Volume 14 Number 19 1986 Nucleic Acids Research

Complete nucleotide sequence of the *Escherichia coli ptr* gene encoding Protease III

Paul W.Finch, Rosemary E.Wilson, 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 14 August 1986; Accepted 12 September 1986

ABSTRACT

The nucleotide sequence of a 3120 bp region of the E. coli chromosome that includes the entire ptr gene has been determined. The proposed coding region for Protease III is 2889 nucleotides long, which would encode a protein consisting of 962 amino acids with a calculated molecular mass of 107,719 daltons. The predicted primary structure of the protein includes a 23-residue signal sequence, cleavage of which would give rise to a mature protein of molecular mass 105,124 daltons. At its 3' end, the ptr gene overlaps the start of the recB coding sequence by 8 bases, suggesting that these genes may form part of an operon.

INTRODUCTION

Protease III is a Mg^{2+}-dependent endopeptidase whose only known biochemical activity is the degradation of small peptides of molecular mass less than 7 kDa, such as insulin and glucagon (1). The enzyme is highly sequence specific, producing cleavages only between tyr-leu and phe-tyr residues of oxidised insulin B chain (1). The purified enzyme consists of a single polypeptide chain of molecular mass 110 kDa (1), which is principally located in the periplasmic space (2). Mutants lacking Protease III activity are not apparently altered phenotypically (3,4) and the precise role of this enzyme in the cell remains to be elucidated.

The structural gene for Protease III, designated ptr, maps at minute 60 on the E. coli linkage map (3) and can be isolated on a 19 kb BamHI-generated DNA fragment (5,6). Characterisation of this fragment (4) has revealed that the ptr gene is located between the recB and recC genes, which code for subunits of Exonuclease V, and that it encodes two polypeptides of molecular masses 110 kDa (Protease III) and 50 kDa (p50), the latter apparently being derived from the N-terminal portion of the ptr coding sequence (4). Peptide mapping of Protease III and p50 by partial S. aureus V-8 proteolysis revealed that the two proteins have sequence homology, suggesting that they are derived from the same reading frame (4).

© IRL Press Limited, Oxford, England.

As part of a project to sequence the complete recB-recC coding region, and as a first stage in an investigation of the possible relationships between Protease III, p50 and Exonuclease V, we have determined the nucleotide sequence of the ptr gene. This shows that there is a single reading frame that must code for both Protease III and p50, and that at its 3' end, the ptr gene overlaps the start of the recB structural gene. Also, there is a potential signal sequence at the N-terminus of the predicted amino acid sequence of the Protease III protein.

MATERIALS AND METHODS

Enzymes and biochemicals

Restriction endonucleases were purchased from NBL Enzymes, New England Biolabs or BCL. DNA polymerase I (Klenow fragment) and T4 DNA polymerase were from Pharmacia. Calf intestinal phosphatase and T4 DNA ligase were from BCL. Deoxynucleoside and dideoxynucleoside triphosphates were from Sigma. Radiochemicals were from Amersham.

DNA sequence analysis

DNA sequence analysis was performed by the dideoxy chain termination method (7) using single-stranded DNA from clones of M13 mp18 and mp19, a synthetic 17 base universal primer, and [α^{35}S] dATP as radiolabel. The nucleotide sequence was determined by electrophoresis through 0.4 mm polyacrylamide buffer gradient gels (8) followed by exposure to Fuji RX X-ray film.

The plasmid pPE37 (9) carries 90% of the ptr gene within a 8.7 kb PstI fragment. The sequence of this fragment was built up by determining the sequences of clones generated by shearing the DNA into random fragments by sonication. These clones were processed as described previously (10).

To determine the last 10% of the sequence, a 3166 bp XhoI fragment which contains the 3' end of ptr (and 60% of the recB coding sequence), was cloned in both orientations into M13 mp19. Four different 17-mer single-stranded DNA primers were synthesised, and used to prime DNA synthesis from sites upstream and adjacent to the sequence to be determined.

The DNA sequence throughout the entire ptr gene was determined on both strands.

Computer programs of Queen and Korn (11), and Staden (12) were used to assemble and analyse the sequence. The molecular weight of the Protease III protein was calculated using the program of Queen and Korn (11).

RESULTS

Nucleotide sequence

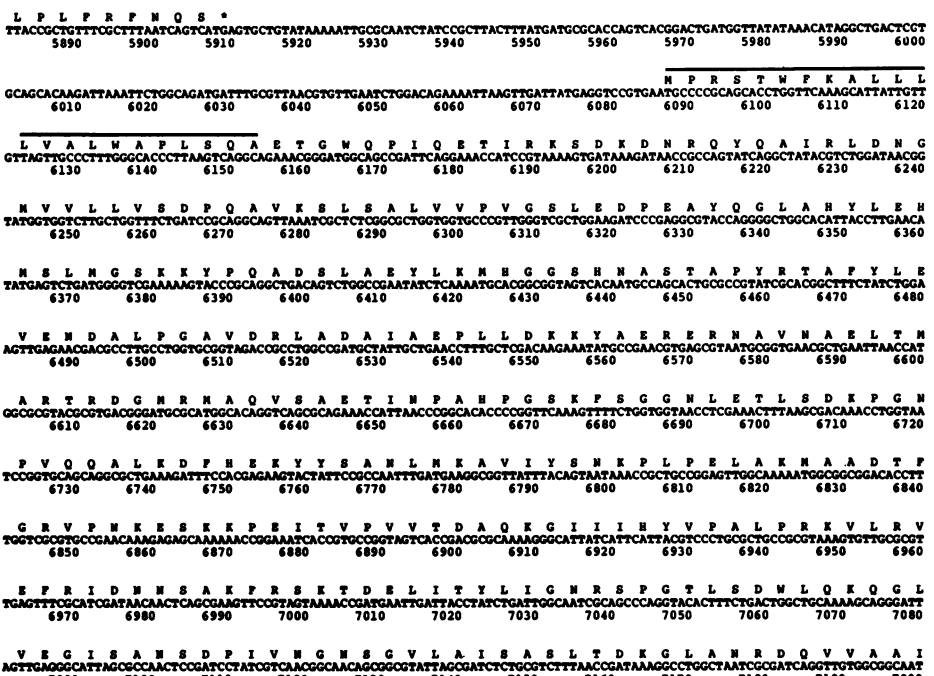
The sequence of a 3120 bp region of the *E. coli* chromosome, including the entire *ptr* gene, is shown in Figure 1. The sequence is continuous with that described previously for the *recC* gene (10) and is also numbered from the unique PstI site in the *thyA* gene. The first 30 nucleotides represent the 3' end of the predicted *recC* coding region (10). In the remaining sequence there is one long open reading frame which begins at the ATG initiation codon at bp 6086 and extends for 2889 nucleotides until the termination codon, TGA, at bp 8974. This is the proposed coding region for Protease III and would direct the synthesis of a protein consisting of 962 amino acids with a calculated molecular mass of 107,719 daltons. A possible ribosome binding site, GAGG (13), precedes the ATG initiation codon by 7 nucleotides. Overlapping the 3' terminus by 8 nucleotides is the start of the *recB* structural gene (14). No long open reading frames are encoded by the opposite DNA strand.

In the 176 nucleotides between the end of the *recC* gene and the proposed start of *ptr*, there is the sequence TTGCGC (bp 5925-5930), followed 17 bp later by the sequence TATGAT (bp 5948-5953), which may act as the -35 and -10 regions, respectively, of the *ptr* promoter. No other sequences could be found in this region which match the canonical promoter sequences, -35 (TTGaca) and -10 (TATAAT) (15), more closely.

Codon usage and amino acid composition

In *E. coli*, 8 rarely-used codons ATA (Ile), TCG (Ser), CAA (Gln), AAT (Asn), CCT and CCC (Pro), ACG (Thr) and AGG (Arg) occur approximately 3 times more frequently in non-coding frames than in the coding frame of efficiently expressed genes (4% vs. 11% and 10%) (16). The codons occur at equal frequency in all 3 reading frames in certain genes which encode products present in only a few copies per cell. In the *ptr* coding sequence, the rare codons occur at a frequency of 8.4% in the coding frame, and 14.1% and 9.2% in the non-coding frames (Table 1). This may be a mechanism for limiting the rate of *ptr* gene 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 efficiently expressed 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 or AA doublets and when



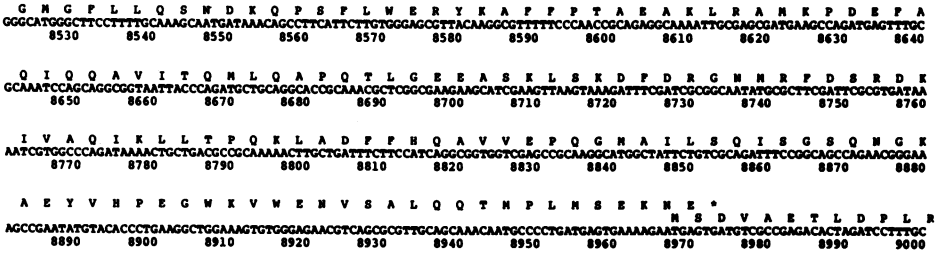


Figure 1
Complete sequence of the ptr gene and its flanking region. The numbering of the nucleotides is from the PstI cleavage site in the thyA gene (10) and is continuous with that described previously for the recC gene (10). The 3' end of the recC gene extends from bp 5881 to 5910. The ptr coding region extends from bp 6086 to bp 8974. The proposed signal peptide of Protease III (residues 1-23) is overlined.

a U follows GC, CG, CC, and GG doublets (17). However, this bias does not exist in genes which code for proteins present in low copy number. In the ptr coding sequence, AU, UA, UU, AA doublets are followed by a C in 47.5% of cases, but also by a U in 52.5% of cases. Similarly, GC, CG, CC and GG doublets are followed by a U in 47% of cases but by C in 53% of cases. This further indicates that the expression of ptr may be regulated at the level of translation.

From the predicted amino acid sequence, Protease III would contain 223

Table 1 Codon usage in the ptr gene.

TTT Phe	17	TCT Ser	7	TAT Tyr	21	TGT Cys	1
TTC Phe	18	TCC Ser	9	TAC Tyr	18	TGC Cys	0
TTA Leu	12	TCA Ser	7	TAA End	0	TGA End	1
TTG Leu	22	TCG Ser	13	TAG End	0	TGG Trp	14
CTT Leu	9	CCT Pro	10	CAT His	6	CGT Arg	21
CTC Leu	11	CCC Pro	9	CAC His	6	CGC Arg	17
CTA Leu	0	CCA Pro	8	CAA Gln	14	CGA Arg	4
CTG Leu	37	CCG Pro	25	CAG Gln	43	CGG Arg	2
ATT Ile	24	ACT Thr	4	AAT Asn	21	AGT Ser	10
ATC Ile	16	ACC Thr	21	AAC Asn	24	AGC Ser	24
ATA Ile	3	ACA Thr	3	AAA Lys	45	AGA Arg	0
ATG Met	31	ACG Thr	11	AAG Lys	17	AGG Arg	0
GTT Val	15	GCT Ala	17	GAT Asp	44	GGT Gly	14
GTC Val	14	GCC Ala	23	GAC Asp	16	GGC Gly	21
GTA Val	10	GCA Ala	25	GAA Glu	29	GGA Gly	2
GTG Val	22	GCG Ala	37	GAG Glu	28	GGG Gly	10

charged residues, consisting of 117 (12.1%) acidic and 106 (11.0%) basic residues. This would give a net charge of -11, indicating that the isoelectric point of Protease III would be slightly acidic. Removal of the signal peptide (see below) would alter the net charge to -13.

Sequence Comparisons

Protease III is known to be located in the periplasmic space (2) and must therefore be exported across the inner cell membrane. In common with other secreted proteins, the mature enzyme would be expected to be derived from a larger precursor protein by proteolytic removal of an N-terminal signal peptide. Comparison of many such signal sequences has revealed certain common features. They are short sequences, typically between 16 and 36 residues, with a basic N-terminus and a central hydrophobic core. In addition, only certain amino acid residues are found to occupy specific positions in the sequence, in particular the -1 and -3 positions from the site of cleavage (18). All of these diagnostic features are seen at the N-terminus of the predicted Protease III amino acid sequence (Figure 1). By comparison of many signal sequences with known cleavage sites, von Heijne (19) has described a weight-matrix method for both estimating the likelyhood that a particular sequence represents a signal peptide and for predicting with a high degree of probability the site of cleavage between the signal peptide and the mature exported protein. Using this method, the proposed Protease III signal sequence gives a score of 12.6, which is comparable with that of two known cleaved and exported *E. coli* proteins, the *phoA* and *bla* gene products (18), which score 11.0 and 10.5 respectively. It is predicted that cleavage of the proposed Protease III signal sequence would occur between residues Ala-23 and Glu-24 (assuming that Met-1 has not been removed from the protein). The molecular mass of the mature Protease III protein would consequently be 105,124 daltons.

At least one *E. coli* protease, La (the lon gene product), is known to be induced as part of the heat-shock response (20,21). No sequence that reasonably fits the consensus for heat-shock promoters (22) could be found in the region 5' to the *ptr* coding sequence.

The Protease III amino acid sequence does not contain the so-called 'catalytic triad' (His, Asp and Ser residues with a fixed spacing within the primary protein structure), or a set of other conserved residues which are characteristically found in serine proteases (23). This is in agreement with experimental data showing that Protease III is not inactivated by known serine protease inhibitors (24).

DISCUSSION

We have determined the nucleotide sequence of a 3120 bp region of the E. coli chromosome that includes the entire ptr gene. The 962 amino acid protein encoded by ptr would have an unmodified molecular mass of 107,719 daltons.

We have identified a sequence at the N-terminus of the predicted Protease III amino acid sequence which contains all of the known features characteristically found in signal sequences. Assuming that the proposed signal sequence is cleaved, the mature Protease III protein located in the periplasmic space would be shorter by 23 residues and would have a molecular mass of 105,124 daltons, in agreement with the values estimated for Protease III from SDS-PAGE (1,4).

It has been reported that the ptr locus encodes two polypeptides of molecular masses 110 kDa and 50 kDa (4). Transposon inactivation studies show that p50 is encoded in the N-terminal portion of the ptr coding sequence (4). Limited proteolysis of the Protease III and p50 proteins indicates that the two proteins share some homology (4), as would be expected if they were derived from the same reading frame. Examination of the nucleotide sequence of the ptr gene fails to reveal a separate open reading frame that would encode a protein of molecular mass 50 kDa. This supports the contention that p50 is derived from the Protease III coding frame (4), but raises a question about the derivation of p50. One possibility is that transcription of ptr occasionally terminates prematurely. Examination of the ptr sequence in the region where transcription would be expected to terminate in order to generate a 50 kDa protein fails to reveal the presence of any obvious Rho-independent transcription terminator structure (a stem and loop structure followed by a run of T residues). However, this does not rule out the possibility that p50 is produced as a result of premature transcription termination. Since little p50 protein is seen in cells that have been osmotically shocked (4), it has been suggested that p50 is not only found but is also predominantly produced in the periplasmic space, and represents a stable proteolytic fragment of Protease III.

Analysis of the sequence 5' to the ptr coding sequence in the recC-ptr intergenic region reveals the presence of a putative promoter sequence. Using the scoring system of Mulligan et al. (25), which is based on comparisons with many known promoters, the homology score for this promoter sequence is 58.3%, which is well above the arbitrary level (45%) set by

Mulligan et al. for promoters able to operate efficiently *in vitro*.

It is intriguing that a gene coding for a protease is located between two genes known to form subunits of an exonuclease involved in DNA repair and genetic recombination. This location may be fortuitous or, alternatively, Protease III may be involved with the control or function of Exonuclease V. A possible relationship between Protease III and Exonuclease V is suggested by the finding that the structural genes for *ptr* and *recB* actually overlap. Combined with the observation (26) that the *recB* gene overlaps at its 3' end with the *recD* gene (which also codes for a subunit of Exonuclease V), there is a real possibility that these three genes form an operon. As few examples exist of operons containing two or more genes of completely unrelated function, the possibility that Protease III and Exonuclease V interact *in vivo* is worthy of further investigation. The observation that Protease III exists predominantly in the periplasmic space suggests that a role for the protease in DNA repair or genetic recombination is unlikely. However, it is possible that the unmodified cytoplasmic (Pro)protease III protein performs some cellular role prior to its export into the periplasm.

Konigsberg and Godson (16) have suggested that there may be a clustering of rare codons in genes forming part of an operon, with the result that transit of ribosomes along the mRNA is temporarily halted. In the *ptr* gene, there appears to be a clustering of rare codons at the 3' end of the gene, which may serve to regulate expression of not only the *ptr* gene, but also the downstream *recB* and *recD* genes.

It has been reported that certain mutations affecting Exonuclease V activity actually appear to map in the *ptr* gene (5). Our finding that the *ptr* and *recB* genes overlap, and may possibly be coordinately controlled, provides a possible mechanism whereby certain mutations in *ptr* could be polar on *recB*.

ACKNOWLEDGMENTS

We thank Jean Wake for typing the manuscript and the Medical Research Council for financial support.

REFERENCES

1. Cheng, Y-S.E. and Zipser, D. (1979) J. Biol. Chem. 254, 4698-4706.
2. Swamy, K.H.S. and Goldberg, A.L. (1982) J. Bacteriol. 149, 1027-1033.
3. Cheng, Y-S.E., Zipser, D., Cheng, C-Y. and Rolseth, S.J. (1979) J. Bacteriol. 140, 125-130.

4. Dykstra, C.C. and Kushner, S.R. (1985) J. Bacteriol. 163, 1055-1059.
5. Dykstra, C.C., Prasher, D. and Kushner, S.R. (1984) J. Bacteriol. 157, 21-27.
6. Umeno, M., Sasaki, M., Anai, M. and Takagi, Y. (1983) Biochem. Biophys. Res. Commun. 116, 1144-1150.
7. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
8. Biggin, M.D., Gibson, T.J. and Hong, C.F. (1983) Proc. Natl. Acad. Sci. USA 80, 3963-3965.
9. Hickson, I.D., Atkinson, K.E., Hutton, L., Tomkinson, A.E. and Emmerson, P.T. (1984) Nucl. Acids Res. 12, 3807-3819.
10. 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.
11. Queen, C. and Korn, L.J. (1984) Nucl. Acids Res. 12, 581-599.
12. Staden, R. (1984) Nucl. Acids Res. 12, 551-567.
13. Steitz, J.A. (1979) In Ribosomes (Chambliss, G., Craven, G.R., Davies, J., Kahan, L. and Nomura, M., eds) pp. 349-399, University Park Press, Baltimore.
14. Finch, P.W., Storey, A., Chapman, K.E., Brown, K., Hickson, I.D. and Emmerson, P.T. (1986) Nucl. Acids Res. (to be submitted).
15. Hawley, D.K. and McClure, W.R. (1983) Nucl. Acids Res. 11, 2237-2255.
16. Konigsberg, W. and Godson, G.N. (1983) Proc. Natl. Acad. Sci. USA 80, 687-691.
17. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
18. von Heijne, G. (1985) J. Molec. Biol. 184, 99-105.
19. von Heijne, G. (1986) Nucl. Acids Res. 14, 4683-4690.
20. Goff, S.A., Casson, L.P. and Goldberg, A.L. (1984) Proc. Natl. Acad. Sci. USA 81, 6647-6651.
21. Phillips, T.A., VanBogelen, R.A. and Neidhardt, F.C. (1984) J. Bacteriol. 159, 283-287.
22. Gayda, R.C., Stephens, P.E., Hewick, R., Schoemaker, J.M., Dreyer, W.J. and Markovitz, A. (1985) J. Bacteriol. 162, 271-275.
23. Woodbury, R.G., Katunuma, N., Kobayashi, K., Titani, K. and Neurath, H. (1978) Biochemistry 17, 811-819.
24. Goldberg, A.L., Swamy, K.H.S., Chung, C.H. and Larimore, F. (1982) Methods in Enzymol. 80, 680-702.
25. Mulligan, E.M., Hawley, D.K., Entriken, R. and McClure, W.R. (1984) Nucl. Acids Res. 12, 789-800.
26. Finch, P.W., Storey, A., Brown, K., Hickson, I.D. and Emmerson, P.T. (1986) Nucl. Acids Res. (to be submitted).