Sequencing and overexpression of the *Escherichia coli* *aroE* gene encoding shikimate dehydrogenase

Ian A. ANTON and John R. COGGINS*

Department of Biochemistry, University of Glasgow, Glasgow G12 8QQ, Scotland, U.K.

The *Escherichia coli* *aroE* gene encoding shikimate dehydrogenase was sequenced. The deduced amino acid sequence was confirmed by *N*-terminal amino acid sequencing and amino acid analysis of the overproduced protein. The complete polypeptide chain has 272 amino acid residues and has a calculated M_r of 29380. *E. coli* shikimate dehydrogenase is homologous to the shikimate dehydrogenase domain of the fungal *arom* multifunctional enzymes and to the catabolic quinate dehydrogenase of *Neurospora crassa*.

INTRODUCTION

Micro-organisms and plants synthesize aromatic amino acids from the common precursor chorismate. The route to chorismate, which is known as the shikimate pathway, is the same in all species so far studied (Haslam, 1974; Weiss & Edwards, 1980). However, there are remarkable differences between species in the organization of the enzyme activities that catalyse the seven universal reactions of the shikimate pathway. In the bacterium *Escherichia coli* all seven shikimate pathway activities are separate enzymes (Berlyn & Giles, 1969; Coggins *et al.*, 1985) encoded by widely scattered genes (Pittard & Wallace, 1966). In contrast, in the fungus *Neurospora crassa* five of the seven shikimate-pathway activities (steps 2–6) occur as a multifunctional protein, the *arom* multifunctional enzyme, consisting of two identical pentafunctional polypeptides encoded by a single gene (Giles *et al.*, 1967; Lumsden & Coggins, 1977, 1978; Gaertner & Cole, 1977; Lambert *et al.*, 1985; Coggins & Boocock, 1986). Similar pentafunctional *arom* proteins are found, for example, in the yeast *Saccharomyces cerevisiae* (Larimer *et al.*, 1983; Duncan *et al.*, 1987) and in the fungus *Aspergillus nidulans* (Charles *et al.*, 1986). This system thus provides a stark example of a phenomenon that is not well understood. Why should a particular pathway involve multifunctional proteins in some species but not in others? Before this problem can be properly addressed all the enzymes must be fully characterized. In the case of the *E. coli* monofunctional enzymes this has proved difficult because of the very low concentrations at which these enzymes occur in the cell (Chaudhuri & Coggins, 1985). Gene cloning and over-expression, however, provide a means of obtaining adequate quantities of all the shikimate-pathway enzymes. In the present paper we report the subcloning, sequencing and overexpression of the *E. coli* *aroE* gene, which encodes shikimate dehydrogenase.

This work also forms part of the groundwork for addressing another problem: how, rather than why, did the *arom* protein arise? The working hypothesis in our laboratory was that this protein evolved by the fusion of genes encoding monofunctional enzymes (Coggins *et al.*, 1985; Chaudhuri & Coggins, 1985; Coggins & Boocock,

1986; Hardie & Coggins, 1986). To test this idea it was necessary to compare the amino acid sequence of the *arom* multifunctional enzyme with the amino acid sequences of a set of corresponding monofunctional enzymes, such as those from *E. coli*. Clear precedents for multifunctional proteins arising by a gene fusion mechanism are known from the tryptophan pathway (Miozzari & Yanofsky, 1979; Zalkin & Yanofsky, 1982; Zalkin *et al.*, 1984). Also, evidence has been presented to demonstrate that the *arom* multifunctional enzyme of *A. nidulans* (Charles *et al.*, 1986) contains a region homologous to the monofunctional *E. coli* 5-enolpyruvylshikimate 3-phosphate synthase (Duncan *et al.*, 1984) and has therefore almost certainly arisen by gene fusion. Our own work has been directed towards making a detailed comparison of the sequence of the *S. cerevisiae* *arom* multifunctional enzyme with all five corresponding monofunctional *E. coli* enzymes (Duncan *et al.*, 1987; Coggins *et al.*, 1987). The sequences of four of these five *E. coli* enzymes have already been reported: 3-dehydroquinate synthase (the *aroB*-gene product) (Millar & Coggins, 1986), 3-dehydroquinase (*aroD*) (Duncan *et al.*, 1986), shikimate kinase (*aroL*) (DeFeyer *et al.*, 1986; Millar *et al.*, 1986) and 5-enolpyruvylshikimate 3-phosphate synthase (*aroA*) (Duncan *et al.*, 1984). Here we report the sequence of the fifth enzyme, shikimate dehydrogenase (the *aroE*-gene product).

The *aroE* gene maps at 72 min on the *E. coli* chromosome (Bachmann, 1983) and it is thought to be expressed constitutively (Tribe *et al.*, 1976). The source of the *aroE* gene cloned in our laboratory was the specialized transducing bacteriophage λ spc1 (Jaskunas *et al.*, 1975). The subcloning of the *aroE* gene from this transducing bacteriophage into the plasmid vector pAT153 (Twigg & Sherratt, 1980) has already been described (Anton & Coggins, 1984).

Shikimate dehydrogenase (dehydroshikimate reductase, shikimate:NADP⁺ oxidoreductase, EC 1.1.1.25) catalyses the NADPH-linked reduction of 3-dehydroshikimate. The *E. coli* K12 enzyme has been purified to homogeneity (Chaudhuri & Coggins, 1985) and is a monomer of M_r 30000. No monofunctional biosynthetic shikimate dehydrogenase has previously been sequenced, although the sequences of the shikimate dehydrogenase

Abbreviation used: ORF, open reading frame.

* To whom reprint requests should be sent.

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession no. Y00710.

domains of the *arom* multifunctional enzymes of *A. nidulans* (Charles *et al.*, 1985) and *S. cerevisiae* (Duncan *et al.*, 1987) have been reported.

MATERIALS AND METHODS

Cloning in M13

The paired vectors M13mp8 and M13mp9 were used (Messing & Vieira, 1982). All manipulations were carried out as described in the M13 Cloning and Sequencing Handbook (Amersham International, Amersham, Bucks., U.K.). The strategy used involved both random cloning of *Hpa*II and *Taq*I fragments of the 1.82 kbp *Hind*III–*Cla*I pIA301 insert (into the *Acc*I site of M13mp9) and directed cloning of defined restriction fragments.

DNA sequence analysis

Construction of recombinant M13 clones allowed the use of the dideoxy chain-termination method for DNA sequencing (Sanger *et al.*, 1977; Messing, 1983). All the components for DNA sequence analysis were purchased in the form of kits from Amersham International. Sequencing reactions were carried out as described in the M13 Cloning and Sequencing Handbook (Amersham International). [α - 35 S]Thiol]dATP (Amersham SJ304) was used throughout as label (Biggin *et al.*, 1983). Electrophoresis was on 6% polyacrylamide gels (20 cm $\times$ 40 cm $\times$ 0.4 cm). Gels were dried on Whatman 3MM paper by using a Bio-Rad model 1125 gel dryer before being autoradiographed at room temperature with Fuji RX film.

General DNA manipulations

All restriction enzymes were purchased from Bethesda Research Laboratories, Paisley, Scotland, U.K., and digestions were performed under the conditions recommended by the manufacturer. Bacteriophage T4 DNA ligase was purchased from Amersham International. Plasmid DNAs were prepared as described in Maniatis *et al.* (1982). The isolation of particular restriction fragments after electrophoresis through low-melting-temperature agarose (Bethesda Research Laboratories) was carried out as described in Maniatis *et al.* (1982). Transformation of bacteria with plasmids was by the method of Dagert & Ehrlich (1979).

Construction of pIA321

The 2.47 kbp *Hind*III–*Bam*HI insert fragment from the second smallest *aroE* subclone pIA303 (Anton & Coggins, 1984) consists of the 1.82 kbp *Hind*III–*Cla*I fragment (which forms the insert of pIA301) plus a *Cla*I–*Bam*HI fragment. To clone the 1.27 kbp *Hind*III–*Hinc*II fragment (see the Results section), a *Hinc*II digest of the purified 2.47 kbp *Hind*III–*Bam*HI fragment was ligated with *Sma*I + *Hind*III-cut pKK223-3 and the ligation mix was used to transform the *aroE* auxotrophic mutant *E. coli* AB2834 (Pittard & Wallace, 1966). Following plating on L agar plus ampicillin (100 μ g/ml) and overnight growth, the ampicillin-resistant transformants were tested for the Aro⁺ phenotype by replica-plating on to minimal medium. Six colonies growing on minimal medium were picked, and, from small cultures grown from each, plasmid DNA

was extracted (Holmes & Quigley, 1981) and analysed by restriction-enzyme digests. All contained the desired 1.27 kbp *Hind*III–*Hinc*II fragment and none contained the very small 0.090 kbp *Hinc*II fragment. This construct is pIA321.

Purification of overproduced *E. coli* shikimate dehydrogenase from *E. coli* AB2834/pIA321

This was done by the method of Chaudhuri & Coggins (1985) but with the modifications given below.

Cells of the overproducing strain were grown in L broth plus ampicillin (100 μ g/ml) in 2-litre conical flasks (each with 500 ml of culture) on an orbital shaker at 37 °C. The *lac* inducer isopropyl β -D-thiogalactopyranoside was added to a final concentration of 0.5 mM 1.25 h after inoculation of the flasks (from small feeder cultures). After a total of 5.25 h growth the cultures had reached stationary phase ($A_{650} = 3.7$ – 3.8) and were removed to ice before being harvested by centrifugation.

The Mono Q fast-protein-liquid-chromatography step required with wild-type cells is not required with the overproducing strain, for which the final step is the ADP–Sephacrose column. Owing to the limited capacity of ADP–Sephacrose (Pharmacia) and the large quantities of overproduced enzyme a 20 ml bed volume was used for this column rather than the 5 ml used by Chaudhuri & Coggins (1985). The storage dialysis buffer used was 50 mM-Tris/HCl, pH 7.5, containing 50 mM-KCl, 0.4 mM-dithiothreitol, 1 mM-benzamidine and 50% (v/v) glycerol.

Shikimate dehydrogenase was assayed in the reverse direction at 25 °C as described by Chaudhuri & Coggins (1985) but with 100 mM-Tris/HCl, pH 9.0, as the buffer rather than 100 mM-Na₂CO₃, pH 10.6. One unit of activity is defined as the amount of enzyme required to catalyse the conversion of 1 μ mol of substrate/min under the defined conditions.

Polyacrylamide-gel electrophoresis

Polyacrylamide-gel electrophoresis in the presence of SDS was performed by the method of Laemmli (1970). Non-denaturing polyacrylamide-gel electrophoresis was done by the method of Davis (1964).

Automatic amino acid sequencing

E. coli shikimate dehydrogenase, reduced and carboxy-methylated as described in Lumsden & Coggins (1978), was sequenced in a Beckman model 890 liquid-phase sequencer as in Smith *et al.* (1982). The phenylthiohydantoin derivatives of liberated *N*-terminal amino acids were separated and identified by chromatography on a Waters 5 μ m-particle-size Resolve C₁₈ reverse-phase column with a pH 5.0 acetate/acetonitrile buffer gradient system as in Carter *et al.* (1983). The initial amount of protein sequencing was 53 nmol and the repetitive yield from residues 1–60 was 94% by regression analysis (correlation coefficient 0.96).

Amino acid analysis

Samples of *E. coli* shikimate dehydrogenase were oxidized with performic acid as in Hirs (1967), then hydrolysed and analysed on an LKB model 4400 amino acid analyser as in Lumsden & Coggins (1978) but with only three time points for hydrolysis (24, 48 and 72 h).

Fig. 1. Sequencing strategy of the 1.82 kbp *HindIII*–*ClaI* fragment carrying the *aroE* gene

Only the relevant *Sau3A*, *HaeIII* and *AluI* restriction sites are shown. The horizontal arrows indicate the direction and extent of sequence determinations. At the top '1000' and '0' indicate bp from the *ClaI* site. Left to right is clockwise on the *E. coli* linkage map (in the direction of *trkA*). The box labelled 'AroE' represents the *aroE* coding sequence. The extent of *N*-terminal amino acid sequence determination is indicated by the black shading.

Table 1. Purification scheme for shikimate dehydrogenase from *E. coli* AB2834/pIA321

The results presented are for a purification starting from 20 g wet wt. of cells.

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	75	20.8	1560	414	31100	19.9	1	100
2. 30–55% satn. (NH ₄) ₂ SO ₄	36	24	864	807	29100	33.6	1.7	94
3. DEAE-Sephacel	33	1.9	63	539	17800	284	14.3	57
4. Sephacryl S-200	15.5	1.05	16.3	670	10400	638	32.1	33
5. ADP-Sephacrose	14.6	0.72	10.5	619	9040	860	43.2	29

RESULTS

DNA sequence analysis

The smallest subclone obtained previously was pIA301 (Anton & Coggins, 1984), which carries the *aroE* gene on a 1.82 kbp *HindIII*–*ClaI* fragment. This fragment was sequenced by using the M13/dideoxy chain-termination method. The overall sequencing strategy is summarized in Fig. 1. Both strands were sequenced in their entirety and all the restriction sites used to generate fragments for sequencing were overlapped by other sequenced fragments.

To locate the *aroE* coding region both strands of the sequence were first translated in all three reading frames. This revealed two large open reading frames (ORFs). The smaller of these has 624 bp between stop codons and 615 bp from the start of the first plausible initiation codon (GTG). The 615 bp stretch encodes a polypeptide with an *M_r* less than 23000. The larger ORF has 846 bp between stop codons, 843 bp from the first GTG codon and 816 bp from the first ATG codon. From the first methionine codon this ORF encodes a polypeptide with 272 amino acid residues and a calculated *M_r* of 29380. This agrees well with the subunit *M_r* value of 30000

determined for *E. coli* shikimate dehydrogenase by SDS/polyacrylamide-gel electrophoresis (Chaudhuri & Coggins, 1985). Hence it appeared that the largest ORF was the one likely to encode shikimate dehydrogenase. To confirm this, and to locate precisely the translational initiation codon, shikimate dehydrogenase was purified and the *N*-terminal amino acid sequence determined.

Purification of overproduced *E. coli* shikimate dehydrogenase

To facilitate purification of large amounts of *E. coli* shikimate dehydrogenase the putative *aroE* ORF was placed under the control of the powerful *tac(I)* promoter (De Boer *et al.*, 1983) in the expression vector pKK223-3 (available commercially from Pharmacia). The first *HincII* site to the right of the *HindIII* site in the sequenced region (see Fig. 1) roughly bisects the second largest ORF and gives rise to a 1.27 kbp *HindIII*–*HincII* fragment that carries the whole of the largest ORF but includes only the latter 250 bases of the second largest ORF. The 1.27 kbp fragment was cloned into pKK223-3 (as described in the Materials and methods section) so that it would be transcribed in the *HincII*-to-*HindIII* direction, the correct orientation for over-expressing the

1	ATGGAACCTATGCTGTTTTTGGTAATCCGATAGCCACAGCAAATCGCCATTCATTTCAT	60
	MetGluThrTyrAlaValPheGlyAsnProIleAlaHisSerLysSerProPheIleHis	
61	CAGCAATTTGCTCAGCAACTGAATATTGAACATCCCTATGGGCGCGTGTGGCACCCATC	120
	GlnGlnPheAlaGlnGlnLeuAsnIleGluHisProTyrGlyArgValLeuAlaProIle	
121	AATGATTTTCATCAACACACTGAACGCTTTCTTTAGTGCTGGTGGTAAAGGTGCGAATGTG	180
	AsnAspPheIleAsnThrLeuAsnAlaPhePheSerAlaGlyGlyLysGlyAlaAsnVal	
181	ACGGTGCCTTTTAAAGAAGAGGCTTTTGGCAGAGCGGATGAGCTTACTGAACGGGCAGCG	240
	ThrValProPheLysGluGluAlaPheAlaArgAlaAspGluLeuThrGluArgAlaAla	
241	TTGGCTGGTGTGTTAATACCTCATGCGGTTAGAAGATGGACGCCTGCTGGGTGACAAT	300
	LeuAlaGlyAlaValAsnThrLeuMetArgLeuGluAspGlyArgLeuLeuGlyAspAsn	
301	ACCGATGGTGTAGGCTTGTTAAGCGATCTGGAACGTCTGTCTTTTATCCGCCCTGGTTTA	360
	ThrAspGlyValGlyLeuLeuSerAspLeuGluArgLeuSerPheIleArgProGlyLeu	
361	CGTATTCTGCTTATCGGCGCTGGTGGAGCATCTCGCGGCGTACTACTGCCACTCCTTTCC	420
	ArgIleLeuLeuIleGlyAlaGlyGlyAlaSerArgGlyValLeuLeuProLeuLeuSer	
421	CTGGACTGTGCGGTGACAATACTAATCGGACGGTATCCCGCGGGAAGAGTTGGCTAAA	480
	LeuAspCysAlaValThrIleThrAsnArgThrValSerArgAlaGluGluLeuAlaLys	
481	TTGTTTGGCAGCACTGGCAGTATTCAAGCGTTGAGTATGGACGAACTGGAAGGTCATGAG	540
	LeuPheAlaHisThrGlySerIleGlnAlaLeuSerMetAspGluLeuGluGlyHisGlu	
541	TTGATCTCATTATTAATGCAACATCCAGTGGCATCAGTGGTGATATCCGGCGATCCCG	600
	PheAspLeuIleIleAsnAlaThrSerSerGlyIleSerGlyAspIleProAlaIlePro	
601	TCATCGCTCATTCACAGGCATTTATTGCTATGACATGTTCTATCAGAAAGGAAAACT	660
	SerSerLeuIleHisProGlyIleTyrCysTyrAspMetPheTyrGlnLysGlyLysThr	
661	CCTTTTCTGGCATGGTGTGAGCAGCGAGGCTCAAAGCGTAATGCTGATGGTTAGGAATG	720
	ProPheLeuAlaTrpCysGluGlnArgGlySerLysArgAsnAlaAspGlyLeuGlyMet	
721	CTGGTGGCAGGCGGCTCATGCCCTTCTCTCTGGCAGGCTTCTGCCTGACGTAGAA	780
	LeuValAlaGlnAlaAlaHisAlaPheLeuLeuTrpHisGlyValLeuProAspValGlu	
781	CCAGTTATAAAGCAATTGCAGGAGGAATTGTCCGCGTGA	819
	ProValIleLysGlnLeuGlnGluGluLeuSerAlaEnd	

Fig. 2. Complete nucleotide sequence of the *E. coli aroE* gene and the corresponding amino acid sequence of *E. coli* shikimate dehydrogenase

Nucleotides are numbered in the 5'-to-3' direction beginning with the first residue of the ATG triplet encoding the *N*-terminal methionine residue. The *N*-terminal protein sequence has also been determined directly (see the Results section).

product of the largest ORF. This construct is pIA321. With this plasmid in *E. coli* AB2834 (*E. coli* AB2834/pIA321) one can obtain crude cell extracts with specific activities of shikimate dehydrogenase up to 300 times higher than in equivalent crude cell extracts of *E. coli* K12 (results not shown).

Strain *E. coli* AB2834/pIA321 provided the starting material for the large-scale purification of shikimate dehydrogenase. The purification is summarized in Table 1. The protocol used was the same as that used for the purification of shikimate dehydrogenase from wild-type *E. coli* K12 (Chaudhuri & Coggins, 1985) with the modifications described in the Materials and methods section. In particular, the Mono Q fast-protein-liquid-chromatography step needed as the final step with *E. coli* K12 was unnecessary. With the overproducing strain, almost homogeneous material was obtained after the penultimate Sephacryl S-200 step. The fully purified enzyme showed a single protein band after polyacrylamide-gel electrophoresis in the presence of SDS, as

well as a single protein band after simple non-denaturing polyacrylamide-gel electrophoresis (gels not shown). With both these gel systems the overproduced and wild-type enzymes were indistinguishable. A yield of 10.5 mg of homogeneous enzyme was obtained from 20 g wet wt. of cells of the overproducing strain.

N-Terminal amino acid sequencing and amino acid analysis of overproduced *E. coli* shikimate dehydrogenase

The *N*-terminal amino acid sequence of overproduced *E. coli* shikimate dehydrogenase was determined in an automated liquid-phase sequencer as described in the Materials and methods section. The identification of the first 30 amino acid residues was unambiguous and the sequence was in complete agreement with the protein sequence deduced from the gene sequence of the largest ORF in the 1.82 kbp *Hind*III-*Cla*I region. This confirmed that this ORF was the *aroE* coding sequence and it also established that translation starts at the first ATG

Table 2. Amino acid composition of overproduced *E. coli* shikimate dehydrogenase compared with the amino acid composition predicted from the *E. coli* *aroE* gene sequence

Amino acid composition (residues/molecule)		
Amino acid residue	From amino acid analysis (based on Lys = 8 residues)	Predicted from the DNA sequence
Cys*	3.00	3
Asx	23.5	24
Met†	5.22	5
Thr‡	12.3	12
Ser‡	17.4	17
Glx	27.9	27
Pro	12.8	13
Gly	26.1	25
Ala	30.8	30
Val§	13.3	13
Ile§	17.6	18
Leu	34.5	35
Trp	N.D.	2
Tyr	5.06	5
Phe	14.5	14
His	7.91	8
Lys	(8)	8
Arg	12.9	13

* Determined as cysteic acid.

† Determined as methionine sulphone.

‡ Experimental values were extrapolated to zero time.

§ Values were calculated by using only data from the 72 h time point.

|| Not determined.

codon. The complete nucleotide sequence of the *aroE* gene and the deduced amino acid sequence for shikimate dehydrogenase are shown in Fig. 2.

A further 30 residues of protein sequence were determined; residues 31 and 35 were almost certainly histidine and arginine respectively, in agreement with the gene sequence, but there was slight doubt owing to the h.p.l.c. retention times of the phenylthiohydantoin derivatives being shorter than usual; residue 42 was probably aspartic acid, but the preceding asparagine residue made unambiguous interpretation difficult; residues 36, 46 and 52 could not be identified. All other residues up to and including residue 60 agreed exactly with the gene sequence.

A comparison of the experimentally determined amino acid composition of overproduced *E. coli* shikimate dehydrogenase with the amino acid composition predicted from the nucleotide sequence is shown in Table 2. The overall agreement between the observed and predicted values is good. It is thus extremely unlikely that there are any frameshift errors in the *aroE* gene sequence.

The codon utilization of the *aroE* gene is shown in Table 3. The pattern is similar to that found for *E. coli* genes in the moderately to weakly expressed category (Grosjean & Fiers, 1982; Gouy & Gautier, 1982). From the bias in its codon utilization the second largest ORF in the 1.82 kbp *HindIII*-*Clal* sequence is probably also a gene (results not shown), though its nature remains to be determined.

Table 3. Codon utilization in the *aroE* gene

	T	C	A	G	
T	{ Phe 10	Ser 2	Tyr 5	Cys 2	T
	{ Phe 4	Ser 4	Tyr 0	Cys 1	C
	{ Leu 4	Ser 2	Term -	Term -	A
	{ Leu 8	Ser 2	Term -	Trp 2	G
C	{ Leu 4	Pro 4	His 5	Arg 3	T
	{ Leu 5	Pro 2	His 3	Arg 5	C
	{ Leu 1	Pro 4	Gln 3	Arg 1	A
	{ Leu 13	Pro 3	Gln 7	Arg 3	G
A	{ Ile 9	Thr 4	Asn 9	Ser 5	T
	{ Ile 6	Thr 3	Asn 2	Ser 2	C
	{ Ile 3	Thr 3	Lys 6	Arg 1	A
	{ Met 5	Thr 2	Lys 2	Arg 0	G
G	{ Val 4	Ala 11	Asp 8	Gly 13	T
	{ Val 0	Ala 3	Asp 5	Gly 7	C
	{ Val 4	Ala 6	Glu 11	Gly 4	A
	{ Val 5	Ala 10	Glu 6	Gly 1	G

DISCUSSION

With the results presented above for *E. coli* shikimate dehydrogenase, all five of the monofunctional *E. coli* enzymes that correspond to activities of the *arom* pentafunctional enzyme have been overexpressed, purified and sequenced (Duncan *et al.*, 1984; Millar & Coggins, 1986; DeFeyer *et al.*, 1986; Millar *et al.*, 1986; Duncan *et al.*, 1986). The complete sequence of the *A. nidulans* (Charles *et al.*, 1986) and the *S. cerevisiae* (Duncan *et al.*, 1987) *arom* polypeptides have also been deduced from the appropriate gene sequences. In Fig. 3 the sequence of *E. coli* shikimate dehydrogenase is compared with the sequences of the shikimate dehydrogenase domains of the two *arom* multifunctional enzymes. All three sequences are homologous (for example 25% of the amino acid residues are conserved between the *E. coli* enzyme and the *S. cerevisiae* domain), which is consistent with the proposal that the two *arom* multifunctional enzymes have mosaic structures consisting of five functional domains each one of which is homologous to the corresponding monofunctional *E. coli* enzyme (Charles *et al.*, 1986; Duncan *et al.*, 1987).

In Fig. 3 the shikimate dehydrogenase sequences are also compared with the sequence of the inducible quinate (shikimate) dehydrogenase of *N. crassa* (the *qa-3*-gene product) (Alton *et al.*, 1982). This catabolic enzyme is also clearly homologous to the three biosynthetic shikimate dehydrogenases. Some 19% of the amino acid residues are conserved between the *E. coli* enzyme and the inducible quinate dehydrogenase, and there are several very convincing long blocks of homology particularly within the first 153 residues of the *E. coli* sequence. The number of identical residues in this N-terminal region is 27%. This is in marked contrast with the situation for the catabolic 3-dehydroquinases of *N. crassa* (Hawkins *et al.*, 1982) and *A. nidulans* (Da Silva *et al.*, 1986), which show no recognizable homology with either the *E. coli* biosynthetic 3-dehydroquinase (Duncan *et al.*, 1986) or with the biosynthetic 3-dehydroquinase domains of the two *arom* multifunctional enzymes (Charles *et al.*, 1985; Duncan *et al.*, 1987). It thus appears that in one catabolic pathway in fungi, the

<i>N. crassa</i> QA3	TRHGYLFGQKLAASMSPLLLHLSIVYSHLSSLN.....WAQLRLDSSPSIPLFLQLAQHP [87]
<i>A. nidulans</i>	PKKFAIFGSSPISQSAAPQLSTTPYLPSSASPIITPAWRRLRTPKMCSSSSAALTSAAPSVTI [1349]
<i>S. cerevisiae</i>	PKELFVVGKPIGHSRSPILHNTGYEILGLP.....HKFDKFEETESAQVLVKEKLLDGK [1358]
<i>E. coli</i> <i>aroE</i>	METVAVFGGPNPIAHSKSPFIHQQFAQQQLNIE.....HPYGRVLA PINDFINTLNAFSA [53]
<i>N. crassa</i> QA3	DFYGAISVTMPHKAIAIIPHLDHILTPEECRDVGA CNTLFLKKTDPATGRRLLYCGAANTDVI [147]
<i>A. nidulans</i>	RSSSTSCPFSTKLPKPPRSSLELTQSFPCRLARTLHH.....AYVGRNTDQWQGMIL [1400]
<i>S. cerevisiae</i>	NFGGAIVTIPTLKLDIMQYMDLELTDAAKVI GAVNTV IPLGN.....KKFKGDNTD [1413]
<i>E. coli</i> <i>aroE</i>	GKGANVTVPFKEEAFARADELTERAALACAVNTLMRLED.....GRLLGDDNTDGVGLLS [108]
<i>N. crassa</i> QA3	SFV.QNVSDPARVYESRPA LVIGGGGAARSAVYALHKKWL GATDIYLVNRRDKSEVDAVIAE [206]
<i>A. nidulans</i>	SLRKAGVYGPKKQDQESALVVGGGGTARAAIYALHN.MGYSPPIYIVGRTTPSSKLENMV.S [1459]
<i>S. cerevisiae</i>	AL...INN GVPEYVGHTAGLVIGAGGTSRAALYALHS.LGCKKIFIIINRTTSKLEPLI.E [1468]
<i>E. coli</i> <i>aroE</i>	DL...ERLSFIRPGLRI L LIGAGGASRGVLLPLLS.LDC.AVTIITNRTVSR AELLA.K [160]
<i>N. crassa</i> QA3	CTERGYGDRLLVHVASVEQAEGLEGPGGAIVACIPDFP PKTEKE.....MLVRRIVETFLM. [260]
<i>A. nidulans</i>	SFPSSYNIRIVEBPSSFFESVPHVAIGTIPADQPIDPTMR ETLCHMFERAQEADAEAVKAI [1318]
<i>S. cerevisiae</i>	SLPSEFNIIIGIESGKSI EEIKEHVGVAVSCVPAADKPLDDDEL L.....SKLEFLLVKG AH [1522]
<i>E. coli</i> <i>aroE</i>	LFAHTGSIQALSMDLELEGHEFDLIINATSSGISGDI PAIPSS.....L [203]
<i>N. crassa</i> QA3	KEEKGA ML E M C Y N P S P F T E L G A L A E H E G G W Q V I L G T E A L I Q Q G I E Q V C I R P E L T L Y C G G Q L N [321]
<i>A. nidulans</i>	EHA PR I L L E M A Y K P S V T P V M T I S O D K Y O W H V V P P G S Q M L V H Q G V A Q F E K V T G F K G P F K A I F [1577]
<i>S. cerevisiae</i>	AA F V P T L L E A Y K P S V T P V M T I S O D K Y O W H V V P P G S Q M L V H Q G V A Q F E K V T G F K G P F K A I F [1582]
<i>E. coli</i> <i>aroE</i>	I H P G I Y C Y D M F Y Q K G K T P F L A W C E Q R G S K R N A D G L G M L V A Q A H A F L L W H G V L P D V E P V I [263]
<i>A. nidulans</i>	ERSLNTGLGSR RRYMRVP GHVAPPSFN [1603]
<i>S. cerevisiae</i>	DAVTK E [1586]
<i>E. coli</i> <i>aroE</i>	KQLQEELSA [272]

Fig. 3. Comparison of the amino acid sequence of *E. coli* shikimate dehydrogenase with the amino acid sequence of other shikimate (quinic) dehydrogenase domains

The sources of the sequences are given in the text. Only perfect matches are boxed. The ADP-binding fingerprint lies between residue 121 and 151 in the sequence of the *aroE*-gene product.

Table 4. Location of the 'fingerprint' characteristic of an ADP-binding $\beta\alpha\beta$ -fold in *E. coli* shikimate dehydrogenase

Finger-print position	Position and residue identity in shikimate dehydrogenase sequence	Comments
1	Arg-121	Obeys fingerprint, running score 1*
2	Ile-122	Obeys fingerprint, running score 2
3	Leu-123	Position not specified in fingerprint
4	Leu-124	Obeys fingerprint, running score 3
5	Ile-125	Position not specified in fingerprint
6	Gly-126	Obeys fingerprint, running score 4
7	Ala-127	Position not specified in fingerprint
8	Gly-128	Obeys fingerprint, running score 5
9	Gly-129	Position not specified in fingerprint
10	Ala-130	Position not specified in fingerprint
11	Ser-131	Should be a glycine in the fingerprint
12	Arg-132	Position not specified in fingerprint
13	Gly-133	Position not specified in fingerprint
14	Val-134	Position not specified in fingerprint
15	Leu-135	Obeys fingerprint, running score 6
16	Leu-136	Position not specified in fingerprint
17	Pro-137	Position not specified in fingerprint
18	Leu-138	Obeys fingerprint, running score 7
19	Leu-139	Position not specified in fingerprint
20	Ser-140	Position not specified in fingerprint
21	Leu-141	Position not specified in fingerprint
22	Asp-142	Variable loop region in fingerprint
23	Cys-143	Variable loop region in fingerprint
24	Ala-144	Variable loop region in fingerprint
25	Val-145	Variable loop region in fingerprint
27†	Thr-146	Position not specified in fingerprint
28	Ile-147	Obeys fingerprint, running score 8
29	Thr-148	Position not specified in fingerprint
30	Asn-149	Should be a small and hydrophobic residue in fingerprint
31	Arg-150	Position not specified in fingerprint
32	Thr-151	Must not be acidic residue in NADP ⁺ -binding protein; obeys fingerprint, giving total score of 9

* The scoring and position numbering of this fingerprint follows that described by Wierenga *et al.* (1986).

† The variable loop in shikimate dehydrogenase appears to be only four residues long and not five as in the model numbering system employed by Wierenga *et al.* (1986); this accounts for the gap in the numbering.

inducible quinate pathway, one of the enzymes (quinate dehydrogenase) is homologous to a biosynthetic enzyme that can catalyse the same reaction whereas a second enzyme (3-dehydroquinase) has no discernible primary structure homology to a biosynthetic enzyme with the same catalytic activity. It therefore appears likely that the fungal catabolic 3-dehydroquinases may have arisen by convergent evolution (Da Silva *et al.*, 1986).

One interesting feature of the sequence of *E. coli* shikimate dehydrogenase is the occurrence between residues 121 and 151 of the amino acid 'fingerprint' characteristic of an ADP-binding $\beta\alpha\beta$ -fold (Wierenga *et al.*, 1986). Following the scoring system of Wierenga *et al.* (1986) *E. coli* shikimate dehydrogenase gives a score of nine (Table 4), which strongly suggests that it contains the recurrent fold that binds the ADP moiety of dinucleotides. This score takes account of the fact that

shikimate dehydrogenase is an NADP⁺-linked enzyme and would therefore not be expected to have an acidic residue at the end of the fingerprint as is found for NAD-linked enzymes (Wierenga *et al.*, 1986). This fingerprint region is particularly well conserved in all the four homologous shikimate (quinate) dehydrogenases (Fig. 3). One interesting feature is the occurrence in the *N. crassa* catabolic quinate dehydrogenase of an aspartic acid residue at the final position in the fingerprint; this catabolic enzyme is NAD⁺-specific (Barea & Giles, 1978), in contrast with the biosynthetic shikimate dehydrogenases, which are NADP⁺-specific and have threonine in this position (Fig. 3).

In many plants the third and fourth steps of the shikimate pathway are catalysed by a bifunctional enzyme having both 3-dehydroquinase and shikimate dehydrogenase activities (Boudet & Lecussan, 1974; Polley, 1978; Mousdale *et al.*, 1987). There are presently no sequence data for the bifunctional plant enzyme, but the fact that the *Pisum sativum* enzyme is a monomer of subunit *M_r* 59000 is consistent with the idea that the plant enzyme consists of two functional domains, one homologous to the monofunctional *E. coli* 3-dehydroquinase, which has an *M_r* of 26377 (Duncan *et al.*, 1986), and one homologous to *E. coli* shikimate dehydrogenase (*M_r* 29380) (Coggins *et al.*, 1985; the present study).

The three-dimensional structures of a number of NADP⁺-linked dehydrogenases are now known and they appear to be more diverse in structure than the known NAD⁺-linked dehydrogenases (Adams *et al.*, 1983). The known sequence of *E. coli* shikimate dehydrogenase, the availability of relatively large amounts of the enzyme and the fact that it is a monomer (Chaudhuri & Coggins, 1985) make this protein of interest to those wishing to investigate further the diversity of structures among NADP⁺-linked dehydrogenases.

We thank Professor J. E. Fothergill, Dr. L. A. Fothergill-Gilmore and Mr. B. Dunbar (University of Aberdeen Protein Sequencing Facility) for help with the amino acid sequencing, Mr. J. Jardine (University of Glasgow) for operating the amino acid analyser and Dr. K. Duncan and Dr. G. Millar for helpful discussions. We are grateful to the University of Glasgow for a studentship (I. A. A.) and the Science and Engineering Research Council for financial support.

REFERENCES

- Adams, M. J., Gover, S., Pickersgill, R. W. & Helliwell, J. R. (1983) *Biochem. Soc. Trans.* **11**, 429–435.
- Elton, N. K., Buxton, F., Patel, V., Giles, N. H. & Vapnek, D. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 1955–1959.
- Anton, I. A. & Coggins, J. R. (1984) *Biochem. Soc. Trans.* **12**, 275–276.
- Bachmann, B. J. (1983) *Microbiol. Rev.* **47**, 180–230.
- Barea, J. L. & Giles, N. H. (1978) *Biochim. Biophys. Acta* **524**, 1–14.
- 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.
- Boudet, A. M. & Lecussan, R. (1974) *Planta* **119**, 71–79.
- 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.
- Charles, I. G., Keyte, J. W., Brammar, W. J., Smith, M. & Hawkins, A. R. (1986) *Nucleic Acids Res.* **14**, 2201–2213.

- Chaudhuri, S. & Coggins, J. R. (1985) *Biochem. J.* **226**, 217–223
- 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
- Coggins, J. R., Duncan, K., Anton, I. A., Boocock, M. R., Chaudhuri, S., Lambert, J. M., Lewendon, A., Mousdale, D. M. & Smith, D. D. S. (1987) *Biochem. Soc. Trans.* **15**, 754–759
- Dagert, M. & Ehrlich, S. D. (1979) *Gene* **6**, 23–28
- Da Silva, A. J. F., Whittington, H., Clements, J., Roberts, C. & Hawkins, A. R. (1986) *Biochem. J.* **240**, 481–488
- Davis, B. J. (1964) *Ann. N.Y. Acad. Sci.* **121**, 404–427
- De Boer, H. A., Comstock, L. J. & Vasser, M. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 21–25
- DeFeyter, R. C., Davidson, B. E. & Pittard, J. (1986) *J. Bacteriol.* **165**, 233–239
- Duncan, K., Lewendon, A. & Coggins, J. R. (1984) *FEBS Lett.* **170**, 59–63
- Duncan, K., Chaudhuri, S., Campbell, M. S. & Coggins, J. R. (1986) *Biochem. J.* **238**, 475–483
- Duncan, K., Edwards, R. M. & Coggins, J. R. (1987) *Biochem. J.* **246**, 375–386
- Gaertner, F. H. & Cole, K. W. (1977) *Biochem. Biophys. Res. Commun.* **75**, 259–264
- Giles, N. H., Case, M. E., Partridge, C. W. H. & Ahmed, S. I. (1967) *Proc. Natl. Acad. Sci. U.S.A.* **58**, 1453–1460
- Gouy, M. & Gautier, C. (1982) *Nucleic Acids Res.* **10**, 7055–7074
- Grosjean, H. & Fiers, W. (1982) *Gene* **18**, 199–209
- Hardie, D. G. & Coggins, J. R. (1986) in *Multifunctional Proteins: Structure and Evolution* (Hardie, D. G. & Coggins, J. R., eds.), pp. 332–344, Elsevier, Amsterdam
- Haslam, E. (1974) *The Shikimate Pathway*, Butterworths, London
- Hawkins, A. R., Reinert, W. R. & Giles, N. H. (1982) *Biochem. J.* **203**, 769–773
- Hirs, C. H. W. (1967) *Methods Enzymol.* **11**, 197–199
- Holmes, D. S. & Quigley, M. (1981) *Anal. Biochem.* **114**, 193–197
- Jaskunas, S. R., Lindahl, L. & Nomura, M. (1975) *Proc. Natl. Acad. Sci. U.S.A.* **72**, 6–10
- Laemmli, U. K. (1970) *Nature (London)* **227**, 680–685
- Lambert, J. M., Boocock, M. R. & Coggins, J. R. (1985) *Biochem. J.* **226**, 817–829
- Larimer, F. W., Morse, C. C., Beck, A. K., Cole, K. W. & Gaertner, F. H. (1983) *Mol. Cell Biol.* **3**, 1609–1614
- 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: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor
- Messing, J. (1983) *Methods Enzymol.* **101**, 20–78
- Messing, J. & Vieira, J. (1982) *Gene* **19**, 269–276
- Millar, G. & Coggins, J. R. (1986) *FEBS Lett.* **200**, 11–17
- Millar, G., Lewendon, A., Hunter, M. & Coggins, J. R. (1986) *Biochem. J.* **237**, 427–437
- Miozzari, G. F. & Yanofsky, C. (1979) *Nature (London)* **277**, 486–489
- Mousdale, D. M., Campbell, M. S. & Coggins, J. R. (1987) *Phytochemistry* **26**, 2665–2670
- Pittard, J. & Wallace, B. J. (1966) *J. Bacteriol.* **91**, 1494–1508
- Polley, L. D. (1978) *Biochim. Biophys. Acta* **526**, 259–266
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
- Smith, M. A., Gerrie, L. M., Dunbar, B. & Fothergill, J. E. (1982) *Biochem. J.* **207**, 253–260
- Tribe, D. E., Camakaris, H. & Pittard, J. (1976) *J. Bacteriol.* **127**, 1085–1097
- Twigg, A. J. & Sherratt, D. (1980) *Nature (London)* **283**, 216–218
- Weiss, U. & Edwards, J. M. (1980) *The Biosynthesis of Aromatic Compounds*, John Wiley and Sons, New York
- Wierenga, K., Terpstra, P. & Hol, W. G. J. (1986) *J. Mol. Biol.* **187**, 101–107
- Zalkin, H. & Yanofsky, C. (1982) *J. Biol. Chem.* **257**, 1491–1500
- Zalkin, H., Paluh, J. L., Cleemput, M., Moye, W. S. & Yanofsky, C. (1984) *J. Biol. Chem.* **259**, 3985–3992

Received 30 March 1987/11 August 1987; accepted 1 September 1987