Volume 14 Number 21 1986

Complete nucleotide sequence of the *Escherichia coli recB* gene

Paul W.Finch, Alan Storey, Karen E.Chapman*, Kate Brown, Ian D.Hickson[1] and Peter T.Emmerson

Department of Biochemistry, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU and [1]Department 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 3543 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 ptr 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'-GCTGGTGG-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 nuclease 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 ptr, 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 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 recB and recC 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-recC intergenic region (17), the recC gene (17) and the complete ptr gene (28). Here, we report the complete nucleotide sequence of the recB gene and discuss sequence homologies between the predicted amino acid sequence of the RecB protein and ATPases, DNA binding proteins, and enzymes involved in DNA repair.

METHODS

Bacterial strains and plasmids

The source of the recB 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^{35}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 SmaI 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 [^{3}H]-λ DNA was measured using nitrocellulose filters, essentially as described (49).

RESULTS

Nucleotide Sequence

The sequence of a 3,960 bp region of the *E. coli* chromosome that carries the entire *recB* gene is shown in Fig. 1. The sequence is numbered from the unique PstI site in the *thyA* gene (17) and is continuous with the numbering we have used for the *recC* (17) and *ptr* genes (28). The putative 3543 bp *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 *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 *recB* gene there are no sequences that reasonably fit the consensus *E. coli* promoter -10 (TATAAT) and -35 (TTGaca) sequences (30).

In the sequence presented in Fig. 1, in addition to the *recB* coding

```
  Q   I   Q   Q   A   V   I   T   Q   M   L   Q   A   P   Q   T   L   G   E   E   A   S   K   L   S   K   D   F   D   R   G   N   M   R   F   D   S   R   D   K
GCAAATCCAGCAGGCGGTAATTACCCAGATGCTGCAGGCACCGCAAACGCTCGGCGAAGAAGCATCGAAGTTAAGTAAAGATTTCGATCGCGGCAATATGCGCTTCGATTCGCGTGATAA
      8650      8660      8670      8680      8690      8700      8710      8720      8730      8740      8750      8760

  I   V   A   Q   I   K   L   L   T   P   Q   K   L   A   D   F   F   H   Q   A   V   V   E   P   Q   G   M   A   I   L   S   Q   I   S   G   S   Q   N   G   K
AATCGTGGCCCAGATAAAACTGCTGACGCCGCAAAAACTTGCTGATTTCTTCCATCAGGCGGTGGTCGAGCCGCAAGGCATGGCTATTCTGTCGCAGATTTCCGGCAGCCAGAACGGAA
      8770      8780      8790      8800      8810      8820      8830      8840      8850      8860      8870      8880

  A   E   Y   V   H   P   E   G   W   K   V   W   E   N   V   S   A   L   Q   Q   T   M   P   L   M   S   E   K   N   E   *
                                                                                               M   S   D   V   A   E   T   L   D   P   L   R
AGCCGAATATGTACACCCTGAAGGCTGGAAAGTGTGGGAGAACGTCAGCGCGTTGCAGCAAACAATGCCCCTGATGAGTGAAAAGAATGAGTGATGTCGCCGAGACACTAGATCCTTTGC
      8890      8900      8910      8920      8930      8940      8950      8960      8970      8980      8990      9000

  L   P   L  [Q   G   E   R   L   I   E   A   S   A   G   T   G   K   T   F   T   I   A   A   L   Y]  L   R   L   L   L   G   L   G   G   S   A   A   F   P   R
GCTTGCCCTTACAGGGTGAGCGCCTGATTGAAGCCTCTGCCGGCACAGGCAAAACCTTTACGATTGCGGCGCTCTATTTGCGCCTGTTACTTGGACTAGGCGGTTCCGCCGCCTTTCCCC
      9010      9020      9030      9040      9050      9060      9070      9080      9090      9100      9110      9120

  P   L   T   V   E   E   L   L   V   V   T   F   T   E   A   A   T   A   E   L   R   G   R   I   R   S   N   I   H   E   L   R   I   A   C   L   R   E   T   T
GCCCGCTGACCGTTGAAGAACTGCTGGTGGTCACCTTTACCGAGGCTGCCACGGCAGAATTGCGCGGTCGTATCCGTAGCAATATCCACGAGTTGCGCATCGCCTGTCTGCGTGAAACCA
      9130      9140      9150      9160      9170      9180      9190      9200      9210      9220      9230      9240

  D   N   P   L   Y   E   R   L   L   E   E   I   D   D   K   A   Q   A   A   Q   W   L   L   L   A   E   R   Q   M   D   E   A   A   V   F   T   I   H   G   F
CCGACAATCCACTGTACGAACGCCTGCTGGAAGAGATCGACGATAAAGCGCAAGCCGCGCAGTGGTTGTTGTTAGCCGAACGGCAGATGGATGAAGCGGCAGTCTTTACTATTCACGGCT
      9250      9260      9270      9280      9290      9300      9310      9320      9330      9340      9350      9360

  C   Q   R   M   L   N   L   N   A   F   E   S   G   M   L   F   E   Q   Q   L   I   E   D   E   S   L   L   R   Y   Q   A   C   A   D   F   W   R   R   H   C
TTTGCCAGCGCATGCTCAACCTGAATGCCTTTGAATCCGGCATGCTGTTTGAGCAGCAGCTGATTGAAGATGAGTCTCTGCTACGCTACCAGGCCTGCGCCGATTTCTGGCGTCGCCACT
      9370      9380      9390      9400      9410      9420      9430      9440      9450      9460      9470      9480

  Y   P   L   P   R   E   I   A   Q   V   V   F   E   T   W   K   G   P   Q   A   L   L   R   D   I   N   R   Y   L   Q   G   E   A   P   V   I   K   A   P   P
GCTACCCGCTGCCGCGTGAAATAGCCCAGGTCGTCTTTGAAACCTGGAAAGGGCCGCAGGCGTTGCTGCGCGATATTAATCGTTATCTGCAAGGCGAAGCGCCGGTTATCAAAGCACCGC
      9490      9500      9510      9520      9530      9540      9550      9560      9570      9580      9590      9600

  P   D   D   E   T   L   A   S   R   H   A   Q   I   V   A   R   I   D   T   V   K   Q   Q   W   R   D   A   V   G   E   L   D   A   L   I   E   S   S   G   I
CGCCCGATGATGAAACGCTGGCTTCCCGTCACGCGCAAATTGTGGCGCGTATTGATACGGTAAAACAGCAGTGGCGCGACGCAGTGGGTGAACTGGATGCGCTGATCGAATCTTCTGGTA
      9610      9620      9630      9640      9650      9660      9670      9680      9690      9700      9710      9720

  D   R   R   K   F   N   R   S   N   Q   A   K   W   I.  D   K   I   S   A   W   A   E   E   E   T   N   S   Y   Q   L   P   E   S   L   E   K   F   S   Q   R
TTGATCGACGCAAGTTTAACCGTAGCAATCAGGCTAAATGGATCGACAAGATCAGCGCCTGGGCAGAAGAAGAGACAAACAGTTATCAGTTGCCGGAGTCGCTGGAAAAATTCTCCCAGC
      9730      9740      9750      9760      9770      9780      9790      9800      9810      9820      9830      9840

  F   L   E   D   R   T   K   A   G   G   E   T   P   R   H   P   L   F   E   A   I   D   Q   L   L   A   E   P   L   S   I   R   D   L   V   I   T   R   A   L
GTTTCTTAGAAGATCGCACGAAGGCCGGGGGGAAACCCCGCGACATCCACTGTTTGAGGCGATCGATCAACTGCTTGCAGAACCATTGTCGATCCGCGATCTGGTGATCACCCGCGCAT
      9850      9860      9870      9880      9890      9900      9910      9920      9930      9940      9950      9960

  A   E   I   R   E   T   V   A   R   E   K   R   R   R   G   E   L   G   F   D   D   M   L   S   R   L   D   S   A   L   R   S   E   S   G   E   V   L   A   A
TGGCTGAGATCCGCGAAACAGTAGCGCGTGAAAAACGCCGCCGTGGCGAATTGGGTTTTGATGACATGTTAAGTCGGCTCGATTCCGCGCTGCGTAGCGAAAGCGGTGAGGTGTTGGCAG
      9970      9980      9990     10000     10010     10020     10030     10040     10050     10060     10070     10080

  A   I   R   T   R   F   P   V   A   M   I   D   E   F   Q   D   T   D   P   Q   Q   Y   R   I   F   R   R   I   W   H   H   Q   P   E   T   A   L   L   L   I
CGGCGATCCGTACGCGATTCCCGGTGGCAATGATCGATGAATTTCAGGATACCGACCCCCAGCAGTACCGAATTTTTCGCCGTATCTGGCACCATCAGCCGGAAACCGCATTGTTGCTAA
     10090     10100     10110     10120     10130     10140     10150     10160     10170     10180     10190     10200

  G   D   P   K   Q   A   I   Y   A   F   R   G   A   D   I   F   T   Y   M   K   A   R   S   E   V   H   A   H   Y   T   L   D   T   N   W   R   S   A   P   G
TTGGCGACCCGAAGCAGGCCATATATGCATTCCGGGGTGCGGATATCTTCACTTATATGAAGGCGCGTAGCGAAGTTCACGCCCACTACACTTTAGACACCAACTGGCGTTCCGCACCAG
     10210     10220     10230     10240     10250     10260     10270     10280     10290     10300     10310     10320

  M   V   N   S   V   N   K   L   F   S   Q   T   D   D   A   F   M   F   R   E   I   P   F   I   P   V   K   S   A   G   K   N   Q   A   L   R   F   V   F   K
GAATGGTGAACAGCGTGAATAAGCTTTTCAGCCAGACTGATGACGCGTTCATGTTTCGCGAAATACCGTTTATTCCAGTGAAATCAGCCGGGAAAAATCAGGCGTTACGTTTTGTATTTA
     10330     10340     10350     10360     10370     10380     10390     10400     10410     10420     10430     10440

  G   E   T   Q   P   A   M   K   M   W   L   M   E   G   E   S   C   G   V   G   D   Y   Q   S   T   M   A   Q   V   C   A   A   Q   I   R   D   W   L   Q   A
AAGGTGAAACACAGCCTGCGATGAAAATGTGGCTGATGGAAGGCGAAAGCTGCGGCGTTGGCGATTATCAAAGTACCATGGCGCAGGTATGTGCTGCGCAAATCCGCGACTGGCTACAAG
     10450     10460     10470     10480     10490     10500     10510     10520     10530     10540     10550     10560

  G   Q   R   G   E   A   L   L   M   N   G   D   D   A   R   P   V   R   A   S   D   I   S   V   L   V   R   S   R   Q   E   A   A   Q   V   R   D   A   L   T
CCGGACAGCGGGGCGAAGCGTTGCTGATGAACGGCGACGACGCGCGTCCGGTGCGTGCTTCGGACATCAGTGTGCTGGTGCGCAGCCGCCAGGAGGCCGCCCAGGTGCGCGATGCCTTAA
     10570     10580     10590     10600     10610     10620     10630     10640     10650     10660     10670     10680

  L   L   E   I   P   S   V   Y   L   S   N   R   D   S   V   F   E   T   L   E   A   Q   E   M   L   W   L   L   Q   A   V   M   T   P   E   R   E   N   T   L
CGTTGCTGGAAATCCCTTCCGTTTACCTTTCGAACCGCGACAGTGTTTTTGAAACTCTGGAAGCGCAGGAAATGCTTTGGTTGTTGCAGGCGGTGATGACGCCCGAACGTGAGAACACCC
     10690     10700     10710     10720     10730     10740     10750     10760     10770     10780     10790     10800

  R   S   A   L   A   T   S   M   M   G   L   N   A   L   D   I   E   T   L   N   N   D   E   H   A   W   D   V   V   V   E   E   F   D   G   Y   R   Q   I   W
TGCGTAGTGCGCTGGCAACGTCAATGATGGGGCTGAACGCGCTGGATATCGAAACGCTGAACAATGACGAACATGCGTGGGATGTGGTAGTCGAAGAGTTCGATGGTTATCGGCAAATCT
     10810     10820     10830     10840     10850     10860     10870     10880     10890     10900     10910     10920

  R   K   R   G   V   M   P   M   L   R   A   L   M   S   A   R   N   I   A   E   N   L   L   A   T   A   G   G   E   R   R   L   T   D   I   L   H   I   S   E
GGCGCAAACGTGGCGTTATGCCGATGCTGCGGGCGCTGATGTCGGCGCGTAACATTGCTGAAAACTTGCTGGCAACGGCAGGCGGTGAGCGGCGTCTTACCGATATCTTGCATATCAGCG
     10930     10940     10950     10960     10970     10980     10990     11000     11010     11020     11030     11040

  L   L   Q   E   A   G   T   Q   L   E   S   E   H   A   L   V   R   W   L   S   Q   H   I   L   E   P   D   S   N   A   S   S   Q   Q   M   R   L   E   S   D
AACTGCTACAAGAAGCCGGCAACGCAGCTGGAAAGTGAACATGCGCTGGTACGCTGGTTATCGCAACATATCCTCGAGCCAGACAGTAATGCCTCCAGCCAACAAATGCGTCTCGAAAGTG
     11050     11060     11070     11080     11090     11100     11110     11120     11130     11140     11150     11160

  K   H   L   V   Q   I   V   T   I   H   K   S   K   G   L   E   Y   P   L   V   W   L   P   F   I   T   N   F   R   V   Q   E   Q   A   F   Y   H   D   R   H
ATAAACATCTGGTGCAGATTGTCACGATCCACAAATCGAAAGGGCTGGAATATCCATTGGTCTGGCTGCCGTTTATCACCAATTTCCGCGTCCAGGAGCAGGCGTTTTATCACGATCGCC
     11170     11180     11190     11200     11210     11220     11230     11240     11250     11260     11270     11280

  S   F   E   A   V   L   D   L   N   A   A   P   E   S   V   D   L   A   E   A   E   R   L   A   E   D   L   R   L   L   Y   V   A   L   T   R   S   V   W   H
ACTCGTTTGAGGCAGTTCTGGATCTTAATGCTGCGCCAGAAAGCGTCGACCTCGCGGAGGCCGAACGTCTGGCGGAAGATCTGCGTTTGCTTTACGTGGCGCTGACACGTTCGGTTTGGC
     11290     11300     11310     11320     11330     11340     11350     11360     11370     11380     11390     11400

  C   S   L   G   V   A   P   L   V   R   R   R   G   D   K   K   G   D   T   D   V   H   Q   S   A   L   G   R   L   L   Q   K   G   E   P   Q   D   A   A   G
ATTGCAGTCTCGGCGTTGCACCGCTGGTGCGCCGTCGTGGCGATAAAAAAGGTGACACCGACGTCCACCAAAGTGCGCTCGGGCGTTTGCTGCAAAAAGGGGAACCGCAAGATGCGGCAG
     11410     11420     11430     11440     11450     11460     11470     11480     11490     11500     11510     11520
```

```
  L  R  T  C  I  E  A  L  C  D  D  D  I  A  W  Q  T  A  Q  T  G  D  N  Q  P  W  Q  V  N  D  V  S  T  A  E  L  N  A  K  T
GGCTTCGCACCTGTATTGAAGCGTTATGCGATGATGATATTGCCTGGCAAACGGCACAAACTGGTGATAACCAACCCTGGCAGGTTAATGATGTTTCTACAGCAGAGCTGAATGCGAAGA
    11530     11540     11550     11560     11570     11580     11590     11600     11610     11620     11630     11640

  L  Q  R  L  P  G  D  N  W  R  V  T  S  Y  S  G  L  Q  Q  R  G  H  G  I  A  Q  D  L  M  P  R  L  D  V  D  A  A  G  V  A
CGTTACAACGATTGCCCGGCGATAACTGGCGCGTCACCAGCTACTCTGGTTTGCAACAGCGTGGTCACGGTATCGCCCAGGATTTGATGCCTCGGCTGGATGTCGATGCTGCAGGCGTTG
    11650     11660     11670     11680     11690     11700     11710     11720     11730     11740     11750     11760

  S  V  V  E  E  P  T  L  T  P  H  Q  F  P  R  G  A  S  P  G  T  F  L  H  S  L  F  E  D  L  D  F  T  Q  P  V  D  P  N  W
CCAGCGTCGTTGAAGAACCGACGTTAACACCACATCAGTTTCCGCGCGGTCGTCACCGGGGACGTTCTTGCACAGTTTGTTGAAGACCTGGATTTTACCCAGCCGGTTGACCCGAACT
    11770     11780     11790     11800     11810     11820     11830     11840     11850     11860     11870     11880

  V  R  E  K  L  E  L  G  G  F  E  S  Q  W  E  P  V  L  T  E  W  I  T  A  V  L  Q  A  P  L  N  E  T  G  V  S  L  S  Q  L
GGGTGCGGGAAAAACTGGAACTCGGCGGCTTTGAATCGCAGTGGGAACCGGTATTGACCGAGTGGATCACGGCTGTCCTCCAGGCACCTCTCAATGAAACCGGCGTAAGCCTGAGTCAAC
    11890     11900     11910     11920     11930     11940     11950     11960     11970     11980     11990     12000

  S  A  R  N  K  Q  V  E  M  E  F  Y  L  P  I  S  E  P  L  I  A  S  Q  L  D  T  L  I  R  Q  F  D  P  L  S  A  G  C  P  P
TTTCCGCCCGCAATAAACAGGTGGAGATGGAGTTTTATCTGCCGATTAGTGAACCGCTTATCGCCAGTCAGCTTGATACGTTAATCCGCCAGTTTGACCCGCTATCCGCAGGCTGCCCGC
    12010     12020     12030     12040     12050     12060     12070     12080     12090     12100     12110     12120

  L  E  F  M  Q  V  R  G  M  L  K  G  F  I  D  L  V  F  R  H  E  G  R  Y  Y  L  L  D  Y  K  S  N  W  L  G  E  D  S  S  A
CGCTGGAGTTCATGCAGGTACGTGGCATGTTAAAAGGCTTTATCGACCTGGTGTTCCGCCACGAAGGGCGTTATTACCTGCTCGACTATAAATCCAACTGGTTGGGTGAAGACAGTTCGG
    12130     12140     12150     12160     12170     12180     12190     12200     12210     12220     12230     12240

  Y  T  Q  Q  A  M  A  A  A  M  Q  A  H  R  Y  D  L  Q  Y  Q  L  Y  T  L  A  L  H  R  Y  L  R  H  R  I  A  D  Y  D  Y  E
CTTACACCCAACAGGCTATGGCAGCGGCAATGCAGGCACACCGCTATGATCTGCAATATCAGCTTTATACCCTGGCGCTGCATCGTTATCTGCGCCATCGCATTGCTGATTACGACTATG
    12250     12260     12270     12280     12290     12300     12310     12320     12330     12340     12350     12360

  H  H  F  G  G  V  I  Y  L  P  L  R  G  V  D  K  E  H  P  Q  Q  G  I  Y  T  T  R  P  N  A  G  L  I  A  L  M  D  E  M  F
AGCACCACTTTGGCGGCGTTATTTATCTGTTCCTGCGTGGCGTTGATAAAGAACATCCGCAACAGGGGATTTACACAACCCGACCCAACGCCGGGTTGATTGCCCTGATGGATGAGATGT
    12370     12380     12390     12400     12410     12420     12430     12440     12450     12460     12470     12480

                                                                  M  K  L  Q  K  Q  L  L  E  A  V  E  H  K  Q  L  R  P  L  D  V  Q  F  A  L  T  V  A  G  D  E
  A  G  M  T  L  E  E  A  *
TTGCCGGTATGACCCTGGAGGAGGCGTAATGAAATTGCAAAAGCAATTACTGGAAGCTGTGGAGCACAAACAGCTACGCCCGCTGGATGTGCAATTTGCCCTGACCGTGGCGGGAGATGA
    12490     12500     12510     12520     12530     12540     12550     12560     12570     12580     12590     12600
```

Figure 1
Nucleotide sequence of the recB gene. The numbering of the nucleotides is from the PstI site within the thyA gene (17) and is continous with that used for the recC (17) and ptr genes (28). The recB gene and its deduced amino acid sequence is proposed to begin at bp 8967. 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 8641 to a termination codon, TGA, at bp 8974 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 recD 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 recB Gene

TTT	Phe	30	TCT	Ser	6	TAT	Tyr	18	TGT	Cys	3
TTC	Phe	14	TCC	Ser	11	TAC	Tyr	12	TGC	Cys	7
TTA	Leu	14	TCA	Ser	3	TAA	End	1	TGA	End	0
TTG	Leu	32	TCG	Ser	11	TAG	End	0	TGG	Trp	23
CTT	Leu	13	CCT	Pro	5	CAT	His	12	CGT	Arg	37
CTC	Leu	12	CCC	Pro	8	CAC	His	17	CGC	Arg	40
CTA	Leu	7	CCA	Pro	9	CAA	Gln	24	CGA	Arg	6
CTG	Leu	66	CCG	Pro	30	CAG	Gln	47	CGG	Arg	9
ATT	Ile	20	ACT	Thr	6	AAT	Asn	15	AGT	Ser	17
ATC	Ile	31	ACC	Thr	26	AAC	Asn	17	AGC	Ser	16
ATA	Ile	3	ACA	Thr	9	AAA	Lys	24	AGA	Arg	0
ATG	Met	31	ACG	Thr	19	AAG	Lys	7	AGG	Arg	0
GTT	Val	17	GCT	Ala	13	GAT	Asp	48	GGT	Gly	19
GTC	Val	15	GCC	Ala	33	GAC	Asp	26	GGC	Gly	28
GTA	Val	9	GCA	Ala	26	GAA	Glu	67	GGA	Gly	4
GTG	Val	21	GCG	Ala	45	GAG	Glu	29	GGG	Gly	13

(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 recB coding frame, and at 13.1% and 9.1% in the non-coding frames (Table 1). The pattern of codon usage within recB 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 AU, UA, UU and AA doublets and when a C follows GC, CG, CC and GG doublets (36,37). However, in weakly expressed genes this bias is not present. In the recB coding sequence, AU, UA, UU and AA doublets are followed by a T in 53% of cases and C in 47%. Similarly, GC, CG, 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 recB.

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.

Protein	Residues	Sequence	Reference
UvrA	24- 45	D K L I V V T G L S G S G K S S L A F D T L	41
	633-654	G L F T C I T G V S G S G K S T L I N D T L	41
UvrB	32- 53	L A H Q T L L G V T G S G K T F T I A N V I	47,48
UvrD	22- 43	R S N L L V L A G A G S G K T R V L V H R I	40
RecA	59- 80	G R I V E I Y G P E S S G K T T L T L Q V I	39
RecB	16- 37	Q G E R L I E A S A G T G K T F T I A A L Y	

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

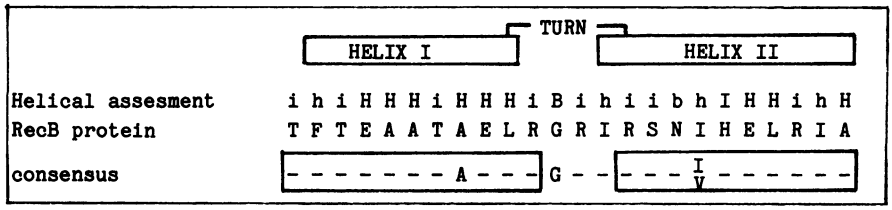
In complexes of DNA with the Cro and cI repressors of bacteriophage lambda, and with the CAP protein of E. coli, many of the DNA contacts are made by two α-helices that are linked by a tight turn (see [42] for a review). This structure is also found in a number of other DNA binding proteins, suggesting 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

Protein	% DNA retained
None	17
RecB	64
RecC	23

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.



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 efficently 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 occurence 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 ptr gene overlaps the proposed start of recB by 8 nucleotides. Furthermore, the recB termination codon overlaps the initiation codon of the downstream recD gene. Thus, the three genes may constitute an operon. Further work will be required to elucidate the mechanisms of transciption 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.

ACKNOWLEDGEMENTS

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

1. Buttin, G. and Wright, M. (1968) Cold Spring Harbour Symp. Quant Biol. 33, 259-264.
2. Oishi, M. (1969) Proc. Natl. Acad. Sci. USA 64, 1292-1299.
3. Goldmark, P.J. and Linn, S.(1970) Proc. Natl. Acad. Sci. USA 67, 434-441.
4. Emmerson, P.T. (1968) Genetics 60, 19-30.
5. Willetts, N.S. and Mount, D.W. (1969) J. Bacteriol. 100, 923-934.
6. Capaldo-Kimball, F. and Barbour, S.D. (1971) J. Bacteriol. 106, 204-212.
7. Rosamond, J., Telander, K.M. and Linn, S. (1979) J. Biol. Chem. 254, 8646-8652.
8. Taylor, A. and Smith, G.R. (1980) Cell 22, 447-457.
9. Ponticelli, A.S., Schultz, D.W.,Taylor, A.F. and Smith, G.R. (1985) Cell 41, 145-151.
10. Taylor, A.F., Schultz, D.W., Ponticelli, A.S. and Smith, G.R. (1985) Cell 41, 153-163.
11. Smith, G.R. (1983) in Lambda II (Hendrix, R.W., Roberts, J.W., Stahl, F.W. and Weisberg, R.A., eds.). pp. 175-209, Cold Spring Harbor, New York.
12. Gillen, J.R. and Clark, A.J. (1974) in Mechanisms In Recombination (Grell, R.F., ed) p123, Plenum Press, New York.
13. Muskavitch, K.M.T. and Linn, S. (1981) in The Enzymes (Boyer, P.D., ed) Vol. 14, pp233-250, Academic Press, New York.
14. Hickson, I.D. and Emmerson, P.T. (1981) Nature 294, 578-580.
15. Sasaki, M., Fujiyoshi, T., Shimada, K. and Takagi, Y. (1982) Biochem. Biophys. Res. Commun. 109, 414-422.
16. Dykstra, C.C., Prasher, D. and Kushner, S.R. (1984) J. Bacteriol. 157,

21-27.
17. Finch, P.W., Wilson, R.E., Brown, K., Hickson, I.D., Tomkinson, A.E. and Emmerson, P.T. (1986) Nucl. Acids Res. 14, 4437-4451.
18. Tomkinson, A.E. (1983) PhD Thesis, University of Newcastle upon Tyne, U.K.
19. Hickson, I.D., Robson, C.R., Atkinson, K.E., Hutton, L. and Emmerson, P.T. (1985) J. Biol. Chem. 260, 1224-1229.
20. Twigg, A.J. and Sherratt, D. (1980) Nature 283, 216-218.
21. Yanisch-Perron, C., Vierra, J. and Messing J. (1985) Gene 33, 103-119.
22. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
23. Biggin, M.D., Gibson, T.J. and Hong, C.F. (1983) Proc. Natl. Acad. Sci. USA 80, 3963-3965.
24. Bankier, A.T. and Barrell, B.G. (1983) Nucleic Acid Biochemistry, (Flavell, R.A., ed) Vol. B5, pp 1-34, Elsevier Scientific Publishers (Ireland) Ltd.
25. Ponaz, M., Solowiejczyk, D., Ballantine, M., Schwartz, E. and Surrey, S. (1982) Proc. Natl. Acad. Sci. USA 79, 4298-4302.
26. Queen, C. and Korn, L.J. (1984) Nucl. Acids Res. 12, 581-599.
27. Staden, R. (1984) Nucl. Acids Res. 12, 551-567.
28. Finch, P.W., Wilson, R.E., Brown,K., Hickson, I.D. and Emmerson, P.T. (1986) Nucl. Acids Res., submitted.
29. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA 21, 1342-1346.
30. Hawley, D.K. and McClure, W.R. (1983) Nucl. Acids Res. 11, 2237-2255.
31. Finch, P.W., Storey, A., Brown, K., Hickson, I.D. and Emmerson, P.T. (1986) Nucl. Acids Res., 14, 8583-8594.
32. Lindahl, T. (1982) Annu. Rev. Biochem. 51, 61-87.
33 Sancar, A., Franklin, K.A. and Sancar, G.B. (1984) Proc. Natl. Acad. Sci. USA 81, 7397-7401.
34. Sancar, G.B., Sancar, A. and Rupp, W.D. (1984) Nucl. Acids Res. 12, 4593-4608
35. Konigsberg, W. and Godson, G.N. (1983) Proc. Natl. Acad. Sci. USA 80, 687-691.
36. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
37. Gouy, M. And Gautier, C. (1982) Nucl. Acids Res. 10, 7055-7074.
38. Umeno, M., Anai, M., Sasaki, M. and Takagi, Y. (1985) J. Biochem. 98, 681-685.
39. Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) EMBO J. 1, 645-951.
40. Finch, P.W. and Emmerson, P.T. (1984) Nucl. Acids Res. 12, 5789-5799.
41. Husain, I., Van Houten, B., Thomas, D.C. and Sancar, A. (1986) J. Biol. Chem. 261, 4895-4901.
42. Pabo, C.O. and Sauer, R.T. (1984) Annu. Rev. Biochem. 53, 293-321.
43. Matthews, B., Ohlendorf, D.H., Anderson, W.F. and Takeda, Y. (1982) Proc. Natl. Acad. Sci. USA 79, 1428-1432.
44. Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M. and Pabo, C.O. (1982) Nature 298, 447-451.
45. Weber, I.T., McKay, D.B. and Steitz, T.A. (1982) Nucl. Acids Res. 10, 5085-5102.
46. Chou, P.Y. and Fassman, G.D. (1978) Adv. Enzymol. Relat. Areas Mol. Biol. 47, 45-148.
47. Backendorf, C., Spaink, H., Barbeiro, A.P. and van de Putte, P. (1986) Nucl. Acids Res. 14, 2877-2890.
48. Arikan, E., Kulkarni, M.S., Thomas, D.C. and Sancar, A. (1986) Nucl. Acids Res. 14, 2637-2650.
49.McKentee, K., Weinstock, G.M. and Lehman, I.R. (1980) Proc. Natl. Acad. Sci. USA 77, 857-861.