changes, stem2 binding does not induce significant conformational changes in Rep, and superposition of uncomplexed and complexed Rep results in an rmsd of 0.58 Å in Ca positions.

Figure 5. The N-Terminal Domain of Rep Specifically Recognizes the Tip of One of the Hairpin Arms
(A) Surface representation of Rep197/stem2. For reference, the two structural elements responsible for RBS binding, αC and the β4-β5 loop, are also indicated.
(B) Bases C6-G10 showing hydrogen bonding interactions. Water molecules are shown as blue spheres.

Description of the Interface
As shown in Figure 5B, stem2 contains a three-membraned loop in which T7–T9 are unpaired and the flanking bases C6 and G10 form a Watson-Crick bp, which stacks on the remaining base pairs of the stem. T7 is flipped out, while T8 and T9 are in conformations that allow continued helical stacking of the stem into the loop in the 5'-3' direction. This differs substantially from the NMR structure of the free AAV2 stem loop (Chou et al., 2000) in which the first T of the loop is folded into the minor groove, and the second and third T bases point in parallel toward the major groove. Since the RBE sequences are identical between AAV2 and AAV5 and the amino acid residues at the Rep197/stem2 interface are highly conserved among serotypes, it is likely that Rep binds a conformation that is not the predominant conformation in solution.

The flipped-out base T7 is inserted into a preformed surface pocket of Rep197 assembled by the convergence of several residues, reminiscent of the mode of substrate binding by methyltransferases and uracil-DNA glycosylases (Roberts and Cheng, 1998). One wall of the pocket is formed by Trp-29 onto which T7 is stacked at a distance of 3.5 Å. All of the potential hydrogen bond donors and acceptors in T7 are used, as T7 forms hydrogen bonds with Nε of Arg-58, the carbonyl oxygen of Val-119, and the amide side chain of Gln-121. The other two T bases in the loop have far fewer protein interactions. T8 is stacked between bases C6 and T9 and forms only a hydrogen bond with Arg-58. T9, sitting in a shallower pocket than T7, is stacked on Tyr-65 at a distance of 3.4 Å and there is a single water-mediated hydrogen bond between O4 and the side chains of Glu-66 and Gln-114. The DNA backbone conformation is sharply twisted between C6 and T7 as indicated by large changes in α and γ backbone torsion angles. Although the C6-G10 bp retains Watson-Crick hydrogen bonding, it deviates substantially from coplanarity with an inclination of 27° and buckle of 18°.

Four residues of Rep197 (Trp-29, Tyr-65, Arg-58, and
Gln-121) form direct side chain contacts with the three T bases in the loop, and the side chains of Arg-58 and Arg-61 also participate in backbone phosphate binding. Among serotypes AAV2-6, Trp-29, and Arg-61 are strictly conserved; Arg-58 and Gln-121 are conservatively substituted by Lys and Arg, respectively, in all other serotypes, while Tyr-65 is the most variable residue, appearing as Thr in AAV2 and Val in AAV3, 4, and 6. This lack of conservation is consistent with residue 65 providing van der Waals contacts and generic bulk to pack against the third base of the loop. To confirm the importance of residues Trp-29, Arg-58, Arg-61, and Tyr-65 in contributing to stem2 binding, we mutated each individually to Ala; Arg-58 and Arg-61 were also mutated to Gln. As shown in Figure 6, none of the point mutants was able to bind stem2.

Significance of the RBE' Sequence
Methylation interference experiments and DNaseI protection assays indicate that AAV2 Rep preferentially contacts one of the two hairpin arms of the ITR and that this arm corresponds to the one furthest away from the trs, whether the ITR is in the flip or the flop orientation (Ashktorab and Srivastava, 1989; Im and Muzyczka, 1989). Ryan et al. (1996) showed that five bases, 5'-CTTTG-3', at the tip of one of the hairpin arms particularly contribute to Rep binding. The low efficiency of trs cleavage on linear duplex substrates compared to that on substrates with intact ITRs is unaffected by the addition of the hairpin arms in trans (Chiorini et al., 1994b), suggesting that the Rep/RBE' interaction stimulates cleavage by imposing a geometric constraint on Rep assembled on the RBS. It is possible that this stimulation is indirect and that the Rep/RBE' interaction is important primarily to generate a single-stranded substrate rather than to stimulate the chemical steps of DNA cleavage (Brister and Muzyczka, 2000). This is consistent with the structure, which shows that stem-loop binding does not induce any conformational changes at the enzyme active site.
The fine tuning of the cleavage activity by the RBE plays an important role in Rep’s ability to discriminate between fully replicated dsDNA viral genomes—which still possess RBS and trs sequences—and genomes whose terminal hairpin arms identify them as unreplicated. Nevertheless, the Rep/RBE interaction must be modulatory rather than essential; otherwise, Rep would not be able to cleave the trs sequence present at the chromosomal AAVS1 site and the virus would be unable to integrate.

Model of Proposed Assembly of Rep on ITRs

For its size (~20 kDa), the AAV Rep endonuclease domain exhibits a remarkable capacity to interact with different DNA substrates. There is an active site for trs cleavage located on one surface, a hairpin binding site on the opposite side, and a region in between that is used for sequence-specific RBS binding. Satisfyingly, Rep molecules bind RBS DNA such that all of the active sites are oriented in the direction of the trs, and the surface that binds stem2 is necessarily oriented toward its hairpin arm substrate. This arrangement also provides an elegant explanation for the results of RBS polarity mutation studies, which demonstrated that various orientations of the RBS sequence do not significantly affect Rep binding, yet only the wild-type orientation is able to direct efficient trs cleavage (Brister and Muzyczka, 2000).

It has been established that DNA unwinding at the trs is catalyzed by the Rep helicase activity, that trs cleavage occurs on ssDNA, and that Rep must remain bound to the RBS to cleave the trs efficiently (Brister and Muzyczka, 2000). AAV2 Rep can hexamerize when bound to dsDNA containing the RBS, and each integral protein-DNA complex can be detected (Smith et al., 1997). The Rep197/RBS26 structure and the binding data in Figure 4 reveal that this reflects Rep N-terminal domains binding to RBS DNA and does not necessarily imply anything about the helicase multimeric state.

Our structure of the Rep197/RBS26 complex suggests that binding of several Rep N-terminal domains to the RBS is the initiating step in the assembly of an active complex on the viral ITR, as shown in Figure 7. We propose that the binding of N-terminal domains stimulates the subsequent assembly of the C-terminal helicase domains into a hexamer at the site where unwinding activity is required to generate ssDNA at the trs. The observation that Rep binds more tightly to the RBS when it is flanked by nonspecific DNA (Chiorini et al., 1994b; Weitzman et al., 1994) is consistent with the need for room to accommodate another DNA binding domain. DNaseI footprinting data indicating protection of the region between the trs and the RBS (Im and Muzyczka, 1989) clearly suggests that this is where the helicase assembles. Once the helicase is assembled, DNA can be unwound, possibly in steps that resemble the “iris” mechanism recently proposed on the basis of the SV40 large T antigen helicase structure (Li et al., 2003). Experimental support for this is provided by the observation that AAV2 Rep68 can bind and unwind fully duplex blunt-ended DNA provided it contains an RBS (Zhou et al., 1999). A particularly appealing aspect of the proposed parallel with the SV40 T antigen origin melting mechanism is that extrusion of ssDNA through a side channel of the helicase would cause the trs sequence to pass close to the active sites of the Rep N-terminal domains during its transit from the interior.

One challenge in reconciling the Rep197/RBS26 structure with the properties of a helicase is that it is difficult to envisage a planar hexameric ring of C-terminal domains physically linked to six bound N-terminal
domains spiraling along the DNA axis, and we assume that DNA bending must occur. We also note that the expected conformation of the three-way junction in the ITR—coaxial stacking between two arms and an antiparallel orientation of the third arm (Altorna et al., 1996; see also Figure 7)—would in principle allow one of the bound Rep197 monomers to simultaneously bind the tip of one of the terminal hairpins. This would provide a mechanism to explain the reported stimulation of trs cleavage by the RBE sequence, as this might introduce asymmetry, leading to enhanced DNA bending and unwinding that might facilitate helicase activity and subsequent cleavage.

The model in Figure 7 is specific for complex assembly at ITRs during viral replication. The role played by the stem loop is not needed in other contexts; for example, at AAVS1 where there are no stem-loop sequences and yet where a Rep complex also assembles to direct integration. It is possible that the N-terminal domains are not irreversibly bound to the RBS and that, upon helicase domain multimerization, N-terminal domains can transiently dissociate from the DNA. Thus, one monomer within the complex might be bound only to the stem loop rather than simultaneously to both the RBS and the hairpin sequence. More intriguingly, liberated N-terminal domains would be free to bind to other DNA molecules. Our identification of two interfaces on the surface of the N-terminal domain capable of binding DNA provides an obvious mechanism for bridging two pieces of DNA that contain RBS sequences or stem-loop sequences, a necessary step during the process of site-specific integration.

Experimental Procedures

Protein Purification and Crystallization

AAV Rep1-197 (Rep197) was purified as previously described (Hickman et al., 2002), dialyzed against 20 mM Tris (pH 7.5), 0.5 M NaCl, and concentrated to ~10 mg/ml. Oligonucleotides were from Integrated DNA Technologies, Inc. (Coraville, IA). Site-specific point mutations were introduced using the QuikChange kit (Stratagene), and the entire coding region was sequenced to confirm that no additional mutations had been introduced. The elution times on size-exclusion chromatography were essentially identical for each of the point mutants and corresponded to that of wild-type Rep197, suggesting that the overall structure of the protein was unperturbed. Histidine-tagged versions of AAV2 Rep68 and AAV2 Rep 1–197 were similarly purified.

For the RBS26 substrate, PAGE-purified oligonucleotides (5'-CAGCCTTGCCTGCTGCGGCTCTGCGT-3' and 5'-CAGGAGCCAGC GAGCGAACGCG-3') were resuspended in TE (10 mM Tris [pH 8], 1 mM EDTA), mixed in a 1:1 molar ratio, heated to 90°C for 10 min and then cooled slowly to 20°C. Rep197 and RBS26 were mixed in a 3:1 molar ratio and dialyzed into 20 mM Tris (pH 7.5), 0.12 M NaCl. Initial crystals were grown at 4°C by the hanging drop method by mixing the protein/DNA complex in a 1:1 ratio with 20% (v/v) PEG 3000, 0.1 M sodium citrate (pH 5.5). Two rounds of microseeding into 14% (v/v) PEG 3350, 0.1 M sodium citrate yielded diffraction-quality crystals.

NaCl. Crystals were grown at 4°C by the hanging drop method by mixing the protein/DNA complex in a 1:1 ratio with 20% (v/v) PEG 3000, 0.1 M sodium citrate (pH 5.5). Two rounds of microseeding into 14% (v/v) PEG 3350, 0.1 M sodium citrate yielded diffraction-quality crystals. Crystals were cryoprotected by transfer into 10 mM Tris (pH 7.5), 60 mM NaCl, 18% (v/v) PEG 3350, 50 mM sodium citrate (pH 5.5), 60% RBS26, 5% (v/v) ethylene glycol, and the ethylene glycol concentration then slowly increased to 20% (v/v). Crystals were flash cooled by immersion in liquid propane. The space group was P2_1_2_1 (a, 78.67 Å; b, 131.71 Å; c, 82.16 Å; β, 112.7°).

For the stem2 substrate, the 15-mer oligonucleotide (5'-CAGCCTTGCCTGCTGCGGCTCTGCGT-3') was resuspended in TE, heated to 90°C for 15 min, and then rapidly cooled on ice. Rep197 and stem2 were mixed in a 1:1.1 molar ratio and dialyzed into 20 mM Tris (pH 7.5), 0.12 M NaCl. Crystals were grown at 4°C by the hanging drop method by mixing the protein/DNA complex in a 1:1 ratio with 20% (v/v) PEG 3000, 0.1 M sodium citrate (pH 5.5). Two rounds of microseeding into 14% (v/v) PEG 3350, 0.1 M sodium citrate yielded diffraction-quality crystals. Crystals were cryoprotected by transfer into 10 mM Tris (pH 7.5), 60 mM NaCl, 18% (v/v) PEG 3350, 50 mM sodium citrate (pH 5.5), 60% RBS26, 5% (v/v) ethylene glycol, and the ethylene glycol concentration then slowly increased to 20% (v/v). Crystals were flash cooled by immersion in liquid propane. The space group was P2_1_2_1 (a, 78.67 Å; b, 131.71 Å; c, 82.16 Å; β, 112.7°).

For the stem2 substrate, the 15-mer oligonucleotide (5'-CAGCCTTTAGGCT-3') was resuspended in TE, heated to 90°C for 15 min, and then rapidly cooled on ice. Rep197 and stem2 were mixed in a 1:1.1 molar ratio and dialyzed into 20 mM Tris (pH 7.5), 0.12 M NaCl. Crystals were grown at 4°C by the hanging drop method by mixing the protein/DNA complex in a 1:1 ratio with 20% (v/v) PEG 3000, 0.1 M sodium citrate (pH 5.5). Two rounds of microseeding into 14% (v/v) PEG 3350, 0.1 M sodium citrate yielded diffraction-quality crystals. Crystals were cryoprotected by transfer into 10 mM Tris (pH 7.5), 60 mM NaCl, 18% (v/v) PEG 3350, 50 mM sodium citrate (pH 5.5), 60% RBS26, 5% (v/v) ethylene glycol, and the ethylene glycol concentration then slowly increased to 20% (v/v). Crystals were flash cooled by immersion in liquid propane. The space group was P2_1_2_1 (a, 78.67 Å; b, 131.71 Å; c, 82.16 Å; β, 112.7°).
ACGGCCAGAGGGCCGT-3'; stem2: 5'-CAGCTCTTTTGAAGCTG-3'; stem3: 5'-GAGTGCCACACTC-3'; trshP: 5'-GAGTGTCAGCTC-3').

For the AAVS1 (5'-TTCGGGTCGCCGGCGCTCGTCTGCGTCGGC-3' and its reverse complement) and AAV2 RBS32 (5'-CCC TCTCTGGGGGCGCTCGTCTCAGTTGAGGC-3' and its reverse complement) assays, oligonucleotides were annealed and binding assessed as described above except that the buffer contained 0.12 M NaCl. Purified AAV2 Rep68 was held constant at 57.8 μM; AAV2 Rep197 at 66.9 μM.

Acknowledgments

We thank P. Musingamiru for the gift of AAV2 Rep197; S. Chacko for helpful discussions; W. Yang and E. Greene for insightful suggestions; N. Leyaronska for assistance with data collection at SERCAT; and T. Chiou for help with Curves. Use of the Advanced Photon Source was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Office of Science under Contract No. W-31-109-Eng-38. This work was supported in part by the NIH Intramural AIDS Targeted Antiviral Program. Coordinates have been deposited with the Protein Data Bank.

Received: July 30, 2003
Revised: December 8, 2003
Accepted: December 9, 2003
Published: February 12, 2004

References


Wonderling, R.S., and Owens, R.A. (1997). Binding sites for adeno-