The Complete Nucleotide Sequence of Coxsackievirus B4 and Its Comparison to Other Members of the Picornaviridae

By OWEN JENKINS, JOHN D. BOOTH, PHILIP D. MINOR AND JEFFREY W. ALMOND

1Department of Microbiology, University of Reading, London Road, Reading RG1 5AQ and 2National Institute for Biological Standards and Control, Holly Hill, London NW3 6RB, U.K.

(Accepted 30 March 1987)

SUMMARY

The genome of the prototype strain of coxsackievirus B4 (J.V.B. Benschoten) has been cloned in Escherichia coli and its complete nucleotide sequence determined. Excluding the poly(A) tract, the RNA genome is 7395 nucleotides in length and appears to encode a single polyprotein of 2183 amino acids. The predicted amino acid sequence of the polyprotein shows close homology (88%) to that of the previously sequenced coxsackievirus B3 and to certain regions of the polyproteins of the polioviruses and human rhinovirus 14. This allows identification of putative polyprotein cleavage signals, antigenic domains and other structural features likely to be important to the biological integrity of the virus.

INTRODUCTION

Coxsackieviruses are members of the enterovirus genus of the family Picornaviridae. They are divided into 23 group A serotypes (CA1 to CA22, CA24) and six group B serotypes (CB1 to CB6). This antigenically diverse group of viruses has a limited host range, but exhibits a broad spectrum of tissue tropism within the natural host and is associated with a correspondingly large variety of clinical illnesses, ranging from mild respiratory infections to severe myocarditis and neurological disorders (Grist et al., 1978; Melnick, 1985). The group B coxsackieviruses and especially CB4 are of particular interest in that they have been associated with type 1 juvenile onset or insulin-dependent diabetes mellitus (Yoon et al., 1979; King et al., 1983; Barret-Connor, 1985; Frisk et al., 1985). The molecular basis of the tissue tropism of these viruses is not well understood, although there is some evidence to suggest that for the picornaviruses generally, a major determinant is the affinity of the viruses for a specific cellular receptor on the surface of the target cell (see Crowell et al., 1981).

Knowledge of the primary structure and genetic organization of picornaviruses has increased dramatically in recent years, and this has provided insights into the mechanisms of replication (Kitamura et al., 1981), genetic relationships (Stanway et al., 1984a), evolution (Toyoda et al., 1984), antigenicity (Minor et al., 1983, 1986) and pathogenicity (Stanway et al., 1984b; Cann et al., 1984; La Monica et al., 1986; Evans et al., 1985). All picornaviruses share common structural features, namely an approximately 30 nm capsid of icosahedral symmetry, made up of 60 copies each of four virus-coded proteins (VP1 to VP4) enclosing a single-stranded, positive-sense RNA genome of approximately 7500 nucleotides. The RNA is polyadenylated at its 3' terminus and has a small protein, VPg, covalently attached to the 5' terminus. The primary translation product of picornavirus RNA is a single large polyprotein which is processed by virus-encoded proteases to yield the mature viral proteins (for a review, see Rueckert, 1985).

Recently the complete nucleotide sequence of CB3 (strain Nancy) has been determined (Lindberg et al., 1987), providing detailed information on coxsackievirus genome organization and allowing a comparison at the nucleotide level with other members of the Picornaviridae.

As part of a study into the molecular basis of serotype diversity, pathogenicity and tissue tropisms of these viruses, we have determined the complete nucleotide sequence of the genome...
of the prototype strain of CB4 (J.V.B. Benschoten). We provide a detailed comparison of this sequence with that of the closely related CB3 and with those of other members of the Picornaviridae. The high degree of similarity to poliovirus (PV) type 1 and rhinovirus 14 (HRV14) also allows interpretation of the predicted amino acid sequence data in relation to the three-dimensional structures recently determined for these two viruses (Hogle et al., 1985; Rossmann et al., 1985).

METHODS

Virus and cells. Coxsackievirus B4 (strain J.V.B. Benschoten) (Dalldorf, 1950) was obtained from the American Type Culture Collection. The virus was propagated in HEp-2c cells and purified on sucrose gradients as previously described (Minor, 1980).

Molecular cloning and nucleotide sequencing. Purified viral RNA (approximately 2 μg) was reverse-transcribed and cloned into Escherichia coli JA221 by the cDNA:RNA hybrid method (Cann et al., 1983; Stanway et al., 1984c) using vector pBR322. Of the recombinants obtained, approximately 1000 were screened by hybridization using 3'-enriched and randomly primed CB4 cDNA probes (Cann et al., 1983). Plasmid DNA isolated from strongly hybridizing colonies was further characterized by restriction enzyme mapping and by cross-hybridization. A set of five overlapping clones were selected which together spanned the genome (data not shown). The sequences of cDNA inserts were determined by the dideoxynucleotide method after generation of random fragments and cloning into M13mp8 as previously described (Stanway et al., 1984b). This method was used to obtain the nucleotide sequence of the majority of the genome. The sequence of the remainder was obtained after cloning specific restriction fragments into M13mp18 or mp19. The whole of the sequence was determined at least twice and approximately 75% of it was obtained in both orientations. Throughout its assembly the sequence was compared to those of other enteroviruses and this provided a useful check on possible frameshift sequencing errors. Where significant differences were observed these were checked in the opposite orientation. The sequence data were assembled and analysed using published computer programs (Staden, 1980).

RESULTS AND DISCUSSION

The complete nucleotide sequence and predicted amino acid sequence of CB4 (strain J.V.B. Benschoten) are shown in Fig. 1. The genome is highly homologous to those of other enteroviruses, suggesting that it has a similar genetic organization. Thus, the genome comprises a 5' non-coding region of 743 nucleotides, a single open reading frame of 6552 nucleotides (2184 codons) and a 3' non-coding region of 100 nucleotides prior to a poly(A) tract. Hence the total size of the CB4 genome excluding the poly(A) tract is 7395 nucleotides (compared with CB3, 7396; PV1, 7433; PV3, 7432; HRV14, 7208 nucleotides) and it has the % base composition A, 28.30; G, 24.84; C, 22.91; T, 23.95.

5' non-coding region

By analogy with the polioviruses, the 5' terminal 743 nucleotides of the CB4 genome are assumed to be non-coding (Kitamura et al., 1981). There are, however, seven potential translation start codons prior to that which initiates the large open reading frame at position 744. Four of these are quickly followed by stop codons. Those at positions 272 and 463 are followed by open reading frames of 66 codons and 110 codons respectively. The latter AUG is also present in CB3 and PV3 but the size of the reading frame is not conserved, being terminated after 76 amino acids at position 686 in CB3 and after 44 amino acids at position 593 in PV3. These similarities may be fortuitous though it is likely that they relate to sequence-dependent functions other than translation. It is considered unlikely that any of these short open reading frames are translated since the corresponding peptides have not been found in PV-infected cells. Furthermore, none of the AUGs have flanking nucleotide sequences favoured by eukaryotic ribosomes for the initiation of protein synthesis (Kozak, 1986).

The 5' non-coding regions of CB4, CB3, PV1, PV3 and HRV14 can be aligned by taking into account several small deletions or insertions to give remarkable sequence homology (Table 1, Fig. 2). The first 10 nucleotides are identical in all these viruses. This conservation of sequence may be important for interactions with proteins involved in the replication of the viral RNA (Lindberg et al., 1987; Hewlett & Florkiewicz, 1980; Toyoda et al., 1984). Other regions of pronounced conservation include nucleotide positions 66 to 85, 446 to 472 and 547 to 567 and
Sequence of coxsackievirus B4

Fig. 1. The complete nucleotide sequence and predicted amino acid sequence of the polyprotein of CB4 (J.V.B. Benschoten). Amino acid differences from the previously sequenced CB3 (Nancy), are circled. Predicted polyprotein cleavage sites are arrowed.

Table 1. Nucleotide sequence homologies between the non-coding regions of CB4 and four other picornaviruses*

<table>
<thead>
<tr>
<th></th>
<th>CB4 : CB3</th>
<th>CB4 : PV1</th>
<th>CB4 : PV3</th>
<th>CB4 : HRV14</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' non-coding</td>
<td>84.4</td>
<td>71.4</td>
<td>71.0</td>
<td>62.8</td>
</tr>
<tr>
<td>3' non-coding</td>
<td>93.8</td>
<td>59.1</td>
<td>59.1</td>
<td>45.2</td>
</tr>
</tbody>
</table>

* Sequence homologies are expressed as percentages. Details of the other virus sequences taken from Lindberg et al. (1987) for CB3, Kitamura et al. (1981) for PV1, Stanway et al. (1984b) for PV3 and Stanway et al. (1984a) for HRV14.

these correspond to those previously described as highly conserved between HRV14 and PV3 (Stanway et al., 1984a). It has been suggested that these conserved sequences may be important for RNA secondary structure which is necessary for an, as yet, unspecified function in replication (Tracy et al., 1985; Toyoda et al., 1984; Newton et al., 1985). Of interest in this respect is a region from nucleotides 10 to 34 which in the polioviruses has the potential to form a stable stem-loop secondary structure (Larsen et al., 1981; Stanway et al., 1983). In the corresponding region, CB4 is not well conserved in primary sequence but can form a similar stem-loop structure composed of exactly the same number of GC and AT base pairs. This stem-loop is likely to have functional importance in replication since it has been shown for PV2 that deletion of base 10 (which destabilizes the structure) leads to a mutant virus with a temperature-sensitive phenotype. This mutation appears to affect the rate of protein synthesis at both permissive and non-permissive temperatures (Racaniello & Meriam, 1987). Interestingly the corresponding regions in HRV14 and HRV2 do not appear to form such structures readily,
suggesting that the function is not indispensable. The hundred or so nucleotides prior to the initiation of translation of the polyprotein, as in the polioviruses, are poorly conserved and it is unlikely that this region has any functional significance, although it has been suggested that it may play a role in the conservation of the length of the 5' non-coding region (Toyoda et al., 1984). This idea does not hold for the rhinoviruses, however, where the region appears to have been completely deleted.

**Translated region**

Translation of the CB4 RNA probably initiates at nucleotide 744. The initiation codon at this position forms part of the sequence AAAAUGG, which is an almost optimal translation initiation sequence for eukaryotic ribosomes (Kozak, 1986). In this frame, there are no termination codons until nucleotide 7293. The region therefore can encode a polyprotein of 2183 amino acids, consistent with the known replication strategy of the picornaviruses (see Rueckert, 1985). The amino acid homology between the predicted proteins of CB4 and those of CB3, PV1, PV3 and HRV14 are shown in Table 3 and represented diagrammatically in Fig. 3. The homology with PV1, for which a detailed genetic map has been determined biochemically (Pallansch et al., 1984; Kuhn & Wimmer, 1987) facilitates the identification of the sites in the CB4 polyprotein at which the virus-encoded proteases are likely to act. These are presented in Table 2 together with the assumed cleavage sites of CB3, PV3, HRV14, and the determined cleavage sites of PV1. With the exception of that at the P2-B/P2-C junction all of the cleavage sites in PV1 are conserved in CB4 and are therefore likely to be utilized. To identify the likely cleavage site between P2-B and P2-C, a comparison with CB3 is helpful. Lindberg et al. (1987)
Table 3. Amino acid sequence homology between the proteins of CB4 and four other picornaviruses*

<table>
<thead>
<tr>
<th>Protein</th>
<th>CB4 : CB3</th>
<th>CB4 : PV1</th>
<th>CB4 : PV3</th>
<th>CB4 : HRV14</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP4</td>
<td>94.2</td>
<td>69.6</td>
<td>71.0</td>
<td>60.8</td>
</tr>
<tr>
<td>VP2</td>
<td>80.5</td>
<td>56.0</td>
<td>56.3</td>
<td>54.6</td>
</tr>
<tr>
<td>VP3</td>
<td>78.2</td>
<td>55.9</td>
<td>55.9</td>
<td>46.8</td>
</tr>
<tr>
<td>VP1</td>
<td>71.2</td>
<td>47.3</td>
<td>44.6</td>
<td>35.7</td>
</tr>
<tr>
<td>Total P1</td>
<td>77.9</td>
<td>54.4</td>
<td>54.1</td>
<td>47.9</td>
</tr>
<tr>
<td>P2-A</td>
<td>91.2</td>
<td>57.8</td>
<td>59.1</td>
<td>44.4</td>
</tr>
<tr>
<td>P2-B</td>
<td>97.9</td>
<td>50.5</td>
<td>51.5</td>
<td>55.7</td>
</tr>
<tr>
<td>P2-C</td>
<td>97.6</td>
<td>62.6</td>
<td>62.9</td>
<td>58.7</td>
</tr>
<tr>
<td>P3-A</td>
<td>92.1</td>
<td>47.1</td>
<td>50.6</td>
<td>48.2</td>
</tr>
<tr>
<td>VPg</td>
<td>90.9</td>
<td>77.3</td>
<td>72.7</td>
<td>46.9</td>
</tr>
<tr>
<td>Protease (P3-C)</td>
<td>95.0</td>
<td>60.7</td>
<td>60.6</td>
<td>52.7</td>
</tr>
<tr>
<td>Polymerase (P3-D)</td>
<td>95.2</td>
<td>74.0</td>
<td>73.9</td>
<td>65.9</td>
</tr>
</tbody>
</table>

* Sequence homologies are expressed as percentages. Details taken from Lindberg et al. (1987) for CB3, Kitamura et al. (1981) for PV1, Stanway et al. (1984b) for PV3 and Stanway et al. (1984a) for HRV14.

have proposed that in CB3 this cleavage occurs either at a tyrosine–glycine pair or a glutamine–asparagine pair which by alignment correspond to those in CB4 at nucleotide positions 4013 and 4037 respectively (Fig. 1, Table 2). In PV1 this cleavage occurs at a glutamine–glycine pair which by alignment corresponds exactly to the glutamine–asparagine pair in CB4. The amino acid sequence following this site is highly conserved between CB4, CB3 and PV1, whereas the amino acid sequence around the tyrosine–glycine pair (although conserved between CB4 and CB3) shows no homology to the corresponding region in the polioviruses. Moreover glutamine–asparagine has been shown by amino acid sequencing to be a cleavage site in HRV2 (Skern et al., 1985). We conclude therefore that the glutamine–asparagine pair probably functions as the cleavage site for P2-B/P2-C in both CB3 and CB4.

Of interest also is the PI/P2 cleavage site occurring between a tyrosine and glycine residue in HRV14, and all serotypes of poliovirus (Table 2). The sequence comparisons indicate that this site is conserved in CB4 and is therefore likely to be used, but that it is not present in CB3 (Fig. 1, Table 2). Four alternative sites have been proposed for the corresponding cleavage in CB3 (Tracy et al., 1985; Lindberg et al., 1987) (Table 2). Unfortunately the two coxsackieviruses are very different in this region and the CB4 sequence does not therefore help to distinguish between the four possibilities for cleavage in CB3.

Cleavage of the poliovirus precursor polypeptide VP0 to give VP4 and VP2 occurs at an asparagine–serine amino acid pair and is thought to be an autocatalytic event (Hogle et al., 1985). This site is also conserved at an equivalent position in the CB4 polyprotein (nucleotide position 950) and is likely therefore to function in a similar manner (Fig. 1).

Verification of these proposed polyprotein processing sites in CB4 would, of course, require N- and C-terminal protein sequence analysis of infected cell and virus structural polypeptides. However, the high degree of homology to the other enteroviruses in these regions of the polyprotein provides persuasive evidence that the cleavage sites identified in Table 2 for CB4 are correct. The sizes of the predicted CB4 proteins are summarized in Table 4.

Structural proteins

The P1 region of the polyprotein forms the precursor to the four structural proteins VP1 to VP4. As expected, and in common with the rhinoviruses and most other enteroviruses, VP4 is the most conserved of the capsid proteins. In PV1 and HRV14, VP4 is not exposed on the outer surface of the virion (Hogle et al., 1985; Rossmann et al., 1985) and does not contribute to antigenic domains important in virus neutralization (Minor et al., 1986). This protein is therefore unlikely to be subjected to immune selection pressure. In the three other structural proteins, VP1, VP2 and VP3, there are regions that are highly conserved, interspersed with regions of substantial divergence (Fig. 3). Regions of highest homology are those likely to be of importance in maintaining the stable secondary structure of the capsid proteins and these
Fig. 3. Comparison between the capsid proteins of CB4 (J.V.E. Benschoten), CB3 (Nancy), PV1 (Mahoney), PV3 (P3/Leon/37) and HRV14. Conserved amino acids are represented by black boxes. Antigenic sites determined for PV3 are shown. Regions identified in the 3-D structure of PV1 as $\alpha$-helices (A), $\beta$-sheets (B) and random coil or loop-out regions (L) are indicated.
Sequence of coxsackievirus B4

Table 4. CB4 proteins and their sizes*

<table>
<thead>
<tr>
<th>Protein</th>
<th>Number of amino acids</th>
<th>Mol. wt. ($\times 10^{-3}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP4</td>
<td>69</td>
<td>7.49</td>
</tr>
<tr>
<td>VP2</td>
<td>261</td>
<td>28.56</td>
</tr>
<tr>
<td>VP3</td>
<td>238</td>
<td>26.43</td>
</tr>
<tr>
<td>VP1</td>
<td>284</td>
<td>31.98</td>
</tr>
<tr>
<td>P2-A</td>
<td>147</td>
<td>16.22</td>
</tr>
<tr>
<td>P2-C</td>
<td>329</td>
<td>37.27</td>
</tr>
<tr>
<td>P3-A</td>
<td>89</td>
<td>9.96</td>
</tr>
<tr>
<td>VPg</td>
<td>22</td>
<td>2.41</td>
</tr>
<tr>
<td>Protease (P3-C)</td>
<td>183</td>
<td>20.33</td>
</tr>
<tr>
<td>Polymerase (P3-D)</td>
<td>462</td>
<td>52.53</td>
</tr>
<tr>
<td>Total polyprotein</td>
<td>2183</td>
<td>243.96</td>
</tr>
</tbody>
</table>

* Based on the cleavage sites proposed in Table 2.

correspond, by alignment with PV1, to $\alpha$-helices and $\beta$-sheets in the three-dimensional structure (Hogle et al., 1985). By contrast, the regions which show least homology correspond to the more flexible loop-out or random coil secondary structures, some of which are known to be antigenic determinants in polioviruses and rhinoviruses (Minor et al., 1986; Rossmann et al., 1985; Skern et al., 1987; Sherry & Rueckert, 1985; Fig. 3). Thus it is likely that in CB4 the amino acids 76 to 87 in VP1, 158 to 164 in VP2 and 58 to 60 and 70 to 71 in VP3 are antigenically important. Although other amino acids are also likely to contribute to antigenicity, the imprecision of the alignment in these regions of random coil makes their identification less reliable.

Non-structural proteins

The P2 region is the part of the polyprotein most highly conserved between the two coxsackieviruses B3 and B4. There are only two amino acid differences between them in the P2-B protein and only eight differences in the P2-C protein (Fig. 1, Table 3). The functions of both of these proteins are unknown, but there is evidence that P2-C forms part of the membrane-bound replication complex (Takegami et al., 1983). Mutations conferring guanidine resistance in PV1 have also been mapped to this protein (Pincus et al., 1986). P2-A has been shown to function as a protease responsible for cleavage of the polyprotein at tyrosine-glycine cleavage sites (Toyoda et al., 1986). As discussed above, the amino acid pair tyrosine-glycine is present at the P1/P2 junction in CB4, PV1, PV3 and HRV14 (Table 2) but is not present in CB3. It is therefore interesting to note that P2-A of CB4 is substantially more homologous to the P2-A of CB3 (91%) even though their substrates are different, than to those of the other viruses (less than 60%) in which the substrate cleavage site is the same.

The P3 region is processed to give P3-AB (which is subsequently processed to provide VPg), a protease (P3-C) and an RNA-dependent RNA polymerase (P3-D). The VPgs of coxsackieviruses B1, B3 and B5 have been compared at the amino acid level (Lindberg et al., 1987). The VPg sequence of CB4 differs from those of CB1 and CB5 by only one amino acid, and differs from the CB3 sequence by two. These changes are conservative. As is the case in all enteroviruses both the protease (P3-C) and polymerase (P3-D) of coxsackieviruses are highly conserved (Fig. 1, Table 3), the polymerase being the most highly conserved protein of those specified by picornavirus genomes (Table 3; Argos et al., 1984).

3' non-coding region

The translation of the polyprotein is terminated at position 7293 by the sequence UAA which is followed by a non-coding region of 100 nucleotides prior to the poly(A) tract (Fig. 1). This region is very highly conserved between CB4 and CB3 (94%, Table 1) and also between the three serotypes of polioviruses (98%, Toyoda et al., 1984) but among the enteroviruses and rhinoviruses as a whole it is far less well conserved than the 5' non-coding region. The function of this 3' non-coding region in picornaviruses has not yet been determined, although it is likely to
Figure 4. (a) Alignment of the 3' non-coding regions of CB4 (J. V. B. Benschoten), CB3 (Nancy) and PV3 (P3/Leon/37). Conserved nucleotides are shown in upper case. (b) Predicted RNA secondary stem–loop structure of this region. Conserved nucleotides underlined in (a) form part of the stem.

be involved in the control of genome replication (Fellner, 1979). The conserved blocks of nucleotides reported previously following comparisons of CB3, PV1 and swine vesicular disease virus (Stalhandske et al., 1984) are also present in CB4 (Fig. 4). A number of different stem–loop secondary structures can be compiled for this region (Ryan, 1985) in which the conserved blocks of nucleotides form part of the stem. Occasionally sequence divergence within the stem is observed (e.g. in enterovirus 70; M. Ryan, unpublished), but base pairing is always maintained by compensatory mutations across the stem. It has been shown that the insertion of an eight nucleotide linker at position 7387 in the 3' non-coding region of PV1 gives rise to a virus with a temperature-sensitive phenotype suggesting that the structure of the region has some essential function. The temperature-sensitive phenotype may result from destabilization of the secondary structure at the restrictive temperature (Sarnow et al., 1986). It is interesting that HRV14 and HRV2 have much smaller 3' non-coding regions than the enteroviruses. The conserved block of nucleotides closest to the poly(A) tract (Fig. 4) is deleted and it is therefore unlikely that the rhinoviruses are able to form similar secondary RNA structures. The significance of this observation is unclear, but it is interesting to note the parallel with the 5' non-coding region as discussed above where the loop formed by nucleotides 10 to 34 is also missing in the rhinoviruses. These features may be consistent differences between the rhinoviruses and enteroviruses (Stanway et al., 1984a).
Sequence of coxsackievirus B4

It is clear that CB4 is very closely related to CB3 and to other members of the enterovirus genus. This high level of homology, particularly with the better studied members of the Picornaviridae such as PV1 and HRV14, allows a comprehensive interpretation of the sequence to be made. Thus, polypeptide cleavage signals can be confidently located and likely antigenic domains tentatively identified. These comparisons are useful in that they also indicate regions of sequence difference which must ultimately determine the characteristic biological properties which differ between picornaviruses. The possibility to manipulate these regions via site-directed mutagenesis of cloned cDNA provides an experimental approach to understanding the molecular basis of picornavirus diversity. Such experiments are in progress.

This work was supported by the Medical Research Council, project grant number G106/253.

REFERENCES


*(Received 6 February 1987)*