Structural and Functional Homology of Parvovirus and Papovavirus Polypeptides

By CAROLINE R. ASTELL, CLIFFORD D. MOL and WAYNE F. ANDERSON

Department of Biochemistry, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5 and Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2H7

(Accepted 17 November 1986)

SUMMARY

We have compared the sequences of the putative polypeptides of the human pathogenic B19 parvovirus with protein sequences in the National Bethesda Research Foundation Library, and have discovered a significant homology between a B19 parvovirus non-structural (NS) protein and the T antigens of polyomaviruses and simian virus 40 (SV40) and the putative E1 proteins of papillomaviruses. The region of highest homology with the papovavirus proteins corresponds to the region that is most highly conserved in the NS1 proteins of several other parvoviruses. Studies with the T antigen of both polyomaviruses and SV40 have implicated this region as having an ATPase activity and nucleotide-binding function.

The Parvoviridae are a family of small viruses with single-stranded, linear DNA genomes approximately 5000 nucleotides in length (Siegl, 1984). The two groups of vertebrate parvoviruses are the dependoviruses [e.g. adeno-associated virus (AAV)] which require a helper adenovirus or herpesvirus for replication, and the autonomously replicating parvoviruses [e.g. minute virus of mice (MVM), H-1] (Siegl, 1984). Human B19 virus, which has recently been characterized as a parvovirus (Summers et al., 1983; Cotmore & Tattersall, 1984), is the causative agent of erythema infectiosum and a type of reticulocytopenic aplastic crisis associated with several types of haemolytic anaemias. This virus is known to cause a transient, post-infection arthropathy in adults and to cross the placenta, eliciting an immune response in the foetus (see Shade et al., 1986 and Cotmore et al., 1986 for references). DNA of the Augusta isolate (B19-Au) has been cloned (Cotmore et al., 1986) and sequenced (Shade et al., 1986). In addition, major structural and non-structural proteins encoded by the human parvovirus B19 genome have been mapped (Cotmore et al., 1986). Organization of the B19 genome is similar to that of the prototype dependovirus genome (AAV2) and the autonomous virus genome (MVM). Coding regions are located exclusively on one strand, the plus strand. While data on the transcription of B19 are not yet available, by analogy with AAV2 and MVM (Carter et al., 1984), there are likely multiple, overlapping transcription units. The nucleotide sequence of the B19 genome suggests four possible promoters (P1, P2, P3 and P4, Fig. 1; Shade et al., 1986). The large open reading frame (ORF) in the right-half of the genome encodes viral structural polypeptides of which two [VP1 (mol. wt. 83000, 83K) and VP2 (58K)] have been identified, while the large ORF in the left-half of the genome most likely encodes three non-structural polypeptides [NS1 (71K), NS2 (63K) and NS3 (52K)] (Cotmore et al., 1986).

Hybridization studies of the B19-Wi clone showed that there is no detectable DNA homology with AAV2 DNA and only weak homology with the autonomous viral genomes such as MVM, H-1, Kilham rat virus and LuIII (Cotmore & Tattersall, 1984). However, when the sequence of the B19 genome became available and putative polypeptide sequences deduced (Shade et al.,
Fig. 1. Molecular organization of the human B19 parvovirus genome. The organization of the plus or coding strand is given. The major left-hand ORF (nucleotides 427 to 2445) encodes non-structural polypeptides NS1, NS2 and NS3, while the major right-hand ORF (nucleotides 2441 to 4787) encodes non-structural polypeptides VP1 and VP2 (Shade et al., 1986; Cotmore et al., 1986). Analysis of the nucleotide sequence has tentatively identified four promoter sequences (P1 to P4). The P1 promoter has now been characterized by in vitro transcription (M. C. Blundell, C. Beard & C. Astell, personal communication). The region within the non-structural protein(s) homologous to the T antigens and E1 proteins of papovaviruses is indicated by a solid box (region B) and hatched box (region C). mu, Map units; nt, nucleotides.

1986), it was possible to investigate whether B19 is related to other parvoviruses by comparing the amino acid sequence of the polypeptides. The DNA SEQNCE program (Delaney, 1982), which identifies perfect matches, detected a highly conserved 145 amino acid sequence in the middle of the putative NS1 genes of B19, AAV2 and MVM (Shade et al., 1986). The level of homology is 51% (B19/AAV2), 41% (B19/MVM) and 51% (MVM/AAV2). Shorter regions within this sequence showed homologies as high as 92% (11 out of 12 amino acids identical). Comparison of the amino acid sequence encoded by the right-hand ORF of B19 detected similar homologies. At the time these studies were done, a search of the National Bethesda Research Foundation Library (NBRFL) data bank failed to identify any significant homologies with other proteins (Shade et al., 1986).

Other computer programs have become available which greatly facilitate the search of databases and also allow matches in which conservative amino acid changes are recognized. One such program is FASTP (Lipman & Pearson, 1985). We have analysed the sequences of the NS1 proteins of the parvoviruses (B19, AAV2 and MVM) using FASTP and shown that the homologies between these proteins are not limited to short regions, but extend throughout most of the proteins (Table 1). The regions of highest homology correspond to the regions identified earlier (Shade et al., 1986). A comparison of the entire NS1 region of AAV2 and B19 (Fig. 2) illustrates the extent of these homologies.

We have also used the FASTP program to search the NBRFL data bank and find significant homology between the putative NS1 protein of B19 virus and other viral proteins including the T antigen of mouse (Soeda et al., 1980) and hamster (Delmas et al., 1985) polyomaviruses and T antigen of simian virus 40 (SV40) (Fiers et al., 1978; Reddy et al., 1978). In addition, there is also significant homology with the putative E1 protein of papillomaviruses [human papillomavirus type 1a, HPV1a (Chen et al., 1982), bovine type 1 (Danos et al., 1983) and the cottontail rabbit (Shope) virus (Schwartz et al., 1983)]. These homologies are summarized in Table 1 and a comparison of the region of highest homology with the T antigen of mouse polyomavirus and the putative E1 protein of HPV1a is illustrated in Fig. 3. In Table 1, the I score (initialized score) indicates the extent of the homology and the O score (optimized score) is that obtained when insertions and deletions are permitted to maximize the homology. It also ignores dissimilar portions of the amino acid sequence outside the optimized alignment region such that they do not affect the score of the alignment (Lipman & Pearson, 1985). An increase in the O score over the I score indicates the homology is likely to be significant. Of particular interest is that the region of highest conserved sequence between the parvovirus NS1 proteins (Fig. 2) is precisely that region homologous with the T antigens and E1 proteins (Fig. 3).
Short communication

b:bl9 left

Fig. 2. Homology of the major non-structural (NS1) proteins of B19 (b:bl9 left) and AAV2 (UYADIA), using the FASTP computer program (Lipman & Pearson, 1985). The colon symbol identifies perfect matches and the single dot identifies analogous matches. The putative NS1 protein of B19 is 671 amino acids long (encoded by nucleotides 436 to 2449) (Shade et al., 1986). The NS1 protein of AAV2 is 621 amino acids long (encoded by nucleotides 321 to 2184) (Srivastava et al., 1983). (Note that the amino acid sequence of AAV2 in the NBRFL assumes an in-frame splice from nucleotides 1906 to 2228, resulting in a protein of only 536 amino acids. For this comparison we have not included the in-frame splice; hence, the AAV2 NS1 protein is 621 amino acids long.)
Fig. 3. Homology of the conserved regions within the NS1 protein of B19, T antigen of mouse polyoma virus (PY-MS), and probable E1 protein of HPV1a. The regions of homology were initially aligned using the program FASTP (Lipman & Pearson, 1985). Subsequently, minor adjustments were made to permit the line up of HPV1a and PY-MS to conform to that published by Clertant & Seif (1984). This adjustment introduces an 11 amino acid gap in the B19 NS1 protein sequence (opposite the 10 amino acid gap in the HPV1a E1 protein sequence). Homology is observed in all three proteins within the B region (likely ATPase functional region) and C region (likely nucleotide binding site): B and C are bracketed and the areas within these regions which are more highly conserved are indicated with a heavy bar. We have retained a solid triangle used by Clertant & Seif (1984) to indicate an identical amino acid and an open triangle to indicate an analogous amino acid in the six papovavirus proteins analysed. In addition, a solid circle indicates that this amino acid is conserved in all six papovavirus proteins as well as the NS1 protein of B19 (Shade et al., 1986). An open circle indicates analogy at this position. A solid diamond indicates that the amino acid is identical in the NS1 proteins of six parvoviruses: B19 (Shade et al., 1986), AAV2 (Srivastava et al., 1983), H-1 (Rhode & Paradiso, 1983), MVM (Astell et al., 1983b), canine parvovirus (Rhode, 1985a), and feline panleukopenia virus (Carlson et al., 1985). It is of interest that there is one conserved lysine (K) residue at position 133 in domain C which may be the δ-NH₂ group in the nucleotide binding site of T antigens that is labelled with oxATP (Clertant et al., 1984).

Studies by Seif (1984) recognized sequence homology between the large tumour antigen of polyomaviruses and the putative E1 protein of papillomaviruses. Clertant & Seif (1984) showed that these regions of homology, located in the carboxy-terminal halves of the proteins, can be divided into four adjacent regions of homology: A, B, C and D. The two more highly conserved regions are B, which is approximately 100 amino acids long, and C, which is approximately 60 amino acids long. An ATPase activity associated with the SV40 large T antigen is inhibited by monoclonal antibodies which bind to a region of the protein extending from amino acids 417 to 509 (Lane & Hoeffler, 1980; Clark et al., 1981, 1983). These residues include much of the B domain (amino acids 381 to 481) as well as the first 20 residues of the C domain (amino acids 487 to 552) (Clertant & Seif, 1984). The binding site of one monoclonal antibody which inhibits ATPase activity, PAb204, has now been mapped to lie within amino acids 448 to 509 (Mole & Lane, 1985) which still includes the COOH end of the B domain as well as the NH₂ end (about 20 amino acids) of the C domain. However, because the binding site for this antibody is now outside the consensus sequence for many ATPases (Walker et al., 1982; and see below), which is located in the B domain of the T antigens, it is clear that an antibody can inhibit activity by inducing conformational changes to the antigen, and not simply by physically blocking the active site. A nucleotide binding site detected by affinity labelling of polyomavirus and SV40 large T antigens is distinct from the ATPase catalytic site, and has been mapped to cyanogen bromide peptides (amino acids 413 to 527 of SV40 T antigen and amino acids 637 to 674 of mouse polyoma virus T antigen) which include the C region (Clertant et al., 1984). Our studies comparing the homologous regions of parvovirus and papovavirus non-structural proteins (Fig.
Table 1. Homology of putative B19 NS protein with other parvovirus NS proteins and papovavirus T antigen and E1 polypeptides using FASTP

<table>
<thead>
<tr>
<th>Polypeptide (a) Parvovirus NS proteins</th>
<th>I score</th>
<th>O score</th>
<th>fold increase</th>
<th>Level of homology and extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probable NS proteins of adeno-associated virus (AAV2)</td>
<td>269</td>
<td>562</td>
<td>2-1×</td>
<td>33-0% identity in 409 aa* overlap</td>
</tr>
<tr>
<td>Probable NS1 protein of minute virus of mice (MVM)</td>
<td>121</td>
<td>263</td>
<td>2-2×</td>
<td>26-9% identity in 290 aa overlap</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Polypeptide (b) Papovavirus T antigens and E1 polypeptides</th>
<th>I score</th>
<th>O score</th>
<th>fold increase</th>
<th>Level of homology and extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large T antigen of hamster polyomavirus</td>
<td>63</td>
<td>81</td>
<td>1-3×</td>
<td>32-0% identity in 50 aa overlap</td>
</tr>
<tr>
<td>Large T antigen of mouse polyomavirus</td>
<td>55</td>
<td>64</td>
<td>1-2×</td>
<td>24-8% identity in 121 aa overlap</td>
</tr>
<tr>
<td>Large T antigen of SV40</td>
<td>40</td>
<td>59</td>
<td>1-5×</td>
<td>23-7% identity in 135 aa overlap</td>
</tr>
<tr>
<td>Probable E1 protein of cottontail rabbit (Shope) papillomavirus</td>
<td>56</td>
<td>108</td>
<td>1-9×</td>
<td>24-8% identity in 121 aa overlap</td>
</tr>
<tr>
<td>Probable E1 protein of human papillomavirus (type 12)</td>
<td>53</td>
<td>73</td>
<td>1-4×</td>
<td>24-1% identity in 133 aa overlap</td>
</tr>
<tr>
<td>Probable E1 protein of human papillomavirus type 1a</td>
<td>51</td>
<td>93</td>
<td>1-8×</td>
<td>21-1% identity in bovine papillomavirus, 190 aa overlap</td>
</tr>
</tbody>
</table>

* aa, Amino acid.

3) reveal that six residues (positions 39, 47, 48, 52, 53 and 80) in region B and four residues (positions 133, 144, 148 and 150) in region C are identical among six papovaviruses and six parvoviruses. In addition, ten residues are analogous (open diamond) in B and six in C (Fig. 3).

The molecular organizations including the entire nucleotide sequences of SV40 and mouse polyomavirus were among the first to be determined (Soeda et al., 1980; Delmas et al., 1985; Fiers et al., 1978; Reddy et al., 1978). T antigens of SV40 and polyomaviruses are multi-functional proteins involved in DNA replication and regulation of transcription. T antigen has been shown to induce cellular DNA synthesis and the expression of cellular proteins including enzymes involved in DNA metabolism (Tooze, 1981). T antigen is necessary for the expression of the transformed phenotype (Tooze, 1981) and more recently ATPase (Clark et al., 1981, 1983) and nucleotide-binding activities (Clertant & Seif, 1984) have been associated with these proteins. A DNA-binding domain of T antigen has also been mapped (Paucha et al., 1986). Nucleotide sequences of several papillomaviruses are now known including the prototype virus bovine papillomavirus type 1 (BPV-1) (Chen et al., 1982). Papillomaviruses all have a double-stranded circular DNA genome (about 8000 bp) and share a common genetic organization (Clertant & Seif, 1984). The largest NS polypeptide, E1, is the most conserved between these viruses and amino acid sequences of the putative E1 proteins show significant homologies in the carboxy-terminal halves, extending over 200 amino acids. These regions are predicted to share similar secondary structures and correspond to the regions of large T antigen of polyomavirus and SV40 sites believed to be the ATPase and nucleotide-binding sites, although the role(s) of the ATPase activity and nucleotide binding site are not clear. E1 protein of BPV-1 is required for plasmid maintenance in transformed cells (Lusky & Botchan, 1984). Although the function of the NS1 proteins of both AAV2 and the autonomous parvoviruses is still uncertain, they seem to be multi-functional polypeptides. In AAV2 the gene product of the major left-hand ORF is necessary for viral DNA replication (Hermonat et al., 1984; Tratschin et al., 1984) and NS1 protein of H-1 parvovirus has been implicated as a transactivator of the P38 promoter used for transcription of mRNA encoding viral structural proteins (Rhode, 1985b). So far, there is no
Table 2. Homology of putative B19 NS protein with other proteins using a sequence comparison program designed to locate structurally similar regions*

(a) Comparison of two sequences

<table>
<thead>
<tr>
<th></th>
<th>B19 NS</th>
<th>MVM NS</th>
<th>AAV NS</th>
<th>HPV E1</th>
<th>Shope E1</th>
<th>SV40 TAg</th>
<th>Mouse TAg</th>
<th>Hamster TAg</th>
</tr>
</thead>
<tbody>
<tr>
<td>B19 NS</td>
<td>3·3 × 10⁵</td>
<td>2·8 × 10¹²</td>
<td>1600</td>
<td>3·7 × 10⁵</td>
<td>690</td>
<td>620</td>
<td>4400</td>
<td></td>
</tr>
<tr>
<td>MVM NS</td>
<td></td>
<td>3·4 × 10⁸</td>
<td>780</td>
<td>150</td>
<td>950</td>
<td>1·7 × 10⁵</td>
<td>6·2 × 10⁴</td>
<td></td>
</tr>
<tr>
<td>AAV NS</td>
<td>2·8 × 10¹²</td>
<td></td>
<td>270</td>
<td>19</td>
<td>170</td>
<td>2800</td>
<td>8500</td>
<td></td>
</tr>
<tr>
<td>HPV E1</td>
<td>1600</td>
<td>780</td>
<td>270</td>
<td></td>
<td>2·9 × 10¹²</td>
<td>250</td>
<td>330</td>
<td>8·4</td>
</tr>
<tr>
<td>Shope E1</td>
<td>3·7 × 10⁵</td>
<td>150</td>
<td>19</td>
<td>2·9 × 10¹²</td>
<td></td>
<td>250</td>
<td>880</td>
<td>20</td>
</tr>
<tr>
<td>SV40 TAg</td>
<td>690</td>
<td>950</td>
<td>170</td>
<td>250</td>
<td>250</td>
<td>1·7 × 10¹²</td>
<td>1·3 × 10¹⁹</td>
<td></td>
</tr>
<tr>
<td>Polyoma TAg (mouse)</td>
<td>620</td>
<td>1·7 × 10⁵</td>
<td>2800</td>
<td>330</td>
<td>880</td>
<td>1·7 × 10¹²</td>
<td>–</td>
<td>9·8 × 10¹⁸</td>
</tr>
<tr>
<td>Polyoma TAg (hamster)</td>
<td>4400</td>
<td>6·2 × 10⁴</td>
<td>8500</td>
<td>8·4</td>
<td>20</td>
<td>1·3 × 10¹⁹</td>
<td>9·8 × 10¹⁸</td>
<td></td>
</tr>
<tr>
<td>E. coli dnaA</td>
<td>420</td>
<td>6·2</td>
<td>18</td>
<td>1·5</td>
<td>0·5</td>
<td>0·5</td>
<td>0·2</td>
<td>0·4</td>
</tr>
<tr>
<td>E. coli recA</td>
<td>170</td>
<td>2·6</td>
<td>16</td>
<td>0·6</td>
<td>46</td>
<td>8·3</td>
<td>8·6</td>
<td>25</td>
</tr>
</tbody>
</table>

(b) Comparison of three sequences

<table>
<thead>
<tr>
<th></th>
<th>B19 NS + SV40 TAg + dnaA 1·1 × 10⁵</th>
<th>B19 NS + SV40 TAg + recA 1·8 × 10⁶</th>
<th>B19 NS + dnaA + recA 5·5 × 10⁴</th>
<th>SV40 TAg + dnaA + recA 1·2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1·1 × 10⁵</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

* Each pair of sequences (a) or set of three (b) was compared using a probe length of 20 residues without any insertions or deletions. The values in the table are the observed frequency of the best score (F_obs) divided by the expected frequency (F_calc) calculated using the matching probability and the amino acid compositions of the pair of proteins being compared (McLachlan, 1971; Bacon & Anderson, 1986). Simultaneously aligning three sequences (b) improves the signal to noise so that the F_obs/F_calc increases if the three sequences are related. For unrelated sequences this ratio is < 1·0 (Bacon & Anderson, 1986).
Short communication 891

evidence to support the proposal that this protein may be the nicking enzyme responsible for resolving dimer replicative form molecules (Astell et al., 1983a), although AAV2 DNA concatemeric intermediates accumulate when AAV polypeptide synthesis is restricted (E. Sebring & J. Rose, personal communication). In this paper, we have shown that the regions of the ATPase (B) and nucleotide binding activities (C) in the papovavirus proteins are precisely those most highly conserved regions in the parvovirus NS1 proteins (Fig. 3). Hence, we predict that at least two additional functions of these parvovirus NS1 proteins are an ATPase activity and nucleotide binding activity. Other similarities between the T antigens and NS1 proteins of paroviruses are that both have a nuclear location and both exist in phosphorylated forms (Tooze, 1981; Paradiso, 1984; Cotmore & Tattersall, 1986). However, one apparent difference is that while the amino-terminal end of the T antigens is involved in cellular transformation, paroviruses are not known to transform cells (Siegl, 1984) and our computer homology studies show no homology with the amino-terminal end of T antigens.

It should be noted that the B regions of the NS1 proteins which are most highly conserved among paroviruses (Fig. 2) and which share homology with papovavirus NS proteins (Fig. 3) include the consensus sequence, GX4GKT/sX6 l/v, recognized by Walker et al. (1982) in several ATP-requiring enzymes. This sequence is found at the active site of many, although not all, ATPases. More recent listings of proteins containing this consensus sequence are given in Husain et al. (1986) and Fry et al. (1986).

Further studies utilizing another sequence comparison program designed to locate structurally similar regions (Bacon & Anderson, 1986) also show that the parvovirus NS1 proteins and papillomavirus E1 proteins all have a very significant homology with the B region of SV40 T antigen (Table 2). Comparison of B19 NS1 with a number of other proteins did not reveal any statistically significant similarity to the helix-turn-helix motif that often occurs in prokaryotic proteins which regulate transcription (Anderson et al., 1982; Sauer et al., 1982; Pabo & Sauer, 1984) or to a consensus β-α-β unit characteristic of nucleotide binding enzymes (Wierenga et al., 1986). This search did, however, reveal that residues 316 to 351 (33 to 71 in Fig. 3) of B19 NS1 that are equivalent to the SV40 T antigen B region, exhibit a statistically significant similarity to residues 160 to 195 of the Escherichia coli dnaA protein and to residues 54 to 89 of the E. coli RecA protein. The significance of this similarity is further supported by the fact that the optimal simultaneous alignment of segments of 20 residues from these three sequences gives a much higher F_{obs}/F_{calc} ratio than any of the pairwise comparisons of these sequences (Table 2). The biological significance of the proposed structural similarity is difficult to assess at this time. However, it is interesting that the dnaA protein is involved in the initiation of replication, as is the SV40 antigen (although it is not reported to have an ATPase activity), and that the recA protein has a DNA-dependent ATPase activity (Weinstock et al., 1981). The similarity between the B19 NS1 protein and the recA protein is greater than that between the SV40 T antigen and recA which was noted before (Seif, 1982) and which is not much above random expectation. It will be of interest to determine whether the B19 NS protein is an ATPase and whether it is involved in the initiation of DNA replication.

We would like to thank J. Hassell for helpful discussions and G. Mauk for his help with running the FASTP program. This work was supported by grants from the MRC (Canada) and BCHCRF (to C.R.A.) and through the MRC Group on Protein Structure and Function (to W.F.A.).

REFERENCES


(Received 14 July 1986)