The overexpression and complete amino acid sequence of *Escherichia coli* 3-dehydroquinase

Kenneth DUNCAN, Subhendhu CHAUDHURI, Murray S. CAMPBELL and John R. COGGINS
Department of Biochemistry, University of Glasgow, Glasgow G12 8QQ, Scotland, U.K.

The enzyme 3-dehydroquinase was purified in milligram quantities from an overproducing strain of *Escherichia coli*. The amino acid sequence was deduced from the nucleotide sequence of the *aroD* gene and confirmed by determining the amino acid composition of the overproduced enzyme and its *N*-terminal amino acid sequence. The complete polypeptide chain consists of 240 amino acid residues and has a calculated subunit M_r of 26377. Transcript mapping revealed that *aroD* is a typical monocistronic gene.

INTRODUCTION

The enzyme 3-dehydroquinase catalyses the third step of the early common pathway (the pre-chorismate pathway) for the biosynthesis of aromatic amino acids. Three classes of biosynthetic 3-dehydroquinases are known. The *Escherichia coli* enzyme is monofunctional (Berlyn & Giles, 1969; Chaudhuri *et al.*, 1986), the plant enzyme occurs on a bifunctional polypeptide chain with shikimate dehydrogenase (Polley, 1978; Koshiba, 1978; Coggins *et al.*, 1985), and the *Neurospora crassa* enzyme occurs on a pentafunctional polypeptide chain that carries the activities of the five central steps of the pre-chorismate pathway (Lumsden & Coggins, 1977, 1978; Gaertner & Cole, 1977; Smith & Coggins, 1983; Lambert *et al.*, 1985; Coggins *et al.*, 1985).

3-Dehydroquinase occurs in very low amounts in *E. coli* and has to be purified over 4000-fold to achieve homogeneity (Chaudhuri *et al.*, 1986). To make a detailed comparison of the monofunctional *E. coli* enzyme with the multifunctional enzymes found in other species milligram quantities of the pure enzyme are required rather than the microgram quantities previously obtained (Chaudhuri *et al.*, 1986). To facilitate such studies we have subcloned the *E. coli* gene (*aroD*) that encodes 3-dehydroquinase from a previously isolated plasmid carrying this gene (Kinghorn *et al.*, 1981) and constructed a strain that overproduces the enzyme. We have also located the *aroD* gene precisely, determined its complete nucleotide sequence, and confirmed the predicted amino acid sequence of 3-dehydroquinase by determination of the *N*-terminal amino acid sequence and the amino acid composition of the overproduced enzyme.

MATERIALS AND METHODS

Reagents

All the reagents, purified proteins and enzymes used in the enzyme purification, except those purified below, were from BDH Chemicals (Poole, Dorset, U.K.), Boehringer Corp. (Lewes, East Sussex, U.K.), Sigma Chemical Co. (Poole, Dorset, U.K.) or Aldrich Chemical Co. (Gillingham, Dorset, U.K.). DEAE-Sephacel and Sephacryl S200 (superfine grade) were obtained from Pharmacia (Milton Keynes, Bucks., U.K.) and Bio-Gel

HTP from Bio-Rad Laboratories (Watford, Herts., U.K.). Ammonium 3-dehydroquinolate was prepared by the method of Grewe & Haendler (1968).

Subcloning

All restriction enzymes were purchased from Bethesda Research Laboratories (Paisley, Scotland, U.K.) and were used under the conditions recommended by the manufacturer. Bacteriophage-T4 DNA ligase was purchased from Amersham International (Amersham, Bucks, U.K.). Subdigestion of restriction fragments and the isolation of DNA fragments for subcloning after electrophoresis through low-melting-temperature agarose (Bethesda Research Laboratories) were carried out as described in Maniatis *et al.* (1982). Bacterial transformation was by the method of Dagert & Ehrlich (1979).

Cloning in M13

All cloning in M13 was carried out as described in the M13 Cloning and Sequencing Handbook (Amersham International, 1983), with the paired vectors M13mp8 and M13mp9 (Messing & Vieira, 1982).

DNA sequence analysis

The reagents and enzyme for DNA sequencing by the dideoxy method (Sanger *et al.*, 1977; Biggin *et al.*, 1983) were purchased in the form of kits from Amersham International and were used in the protocols described in the M13 Cloning and Sequencing Handbook (Amersham International, 1983). The label used throughout was [α - ^{35} S]thio[dATP (Amersham SJ304). Electrophoresis was in 8 M-urea/6% (w/v) polyacrylamide linear- or buffer-gradient gels (0.4 cm $\times$ 40 cm $\times$ 20 cm).

Transcript mapping

RNA was prepared from *E. coli* AB2848/pKD201 by the hot-phenol method, as described in Duncan & Coggins (1986). The synthetic primer (complementary to sequence positions 729–753) was generously given by Searle Research and Development (High Wycombe, Bucks., U.K.). RNA (approx. 15 μ g) was annealed for 1 h at 55 °C to primer DNA (2.5 ng) in 10 mM-Tris/HCl buffer, pH 7.0, containing 10 mM-MgCl₂ (10 μ l). To this was added 1 μ l of [α - ^{35} S]thio[dATP, 20 units of avian-myeloma-virus reverse transcriptase (Northumbria

Fig. 1. (a) Restriction map of plasmid pJKK12 and (b) sequencing strategy of the 1.8 kb *ClaI* genomic insert in plasmid pKD201

(a) Enzyme cleavage sites in the insert shown above the line are from Kinghorn *et al.* (1981). The bar labelled 'pKD201' shows the location of the *ClaI* fragment, subcloned in plasmid pKD201. (b) The box indicates the location of the *aroD* coding sequence; the solid black portion is that part of the sequence confirmed by *N*-terminal protein sequencing of 3-dehydroquinase. Arrows indicate the direction and extent of sequences.

Biologicals Ltd., Cramlington, Northumberland, U.K.) and dCTP, dGTP and dTTP, each to a final concentration of 1 mM. After incubation for 30 min at 30 °C, 4 µl of formamide/dye mixture was added to each sample before electrophoresis in an 8 M-urea/6% acrylamide sequencing gel.

Purification of 3-dehydroquinase

E. coli strain AB2848/pKD201 was grown at 37 °C in 3-litre stirred cultures with vigorous aeration on minimal medium (M9 medium containing 0.2% glucose) (Maniatis *et al.*, 1982). Cells were grown into stationary phase (A_{650} 1.3–1.5) and then harvested by centrifugation and stored at –20 °C.

The cells were broken and the enzyme purification was carried out by using essentially the same procedure as described previously for the wild-type enzyme (Chaudhuri *et al.*, 1986), except that the final step (chromatography on a Mono Q anion-exchange column) was not required.

Assay of 3-dehydroquinase activity

The enzyme was assayed at 20 °C in 100 mM-potassium phosphate buffer, pH 7, containing 0.33 mM-ammonium dehydroquininate. Formation of 3-dehydroshikimate was monitored at 234 nm in a total volume of 1 ml ($\epsilon = 1.2 \times 10^4 \text{ M}^{-1} \cdot \text{cm}^{-1}$).

Determination of protein

Protein was determined by the method of Bradford (1976), with bovine serum albumin as standard.

Polyacrylamide-gel electrophoresis

Polyacrylamide-gel electrophoresis in the presence of SDS was performed according to the method of Laemmli (1970), with 3.5% stacking gel and 12.5% running gel. Gels were stained for protein with AgNO_3 (Eschenbruch & Bürk, 1982).

Automatic amino acid sequence determination

This was carried out as previously described (Duncan *et al.*, 1984), with a Beckman model 890 liquid-phase sequencer. The phenylthiohydantoin samples were analysed by chromatography on a Waters Resolve C₁₈ reverse-phase column with a pH 5.0 acetate/acetonitrile buffer system (Carter *et al.*, 1983).

Amino acid analysis

Samples of performic acid-oxidized 3-dehydroquinase (Hirs, 1967) were hydrolysed and analysed on an LKB model 4400 amino acid analyser as described previously (Lumsden & Coggins, 1978).

RESULTS

Subcloning of the *aroD* gene

The plasmid pJJK12 (Kinghorn *et al.*, 1981) carries 5.6 kb of *E. coli* genomic DNA, inserted into the *Pst*I site of plasmid pBR322. As this is sufficiently large to encode several genes, it was decided to subclone *aroD* on to a smaller DNA fragment, suitable for DNA sequence analysis. Previous attempts at this had been unsuccessful, but did yield the information that the *aroD* gene spans one of the two *Bgl*II sites located on the insert (Kinghorn *et al.*, 1981). A number of restriction-enzyme target sites were mapped on plasmid pJJK12, and two *Cla*I fragments of 1.8 kb and 2.2 kb were identified that each had a central *Bgl*II site (Fig. 1). The 1.8 kb and 2.2 kb *Cla*I fragments were isolated and cloned into the *Cla*I site of plasmid pAT153 (Twigg & Sherratt, 1981). A recombinant carrying the 1.8 kb *Cla*I fragment (designated pKD201) was able to complement the *aroD* mutation in *E. coli* AB2848 (Pittard & Wallace, 1966). A crude cell extract of *E. coli* AB2848/pKD201 was prepared by sonication as described in Duncan & Coggins (1986). The specific activity of 3-dehydroquinase was found to be 3.0 units/mg of protein, approx. 100-fold higher than that found in wild-type *E. coli* K12 (0.032 unit/mg). This level of overexpression is consistent with the known copy number of plasmid pAT153 (approx. 100).

DNA sequence analysis

The genomic DNA encoding *aroD* was isolated after electrophoresis of a *Cla*I digest of plasmid pKD201 through low-melting-temperature agarose. Portions were subdigested (without further purification) with *Taq*I and *Hpa*II, revealing four *Taq*I and five *Hpa*II sites within the insert. The partial restriction mapping technique of Smith & Birnstiel (1976) was employed to locate these sites (results not shown); they were found to be distributed in such a way that the entire 1.8 kb sequence could be determined by analysis of cloned *Taq*I and *Hpa*II fragments, along with the sequence either side of the unique *Bgl*II site.

Plasmid pKD201 was digested with the enzymes *Eco*RI and *Hind*III, which cleave at sites flanking the *Cla*I insert. The insert fragment was cloned into the paired vectors M13mp8 and M13mp9, and the DNA sequence at either end of the insert was obtained. A *Bgl*II-*Bam*HI digest of plasmid pKD201 gives two fragments; these were cloned into the *Bam*HI site of M13mp8, and the sequences either side of the *Bgl*II site were determined. The *Cla*I fragment was isolated from plasmid pKD201 and subdigested with *Taq*I and *Hpa*II. Fragments were separated by electrophoresis on low-melting-temperature agarose, and each band was excised. The DNA was extracted and purified, then ligated to *Acc*I-cleaved vector M13mp8 (end fragments were also cloned, as *Cla*I produces the same 'sticky ends' as do *Taq*I and *Hpa*II). The complete sequence was determined by the dideoxy method (Sanger, 1981) and compiled by using the Staden (1980) programs as described previously (Duncan *et al.*, 1984). The sequencing strategy is summarized in Fig. 1 and the sequence is shown in Fig. 2.

Both strands of the DNA sequence were translated and ORFs identified. Only one ORF was found that was sufficiently long to encode an enzyme of the *M_r* of *E. coli*

<u>ClaI</u>	
1 ATCGATGCCGAACAACAACCACTGGCGTGTGAATATGTTGATGAATTAACACAGCTGCCAA	60
SerMetPro----->ORF	
61 ACTGGTGGGGCCATCAACACCATTGTTAATGATGATGGCTATCGCTGGGTATAACAC	120
121 CGACGGCAGCGGCCATATTGCGCCATTAAAGAGAGCGGTTTGTATCAAAAGGCAAAAC	180
181 GATGGTCTGTTAGGGGCGGTGGTGCCTCAACGGCAATGGCGCGAGGGGCAATTGA	240
241 AGGTTTAAAGAAATTAACCTCTTAAACCGTCGGGATGAGTCTTCGATAAAGCCCTCGC	300
301 CTTGCGCAGCGGTTAATGAAAAACCGATTGTGTGTCGTCACGGTCACCGATCTCGCCGAT	360
361 CAGCAAGCCTTGTCTGAAGCCCTGGCTCCGCGACATTTTAAACCAATGGCAAAAAGT	420
421 GGGTATGAAACCTTGAGAATGAATCATTGGTTAATGATATCGATCTGTTACATCCGGGA	480
481 CTTCTGCTCACTGAACTGGCGGTGATAACCCGATATGACGAAGTTATTGCAGCAGCGG	540
ORF----->ThrArgIleEnd	
-35 TTGACA <17bp> -10 TATAAT Consensus -35 -10B -10A *	
541 CAACAAGCTGGTTGCAAAACGATTGATGGATACGGCATGTTGTTGGCAAGGGGCTGAA	600
601 CAGTTCACATTATGGACTGGCAAGATTTCCTCTGGAATATGTTAAACAGGTCATGGGG	660
661 TTCGTTGCTGACAGGCTGACCGCGTCAGAAAGGCTAAAGAAATGAAACCGTAACGTGA	720
<u>BglII</u> RBS <i>aroD</i> coding region MetLysThrValThrVal	
721 AAAGATCTCGTCATTGGTACGGGCGCACCTAAATCATCGTCTGCTGATGGCGGAAGAT	780
LysAspLeuValIleGlyThrGlyAlaProLysIleIleValSerLeuMetAlaLysAsp	
781 ATCGCCAGCGTGAATCCGAAGCTCTCGCCTATCGTGAAGCGGACTTTGATATTCTGGAA	840
IleAlaSerValLysSerGluAlaLeuAlaTyrArgGluAlaAspPheAspIleLeuGlu	
841 TGGCGTGTGGACCACTATGCCGACCTCTCAATGTGGAGTCTGTCTATGGCGGACGAAAA	900
TrpArgValAspHisTyrAlaAspLeuSerAsnValGluSerValMetAlaAlaLys	
901 ATTCTCGTGAGACCATGCCAGAAAAACCGCTGCTGTTTACCTCCGAGTGCCAAAGAA	960
IleLeuArgGluThrMetProGluLysProLeuLeuPheThrPheArgSerAlaLysGlu	
961 GCGGCGGACAGCGGATTTCCACCGAGGCTTATTATTCAGTCTGTCGACCATCGAC	1020
GlyGlyGluGlnAlaIleSerThrGluAlaTyrTyrCysThrHisArgAlaAlaIleAsp	
1021 AGCGGCTCGTGTATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT	1080
SerGlyLeuValAspMetIleAspLeuGluLeuPheThrGlyAspAspGlnValLysGlu	
1081 ACCGTCGCTACGCCACGCGCATGATGTGAAGTAGTCTGATGCAACCATGACTTCCAT	1140
ThrValAlaTyrAlaHisAlaHisAspValLysValValMetSerAsnHisAspPheHis	
1141 AAAACGCGGAGCGGCAAGAAATCATTGCCGCTCGCGCAAAATGCAATCCTTCGACGCC	1200
LysThrProGluAlaGluGluIleIleAlaArgLeuArgLysMetGlnSerPheAspAla	
1201 GATATTCTAAGATTGCGCTGATGCGGCAAGTACCAGCATGTGCTGACGCTGCTTGGC	1260
AspIleProLysIleAlaLeuMetProGlnSerThrSerAspValLeuThrLeuLeuAla	
1261 GCGACCTCGAGATGCAGGACGAGTATGCGCATCGTCCAATTATCAGCATGTCGATGGCA	1320
AlaThrLeuGluMetGlnGluGlnTyrAlaAspArgProIleIleThrMetSerMetAla	
1321 AAAAATGGCGAAATTTCTCGTCTGGCTGGTGAAGTATTGGCTCTGGCGGCAACTTTGG	1380
LysThrGlyGluIleSerArgLeuAlaGlyGluValPheGlySerGlyGlyAsnPheTrp	
1381 TGGGTAAAAAAGCGTCTCGCGCAGGCAATCTCGGTAATGATTGCGCAGCGTATT	1440
CysGlyLysLysSerValCysAlaArgAlaAsnLeuGlyLysEnd	
T	
1441 AACTATTTTACACCGGATAAGCAATAATATTTCCGCGGGAACACCTCCCGCGCAAC	1500
1501 TAAAAATATATTCAATCGTATTTAATAAAAAATATTTGCTGAGTCTCTGTGCGTAAATTC	1560
1561 TCCATTGGCGTAGGGAAATCAGATCTGAATCAGGAATTAACAACTGAAACCTGAAAAAC	1620
1621 CACCTCGTATTAATGACGAGTGGCGGCTCTGCGCACAGGAAGCGGTAATTATATTC	1680
1681 CCGACGAAGCAACACTTTGTGTGTTAGCGGCTGGCGCGGTATTCTGGAAGCCACACGT	1740
<u>ClaI</u>	
1741 TAATTACTGCTCTTGCTGATAAATAAACAGACATCAACACCAAGTAATTTATCGAT	1798

Fig. 2. Complete DNA sequence of the 1.8 kb *Cla*I insert in plasmid pKD201, and the *aroD* coding sequence

RBS indicates the probable ribosome-binding site of the gene. -10A and -10B indicate possible promoter -10 regions; -35 indicates the possible promoter -35 region.

* Indicates the transcription start site.

3-dehydroquinase; this ORF begins at position 700, ends at position 1422, and spans the *Bgl*II site (Fig. 2). The program TESTCODE (Devereux *et al.*, 1984) was applied to both strands of the DNA sequence.

Fig. 3. Application of Fickett's 'Testcode' statistic to the sequence of the 1.8 kb *Clal* fragment

The resulting plot is divided into three regions for which the statistic makes a prediction. Region A is the region within which 95% of coding regions fall and within which 95% of the non-coding regions do not fall. Region C is the opposite. The middle ground (B) is the 'window of vulnerability' where the statistic does not have a significant ability to predict coding and non-coding regions.

TESTCODE searches for genes by plotting the running value of Fickett's 'Testcode' statistic along the length of the sequence. The test depends on finding statistical order in the base sequence within protein-coding regions. This order is based on the non-random utilization of codons and provides an objective means of identifying protein-coding sequences (Fickett, 1982). The resulting plot for the ORF-encoding strand is shown in Fig. 3. The ORF has a 'Testcode' statistic that falls into the region where 95% of the coding regions fall; this indicates that the ORF encodes a protein that must, in this case, be the *aroD*-gene product, 3-dehydroquinase. It may be noted that there is a second ORF, which begins at the *Clal* site and runs to position 516 of the sequence (Fig. 2). The 'Testcode' statistic predicts that this is also likely to be protein-coding, i.e. the 3'-end of a gene. The identity of this gene is unknown. A similar plot for the complementary DNA strand (not shown) did not predict any significant protein-coding regions. The location of the *aroD* gene was confirmed by *N*-terminal protein sequence analysis on a sample of purified 3-dehydroquinase. The *aroD*-gene coding sequence begins with the ATG/Met at position 703 and ends with the TGA/stop codon at position 1423. The gene encodes a polypeptide of 240 amino acid residues, the sequence of which is shown in Fig. 2.

Purification of overproduced 3-dehydroquinase

The enzyme was purified to homogeneity as described previously for the wild-type enzyme (Chaudhuri *et al.*, 1986), except that the final purification step (chromatography on a Mono Q anion-exchange column) was found to be unnecessary. A purification scheme is shown in Table 1, and the elution profile for the final step in the

purification (chromatography on Sephacryl S200) is shown in Fig. 4; the constant specific activity throughout the protein peak in this profile together with the exact coincidence of the protein and activity peaks was the first evidence of the success of the purification. The progress of the purification was monitored by SDS/polyacrylamide-gel electrophoresis (Fig. 5); the purified enzyme showed a single protein band, which had the same mobility as the single protein band seen in a sample of 3-dehydroquinase purified from a wild-type K12 strain of *E. coli* (Fig. 5) (Chaudhuri *et al.*, 1986).

The yield of pure protein was 7 mg from 20 g of *E. coli* AB2848/pKD201 cells, and the specific activity of the purified enzyme was 347 units/mg of protein. This compares with the 70 μ g of enzyme of similar specific activity obtained from 20 g of 'wild-type' cells (Chaudhuri *et al.*, 1986).

The *N*-terminal sequence of 3-dehydroquinase was established by using a liquid-phase sequencer (see Table 2). The first 17 amino acid residues were identified unambiguously; residue 18 could not be identified. Residues 19 and 20 were identified as isoleucine and valine, respectively and residue 21 was tentatively identified as serine. These results are in complete agreement with the amino acid sequence predicted from the DNA sequences (see Fig. 2).

The amino acid composition is shown in Table 3 alongside that predicted from the DNA sequence; the two sets of data are in good agreement. The subunit M_r calculated from the predicted amino acid sequence is 26377. This compares favourably with the estimated M_r of 29000 obtained for the purified enzyme by SDS/polyacrylamide-gel electrophoresis (Fig. 5). The codon utilization for *aroD* is shown in Table 4. The pattern

Table 1. Purification scheme for 3-dehydroquinase from plasmid-pKD201-transformed *E. coli* AB2848

The results presented are for a typical purification starting from 20 g of *E. coli* cells (see the Materials and methods section).

Step	Volume (ml)	Concn. of protein (mg/ml)	Total protein (mg)	Activity (units/ml)	Total activity (units)	Specific activity (units/mg)	Purification (fold)	Yield (%)
1. Crude extract	66	14	924	97	6400	6.9	1	100
2. 55-75%-satn. $(\text{NH}_4)_2\text{SO}_4$	32	7.8	250	165	5280	21.3	3.4	82
3. DEAE-Sephacel	49	0.76	37.2	103	5050	135	19.6	79
4. Hydroxyapatite	53	0.30	15.9	75	3980	250	36.2	61
5. Sephacryl S-200	20.5	0.37	7.50	128	2600	347	54	41

Fig. 4. Chromatography of overproduced *E. coli* 3-dehydroquinase on Sephacryl S200 (step 5 of purification scheme)

Enzyme from step 4 (3980 units in 53 ml) was concentrated and applied to a Sephacryl S200 column (85 cm $\times$ 2.1 cm) and eluted as described in the Materials and methods section (pool I fractions 145–148 and pool II fractions 149–162). $\circ$, Protein (A_{280}); $\bullet$, 3-dehydroquinase activity (units/ml); $\blacktriangle$, specific activity (units/ A_{280}).

of utilization is similar to that observed for the *aroA* (Duncan *et al.*, 1984; Duncan & Coggins, 1986), *aroB* (Millar & Coggins, 1986) and *aroE* (Anton, 1985) genes, and for *E. coli* genes in the 'moderately to weakly' expressed class (Gouy & Gautier, 1982; Grosjean & Fiers, 1982). It has been proposed that the efficiency of translation of an *E. coli* mRNA is affected by the frequency of use of codons in a degenerate set that optimizes the codon-anticodon interaction. The ratio of optimal/suboptimal codons in *aroD* is 0.81, and that of the 'moderately to weakly' expressed class (data in Table III of Grosjean & Fiers, 1982) is 0.82.

Transcript mapping

No 'regulatory' mutants of *aroD* have been described, and there is no evidence in the literature that *aroD* forms part of an operon, or that *aroD* expression is modulated in any way, for example, by a repressor. However, this was also thought to be true for the *E. coli aroA* gene until it was shown that *aroA* is co-expressed with the *serC* gene (Duncan & Coggins, 1986). In order to investigate the nature of the *aroD* cistron, the transcription start site was located by using the reverse-transcriptase run-off procedure (Fig. 6). Several bands are observed in the run-off transcript track. This could be due to a number of factors, including multiple transcription initiation sites, premature termination of reverse transcription, and mRNA degradation. Attention is drawn to the two most prominent bands, labelled A and B. Band A maps the most likely true transcription start site, as it is the longest transcript. This transcript maps to the G residue at

Fig. 5. Purification of overproduced 3-dehydroquinase from *E. coli*

This photomontage of two 12.5% polyacrylamide gels run in the presence of SDS and stained by the AgNO_3 method monitors the purification of 3-dehydroquinase. The numbers on the right-hand side refer to the positions of marker proteins used for M_r calibration: 1, bovine serum albumin (M_r 68 000); 2, rabbit muscle fructose-bisphosphate aldolase (M_r 40 000); 3, sperm-whale myoglobin (M_r 17 200). Panel (a): track A, crude extracts (6 μg of protein); track B, 55–75% -satn. $(\text{NH}_4)_2\text{SO}_4$ fraction (5 μg); track C, enzyme eluted from DEAE-Sephacel column (2.5 μg); track D, enzyme eluted from hydroxyapatite column (2.3 μg). Panel (b): track A, enzyme eluted from Sephacryl S200 column (1.7 μg); track B, enzyme eluted from Sephacryl S200 column (0.5 μg); track C, pure 3-dehydroquinase from wild-type *E. coli* strain A. T. C. C. 14948 (0.6 μg).

position 588 and is preceded by two sequences that are homologous to the consensus *E. coli* promoter -10 region (Hawley & McClure, 1983). Fig. 2 shows these sequences as ' -10A ' and ' -10B '. Upstream of the -10 region, at a distance of 17 bp, the -35 region of the promoter is normally found. No sequence homologous to the consensus *E. coli* promoter -35 sequence (Hawley & McClure, 1983) is found upstream of either -10 region at the normal distance. The best -35 match is probably the sequence TTGCAA at position 552–557, which results in a 'non-ideal' spacing of 20 nucleotides from ' -10A '. ' -10A ' is, however, the more likely -10 region as the transcription start site is five nucleotide residues downstream of the final T of the -10 sequence, and this is the most commonly found spacing.

Under certain conditions (for example increased RNA concentration or shorter incubation time), the band marked B becomes very prominent. The G residue at position 679 mapped by band B is not preceded by any sequences homologous to the *E. coli* promoter sequence. It is likely that this band is due to the reverse transcriptase being unable to continue because of secondary structure in the RNA. There is a short inverted repeat, positions 667–677, which will be capable of forming a stem-loop, causing premature termination of the transcripts. It is occasionally evident from the sequencing ladder alongside that Klenow fragment of DNA

polymerase I also has problems with secondary structure in this region. A band can sometimes be seen in all four tracks at position 678, indicating that the polymerase stops before the stem-loop, and this is followed by a 'compression' in the G and C tracks. The other bands observed are lower in intensity and map to regions preceding weak secondary structure, or before runs of G or C residues. However, we cannot exclude the possibility that there is more than one sequence capable of acting as a promoter, although no sequences with strong homology to the consensus promoter are observed at the appropriate distance upstream of these sites.

DISCUSSION

Recent work in this laboratory has shown that *E. coli* 3-dehydroquinase is a simple dimeric protein (Chaudhuri *et al.*, 1986). This enzyme is mechanistically and structurally very similar to the biosynthetic 3-dehydroquinase component of the *arom* multifunctional enzyme of *N. crassa* (Chaudhuri *et al.*, 1986). The availability of 3-dehydroquinase in large quantities will allow, for the first time, structural studies on this enzyme, leading to the determination of the arrangement of the amino acid residues at the active site of the protein. It is known, for example, that the mechanism of the reaction catalysed by

Table 2. N-Terminal amino acid sequence of *E. coli* 3-dehydroquinase

The sequence was determined as described in the Materials and methods section. A 10 nmol portion of reduced and carboxymethylated protein was loaded on the sequencer.

Residue no.	Amino acid phenylthiohydantoin derivative identified	Yield (nmol)
1	Met	6.4
2	Lys	4.7
3	Thr	2.5
4	Val	5.3
5	Thr	(Trace)
6	Val	3.5
7	Lys	1.0
8	Asp	1.5
9	Leu	1.4
10	Val	0.5
11	Ile	0.5
12	Gly	0.9
13	Thr	0.7
14	Gly	0.2
15	Ala	0.1
16	Pro	0.4
17	Lys	0.1
18	*	—
19	Ile	(Trace)
20	Val	(Trace)

* No amino acid phenylthiohydantoin derivative could be identified.

Table 4. Codon utilization for the *aroD* gene

	T		C		A		G		
T	Phe	5	Ser	3	Tyr	5	Cys	0	T
	Phe	3	Ser	5	Tyr	1	Cys	3	C
	Leu	1	Ser	0	Term.	—	Term.	—	A
	Leu	1	Ser	2	Term.	—	Trp	2	G
C	Leu	1	Pro	2	His	4	Arg	7	T
	Leu	5	Pro	0	His	2	Arg	2	C
	Leu	0	Pro	2	Gln	2	Arg	0	A
	Leu	11	Pro	3	Gln	4	Arg	0	G
A	Ile	9	Thr	3	Asn	2	Ser	2	T
	Ile	7	Thr	8	Asn	2	Ser	4	C
	Ile	0	Thr	0	Lys	16	Arg	0	A
	Met	11	Thr	4	Lys	1	Arg	1	G
G	Val	2	Ala	3	Asp	11	Gly	5	T
	Val	6	Ala	13	Asp	6	Gly	8	C
	Val	4	Ala	6	Glu	11	Gly	0	A
	Val	5	Ala	7	Glu	7	Gly	0	G

the enzyme involves a covalent interaction between the substrate 3-dehydroquinone and a lysine residue (Butler *et al.*, 1974; Chaudhuri *et al.*, 1986), but the precise location of this lysine residue and the histidine residue also implicated in the active site (Chaudhuri *et al.*, 1986) remain to be established.

DNA sequence analysis has shown that the *aroD* gene is monocistronic. A second ORF in the sequenced DNA, which runs in the same direction as *aroD* and has been

Table 3. Amino acid composition *E. coli* 3-dehydroquinase

The experimentally determined amino acid composition of *E. coli* 3-dehydroquinase is compared with the amino acid composition deduced for 3-dehydroquinase from the *E. coli aroD*-gene sequence. Samples were analysed in duplicate after hydrolysis of performic acid oxidized-protein with 6 M-HCl at 105 °C for 24, 48, 72 and 96 h. The eight experimental values were simply averaged, except where indicated in the footnotes. Abbreviation: N.D., not determined.

Amino acid	Composition (residues/molecule)	
	Relative amino acid composition based on Leu = 19 residues/molecule	Theoretical amino acid composition predicted from the DNA sequence
Asx	20.0	21
Thr*	15.0	15
Ser*	15.8	16
Glx	27.0	24
Pro	N.D.	7
Gly	13.0	13
Ala	29.9	29
Cys†	3.2	3
Val	15.4	17
Met‡	9.3	11
Ile	14.5	16
Leu	19.0	19
Tyr	4.3	6
Phe	7.0	8
His	5.3	6
Lys	13.8	17
Arg	8.7	10
Trp	N.D.	2

* 24 h time points only.

† Determined as cysteic acid.

‡ Determined as methionine sulphone.

Fig. 6. Transcript mapping of the *aroD* promoter

Autoradiograph of a 6% polyacrylamide sequencing gel showing the result of a reverse-transcriptase run-off mapping of the *aroD* promoter (track 5). Tracks 1-4 are the A, C, G and T tracks respectively of a sequencing ladder, synthesized from the same primer. A and B are discussed in the text.

shown to have a high probability that it is protein-coding, ends 187 bp before the *aroD* translational start point. That *aroD* is not co-expressed with this gene is shown by the location of the transcription start site between the two ORFs; it is preceded by the elements normally associated with an *E. coli* promoter. No regulatory mutants of *aroD* have been described; the gene is thought to be expressed constitutively (Tribe *et al.*, 1976).

The amino acid sequence of *E. coli* 3-dehydroquinase has been compared with that of a second 3-dehydroquinase, the product of the *N. crassa qa-2* gene (Alton *et al.*, 1982; M. E. Case, personal communication). The *qa-2* gene is a member of the inducible *qa* gene cluster, which in some species of fungi is responsible for the degradation of quinate (Giles *et al.*, 1967, 1985). No significant homology can be detected between the enzymes at the amino acid sequence level. Although the enzymes catalyse the same chemical reaction, they play different roles in cellular metabolism. *E. coli* 3-dehydroquinase fulfils a biosynthetic function whereas the inducible *N. crassa* 3-dehydroquinase has a degradative function. Thus either these enzymes must have diverged greatly from a distant ancestral sequence or they must have arisen completely independently.

Several lines of evidence suggest that the monofunctional *E. coli* enzymes corresponding to the five activities found on the *arom* multifunctional enzyme of *N. crassa* are structurally related to functional domains found within the *arom* multifunctional enzyme (Coggins *et al.*, 1985; Coggins & Boocock, 1986). Comparison of the *E. coli* 3-dehydroquinase sequence and with the partial sequence of the *arom* gene of *Aspergillus nidulans*, which is known to include the 3-dehydroquinase domain (Charles *et al.*, 1985), reveals significant homologies.

We thank Bryan Dunbar, Linda Fothergill-Gilmore and John Fothergill for assistance with the amino acid sequencing, which was carried out in the S.E.R.C.-funded protein sequencing facility at Aberdeen University. We also thank Jim Jardine (University of Glasgow) for carrying out the amino acid analyses. We are indebted to Mark Edwards of Searle Research and Development (High Wycombe, Bucks., U.K.) for the oligonucleotide primer, to Sydney Kushner (University of Georgia, Athens, GA, U.S.A.) for plasmid pJKK12 and to Mary Case (University of Georgia, Athens, GA, U.S.A.) for providing the *qa-2* sequence before publication. This work was supported by grants to J. R. C. from the Science and Engineering Research Council (S.E.T. C.).

REFERENCES

- Alton, N. K., Buxton, F., Patel, V., Giles, N. H. & Vapnek, D. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 1955-1959
- Amersham International (1983) *M13 Cloning and Sequencing Handbook*, Amersham International, Amersham
- Anton, I. A. (1985) Ph.D. Thesis, University of Glasgow
- Berlyn, M. B. & Giles, N. H. (1969) *J. Bacteriol.* **99**, 222-230
- Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963-3965
- Bradford, M. M. (1976) *Anal. Biochem.* **72**, 248-254
- Butler, J. R., Alworth, W. L. & Nugent, M. J. (1974) *J. Am. Chem. Soc.* **96**, 1617-1618
- Carter, P. E., Dunbar, B. & Fothergill, J. E. (1983) *Biochem. J.* **215**, 565-571
- Charles, I. G., Keyte, J. W., Brammar, W. J. & Hawkins, A. R. (1985) *Nucleic Acids Res.* **13**, 8119-8128
- Chaudhuri, S. & Coggins, J. R. (1985) *Biochem. J.* **226**, 217-223
- Chaudhuri, S., Lambert, J. M., McColl, L. A. & Coggins, J. R. (1986) *Biochem. J.*, in the press
- Coggins, J. R. & Boocock, M. R. (1986) in *Multifunctional Proteins: Structure and Evolution* (Hardie, D. G. & Coggins, J. R., eds.), pp. 259-281, Elsevier, Amsterdam
- Coggins, J. R., Boocock, M. R., Campbell, M. S., Chaudhuri, S., Lambert, J. M., Lewendon, A., Mousdale, D. M. & Smith, D. D. S. (1985) *Biochem. Soc. Trans.* **13**, 299-303
- Dagert, M. & Ehrlich, S. D. (1979) *Gene* **6**, 23-28

- Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395
- Duncan, K. & Coggins, J. R. (1986) *Biochem. J.* **234**, 49–57
- Duncan, K., Lewendon, A. & Coggins, J. R. (1984) *FEBS Lett.* **170**, 59–63
- Eschenbruch, M. & Bürk, R. R. (1982) *Anal. Biochem.* **125**, 96–99
- Fickett, J. W. (1982) *Nucleic Acids Res.* **10**, 5303–5318
- Gaertner, F. H. & Cole, K. W. (1977) *Biochem. Biophys. Res. Commun.* **75**, 259–264
- Giles, N. H., Partridge, C. W. H., Ahmed, S. I. & Case, M. E. (1967) *Proc. Natl. Acad. Sci. U.S.A.* **58**, 1930–1937
- Giles, N. H., Case, M. E., Baum, J., Geever, R., Huiet, L., Patel, V. & Tyler, B. (1985) *Microbiol. Rev.* **49**, 338–358
- Gouy, M. & Gautier, C. (1982) *Nucleic Acids Res.* **10**, 7055–7074
- Grewe, R. & Haendler, H. (1968) *Biochem. Prep.* **12**, 21–26
- Grosjean, H. & Fiers, W. (1982) *Gene* **18**, 199–209
- Hawley, D. K. & McClure, W. R. (1983) *Nucleic Acids Res.* **11**, 2237–2255
- Hirs, C. H. W. (1967) *Methods Enzymol.* **11**, 197–199
- Kinghorn, J. R., Schweizer, M., Giles, N. H. & Kushner, S. R. (1981) *Gene* **14**, 73–80
- Koshiba, T. (1978) *Plant Cell Physiol.* **20**, 667–670
- Laemmli, U. K. (1970) *Nature (London)* **227**, 680–685
- Lambert, J. M., Boocock, M. R. & Coggins, J. R. (1985) *Biochem. J.* **226**, 817–829
- Lumsden, J. & Coggins, J. R. (1977) *Biochem. J.* **161**, 599–607
- Lumsden, J. & Coggins, J. R. (1978) *Biochem. J.* **169**, 441–444
- Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning*, p. 170, Cold Spring Harbor Laboratory, Cold Spring Harbor
- Messing, J. & Vieira, J. (1982) *Gene* **19**, 269–276
- Millar, G. & Coggins, J. R. (1986) *FEBS Lett.* **200**, 11–17
- Pittard, A. J. & Wallace, B. J. (1966) *J. Bacteriol.* **91**, 1494–1508
- Polley, L. D. (1978) *Biochim. Biophys. Acta* **526**, 259–266
- Sanger, F. (1981) *Science* **214**, 1205–1210
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
- Smith, H. O. & Birnstein, M. L. (1976) *Nucleic Acids Res.* **3**, 2387–2398
- Smith, D. D. S. & Coggins, J. R. (1983) *Biochem. J.* **213**, 405–415
- Staden, R. (1980) *Nucleic Acids Res.* **8**, 3673–3694
- Tribe, D. E., Camakaris, H. & Pittard, A. J. (1976) *J. Bacteriol.* **127**, 1085–1097
- Twigg, A. J. & Sherratt, D. J. (1981) *Nature (London)* **283**, 216–218

Received 30 January 1986/14 April 1986; accepted 6 May 1986