Complete nucleotide sequence of the *Escherichia coli recB* gene

Paul W. Finch, Alan Storey, Karen E. Chapman*, Kate Brown, Ian D. Hickson¹ and Peter T. Emmerson

Department of Biochemistry, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU and ¹Department of Clinical Oncology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK

Received 10 September 1986; Accepted 14 October 1986

**ABSTRACT**

The complete nucleotide sequence of the *Escherichia coli recB* gene which encodes a subunit of the ATP-dependent DNase, Exonuclease V, has been determined. The proposed coding region for the RecB protein is 3343 nucleotides long and would encode a polypeptide of 1180 amino acids with a calculated molecular weight of 133,973. The start of the recB coding sequence overlaps the 3' end of the upstream pfr gene, and the recB termination codon overlaps the initiation codon of the downstream recD gene, suggesting that these genes may form an operon. No sequences which reasonably fit the consensus for an *E. coli* promoter could be identified upstream of the proposed recB translational start. The predicted RecB amino acid sequence contains regions of homology with ATPases, DNA binding proteins and DNA repair enzymes.

**INTRODUCTION**

The recB and recC genes of *Escherichia coli* code for subunits of Exonuclease V (1-3), which is required for genetic recombination, efficient repair of DNA and maintenance of cell viability (4-6). The enzyme unwinds double-stranded DNA to produce single-stranded loops (7,8) which are cleaved predominantly adjacent to Chi sequences, (5'-OCTGGTGG-3') (9,10), known to locally stimulate genetic recombination (see [11] for a review) via the RecBC pathway (12). The enzyme possesses a number of other activities including exonuclease activity on single- and double-stranded DNA, and endonuclease activity on single-stranded DNA. Both the unwinding and the nuclelease activities of the enzyme require concomittant hydrolysis of ATP (see [13] for a review).

The recB and recC genes have been cloned and their products identified as proteins of approximately 135 kDa and 125 kDa respectively (14-16). The genes are physically closely linked (4,5) and can be isolated on a 19 kb BamHI fragment of the *E. coli* chromosome (15,16). Maxicell analysis of recombinant plasmids containing this fragment has demonstrated that pfr, the structural gene for Protease III, lies between recC and recB (16).
An understanding of the mechanisms of action of the individual components of Exonuclease V will depend on an analysis of the specific interactions between the different subunits of the enzyme both with each other and with DNA. For such studies, a knowledge of the primary sequences of the individual proteins will be necessary. As a first step in this study we have determined the sequence of the region of the E. coli chromosome between thyA and argA, which includes the reoB and reoC genes. Analysis of this sequence should also give an insight into possible mechanisms by which the expression of these genes is controlled. We have previously determined the entire sequence of the thyA-reoC intergenic region (17), the reoC gene (17) and the complete pfr gene (28). Here, we report the complete nucleotide sequence of the reoB gene and discuss sequence homologies between the predicted amino acid sequence of the ReoB protein and ATPases, DNA binding proteins, and enzymes involved in DNA repair.

**Methods**

**Bacterial strains and plasmids**

The source of the reoB gene was either pPE399 (18,19) which carries the gene on a 7 kb XhoI fragment of chromosomal DNA cloned into the vector pAT153 (20), or pIDH201 which carries a 19 kb BamHI fragment of chromosomal DNA containing the entire thyA-argA region of the chromosome cloned into pBR328. JM105 was used as a host for the phage cloning vectors M13 mp18 and mp19 (21), and their recombinants.

**DNA sequence analysis**

DNA sequence analysis was performed by the dideoxy chain termination method (22) using single-stranded DNA from clones of M13 mp18 and mp19, a synthetic 17 base universal primer and [α<sup>35</sup>S]dATP (Amersham) as radiolabel. The nucleotide sequence was determined by electrophoresis through 0.4 mm polyacrylamide buffer gradient gels (23) followed by exposure to Fuji RX X-ray film.

Initially, the sequence was built up by determining the sequences of pPE399 restriction fragments cloned into M13 mp8 or M13 mp9 RF DNA. Further clones were generated by using the enzyme Bal-31 to delete increasingly large DNA fragments from the region to be sequenced, in order to bring more distant sequences within range of the universal primer (25). Shotgun clones of the 3.0 and 3.6 kb PstI fragments of pIDH201 were also generated by randomly shearing the DNA by sonication. Fragment ends were repaired using T4 DNA polymerase and dNTPs, and then cloned into Smal cleaved, alkaline phosphatase.
treated M13 mp18 RF DNA as previously described (17,24). The complete sequence was determined on both strands.

Computer programs developed by Queen and Korn (26) and Staden (27) were used to assemble and analyse the sequence. Molecular weights calculated by these programs differed slightly. Those reported in this paper were calculated according to (26).

DNA Binding

The RecB and RecC proteins were purified as described previously (19). Binding of these proteins to heat-denatured [3H]-\(\lambda\) DNA was measured using nitrocellulose filters, essentially as described (49).

RESULTS

Nucleotide Sequence

The sequence of a 3,960 bp region of the \(E. \text{coli}\) chromosome that carries the entire \(\text{recB}\) gene is shown in Fig. 1. The sequence is numbered from the unique PstI site in the \(\text{thrA}\) gene (17) and is continuous with the numbering we have used for the \(\text{recC}\) (17) and \(\text{pil}\) genes (28). The putative 3543 bp \(\text{recB}\) coding sequence begins at the ATG initiation codon at bp 8967 and continues until the TAA termination codon at bp 12509. This would direct the synthesis of a polypeptide of 1180 amino acids with a calculated molecular weight of 133,974. The ATG initiation codon is preceded 8 bp upstream by the sequence GAG, which is homologous to part of the consensus ribosome binding sequence (29).

Assignment of the start of the \(\text{recB}\) coding sequence to the ATG at bp 8967 and not that at bp 9327 was by two criteria. Firstly, initiation at bp 9327 would give a RecB protein with a molecular weight of 120,688 which is less than that observed by SDS-PAGE (14-16) and also less than that of the RecC protein determined from its nucleotide sequence (17). However, on SDS-PAGE, the RecB protein has always been found to have a higher molecular weight than the RecC protein (14-16). Secondly, the RecB protein is known to be a DNA-dependent ATPase (19), and the only sequence homologous to the consensus for both ATP binding proteins and DNA binding proteins in the predicted RecB amino acid sequence is found in the region encoded between these two ATG start codons (see below).

In the 326 nucleotides preceding the \(\text{recB}\) gene there are no sequences that reasonably fit the consensus \(E. \text{coli}\) promoter -10 (TATAAT) and -35 (TTGaca) sequences (30).

In the sequence presented in Fig. 1, in addition to the \(\text{recB}\) coding
Fig. 1
Nucleotide sequence of the recB gene. The numbering of the nucleotides is from the PstI site within the thrA gene (17) and is continuous with that used for the recF (17) and ptr genes (28). The recB gene and its deduced amino acid sequence is proposed to begin at bp 8,967. The coding sequence for the C-terminus of protease III extends from bp 8,641 to bp 8,974, and the coding sequence for the N-terminus of the RecD protein extends from bp 12,509 to bp 12,600. The region of the RecB amino acid sequence that is homologous to the consensus found in other ATPases (residues 23 to 37) is boxed.

sequence, there are two other open reading frames. The first extends from bp 8,641 to a termination codon, TGA, at bp 8,974 and therefore overlaps the proposed recB translational start by 8 nucleotides (including the termination codon). This reading frame is the coding sequence for the C-terminal portion of Protease III (28). The second open reading frame, which extends from the ATG initiation codon at bp 12,509 and continues until bp 12,600, overlaps the recB termination codon by 1 nucleotide. This is the proposed start of the recB gene encoding the α subunit of Exonuclease V discussed in the accompanying paper (31).

Codon Usage and Amino Acid Composition

The RecB protein is present in low copy number in the cell (8,19), an apparently common feature of DNA repair enzymes in E. coli (32-34). In efficiently expressed genes rare codons normally occur at a level of 4% in the coding frame versus 11% and 10% in the non-coding frames, whilst in genes which code for low copy number proteins the rare codons are found in equal frequency in all three reading frames (35). The rare codons, which are ATA
Table 1
Codon Usage in the repB Gene

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino Acid</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT Phe</td>
<td>30</td>
<td>TCT Ser</td>
</tr>
<tr>
<td>TTC Phe</td>
<td>14</td>
<td>TCG Ser</td>
</tr>
<tr>
<td>TTA Leu</td>
<td>14</td>
<td>TCA Ser</td>
</tr>
<tr>
<td>TTG Leu</td>
<td>32</td>
<td>TGG Ser</td>
</tr>
<tr>
<td>CTT Leu</td>
<td>13</td>
<td>CCT Pro</td>
</tr>
<tr>
<td>CTC Leu</td>
<td>12</td>
<td>CCC Pro</td>
</tr>
<tr>
<td>CTA Leu</td>
<td>7</td>
<td>CCA Pro</td>
</tr>
<tr>
<td>CTG Leu</td>
<td>66</td>
<td>CCG Pro</td>
</tr>
<tr>
<td>ATT Ile</td>
<td>20</td>
<td>ACT Thr</td>
</tr>
<tr>
<td>ATC Ile</td>
<td>31</td>
<td>ACC Thr</td>
</tr>
<tr>
<td>ATA Ile</td>
<td>3</td>
<td>ACA Thr</td>
</tr>
<tr>
<td>ATG Met</td>
<td>31</td>
<td>ACG Thr</td>
</tr>
<tr>
<td>GCT Val</td>
<td>17</td>
<td>GCA Ala</td>
</tr>
<tr>
<td>GTC Val</td>
<td>15</td>
<td>GCC Ala</td>
</tr>
<tr>
<td>GTA Val</td>
<td>9</td>
<td>GGA Glu</td>
</tr>
<tr>
<td>GTG Val</td>
<td>21</td>
<td>GCG Ala</td>
</tr>
</tbody>
</table>

(Ile), TCG (Ser), CAA (Gln), AAT (Asn), CCT and CCC (Pro), ACG (Thr) and AGG (Arg), occur at a frequency of 7.2% in the repB coding frame, and at 13.1% and 9.1% in the non-coding frames (Table 1). The pattern of codon usage within repB appears therefore to be indicative of an intermediate level of translation.

The level of expression of a gene can also be correlated with the choice between U and C in codon position 3. A preference exists in well expressed E. coli genes for nucleotides in the 'wobble' position that yield a codon-anticodon binding interaction of intermediate strength. This interaction is optimised when a C follows AD, DA, DD and AA doublets and when a C follows GC, CO, CC and GG doublets (36,37). However, in weakly expressed genes this bias is not present. In the repB coding sequence, AD, DA, DD and AA doublets are followed by a T in 53% of cases and C in 47%. Similarly, GC, CO, CC and GG doublets are followed by a T in 40% of the cases and by a C in 60%. This indicates that the efficiency of translation may be decreased in repB.

From the predicted amino acid sequence, the RecB protein consists of 123 (10.4%) basic residues and 170 (14.4%) acidic residues representing a net charge of -47, consistent with its acidic isoelectric point of approximately 5.6 (38, our unpublished results).

Identification of a Putative ATP Binding Site in the RecB Protein

Walker et al. (39) identified a conserved sequence that is present in a number of adenine nucleotide binding proteins, such as ATPases. Similar
Table 2
Alignment of putative ATP binding sequences in the RecB protein and other E. coli DNA repair enzymes. Identical or similar residues are boxed.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Residues</th>
<th>Sequence</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uvra</td>
<td>24-45</td>
<td>DKLIVVTGGLSGGKSSLAFDTL</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>633-654</td>
<td>LAHTLLGVTSGKTFTIANVI</td>
<td>41</td>
</tr>
<tr>
<td>Uvrb</td>
<td>32-53</td>
<td>RSNLLVLAGAGSGKTTRVLVHR</td>
<td>47, 48</td>
</tr>
<tr>
<td>Uvrd</td>
<td>22-43</td>
<td>GRIVEIYGPESSGKTTLTLQVI</td>
<td>40</td>
</tr>
<tr>
<td>Reca</td>
<td>59-80</td>
<td>QTGKTPTI AALY</td>
<td>39</td>
</tr>
<tr>
<td>Recb</td>
<td>16-37</td>
<td>QGERLIEASAGTGKTPTI AALY</td>
<td>40</td>
</tr>
</tbody>
</table>

sequences have been found in a number of E. coli ATPases involved in DNA repair including the RecA (39), the UvrD (40), and the Uvra proteins (41). The RecB protein has DNA-dependent ATPase activity (19) and might be expected therefore to have an ATP recognition site. The sequence of the RecB protein from residues 23 to 37 shows homology to the consensus sequence (Table 2). The homology is particularly strong between the RecB and Uvrb sequences.

Identification of a Possible DNA Binding Site in the RecB Protein

In complexes of DNA with the Cro and CI repressors of bacteriophage lambda, and with the CAP protein of E. coli, many of the DNA contacts are made by two α-helices that are linked by a tight turn (see [42] for a review). This structure is also found in a number of other DNA binding proteins, suggesting that they too use helix-turn-helix structures for DNA interactions (43-45). In filter-binding assays, we find that the RecB protein, but not the RecC protein, binds to single-stranded DNA (Table 3). Using Chou and Fasman rules (46) it is possible to predict a helix-turn-helix structure from residues 63 to 86 of the RecB amino acid sequence. This region contains the same pattern of conserved residues and residue types that have been suggested by Pabo and Sauer (42) to be involved in the interaction with DNA (Table 4).

Table 3
Binding to single-stranded DNA

<table>
<thead>
<tr>
<th>Protein</th>
<th>% DNA retained</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>17</td>
</tr>
<tr>
<td>RecB</td>
<td>64</td>
</tr>
<tr>
<td>RecC</td>
<td>23</td>
</tr>
</tbody>
</table>

Reaction mixtures containing 0.1 µg of either RecB or RecC protein and 100 µM ATPγS. Following incubation at 37°C for 10 minutes, samples were applied to nitrocellulose filters, washed and the bound radioactivity determined.
Table 4
Putative DNA binding site in the RecB protein.

<table>
<thead>
<tr>
<th>Helical assessment</th>
<th>HELIX I</th>
<th>TURN</th>
<th>HELIX II</th>
</tr>
</thead>
<tbody>
<tr>
<td>RecB protein</td>
<td>i h i H H H i H i B i i b h h H H i h H</td>
<td>T F T E A A T A E L R G I R S N I H E L R I A</td>
<td></td>
</tr>
<tr>
<td>consensus</td>
<td>---------</td>
<td>------</td>
<td>---------</td>
</tr>
</tbody>
</table>

H - strong helix former, h - helix former, I - weak helix former, i - indifferent helix former, B - strong helix breaker, b - weak helix breaker.

Sequence homology between the RecB protein and DNA Repair enzymes

There are two regions of the RecB protein sequence (residues 516-533 and 557-574) which are homologous with the regions of the UvrB (residues 650-667) and UvrC proteins (residues 199-216) designated Domain-2 (47,48). Also, the previously published RecC protein sequence (17) between residues 703 and 708 has some homology to Domain-1 of the UvrB and UvrC proteins (47,48). There are also regions of homology between the predicted sequence of the RecB protein and the UvrD protein (40), in addition to that at the ATP binding sequence, but further work will be required to assess their significance, if any.

DISCUSSION

We have determined the complete nucleotide sequence of the recB gene and shown that it would encode a polypeptide 1180 amino acids long of molecular weight of 133,973, in agreement with the values of 135 - 140 kDa estimated from SDS PAGE (14-16).

Several features of the recB gene sequence may contribute to low intracellular level of the RecB protein (8,19). Immediately preceding the coding sequence, only the triplet GAG is homologous to the ribosome binding site consensus sequence, AGGAGGT (29). Rare codons occur within recB at higher level than in most efficiently expressed genes although not at the frequency found in other genes coding for low copy number proteins. Thus, a combination of a relatively inefficient ribosome binding site, an intermediate level of occurrence of rare codons and no apparent bias towards the use of codons that give intermediate levels of codon-anticodon interactions within the recB coding sequence, might limit the rate of translation.

The S1 mapping experiments of Sasaki et al. (15) indicate that transcription of recB is initiated 1.5 kb upstream of the HindIII site (bp...
10341), approximately 130 nucleotides preceding the initiation codon. However, there is no readily identifiable promoter sequence in this region. Furthermore, in preliminary S1 mapping experiments we find that a 475 bp PstI-BstEII fragment (bp 8672 to bp 9147) is protected by total cellular RNA against nuclease digestion (results not shown).

The distal end of the pII gene overlaps the proposed start of regB by 8 nucleotides. Furthermore, the regB termination codon overlaps the initiation codon of the downstream regD gene. Thus, the three genes may constitute an operon. Further work will be required to elucidate the mechanisms of transcription of these genes.

The deduced RecB amino acid sequence contains a consensus ATP binding site (39), and a predicted helix-turn-helix structure implicated in DNA binding (42), in agreement with the experimental observations that the RecB protein has DNA-dependent ATPase activity (19) and binds tightly to single stranded DNA.

ACKNOWLEDGMENTS

We thank Rodger Staden for the gift of his computer programs. This work was supported by the Medical Research Council.

*Present address: MRC Brain Metabolism Unit, Royal Edinburgh Hospital, Edinburgh EH10 5HF, UK

REFERENCES
21-27.