STUDIES ON THE *aroB*, *aroC* AND *aroL* GENES OF *ESCHERICHIA COLI*

Gary Millar

Submitted for the degree of Doctor of Philosophy in the Faculty of Medicine, University of Glasgow.

Biochemistry Department,
University of Glasgow

October, 1986

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Contents — Page No.

List of Figures

List of Tables

Acknowledgements

I would like to thank the following people who have contributed towards the work described in this thesis:

Professor R.M.S. Smellie for making available the facilities of the Biochemistry Department. Professor J.R. Coggins for his guidance, encouragement, boundless enthusiasm and financial support. Everyone on D-floor for their help, particularly Ken and Ian who showed the way. Dr Stevely who acted as my auditor. Mick, Ian and Mark, (ex) Searle for their help, oligos and free meals. The Medical Faculty of the University of Glasgow for their financial support. Mrs Peedle for her skill and patience in typing this thesis. Mr Ian Ramsden at Medical Illustration for skilled artwork and photography.

And Elizabeth, for too much.

Abbreviations

The abbreviations used in this thesis are as set out in 'Instructions to Authors', Biochemical Journal (1985) 225, 1-26, except the following:

amp	Ampicillin
bla	β-lactamase
βME	2-mercaptoethanol
bisacrylamide	N,N'-methylene bisacrylamide
ca.	approximately
ccc.	covalently closed circular
(k)Da	(kilo)daltons
DAHP	3-deoxy-D-arabinoheptulosonate-7-phosphate
DHQ	3-hydyroquinate
DHS	3-dehydroshikimate
d.s.	double stranded
dpm	disintegrations per minute
DTT	dithiothreitol
E0	DAHP synthase EC 4.1.2.15
E1	DHQ synthase EC 4.6.1.3
E2	3-dehydroquinase EC 4.2.1.10
E3	shikimate dehydrogenase EC1.1.1.25
E4	shikimate kinase EC 2.7.1.71
E5	EPSP synthase EC 2.5.1.19
E6	chorismate synthase EC 4.6.1.4
EPSP	5-enolpyruvylshikimate-3-phosphate
FPLC	fast protein liquid chromatography

IPTG	isopropyl-β-D-thiogalactoside
(k)bp	(kilo)base pairs
LA	L agar
LB	L broth
MM	Minimal medium
$\underline{M}_r$	Molecular mass (relative)
m.w. (mol. wt.)	Molecular weight
ORF	Open reading frame
phe	L-phenylalanine
RF	Replicative form
RNAse	Ribonuclease
SDS	Sodium dodecyl sulphate
SDS PAGE	Polyacrylamide gel electrophoresis in the presence of SDS
ss	Single stranded
SSC	Standard saline citrate
TEMED	N,N,N',N'-tetramethylethylene diamine
tet	Tetracycline
trp	L-tryptophan
tyr	L-tyrosine
u(units)	units of enzyme activity
X-gal	5-bromo-4-chloro-3-indolyl-β-galactoside

Summary

The *arom* pentafunctional enzyme found in fungi catalyses five consecutive reactions of the shikimate pathway. The same aromatic biosynthetic functions in bacteria are achieved by five monofunctional enzymes, products of individual unlinked genes. The evolutionary implications and questions raised by these two extremes of structural organisation can be addressed by comparison of the primary structures of the five monofunctional activities with that of the multifunctional enzyme. This thesis is concerned with contributing some of the required bacterial data.

The complete amino acid sequence of the *Escherichia coli* 3-dehydroquinate synthase (*aroB* gene product) has been determined by a combined nucleotide and direct amino acid sequencing strategy. The *aroB* gene was sub-cloned from the plasmid pJB14, its nucleotide sequence was determined and an overexpressing (inducible) strain constructed. The gene product DHQ synthase was purified from this strain and its N-terminal amino acid sequence determined. *E.coli* DHQ synthase is 362 amino acids long with a calculated M_r of 38,880. Analysis of the *aroB* nucleotide sequence and its 5' and 3' flanking regions has identified the *aroB* promoter and possible 3' terminator sequences.

The *aroL* gene encoding the *tyrR* regulated shikimate kinase II was cloned from *E.coli* K12. The *aroL* gene has been identified by nucleotide sequencing and direct N-terminal amino acid sequencing. Construction of overexpressing

strains has allowed purification (for the first time) of a monofunctional shikimate kinase. E.coli shikimate kinase II is monomeric with a calculated M_r of 18,937. The amino acid sequence contains a region with some homology to sequences found in other kinases and ATP-requiring enzymes. Transcript mapping has identified a possible operator sequence overlapping the aroL promoter which could constitute the tyrR repressor binding site.

The aroC gene encoding chorismate synthase has been cloned from E.coli. The construction of overexpressing strains has allowed purification for the first time of E.coli chorismate synthase. The aroC gene has been identified by nucleotide sequencing and confirmed by N-terminal amino acid sequencing of the purified protein. E.coli chorismate synthase is 357 amino acids long with a calculated M_r of 38,183. Analysis of the 3' flanking sequences has identified possible terminator elements.

The bacterial sequences for the aroB and aroL gene products have been compared with the S.cerevisiae and A.nidulans arom amino acid sequences. Discrete non-overlapping regions of the fungal polypeptide are homologous with both E.coli DHQ synthase (aroB), shikimate kinase (aroL) and the remaining aroA, aroD and aroE gene products. This evidence supports the hypothesis which forms the basis of this study: that the fungal arom multifunctional enzyme is a mosaic of five independently folded 'domains'.

CHAPTER 1

INTRODUCTION

1.1 General Introduction

In micro-organisms and plants the biosynthesis of aromatic compounds proceeds via the shikimate pathway (Haslam, 1974; Weiss & Edwards, 1980). Higher organisms lack this biosynthetic capability and require a dietary intake of at least two of the three aromatic amino acids tyrosine, phenylalanine and tryptophan.

The first seven steps on this pathway convert the intermediates of carbohydrate metabolism, erythrose-4-phosphate and phosphoenolpyruvate, via shikimate to chorismate (Figure 1.1). Chorismate is a common precursor of not only the three aromatic amino acids but also of a number of metabolically and structurally important aromatic compounds.

Although of intrinsic interest, the shikimate or common (aromatic) pathway is best characterised by the diversity of structural forms which its component activities adopt in different species. It is this striking diversity which, almost paradoxically, is the unifying theme and common thread running through this thesis.

Perhaps the two most polarised examples of differential structural organisation in shikimate pathway activities can be observed in comparing Escherichia coli and Neurospora crassa. In the enteric bacterium (E.coli) all seven common pathway enzyme activities are separable (Berlyn & Giles, 1969) and their genes are scattered around the bacterial chromosome (Pittard & Wallace, 1966). In marked contrast, the fungal (N.crassa) aromatic biosynthetic pathway includes a component

pentafunctional polypeptide - the *arom* multifunctional enzyme - catalysing five of the seven common pathway activities, and is the product of a single gene (Lumsden & Coggins, 1977; Gaertner & Cole, 1977; Giles *et al.*, 1967a).

The shikimate pathway therefore provides an excellent model system through which a number of pertinent questions regarding the evolution of biosynthetic pathways/multifunctional proteins can be experimentally addressed. The occurrence of multifunctional proteins in only some biosynthetic routes of a given organism; the occurrence of multifunctional proteins on a biosynthetic route in some species but not others; the evolutionary origin of biosynthetic routes in general and multifunctional proteins in particular, are all examples of as yet unexplained phenomena. It is the intention of this study, as part of a larger group effort, to contribute towards a greater understanding of the existence, within the aromatic biosynthetic pathway, of such a variety of structural forms and to assess the significance of any common structural motifs.

1.2 The shikimate pathway: an overview

1.2.1 Pathway intermediates

In all organisms that do not rely on an external supply of aromatic amino acids the common pathway of chorismate biosynthesis is the major synthetic route to homocyclic aromatic compounds. The shikimate pathway takes its name from the central intermediate shikimic acid whose carbon skeleton is derived from the condensation of molar equivalents

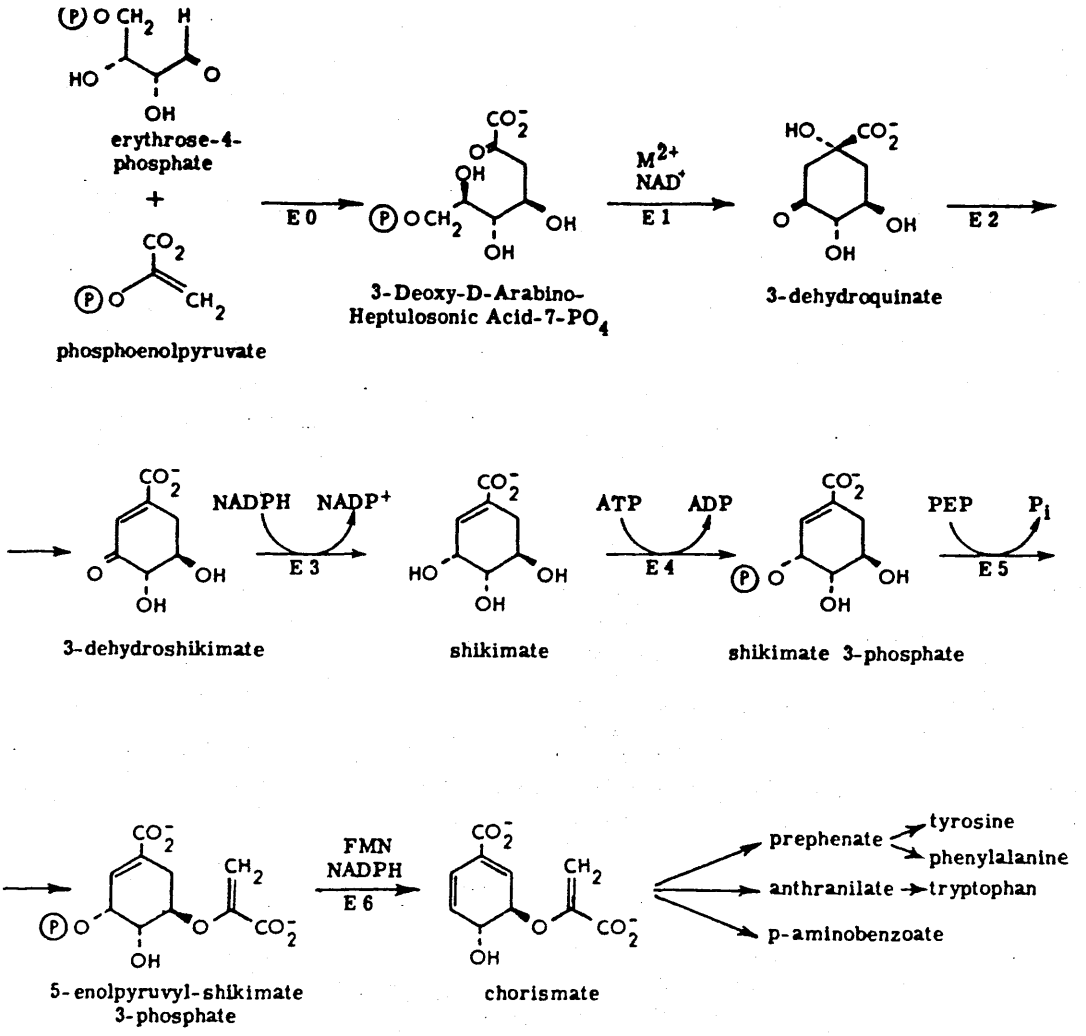


Figure 1.1: The shikimate pathway.

Activities:

- E0 3-deoxy-D-arabinoheptulosonate 7-phosphate synthase.
- E1 3-dehydroquinate synthase
- E2 3-dehydroquinase
- E3 shikimate dehydrogenase
- E4 shikimate kinase
- E5 5-enolpyruvylshikimate 3-phosphate synthase
- E6 chorismate synthase

of erythrose-4-phosphate and phosphoenolpyruvate. The subsequent addition of a second molecule of phosphoenolpyruvate during the penultimate pathway step completes the carbon skeleton of chorismic acid. Most of the chemical intermediates on the shikimate pathway were isolated and identified more than two decades ago (Levin & Sprinson, 1964; Davis, 1965; Gibson & Pittard, 1968). It is only in the past 10 years that the enzymology of the pathway has reached a similar level of understanding.

1.2.2 Utilisation of chorismate

Chorismate is perhaps one of the most versatile chemical intermediates in primary metabolism (Figure 1.2). Many pathways diverge from chorismate to yield not only the three aromatic amino acids but also ubiquinone, plastoquinone, folic acid and vitamin K. In plants phenylalanine is a precursor of the lignins which provide much of the tensile strength of woody tissues. The "terminal pathways" from chorismate to the three aromatic amino acids exhibit an even greater species-specific diversity in structural organisation and are discussed below (Section 1.6).

1.2.3 Organisation of this introduction

This introduction considers the organisation of the shikimate pathway in different species. A detailed account of the enzymology of the seven pre-chorismic acid activities of the common pathway (whether mono- or multifunctional) is presented together with a discussion of the genetic organisation, where known, of their respective genes.

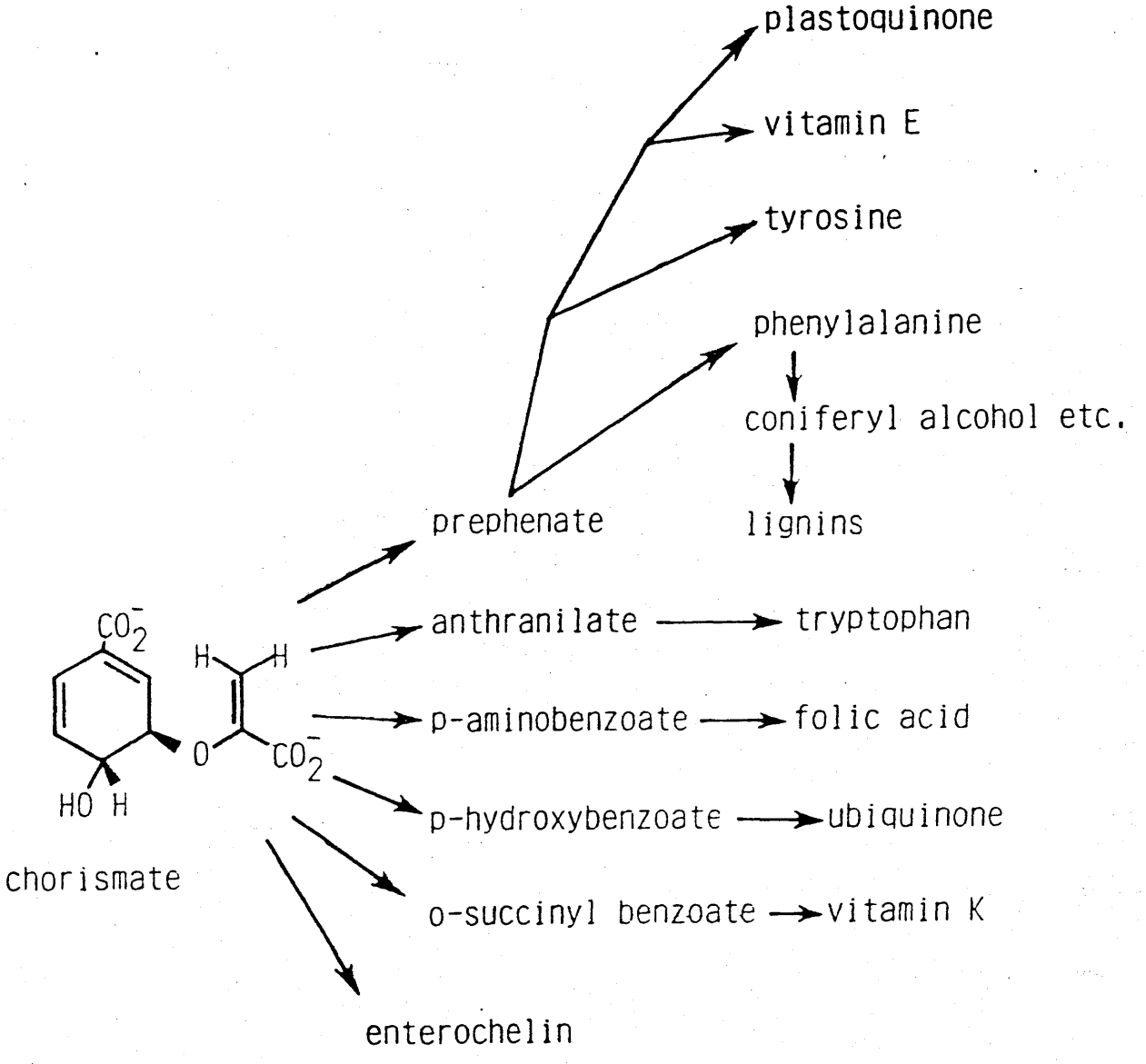


Figure 1.2: Utilisation of chorismate as a precursor for aromatic biosynthesis.

Some aspects of multifunctional proteins, and how our thinking of their origin and function can be influenced by examination of the shikimate pathway, is also presented.

1.3 The shikimate pathway: bacterial organisation

1.3.1 Genetic analysis of the aro genes

With the exception of shikimate kinase, see below, mutants auxotrophic for each of the pathway steps (encoded by aroA-aroH, see Figure 1.3) have been characterised for E.coli and their genes mapped (Pittard & Wallace, 1966; Wallace & Pittard, 1967). As can be seen in Figure 1.3 the common pathway enzymes in E.coli are encoded by widely scattered genes (Bachmann, 1983). A similar pattern of chromosomal distribution is observed in Salmonella typhimurium (Gollub et al., 1967; Sanderson & Roth, 1983). Mutants auxotrophic for shikimate kinase have also been isolated in Bacillus subtilis (Nakatsukasa & Nester, 1972), and in this gram positive bacterium there is some genomic clustering of the aro genes (Henner & Hoch, 1980).

The inability to isolate shikimate kinase mutants of E.coli and S.typhimurium by selecting aromatic auxotrophs has been rationalised in terms of multiple (two or more) shikimate kinase isoenzymes. The identification by Berlyn & Giles (1969; see below) of two peaks of shikimate kinase activity after density gradient centrifugation of extracts of E.coli and S.typhimurium supports this hypothesis. Ely & Pittard (1979) were able to map and mutationally defined the aroL gene, encoding E.coli shikimate kinase II,

to minute 9 of the chromosome. The structural gene for the other isoenzyme in E.coli remains unmapped. The aroL gene has been shown to be transcriptionally regulated (Ely & Pittard; see Section 1.3.3). There are also three isoenzymes for DAHP synthase in E.coli encoded by the unlinked aroF, aroG and aroH genes. Each of these genes is regulated by repression and their encoded protein subject to end-product inhibition (see Section 1.3.3).

1.3.2 Separability of shikimate pathway activities

Berlyn & Giles (1969) examined several species of bacteria, including E.coli, S.typhimurium, B.subtilis and A.aerogenes, for physical aggregation of the five central common pathway activities. Sucrose density gradient centrifugation of cell extracts taken from these bacteria displayed five independently separating common pathway activities. For E.coli and S.typhimurium two peaks of shikimate kinase activity were resolved.

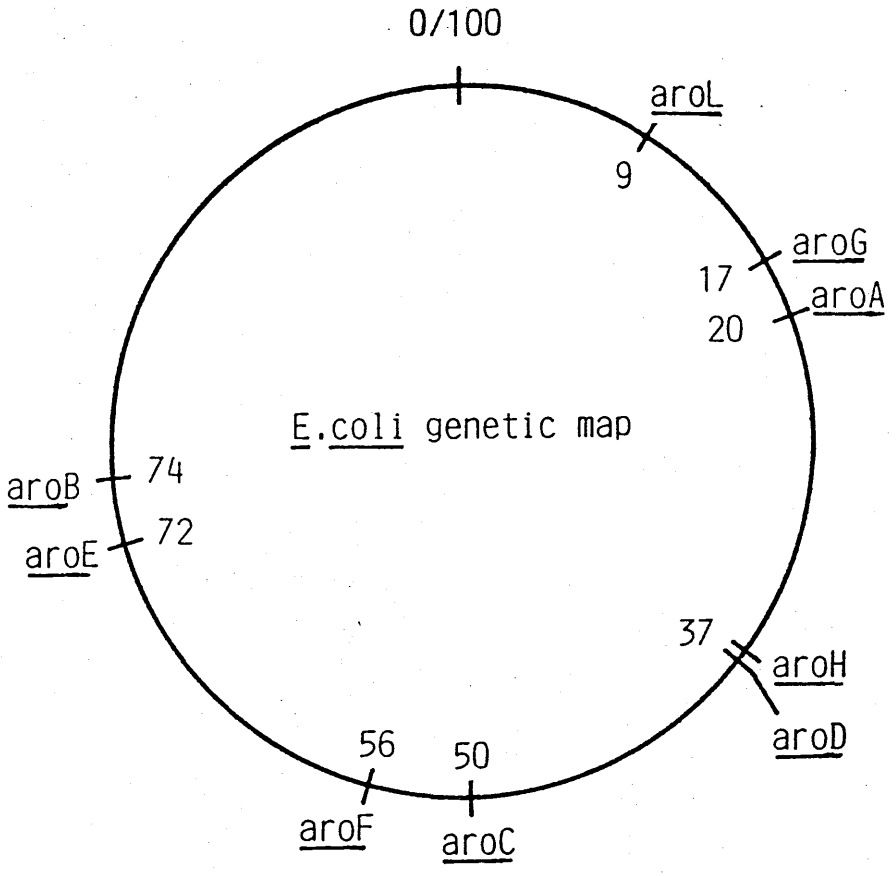
1.3.3 Regulation of expression of bacterial aro genes

The regulation of common and terminal pathway gene expression has been most extensively studied in E.coli. Control of aromatic biosynthesis is exerted on the first and fifth steps of the common pathway. Expression of aroF or aroG is repressed by the tyrR protein complexed with tyrosine or phenylalanine plus tryptophan, respectively, as corepressor(s) (Doy & Brown, 1965; Brown, 1968; Wallace & Pittard, 1969; Brown & Somerville, 1971; Im et al., 1971;

Camakaris & Pittard, 1973). The aroH (tryptophan-repressible) is regulated by the trpR repressor (Brown, 1968). Expression of aroL, encoding shikimate kinase II, is repressed by the tyrR protein complexed with tyrosine or tryptophan as corepressor (Ely & Pittard, 1979). Tribe et al. (1976) have shown that the remaining common pathway genes are expressed constitutively.

A number of genes on the terminal pathways which extend from chorismate to the aromatic end products (Figure 1.2) have been shown to be transcriptionally regulated. The bifunctional enzymes chorismate mutase/prephenate dehydratase and chorismate mutase/prephenate dehydrogenase are encoded by the genes pheA and tyrA respectively. The pheA gene has been sequenced and shown to contain a phenylalanine-rich leader sequence (Zurawski et al., 1978). This sequence has been identified within a potential attenuator structure implying that expression of pheA is intimately coupled to cellular levels of $tRNA^{phe}$. The pheA gene is also regulated by repression through the pheR repressor (Gowrishankar & Pittard, 1982).

In addition to pheA, the tyr operon is also located near 57 minutes on the E.coli chromosome but is expressed from the opposite strand. The tyr operon contains the aroF and tyrA genes and its expression is repressed by the tyrR protein complexed with tyrosine (and possibly phenylalanine) (Brown & Somerville, 1971; Im et al., 1971; Mattern & Pittard, 1971; Camakaris & Pittard, 1973). Camakaris et al. (1983) have noted that in a repressor deficient background (tyr^-) expression of aroF is elevated by decreasing the levels of



Pathway step	gene	activity
1 (E0)	aroF	DAHP synthase (tyr-repressible)
1 (E0)	aroG	DAHP stnthase (phe-repressible)
1 (E0)	aroH	DAHP synthase (trp-repressible)
2 (E1)	aroB	DHQ synthase
3 (E2)	aroD	3-dehydroquinase
4 (E3)	aroE	shikimate dehydrogenase
5 (E4)	aroL	shikimate kinase
6 (E5)	aroA	EPSP synthase
7 (E6)	aroC	chorismate synthase

Figure 1.3: Chromosomal location of the E.coli aro genes.

$tRNA^{trp}$, suggesting an attenuation-like control system. The *aroF* gene has been sequenced (Hudson & Davidson, 1984) and potential operator sequences mediating repressor binding (*tyrR*), mutationally defined (Garner & Herrmann, 1985).

The final reactions in the three-step conversions of chorismate to either tyrosine or phenylalanine are catalysed by a single enzyme, the aromatic aminotransferase encoded by the *tyrB* gene (Figure 1.4). Expression of *tyrB* may be repressed by tyrosine (Silbert *et al.*, 1962) mediated through *tyrR* binding (Wallace & Pittard, 1969). The *E.coli tyrB* gene has been sequenced (Fotheringham *et al.*, 1986), but comparisons with the upstream sequences of *aroG* and *aroF* (Davies & Davidson, 1982; Hudson & Davidson, 1984) genes reveals no striking homologies with the mutationally-defined *tyrR* repressor binding site (Garner & Herrmann, 1985).

1.3.4 Regulation of the shikimate pathway at the enzyme level

As with regulation of gene expression, control of aromatic biosynthesis in *E.coli* at the enzyme level is concentrated on the first committed reaction which is catalysed by three isoenzymes of DAHP synthase. Nearly all of the enzymes of the terminal pathways in *E.coli* are also subject to some form of regulation but discussion here is confined to the pre-chorismate pathway activities.

In *E.coli* the three isoenzymes of DAHP synthase are each inhibited by either tyrosine, tryptophan or phenylalanine (Gibson & Pittard, 1968). None of the remaining *E.coli*

shikimate pathway enzymes are inhibited by the aromatic amino acids or indeed by chorismate or prephenate. The differences in cellular activities of the three DAHP synthase isoenzymes is dramatic. The phenylalanine-repressible activity specifies more than 70% of the total activity while the remainder is composed mainly of the tyrosine-sensitive isoenzyme with only a minor contribution from the third isoenzyme (McCandliss *et al.*, 1978). The existence of three differentially-regulated isoenzymes enables the bacterium to exert multivalent control on aromatic biosynthesis whilst allowing sufficient residual activity to provide for essential aromatic co-factor synthesis.

In *B. subtilis* 168 a single DAHP synthase occurs as a trifunctional complex with chorismate mutase and shikimate kinase (Nakatsukasa & Nester, 1972). The kinase is only active when complexed as the aggregate form (Huang *et al.*, 1975). In addition, the shikimate kinase moiety is subject to specific inhibition by chorismate, prephenate and possibly ADP. It has been postulated that this inhibition by chorismate and/or prephenate is mediated through allosteric binding of these metabolites at the chorismate mutase catalytic site (Huang *et al.*, 1975). In the same way the DAHP synthase activity is inhibited by chorismate and prephenate. The chorismate mutase active site therefore exhibits a functional duality where binding of effectors can result in the inhibition of other activities in the complex.

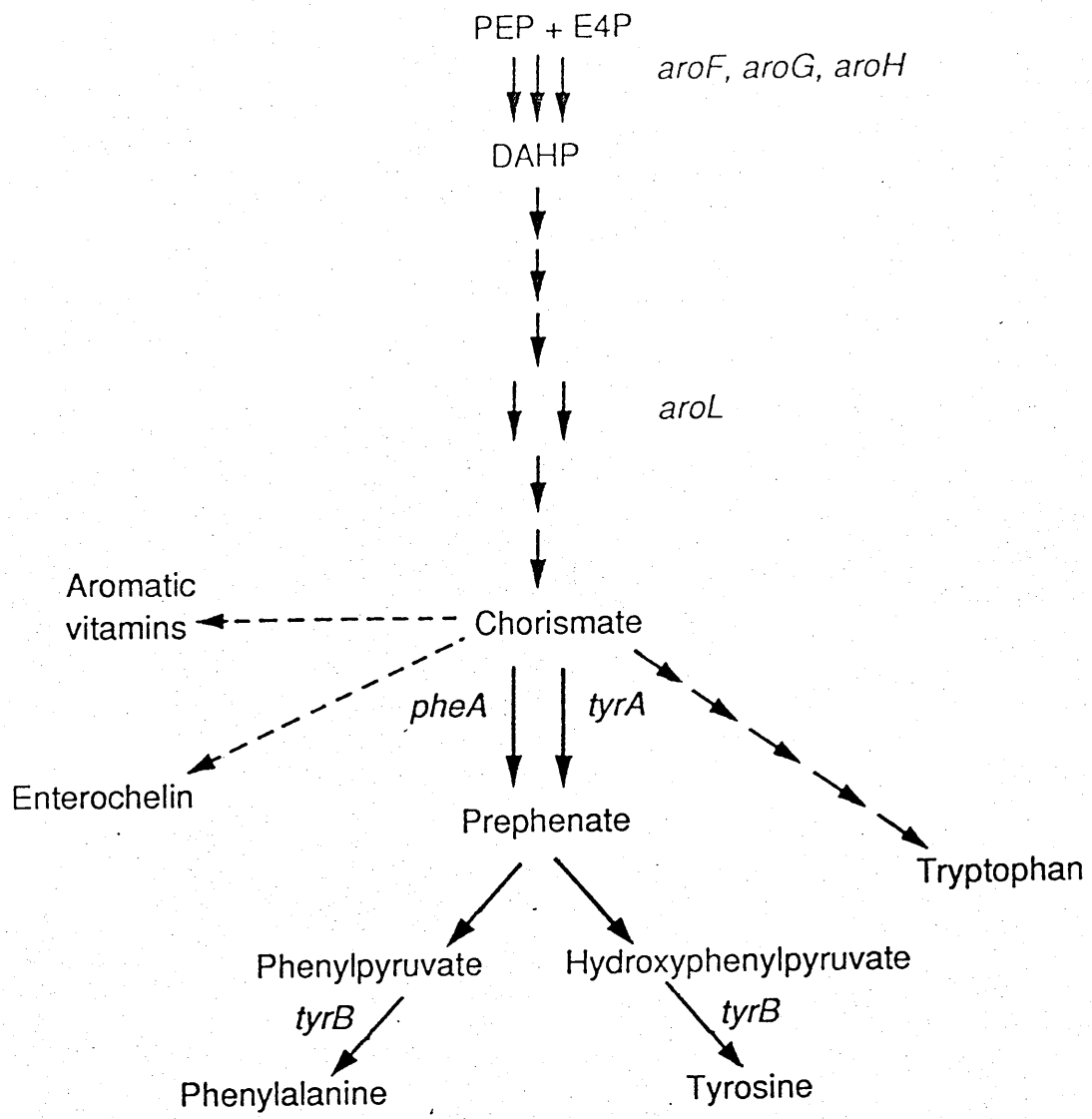


Figure 1.4: Outline of the pathways of aromatic biosynthesis in E.coli. Shown are the major regulated genes. Abbreviations are as discussed previously. Isoenzymes are indicated by multiple arrows.

1.4 The shikimate pathway: fungal organisation

1.4.1 Separability of fungal common pathway activities

The general separability of the bacterial shikimate pathway enzymes (see Section 1.3.2) contrasts with the situation in fungi where five of the seven common pathway activities exhibit a higher order of structural organisation. Giles *et al.* (1967a) first demonstrated for *Neurospora crassa* that the five consecutive shikimate pathway activities, (see Figure 1.1 for numerology), DHQ synthase (E1) to EPSP synthase (E5), cosedimented during sucrose density gradient centrifugation. This work was extended by Ahmed & Giles (1969) to six other species of fungi, including *Aspergillus nidulans*, and showed that the same five common pathway activities (E1 — E5) were physically associated. A partial aggregation of three of the five activities intimately associated in *N. crassa* was demonstrated for the yeast *Saccharomyces cerevisiae* (De Leeuw, 1967). Subsequent genetic data (Section 1.4.2) indicated that the two excluded activities were missing as a result of the extraction procedure.

The *Neurospora* enzymes were found to co-purify (Burgoyne *et al.*, 1969; Gaertner, 1972) and estimates of M_r 230,000 and M_r 293,000 made for the intact complex. Several discrete polypeptides were observed by both groups and interpreted on the basis of functional sub-assemblies of a multienzyme complex. It was not until a rigorous anti-protease strategy was employed throughout the enzyme preparation that an intact

pentafunctional polypeptide of M_r 165,000 was observed (Lumsden & Coggins, 1977; Gaertner & Cole, 1977). The Neurospora arom multifunctional complex was shown by cross-linking to be a dimer of sub-unit M_r 165,000 (Lumsden & Coggins, 1977) and by peptide mapping to be a homodimer (Lumsden & Coggins, 1978).

1.4.2 Genetic analysis of fungal arom genes

(1) Early genetic analysis of the Neurospora locus defining several of the shikimate pathway activities is best re-interpreted with the knowledge that steps E1 to E5 (see Figure 1.1) occur as a multifunctional protein in this fungus (Lumsden & Coggins, 1977; Gaertner & Cole, 1977).

Gross & Fein (1960) first reported the possible existence of gene cluster in Neurospora when they detected arom mutations in a number of distinct complementation groups mapping in a cluster on linkage group II. Giles et al. (1967a) reisolated some 500 new arom mutants for N.crassa and classified these on the basis of heterokaryon complementation analysis and direct enzymatic assay. Two distinct groups of arom mutants were identified by this procedure. In the first class five complementation groups, each associated with a defect in one of the five central shikimate pathway activities, were found to map in the same tight cluster identified by Gross & Fein (1960). Genetic mapping of the mutants indicated that each mapped in a distinct, localised part of the arom region. A second class of polyaromatic auxotrophs exhibited biochemical pleiotropy (loss of more

than one activity) and displayed distinct polarity in their complementation responses. Some of the polarity mutants were completely noncomplementing mutants and were found to map asymmetrically at one end of the region among arom-2 (DHQ synthase) mutants.

This latter type of polarity mutant provided the first evidence that the cluster constituted a supragenic functional unit. Some of these arom-2 and other arom polarity mutants were found to be suppressible by a Neurospora nonsense suppressor and were interpreted in terms of being nonsense mutations (Giles et al., 1967a). The sedimentation characteristics of a number of these suppressible polarity mutants revealed smaller protein "aggregates" and were wrongly interpreted as aggregation-deficient mutations rather than simply as containing truncated multifunctional polypeptides.

(2) A similar approach was used to analyse the arom-1 locus defining several shikimate pathway activities in the yeast S. cerevisiae (De Leeuw, 1967). Again mutations in several distinct arom complementation groups mapped tightly in a cluster. Similarly, pleiotropic mutants exhibiting polarity in complementation challenges suggested a unique linear order of activities. However completely non-complementing suppressible polarity mutants were found to map in the E2 (dehydrogenase) sub-region of the "gene cluster" indicating a different order of activities from that observed for N.crassa (Giles et al., 1967a). Subsequent DNA sequencing

and further mutational analysis (Duncan et al., 1986a; Section 1.4.3) has shown this observation to be incorrect.

(3) Strauss (1979) isolated a series of aro mutations in the fission yeast Schizosaccharomyces pombe and through precise genetic analysis established that the clustering of 'genes' for aromatic biosynthetic activities in the aro3 region of S.pombe resembled that of N.crassa and S. cerevisiae. The aro3 region also consisted of five complementation groups and from the observed polarity effects, Strauss (1979) predicted that these five complementation groups were expressed (transcribed) continuously.

(4) A parallel study of A.nidulans demonstrated that the organisation of the arom locus observed in N.crassa, S. cerevisiae and S.pombe, and defined by biochemical and genetic approaches, was identical in this fungus too (Roberts, 1969).

1.4.3 Cloning of the fungal arom genes

By selecting activities capable of complementing E.coli aro mutations, a gene responsible for biosynthesis of aromatic amino acids in S.pombe (aro3) has been cloned (Nakanishi & Yamamoto, 1984). The entire aro3 locus has been isolated as a set of plasmids containing overlapping genomic fragments which show differential ability to complement E.coli aroD and aroE mutant strains. In addition, transformation of mutants defective in the five distinct sub-regions (A to E) of the S.pombe aro3 locus, as defined by Strauss (1979), has allowed the assignation of these sub-regions to the individual

plasmids. The entire *aro3* locus has been shown to be transcribed as a 4.5 kbp mRNA transcript and the order of activities (A — E sub-regions) has established that *S.pombe* and *N.crassa* are identical in terms of their enzymatic coding activities.

Catcheside *et al.* (1985) have constructed a hybrid phage carrying a 4 kbp HindIII fragment of *N.crassa* genomic DNA capable of complementing an *E.coli aroD* mutant. Discrimination between the catabolic dehydroquinase gene (*qa-2*) and the biosynthetic dehydroquinase gene (*arom-9*, part of the complex *ARO* locus) was based on the HindIII restriction pattern and thermostability of the encoded dehydroquinase. Two classes of phage capable of complementing *aroD* mutants of *E.coli* and *N.crassa* (*arom-9*) were isolated; one apparently contained a reisolate of the 2.0 kbp HindIII fragment carrying the *qa-2* gene cloned previously by Vapnek *et al.* (1977). The other isolate (4 kbp HindIII clone) transformed *Neurospora arom-9* mutants to prototrophy, did not cross-hybridise with the other isolate (2.9 kbp Hind III) and encoded a thermolabile dehydroquinase activity. It was concluded that it contained at least the *arom-9* region of the pentafunctional *ARO* locus.

The biosynthetic dehydroquinase function of the pentafunctional *AROM* locus of *A.nidulans* was cloned as a 1.9 kbp HindIII genomic fragment by a similar direct complementation procedure (Kinghorn & Hawkins, 1982). The isolated HindIII fragment was later used as a DNA probe to locate larger *AROM*-containing recombinant phages in a hybrid *A.nidulans* DNA λ gene bank (Charles *et al.*, 1985). Two recombinant

phage overlapping the 5' and 3' ends of the 1.9 kbp HindIII region were isolated and the DNA sequence of 6.5 kbp of A.nidulans genomic DNA determined (Charles et al., 1986). A single open-reading frame of 4,812 bp was identified in the same orientation and phase as determined for the region encoding the dehydroquinase function (Charles et al., 1985). The inferred molecular weight of the AROM-polypeptide encoded by this sequence is 175 kDa. Complementation of only E.coli aroD mutations has been reported for this clone but evidence presented later strongly suggests that this protein sequence is that of the A.nidulans arom multifunctional protein.

The ARO1 gene of S.cerevisiae has been isolated by complementation in Saccharomyces and E.coli from a yeast DNA library of BamHI fragments. The ARO1 gene was originally isolated as a 17.2 kbp BamHI fragment (Larimer et al., 1983) capable of integration at the aro1 locus which implies physical identity. A Sau3A sub-clone (6.2 kbp) further located the ARO1 region and was able to complement both missense and nonsense alleles of ARO1 in yeast cells and four separate aro defects in E.coli (B, D, E, A) indicating functional identity (Larimer et al., 1983). Yeast aro1 mutants transformed with the ARO1 episome overexpress (10 fold) the normal levels of the five ARO1 enzymes and contain elevated levels of the ARO1 protein (Larimer et al., 1983). Deletion and insertional mutations followed by transformation of the available E.coli aro mutants (aroA, aroB, aroD and aroE) has further defined sub-regions and the order of activities within this clone (Duncan et al., 1986a). The

complete nucleotide sequence of the 6.2 kbp region of S.cerevisiae cloned by Larimer et al. (1983) has now been determined. A single open-reading frame encoding a 1588 amino acid residue polypeptide has been identified (M_r 174,555) (Duncan et al., 1986a). Functional regions within the polypeptide chain, as defined by deletion/insertion analysis and comparative studies with the E.coli aro genes are discussed in Chapter Six.

It is interesting to note that all four fungal arom genes which have been cloned, A.nidulans, N.crassa, S.cerevisiae and S.pombe appear to lack introns by virtue of their functional expression in E.coli. This evidence is largely circumstantial for N.crassa and S.pombe but supported by sequence analysis for the other two fungal genes. No introns have been identified in the S.cerevisiae or A.nidulans sequences but the possibility of small in-phase introns lacking both termination codons and consensus boundary signals cannot be excluded.

1.5 The shikimate pathway in plants

1.5.1 Enzyme separability

The organization of polyaromatic biosynthetic enzymes in a variety of photosynthetic organisms has been examined (Berlyn et al., 1970). Sucrose density centrifugation of cell extracts was used to estimate the molecular weights and determine possible physical aggregation of the enzymes of the shikimate pathway in Anabaena variabilis, Chlamydomonas reinhardi, Euglena gracilis, Nicotiana tabacum and

Physcomitrella patens. In A. variabilis, as in other prokaryotes (Berlyn & Giles, 1969), the five enzymes catalysing steps 2 to 6 of the shikimate pathway are separable. Separation of these enzymes was also observed for the eukaryotes C.reinhardi, P.patens and N.tabacum except that the dehydroquinase (E2) and shikimate dehydrogenase (E3) activities were clearly associated (Berlyn et al., 1970). In Euglena gracilis all five activities cosediment as a large aggregate (approximately 120 kDa) but dissociation of this aggregate into smaller (c.a. 60 kDa) components was also observed. In an extension of this work, Patel & Giles (1979) purified the Euglena complex 2000-fold although they did not demonstrate (by SDS PAGE) that the preparation was homogeneous. Although not definitively proven, the occurrence of a large aggregate and the ability to isolate active fragments of such a complex strengthens the analogy with the N.crassa arom multifunctional protein.

1.5.2 E2/E3: multifunctional protein or multienzyme complex?

Multifunctionality of plant shikimate pathway enzymes is not confined to the Euglena arom complex. As discussed above (Section 1.5.1), several photosynthetic organisms contain dehydroquinase (E2) and shikimate dehydrogenase (E3) activities which cosediment in sucrose gradients (Berlyn et al., 1970). Boudet & Lécussan (1974) demonstrated that under a variety of separation techniques E2 and E3 activities co-purified from a range of higher plants including Zea mays L., mung bean (Phaseolus mungo) and pea (Pisum sativum L.).

A bifunctional protein with E2 and E3 enzymatic activities was first demonstrated in P.patens (Polley, 1978). A single polypeptide with M_r 48,000 (as judged by SDS PAGE) was isolated from this lower plant with E2 and E3 purifying in constant activity ratio. Mousdale & Coggins (1984) have demonstrated that for pea seedlings the shikimate pathway activities are plastidic and the E2 and E3 activities occur on a bifunctional protein of M_r 59,000 (Mousdale et al., 1986).

1.5.3 EPSP synthase

Much of the recent interest in plant shikimate pathway enzymology is as a result of the observation that the herbicide Glyphosate (N- phosphonomethyl-glycine) interferes with aromatic amino acid biosynthesis by inhibiting the enzyme EPSP synthase (E5) (Steinrücken & Amrhein, 1980). Glyphosate is a potent broad-spectrum herbicide which inhibits the growth of both weed and crop species. Several groups have tried to introduce herbicide tolerance into plants and two main approaches have been used.

Comai et al. (1983, 1985) have cloned a glyphosate-resistant aroA allele from S.typhimurium (aroA encodes EPSP synthase) and have established that a single amino acid substitution in EPSP synthase is sufficient to confer glyphosate resistance. Subsequent expression of this mutant bacterial allele in Nicotiana tabacum again confers a tolerant phenotype.

Shah *et al.* (1986) have engineered herbicide tolerance, also in transgenic plants, by introducing into *Petunia hybrida* a chimeric *Petunia* EPSP synthase gene under the control of the cauliflower mosaic virus 35S promoter. The high levels of EPSP synthase expression obtained under these conditions allows growth in the presence of glyphosate. The *Petunia* EPSP synthase has a 72 amino acid leader peptide which directs the 444 amino residue mature enzyme to the chloroplast (Shah *et al.*, 1986). It is therefore surprising that a non-targetted glyphosate-tolerant EPSP synthase (c.f. Comai *et al.*, 1983, 1985) is able to confer herbicide resistance *in vivo*. Speculation regarding an additional cytosolic shikimate pathway or diffusion of pathway intermediates (and inhibitors) across the chloroplast membrane(s) remains unsubstantiated.

1.6 The terminal pathways

1.6.1 The enzymes and genes of tryptophan biosynthesis

The biosynthesis of tryptophan from chorismate proceeds in all organisms as shown in Figure 1.5. Between different genera there is a striking diversity with respect to the molecular nature of the individual *trp* biosynthetic enzymes (reviewed by Crawford, 1975). Some chemically distinct reactions in tryptophan biosynthesis are catalysed by multifunctional proteins whether the steps are consecutive or not. Similarly reactions catalysed in one organism by a single gene product is catalysed in another organism by a multimeric protein encoded by two genes.

Figure 1.5 (facing)

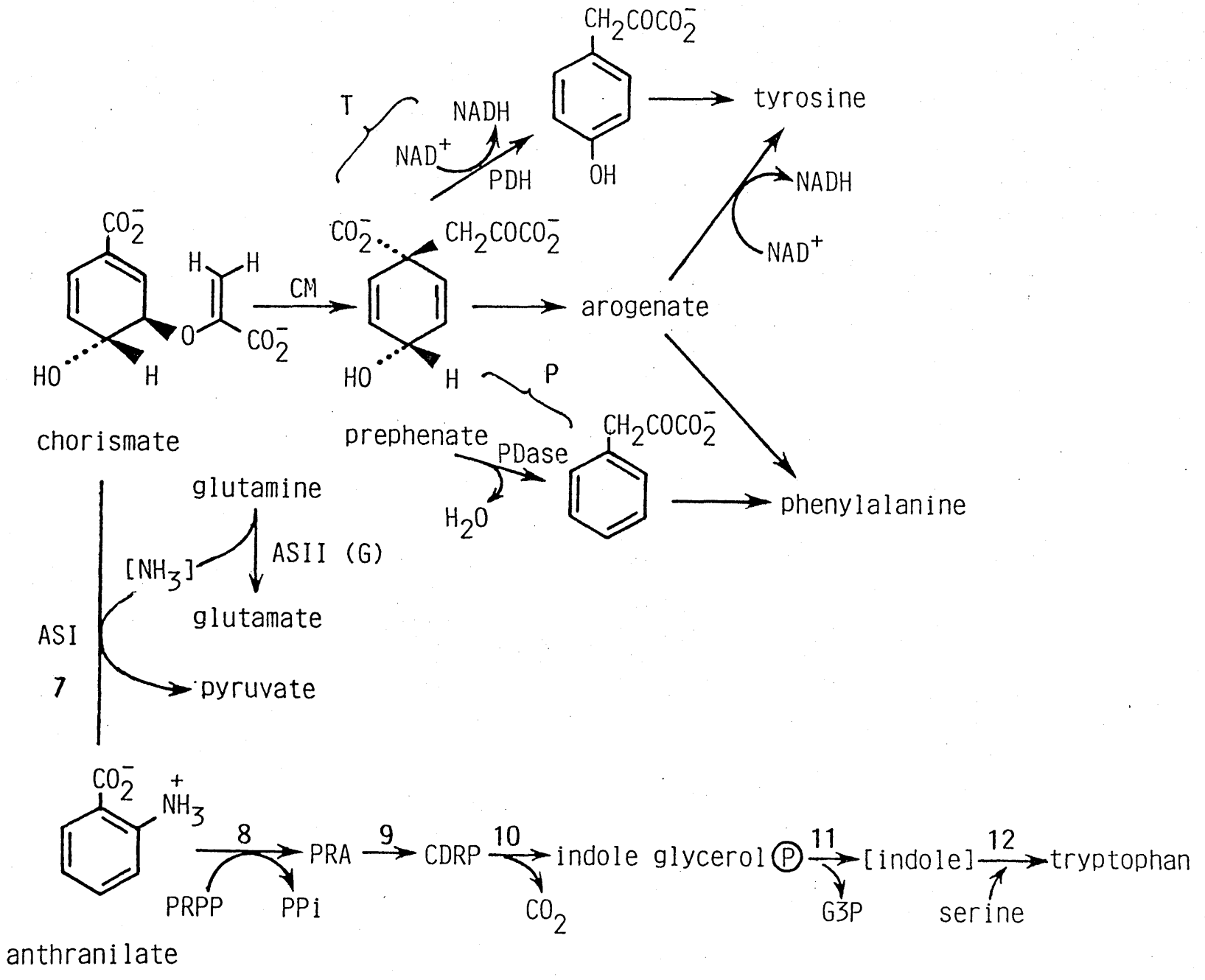
The biosynthetic pathways leading from chorismate to the aromatic amino acids phenylalanine, tyrosine and tryptophan. The numbering of activities is as appears in Figure 1.6.

Enzymes

ASI (7)	Anthranilate synthase, catalytic subunit.
ASII (G)	Anthranilate synthase, glutaminase subunit.
8	Anthranilate phosphoribosyl transferase.
9	phosphoribosyl anthranilate isomerase
10	indole glycerol phosphate synthase
11	tryptophan synthase, step 1
12	tryptophan synthase, step 2
T,P	The bifunctional T- and P-proteins discussed in Section 1.6.2.
CM	chorismate mutase
PDH	prephenate dehydrogenase
PDase	prephenate dehydratase

Intermediates

PRA	phosphoribosyl anthranilate
CDRP	(O-carboxyphenylamino)-1-deoxyribulose-5-phosphate
PRPP	phosphoribosyl pyrophosate
G3P	glyceraldehyde-3-phosphate.


T
NADH
NAD+
PDH
CH2COCO2−
OH
tyrosine
NADH
NAD+
CO2−
H
H
CM
O
CO2−
HO
H
CO2−
CH2COCO2−
HO
H
arogenate
chorismate
P
prephenate
PDase
H2O
CH2COCO2−
phenylalanine
glutamine
[NH3]
ASII (G)
glutamate
ASI
7
pyruvate
CO2−
+
NH3
8
PRPP
PPi
PRA
9
CDRP
10
CO2
indole glycerol Ⓟ
11
G3P
[indole]
12
serine
tryptophan
anthranilate

The two initial reactions specific to the tryptophan pathway are catalysed by the enzymes anthranilate synthase (AS) and anthranilate-5-phosphoribosylpyrophosphate phosphoribosyl transferase (PRT) (reactions 7 and 8 respectively in Figure 1.5). In many bacterial genera AS is an enzyme complex containing two non-identical polypeptides ASI and ASII. Component I alone can convert chorismate to anthranilate using NH_3 but not L-glutamine as the amino donor. Component II contributes glutaminase activity to the complex allowing the use of L-glutamine as amino donor. In E.coli and S.typhimurium the ASII component is fused to the PRT activity and for E.coli this bifunctional polypeptide is the product of the trpD gene. The two enzyme activities of ASII of E.coli and S.typhimurium are associated with different segments of their respective polypeptide chains. It has been suggested, through comparison with the monofunctional glutaminase and PRT polypeptides of Serratia marcescens, that the glutaminase and PRT bifunctional gene of E.coli and S.typhimurium arose by fusion of two S.marcescens-like genes (Miozzari & Yanofsky, 1979).

Whereas in S.marcescens, B.subtilis and P.putida the glutaminase of AS resides within a separate protein with no other function, the N.crassa glutaminase activity is part of a multifunctional protein. Unlike E.coli, the N.crassa glutaminase is fused to the enzymes phosphoribosylanthranilate isomerase and indoleglycerolphosphate synthase which convert phosphoribosyl anthranilate to indoleglycerolphosphate. The S.cerevisiae ASII component is fused only to indoleglycerol phosphate synthase (Zalkin, 1980). Also in yeast the third

and fourth steps of the biosynthetic pathway are catalysed by separate proteins coded for by unlinked genes. In E.coli and many other prokaryotes both of these reactions are catalysed by a single bifunctional protein encoded by the trpC gene (Crawford, 1975).

The converse is true for tryptophan synthase (Section 1.9.4), where the fungal enzyme consists of a fusion of α and β chain elements (Matchett & De Moss, 1975) which are separate polypeptides in E.coli (summarised in Figure 1.6).

Crawford & Stauffer (1980) have reviewed the regulation of tryptophan biosynthesis. For E.coli two mechanisms exist to control transcription of the structural genes of the trp operon. Firstly, binding of the trpR repressor complexed with tryptophan to the operator site regulates transcription of the operon (Platt, 1978). Secondly, transcription and translation of the trp mRNA are intimately coupled allowing an attenuation mechanism to control trp operon expression by responding to cellular $tRNA^{trp}$ levels (Yanofsky, 1981).

1.6.2 Tyrosine and Phenylalanine biosynthesis

A full account of tyrosine (Camakaris & Pittard, 1983) and Phenylalanine (Garner & Herrmann, 1993) biosynthesis is given in Herrmann & Somerville (1983).

Chorismate is converted to phenylalanine in three steps via either phenylpyruvate or arogenate (Figure 1.5) depending upon the organism (Jensen & Pierson, 1975). The enteric bacteria E.coli, S.typhimurium and Klebsiella pneumoniea use

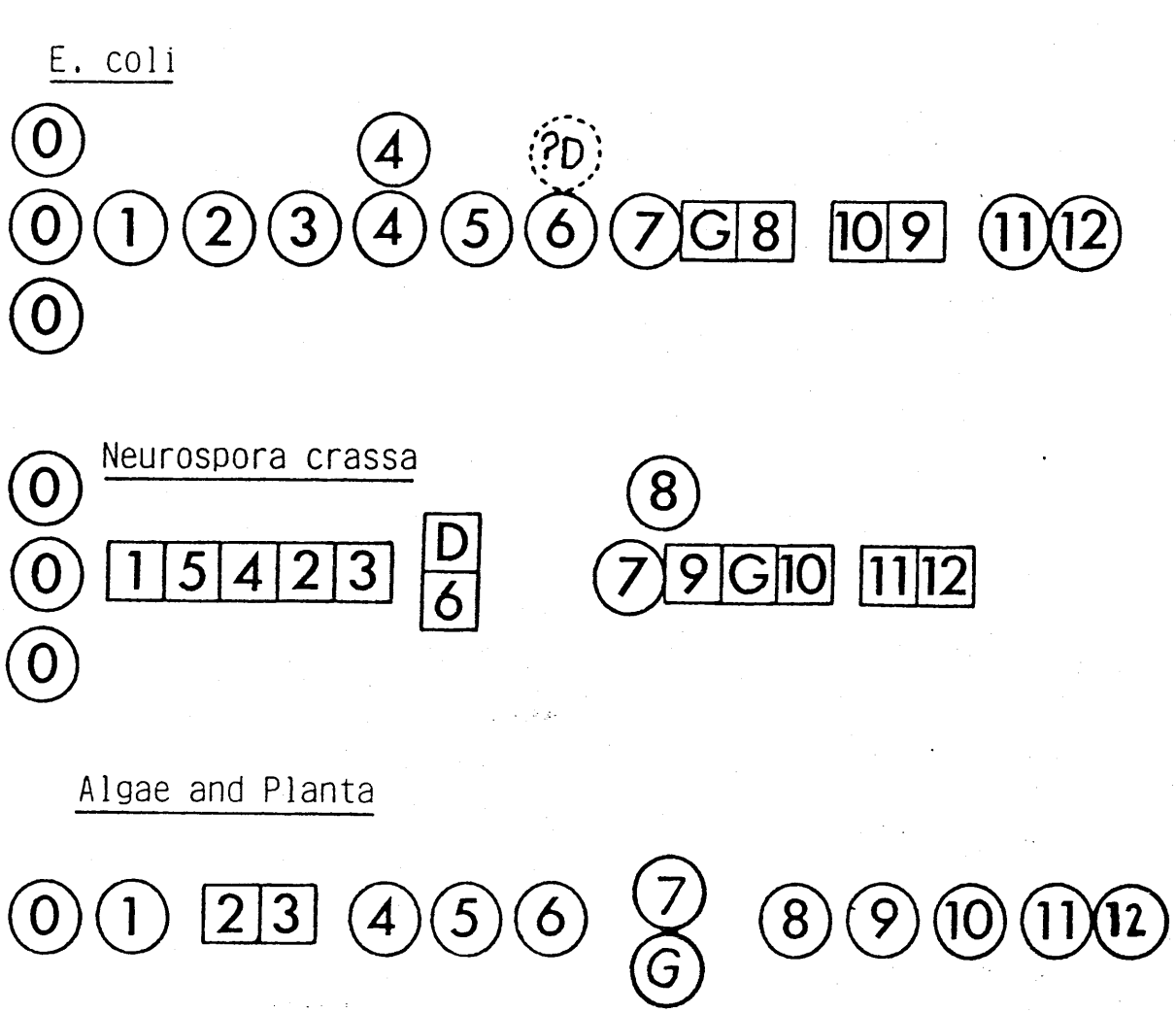


Figure 1.6: A summary of the different structural organisations of the chorismate and tryptophan aromatic biosynthetic pathways found in different species. The numbering of activities is as in Figures 1.1 and 1.5.

Multifunctional enzymes are indicated by joined rectangles. Monofunctional enzymes are represented by circles (isoenzymes are multiple circles). Joined circles indicate multienzyme complexes.

'D' - diaphorase activity of chorismate synthase
'G' - Glutaminase activity of anthranilate synthase.

the phenylpyruvate pathway and have a bifunctional chorismate mutase/prephenate dehydratase (P-protein).

For tyrosine biosynthesis these organisms utilise the hydroxyphenylpyruvate pathway which also has a bifunctional enzyme, in this case chorismate mutase/prephenate dehydrogenase (T-protein).

Neurospora and Saccharomyces use the phenylpyruvate route to phenylalanine but the P-protein activities reside on separate polypeptide chains. In B.subtilis a single chorismate mutase provides prephenate for both prephenate dehydratase and prephenate dehydrogenase. Ahmad & Jensen (1986) have traced the evolutionary history of the two bifunctional proteins (P- and T-protein) that participate in aromatic biosynthesis in some Gram-negative bacteria. They have proposed that although the two bifunctional enzymes accomplish exactly analogous roles in the short phenylalanine and tyrosine branches of aromatic amino acid biosynthesis, the P-protein is of ancient origin whilst the T-protein is of recent origin.

1.7 Catabolic quinic acid pathway in N.crassa

Both N.crassa (Giles et al., 1967b) and A.nidulans (Ahmed & Giles, 1969) have two enzymatic activities that can metabolize dehydroquinate: (1) a heat stable inducible catabolic dehydroquinase and (2) a heat labile constitutive biosynthetic dehydroquinase which is part of the arom multi-functional polypeptide. In N.crassa, which has been characterised better both genetically and biochemically, and

to which discussion will be confined, the catabolic dehydroquinase is the product of the qa-2 gene. This structural gene is a member of the qa cluster comprising five structural and two regulatory genes (Huiet, 1984).

If N.crassa is grown on quinic acid, enzymes necessary for quinate catabolism are induced. The induced enzymes, quinate (shikimate) dehydrogenase (qa-3), catabolic dehydroquinase (qa-2) and 3-dehydroshikimate dehydratase (qa-4) (Figure 1.7), are responsible for the first three steps of quinate and shikimate metabolism. The catabolic dehydroquinase has been purified to homogeneity and is a dodecamer of M_r 20,000 (Chaudhuri & Coggins, 1982).

The three qa structural genes are regulated at the level of transcription (Patel et al., 1981). Huiet (1984) has shown that two regulatory genes, qa-1S and qa-1F encode a repressor and activator protein respectively, both of which control qa gene expression.

Kinetic parameters measured for both the synthetic and degradative dehydroquinase activities (Chaudhuri & Coggins, 1982) of Neurospora has shown that the biosynthetic activity has a much lower K_m for DHQ (5 μM as opposed to 170 μM for the catabolic activity). This lower K_m may ensure that DHQ levels are kept low to avoid induction of the qa cluster which is only required when N.crassa is grown on quinic acid.

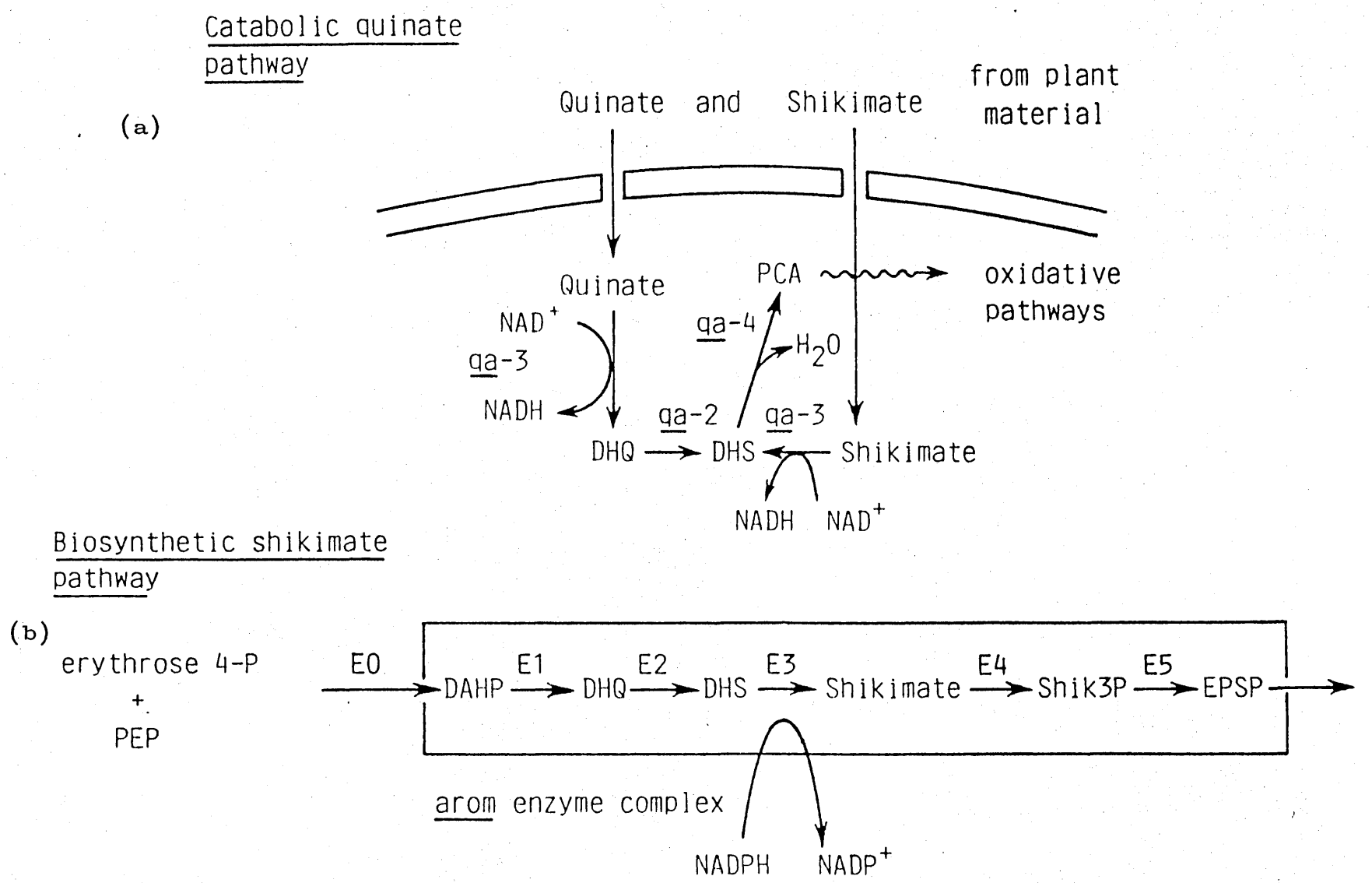


Figure 1.7: Interaction between the (a) catabolic quinate pathway and (b) biosynthetic shikimate pathway in *N. crassa*. Activities are: (a) as described in Section 1.7; (b) as in Figure 1.1. PCA is protocatechuate.

1.8 The enzymes of the shikimate pathway

1.8.1 DAHP synthase

The first committed step of the shikimate pathway is the condensation of phosphoenolpyruvate and erythrose-4-phosphate to form the seven carbon compound 3-Deoxy-D-*arabino*-heptulosonate 7-phosphate (DAHP). This reaction is catalysed by the enzyme DAHP synthase (EO, Figure 1.1).

Wild-type *E.coli* produces three DAHP synthase isoenzymes: phenylalanine-sensitive DAHP synthase (Phe), tyrosine-sensitive DAHP synthase (Tyr), and tryptophan-sensitive DAHP synthase (Trp) (Day & Brown, 1965), which are encoded by the unlinked *aroG*, *aroF* and *aroH* genes respectively. As described earlier (Sections 1.3.3 and 1.3.4), the DAHP-synthases of *E.coli* are feedback inhibited by the aromatic end products and their genes transcriptionally controlled by the *tyrR* and *trpR* repressors. The three isoenzymes have all been purified to homogeneity (Herrmann, 1983) and the nucleotide sequences of their respective genes determined (Shultz *et al.*, 1984; Davies & Davidson, 1982; Zurawski *et al.*, 1981), although only the first 36 N-terminal residues and 117 C-terminal residues of DAHP synthase (Trp) can be deduced from the data of Zurawski *et al.* (1981).

The lack of a complete *aroH* sequence prevents a full analysis of the relatedness of the three isoenzymes. The amino acid sequences predicted from the *aroF* and *aroG* nucleotide sequences are highly homologous. These proteins contain 356 and 360 amino acids residues respectively and

have been shown to exhibit 63% homology, while their genes are 57% homologous (Hudson & Davidson, 1984). Comparison with the incomplete aroH sequence still exhibits large regions of identity between all three isoenzymes (Hudson & Davidson, 1984). It would appear likely that these three genes have evolved by gene duplication. However the stretches of non-similarity between all three isoenzymes has prompted other workers to suggest that evolutionary reshuffling of functionally defined sequences encoding domains specifying, for example, E-4-P or PEP binding, could represent the true origin of the three E.coli isoenzyme genes (Herrmann, 1983). Both hypotheses require subsequent conjecture on transposition of these genes to explain their scattered chromosomal locations.

N.crassa, like E.coli, has three DAHP isoenzymes that are each inhibited by one of the three aromatic amino acids. The tryptophan-sensitive DAHP synthase has been purified to homogeneity (Nimmo & Coggins, 1981).

Bacillus subtilis contains only a single DAHP synthase. In B.subtilis 168 DAHP synthase and chorismate mutase form a bifunctional enzyme which, in concert with a shikimate kinase subunit, forms a highly regulated trifunctional complex (Huang et al., 1975, Section 1.3.4).

Brevibacterium flavum DAHP synthase purified to homogeneity is a tetramer of subunit molecular weight 55 kDa (Sugimoto & Shiio, 1980). As in B.subtilis, DAHP synthase and chorismate mutase form a bifunctional enzyme, but unlike B.subtilis a second distinct polypeptide is required for activity.

1.8.2 Dehydroquinate synthase

3-Dehydroquinic acid (DHQ) is the first of six consecutive alicyclic intermediates in the common pathway. The reaction converting DAHP to DHQ is catalysed by dehydroquinate synthase (DHQ synthase, E1; Figure 1.1).

Early work on the isolation of E.coli B DHQ synthase by Maitra & Sprinson (1978) reported a homogenous preparation of M_r 57,000, although SDS PAGE confirmation was not presented as evidence. The subsequent cloning of the E.coli K12 aroB gene encoding DHQ synthase (Duncan & Coggins, 1983) and the purification of E1 from an overproducing strain (Frost et al., 1984) suggested that the actual subunit molecular weight was in the region of 36-40 kDa. The complete amino acid sequence of the E.coli DHQ synthase has been determined (this study; Millar & Coggins, 1986) and this established the M_r as 38,880.

The B.subtilis enzyme has a molecular weight of 17,000 and forms a trifunctional complex with chorismate synthase and a flavin reductase (Hasan & Nester, 1978a). Association with chorismate synthase is required for enzyme activity of the DHQ synthase and like E.coli DHQ synthase (Maitra & Sprinson, 1978; Srinivasan et al., 1963) the Bacillus activity requires a divalent transition metal cation and catalytic amounts of NAD^+ (Hasan & Nester, 1978b).

The E1 activity of N.crassa is part of the pentafunctional arom complex and is absolutely dependent upon Zn^{2+} for activity (Lambert et al., 1985). The E1 activity of Phaseolus mungo has been purified to homogeneity and it too has a requirement for NAD^+ and an essential cation (Yamamoto, 1980). The native

molecular weight of the mung bean enzyme was estimated to be 67 kDa by gel filtration chromatography and the subunit M_r to be 43,000 by SDS PAGE (Yamamoto, 1980).

1.8.3 3-dehydroquinase

Dehydroquinate dehydratase (3-dehydroquinase; E2, Figure 1.1) catalyses a stereospecific dehydration of DHQ to yield dehydroshikimate (DHS). The monofunctional E.coli E2 activity has been purified from wild-type E.coli (Duncan et al., 1986b) and from an overproducing strain. The E.coli aroD gene encoding 3-dehydroquinase has been cloned (Kinghorn et al., 1981) and its DNA sequence determined (Duncan et al., 1986b). Purification of the cloned enzyme from an over-producing strain and determination of the N-terminal amino acid sequence, has confirmed the DNA sequencing which predicted a molecular weight of 28 kDa.

The E.coli E2 activity is unusual in one respect with regards to the other four "arom" activities (E1, E3-E5) of the E.coli shikimate pathway. E.coli dehydroquinase is a dimeric enzyme whilst the remaining four activities, corresponding to the arom multifunctional enzyme, are monomeric (Frost et al., 1984; Anton & Coggins, 1986; Millar et al., 1986b; Duncan et al., 1984b).

The B.subtilis 168 aroC gene encoding dehydroquinase has been cloned (Warburg et al., 1984) by complementation in both E.coli and B.subtilis mutants. The dehydroquinase structural gene for this Gram positive bacterium is tightly linked to a gene of the serine biosynthetic pathway (the ser-22 locus,

mutants of which specify an unknown enzyme deficiency). This gene does not complement a serC mutant of E.coli (or serA or serB mutants either) but its close association with an aromatic gene is similar to the serC-aroA operon (described in detail in Section 1.8.6). B.subtilis however lacks enterochelin so the observed tight linkage between serine and aromatic pathways, in this case, is mysterious (see Section 1.8.6).

Mousdale et al. (1986) have shown that the E2 and E3 (shikimate dehydrogenase, Section 1.8.4) activities of pea seedlings are a single bifunctional enzyme of M_r 59,000. This is consistent with the observations of Polley (1978) who purified to homogeneity a similar bifunctional enzyme from Physcomitrella patens, and of Berlyn et al. (1970) who studied the separability of shikimate pathway enzymes in a number of photosynthetic organisms (see Section 1.5.2).

The 3-dehydroquinase activities of N.crassa has been discussed previously (see Section 1.7). The constitutive biosynthetic E2 activity of N.crassa shares one feature in common with the E.coli enzyme. Both E2 activities can be inhibited by 'substrate trapping', specifically treatment with sodium borohydride in the presence of DHQ results in the formation of a stable covalent intermediate (Chaudhuri et al., 1986). A Schiff base is formed between the ϵ-NH_2 group of the active site lysine residue of the E2 of both E.coli and N.crassa. This observation has been exploited for E.coli and N.crassa (S. Chaudhuri & Coggins, unpublished work) to isolate radioactively labelled active site peptides

from both sources. This is discussed further in Chapter Six in identifying the E2 domain of the yeast arom multifunctional enzyme.

1.8.4 Shikimate dehydrogenase

Shikimate dehydrogenase (E3, Figure 1.1) catalyses the reduction of DHS to shikimic acid. Shikimate dehydrogenase is encoded by the E.coli aroE gene and is NADP-specific (Dansett & Azerad, 1974). The wild-type E.coli E3 activity has been purified to homogeneity and is monomeric with a subunit M_r 31,000 (Chaudhuri & Coggins, 1985). The E.coli aroE gene has been cloned (Anton & Coggins, 1983), its nucleotide sequence determined, and substantiated by N-terminal sequencing of an overproduced enzyme (Anton & Coggins, 1986). The E.coli E3 activity is unusual in being the only identified monomeric, biosynthetic dehydrogenase to date.

1.8.5 Shikimate kinase

Shikimate kinase (E4, Figure 1.1) phosphorylates shikimate to give shikimate-3-phosphate. As detailed earlier (Section 1.3.1) E.coli (Berlyn & Giles, 1969) and S.typhmurium (Morell & Sprinson, 1968) each contain two isoenzymes, shikimate kinase I and II. In B.subtilis a single shikimate kinase of M_r 10,000 occurs as a multienzyme complex with a bifunctional DAHP synthase (Section 1.8.1) - chorismate mutase, the kinase component has been purified to homogeneity and shown to be active only in the complex (Huang et al., 1975). Shikimate kinase partial purifications have been reported from Phaseolus mungo (Koshiba, 1979) and sorghum (Bowen & Kosugo, 1979).

The E.coli shikimate kinase isoenzymes are differentially expressed. Kinase I is expressed constitutively while kinase II is transcriptionally regulated by the tyrR repressor (Ely & Pittard, 1979; Section 1.3.3). Both the S.typhimurium and E.coli isoenzymes can be separated by ion-exchange chromatography (Morell & Sprinson, 1967; Ely & Pittard, 1979) but both E.coli forms coelute in gel filtration chromatography with an apparent M_r of 20,000 (Ely & Pittard, 1979). The E.coli aroL gene encoding the tyrR-regulated shikimate kinase II has been cloned and its DNA sequence determined (this study, Millar et al., 1986b; Defeyter & Pittard, 1986).

N.crassa has a single shikimate kinase as part of the pentafunctional arom complex. This activity is very readily inactivated by limited proteolysis (Smith & Coggins, 1983).

1.8.6 EPSP synthase

Shikimate 3-phosphate is converted to 5-enolpyruvyl-shikimate-3-phosphate (EPSP) by the enzyme EPSP synthase (E5, Figure 1.1). During the course of the reaction, the enolpyruvyl moiety of phosphoenolpyruvate is transferred unchanged to the aromatic nucleus and inorganic phosphate is released. The monofunctional E.coli EPSP synthase was first purified by Lewendon & Coggins (1983) and was shown to be monomeric with a M_r 46,000 (Duncan et al., 1984b). The observation by Amrhein et al. (1980) that the EPSP synthase of Aerobacter aerogenes was inhibited by the herbicide Glyphosate (Section 1.5.3) has resulted in a stimulation in work on this enzyme, particularly on the plant form (Section 1.5.3).

The E.coli aroA gene encoding EPSP synthase has been cloned (Duncan & Coggins, 1983) and its nucleotide sequence determined (Duncan et al., 1984b). Similarly the EPSP synthase gene of S.typhimurium has been sequenced (Stalker et al., 1985). Comparison of both aroA alleles shows 21% divergence in the nucleotide sequence and 11% difference in the amino acid sequence. In both E.coli and S.typhimurium the aroA gene is the promoter-distal gene of a two gene operon (Duncan & Coggins, 1986; Hoiseth & Stocker, 1985). The other gene in the operon is serC encoding the enzyme of serine biosynthesis phosphoserine-aminotransferase. This is the first example in E.coli of a mixed function operon encoding genes of two distinct biosynthetic pathways. It would appear that chorismate and serine (the two pathway end-products) are required in equimolar amounts for the biosynthesis of enterochelin (Figure 1.2). This iron chelator is essential for bacterial viability due to its iron-scavaging properties and co-ordinate production of pathway intermediates for its own biosynthesis is ensured by expressing some of the genes required on the same operon.

The Petunia hybrida EPSP synthase gene has recently been cloned and sequenced (Shah et al., 1986). The gene contains eight introns and is 48% homologous with the E.coli enzyme at the amino acid level.

EPSP synthase has been purified from Pisum satirum (Mousdale & Coggins, 1984) where like P.hybrida and E.coli it is monomeric with an M_r 46-49,000.

1.8.7 Chorismate synthase

Perhaps the most enigmatic enzyme of the shikimate pathway is chorismate synthase (E6, Figure 1.1) which catalyses the elimination of orthophosphate from EPSP and in doing so introduces the second double bond of the aromatic ring.

Chorismate synthases studied so far all require reduced flavin nucleotides for activity. In *N.crassa* the chorismate synthase and flavin reductase (diaphorase) activities both reside on the same ca. 55 kDa polypeptide chain (Welch *et al.*, 1974). In *B.subtilis* the chorismate synthase subunit (M_r 24,000) is part of a trifunctional complex also containing DHQ synthase and NADPH-dependent flavin reductase activities (Hasan & Nester, 1978a,b). The *E.coli* chorismate synthase is less specific in its reduced flavin requirements accepting $FADH_2$ or NADH but is sensitive to molecular oxygen and up until now has only been assayed under a N_2 atmosphere (Morell *et al.*, 1967).

1.9 Evolution of the *arom* multifunctional enzyme

1.9.1 Multifunctional proteins - occurrence

Multifunctional enzymes are a class of macromolecules found in all types of organisms, from bacteria, and plants to higher eukaryotes but appear to be particularly prevalent in the amino acid biosynthetic pathways of prokaryotes and fungi (Kirschner & Bisswanger, 1976; Schmincke-Ott & Bisswanger, 1980). All known multifunctional proteins, with the exception of artificially created fusion-proteins (Kania & Müller-Hill,

(1980), catalyse either consecutive reactions in a biosynthetic pathway or at least reactions from a single metabolic pathway. To highlight the diversity of cellular processes in which multifunctional proteins are involved, several examples will be discussed in this section and later in Chapter Six. The intention of this section is to introduce the arom multifunctional enzyme as a model system for the purpose of general discussion on functional and evolutionary aspects of multifunctional proteins.

1.9.2 The arom multifunctional enzyme

As described in Section 1.4, the arom complex of fungi (and possibly Euglena gracilis) is a pentafunctional enzyme catalysing five consecutive reactions of the common pathway of aromatic amino acid biosynthesis (Giles et al., 1967a; Lumsden & Coggins, 1977, 1978). The most intriguing question regarding the fungal arom complex, and one which can be extrapolated to other multifunctional proteins, is the degree of relatedness between the corresponding monofunctional activities (where they exist) and the multifunctional protein. Put another way, are multifunctional proteins a mosaic of independently folded domains, derived from monofunctional enzymes, with each domain catalysing a separate reaction?

Giles et al. (1967a) demonstrated the use of fine genetic mapping of the N.crassa complex arom locus to partly answer this question. Pleiotropic arom mutations which affect more than one catalytic activity and which can be suppressed by a nonsense suppressor were all found to map in the "N-terminal"

end of the arom transcription unit. As detailed in Section 1.4, these completely noncomplementing mutants and the pattern of complementation of other mutations which mapped in discrete, non-overlapping segments of the arom locus and which affected only one activity, established genetic evidence for a mosaic of independent domains along the polypeptide chain.

Biochemical evidence for the domain structure of the N.crassa arom complex is comprised of chemical modification and limited proteolysis of the purified enzyme. The arom dehydroquinase activity contains an active site lysine residue which can be irreversibly blocked by substrate trapping (Smith & Coggins, 1983; Section 1.8.3). In this way the arom dehydroquinase activity can be abolished without affecting the remaining four activities. Similarily the arom shikimate dehydrogenase activity can be specifically inactivated by chemical modification with formaldehyde and sodium borohydride (Smith, 1980; Lumsden et al., 1986), thus providing evidence that these active sites of the arom multifunctional polypeptide are spatially distinct.

If such a hypothesis is correct then the isolation of active proteolytic fragments of the intact complex retaining some activity should be a realistic proposal. By treating purified intact arom with proteases, Smith & Coggins (1983) demonstrated that the dehydroquinase and shikimate dehydrogenase (E2 and E3) activities could be isolated as a bifunctional fragment of M_r 68,000 (intact arom has a M_r 165,000). Further proteolysis demonstrated that this fragment could

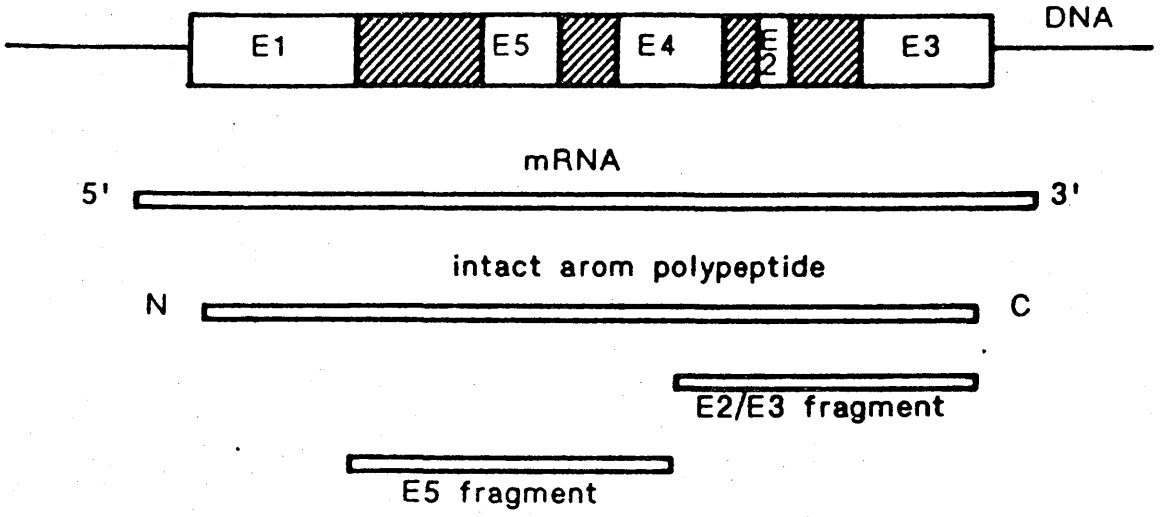


Figure 1.8: The N. crassa arom multifunctional enzyme. Enzyme activities are numbered as in Figure 1.1. The upper boxed figure diagramatically summarises the data of Giles et al., 1967a and Rines et al. 1969, which identified five non-overlapping regions within the arom locus, mutations in which affected one of the five arom activities. The lower figures summarise the proteolysis data of Boocock, 1983; and Smith & Coggins, 1983 in which active fragments containing either E5 or E2 and E3 activities were isolated.

be further shortened to 63 kDa whilst still retaining E2 and E3 activity (and presumably domains) and that another non-overlapping 74 kDa fragment carrying only EPSP synthase (E5) activity could be identified (see Figure 1.8). This provides suggestive evidence for a multidomain structure of the *arom* multifunctional enzyme, but what of its relationship to the monofunctional *E.coli* activities which are all separable (Berlyn & Giles, 1969) and products of individual genes (Wallace & Pittard, 1966)? This fundamental question is tackled later in Chapter Six.

1.9.3 Advantages of multifunctional organisation

Having the functions of several enzyme activities covalently associated and encoded by a single gene ensures that they are co-ordinately expressed and that their *expression* can be similarly co-ordinately regulated. The covalent attachment of domains may allow novel or even essentially spatial assemblies of activities which may, in a multienzyme complex, prove sterically or thermodynamically unstable. Several other reasons for the existence of multienzyme complexes and multifunctional enzymes have been proposed including catalytic facilitation; substrate channeling; protection of unstable intermediates (Coggins & Hardie, 1986 and references therein). However little hard factual data, with the notable exception of *N.crassa* tryptophan synthase which has been shown to channel indole (Matchet, 1974), exist to validate these claims. Giles (1978) has proposed that the *N.crassa arom* may confer protection for catalytic intermediates in so far as preventing induction of the competing catabolic

quinate pathway (Section 1.7) by one of the multifunctional enzyme intermediates. Perhaps the most controversial hypothesis is that there is no reason at all but rather that multi-functional proteins, where they occur in contrast to monofunctional activities of other species, represent the most ancient form and that the split genes/enzymes are derived from them and not vice versa (see Chapter Six).

1.9.4 Other multifunctional enzymes

(A) Depending on the organism, fatty acid synthases (FAS) exhibit various patterns of both structural and functional variation. In most eubacteria and in chloroplasts there are eight structurally independent and monofunctional FAS components - Type II FAS (Block & Vance, 1977). In eukaryotes the enzymes of fatty acid biosynthesis exist not as discrete, separable polypeptides but as enzyme conjugates, both multi-enzyme complexes and multifunctional proteins - type I FAS. In fungi, FAS consists of a heteroligomer of two multifunctional enzymes arranged as $\alpha_6\beta_6$ and catalysing three and four reactions each, respectively (Schweizer et al., 1978). In most other eukaryotes (but not plants) FAS consists of a homodimer of two multifunctional proteins each catalysing all seven functions. There are however fundamental differences between the fungal and vertebrate type I FAS, in particular the active sites of both enzymes have been mapped and their order is different (McCarthy & Hardie, 1984).

Speculation regarding the evolution of the type I FAS tends to suggest a gene fusion mechanism of monofunctional domains. On the one hand the sequence similarities observed

at the type I FAS active sites and monofunctional analogues, and the ability to obtain proteolytic fragments of multifunctional FAS containing active domains (Hardie & McCarthy, 1986) would support the fusion hypothesis. Alternatively the clear differences observed between the type I FAS of vertebrates and fungi demands that if gene fusion took place then it did so convergently and not successively in generating these two type I FAS.

The pentafunctional FAS I gene (β subunit) of S.cerevisiae has recently been cloned and sequenced (Schweizer et al., 1986). The order of catalytic domains along the polypeptide has been unequivocably demonstrated by complementation of defined fas I mutants with overlapping FAS I subclones. Significant sequence homologies exist between the acyl transferase active sites of yeast and animal FAS and have prompted speculation that the fungal and vertebrate FAS (or at least parts thereof) may have evolved from a common ancestor (Schweizer et al., 1986). Such an evolutionary route would require, in order to explain the mutually inverted order of transacylation sites in both types of FAS, that insertion of FAS 2 (α subunit) domains into FAS I took place. The identification of a large unassigned inter-domain region in the yeast FAS I gene may indicate the presence of DNA sequences into which such insertion could take place without disrupting other functions (Schweizer et al., 1986).

(B) The organisation of the genes and enzymes involved in histidine biosynthesis in yeast and the enteric bacteria is strikingly different. In S.cerevisiae and other fungi the genes are not clustered in an operon, as in E.coli and S.typhimurium, but rather are located on different chromosomes. This is with the notable exception of the His4 gene. This gene encodes a multifunctional protein which catalyses four non-consecutive steps in histidine biosynthesis - the 2nd, 3rd, 10th and 11th reactions (Donahue et al., 1982). The order of domains in the His4 protein is 3rd (His4A), 2nd (His4B), 10th and 11th (His4C). The corresponding functions in enteric bacteria are coded for by the second (hisD) and eighth (hisIE) genes of the operon, each encoding a bifunctional enzyme. The nucleotide sequence of the E.coli hisD and the E.coli and S.typhimurium hisIE genes have recently been reported (Chiariotti et al., 1986). Comparison with the S.cerevisiae His4 sequence (Donahue et al., 1982) reveals an overall homology of 38% between the functional domains of hisI/His4A and hisE/His4B as well as the two functional domains of hisD/His4C. Although this overall homology level is modest (38%), there is striking similarity when hydropathic and predicted secondary structures are compared (Bruni et al., 1986). The S.cerevisiae His4 gene is clearly a fusion of several different functional domains which in E.coli and S.typhimurium are encoded by two different genes.

(C) The S.cerevisiae TRP5 gene encodes tryptophan synthase which in E.coli and S.typhimurium is a tetramer of $\alpha_2\beta_2$ form, each subunit the product of a single gene and encoding a partial reaction. The trpA and trpB genes, encoding the α and β subunits respectively, have been sequenced in both E.coli and S.typhimurium (Nichols & Yanofsky, 1979; Crawford et al., 1980). The availability of the DNA sequence of the yeast TRP5 gene (Zalkin & Yanofsky, 1982) allows a direct comparison with the bacterial sequences. The sequence of the bifunctional yeast enzyme is homologous with the α and β chains of E.coli. The primary structure of the yeast enzyme suggest an αβ chain fusion although this is the reverse of the chromosomal order found in the bacterial operon. In E.coli a single base addition or deletion mutation in the intercistronic trpB-TrpA region would fuse the subunits in βα chain order but no bifunctional enzyme with this order of domains has been found.

(D) The ADE3 gene of S.cerevisiae encodes C_1-tetrahydrofolate synthase. This multifunctional protein catalyses three sequential tetrahydrofolate interconversions (Paukert et al., 1977). Alternative organisations of the three activities is observed in prokaryotes ranging from E.coli, which lacks one activity but retains the other two as a bifunctional enzyme (Dev & Harvey, 1978), to Clostridium formicoaceticum which has three separable enzymes (Clark & Ljungdahl, 1982). In support of the multidomain character of the yeast ADE3 complex are limited proteolysis studies in which an active fragment

can be separated by trypsin digestion (Paukert et al., 1977). In addition mutations in the yeast ADE3 which inactivate only one activity without affecting the other activities suggests that the C_1-THF synthase has independent functional regions (McKenzie & Jones, 1977).

(E) In Escherichia coli aspartokinase I-homoserine dehydrogenas I and aspartokinase II-homoserine dehydrogenase II are two bifunctional enzymes that catalyse the first and third reactions of the common pathway of threonine and methionine biosynthesis. The sequence of the corresponding genes, thrA and met L, has been reported (Katinka et al., 1980; Zakin et al., 1983) and comparison of the two amino acid sequences has established that they are derived from a common ancestor. Furthermore an extensive amino acid sequence comparison of both molecules revealed that each polypeptide has four internal motifs homologous, in a pair-wise manner, with that of the other isoenzyme. Ferrara et al. (1984) have, on this basis, speculated that each of these two double-motifs evolved independently of each other and duplicated prior to the suspected gene fusion event which gave rise to the ancestral progenator of the two bifunctional proteins. The thrA and metL genes and their encoded proteins thus provide an example of gene duplication, gene fusion and sequence divergence.

This short summary of a few examples of multifunctional proteins is by no means exhaustive of the plethora of information available on this intriguing class of biological

catalysts. What it was intended to do was to highlight the diversity of multifunctional proteins and also to begin to consider some of the possible evolutionary routes responsible for their genesis. In this respect the yeast tryptophan synthase is clearly an example of gene fusion of E.coli - like α and β subunits, or is it? Why is the order always αβ, could the prokaryotic arrangement represent scission of an ancest ral multifunctional enzyme? The organisation of enzymes of C_1-THF synthase in eubacteria and eukaryotes could be explained assuming that the progenitor of both had a trifunctional complex which has been retained in the eukaryotes. Scission and divergence of this complex could explain the bacterial organisation as easily as any attractive gene fusion model progressing in the opposite direction.

Multifunctional proteins are thought to arise primarily from the fusion of genes for the corresponding monofunctional proteins. However, only very careful sequence comparisons can distinguish gene fusion and gene scission events. Recent evidence on the antiquity of eukaryotic introns (Gilbert et al., 1986; see Chapter Six) may require a reappraisal of our, sometimes unfounded, assumptions that prokaryotes preceded eukaryotes. Similarly, at what level of sequence identity does convergent evolution stop and divergent evolution begin? How can we distinguish between chance convergence towards any given primary structure (specifying the necessary, biologically-active 3-dimensional conformation) responsible for a given enzyme activity and bona fide sequence homology?

Whilst the question regarding evolutionary direction (fusion or scission) is open to interpretation, the difference between convergence and divergence is directly answerable. In this respect the *arom* multifunctional enzyme of *N.crassa* and yeast provides an excellent model system. The wealth of genetic, biochemical and now sequence data will allow examination of each of the domain activities and comparison with their monofunctional counterpart.

1.10 Objectives of this project

The genes and enzymes involved in polyaromatic amino acid synthesis differ significantly in their arrangement between bacteria and fungi. The occurrence of multifunctional *arom* polypeptides in fungi catalysing five central steps of the shikimate pathway not only contrasts sharply with the separate bacterial enzymes, but is a reversal of the relationship of other functionally related genes (Demerec, 1964), and as such raises important evolutionary questions.

A detailed comparison of the *E.coli* and yeast (*S.cerevisiae*) shikimate pathway activities has been undertaken in this laboratory. This will allow the assessment of the degree of relatedness between the mono- and multifunctional proteins. In addition, the conflicting hypotheses of gene fusion or gene scission (as applied to the *arom* enzyme evolutionary route) may be directly answerable and provide a more informative and general model for the evolution of multifunctional proteins.

In this respect, this thesis details the (sub)cloning of the E.coli aroB (Chapter 3), aroL (Chapter 5) and aroC (Chapter 4) genes, their sequence analysis and characterisation of encoded protein products. The information gained will allow further discussion on the evolutionary origin of the yeast arom multifunctional enzyme.

CHAPTER 2

MATERIALS AND METHODS

2.1 Materials

2.1.1 Fine chemicals

Amberlite MB3, acrylamide, NN'methylene bisacrylamide, p-aminobenzoic acid, ammonium sulphate (enzyme grade), bromophenol blue, caesium chloride, 99% formic acid (Analar), β-glycerophosphate, p-hydroxybenzoic acid, conc. hydrochloric acid (Aristar), 30% hydrogen peroxide, L-leucine, 2-mercapto-ethanol, L-phenylalanine, polyethylene glycol 6000, L-proline, N,N,N',N'-tetramethylene diamine (TEMED), SDS, L-tyrosine, and xylene cyanol were obtained from BDH Chemicals, Poole, Dorset, U.K.

Agarose, low melting temperature (LMT) agarose, Isopropyl-β-D-thiogalactoside (IPTG), phenol (Ultrapure), urea (Ultrapure), and 5-bromo-4-chloro-3-indolyl-β-galactoside (X-gal) were obtained from BRL, Gibco Ltd., Paisley, U.K.

NADH(Na^+ salt, Grade II), PEP(K^+ salt), NAD^+ (free acid, Grade I), Tris, Triethanolamine HCl (TEA HCl), and dithiothreitol (DTT) were obtained from Boehringer Corp., Lewes, East Sussex, U.K.

Bactotryptone, yeast extract and "Bactoagar" (agar) were obtained from Difco, Detroit, U.S.A. Oxoid No.1 agar (Oxoid) was obtained from Oxoid Ltd., London, U.K.

ATP, ampicillin, tetracycline, kanamycin sulphate poly-vinylpyrrolidone chloramphenicol, ethidium bromide Ficoll, and Coomassie Brilliant Blue G-250 were obtained from Sigma Chemical Co., Poole, Dorset, U.K.

3-Dehydroquinate (ammonium salt), and DAHP were the generous gifts of Dr S. Chaudhuri and Dr J. Lambert respectively. Shikimic acid was obtained from the Aldrich Chemical Co., Gillingham, Dorset, U.K.

All other reagents used were of the highest grade commercially available. All solutions were prepared, where appropriate, with glass-distilled water.

2.1.2 Chromatographic media

DEAE-Sephacel, Sephadex G-50 (medium grade), phenyl-Sepharose CL-4B, and Sephacryl S200 superfine were obtained from Pharmacia (GB) Ltd., London, U.K. Hydroxylapatite Bio-Gel HTP was obtained from Bio-Rad Laboratories, Richmond, CA, U.S.A. Procion Red HE3B was obtained from Amicon Corp., Lexington, Mass., U.S.A.

Prepacked Mono-Q (HR 5/5) and Superose-12 columns were obtained from Pharmacia (GB) Ltd., London, U.K.

2.1.3 Enzymes

The following enzymes were obtained from Boehringer Corp., Lewes, East Sussex, U.K.:-

aldolase (EC 4.1.2.13) from rabbit muscle,

alkaline phosphatase (EC 3.1.3.1) from calf intestine (and further purified by I. Anton, unpublished),

carbonic anhydrase (EC 4.2.1.1) from bovine erythrocyte,

deoxyribonuclease I (EC 3.1.4.5) from bovine pancreas,

glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1.12) from rabbit muscle,

glutamate dehydrogenase (EC 1.4.1.3) from beef liver,
malate dehydrogenase (EC 1.1.1.37) from pig heart,
lactate dehydrogenase (EC 1.1.1.27) from rabbit muscle/
pyruvate kinase (EC 2.7.1.40) from rabbit muscle
mixed suspension,

Bovine serum albumin, Ribonuclease A (RNase A), lysozyme, chicken ovalbumin, and horse heart cytochrome C were obtained from Sigma Chemical Co., Poole, Dorset, U.K.

Bacteriophage T4 DNA ligase, nuclease-free bovine serum albumin, and all restriction enzymes were obtained from BRL, Gibco Ltd., Paisley, U.K. Avian myeloblastosis virus (AMV) RNA dependant DNA polymerase (RTase) was obtained from NBL, Cramlington, Northumbria, U.K. Klenow fragment of E.coli DNA polymerase I was obtained from Amersham International plc, Amersham, U.K. (see Section 2.17).

2.2 Bacterial strains

The bacterial strains used in this study are shown in Table 2.1. Many were a gift of Dr M.G. Edwards, G.D. Searle R & D, High Wycombe, Bucks., U.K.

2.3 Plasmids

Plasmids constructed during the course of this study are shown in Table 2.3. Plasmids gifted or otherwise obtained are shown in Table 2.2. Recombinant M13mp8/9 clones are not specified individually but discussed in the context of their construction (Chapters 3, 4 and 5). Expression vectors pKK223/3 and pIH223/3 are described in Chapters 3 and 5.

organism	genotype	source/reference
E.coli K12	wild type ATCC 14948, F^-, λ lysogenic, Lederberg strain W3100.	American Type Culture Collection (Rockville, Maryland U.S.A)
E.coli HB101	F^-, *pro*, *leu*, *thi*, *lacY*, *hsdR*, *endA*, *recA*, *rpsL20*, *ara14*, *galK2*, *xyl5*, *mtl1*, *supE44*	Bolivar & Backman (1979)
E.coli JM101	Δ(*lac pro*), *thi*, *supE*, F'*traD36*, *proAB*, $lacI^q$, *lacZ*ΔM15	Messing *et al.* (1981)
E.coli HW0927	*proC32*, *metE70*, *trpE38*, *ara-14*, *lacZ36*, *mtl-1*, *xyl5*, *thi1*, *purE42*, *recA1*, *a216*, *tsx-67*, *rps1109*, *tonA*, *supE*.	Searle
E.coli HW1045	*tyrR*, *tyrA*, *trpR*: Kan^r	Searle
E.coli HW87	*araD* 139 (*ara-leu*)Δ7697, *lacZIPOZY*, Δ74, *galU*, *galK*, *hsdR*, *rpsL*, *srl*, *recA56*	Searle
E.coli HW1111	$lacI^qL8$, *tyrA* : Kan^r	Searle
E.coli AB2826	*aroB*, *supE42*, λ^-	CGSC:- (*E.coli* Genetic Stock Centre Dept. of Human Genetics, Yale University, New Haven, U.S.A.) Pittard & Wallace (1966)
E.coli AB2849	*aroC*	CGSC
E.coli TG1	As JM101 but also *hsd* Δ5 (EcoK r^-m^-)	T. Gibson (unpublished)

Table 2.1: Bacterial strains used.

Plasmid	Markers	Source
pKD106	*amp*^{r}, *aroB*$^+$	Duncan & Coggins (1983)
pLC29-47	*aroB*$^+$, *mrcA*$^+$	Takeda *et al*. (1981)
pJB14	*aroB*$^+$, *amp*^{r}	Frost et al. (1984)
pKK223/3	*amp*^{r}	J. Brosius, unpublished
pIH223/3	*amp*^{r}	I. Hunter, unpublished
pMH423	*proC*$^+$, *amp*^{r}, *tet*^{r}	M. Hunter, Searle
pLC33-1	*aroC*$^+$	Clarke & Carbon (1976)
pAT153	*amp*^{r}, *tet*^{r}	Twigg & Sherrat (1980)

Table 2.2: Plasmids used during this study.

Plasmid	Markers	Derivation	Vector
pGM107	*aroB*$^+$, *amp*^{r}	1.65 kbp *Eco*RI subclone of pJB14	pKK223/3
pGM108	*aroB*$^+$, *amp*^{r}, *tet*^{r}	1.65 kbp *Eco*RI subclone of pJB14	pAT153
pGM424	*amp*^{r}, (*aroL*$^+$)	2.7 kbp *Bam*HI subclone of pMH423	pAT153
pGM63A	*amp*^{r}, *proC*$^+$	*Bam*HI deleted pMH423	pAT153
pGM425	*amp*^{r}, (*aroL*$^+$)	*Pvu*II deletion of pGM424	pAT153
pGM429	*amp*^{r}, (*aroL*$^+$)	2.5 kbp *Bam*HI subclone of pMH423	pAT153
pGM430	*amp*^{r}, (*aroL*$^+$)	*Pvu*II deletion of pGM429	pAT153
pGM450	*amp*^{r}, (*aroL*$^+$)	1.3 kbp *Bam*HI subclone of pGM430	pIH223/3
pGM601	*aroC*$^+$, *amp*^{r}	5.2 kbp *Cla*I subclone of pLC33-1	pAT153
pGM602	*aroC*$^+$, *amp*^{r}	*Sal*I deletion of pGM601	pAT153
pGM603	*aroC*$^+$, *amp*^{r}	*Nru*I deletion of pGM602	pAT153
pGM604	*aroC*$^+$, *amp*^{r}	*Eco*RV deletion of pGM602	pAT153
pGM605	*aroC*$^+$, *amp*^{r}	1.7 kbp *Hinc*II/ *Nru*I subclone of pGM602	pKK223/3

Table 2.3: Plasmids constructed during this study.

2.4 Growth Media

2.4.1 Rich media

The constituents of 'rich' media (e.g. L-broth, 2x TY, L-agar) are shown in Table 2.4. The abbreviations LB and LA are used extensively in the text for L-broth and L-agar. Where appropriate, glucose (20% w/v) was autoclaved separately (5 psi for 50 minutes) and added later to the correct final concentration.

2.4.2 Minimal medium

Minimal medium (abbr. MM or M9S) was prepared as shown in Table 2.4. For solid MM a two-fold concentrated solution (2 x M95) was prepared and autoclaved separately from the agar (Oxoid). After cooling (to 55°C) the two solutions were mixed to the correct final concentration.

For both liquid and solid MM, 100 mM $CaCl_2$ and 20% (w/v) glucose were autoclaved separately and added later to the correct concentration. Vit B_1 (thiamine hydrochloride) 2 mg/ml was filter sterilized by passage through a 0.22 µM filter and added later to cooled media (see Table 2.4).

2.4.3 Supplements to growth media

Antiobiotic supplements (ampicillin, tetracycline, Kanamycin and chloramphenicol) were prepared as concentrated stock solutions, filter sterilized and added either to cooled (55°C) solid media or immediately before use (liquid media). The concentrated stock solutions and final amounts used are summarised in Table 2.5.

Medium (Liquid)	Constituents (l^{-1})
Minimal Medium (MM)	1g NH_4Cl 0.13g $MgSO_4 \cdot 7H_2O$ 3g KH_2PO_4 6g Na_2HPO_4 (+ 1 ml 100 mM $CaCl_2$) (+22 mg thiamine hydrochloride) (+ 10 ml 20% glucose)
L-Broth (LB)	10g bactotryptone 10g NaCl 5g yeast extract (+ 5 ml 20% glucose)
2 TY	16g bactotryptone 5g NaCl 10g yeast extract
Medium (Solid)	**Constituents (l^{-1})**
Minimal Agar	as MM, + 15g agar (Oxoid)
L-Agar	as LB, + 15g agar
H-agar	10g bactotryptone 8g NaCl 12g agar
H-top agar	10g bactotryptone 8g NaCl 8g agar

Table 2.4: Growth Media.

Supplement	Final Concentration	Stock Concentration
Ampicillin	100 µg/ml	25 mg/ml
Tetracycline	15 µg/ml	12.5 mg/ml (ethanol)
Kanamycin	50 µg/ml	25 mg/ml
Chloramphenicol	170 µg/ml	34 mg/ml (ethanol)
Aromatic compounds:		
(a) L-tyrosine	80 µg/ml	300 µg/ml
(b) L-phenylalanine	80 µg/ml	300 µg/ml
(c) L-tryptophan	40 µg/ml	150 µg/ml
(d) p-aminobenzoic acid	0.32 µg/ml	1.2 µg/ml
(e) p-hydroxybenzoic acid	0.32 µg/ml	1.2 µg/ml
Amino acids:		
(a) L-leucine	100 µg/ml	1 mg/ml
(b) L-proline	150 µg/ml	1 mg/ml

Table 2.5: Supplements to growth media.

Nutritional supplements (aromatics, amino acids) were prepared separately and added, in a similar fashion, to the correct final concentration.

2.5 General Methods

2.5.1 pH measurements

The pH of solutions was measured using a Radiometer pH meter and combination electrode at room temperature.

2.5.2 Conductivity measurements

Conductivity measurements were made (at $4^{o}C$) using a Radiometer conductivity meter type CDM2e (Radiometer, Copenhagen, Denmark).

2.5.3 Protein estimation

Protein estimation was made by the method of Bradford (1976) using bovine serum albumin as standard.

2.5.4 Acid-washed glassware

All glassware used for protein chemistry (e.g. amino acid analysis, N-terminal sequencing) was washed overnight in concentrated nitric acid. The acid-washed glassware was rinsed extensively with water and baked in an oven before use.

2.5.5 Microbiological techniques

(a) Microbiological safety

The recommendations as outlined in "Guidelines for Microbiological Safety" were followed.

(b) Bacterial strain storage

Several methods were employed for the preservation of strains used in this study. Long term storage was as a 50% (v/v) glycerol/LB solution at -20°C or as a stab (LB) at room temperature. Short term storage was on a suitable sealed plate at 4°C or as a 10 ml LB culture at 4°C. All strains were regularly tested for genetic markers and/or plasmids. Long term storage of plasmid was usually in a recA background (e.g. E.coli HB101).

(c) Bacterial growth

Bacterial growth was monitored at 650 nm (light-scattering) using a Gilford-Unicam model 252 spectrophotometer. Samples were diluted where necessary (in the appropriate medium) to maintain an absorbance value of less than 0.7-0.8 (upper threshold of linear range). Where an antibiotic or nutritional selection was available, it was used (except during the expression phase of transformed cells). Cell growth was at 37°C.

(d) Replica Plating

For a defined number of colonies, sterile toothpicks were used to transfer bacteria between selective plates. For whole plate transfer a pedestal and sterile velvets were used.

2.6 Preparation of crude extracts of E.coli

2.6.1 Sonicated crude extracts

100 ml MM cultures were harvested by centrifugation (Section 2.23.1) and resuspended in 5 ml of the appropriate buffer (see below). Sonication was used to break cells prior to high speed centrifugation. A soniprobe type 1130A (Dawe Instruments Ltd.) was used, sonication was at 80 watts for 4 x 30 secs. with 30 secs. rest intervals. The soniprobe housing was kept cool during sonication in an ice-water slurry. The sonicated cell mixtures (now less turbid) were centrifuged at 100,000g for 2 hours at 4°C, the supernatant was retained for enzyme assay and SDS PAGE analysis.

2.6.2 Extracts prepared using the French Pressure Cell

E.coli cells (wet paste) were resuspended in an appropriate volume of the specified buffer (see below) and broken by two passages through a French Pressure Cell at 98MPa (14,300 psi) (internal pressure). The French Pressure cell was cooled in ice prior to use.

For E.coli DHQ synthase determinations the following buffer was used for cell resuspension:

(i) Sonicated extracts -

200mM-KH_2PO_4 (pH 7.0), 0.2M-KCl, 2mM-$MgCl_2$, 1mM-βME

(ii) French Pressure Cell extracts -

10mM β-glycerophosphate (pH 6.6), containing 0.25mM-$COCl_2$, 0.5mM-NAD^+.

For E.coli shikimate kinase determinations the following buffer was used for cell resuspensions (for both types of cell breakage):

0.2M-Tris/HCl (pH 7.5), containing 0.2M-KCl, 5mM-$MgCl_2$, 0.4mM-DTT.

2.7 Enzyme assays

Spectrophotometric assays were carried out at 25°C in a final volume of 1 ml. A Gilford Unicam SP500 spectrophotometer interfaced with a Gilford photoelectric detector and chart recorder was used. One unit of enzyme activity is defined as the amount of enzyme required to catalyse the conversion of 1 µmole of substrate to product per minute.

(1) 3-Dehydroquinase

The appearance of DHS was measured spectrophotometrically at 234 nm (ε = 12,000 $M^{-1}cm^{-1}$). E.coli cell extracts were assayed in 0.1M KP_i (pH 7.0) buffer by initiating with 0.2mM DHQ.

(2) Shikimate kinase

Shikimate kinase activity was assayed spectrophotometrically at 340 nm (ε = -6,200$M^{-1}cm^{-1}$) by monitoring the disappearance of NADH. This was achieved by coupling the release of ADP to the pyruvate kinase and lactate dehydrogenase reactions (G.A. Nimmo, unpublished work). Final assay concentrations were:- 50mM-Triethanolamine/HCl/KOH(pH 7.0), 50mM-KCl, 2.5mM-$MgCl_2$, 0.1mM-NADH, 1mM-PEP (neutralised with KOH), 2.5mM-ATP (neutralised with KOH) and pyruvate kinase (3U/ml)/lactate dehydrogenase (2.5U/ml). E.coli extract

assays were initiated with 1mM-shikimic acid (final conc.). Correction for NADH oxidase blank rate is discussed in detail in Chapter Five.

(3) 3-dehydroquinate synthase

3-Dehydroquinate synthase was assayed spectrophotometrically at 234 nm (ε = 12,000$M^{-1}cm^{-1}$) by coupling the release of DHQ to the 3-dehydroquinase reaction and measuring the appearance of DHS. The final assay concentrations were:- 0.1M-glycine/KOH (pH 8.4), 0.2mM-$CoCl_2$, 50μM-NAD^+, 0.4 units partially purified E.coli 3-dehydroquinase. The assay was initiated by addition of DAHP to 400μM (final conc.). Correction for a rapid blank rate is discussed in detail in Chapter Three. Partially purified E.coli 3-dehydroquinase was a gift of Mrs S. Muir.

Checks on the authenticity of the observed DAHP-initiated DHQ synthase activity were as follows:

(a) Co^{2+} dependence

Endogenous Co^{2+} was removed from samples by incubation at 25°C (3 minutes) in 10mM-EDTA (pH 7.0). The treated sample showed no DHQ synthase activity under the standard assay conditions, unless Co^{2+} was added back.

(b) NADH inhibition

NADH is a known inhibitor of E.coli DHQ synthase activity (Srinivasan et al., 1963). Addition of NADH to 50μM final concentration decreases the rate of DAHP-initiated DHQ synthase activity.

2.8 Polyacrylamide gel electrophoresis in the presence of SDS (SDS PAGE)

SDS PAGE was performed by the method of Laemmli (1970) in a slab gel apparatus (Raven Scientific Ltd., Haverhill, Suffolk, U.K.). Separation gels were either 10%, 12.5% or 15% (w/v) acrylamide (at an acrylamide:bis-acrylamide ratio of 30:0.8). The gel buffer was 375mM-Tris/HCl pH 8.8, containing 0.1% (w/v) SDS and polymerisation was initiated by addition of 0.033% (v/v) N,N,N',N',-tetramethylene diamine, 0.05% (w/v) ammonium persulphate (final concentrations). Stacking gels of 3% (w/v) acrylamide were overlaid upon the separation matrix. The stacking gel buffer was 125mM-Tris/HCl pH 6.8, containing 0.1% (w/v) SDS. Polymerisation was induced by addition of 0.067% N,N,N',N',-tetramethylene diamine and 0.1% (w/v) ammonium persulphate (final concentration)

SDS PAGE was carried out at room temperature, and samples prepared by addition of 2% SDS (w/v), 1% (v/v) 2-mercaptoethanol, 5% (v/v) glycerol, 0.01% (w/v) bromophenol blue (final concentrations) followed by heating at 100°C for 3 minutes. The well buffer contained 3g/litre Tris base, 14.4g/litre glycine and 0.1% (w/v) SDS.

Following electrophoresis, gels were stained for protein using 0.1% (w/v) Coomassie Brilliant Blue G250, in 10% (v/v) acetic acid, 50% (v/v) methanol. Staining was for 1 hour at 45°C. Gels were destained overnight (at 40°C) in 10% (v/v) acetic acid, 10% (v/v) methanol.

2.9 Digestion of DNA with restriction endonucleases

Restriction digests of DNA were carried out as described by Maniatis *et al.* (1982). Four buffers were used throughout:

Low salt: 10mM-Tris/HCl pH 7.5, 10mM-$MgCl_2$, 1mM-DTT, 0.1 mg/ml BSA.

Medium salt: 50mM-NaCl, 10mM-Tris/Hcl, pH 7.5, 10mM-$MgCl_2$, 1mM-DTT, 0.1 mg/ml BSA.

High salt: 100mM-NaCl, 10mM-Tris/HCl, pH 7.5, 10mM-$MgCl_2$, 1mM-DTT, 0.1 mg/ml BSA.

*Sma*I salt: 20mM-KCl, 10mM-Tris/HCl, pH 8.0, 10mM-$MgCl_2$, 1mM-DTT, 0.1 mg/ml BSA.

Each buffer was prepared as a 10x stock solution (without BSA), sterile (nuclease-free) BSA at 1 mg/ml was added to each digest to the correct final concentration. Analytical digests were done in a volume of 20 µl, at the temperature recommended by the supplier (usually 37°C). Preparative digests were carried out in whatever final volume was required (not greater than 100 µl). When DNA was digested with two restriction enzymes, the endonuclease requiring the lower salt buffer was used first. After the recommended duration of digestion the salt concentration was adjusted and the second enzyme added.

2.10 Agarose gel electrophoresis of DNA

DNA was separated at room temperature on horizontal submerged agarose gels as described by Maniatis *et al.* (1982). The buffer system used was the Tris-borate (TBE) version. For accurately sizing restriction fragments of 0.3 - 3 kbp

a 1% (w/v) agarose gel was used. Increased resolution of DNA fragments >3 kbp or <0.5 kbp was achieved on 0.8% (w/v) or 2% (w/v) agarose gels respectively. Samples for agarose gels were prepared by addition of 5% (w/v) sucrose, 0.5% bromophenol blue (final concentrations) from a 10-fold concentrated stock solution. Ethidium bromide (0.5 µg/ml) was added to both gel and buffer, and stained DNA bands visualised on a long wave u.v. transilluminator (U.V. Products Inc.).

Low melting temperature (LMT) agarose gels (see Section 2.13.2) were run in an identical fashion except at 4°C.

Known restriction fragment markers, most commonly λcI85757/HindIII (and/or EcoRI) and pAT153/Hinf I (and EcoRI), were run alongside the unknown fragments.

2.11 Small scale preparation of plasmid DNA

The method of Holmes & Quigley (1981) was used for small scale plasmid preparations. A 10 ml LB overnight culture was harvested (MSE18 9,000xg, 3 minutes) and the cell pellet resuspended in 700 µl of lysis buffer (10mM-Tris/HCl, pH 8.0, 50mM-EDTA, 8% sucrose, 0.5% Triton X-100). Lysozyme was added (50 µl of a 10 mg/ml solution) and the mixture boiled for 45 seconds followed by centrifugation in a microfuge. The supernatant was transferred to a fresh microfuge tube and plasmid DNA precipitated by addition of an equal volume of isopropanol. After chilling at -20°C for 30 minutes, plasmid DNA was recovered by centrifugation (10 minutes). The plasmid DNA pellet was resuspended in 50 µl 10mM Tris/HCl,

pH 7.6, 1mM EDTA (TE). A 5 µl aliquot was sufficient for a single restriction digest. The pellet also usually contained much RNA which obscured DNA of 400 bp or less in subsequent agarose gels. To overcome this problem, digests were occasionally supplemented with DNAse-free RNAse A.

2.12 Large scale preparation of plasmid DNA

The alkali lysis method of Birnboim & Daly (1979), as modified by Maniatis et al. (1982), was used routinely to prepare milligram quantities of pure plasmid DNA.

(a) 25 ml of LB medium, containing the appropriate antibiotic, was inoculated with 0.1 ml of an overnight culture of plasmid transformed cells. This culture was incubated at 37°C with vigorous shaking until the A_{650} reached 0.65 (late logarithmic phase). A prewarmed 500 ml LB medium (+ antibiotic) was inoculated with the entire late log phase culture (25 ml) and incubated at 37°C for exactly 2.5 hours. Chloramphenicol was then added to a final concentration of 170 µg/ml and incubation continued for a further 16 hours.

(b) The bacterial cells were harvested by centrifugation at 4,000g for 10 minutes at 4°C (MSE 18). The cell pellet was washed once with 100 ml ice-cold 0.1M-NaCl, 10mM-Tris/HCl, pH 7.5, 1mM-EDTA and resuspended in 9 ml of 50mM-glucose, 25mM-Tris/HCl, pH 8.0, 10mM-EDTA. Freshly prepared lysozyme, 1 ml of a 50 mg/ml solution, in the above glucose-containing buffer, was added and the suspension left at 25°C for 5 minutes. 20 mls of freshly prepared 0.2M-NaOH, 1% (w/v) SDS

was added, the suspension was inverted gently to mix the contents and left on ice for 10 minutes. Lysis was achieved by the addition of 15 ml of 5M-potassium acetate pH 4.8, followed by mixing and incubation on ice for 10 minutes. Bacterial cell debris was removed by centrifugation (Prepspin 50 39,000xg, 30 minutes, 4°C) and nucleic acids precipitated from the resultant supernatant by addition of 0.6 volumes of isopropanol. After 15 minutes at 25°C the DNA was recovered by centrifugation (Prepspin 50 12,000xg, 30 minutes, 25°C). The pellet was washed with 70% (v/v) ethanol, dried briefly under vacuum and resuspended in 8 mls of TE buffer. RNAse A (preheated at 100°C for 10 minutes) and RNAse T_1 (preheated at 80°C for 10 minutes) were each added to a final concentration of 100 µg/ml and 100U/ml respectively and the solution incubated at 37°C for 30 minutes.

(c) The volume of the RNAse-treated DNA solution was measured carefully and exactly 1g of solid CsCl added for every millilitre of solution. The volume was remeasured and exactly 0.8 ml of a solution of ethidium bromide (10 mg/ml in H_2O) added for every 10 ml of CsCl solution. The final density of the solution was 1.55g/ml.

The CsCl-ethidium bromide solution was transferred into two 10 ml polycarbonate tubes suitable for a Beckman Ti 70.1 rotor and centrifuged for at least 36 hours (45,000 rpm, 20°C) in a Beckman L8-M ultracentrifuge.

Only long-wave u.v. light was used to visualise the plasmid DNA band which migrated ahead of the host chromosomal DNA. Plasmid DNA was removed by inserting a needle into the

top of the centrifuge tube and by pumping from below the desired DNA band. Ethidium bromide was removed by several extractions with TE saturated-1-butanol. The aqueous DNA solution was concentrated by extraction with non-saturated 2-butanol until the volume was approximately 1 ml. The solution was finally dialysed against multiple changes of 1 litre of TE buffer at 4°C. The DNA concentration was determined spectrophotometrically using the relationship (A_{260} = 1.0 = 50 µg/ml DNA). The integrity and purity of the DNA preparation was examined by agarose gel electrophoresis,

2.13 Extraction and purification of DNA samples

DNA was routinely purified by phenol and chloroform extraction followed by ethanol precipitation as follows.

2.13.1 Extraction of digestion products

Restriction digests and other solutions containing DNA were deproteinised by phenol extraction. The volume was adjusted to 100 µl (usually) and an equal volume of TE-saturated phenol added. The mixture was mixed by vortexing and centrifuged at 12,000 rpm for 3-5 minutes. The upper aqueous phase was removed to a fresh microfuge tube and the process repeated twice more. Traces of phenol were finally removed by extraction with 1 ml of TE-saturated chloroform in an identical manner. The final aqueous DNA pool was made 0.3M-sodium acetate and 2.5 volumes of ice-cold ethanol added. The mixture was cooled at -20°C overnight or at -70°C for 1 hour and the DNA recovered by centrifugation (12,000 rpm, 25°C, 10 minutes) in a microfuge.

2.13.2 Recovery of DNA from LMT agarose

DNA was visualised by ethidium bromide staining and low energy long-wave u.v. transillumination. The desired DNA band was excised, placed in a 1.5 ml microfuge tube and 200 µl 20mM-Tris/HCl, pH 7.5, 1mM-EDTA added. The gel slice was melted at 65°C for 15 minutes and phenol extracted. The first round extraction was back-extracted with 200 µl 20mM-Tris/HCl, pH 7.5, 1mM-EDTA and the aqueous samples pooled. The DNA was further purified by two rounds of phenol and 2 rounds of chloroform extraction. DNA was recovered by ethanol precipitation as detailed above (Section 2.13.1). DNA purified in this way was sufficiently pure for cloning and in vitro labelling by nick-translation (See Section 2.21).

2.14 Dephosphorylation of DNA

To prevent self-ligation, the terminal 5' phosphates of plasmid/bacteriophage vector DNA were removed by calf intestinal alkaline phosphatase (Chaconas & van de Sande, 1980) as described in Maniatis et al. (1982).

Phenol/chloroform purified DNA was ethanol precipitated and resuspended in 0.05M-Tris/HCl, pH 9.0, 1mM-$MgCl_2$, 0.1mM-$ZnCl_2$. Calf intestinal phosphatase (CiPase) was added (0.01 units) and the reaction mixture incubated at 37°C for 15 minutes. A second aliquot of CiPase was added and incubation continued for 30 minutes. The enzyme was inactivated by heating at 68°C for 15 minutes in the presence of (final concentrations) 10mM-Tris/HCl, pH 8.0, 0.1M-NaCl, 1mM-EDTA, 0.5% (w/v) SDS. Finally the dephosphorylated DNA was deproteinised by phenol/chloroform extraction and ethanol precipitation.

2.15 Ligations

(a) Ligations were performed, using bacteriophage T4 DNA ligase, in a final volume of 20 µl containing

66mM-Tris/HCl, pH 7.6

6.6mM-$MgCl_2$

0.5mM-ATP

10mM-DTT

ca. 100ng vector DNA

ca. 20ng insert DNA

The amount of DNA ligase added and the temperature and duration of the reaction depended upon the nature of the cohesive (or non-) ends. For sticky ends 0.4U, 15°C for 6 hours, for blunt ends 1.0U, 4°C for 12-16 hours.

(b) Recircularisation of plasmid DNA was carried out in an identical manner as detailed above except after 3 hours the volume was increased by 100% to encourage recircularisation ligations rather than concatemer formation.

2.16 Transformation of *E.coli* with plasmid DNA

Competent *E.coli* cells were prepared by a modified version of the transformation protocol of Dagert & Ehrlich (1979). 0.5 ml of an overnight 10 ml LB culture was used to inoculate 50 ml of LB medium. This fresh culture was incubated at 37°C for 4-5 hours (A_{650} = 0.4), the cells cooled on ice for 10 minutes and gently harvested (MSE18 7,000xg, 3 minutes). The cells were resuspended in 20 mls ice-cold sterile 50mM-$CaCl_2$ and left on ice for at least 20 minutes. The fragile competent cells were harvested and resuspended in 1 ml sterile 50mM-$CaCl_2$ (ice-cold).

Transformations were carried out in plastic 5 ml sterile Bijoux tubes. Ligation mix refers to any DNA being used in the transforming process and usually was a ligation mixture. The ligation mix was added to 100 µl of competent cells and the mixture incubated on ice for 10 minutes. The DNA/cell mixture was then heat-shocked at 37°C for 5 minutes. 2 ml of pre-warmed (37°C) LB was added and an expression period of 1.5 hours (in the absence of selective antibiotic) allowed (at 37°C without shaking) prior to plating.

2.17 M13/dideoxy DNA sequencing

All of the DNA sequencing carried out during the course of this study utilised the M13/dideoxynucleotide method (Sanger, 1981; Messing et al., 1983). The incorporated radionuclide was $[\alpha^{35}S]$ dATPαS and the buffer gradient gel system described by Biggin et al. (1983) was used extensively. The protocols used to generate recombinant M13, to prepare single strand DNA templates and to sequence these are described in the "M13 cloning and sequencing handbook" (Amersham International plc, 1983). The cloning manual was supplied with the M13 Sequencing Kit N.4502, and the methodologies described therein proved invaluable in both their application and reproducibility.

2.17.1 Preparation of M13 DNA (RF)

The paired vectors M13mp8 and M13mp9 (Figure 2.1) were used (Messing & Vieira, 1982). Large scale stocks of the double-stranded replicative form (RF) of both phage derivatives were prepared from cleared lysates (Section 2.12(b)) of 500 ml

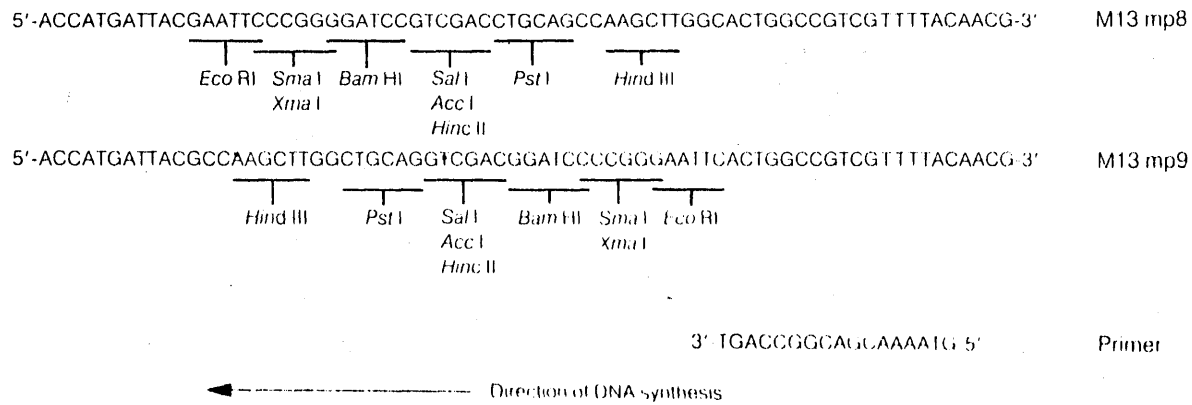


Figure 2.1: The paired M13-derivatised vectors mp8 and mp9 (Messing, 1983). Only the polylinker region and region complementary to the universal primer are shown. Note the inverted orientation of polylinkers in both vectors.

2 TY cultures of E.coli JM101 (transformed with either phage) using the CsCl-ethidium bromide gradient method described above.

2.17.2 Preparation of RF DNA for cloning

1 µg aliquots of RF DNA (both M13mp8 and M13mp9) were digested with the appropriate restriction enzymes required for the cloning procedure (see Figure 2.2). The extent of digestion was checked by agarose gel electrophoresis. After complete digestion the DNA was phenol/chloroform purified free of enzyme and ethanol precipitated. Stocks of digested vector were stored at 10 µg/ml in TE buffer at -20°C.

2.17.3 M13 ligations

These were exactly as described in Section 2.15.1. A molar ratio of 5:1 (insert:vector) DNA was maintained (100 ng:20 ng) in all ligations.

2.17.4 Transformation of E.coli JM101 (TG1) with M13 DNA

A single colony of E.coli JM101 (latterly TG1) from a stock MM plate was used to inoculate 10 ml of 2TY which was incubated overnight at 37°C. 2 ml of this fresh overnight culture was used to inoculate 40 mls 2TY which was incubated at 37°C for 3 hours. Simultaneously 0.1 ml of the overnight culture was used to inoculate 10 mls of 2TY which was incubated (not shaken) at 37°C to provide fresh cells.

Competent cells were made from the 40 ml 2TY culture exactly as described in Section 2.16 except the final $CaCl_2$-treated pellet was resuspended in 4 ml of 50mM-$CaCl_2$. The entire ligation mix (normally 20 µl) was added to 300 µl of competent cells and the mixture incubated on ice for at

Figure 2.2.

Restriction enzyme sites in M13	Restriction enzyme sites for generating fragments to be cloned	
1. *Eco* RI	*Eco* RI	↓ GAATTC CTTAAG ↑
	Eco RI*	↓ NAATTN NTTAAN ↑
2. *Bam* HI	*Bam* HI	↓ GGATCC CCTAGG ↑
	Bgl II	↓ AGATCT TCTAGA ↑
	Bcl I	↓ TGATCA ACTAGT ↑
	Sau 3A	↓ NGATCN NCTAGN ↑
	Xho II	↓ PuGATCPy PyCTAGPu ↑
3. *Pst* I	*Pst* I	↓ CTGCAG GACGTC ↑
4. *Hind* III	*Hind* III	↓ AAGCTT TTCGAA ↑
5. *Xma* I	*Xma* I	↓ CCCGGG GGGCCC ↑
6. *Hinc* II	Any blunt-ended fragment	
7. *Sma* I	Any blunt-ended fragment	

least 45 minutes. The cells were then heat shocked at 42°C for 3 minutes and returned to ice. The following (as a mixture prepared 10 minutes previously) were added to each transformation mix:

40 µl	100mM-IPTG
40 µl	2% (w/v) X-gal (in diethylformamide)
200 µl	fresh E.coli JM101 cells
3 ml	Molten H-top agar (42°C)

and the contents plated on a prewarmed (37°C) H plate.

2.17.5 Preparation of single-stranded template DNA

The M13 cloning system allows the easy identification of recombinant phage by the inactivation of the β-galactosidase marker. The simple plaque assay system can be used to differentiate recombinants (colourless plaques) from non-recombinants (blue plaques).

A 10 ml 2TY overnight culture of E.coli JM101 was used to inoculate fresh 2TY medium at a ratio of 1 ml:100 ml. A single plaque was lifted using a sterile Eppendorf pipette tip and inoculated into 1.5 ml of the low density E.coli JM101 culture. This culture was shaken (37°C) for 5 hours and then centrifuged for 5 minutes in a microfuge. The supernatant was carefully transferred to a fresh Eppendorf tube and recentrifuged (5 minutes, microfuge) to remove any residual cells. The supernatant was added to 200 µl PEG/NaCl (20% (w/v) polyethylene glycol 6000/2.5M-NaCl), mixed thoroughly and left for 15 minutes at room temperature. Precipitated intact phage particles were harvested by centrifugation (5 minutes, microfuge), and residual PEG/NaCl

removed by a second centrifugation step (2 minutes, microfuge) and aspiration. The viral pellet was resuspended in 100 µl TE buffer, 50 µl phenol added, the mixture vortexed for 20 seconds (at least) and left at room temperature for 15 minutes. The aqueous layer was transferred to a fresh microfuge tube and extracted with 1 ml chloroform. 10 µl of 3M-sodium acetate pH6.0 and 250µl ice-cold ethanol were added to the chloroform-extracted solution and the DNA precipitated at -20°C overnight. The DNA was recovered by centrifugation (10 minutes, microfuge) and the template DNA resuspended in 50 µl TE buffer. The integrity and purity of the template preparation was examined by running an aliquot on a 1% agarose gel. Template DNA was stored at -20°C.

2.17.6 Annealings

5 µl of template DNA, prepared as described above, was annealed for 2 hours at 55°C in a mixture also containing:

1 µl M13 sequencing primer (see below)

1.5 µl 100mM-Tris/HCl, pH 8.5, 100mM-$MgCl_2$ (10 x Klenow Buffer)

2.5 µl distilled water.

The M13 sequencing primer employed in this work was a 17mer with the sequence 5'-GTA AAA CGA CGG CCA GT-3'. Annealings were performed in bulk and stored at -20°C for up to 1 month.

2.17.7 Sequencing reactions

The annealed primer/template mix was thawed and 1.5 µl (15µCi) [α-^{35}S] dATPαS at >600 Ci/mmol (Amersham SJ.304) and

1 unit of Klenow fragment added. After thorough mixing, 2.5 µl aliquots of this mix was spotted just inside the rim of four (A,C,G,T) uncapped Eppendorf tubes. 2 µl of the relevant ddNTP/dNTP mix (prepared as described on page 32 of the Amersham Manual, see below) were added to the individual tubes and the contents mixed by a brief spin in the microfuge. The sequencing reactions were performed at ambient temperature and after twenty minutes 2 µl of a 0.5mM uniform chase solution of all four dNTP's added in a similar fashion. Following the 15 minute chase the sequencing reaction was stopped by addition of 4 µl of formamide dye mix (96% (v/v) deionised formamide, 0.03% (w/v) bromophenol blue, 0.03% (w/v) xylene cyanol FF, 20mM-EDTA).

2.17.8 Reaction mixes (composition)

(i) deoxyNTP mixes (A^o,C^o,G^o,T^o) for [α-^{35}S] dATPαS

(volumes are in µl)

	A^o	C^o	G^o	T^o
0.5mM dCTP	20	1	20	20
0.5mM dGTP	20	20	1	20
0.5mM dTTP	20	20	20	1
TE buffer	20	20	20	20

(ii) dideoxyNTP mixes

0.1mM ddATP

0.1mM ddCTP

0.3mM ddGTP

0.5mM ddTTP

(iii) dNTP/ddNTP mixes

To each dNTP mix (N^o), an equal volume of the corresponding ddNTP mix was added.

2.17.9 Polyacrylamide gel electrophoresis

The nested set of primer extended DNA fragments produced either in the M13/dideoxy DNA sequencing reaction or in the primer extension reverse run-off transcript mapping method (see Section 2.19) were resolved by electrophoresis on thin polyacrylamide gels (20 x 40 x 0.4 cm).

Gels were either linear or buffer gradient (Biggin et al., 1983) and were composed of the constituents detailed in Table 2.6. Samples (in formamide dye mix) were prepared by heating at 95-100°C for exactly 3 minutes during which time the gel slots were thoroughly cleared of unpolymerised acrylamide and urea. The heat-denatured samples were loaded immediately on to the gel and electrophoresis carried out at 28mA (constant current). The duration of electrophoresis varied with the amount of DNA sequence desired. A typical linear or buffer gradient gel took 2 - 2.5 hours to run (dye front reaches anode) but longer electrophoretic separations (4.5 hours linear, 3 hours buffer gradient) were also performed. A total of approximately 400 nucleotides of sequence information could be deduced from a combined linear/buffer gradient separation (2 gels).

Gels were preheated (linear only) for 30 minutes at 30 mA before electrophoresis and fixed in 10% (v/v) acetic acid, 10% (v/v) methanol after. Gels were dried onto a

sheet of Whatman 3MM paper using a Bio-rad model 1125 gel drier. The dried gel was autoradiographed (24-48 hours) using Fuji RX film at room temperature.

2.17.10 A-track analysis

Large numbers of recombinant clones were selectively screened for unique classes of sequence (inserts) by A-tracking. For each ten clones the following priming mix was prepared:

4 µl	M13 primer
6 µl	10 x Klenow buffer (see Section 2.17.6)
12 µl	distilled water

2 µl of this priming mix was annealed to 2 µl of each template as described previously. 3 µl of label (dATPαS), 16 µl dATP/ddATP and 2 units of Klenow fragment were mixed and 2 µl of mix added to the annealed primer/template. A 1 µl chase (15 minutes) followed the 20 minute sequencing reaction. Reactions were stopped by the addition of 1 µl of formamide dye mix and the samples heated and loaded as for normal sequencing gels.

2.17.11 In vitro preparation of M13RF DNA

Primer and template were annealed as described in Section 2.17.6. One unit of Klenow fragment and 1 µl of chase (0.5mM dNTPs) solution were added and the mixture incubated at ambient temperature for 30 minutes. The reaction was terminated by heating at 70°C for 10 minutes.

	Linear	buffer gradient	
		upper	lower
40% acrylamide	(ml) 7.5	6	1.12
10 x TBE	(ml) 5.0	2	1.88
sucrose	(g) -	-	0.75
urea	(g) 21	19.2	3.8
bromophenol blue (0.01g/ml)	(ml) -	-	0.08
distilled water	(ml) 37.2	31.8	4.4
TEMED	(µl) 50	80	15
AMPS (%)	(µl) 300(10%)	80(25%)	15(25%)

10 x TBE: 108g/l Tris base

(pH 8.3) 55g/l boric acid

9.3g/l EDTA.$2H_2O$

40% acrylamide: 38.2 acrylamide:bisacrylamide

TEMED: N,N,N',N'-tetramethylethylenediamine

AMPS: ammonium persulphate

Table 2.6: Sequencing acrylamide gel constituents.

2.17.12 Clone turn-around

Double-stranded RF DNA was prepared *in vitro* and the concentration of the solution adjusted to 1x High salt restriction enzyme buffer. The newly formed RF DNA was digested with *Eco*RI and *Hind*III and the products separated on a 1% LMT agarose gel. The appropriate band was excised, its DNA purified and ligated to the complementary vector, prepared as a *Eco*RI/*Hind*III digested stock (Sections 2.17.2 to 2.17.4).

2.18 DNA sequence data analysis

2.18.1 Compilation of DNA sequences

Primary sequence data (recorded from the relevant autoradiogram) was entered into a Digital PDP11-34 computer using the program BATIN (Staden, 1980). A sequence file was defined as the sequence located between restriction sites employed in the construction of the recombinant M13 clones. This rationale made no allowance for discrimination between genuine partial digest products and fortuitous multiple cloning events in the random cloning protocols used in Chapters 3-5. Thus each file was listed separately and treated as the result of a unique ligation event.

2.18.2 Manipulation of DNA sequences

The program TRNTRP was used to translate DNA sequences in any (or all) desired reading frames (Staden, 1978). The program SQRVCM was used to obtain the complementary strand of any given DNA sequence (Staden, 1980). Files containing complementary strands were compared using the program TTEM,

a version of DBCOMP (Staden, 1980), developed by Dr R. Eason (Department of Biochemistry, University of Glasgow). These programs were run on the departmental Digital PDP11-34 computer.

2.18.3 WISGEN

The University of Wisconsin Genetics Computer Group WISGEN package (Devereux et al., 1984) was accessed on the ERCC (Edinburgh Regional Computing Centre) Digital VAX 11/750 computer and the following programs used.

(i) TESTCODE: helped to identify coding regions by plotting the similarity of the pattern of codons in the three forward reading frames of a nucleotide sequence to a known pattern of codon choices.

(ii) BESTFIT: found the best region of similarity between two peptide sequences and where necessary inserted gaps, thereby incurring a negative quality factor, to optimise such alignments.

(iii) GAP: produced an optimal alignment between two sequences by inserting gaps in either one as necessary. The GAP output was used to run the PRETTY program.

(iv) PRETTY: wrote out a number of sequences with their columns aligned. PRETTY did not align sequences on the basis of similarity but simply formatted the GAP program outputs.

(v) CHOUFAS: made secondary structure predictions of a peptide sequence using the rules of Chou & Fasman (1978).

(vi) PUBLISH: arranged sequence figures in a format acceptable for publication.

(vii) WORDSEARCH: found segments of similarity between two

peptide sequences by finding regions with an unusual frequency (density of short perfect matches.

2.18.4 ANALYSEQ

The ANALYSEQ program (Staden, 1984) was run on the Digital VAX computer at G.D. Searle Research and Development, High Wycombe, Bucks., U.K. The ANALYSEQ program identified possible coding regions on the basis of codon preference and identified possible transcriptional and translational regulatory sequences by homology.

2.19 Transcript Mapping

2.19.1 Preparation of RNA from E.coli

RNA was prepared from exponentially growing E.coli cells by the method of Aiba et al. (1981) modified as detailed below. 100 ml of LB medium (supplemented with the appropriate antibiotic) was inoculated with 0.1 ml of a fresh overnight culture of E.coli cells. The culture was incubated at 37°C until the A_{650} = 0.25 - 0.30. The cells were harvested and resuspended in 5 mls 0.02M-sodium acetate, pH 5.5, 0.5% (w/v) SDS, 1mM-EDTA. Phenol saturated with this buffer, and pre-heated to 60°C, was added (10 mls) and the mixture incubated at 60°C for 10 mins. Following centrifugation (7,000xg, 5 minutes) the aqueous layer was re-extracted with phenol at 60°C for 5 minutes. The aqueous phase from a second centrifugation step was ethanol precipitated (see Section 2.13.1) and the nucleic acid recovered by centrifugation (10 minutes, 12,000 rpm). The pellet was resuspended in 400 µl of 0.02M-sodium acetate, pH 5.5, 0.5% (w/v) SDS,

1mM-EDTA and ethanol precipitated. This procedure was repeated once more and the final precipitate dissolved in TE buffer at 1-2 mg/ml and an aliquot examined by 1% agarose gel electrophoresis.

2.19.2 Synthetic oligonucleotides

These were the kind gift of Dr M.G. Edwards, G.D. Searle Research and Development, High Wycombe, Bucks., U.K., and were stored at -20°C as concentrated stock solutions.

2.19.3 Primer extension synthesis (reverse run-off)

10 µg of freshly prepared RNA was annealed with 2.5 ng of the appropriate oligonucleotide primer at 55°C for 30 minutes in 10mM-Tris/HCl, pH 8.5, 10mM-$MgCl_2$. Primer extension synthesis of the reverse run-off products was carried out by adding 10 µCi $[\alpha\ ^{35}S]$ dATPαS and 0.1mM (each) dGTP, dCTP and dTTP. AMV reverse transcriptase (23 units) was added and the mixture incubated at 30°C for 30 minutes. Run-off products were analysed using polyacrylamide sequencing gels.

2.20 Southern hybridisation conditions

2.20.1 Blot transfer of DNA to nitrocellulose

0.8% agarose gels containing DNA were blotted on to nitrocellulose (Schlucher & Schnell BA85) by the method of Southern (1975), as described in detail in Maniatis et al. (1982).

2.20.2 Simultaneous transfer to two nitrocellulose filters

Duplicate nitrocellulose filters were prepared from a single gel as described by Maniatis et al. (1982). The methodology involved in processing the gel prior to blotting (e.g. depurination, neutralisation) were exactly as described above. During the actual blotting the transfer buffer was supplied only by the liquid in the agarose gel itself.

2.20.3 Prehybridisation and hybridisation conditions

(i) Prehybridisation was carried out at 68°C in 0.2 ml buffer/cm^2 of filter. The prehybridisation buffer contained: 6 x SSC, 0.5% (w/v) SDS, 5 x Denharts solution, 100 µg/ml denatured salmon sperm DNA.

(ii) Hybridisation was carried out in the same buffer (50 µl/cm^2 of filter) including 0.01M-EDTA, and ^{32}P-labelled, heat-denatured DNA probe.

High stringency hybridisations were carried out at 68°C overnight (18 hours). Low stringency hybridisations were carried out at 55°C for no more than 12 hours.

Note:

SSC: 0.15M-NaCl, 0.015M-sodium citrate.

50 x Denharts solution: 1% (w/v) Ficoll, 1% (w/v) polyvinylpyrroline, 1% (w/v) BSA.

2.20.4 High stringency washes

After the required hybridisation period the filter was immersed in 200 ml 2 x SSC, 0.5% (w/v) SDS at room temperature. After five minutes the filter was removed and immersed in 200 ml 2 x SSC, 0.1% (w/v) SDS and incubated at room temperature for 15 minutes. The filter was subsequently washed in 200 ml 0.1 x SSC, 0.5% (w/v) SDS for 2 hours at 68°C.

2.20.5 Low stringency washes

Were as described above except the "high temperature wash" was carried out at the hybridisation temperature (55°C) in 2 x SSC, 0.5% (w/v) SDS for 2 hours.

2.20.6 Autoradiography of filters

After washing the filter was dried on a sheet of Whatman 3MM paper, wrapped in Saran Wrap and an X-ray film applied (Fuji RX). Autoradiograph was performed at -70°C with an intensifying screen.

2.21 Nick-translation of DNA samples

2.21.1 Nick-translation reaction conditions

DNA samples (0.5 - 1.0 µg) were radioactively labelled in vitro using the Amersham Nick Translation Kit (Catalogue No. N.5000). The DNA to be labelled was phenol/chloroform extracted and ethanol precipitated prior to nick-translation, in order to remove contaminants like agarose known to inhibit the polymerase activity. The DNA was recovered by centrifugation and resuspended in 10 µl of TE buffer. The nick translation

was carried out in a final volume of 100 µl containing (final concentrations):

50mM-Tris/HCl, pH 7.8

10mM-$MgCl_2$

0.1mM-DTT

50 µg/ml BSA

5 units DNA polymerase I (holoenzyme)

100 pg DNase I

10 µl $[\alpha^{32}P]$ dCTP solution (see below)

20 µM each dATP, dGTP, dCTP, dTTP

Nick translation reactions were carried out at 15°C for 4 hours.

For hybridisations against total genomic DNA ('genomic blots') higher specific activity 3000 Ci/mmol $[\alpha^{32}P]$ dCTP (Code PB. 10205) was used. For hybridisations against cloned fragments of DNA the incorporated radionuclide was $[\alpha^{32}P]$ dATP specific activity > 400 Ci/mmol (Code PB. 10164).

2.21.2 De-salting conditions

Unincorporated radionuclide was removed by gel exclusion chromatography on a 10 cm x 1 cm Sephadex G-50 column. The reaction mixture was loaded directly on to a 10 cm column of Sephadex G-50 equilibrated with 150mM-NaCl, 10mM-EDTA, 0.1% (w/v) SDS, 50mM-Tris/HCl, pH 7.5. Ten drop (400 µl) fractions were collected manually and monitored for radioactivity. The labelled DNA was eluted first, followed by a one fraction 'trough' in radioactivity, and then a large 'peak' of unincorporated radionuclide. The radiolabelled DNA

was recovered from the pooled fractions by ethanol precipitation. The DNA isolated by this procedure was suitable for use as a probe in hybridisation experiments.

2.22 Quantitation of ^{32}P incorporation into DNA samples (Note pg 83)

A 1 µl aliquot of reaction mixture (see Section 2.21.1) was removed and mixed with 200 µl of distilled water. 20 µl of the resulting solution was transferred to a tube containing 50 µl of carrier DNA. 1.5 ml of ice-cold 10% (w/v) trichloroacetic acid (TCA) was added to the carrier DNA/labelled DNA mix and the mixture incubated on ice for 15 minutes. The labelled and carrier DNA co-precipitated under these conditions and were collected by vacuum filtration on a glass fibre (Whatman GF/C) filter disc. The filter disc was washed thoroughly (TCA) and dried using an infra-red lamp. The dried filter was counted by liquid scintillation using 5 ml ECOSCINT (National, Diagnostics) as fluor. The approximate specific activity of the labelled DNA was calculated.

2.23 Enzyme preparations

2.23.1 Growth of cells

For enzyme preparations from cells transformed with *tac* plasmid constructs the cell growth conditions (including induction conditions) are detailed in the relevant chapters (3 and 4). Growth of *E.coli* HB101 or *E.coli* cells transformed with pAT153-derived constructs was as follows.

(i) A fresh overnight 50 ml LB culture was used to inoculate (1 ml) a 50 ml MM culture (supplemented where necessary with the appropriate antibiotic or nutritional additive). This

culture was incubated overnight at 4°C.

(ii) Eight 50 ml MM sub-cultures were maintained from the overnight MM culture and were incubated overnight at 37°C.

(iii) Large scale growth of cells was in 4 x 3 litre MM in 10 litre flasks. Each 3 litre MM (+ supplements) culture was inoculated with 2 x 50 ml MM overnight cultures. The large cultures (3 litres) were stirred constantly by bar magnets and aerated with compressed air at a flow rate of 400 ml/min. The cells were grown to late logarithmic phase before harvesting by centrifugation (8,000 x g, 15 minutes, 4°C). The cells were stored as a paste at -20°C. A yield of 20-30g E.coli cells (wet weight) was obtained from 12 litres of MM. E.coli cells were grown by Mr J. Greene.

2.23.2 E.coli DHQ synthase preparation

All steps after cell breakage were performed at 0-4°C.

(i) Extraction and centrifugation

E.coli AB2826/pGM107 (20g) were passed twice through a pre-cooled French pressure cell at 98 MPa (14,300 psi) (internal pressure), and were then extracted with 30 ml 10mM-β-glycerophosphate, pH 6.6, containing 0.25mM-$CoCl_2$, 0.5mM-NAD^+ (buffer A). Deoxyribonuclease I (0.5 mg) was added and the extract stirred for 1 hour at 4°C. The extract was then centrifuged (MSE18 18,000 rpm, 30 minutes, 4°C) and the supernatant (the crude extract) was dialysed against buffer A (1 litre) for 3 hours.

(ii) Hydroxylapatite chromatography

The dialysed crude extract was loaded on to a 400 ml (25 cm x 4 cm diameter) hydroxylapatite column pre-equilibrated

in buffer A. The column was washed with buffer A (400 ml), and the protein was eluted with a linear gradient of buffer A (2 litres) plus 2 litres of 75mM-potassium phosphate, pH 6.6, containing 10mM-β-glycerophosphate, 0.25mM-$CoCl_2$, 0.5mM-NAD^+ (buffer B). The flow rate was 50 ml/hr and 13.5 ml fractions were collected. Active fractions (containing DHQ synthase activity) were pooled and concentrated by vacuum dialysis. The pooled solution was dialysed overnight against 2 litres 25mM-potassium phosphate, pH 6.6, containing 0.25mM-$CoCl_2$, 0.5mM-NAD^+ (Buffer D).

(iii) Dyematrix Red A (Procion Red) chromatography

A Dyematrix Red A - agarose gel column (20 cm x 2.5 cm diameter) was washed with 10 column volumes (1 litre) of 8M urea at room temperature and then washed with 10 column volumes 10mM-potassium phosphate pH 6.6, containing 0.25mM-$CoCl_2$, 0.5mM-NAD^+ (Buffer C). The column was pre-equilibrated by washing with 10 column volumes buffer D. The pooled hydroxyl-apatite fractions were applied to the column and washed with 100 ml buffer D. Protein was eluted with a linear gradient of buffer D (400 ml) and 400 ml 150mM-potassium phosphate pH 6.6, containing 0.25mM-$CoCl_2$, 0.5mM-NAD^+ (Buffer E). The flow rate throughout was 38 ml/hour and 14 ml fractions were collected. Fractions containing DHQ synthase activity were pooled, concentrated by vacuum dialysis and dialysed into buffer E containing 50% (v/v) before long-term storage at -20°C.

2.23.3 FPLC separation of E.coli shikimate kinase

E.coli cells (see Chapter 4) were disrupted (7g, wet weight) by two passages through a pre-cooled French Pressure cell at 98MPa (internal pressure) and extracted into 15 ml 50mM-Tris/HCl, pH 7.5, containing 5mM-$MgCl_2$. 5 mg of protein from the supernatant of a high speed centrifugation (100,000xg, 2 hours, 4°C) step, was subjected to chromatography on an FPLC Mono-Q (Pharmacia) column. The column had previously been pre-equilibrated with 50mM-Tris/HCl, pH 7.5, 5mM-$MgCl_2$ and protein was eluted with a linear gradient of KCl (0-300mM) in 50mM-Tris/HCl, pH 7.5, 5mM-$MgCl_2$. A constant flow rate of 1 ml/min was used throughout and 1 ml fractions collected. Fractions containing shikimate kinase activity were identified by the coupled spectrophotometric assay.

2.23.4 E.coli shikimate kinase preparation

All steps after cell breakage were carried out at 0-4°C.

(i) Extraction and centrifugation

E.coli HW1111/pGM450 cells (17g) were disrupted by passage (twice) through a pre-cooled French Pressure cell (98MPa), and were then extracted with 30 ml 50mM-Tris/HCl, pH 7.5, containing 50mM-KCl, 5mM-$MgCl_2$, 0.4mM-DTT (buffer A). Deoxyribonuclease I (0.5 mg) was added and the extract stirred at 4°C for 90 minutes. The supernatant of a subsequent high speed centrifugation (100,000xg, 90 minutes, 4°C) step was dialysed against 4 litres of buffer A (4°C, 4 hours). The resulting solution was termed the crude extract.

(ii) DEAE-Sephacel chromatography

The dialysed crude extract was loaded (flow rate 50 ml/hour) on to a DEAE-Sephacel column (10 cm x 3.5 cm diameter) pre-equilibrated in buffer A. The column was washed with buffer A (350 ml) and protein eluted with a linear gradient (600 ml) of 50-300mM KCl in buffer A. The flow rate for the wash and gradient step was 60 ml/hour and 14 ml fractions were collected. Fractions containing shikimate kinase activity were pooled.

(iii) Phenyl-Sepharose chromatography

Solid $(NH_4)_2SO_4$ was added to the DEAE-Sephacel pool to 30% saturation (164g/litre). The solution was stirred (30 minutes, 4°C) and then centrifuged at 20,000xg for 15 minutes. The supernatant was loaded on to a Phenyl-Sepharose CL-4B column (4 cm x 2 cm diameter) pre-equilibrated in 100mM-Tris/HCl, pH 7.5, containing 1.2M-$(NH_4)_2SO_4$, 0.4mM-DTT (buffer B). The column was washed overnight with the same buffer (flow rate 20 ml/hour, collecting 10 ml fractions). The following day protein was eluted with a decreasing linear gradient (400 ml) of 1.2M - 0M$(NH_4)_2SO_4$ in 100mM-Tris/HCl, pH 7.5, containing 0.4mM-DTT. The flow rate was 20 ml/hour and 10 ml fractions were collected. At the end of the gradient the column was washed with a further 250 ml of 100mM-Tris/HCl, pH 7.5, containing 0.4mM-DTT until the residual shikimate kinase activity was eluted.

(iv) Abortive phosphocellulose chromatography

The pooled fractions from step (iii) were dialysed against 4 litres of 10mM-potassium phosphate, pH 6.5, containing 0.4mM-DTT (buffer E) overnight at 4°C. The

shikimate kinase activity failed to bind to a phosphocellulose column (20 cm x 1 cm diameter) equilibrated in buffer E and active fractions (flow-through) were pooled and concentrated by vacuum dialysis prior to gel permeation chromatography.

(v) Gel filtration on Sephacryl-S200

The recovered material from the failed phosphocellulose column was dialysed overnight (20 hours) against 1 litre of buffer A containing 10% (v/v) glycerol. The glycerol pool (20 ml) was divided into two equal aliquots and each applied to a Sephacryl S-200 column (85 cm x 2.5 cm diameter) pre-equilibrated in buffer A. The protein was eluted with buffer A; the flow rate was 10 ml/hour and 4 ml fractions were collected. Fractions containing shikimate kinase activity were examined by SDS PAGE and pure, active fractions pooled. This final pool was dialysed against 1 litre of buffer A containing 50% (v/v) glycerol and the enzyme stored at -20°C.

2.23.5 Superose 12 FPLC separation of shikimate kinase

A Superose 12 gel permeation column (Pharmacia) was equilibrated with 50mM-Tris/HCl, pH 7.5, containing 150mM-KCl, 0.4mM-DTT (Buffer A) on a Pharmacia FPLC apparatus. The column eluate was monitored at 280 nm and the column calibrated with the following proteins:

chicken ovalbumin M_r 45,000

E. coli DHQ synthase M_r 38,900

bovine erythrocyte carbonic anhydrase M_r 29,000

sperm whale myoglobin M_r 17,200

horse heart cytochrome C M_r 12,800

Purified *E. coli* shikimate kinase (10 units) was applied to the column and eluted in buffer A at a flow rate of

0.5 ml/minute, fraction size 0.25 ml). Fractions containing shikimate kinase were verified by assay.

2.24 Performic acid oxidation and amino acid analysis

Approximately 30 nmoles of purified protein was dialysed exhaustively against multiple changes of 0.5% (w/v) ammonium bicarbonate and then lyophilised. The protein was resuspended in 2 ml of distilled water and lyophilised again. Samples of protein were performic acid oxidised (Hirs, 1967) prior to amino acid analysis. The lyophilised protein was resuspended in 420 µl 70% (v/v) formic acid and 80 µl of methanol was added. 500 µl of freshly prepared performic acid reagent (95 volumes of formic acid mixed with 5 volumes of 30% H_2O_2 (w/v), heated at 25°C for 2 hours) was added to the protein and the mixture incubated at -5°C for 2 hours. The reaction was terminated by dilution with 8 volumes of distilled water and lyophilisation. The oxidised protein was resuspended in distilled water, lyophilised again and finally resuspended in 200 µl of 99% (v/v) formic acid. 50 nmoles of DL-*nor*-leucine was added as an internal standard and the final volume adjusted to 2 ml with distilled water. The sample was divided equally among 4 pyrex test-tubes and freeze-dried.

Each tube was freeze dried again and 500 µl of 6M-HCl (Aristar) containing 0.17% (v/v) 2-mercaptoethanol added to each. The tubes were sealed under vacuum and hydrolysis performed at 105°C for 24, 48, 72 and 96 hours. The tubes were opened and desiccated over c.H_2SO_4 (30 ml) and NaOH pellets. The samples were resuspended in 500 µl of distilled water, desiccated as before and resuspended in 125 µl

distilled water. Amino acid analyses (duplicates) were carried out on an LKB 4400 amino acid analyser. Analyses were performed by Mr J. Jardine at the Biochemistry Department, University of Glasgow.

2.25 Carboxymethylation and N-terminal protein sequencing

Carboxymethylation of cysteine residues was carried out essentially as described by Lumsden & Coggins (1978). 100 nmoles of protein was dialysed exhaustively against multiple changes of 0.5% (w/v) ammonium bicarbonate. The sample was freeze-dried, resuspended in 2 mls distilled water and freeze-dried again. The protein was dissolved in 2 ml of 0.1M-Tris/HCl, pH 8.2, 8M-urea, 2mM-DTT and incubated in the dark, at room temperature, under N_2. The solution was then made 15mM in iodoacetic acid and incubated for a further hour. The reaction was terminated by addition of excess DTT (30mM) and the alkylated protein dialysed against 0.5% (w/v) ammonium bicarbonate. The carboxymethylated protein was lyophilised and stored at -20°C.

(ii) The automated N-terminal amino acid sequencing was carried out at the SERC funded protein sequencing facility at Aberdeen University in collaboration with J.E. Fothergill, L.A. Fothergill-Gilmore and B. Dunbar. Analysis was carried out on a Beckman Model 890 Liquid phase sequencer (Smith et al., 1982). The phenylthiohydantoin samples were identified by chromatography on a pre-calibrated Waters Resolve C_{18} reverse phase h.p.l.c. column with a pH 5.0 acetate-acetonitrile buffer system (Carter et al., 1983). S- methylcysteine

was used as an internal standard.

2.26 In vitro DNA-directed coupled transcription-translation analysis (IVT)

2.26.1 IVT

The bacterial cell-free coupled transcription-translation system used was supplied in kit form (Amersham No. N.380). 2.5 µg of ccc. plasmid DNA was incubated at 37°C for 60 minutes in a final volume of 30 µl containing 28 µCi L-[^{35}S] methionine, an S-30 extract prepared from E.coli MRE 600, and an amino acid (-methionine)/nucleotide supplement. Exact concentrations of the various amino acid/nucleotide/S-30 extract components was not available but follows that first described by De Vries and Zubay (1967). After the allotted incubation time the reaction was chased with a methionine chase solution (37°C, 5 minutes) to complete protein chains which might have been prematurely terminated due to the limiting concentration of radioactive methionine. Proteins synthesised in this reaction were analysed by SDS PAGE, as described earlier (see Section 2.8), the gels dried under vacuum and radiolabelled protein bands detected by auto-radiography at -70°C.

2.26.2 Monitoring incorporation of L-[^{35}S] methionine

1 µl of the reaction mixture described above was added to 500 µl 1M-NaOH and incubated at 37°C for 15 minutes. The sample was cooled on ice and 3 ml ice-cold 25% (w/v) trichloroacetic, containing 1 mg/ml casein hydrolysate

(DIFCO) added. The mixture was incubated at 0°C for 30 minutes. Precipitated proteins were collected by vacuum filtration through glass fibre (Whatman GF/C) disc filters which were then washed extensively with 5% (w/v) TCA followed by ethanol. The discs were dried, added to 10 ml ECOSCINT fluor and radioactivity determined by liquid scintillation counting.

Note For ^{32}P quantitation, cpm were converted to dpm by assuming a efficiency of 80%

CHAPTER 3

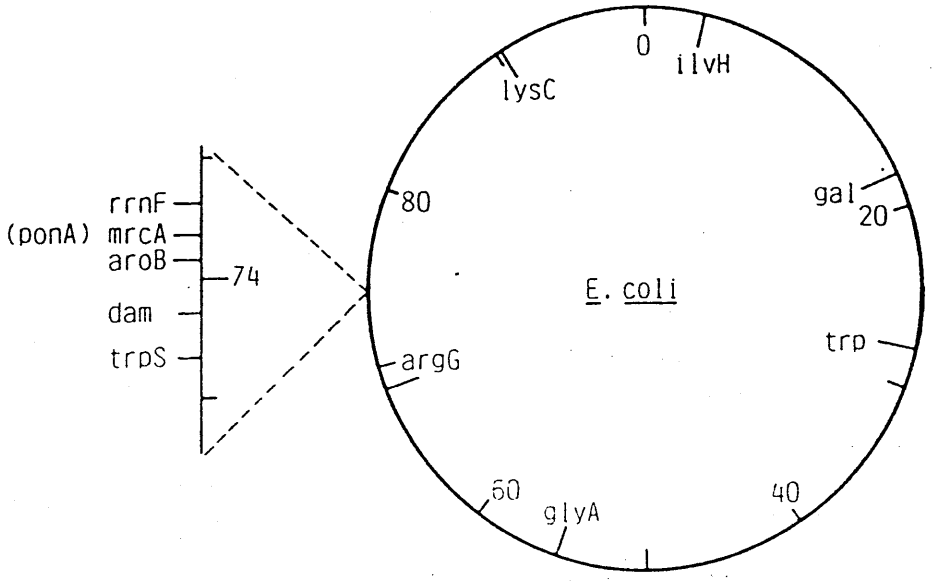
STUDIES ON THE aroB GENE OF E.COLI K12 ENCODING 3-DEHYDROQUINATE SYNTHASE

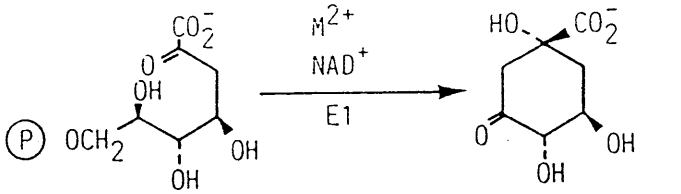
3.1 Introduction

3.1.1 The enzyme

3-Dehydroquinate synthase (E.C.4.6.1.3) catalyses the cyclisation of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAHP) to 3-dehydroquinate (Srinivasan *et al.*, 1963). The formation of this important metabolic intermediate constitutes the second step of the common aromatic amino acid biosynthetic pathway (Fig. 3.1).

Despite the low levels at which this enzyme naturally occurs, purification to homogeneity of monofunctional DHQ synthase activities from *Escherichia coli*, *Bacillus subtilis* and mung bean have been reported (Maitra & Sprinson, 1978; Frost *et al.*, 1984; Hasan & Nester, 1978b; Yamamoto, 1980). The purification of DHQ synthase from *E.coli* B required a 3,000 fold enrichment (Maitra & Sprinson, 1978). Estimation of the apparent molecular weight under native gel filtration conditions and by denaturing SDS PAGE suggested that the enzyme was monomeric with sub-unit molecular weight of c.57,000 (Maitra & Sprinson, 1978). In *B.subtilis* DHQ synthase is associated with two other activities in a multi-enzyme complex and has a subunit molecular weight of 17kDa (Hasan & Nester, 1978b). Yamamoto (1980) has purified DHQ synthase from mung bean. Like *E.coli* (Maitra & Sprinson, 1978) the mung bean DHQ synthase activity requires both NAD^+ and a divalent transition metal cation for activity (Yamamoto, 1980).





3-Deoxy-D-Arabino-Heptulosonic Acid-7-PO_4

3-dehydroquinate

Figure 3.1: The chromosomal location of the E.coli aroB gene encoding 3-dehydroquinate synthase (E1).

Recently the isolation of a monomeric DHQ synthase from E.coli K12 has also been reported (Frost et al., 1984). Purification of the enzyme from a wild-type K12 strain and from an overproducing strain yielded a single activity with the same sub-unit molecular weight of ca. 40,000. The identical electrophoretic mobility of DHQ synthase from both sources strongly suggests that the plasmid-encoded and the chromosomally-encoded enzymes are the same (Frost et al., 1984). The specific activity of the E.coli K12 DHQ synthase activity was 44 units/mg which is twelve fold greater than that reported for the E.coli B activity (Maitra & Sprinson, 1978). The primary structure of the E.coli aroB gene product, DHQ synthase was not determined and the serious discrepancy between the molecular weights of the E.coli enzyme estimated by the two groups needed to be resolved. This would have to be completed before any questions regarding comparative or mechanistic aspects of DHQ synthase activity could be addressed.

3.1.2 The aroB gene

The E.coli aroB gene encoding DHQ synthase maps at minute 74 on the chromosome (Figure 3.1) (Bachmann, 1983). Takeda et al. (1981) have shown that the Clarke & Carbon E.coli gene bank (Clarke & Carbon, 1976) contains a recombinant plasmid, pLC 29-47, which carries aroB. This hybrid-ColE1 plasmid has been the starting material for two independent studies on the expression of the aroB gene.

3.2 Previous work on the aroB gene

3.2.1 Plasmid pkD106 (Duncan & Coggins, 1983)

A comparison of the restriction enzyme analysis of pLC 29-47 (Takeda et al., 1981) with the restriction map of ColE1 (Dougan et al., 1978) showed that most of the E.coli genomic material carried on pLC 29-47 was located within two PstI sites separated by 7.2 kbp (Duncan & Coggins, 1983) (Figure 3.2). The auxotrophic requirements of E.coli AB2826 (aroB) could be relieved when the cells were transformed with pLC 29-47. Levels of DHQ synthase activity were elevated ten-fold relative to wild-type (E.coli K12) levels in crude extracts of pLC 29-47-containing strains (Duncan & Coggins, 1983).

The 7.2 kbp PstI fragment of genomic insert of pLC 29-47 when sub-cloned into the high copy-number plasmid vector pAT153 (Twigg & Sherratt, 1980) retained the ability to complement the nutritional requirements of E.coli AB2826 (aroB). Furthermore a 3.6 kbp sub-clone of this genomic region, bounded by restriction sites EcoRI and BGlII, was also able to confer a prototrophic phenotype upon the E.coli aroB mutant. E.coli AB2826 transformed with either the larger 7.2 kbp (pKD101) or the smaller 3.6 kbp (pKD106) aroB-containing clone could overexpress DHQ synthase activity. The levels of overexpression for both plasmid transformed strains were approximately 20-fold, relative to the wild-type enzyme level (Duncan & Coggins, 1983).

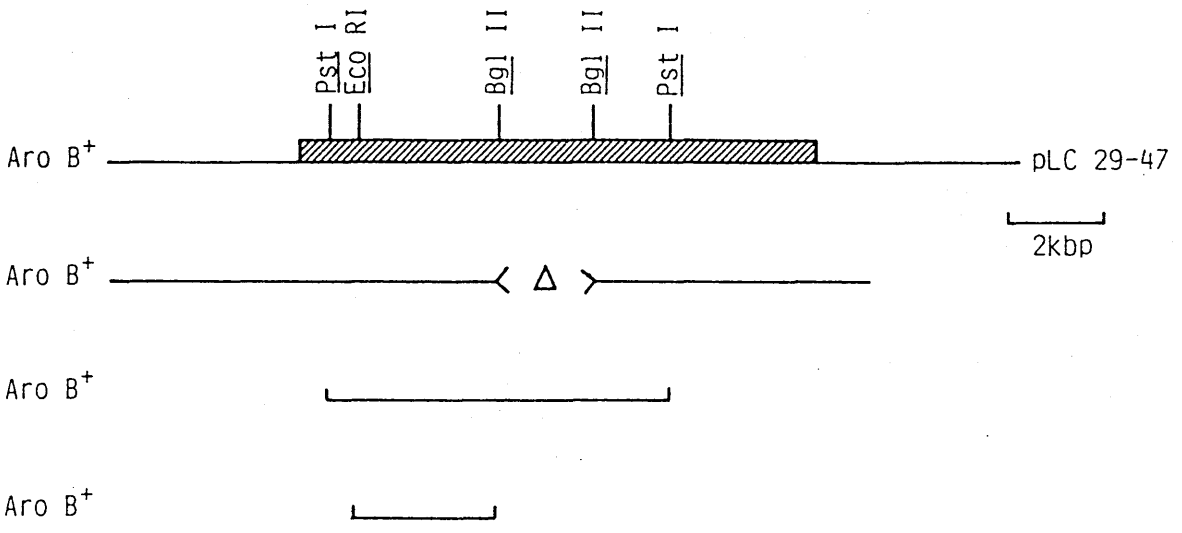


Figure 3.2: Deletion and subcloning analysis of plasmid pLC29-47 defining a putative aroB region. Adapted from Duncan & Coggins, 1983.

3.2.2 Plasmid pJB14 (Frost et al., 1984)

The aroB gene carried by pLC 29-47 was independently cloned by a different route by Frost et al. (1984). Both the cloning strategy and choice of vector was significantly different from that employed to generate pKD106 (Duncan & Coggins, 1983).

The plasmid vector used was the tac expression vector pKK223/3 (Brosius, J., unpublished results; kindly provided by Professor J.R. Knowles, this plasmid is commercially available from Pharmacia). The structure of pKK223/3 is shown in Figure 3.3. This plasmid vector is an example of an inducible expression vector designed to generate very high levels of expression of foreign genes in E.coli. The plasmid is pBR322 (Bolivar et al., 1977) derivative and contains the pBR322 origin of replication. It also contains the powerful trp-lac (tac) promoter (de Boer et al., 1983) which comprises of the -35 region of the trp promoter and and -10 region, operator and ribosome binding site of the lac UV5 promoter. In a lac I^q host strain the tac promoter is repressed and can be derepressed when necessary by addition of an inducer such as isopropyl β-D-thiogalactoside (IPTG). Immediately adjacent to the tac promoter is the multiple cloning region or 'polylinker' derived from pUC-8 (Vieira & Messing, 1982), which facilitates the positioning of genes behind the promoter and ribosome binding site. The stability of the host-vector system is maintained by the location downstream of the 'polylinker' of the dual E.coli rrnB ribosomal RNA transcription terminators (Brosius

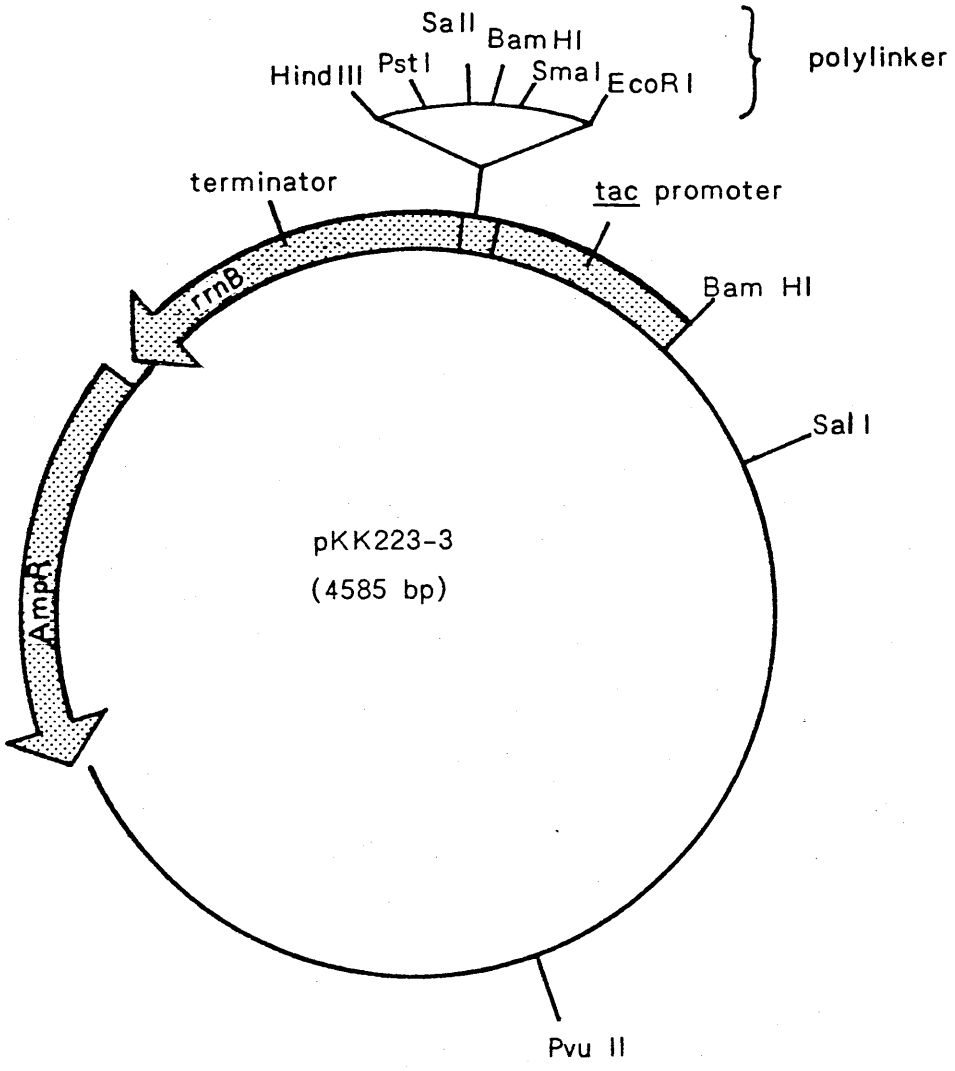


Figure 3.3: Expression plasmid pKK223/3.

et al., 1981a; Brosius et al., 1981b).

Plasmid pJB14 was constructed by insertion of a 2.5 kbp MspI partial digestion product of pLC 29-47 into the unique but modified EcoRI site in the vector. Plasmid pKK223/3 was linearized by digestion with EcoRI; the 5' protruding ends filled in by DNA polymerase, and ClaI linkers ligated. Following digestion with ClaI the vector was used to clone the aroB gene from pLC 29-47. Partial digestion of pLC29-47 was carried out with the four-base recognizing enzymes TaqI and MspI and restriction fragments of 2-4 kbp were recovered from a LMT agarose gel. The vector ClaI cohesive ends are complementary to those generated by digestion with either TaqI or MspI. The gel-purified DNA fragments were ligated into the ClaI site (formerly EcoRI site) of pKK223/3, and recombinants carrying an intact aroB gene identified by relief of auxotrophy of E.coli AB2826.

Only MspI-derived constructs gave rise to $AroB^+$ phenotypes, no colonies with the Amp^r $AroB^+$ phenotype were obtained from insertion of TaqI fragments. One $AroB^+$ recombinant pJB14 which contained a 2.5 kbp MspI insert had approximately 1000 times more DHQ synthase activity when fully induced, compared with E.coli K12 levels (Frost et al., 1984).

Both plasmids, pKD106 and pJB14, share the ability to relieve the auxotrophic requirements of E.coli AB2826 (aroB). Both are derived from pLC29-47 and therefore a comparison of their respective genomic inserts was the first step in elucidating the structure of the aroB gene.

3.3 Structural organisation of the aroB gene

3.3.1 Comparative aspects of pKD106 and pJB14

Large scale preparations of pJB14 and pKD106 were made (Section 2.12) and a detailed restriction map of both plasmids was constructed. A series of single and double restriction enzyme digests were performed and systematically a restriction profile was built up. It became obvious that the genomic inserts of both plasmids shared a number of properties. Several restriction enzyme sites, both numerically and in terms of spatial arrangement, were common to both. In particular, digestion with HincII gave rise to bands at ca. 560 bp and ca. 280 bp for both plasmids (Figure 3.4(a)).

A comparison of the AvaI restriction enzyme profile of plasmids pLC29-47, pJB14 and pKD106 revealed a common ca. 800 bp fragment located within the genomic insert of each plasmid construct (Figure 3.4(b)).

In the construction of pJB14 (Frost et al., 1984) the EcoRI site of the vector pKK223/3 (Figure 3.3) had been duplicated to flank the ClaI linker recognition sequence. The 2.5 kbp MspI insert of pJB14 is therefore flanked, on either side, by an intact EcoRI site. Digestion of pJB14 with EcoRI gave rise to 3 distinct DNA fragments of 4.6 kbp, 1.65 kbp and 0.85 kbp respectively (Figure 3.4(c)). This indicated that the genomic insert of pJB14 contained a single EcoRI site, asymmetrically located 1.65 kbp and 0.85 kbp from the MspI cloning ends (Figure 3.5).

Figure 3.4 (facing)

(a-c) Agarose gel profiles of *aroB*-plasmid restriction digests. Marker sizes (in bp) are shown on either side of gels. All digests were carried out on 0.3-1.0 µg of DNA under conditions recommended by the suppliers.

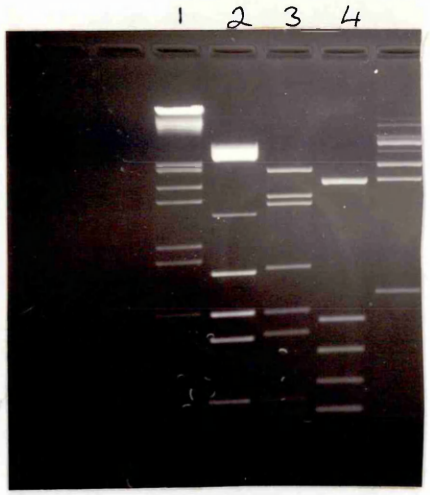
(a) Track 1 *EcoRI*/*HindIII* digest of λDNA (marker).
Track 2 *EcoRI*/*HincII* digest of pKD106.
Track 3 *PvuII*/*HincII* digest of pJB14.
Track 4 *HinfI* digest of pAT153 (marker).

(b) Track 1 *AvaI* digest of pLC29-47.
Track 2 *AvaI* digest of pKD106.
Track 3 *AvaI* digest of pJB14.

(c) Track 1 *EcoRI*/*HindIII* digest of λDNA (marker).
Track 2 *EcoRI* digest of pJB14.
Track 3 *EcoRI*/PvuII digest of pJB14.
Track 4 *HinfI* digest of pAT153 (marker).

Addendum

In this figure, and all others hereafter involving molecular size markers, the marker sizes are not aligned with figure. Rather, they are arranged in the order in which they appear in the 'ladder'.



bp

1900
1570
1320
930
840
580

bp

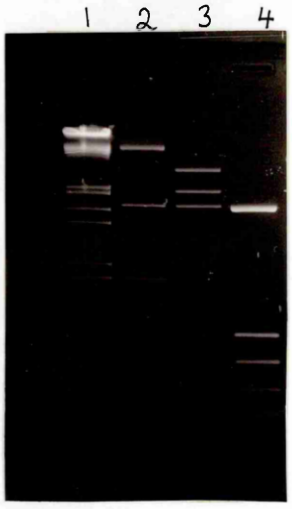
1631

517
396
298
220

(a)

800

(b)



1900
1570
1320
930
840
580

1631

517
396
298
220

(c)

An examination of the detailed restriction enzyme maps of both pKD106 and pJB14 (Figure 3.5) revealed a number of striking similarities. The distribution of specific restriction enzyme sites on the 1.65 kbp EcoRI fragment of pJB14 was identical to the pattern observed within the EcoRI - proximal 1.65 kbp of the 3.6 kbp EcoRI-BglII genomic insert of pKD106 (Figure 3.5).

3.3.2 Preparation of nick-translated aroB probes

Two aroB-specific probes were prepared to ascertain the degree of similarity by hybridisation criteria of the 1.65 kbp region common to both pKD106 and pJB14. A 0.8 kbp AvaI fragment from within the 1.65 kbp 'common' region was selected for both plasmids (pKD106 and pJB14) as the DNA probe (Figure 3.4(b), Figure 3.5 aroB probe (a)).

2 µg of pKD106 and pJB14 were each digested with AvaI and the products separated electrophoretically on a 1% LMT agarose gel. The 0.8 kbp band was excised and purified by phenol/chloroform extraction and ethanol precipitation (Section 2.13.2). This DNA was radioactively labelled *in vitro* using the nick-translation (Kelly *et al.*, 1970) reaction catalysed by *E.coli* DNA polymerase I. Approximately 1 µg of each DNA was labelled and separated from unincorporated radionuclide as described in Sections 2.21.1 and 2.21.2. Radioactively labelled probes were ethanol precipitated twice more and used immediately.

A second probe (Figure 3.5 aroB probe (b)) was prepared as follows. 2 µg of pJB14 was digested with EcoRI and the

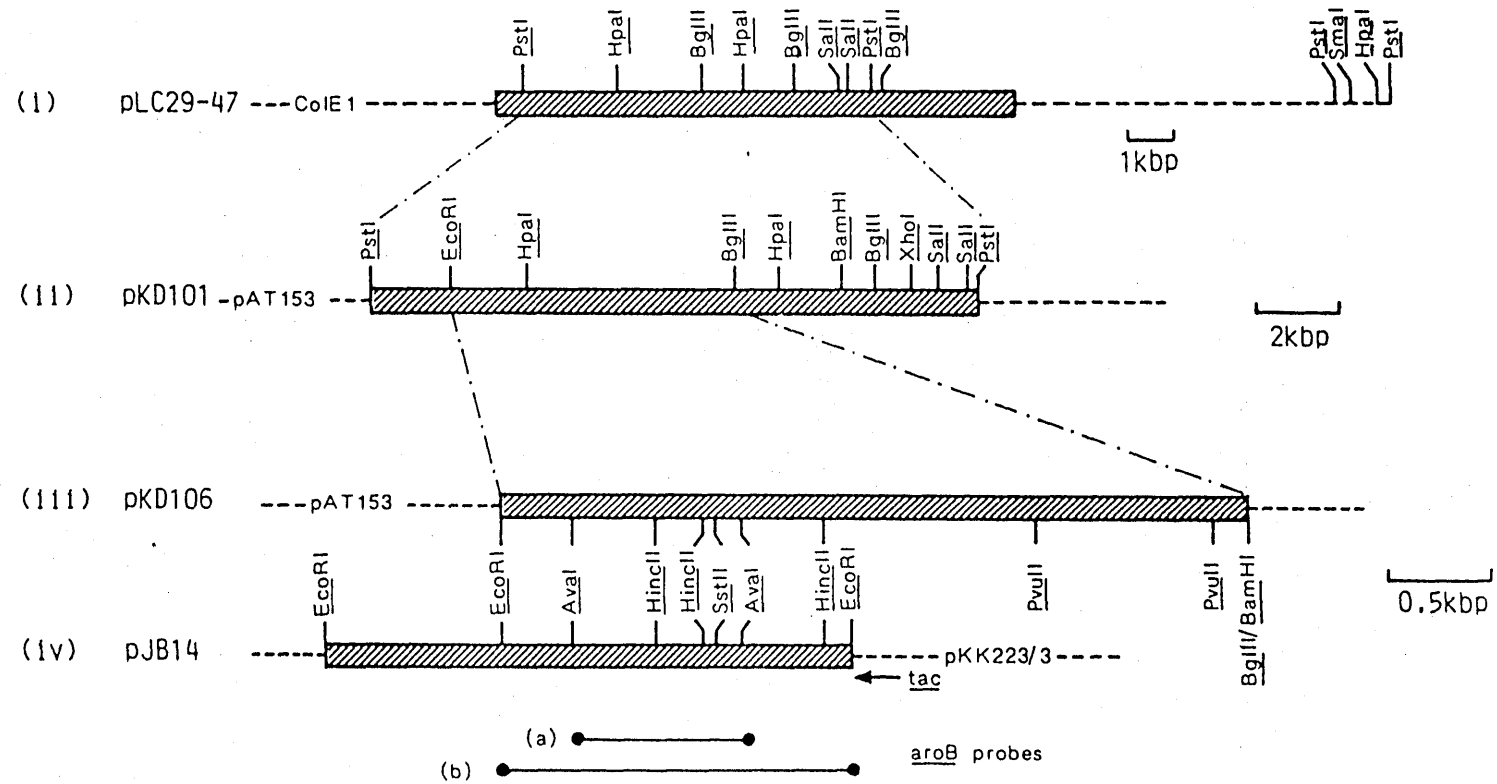


Figure 3.5: Comparative restriction enzyme profiles of the AroB$^+$ plasmids pLC29-47 (i), pKD101 (ii), pKD106 (iii), and pJB14 (iv). Genomic E.coli DNA is boxed and hatched. Plasmid vector DNA is indicated by (----). The regions used as aroB probes (a) & (b) are indicated and discussed in the text.

1.65 kbp EcoRI fragment (Figure 3.4(c)) similarily purified and labelled with dCTP 5'- $[\alpha\text{-}^{32}P]$. The reaction conditions were as described above except higher specific activity dCTP 5'- $[\alpha\text{-}^{32}P]$ (Section 2.21.1) was the incorporated radionuclide.

The specific activity of the ^{32}P labelled probes was determined by TCA precipitation and liquid scintillation counting (Section 2.22).

Probe	sp. activity (dpm/µg)
$^{+}Ava_{106}$	3.2×10^6
$^{+}Ava_{14}$	3.6×10^6
$^{*}Eco_{14}$	1.6×10^8

($^{+}$Figure 3.5 aroB probe (a) prepared from either pKD106 or pJB14 *Figure 3.5 aroB probe (b) prepared from pJB14).

3.3.3 Southern hybridisation analysis of the cloned aroB gene

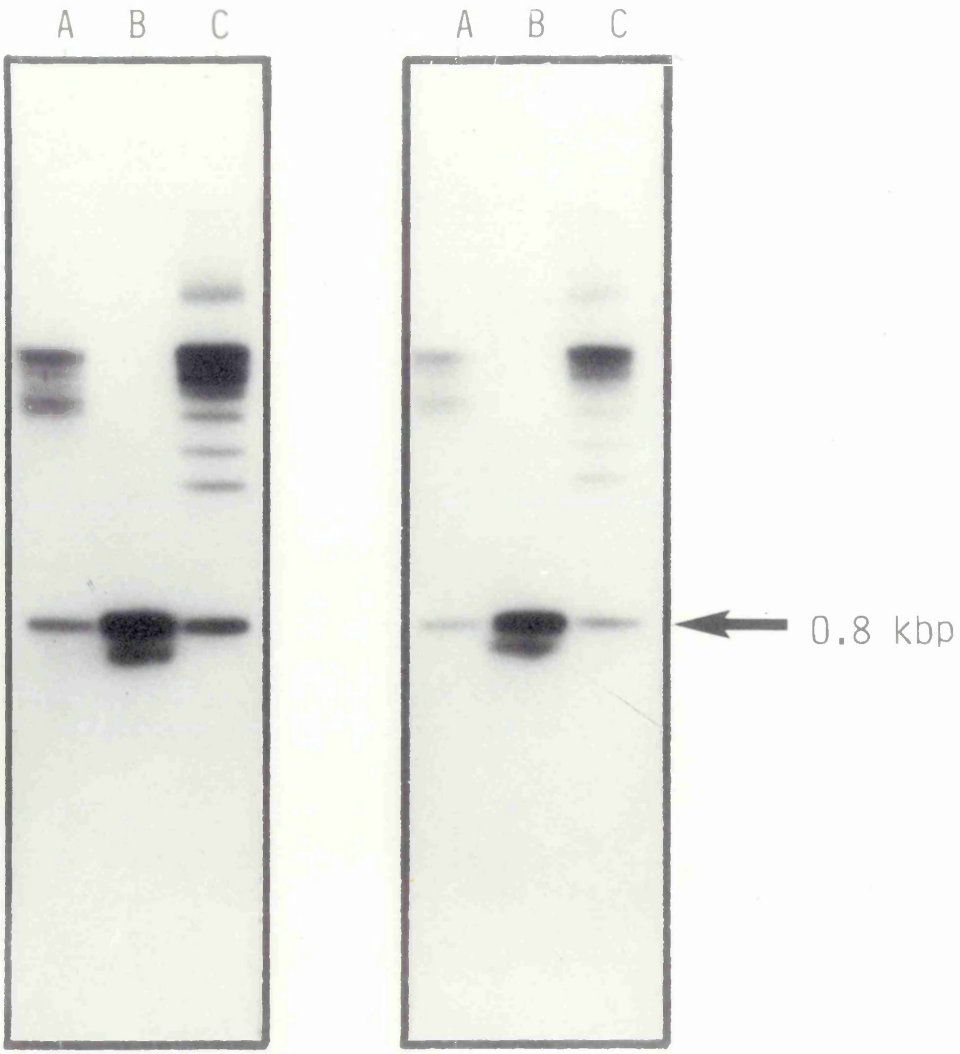
(a) Plasmids pLC29-47, pJB14 and pKD106 were each digested with AvaI and the products separated by gel electrophoresis on 0.8% agarose. Ethidium Bromide staining and U.V. visualisation revealed that only the pKD106 digest had gone to completion (Figure 3.6(a)). The gel was treated as described in Section 2.20.1 prior to blot transfer overnight. The DNA was simultaneously transferred to two sections of nitrocellulose filter (Section 2.20.2). The nitrocellulose was subsequently washed briefly in 2 x SSC, dried at 25°C for 1 hour and baked at 80°C in vacuo for 3 hours.

Each filter was sealed separately in a heat-sealable polythene bag containing 0.2 ml/cm^2 prehybridisation solution

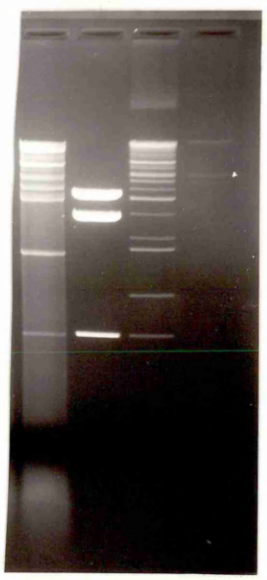
Figure 3.6 (facing):

Lower: (a) 0.8% ethidium bromide stained agarose gel profile of *AvaI* digested:- pLC29-47 (A); pKD106 (B) and pJB14 (C). Note how only the pKD106 digest has gone to completion.

Upper: (b) Southern hybridisation of the cloned *aroB* gene. The gel shown in Figure 3.6(a) below was simultaneously transferred to two nitrocellulose filters and probed with the 0.8 kbp *AvaI* fragment of pKD106 ($\underline{Ava}_{106}$) or pJB14 ($\underline{Ava}_{14}$) as discussed in Section 3.3.3. Tracks A, B & C are as in Figure 3.6(a).



A B C



(a)

-0.8 kbp

(Section 2.20.3). Prehybridisation was at 68°C for 5 hours in a shaking water bath. Hybridisation was carried out as follows: prehybridisation solution was removed and replaced wtih 50 $\mu l/cm^2$ hybridisation solution (Section 2.20.3). Hybridisation solution contained the heat-denatured (100°C, 3 mins) DNA probe. The filter bag was resealed and hybridisation performed overnight at 68°C in a shaking water bath. The following day filters were removed and washed under high stringency conditions (Section 2.20.4). At no time during the washing procedure were filters allowed to dry out. Finally after drying the filters were autoradiographed at -70°C using an intensifying screen.

(b) The hybridisation signal obtained using probes $\underline{Ava}_{106}$ and $\underline{Ava}_{14}$ were identical (Figure 3.6(b)). Both probes cross-hybridise with the other two putative *aroB*-plasmids. The strongest hybridisation signal in both cases is with the 0.8 kbp fragment produced by *Ava*I digestion of pKD106. The intensity of the hybridisation signal can be directly correlated with the *amount* of DNA loaded on the gel (Figure 3.6(a)), where pKD106 is overloaded with respect to pLC29-47 and pJB14. In addition the pattern of partial digestion products of *Ava*I digests of pJB14 and pLC29-47 also provide identical hybridisation patterns (Figure 3.6(b)).

These results, obtained using high stringency hybridisation conditions, strongly suggest that the 0.8 kbp *Ava*I region of the common 1.65 kbp overlap of plasmids pKD106 and pJB14 are identical. The similar pattern of restriction sites within this region of both clones (Figure 3.5) supports this hypothesis. The location of the *aroB* structural gene between

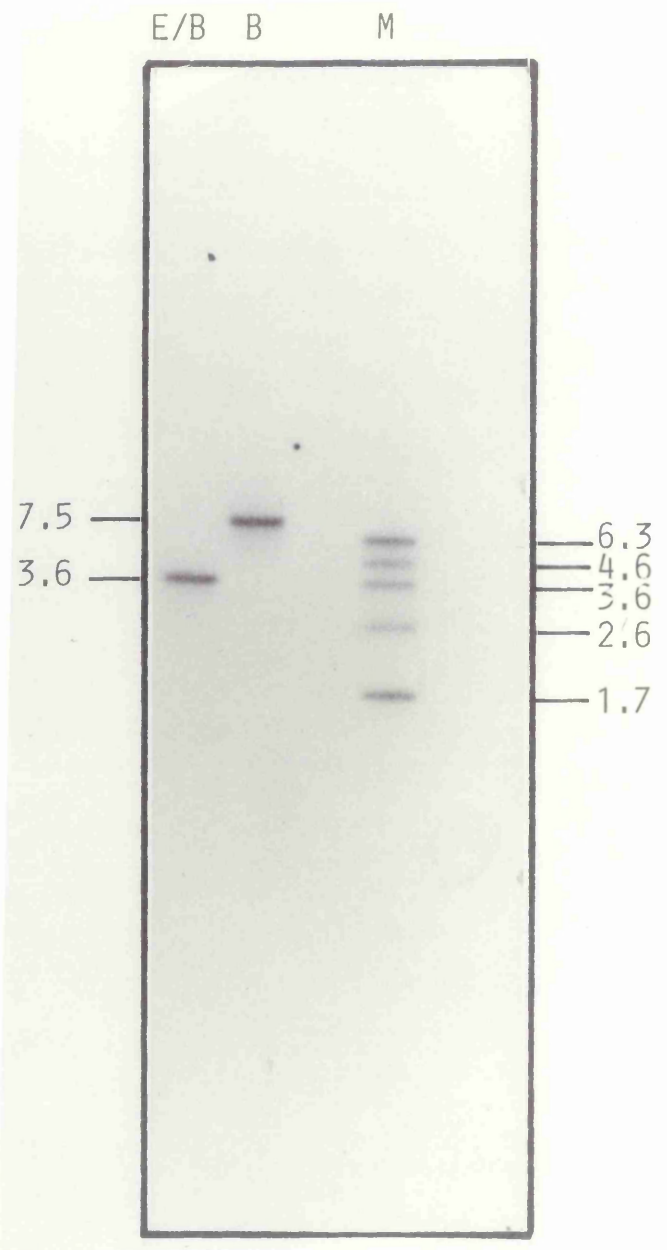


Figure 3.7: Southern hybridisation of the genomic E.coli aroB gene. DNA was digested with BglII (B) or EcoRI and BglII (E/B) and probed as described in Section 3.3.4. Marker (M) fragments of the shown molecular size (kbp) were also detected.

the 1.65 kbp-separated EcoRI sites of pJB14 was further examined by sub-cloning.

3.3.4 Southern hybridisation analysis of the genomic aroB gene

(a) E.coli K12 high molecular weight DNA (a gift of Mr S. Granger) was digested (3 µg) with EcoRI or with EcoRI and BglII. The products were separated electrophoretically on a 0.8% agarose gel. In a separate track a series of restriction digest fragments of pJB14, at very low concentration, were also electrophoresed. These digest products served as useful size markers for the hybridisation profile. The gel was treated and a blot transfer set up as described in Section 3.3.3(a). Hybridisation and prehybridisation conditions were as described (Section 3.3.3(c)) and the nitrocellulose filter probed with ^{32}P-labelled 1.65 kbp EcoRI fragment of pJB14 prepared as described in Section 3.3.2 (Figure 3.5 aroB probe (b)). Following overnight hybridisation the filter was washed under high stringency conditions (Section 2.20.4) and the hybridisation signal detected by autoradiography.

(b) The hybridisation pattern observed when the 1.65 kbp EcoRI fragment of pJB14 (Figure 3.5 aroB probe (b)) was used to probe the genomic structural organisation of the aroB gene is shown in Figure 3.7. The signal produced indicates that this 1.65 kbp fragment, originally derived as a 2.5 kbp MspI partial digest (Frost et al., 1984, Section 3.2.2), is located on a 3.6 kbp BglII/EcoRI region, and on a larger (>6 kbp) BglII region of the chromosome. Examination of the

restriction map of pLC29-47 (Takeda et al., 1981, Figure 3.2) supports these data. The 3.6 kbp EcoRI/BglII fragment of pLC29-47 cloned in pKD106 (Duncan & Coggins, 1983) is located between one end of the plasmid vector cloning end of pLC29-47 and an internal BglII site some 5 kbp away (Figure 3.2). The observed hybridisation signal at >6 kbp (Figure 3.7) suggests that in the construction of pLC29-47 (Clarke & Carbon, 1976), the left hand end of the genomic insert was generated from a site 1 kbp inside this distal BglII site.

The justification for examining the chromosomal organisation of the aroB gene is straightforward. Plasmids pLC29-47, pKD106 and pJB14 were all selected for complementation of aroB$^-$ E.coli AB2826 (recA$^+$). It is vital to ensure that the integrity of the chromosomal aroB structural organisation has been maintained on these plasmids. Recombination of any of these putative aroB constructs with the defective chromosomal marker of E.coli AB2826 could very easily be identified in retransformation and overexpression controls (Section 3.4.2). However more subtle rearrangements adjacent to the aroB gene, not directly affecting expression (complementation) per se, may have gone undetected if not examined.

The genomic Southern hybridisation results suggest that the structural arrangement of the 3.6 kbp EcoRI/BglII fragment and, by definition, the related 1.65 kbp EcoRI fragment are genuine reflections of the genomic organisation. The activity carried within this 1.65 kbp region which is

capable of complementing an AroB$^-$ auxotroph (putatively the *aroB* structural gene) is a *bona fide* region of the *E.coli* K12 chromosome at around minute 74.

3.4 Sub-cloning of the *aroB* gene

3.4.1 Construction of pGM107 and pGM108

Plasmid pJB14 (2 μg) was digested with *Eco*RI and the products separated on 1% LMT agarose gel. The 1.65 kbp fragment (Figure 3.4(c)) was excised and the DNA phenol/chloroform purified. This fragment was ligated into the *Eco*RI site of dephosphorylated pAT153 or pKK223/3 (Sections 2.14 and 2.15). The ligation mix was used to transform competent *E.coli* AB2826 and the transformation mix plated on LA +*amp* (Section 2.16). Following overnight growth at 37°C, colonies were replica plated on to minimal medium and incubated overnight at 37°C. The results of these transformations are summarised in Table 3.1.

All of the colonies with the AroB$^+$ phenotype (growth on minimal medium) were retested for *amp*^{r}. In total 105 AroB$^+$ colonies were replica-plated on to LA +*amp*, minimal medium and, to distinguish pKK223/3- and pAT153-derived clones, LA + *tet*. As expected all the putative positive clones resulting from cloning into pKK223/3 ('pGM107') were *amp*^{r} *tet*^{s} AroB$^+$. The pAT153-derived clones ('pGM108') were *amp*^{r} *tet*^{r} AroB$^+$ (Table 3.1).

Ligation	No. of Colonies on LA + amp	Replica plating	
		No. of Colonies on MM	No. of Colonies on LA + tet
'pGM107'	ca. 200	42	0
'pGM108'	ca. 150	63	63
ligation control	ca. 600	0	ca. 600
no DNA	0 (-amp)	0	not done
10ng pAT153	ca. 1,000	0	ca. 1,000

Table 3.1: Transformation and complementation of E.coli AB2826.

3.4.2 Controls in relief of auxotrophy selection

In selecting for complementation (rescue) of an auxotrophic marker a number of vital controls must be included.

(1) Retransformation The plasmid of interest may indeed have originally carried an intact copy of the gene sought. However in RecA$^+$ strains (e.g. E.coli AB2826) the possibility of a recombination event between the defective chromosomal marker and the functional plasmid-borne determinant cannot be ignored. Complementation under these circumstances would be by expression of a now intact chromosomal allele. This possibility can be excluded by retransforming the auxotroph with a crude preparation of the recombinant plasmid. It is necessary to establish that the ability to complement the auxotrophic requirements of the mutant strain is a plasmid-encoded (transformation dependent) entity by passage through several such challenges.

(2) Reversion The frequency of reversion to prototrophy of the aro$^-$ strains used in this laboratory is sufficiently low at $<10^{-9}$ (I.A. Anton, personal communication) to effectively exclude this possibility. However reversion (no DNA) and ligation (only vector DNA) controls have been regularly performed in parallel with (re)transformation experiments. The reversion controls are plated directly on to LA -amp and on to minimal medium. On no occasion to date has a genuine E.coli colony appeared on either type of plate even after prolonged incubation. The ligation controls are plated on to LA+amp and the resulting colonies replica plated on to minimal medium. If the ligation control is

such that no colonies are expected on LA+amp (e.g. very efficient dephosphorylation of vector 5'-phosphate groups or non-cohesive ends) then the competent cells are transformed with ccc.pAT153. To date only amp^r Aro^- phenotypes have been observed.

(3) Overexpression The nature of the mutations in the aro^- strains is largely undefined. As a consequence the possibility exists that the 'complementation' observed could result from suppression of the mutant phenotype by a plasmid encoded suppressor. If the mutation in the aroB gene is a nonsense mutation then a clone carrying a nonsense suppressor could theoretically restore the wild-type phenotype. One can exclude this possibility by establishing that transformed strains overexpress enzyme activity (Section 2.6). The demonstration that the level of overexpression is related to plasmid copy number strongly indicates that the structural gene, encoding the enzyme activity in question, has been cloned.

3.4.3 Characterisation of pGM107 and pGM108

Ten colonies of each of the putative pGM107 (1.65 kbp EcoRI clone in pKK223/3) and pGM108 (1.65 kbp EcoRI clone in pAT153) were selected for further study. Rapid preparations of plasmid DNA were made from 10 ml LB+amp overnight cultures of each (Section 2.11). Digestion with EcoRI showed that each contained a plasmid with a 1.65 kbp insert at the EcoRI site. Three pGM107 and three pGM108 isolates were characterised further. The asymmetric location of the pair of AvaI sites within the 1.65 kbp insert was used to orientate the clones.

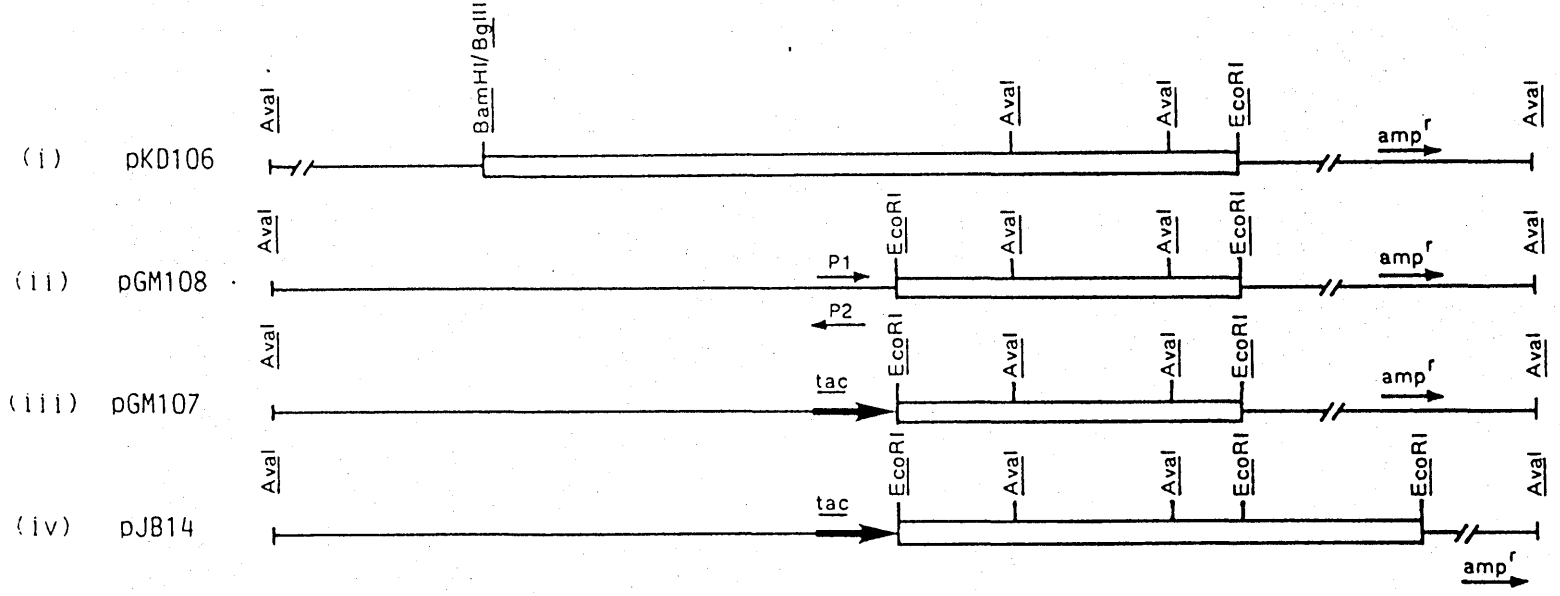


Figure 3.8: AroB$^+$ plasmids pGM107 and pGM108. E.coli genomic DNA is boxed and plasmid vector DNA represented by solid lines. The orientation w.r.t. the vector promoter and/or antibiotic resistance gene is indicated.

Plasmids pGM107(a) and pGM107(c) both contained a 1.65 kbp EcoRI insert in the opposite orientation, with respect to the vector tac promoter, to pJB14. Plasmid construct pGM107(b) had a 1.65 kbp insert at the EcoRI site of pKK223/3 in the same orientation as pJB14. Plasmids pGM108(b) and pGM108(c) had a 1.65 kbp EcoRI insert in pAT153 in the opposite orientation, w.r.t. the vector tet P2 promoter, to pKD106. Isolate pGM108(a) contained the same 1.65 kbp EcoRI insert in pAT153 in the same orientation as pKD106 (Figure 3.8). Recombinant plasmids pGM107(b) and pGM108(a) were selected for further study and designated pGM107 and pGM108 respectively.

3.4.4 Level of 3-dehydroquinate synthase activity in crude extracts

E.coli K12, E.coli AB2826 and E.coli AB2826 transformed with the plasmids pKD106, pJB14, pGM107 and pGM108 were examined for their relative levels of DHQ synthase activity.

Crude extracts were prepared from stationary phase cultures (A_{650} = 0.9 - 1.0) as described in Section 2.6. Cells were grown in 100 ml minimal medium containing ampicillin (50 µg/ml), with the exception of E.coli K12 and E.coli AB2826. The former was grown on minimal medium alone and the latter on minimal medium containing the aromatic supplement (Section 2.4.3).

DHQ synthase activity was measured by coupling the release of DHQ to 3-dehydroquinase which converts DHQ to 3-dehydroshikimate. DHS absorbs strongly at 234 nm (Section 2.7.1) and therefore spectrophotometric quantitation of the

DAHP-initiated change in absorbance at 234 nm can be used to calculate the level of DHQ synthase activity. At 234 nm there is a moderately high background absorbance increase which is independent of DAHP addition. This must be subtracted to give the true enzyme activity. The low levels at which DHQ synthase occurs in E.coli K12 make accurate measurement of the wild-type level difficult. However DAHP-dependent DHQ synthase rates can be calculated for the wild-type enzyme by carefully diluting enzyme samples until the signal:noise ratio is favourable. Once a reproducible and statistically valid rate is obtained for the wild-type level then levels of DHQ synthase overexpression can be quantitated for transformed strains.

Checks on the authenticity of the observed DHQ level (for example its metal dependence, as discussed in Section 2.7.1) were routinely carried out. Table 3.2 summarises the results obtained.

All of the cultures were inoculated with a fresh overnight culture to a final A_{650} = 0.05. In addition duplicate E.coli AB2826/pJB14 and AB2826/pGM107 cultures were maintained. At A_{650} = 0.25 IPTG was added to one of each of these cuplicates for a final concentration of 5×10^{-4}M. The growth rates of all of the cultures were similar. Addition of IPTG had no effect on the growth rates of E.coli AB2826/pGM107 or AB2826/pJB14 compared with their respective control cultures.

No DHQ synthase activity could be detected in E.coli AB2826. In the plasmid transformed strains, the level of

E.coli strain	Crude extract DHQ synthase s.a (units/mg x 10^5)	Crude extract 3-dehydroquinase s.a (units/mg x 10^5)	Level of overexpression w.r.t. K12
AB2826	0	5.0	N/D
K12	1.7	6.4	1
AB2826/ pKD106	33	7.2	19.4
AB2826/ pJB14	63	6.1	37.0
AB2826/ pGM107	65	8.5	38.2
AB2826/ pGM108	27	7.4	15.8
AB2826/ pGM107 (+ IPTG)	98	6.9	57.6
AB2826/ pJB14 (+ IPTG)	102	8.8	60.0

NOTES: 1. 3-hydroquinase activity was measured as an internal control.

2. (+ IPTG) indicates addition of IPTG to 5 x 10^{-4}M final concentration at A_{650} = 0.25.

Table 3.2: Overexpression of DHQ synthase activity.

DHQ synthase activity was elevated 15- to 60-fold with respect to the level in wild-type E.coli K12. The level of 3-dehydroquinase activity was measured as an internal control for each culture. The observed 3-dehydroquinase specific activity varied over a narrow range, typically 5.0 - 8.8 x 10^{-5} units/mg.

The pAT153-derived aroB constructs, pKD106 and pGM108, had similar levels of overexpression (19.4 and 15.8-fold respectively). This level is consistent with that observed by Duncan & Coggins (1983) for E.coli AB2826/pKD106. The tac-aroB constructs, pJB14 and pGM107, exhibited 37- and 38.2-fold (non-inducing conditions) and 60- and 57.6-fold (partially inducing conditions) elevated levels of overexpression respectively. This IPTG-induced four-fold elevation, and IPTG-independent doubling of levels of activity relative to their pAT153-aroB construct counterparts, pKD106 and pGM108, illustrates two main points. Firstly addition of IPTG at early to mid-logarithmic growth phase (A_{650} = 0.25) resulted in a two-fold increase in DHQ synthase overexpression. These partially inducing conditions demonstrate that the inducible tac promoter is functional in these constructs. Secondly the levels of overexpression under non-inducing conditions were also two-fold higher in tac constructs (pJB14 and pGM107) than levels observed for pAT153-aroB clones pKD106 and pGM108. This result suggests that in E.coli AB2826 transformed with the tac-aroB plasmids the lac repressor is being titrated out by the large number of lac operators thus allowing escape synthesis of DHQ synthase from the tac promoter.

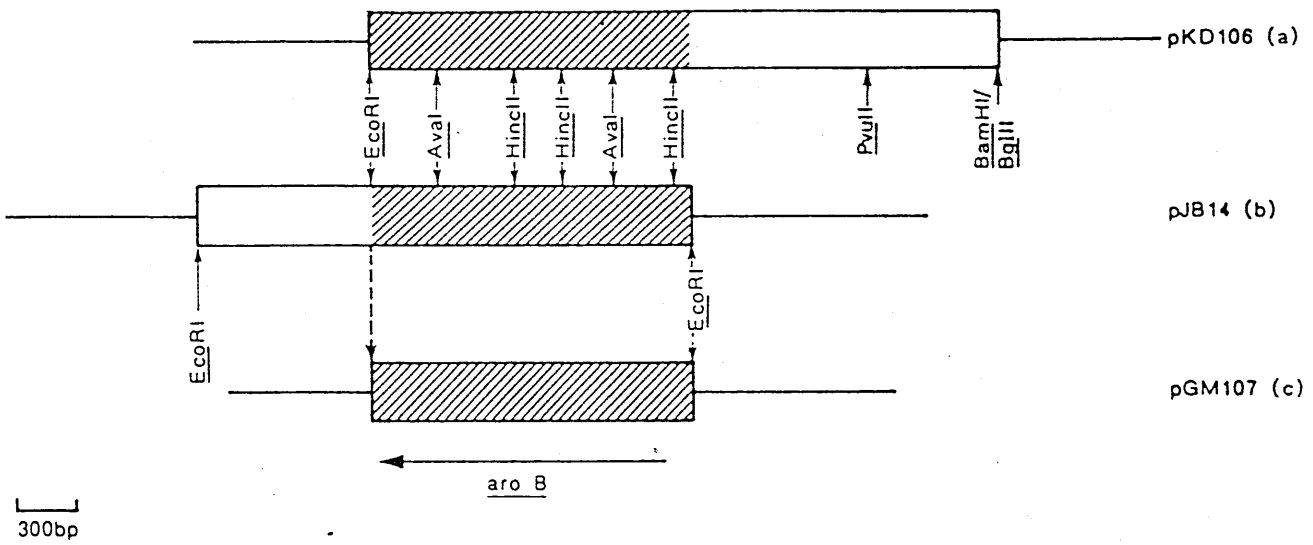


Figure 3.9: Subcloning of the aroB gene.

E.coli genomic DNA is indicated by boxing, plasmid vector DNA shown as solid lines. The region common to both (a) pKD106 and (b) pJB14, and subcloned in (c) pGM107 is shown by hatching. The direction of expression of aroB is indicated.

3.4.5 Conclusion

Sub-cloning of the 1.65 kbp EcoRI fragment of pJB14 to generate plasmids pGM107 and pGM108 has identified the putative aroB coding region. Removal of the non-overlapping 2 kbp of genomic DNA found in pKD106 or the 0.85 kbp of E.coli DNA found in pJB14 does not affect either complementation or the level of DHQ synthase overexpression (Figure 3.9).

The direction of transcription of aroB has been established by cloning the gene downstream of a tac promoter, and inducing increased expression with IPTG. The sub-cloning figure strategy is outlined in Figure 3.9. Relevant restriction sites are shown for comparative purposes.

3.5 Identification of protein products of the cloned insert of pGM107

3.5.1 Potential problems in the interpretation of sub-cloning results

The sub-cloning of a 1.65 kbp EcoRI fragment to give pGM107 (Section 3.4) was definitive only in identifying the limit similarity between pKD106 and pJB14 required to maintain complementation of E.coli AB2826. It is possible that either removal of the 2 kbp fragment unique to pKD106 or the 0.85 kbp fragment unique to pJB14 may result in production of an altered polypeptide. The former is unlikely since expression of pGM107 in tac vector pKK223/3 (Section 3.4.4) identifies the sequences unique to pKD106 as being 5' to the aroB coding region. Truncation of the gene at this end would almost certainly affect expression and the resulting poly-

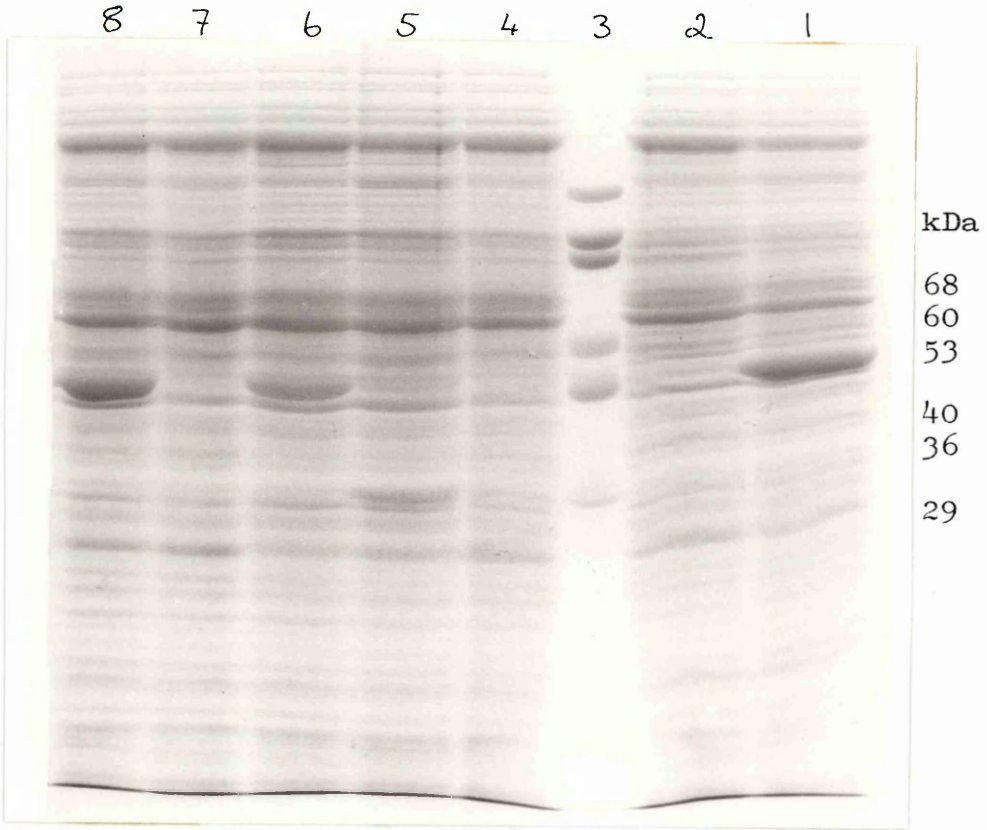


Figure 3.10: Coomassie stained, 12.5% SDS PAGE analysis of E.coli crude extracts as described in Section 3.5.2. E.coli AB2826 (ca. 50 µg protein) transformed with pGM107 (track 1); pKD106 (track 6); pJB14 (track 8). Untransformed E.coli AB2826 (tracks 4 & 7); E.coli K12 (track 2). Track 5 is a crude extract of E.coli AB2826 transformed with a non-aroB plasmid. Track 3 shows molecular weight marker proteins, sizes shown at the side.

peptide would lack part of the N-terminal sequence and possibly have great difficulty in folding properly.

Truncation arising from removal of the 0.85 kbp fragment unique to pJB14 could produce a C-terminal 'clipped' protein. Complementation may still be achieved but by an altered and kinetically less competent polypeptide. The overexpression results in crude extracts tend to dismiss this possibility (Section 3.4.4, Table 3.2), however analysis by SDS PAGE of the crude extracts was undertaken to confirm that an overproduced protein of the same molecular weight is still expressed in successively smaller subclones (pKD106, pJB14, pGM107).

3.5.2 SDS PAGE analysis of crude extracts

Samples from several of the crude extracts described in Section 3.4.4 were analysed by denaturing polyacrylamide gel electrophoresis (SDS PAGE). Samples of crude extract (containing approximately 50 µg protein) of *E.coli* K12, *E.coli* AB2826, *E.coli* AB2826/pJB14, *E.coli* AB2826/pGM107 and *E.coli* AB2826/pKD106 were separated on a 12.5% polyacrylamide gel in the presence of SDS (Section 2.8). The gel was stained with Coomassie Blue, destained overnight and photographed (Figure 3.10).

A band at M_r 36-38,000 can be seen in tracks 1, 6 and 8 (Figure 3.10) corresponding to the crude extracts of strains transformed with plasmids pGM107, pKD106 and pJB14 respectively. No such heavily stained band is present in tracks 2, 4, 5 or 7. Tracks 4 and 7 are samples of untransformed *E.coli* AB2826.

Track 2 is E.coli K12 and track 5 a crude extract of a non-aroB plasmid transformed E.coli AB2826.

It seems very likely that the heavily stained protein band seen in Figure 3.10 is E.coli DHQ synthase. Also it is clear that the overexpressed polypeptides in cells transformed with plasmids pKD106, pJB14 and pGM107 have the same molecular weights in all three cases. Frost et al. (1984) purified DHQ synthase from a pJB14 transformed strain with a subunit M_r 38,000-40,000. The overproduced protein band in Figure 3.10 has a similar molecular weight.

3.5.3 In vitro transcription-translation of pGM107

The bacterial cell-free coupled transcription-translation system first described by De Vries and Zubay (1967) was used to analyse the protein products of the cloned insert of pGM107. Expression in vitro of genes contained on a bacterial plasmid can proceed provided that the relevant control features are present (e.g. Pribnow box for initiation of transcription, Shine-Dalgarno sequence for translation). Plasmids pGM107 and pAT153 were analysed using the DNA-directed coupled transcription-translation system described in Section 2.26.

2.5 µg of pAT153 and pGM107 were each used as template for the incorporation of radioactive label (L-[^{35}S] methionine) into protein (Section 2.26.1). The degree of incorporation was monitored using TCA precipitation of protein products and scintillation counting (Section 2.26.2). A sample of each was analysed by SDS PAGE and autoradiography (Figure 3.11).

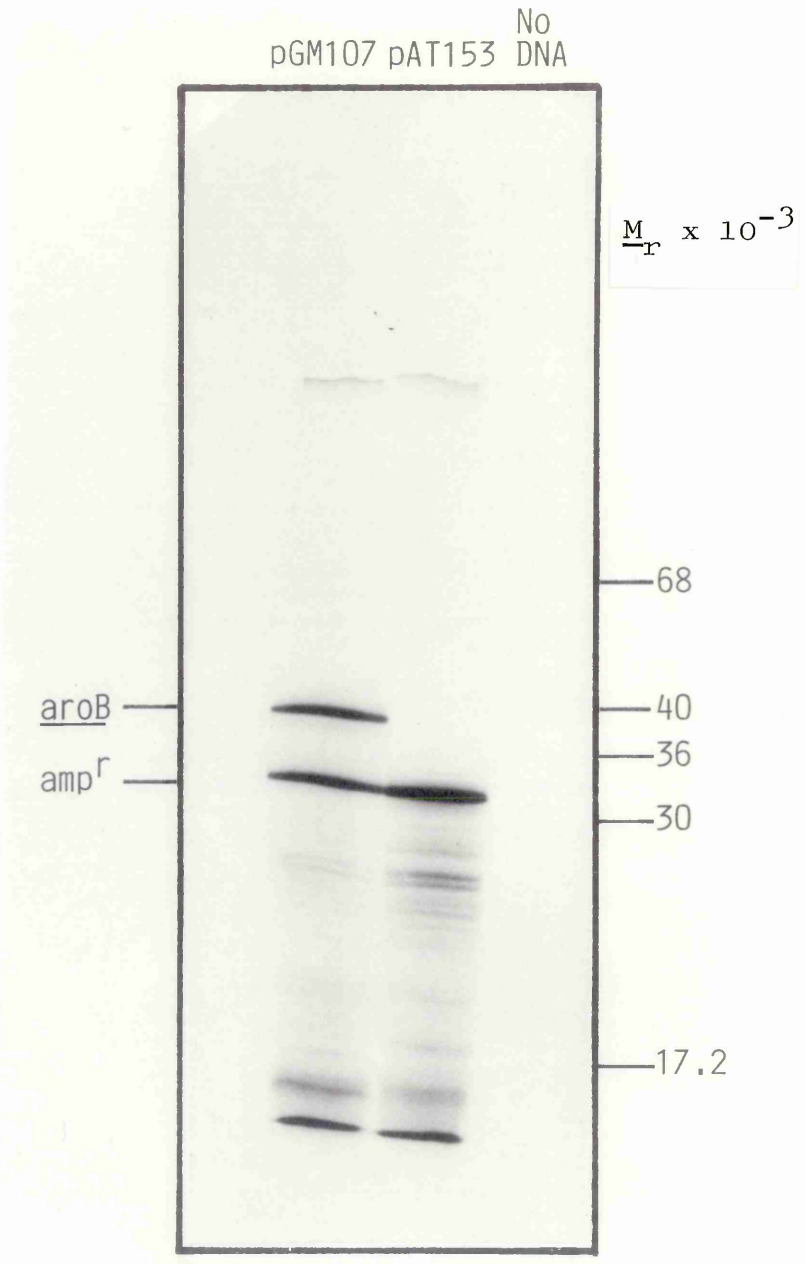


Figure 3.11: In vitro expression of pGM107 and pAT153 as described in Section 3.5.3. Incorporation of ^{35}S-Met into DHQ synthase (aroB) and vector-encoded β-lactamase (amp^r), as detected in this autoradiogram, is shown. Molecular weights ($\times 10^{-3}$) are indicated.

The 12.5% polyacrylamide gel used to fractionate the protein products (gel not shown) was aligned with the resultant autoradiograph. Molecular weight markers on the gel were marked off to calibrate the autoradiograph (Figure 3.11).

A sample containing no DNA was used as an internal control of the cell-free system itself. No bands are visible in this control lane of Figure 3.11. Plasmid encoded protein products are clearly visible in lanes marked pAT153 and pGM107. A comparison of pAT153- and pGM107-directed protein synthesis reveals only one difference, a band migrating at a molecular weight between 36 and 40,000.

Plasmid pGM107 is a pKK223/3-*aroB* construct (Section 3.4.1). A comparison of the pattern of protein products expressed by vector sequences in pGM107 and pAT153 show how closely related the *tac* expression vector is to pAT153. It would appear therefore that the cloned insert of pGM107 is expressing a protein of M_r 37-39,000. A band can be observed in both pGM107 and pAT153 tracks in Figure 3.11 at M_r ca. 30,000. This is the beta-lactamase protein, the product of the vector-contained *bla* gene which confers ampicillin resistance to recipient cells.

3.5.4 Conclusion

The cloned insert of pGM107, a 1.65 kbp *Eco*RI region of DNA common to plasmids pLC29-47, pKD106 and pJB14, encodes a polypeptide of molecular weight ca. 38,000.

3.6 DNA sequence analysis of the aroB gene

3.6.1 Sequencing strategy

3.6.1A What to sequence and how?

Evidence presented in Sections 3.4 and 3.5 suggested that the 1.65 kbp EcoRI fragment which forms the cloned insert of pGM107 contains the aroB gene. In order to determine the protein coding sequence of the aroB gene this region of DNA was fully sequenced. The method employed was the M13/dideoxy sequencing strategy. Techniques utilised in the construction of recombinant M13 clones, template preparation and sequencing reactions are dealt with in Section 2.17.1 to 2.17.12.

The 1.65 kbp EcoRI insert of pGM107 was sequenced by breaking it down into smaller, more manageable sections. These were individually sequenced and gradually the jig-saw was reassembled to reveal the entire DNA sequence. A pre-requisite of this, and any, DNA sequencing project is that the DNA must be fully sequenced on both strands. In addition, in rebuilding a framework of small sequences, all restriction sites used in cloning must be overlapped to confirm un-ambiguously the continuity of the final sequence.

3.6.1B Outline of the sub-cloning approach used

The 1.65 kbp insert in pGM107 was examined for convenient restriction sites that could readily be used in sub-cloning (Figure 3.5). The centrally-located pair of AvaI sites appeared ideally suited for sub-cloning. However after several unsuccessful attempts to clone this fragment into M13 it was assumed that the specificity of the AvaI recognition

sequence(s) was incompatible with the M13 RF AvaI sequence. AvaI has a degenerate recognition sequence and cleavage with AvaI can generate combinations of cohesive termini which are non-ligatable. Subsequent sequence analysis showed this to be the case for ligation of the aroB AvaI sites into M13 cleaved with AvaI (Section 3.6.7).

The entire 1.65 kbp EcoRI fragment could be readily cloned, as could the smaller HincII fragments (ca. 250 bp and 550 bp) which have blunt-ended termini. It was decided therefore to sequence in from either end of the EcoRI fragment and to sequence the internal region of the 1.65 kbp fragment included within the two HincII fragments. This would define the limits of the area to be sequenced and provide an internal reference point from which to 'walk-out' to either end. The completion of the sequence would then depend upon finding other convenient internal sites useful in cloning.

The direction of expression of the aroB gene in pGM107 was known from the sub-cloning results (Section 3.4). For this reason the orientation of the protein product, DHQ synthase, can be correlated with the restriction mapping data. Therefore it is convenient to refer to the 'N-terminal EcoRI site' or '5'-flanking EcoRI site' instead of the tac-proximal EcoRI site. Similarly the 'C-terminal EcoRI site' and '3'-flanking EcoRI site' refer to the tac-distal EcoRI site.

The restriction enzymes HpaII, TaqI and Sau3A all recognise different four-base sequences. The statistical probability of such four-base sequences occurring within a given fragment of DNA is obviously higher than that for

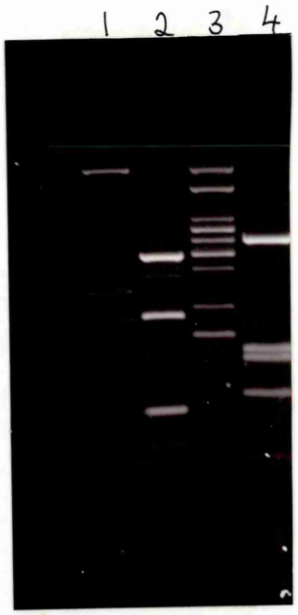


Figure 3.12: 2% Agarose gel profile of secondary restriction digests of the 1.65 kbp cloned insert of pGM107. DNA was prepared for digestion as described in the text (Section 3.6.1C), and digested with TaqI (Track 2), Sau3A (Track 3) or HpaII (Track 4). Size markers from an EcoRI/HinfI pAT153 (Track 1) digest are shown for comparison.

the occurrence of six-base restriction enzyme recognition sites. The occurrence of HpaII, TaqI and Sau3A within the 1.65 kbp EcoRI clone was examined with a view to using these sites to clone small fragments for sequencing.

3.6.1C Distribution of HpaII, TaqI and Sau3A sites within the cloned insert of pGM107

10 µg of pGM107 was digested with EcoRI and the products separated by electrophoresis on a 1% LMT agarose gel. The 1.65 kbp band, corresponding to the cloned insert, was excised and the DNA purified. This DNA preparation was divided three ways and digested with either HpaII, TaqI or Sau3A. The products were analysed by electrophoresis of 50% of the digestion products on a 2% agarose gel. The remaining 50% of each digest was phenol/chloroform extracted and ethanol precipitated prior to use in a subsequent cloning step (Sections 3.6.3 to 3.6.5).

Examination of the digestion patterns (Figure 3.12) indicated that there were internal sites for all three enzymes. Digestion with HpaII produced five distinct bands (Figure 3.12) suggesting four internal HpaII sites. The size distribution ranged from ca. 70 bp to ca. 800 bp. Two of these five bands would be expected to contain an EcoRI sticky-end and it would be impossible to clone these fragments intact into vector specifically prepared to accept HpaII cohesive ends (AccI cut mp8).

The interpretation of the digestion pattern produced by TaqI was more complicated (Figure 3.12). There were

prominent bands corresponding to fragments of ca. 700 bp, ca. 400 bp and ca. 200 bp but this left ca. 350 bp unaccounted for. There may be a number of very small TaqI fragments located either internally or close to the EcoRI ends. Similarly the digestion with Sau3A (Figure 3.12) suggested two, possibly three, internal Sau3A sites. In deciding which of these enzymes to use to generate small sub-regions to be sequenced a number of factors were considered.

Firstly at least two mini-libraries of clones derived from 2 of the 3 four-cutter enzymes would have to be constructed to achieve sufficient overlaps. Secondly the construction of the third mini-library of recombinant clones should greatly facilitate the identification of overlapping sequences. The benefits to be gained from cloning all three mixtures seemed to outweigh the expenditure of time and effort involved in doing so. Therefore all three four-base recognising enzymes were used to construct mini-libraries of recombinant M13 clones (Sections 3.6.3 to 3.6.5), each containing different sub-sections of the 1.65 kbp genomic material cloned in pGM107.

3.6.2 1st round of DNA sequencing (non-random)

3 µg of pGM107 was digested with either EcoRI or HincII and the products separated on a 1.5% LMT agarose gel. A 1.65 kbp EcoRI fragment and a 280 bp HincII fragment were excised and the DNA purified from each band (Section 2.13.2). These fragments were cloned directly into EcoRI-cut or SmaI-cut M13 mp8 RF vector DNA respectively. Recombinant M13

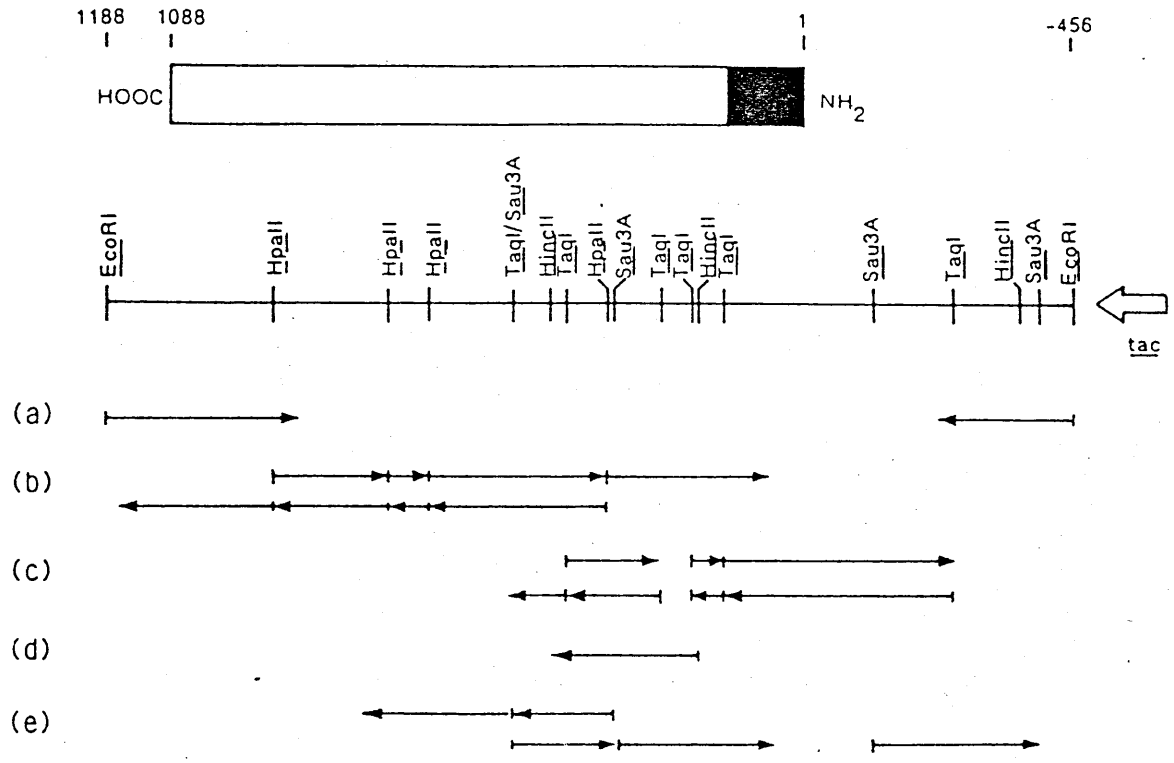


Figure 3.13: DNA sequencing strategy for the aroB gene. Each arrow indicates the direction and extent of a sequenced M13 recombinant clone. The DHQ synthase coding region is boxed and the region confirmed by amino acid sequencing filled. (a) → (e) refer to the various classes of sub-clones described in the text.

clones were identified and single-stranded template DNA prepared for six separate recombinants of each class (Sections 2.17.1 to 2.17.5). All twelve clones were sequenced (Sections 2.17.6 to 2.17.9).

All six HincII recombinants were in the same orientation and the entire 248 bases could be unambiguously identified. The sequence revealed three internal TaqI sites, one HpaII site and one Sau3A site. This region therefore served as an ideal internal foot-hold from which to extend the sequence data.

There were two classes of EcoRI sequences within the six EcoRI recombinants, clearly resulting from the insertion in opposite orientation of the cloned fragment. Sequence analysis revealed a HincII site 89 bases from the EcoRI site in one of these classes of recombinant. This helped to orientate the sequence data as this HincII site was known to be located closest to the tac promoter (Figure 3.8). Also within this recombinant clone were Sau3A and TaqI sites 62 bases and 200 bases respectively from the EcoRI site. The other class of EcoRI recombinant was sequenced for 330 bases in total. Examination of the sequence revealed only a single HpaII site 281 bases from the EcoRI end. There were no TaqI or Sau3A sites in this region (Figure 3.13a).

The sequence information gained from this defined cloning procedure provided enough data to orientate the sequence with respect to the known restriction map (Figure 3.8). In addition sites for the 4 base-cutter enzymes to be used for the bulk of the sequencing were identified close to either end and within the 1.65 kbp region.

3.6.3 2nd round of DNA sequencing (HpaII fragments)

The unfractionated mix of HpaII digestion products of the 1.65 kbp EcoRI insert in pGM107 (3.6.1C) was cloned into AccI-cut M13 mp8 RF. Recombinant M13 clones were identified and their DNA prepared for sequencing (Sections 2.17.1 to 2.17.5). In total 24 DNA templates were made. As a primary screening procedure, A-track analysis (Section 2.17.10) identified at least 8 different A-track sequencing patterns. Representative templates from each of these eight classes of HpaII clones were sequenced fully. The DNA sequences obtained were compiled and compared with the sequences already analysed (Section 3.6.2).

Definitive sub-cloning (Sections 3.4.1 to 3.4.5) had identified the direction of transcription of the aroB gene relative to the EcoRI sites used to construct pGM107. The EcoRI site located at the C-terminal end of the encoded polypeptide was 281 bases from an internal HpaII site (Section 3.6.2). In theory therefore this 281 bp EcoRI/HpaII fragment could only be cloned into AccI-cut mp8 if an internal EcoRI concatemer is formed to yield HpaII cohesive ends. One of the 8 classes of HpaII recombinants analysed was the result of such a ligation event. This allowed the sequences obtained from the C-terminal EcoRI site (Section 3.6.2) to be directly verified on its opposite strand (Figure 3.13b).

The distribution of HpaII sites (Figure 3.12) suggested four internal HpaII sites. Complete digestion

with HpaII should therefore produce three fragments with HpaII cohesive ends. Cloning of each of these (in both orientations) would produce only six classes of recombinant M13 clones. The EcoRI concatemer described above forms a seventh class of unique recombinant clone. The eighth class of HpaII-derived clone observed in A-track analysis must also result from a similar EcoRI concatemerisation event.

Compilation of the HpaII sequence data was completed by identifying the three HpaII fragments and obtaining the sequences on both strands. The exact orientation and spatial arrangement of these HpaII fragments could not unambiguously be ascribed from the available sequence data. However this round of sequencing had established more than 50% of the desired sequence.

Interestingly all 4 HpaII sites within the 1.65 kbp cloned insert of pGM107 are located close to the 'C-terminal' EcoRI site. The sequence information obtained up to this stage therefore was heavily representative of only one half of the genomic insert of pGM107 (Figure 3.13b).

From the known location of a HpaII site within the 248 bp HincII fragment (Section 3.6.2) the second class of EcoRI concatemer was identified. The sequence information obtained from this eighth and final class of HpaII recombinant identified a TaqI site 36 bases upstream of HincII site (36 bases closer to the tac proximal EcoRI site).

3.6.4 3rd round of DNA sequencing (TaqI fragments)

The distribution of TaqI sites in the cloned insert of pGM107 was less well defined (Section 3.6.1C). As a result the expected number of variant class of recombinant M13 clones could not be confidently predicted.

The unfractionated TaqI mix prepared as described in Section 3.6.1C was cloned directly into AccI-cut M13 mp8. A total of 36 template DNA samples were prepared for DNA sequence analysis. Only seven distinct sequence classes were found.

Five of the classes contained sequence information already obtained from the HpaII clones (Section 3.6.3) and the HincII clones (Section 3.6.2). The DNA sequence was useful however in overlapping two HpaII fragments and in verifying some DNA sequence on its opposite strand (Figure 3.13c). The remaining two classes of TaqI clones were the complementary strands of a 404 bp fragment. This extended from the TaqI site located 36 bp upstream of the HincII site (Section 3.6.3) to the TaqI site found 200 bp from the 'N-terminus' EcoRI site (Section 3.6.2). In total therefore, at this stage, all 1644 bp of the 1.65 kbp EcoRI insert of pGM107 had been identified on at least one strand.

3.6.5 4th round of DNA sequencing (Sau3A fragments)

In order to complete the sequence of the 1644 bases on both strands and to overlap all the restriction sites, a final round of sequencing was carried out. The Sau3A digest mix (Section 3.6.1C) was cloned into BamHI-cut M13 8RF and DNA template prepared from 22 recombinant M13

plaques. The sequence information obtained provided the last elements of the sequence framework. The small (74 bp) HpaII fragment (Figure 3.13b) was overlapped at both ends. The DNA sequence at the 5' end (relative to the direction of aroB expression) of the clone was completed to within 62 bp of the EcoRI site (Figure 3.13e).

3.6.6 Compilation of sequence data

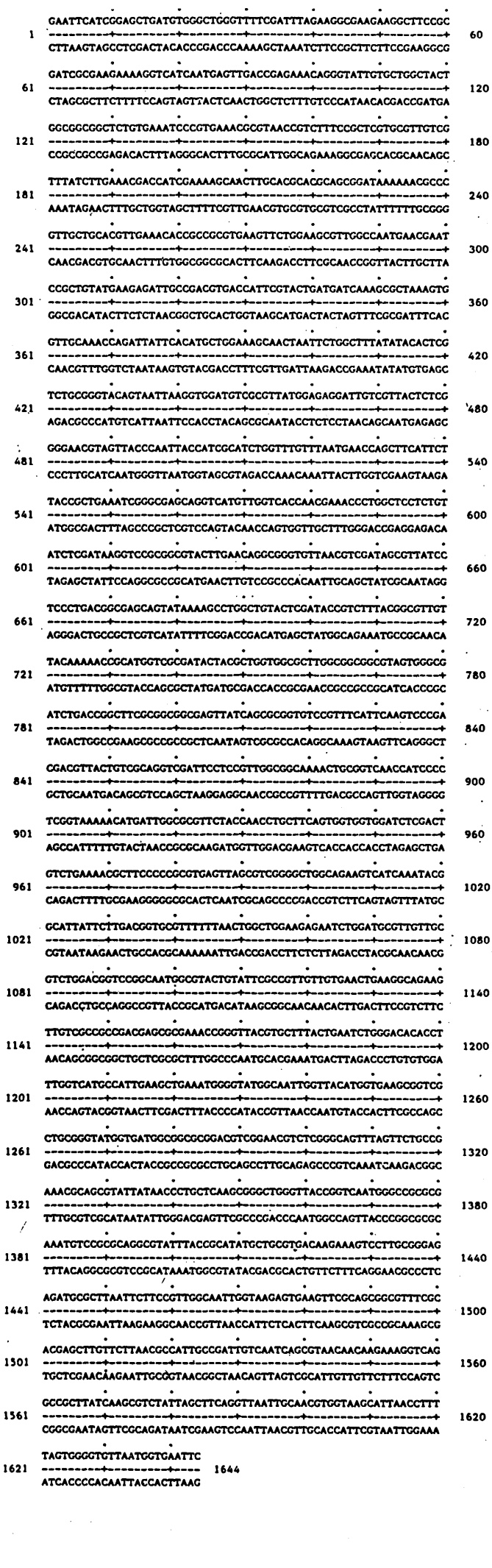
DNA sequences were entered as separate files into a Digital PDP 11-34 computer using a modified version of the Staden program (Staden, 1980) BATIN (Section 2.18.1). It is especially important to recognise sequences arising from cloned partial digestion products or multiple condensers. When sequencing a clone which contains an internal cloning site it is impossible to distinguish between these two possibilities. Hence whenever an internal HpaII, TaqI or Sau3A site was found a new separate sequence file was created.

3.6.7 Complete DNA sequence of the 1.65 kbp EcoRI genomic insert of pGM107

The complete sequence of the genomic insert of pGM107 is shown in Figure 3.14. The total length is 1644 bp, including the EcoRI recognition sites at either end. Only 62 bp at the tac-proximal EcoRI end of the sequence were determined on one strand. The sequence around this region was obtained from at least five different clones and is unambiguous. The distribution of AvaI, HincII, SstII and EcoRI sites within the sequence is identical to the restriction mapping data (Figure 3.5).

Figure 3.14 (facing)

The complete double-strand DNA sequence of the 1644 bp. EcoRI insert of E.coli genomic DNA in pGM107. The EcoRI site at position 0 is nearest the tac promoter in construct pGM107 (Figure 3.8). The sequencing strategy has been shown previously in Figure 3.13.



3.6.8 Analysis of the DNA sequence for protein coding regions

Although the direction of expression of *aroB* was known from sub-cloning (Section 3.4.1), both strands of the DNA sequence were examined for protein coding regions. The program TRNTRP (Staden, 1978, Section 2.18.2) was used to translate DNA sequence in all three reading frames for both strands.

A single large open reading frame (ORF) was identified which extended from position 440 to 1544. At position 456 was an ATG codon which could potentially be a translational initiation codon. No other ORF's capable of encoding a protein of more than 5,000 molecular weight were found. This putative *aroB* gene is positioned in the expected orientation as inferred from the sub-cloning data. When translated from the first ATG codon at position 456, a polypeptide of 362 amino acid residues and M_r 38,880 was predicted. Immediately upstream, at position 445, is the sequence GGTGG which exhibits homology with the consensus sequence for the *E.coli* ribosome binding site, GGAGG (Shine & Dalgarno, 1975). Interestingly this ORF is fully 456 bp downstream of the *tac* promoter in construct pGM107 (Figure 3.13).

3.6.9 Testcode analysis of the ca. 39 kD ORF

A quantitative assessment of the likelihood of this ORF being truly protein coding was obtained using the program TESTCODE (Fickett, 1982) in the WISGEN package (Section 2.18.3).

TESTCODE searches for genes by plotting the running value of Fickett's 'Testcode' statistic (Fickett, 1982)

along the length of the sequence. The test depends upon finding statistical order in the base sequence within protein-coding regions. This order is based on the non-random utilisation of codons and provides an objective means of identifying protein-coding sequences (Fickett, 1982). One consequence of the non-random use of codons is that nucleotides tend to be repeated with a periodicity of three in protein coding sequences (Fickett, 1982). TESTCODE can plot a measure of this 'period three constraint' of each region of a DNA molecule. The statistic is independent of the reading frame and is based on measurement of the 'period three constraint' in the entire Los Alamos Sequence Library for regions thought to be coding and non-coding.

Statistical analysis of DNA sequences, like the TESTCODE algorithm, are valuable but not definitive tools for identifying genes. TESTCODE provides supplementary rather than fundamental evidence and should always be evaluated as such.

Figure 3.15 shows a TESTCODE analysis of the putative aroB coding strand (mRNA-like strand). The plot is divided into three regions for which the statistic makes prediction. The top panel (A) is the region within which 95% of coding regions fall and within which 95% of non-coding regions do not. The bottom panel (C) is the opposite. The middle panel (B) is a region where the statistic makes no prediction. Clearly the program predicts that this ORF is very likely (95% probability) to be protein coding. Conversely Figure 3.16 shows a TESTCODE plot of the complementary strand where

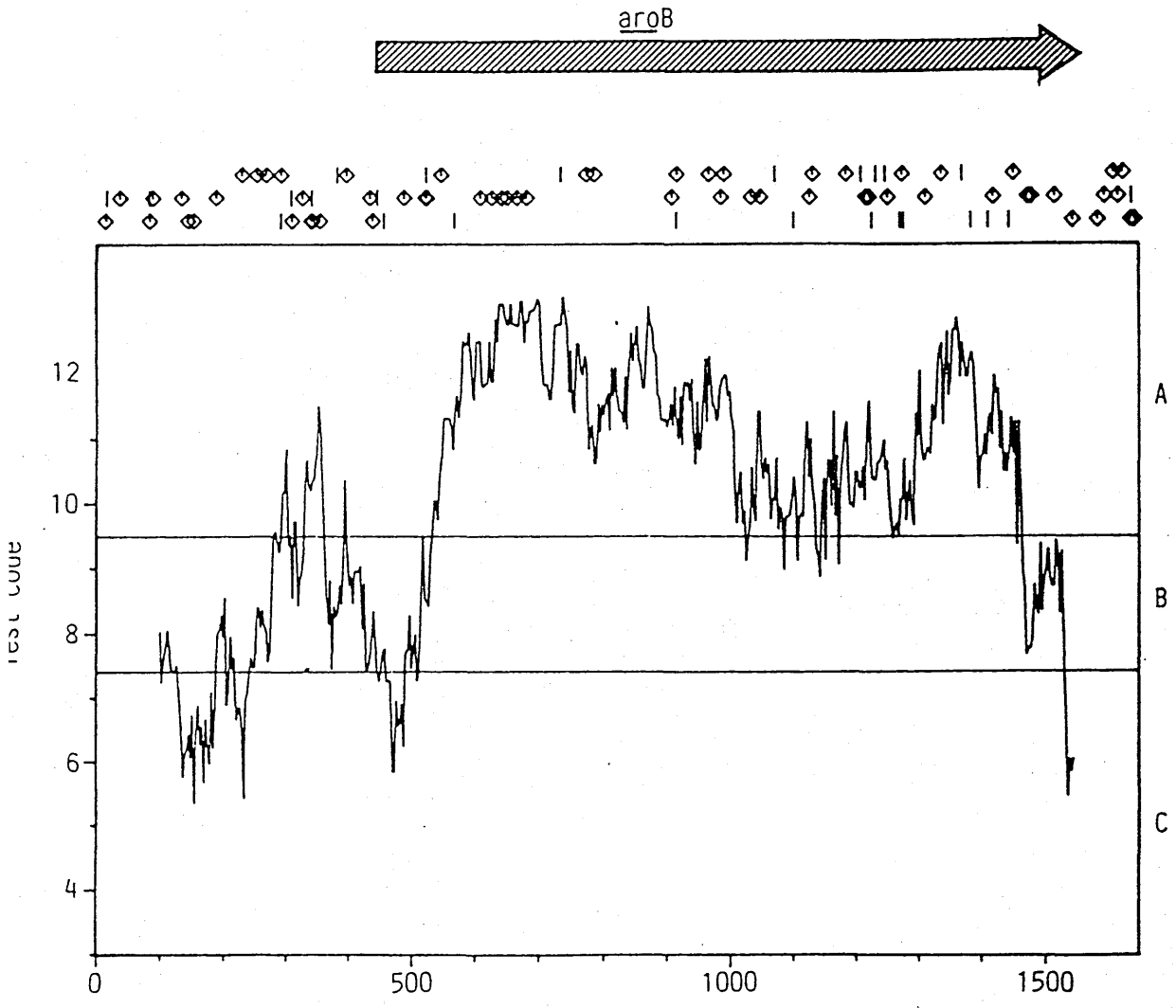


Figure 3.15: TESTCODE analysis of the aroB sense strand. Stop codons are indicated (◇), methionine codons indicated (1). Distance in bp is shown along the abscissa and TESTCODE values along the ordinate. The significance of panels A, B & C is discussed in the text.

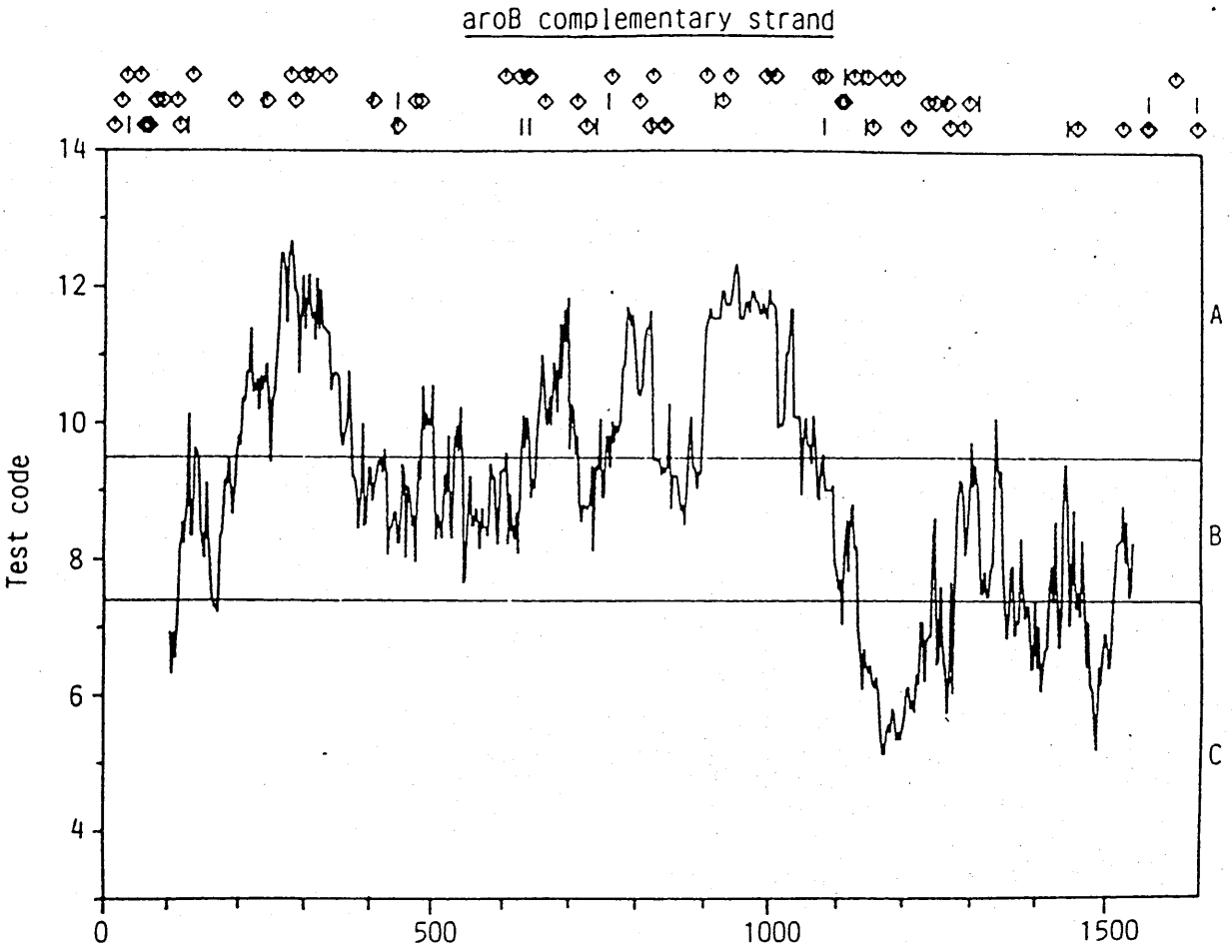


Figure 3.16: TESTCODE analysis of the aroB complementary (non-sense) strand. Symbols and values are as previously described (Figure 3.15).

Figure 3.17 (facing)

The *E. coli* DHQ synthase (*aroB*) coding region. Note the position of the proposed ribosome binding site (RBS). The protein coding sequence was derived from the primary DNA sequence data using the TRN TRP program.

```
   1  GAATTCATCGGAGCTGATGTGGGCTGGGTTTTCGATTTAGAAGGCGAAGAAGGCTTCGG
      .         .         .         .         .         .
  60  CGATCGCGAAGAAAAGGTCATCAATGAGTTGACCGAGAAACAGGGTATTGTGCTGGCTAC
      .         .         .         .         .         .
 120  TGGCGGCGGCTCTGTGAAATCCCGTGAAACGCGTAACCGTCTTTCCGCTCGTGCGTTGTC
      .         .         .         .         .         .
 180  GTTTATCTTGAAACGACCATCGAAAAGCAACTTGCACGCACGCAGCGGATAAAAAACGCC
      .         .         .         .         .         .
 240  CGTTGCTGCACGTTGAAACACCGCCGCGTGAAGTTCTGGAAGCGTTGGCCAATGAACGAA
      .         .         .         .         .         .
 300  TCCGCTGTATGAAGAGATTGCCGACGTGACCATTCGTACTGATGATCAAAGCGCTAAAGT
      .         .         .         .         .         .
 360  GGTTGCAAACCAGATTATTCACATGCTGGAAAGCAACTAATTCTGGCTTTATATACACTC
      .         .         .         .         .         .
 420  GTCTGCGGGTACAGTAATTAAGGTGGATGTCGCGTTATGGAGAGGATTGTCGTTACTCTC
                          RBS          MetGluArgIleValValThrLeu [8]
      .         .         .         .         .         .
 480  GGGGAACGTAGTTACCCAATTACCATCGCATCTGGTTTGTTTAATGAACCAGCTTCATTC
      GlyGluArgSerTyrProIleThrIleAlaSerGlyLeuPheAsnGluProAlaSerPhe [28]
      .         .         .         .         .         .
 540  TTACCGCTGAAATCGGGCGAGCAGGTCATGTTGGTCACCAACGAAACCCTGGCTCCTCTG
      LeuProLeuLysSerGlyGluGlnValMetLeuValThrAsnGluThrLeuAlaProLeu [48]
      .         .         .         .         .         .
 600  TATCTCGATAAGGTCCGCGGCGTACTTGAACAGGCGGGTGTTAACGTCGATAGCGTTATC
      TyrLeuAspLysValArgGlyValLeuGluGlnAlaGlyValAsnValAspSerValIle [68]
      .         .         .         .         .         .
 660  CTCCCTGACGGCGAGCAGTATAAAAGCCTGGCTGTACTCGATACCGTCTTTACGGCGTTG
      LeuProAspGlyGluGlnTyrLysSerLeuAlaValLeuAspThrValPheThrAlaLeu [88]
      .         .         .         .         .         .
 720  TTACAAAAACCGCATGGTCGCGATACTACGCTGGTGGCGCTTGGCGGCGGCGTAGTGGGC
      LeuGlnLysProHisGlyArgAspThrThrLeuValAlaLeuGlyGlyGlyValValGly [108]
      .         .         .         .         .         .
 780  GATCTGACCGGCTTCGCGGCGGCGAGTTATCAGCGCGGTGTCCGTTTCATTCAAGTCCCG
      AspLeuThrGlyPheAlaAlaAlaSerTyrGlnArgGlyValArgPheIleGlnValPro [128]
      .         .         .         .         .         .
 840  ACGACGTTACTGTCGCAGGTCGATTCCTCCGTTGGCGGCAAAACTGCGGTCAACCATCCC
      ThrThrLeuLeuSerGlnValAspSerSerValGlyGlyLysThrAlaValAsnHisPro [148]
      .         .         .         .         .         .
 900  CTCGGTAAAAACATGATTGGCGCGTTCTACCAACCTGCTTCAGTGGTGGTGGATCTCGAC
      LeuGlyLysAsnMetIleGlyAlaPheTyrGlnProAlaSerValValValAspLeuAsp [168]
      .         .         .         .         .         .
 960  TGTCTGAAAACGCTTCCCCCGCGTGAGTTAGCGTCGGGGCTGGCAGAAGTCATCAAATAC
      CysLeuLysThrLeuProProArgGluLeuAlaSerGlyLeuAlaGluValIleLysTyr [188]
      .         .         .         .         .         .
1020  GGCATTATTCTTGACGGTGCGTTTTTTAACTGGCTGGAAGAGAATCTGGATGCGTTGTTG
      GlyIleIleLeuAspGlyAlaPhePheAsnTrpLeuGluGluAsnLeuAspAlaLeuLeu [208]
      .         .         .         .         .         .
1080  CGTCTGGACGGTCCGGCAATGGCGTACTGTATTCGCCGTTGTTGTGAACTGAAGGCAGAA
      ArgLeuAspGlyProAlaMetAlaTyrCysIleArgArgCysCysGluLeuLysAlaGlu [228]
      .         .         .         .         .         .
1140  GTTGTCGCCGCCGACGAGCGCGAAACCGGGTTACGTGCTTTACTGAATCTGGGACACACC
      ValValAlaAlaAspGluArgGluThrGlyLeuArgAlaLeuLeuAsnLeuGlyHisThr [248]
      .         .         .         .         .         .
1200  TTTGGTCATGCCATTGAAGCTGAAATGGGGTATGGCAATTGGTTACATGGTGAAGCGGTC
      PheGlyHisAlaIleGluAlaGluMetGlyTyrGlyAsnTrpLeuHisGlyGluAlaVal [268]
      .         .         .         .         .         .
1260  GCTGCGGGTATGGTGATGGCGGCGCGGACGTCGGAACGTCTCGGGCAGTTTAGTTCTGCC
      AlaAlaGlyMetValMetAlaAlaArgThrSerGluArgLeuGlyGlnPheSerSerAla [288]
      .         .         .         .         .         .
1320  GAAACGCAGCGTATTATAACCCTGCTCAAGCGGGCTGGGTTACCGGTCAATGGGCCGCGC
      GluThrGlnArgIleIleThrLeuLeuLysArgAlaGlyLeuProValAsnGlyProArg [308]
      .         .         .         .         .         .
1380  GAAATGTCCGCGCAGGCGTATTTACCGCATATGCTGCGTGACAAGAAAGTCCTTGCGGGA
      GluMetSerAlaGlnAlaTyrLeuProHisMetLeuArgAspLysLysValLeuAlaGly [328]
      .         .         .         .         .         .
1440  GAGATGCGCTTAATTCTTCCGTTGGCAATTGGTAAGAGTGAAGTTCGCAGCGGCGTTTCG
      GluMetArgLeuIleLeuProLeuAlaIleGlyLysSerGluValArgSerGlyValSer [348]
      .         .         .         .         .         .
1500  CACGAGCTTGTTCTTAACGCCATTGCCGATTGTCAATCAGCGTAACAACAAGAAAGGTCA
      HisGluLeuValLeuAsnAlaIleAlaAspCysGlnSerAlaEnd [362]
      .         .         .         .         .         .
1560  GGCCGCTTATCAAGCGTCTATTAGCTTCAGGTTAATTGCAACGTGGTAAGCATTAACCTT
      .         .         .
1620  TTAGTCGGGTGTTAATGGTGAATTC  1644
```

no long ORFs are present and the statistic fails to identify any protein coding regions.

3.6.10 Predicted amino acid sequence of aroB coding region

The complete nucleotide sequence of the putative aroB gene is shown in Figure 3.17. The predicted polypeptide extending from positions 456 to 1544 has a molecular weight of 38,880. This is in excellent agreement with the value obtained by Frost et al. (1984) for the sub-unit size of DHQ synthase. It is also consistent with the value obtained for the major protein product of constructs pJB14 and pGM107 (Sections 3.5.2, 3.5.3).

3.6.11 Codon utilisation of the aroB gene

The codon utilisation for the open reading frame provisionally identified as the aroB gene is shown in Table 3.3. Discussion of the expected biases in codon usage for weakly expressed E.coli genes and comparison with codon usage pattern in the aroB and in other shikimate pathway genes in E.coli is considered in detail in Chapter 4 (Section 4.5.4).

Briefly the codon bias observed for the putative aroB gene conforms with that expected for a weakly expressed E.coli gene.

3.7 Purification and characterisation of DHQ synthase from the overproducing strain E.coli AB2826/pGM107

3.7.1 Background

The identification of an ORF, within the cloned insert of pGM107, capable of encoding a 39 kDa protein, which has a

	U	aroB	C	aroB	A	aroB	G	aroB	
U	Phe	6	Ser	2	Tyr	5	Cys	5	U
	Phe	4	Ser	3	Tyr	4	Cys	0	C
	Leu	10	Ser	3	ochre	1	opal	0	A
	Leu	6	Ser	5	amber	0	Trp	2	G
C	Leu	8	Pro	3	His	5	Arg	9	U
	Leu	8	Pro	2	His	2	Arg	8	C
	Leu	0	Pro	2	Gln	4	Arg	0	A
	Leu	17	Pro	9	Gln	8	Arg	2	G
A	Ile	12	Thr	8	Asn	5	Ser	4	U
	Ile	3	Thr	8	Asn	6	Ser	3	C
	Ile	1	Thr	0	Lys	8	Arg	0	A
	Met	10	Thr	7	Lys	5	Arg	1	G
G	Val	8	Ala	8	Asp	9	Gly	11	U
	Val	15	Ala	6	Asp	6	Gly	14	C
	Val	3	Ala	5	Gln	16	Gly	2	A
	Val	6	Ala	20	Gln	8	Gly	7	G

Table 3.3: aroB codon utilisation.

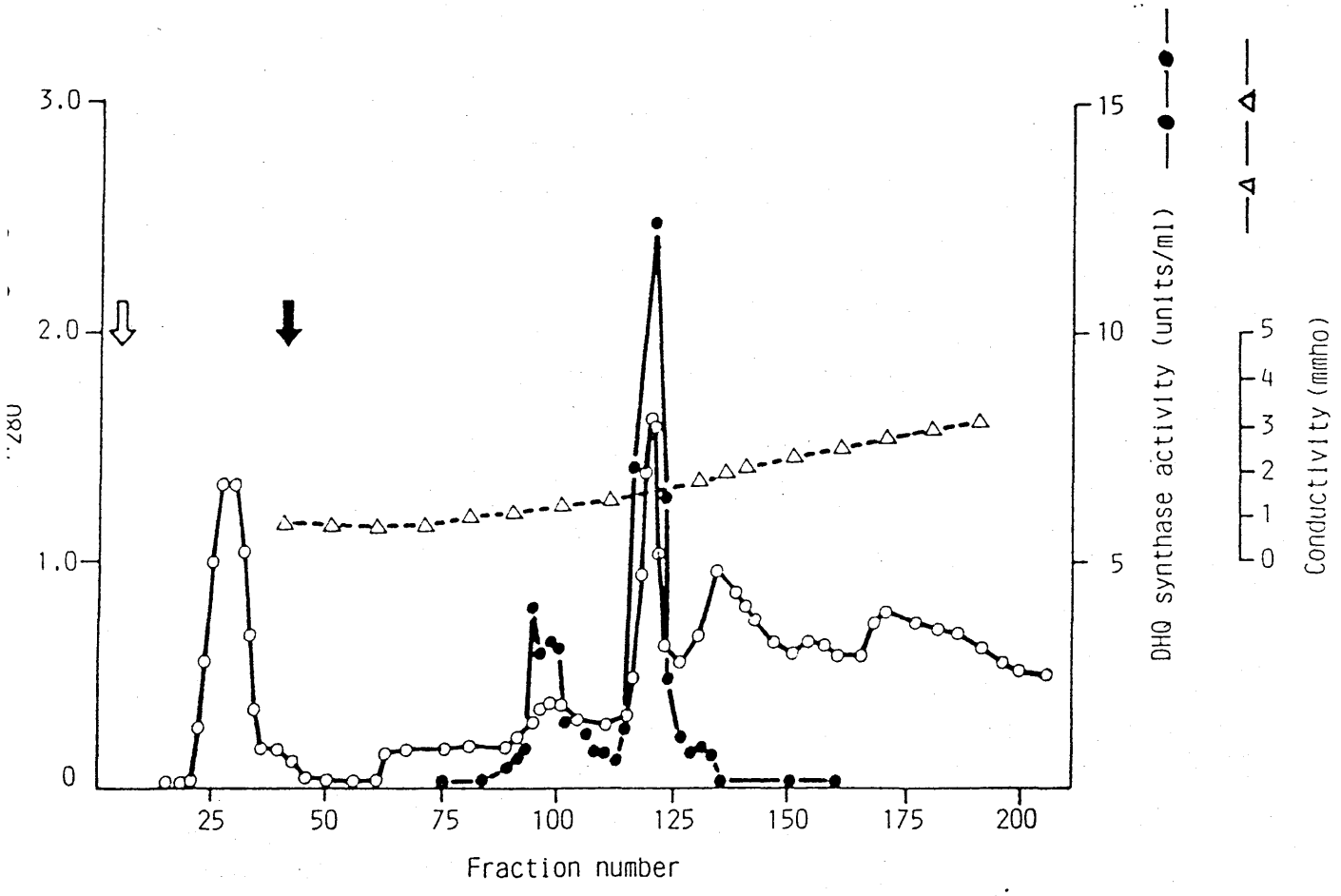
similar electrophoretic mobility to a sample of authentic E.coli DHQ synthase, strongly suggests but does not prove that the aroB gene has been cloned. The need to demonstrate that this predicted polypeptide is in fact DHQ synthase still exists. Also the exact translational initiation site within the ORF must be identified. The definitive location of the E.coli aroB gene can be deduced by firstly purifying the encoded protein, DHQ synthase. Subsequent N-terminal amino acid sequence analysis would then confirm the hypothesis.

Frost et al. (1984) have described a two step purification protocol for DHQ synthase being expressed at very high levels. Since plasmid pGM107 has the aroB gene positioned downstream of powerful tac promoter, the same approach was used (Section 2.23.2) to purify DHQ synthase.

3.7.2 Growth of cells

E.coli AB2826/pGM107 cells were grown on L broth supplemented with ampicillin (Section 2.4.3). An overnight culture of cells (120 ml) was used to inoculate 8 x 500 ml cultures in 2 litre shaking flasks. After one hour incubation, inducer (IPTG) was added to a final concentration of 5×10^{-4}M and incubation continued for a further 9 hours. The cells were harvested and a total of 39 g (wet weight) of cell paste was obtained from the 4 litres of culture.

Figure 3.18: Chromatography of E.coli DHQ synthase on Hydroxylapatite, step 2 of purification scheme detailed in Section 2.23.2 (Table 3.4).



Dialysed crude extract containing 3900 units in 100 ml of Buffer A (Section 2.23.2) was loaded on to a 400 ml Hydroxylapatite column. Following a 400 ml wash (⇩) of buffer A, a linear gradient of (2 litres) Buffer A and (2 litres) Buffer B (Section 2.23.2) was applied (⬇). Flow rate 50 ml/hr, 13.5 ml fractions throughout.

3.7.3 Purification of DHQ synthase

Purification details are given in Chapter Two (Section 2.23.2). DHQ synthase was purified from 20 g (wet weight) of *E.coli* AB2826/pGM107 cells as prepared in Section 3.7.2. A summary of the purification scheme is shown in Table 3.4.

The cells were broken by two passages through a French Pressure cell. Following deoxyribonuclease I treatment, cellular debris was removed by centrifugation. The soluble protein fraction was dialysed for 1 hour against the appropriate buffer (A) after which the conductivity was measured. The protein-containing pool was diluted with the appropriate buffer (A) until its conductivity was 0.9 mmho. This matched the conductivity (0.9 mmho) of the hydroxylapatite column buffer. The resultant soluble protein pool constituted the crude extract.

The first chromatographic step was fractionation on a hydroxylapatite column (Figure 3.18). DHQ synthase activity eluted in a minor pool at the start of the shallow (increasing) salt gradient and later in a major pool at a conductivity of ca. 2 mmho. The peak fractions were pooled and a total of 6,000 enzyme units recovered. This represented an apparent *increase* of 50% of the total activity loaded from the crude extract (3,900 enzyme units). This anomaly is consistent with the data of Frost *et al.* (1984), who also recorded the same phenomenon. The spectrophotometric assay procedure, which involves measuring absorbance changes at 234 nm, is complicated in crude extracts by high backgrounds and blank rates. This makes the precise estimation of enzyme activity

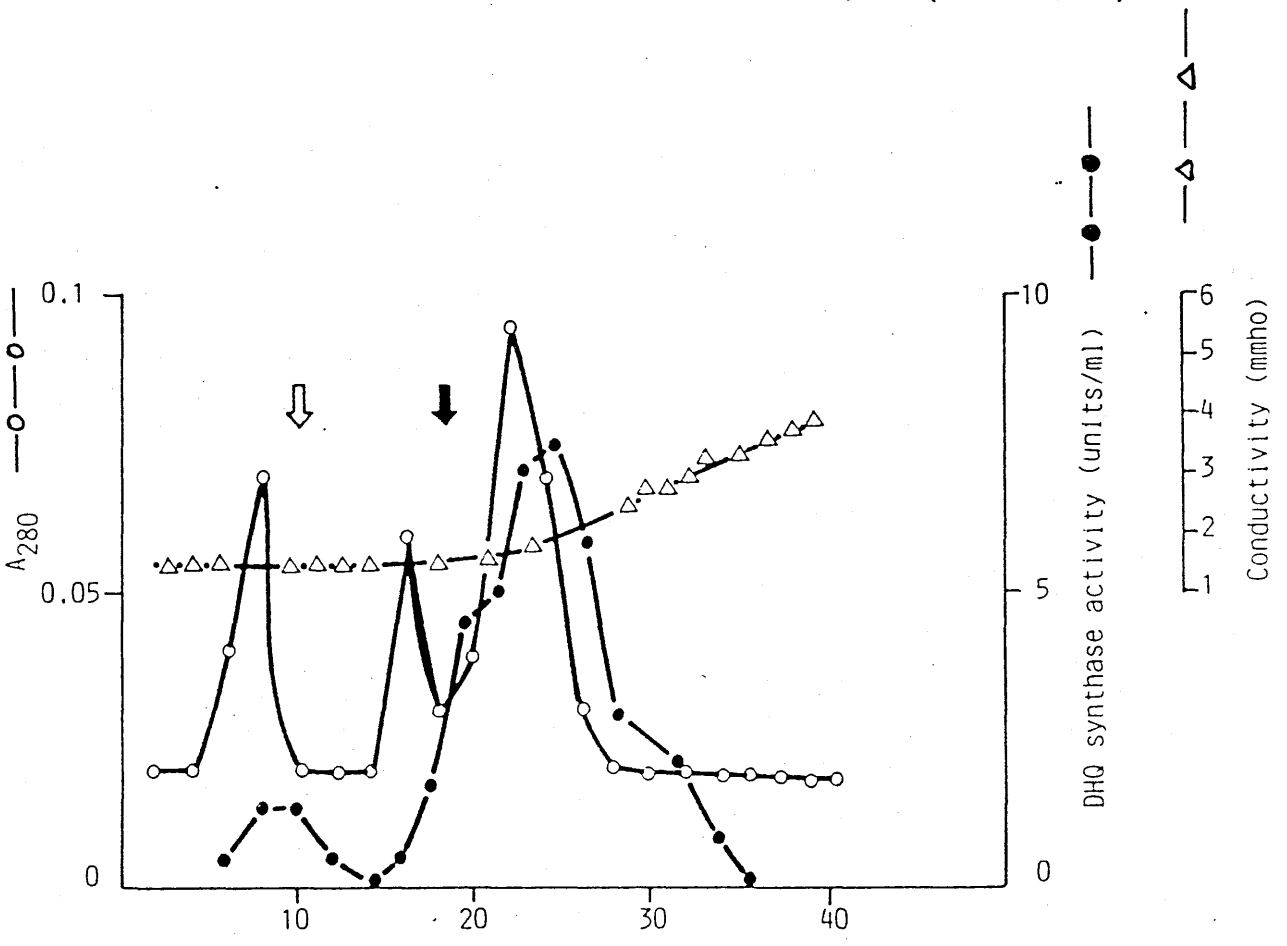
at this stage very difficult. A similar rationale may explain the observed 20-30 fold overproduction of DHQ synthase activity in aroB-plasmid transformed strains (Section 3.4.4). These plasmids have a copy number of 50-100 per cell but the level of overexpression is ca. 2-fold lower than expected.

The next step was chromatography on a Dyematrix Red (Procion Red) column (Figure 3.19). Due to the amounts of activity present (6,000 enzyme units) and the small size and limited binding capacity of the column used, only one third (2,000 enzyme units) of the active pool was subject to chromatographic separation. The remainder (4,000 enzyme units) was stored as a 50% glycerol solution at -20°C.

Procion Red affinity chromatography is useful in binding and separating $NAD(P)^+$ -linked enzyme. More usually, once bound, these enzymes are eluted with their pyridine nucleotide co-factor. In the case of E.coli DHQ synthase NAD^+ is present throughout binding to and elution from the column. Elution is achieved by applying an increasing salt gradient.

The peak fractions were pooled from the Procion Red column and contained 990 units. The specific activity of the protein pool was 41.3 enzyme units/mg. At this stage the pool was concentrated by vacuum dialysis and dialysed into long term storage buffer containing 50% glycerol. A total amount of 24 mg of protein and 990 enzyme units of activity were present. This represented a 32-fold purification of DHQ synthase and a yield of 25.4%. This yield was

Figure 3.19: Dyematrix Red A (Procion Red) Chromatography of E.coli DHQ synthase. Step 3 of purification scheme outlined in Section 2.23.2 (Table 3.4).



Dialysed Hydroxylapatite pool containing 2000 units in 105 ml Buffer D was applied to the 100 ml column and washed with (⇩) 100 ml Buffer D. A linear gradient of 400 ml Buffer D and 400 ml Buffer E was applied (Section 2.23.2; ⬇). Flow rate 38 ml/hr fraction size 14 ml throughout.

calculated on the original amount of activity present (3,900 enzyme units). Only 33% of the hydroxylapatite pool was carried forward so the true yield is really 3 x 25.4% (76.2%).

The specific activity of the enzyme preparation (41.3 units/mg) closely approached the value observed by Frost et al. (1984) in their purification scheme for DHQ synthase. Similarly their yield (86%) and purification factor (27-fold) values are both very similar. The degree of purity of DHQ synthase was examined by SDS PAGE.

Figure 3.20 shows a 12.5% polyacrylamide gel (containing SDS) electrophoretic separation of crude extract, hydroxylapatite and Procion Red pools of DHQ synthase activity. Clearly after the first step, hydroxylapatite separation, the enzyme is almost pure (> 90%). Certainly after Procion Red chromatography the enzyme preparation is pure. The apparent purity of the hydroxylapatite pool raises two questions. Although the peak fractions after hydroxylapatite do contain some contaminant bands (especially at very low molecular weight), none are really intense enough to account for the three peaks of protein (A_{280}) seen on the Procion Red column profile (Figure 3.19). And secondly the loss of > 50% of the loaded activity (2,000 enzyme units) on Procion Red and the increased specific activity (26.1 units/mg $\longrightarrow$ 41.3 units/mg) are perhaps unusual for an almost pure (> 90%) loading. Certainly it would appear that a large amount of the protein eluted from hydroxylapatite is inactive DHQ

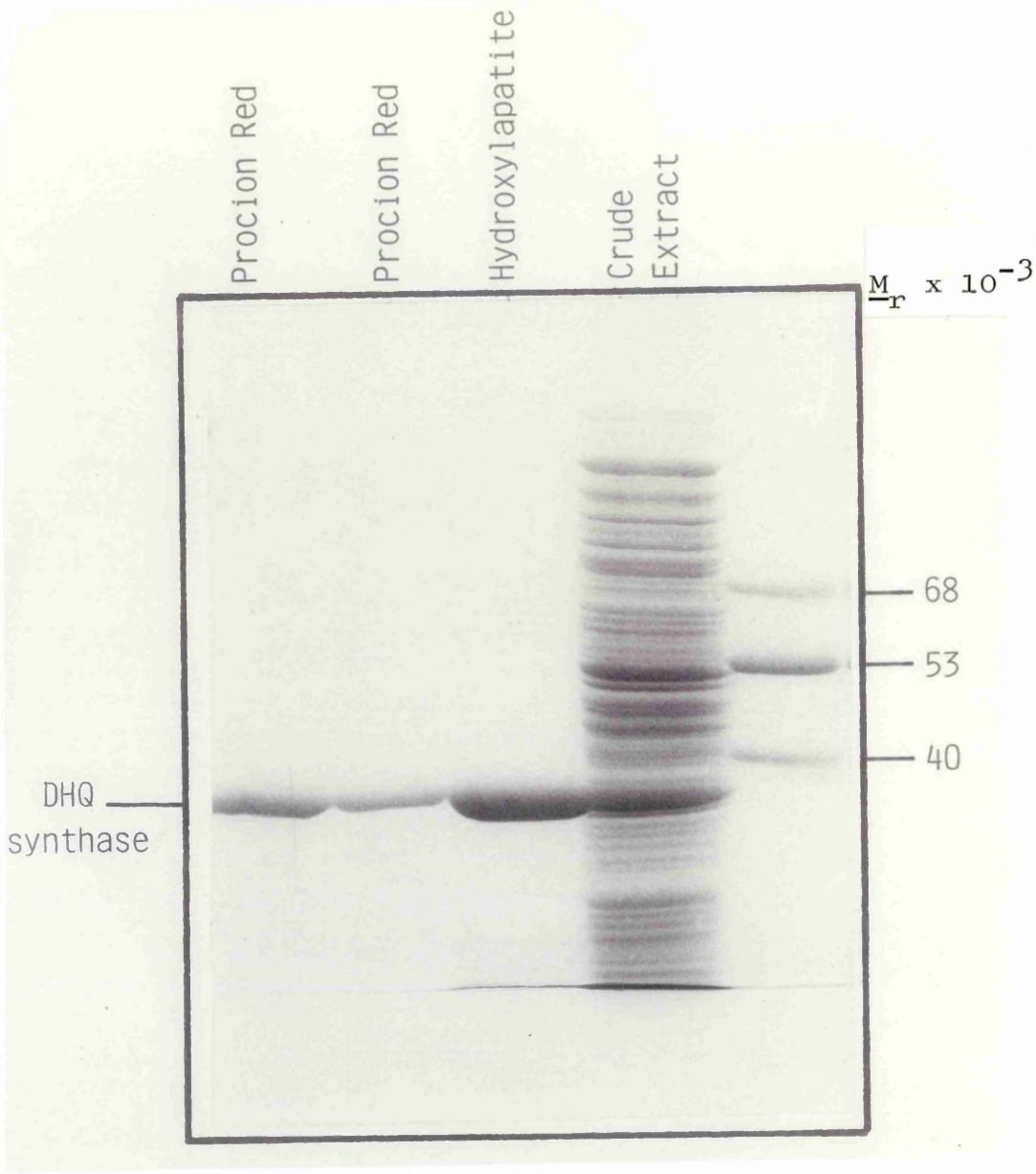


Figure 3.20: Coomassie stained 12.5% SDS PAGE of each step in the E.coli DHQ synthase (overexpressed) purification. Crude extract (ca. 150 µg), Hydroxylapatite (ca. 50 µg) and Procion Red (ca. 15 µg & 30 µg) samples were analysed. Molecular weights are shown.

Table 3.4: The purification of 3-dehydroquinate synthase from 20g of E.coli AB2826/pGM107

Step		Vol (ml)	Protein (mg/ml)	Activity (units/ml)	Total activity (units)	Yield (%)	Specific activity (units/mg)	Purification (fold)
1.	Crude extract	100	30	39	3900	100	1.3	1
2.	Hydroxylapatite	135	1.7	44.4	6000	154 [1]	26.1	20
3.	Procion red	100	0.24	9.9	990	25.4 [1]	41.3	32

Note

[1] Only one-third of the Hydroxylapatite pool was carried forward, therefore yield shown for step 3 is one-third of its true value.

synthase upon chromatographic separation on Procion Red. A way of confirming this would be to run some of the non-active peak fractions of Procion Red chromatography on SDS PAGE beside the active fractions. However since this observation was made retrospectively the said inactive fractions were no longer available.

The level of DHQ synthase activity in the crude extract sample (Figure 3.20) is in the range of ca. 5% of the total soluble protein. The IPTG-induced cells must be producing nearly 1,000 times as much DHQ synthase as the wild-type E.coli K12. The apparent molecular weight of DHQ synthase is ca. 38,000 by SDS PAGE criterion (Figure 3.20). This agrees well with both the values obtained for DHQ synthase from E.coli K12 and a similar tac-aroB overproducing strain (Frost et al., 1984). It also agrees well with the value predicted from the nucleotide sequence of the aroB gene of 38,880 (this study, Section 3.6.10, Millar & Coggins, 1986). As yet a comparison of the kinetic parameters and peptide maps of the overproduced and K12 DHQ synthases, as in Duncan et al. (1984), has not been done.

3.7.4 Determination of the N-terminal amino acid sequence of DHQ synthase

The N-terminal amino acid sequence of the overproduced E.coli DHQ synthase was determined by automatic protein sequencing. The work was carried out in collaboration with J.E. Fothergill, L.A. Fothergill-Gilmore, and B. Dunbar using the Beckman 890C automatic liquid phase sequencer at Aberdeen University. The sequence is shown in Figure 3.21

Residue Number	PTH-amino acid identified	Yield (nmoles)
1	Met	3.8
2	Glu	5.5
3	Arg	2.9
4	Ile	4.0
5	Val	4.6
6	Val	4.3
7	Thr	0.9
8	Leu	3.7
9	Gly	3.4
10	Glu	3.5
11	Arg	1.8
12	Ser	0.4
13	Tyr	2.3
14	Pro	1.6
15	Ile	1.8
16	Thr	0.4
17	Ile	1.3
18	Ala	1.7
19	Ser	0.2
20	Gly	1.1
21	Leu	1.5
22	Phe	1.5
23	Asn	1.1
24	Glu	1.4
25	Pro	0.7
26	Ala	1.0
27	Ser	0.3
28	Phe	0.9
29	Leu	0.6
30	Pro	0.5
31	Leu	0.6
32	Lys	0.4
33	Ser	0.2
34	Gly	0.6
35	Glu	0.7
36	Gln	0.5
37	Val	0.7
38	Met	0.5
39	Leu	0.4
40	Val	0.5
41	-	-
42	Asn	0.5
43	Glu	0.5
44	-	-
45	Leu	0.4

Figure 3.21: The N-terminal amino acid sequence of over-produced 3-dehydroquinate synthase. The sequence was determined on a liquid phase sequencer. The repetitive yield from residue 1 to 45 was 95%. Residues 41 and 44 could not be unambiguously identified.

and the methodologies employed detailed in Section 2.25.

The sequence of residues 1-45 was determined before the build up of background made unambiguous identification of PTH-amino acid derivatives impossible. Residues 41 and 44 were only tentatively identified as Threonine, but all the other identifications were unambiguous.

A comparison of the experimentally-deduced N-terminal amino acid sequence of the purified DHQ synthase with the amino acid sequence predicted from the DNA sequence (Section 3.6.10, Figure 3.17) shows that they are identical. The 45 amino acid residues (including Thr at positions 41 and 44) agree exactly with the proposed aroB coding region. The Methionine codon at position 456 in the nucleotide sequence (Figure 3.14) is the translational initiation codon.

3.7.5 Amino acid composition of the purified DHQ synthase

The amino acid composition of an acid hydrolysate of purified DHQ synthase was determined as described in Section 2.24. The constituent amino acids were normalised against a value of Leu = 49 residues per DHQ synthase subunit (predicted from the DNA sequence data). The values obtained are shown in Table 3.5.

Overall the agreement between the experimentally-obtained and the DNA sequence-derived compositions is very good. It would therefore appear that the amino acid sequence of the DHQ synthase as shown in Figure 3.17 can be directly correlated with and substantiated by the amino acid composition of the purified, overproduced enzyme.

E.coli DHQ synthase (aroB)

Residue	Relative amino acid composition (Leu = 49)	Predicted amino acid composition
Cys [c]	5.1	5
Asx	26.1	26
Met [a]	9.2	10
Thr [b]	16.6	18
Ser [b]	21.1	20
Glx	40.5	36
Pro	16.5	16
Gly	36.5	34
Ala [b]	36.8	39
Val	30.0	32
Ile	13.7	16
Leu	49	49
Tyr	6.1	9
Phe	9.8	10
His	6.7	7
Lys	12.8	13
Arg	20.6	20
Trp	nd	2

Table 3.5: Amino acid composition of E.coli DHQ synthase compared with that predicted from the DNA sequence of the aroB gene.

a. Determined as methionine sulphone
b. Experimental values at t = 0
c. Determined as cysteic acid.

3.7.6 Conclusion

E.coli DHQ synthase has been purified from an over-producing strain, *E.coli* AB2826/pGM107. The purified enzyme has an apparent molecular weight of ca. 38,000, by SDS PAGE criterion (Figure 3.20), which agrees well with the DNA sequence-predicted M_r 38,880.

The translational initiation site has been unambiguously assigned by direct N-terminal amino acid sequence of the purified enzyme. The first 45 residues determined agree exactly with the expected sequence. As a final check on the validity of the coding region, the amino acid composition of purified enzyme was deduced. Again the agreement is complete between the predicted and observed values. This ORF is therefore the DHQ synthase coding sequence.

3.8 Transcript Mapping studies on the *aroB* gene

3.8.1 Preparation of RNA

Total cellular RNA was prepared from exponentially growing cultures of *E.coli* AB2826 transformed with either plasmid pGM107 or pGM108. The method employed is a modified version of that described by Aiba *et al.* (1981), as outlined in Section 2.19.1. The integrity of the RNA preparation was examined by electrophoresis of an aliquot of each on a 1% agarose gel (Section 2.10). Visualisation of the ethidium bromide stained nucleic acids under U.V. irradiation revealed a smearing effect with major bands corresponding to the major ribosomal RNA species (Figure 3.22).

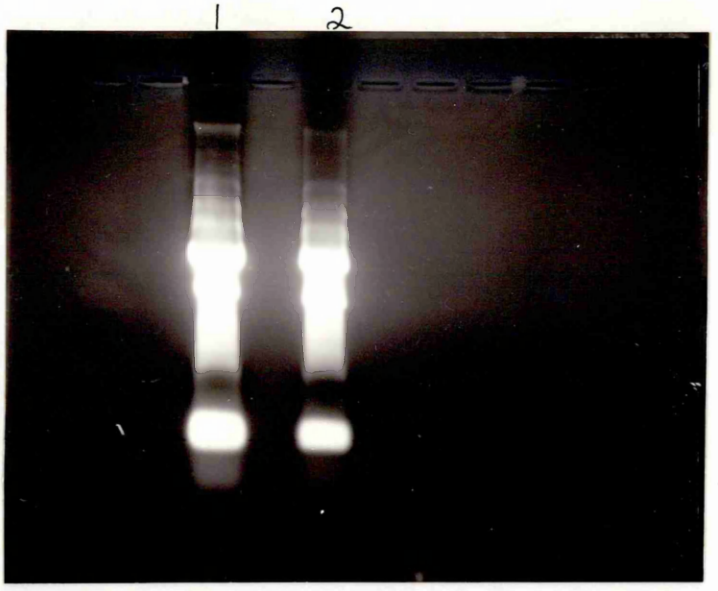


Figure 3.22: 1% Agarose gel profile of total RNA (ca. 10 µg) prepared from E.coli AB2826/pGM107 (Track 1) and E.coli HB101/pGM425 (Track 2). The major ribosomal RNA species serve as size markers.

3.8.2 Primer extension of oligonucleotide PHE128

A synthetic oligonucleotide primer, 25 nucleotides long (sequence 5'-GGT AAT TGG GTA ACT ACG TTC CCC G-3'; PHE 128) was a generous gift of Dr M.G. Hunter, Searle Research and Development. This 25-mer is complementary to nucleotides 24-48 (Figure 3.17) in the *aroB* coding sequence. The transcriptional start site of the *aroB* mRNA was examined by primer extension of this oligonucleotide (PHE 128) as described in Section 2.19.3.

The primer extension ('reverse run-off') analysis was carried out on RNA prepared from both pGM107 and pGM108 transformed AB2826 cells. In construct pGM107 the *aroB* gene has been cloned downstream of the powerful *tac* promoter.

To ensure that identification of the natural *aroB* promoter is not confused with transcripts originating at or near the *tac* (vector) promoter, a comparison of pGM107 (*tac*) and pGM108 (pAT153) PHE 128-initiated reverse run-off products was made (Figure 3.23(b)).

Primer extension transcript mapping analysis of the *aroB*-encoding mRNA (pGM107) using PHE 128 is shown in Figure 3.23(a). PHE 128 was also annealled to a recombinant M13 template and used as a sequencing primer to give a convenient size ladder. Secondary structure formation at the 5' end of the mRNA results in a number of premature terminations of run-off transcripts (see for example position (ii) in Figure 3.23(a)). However the longest extension band (position (i)

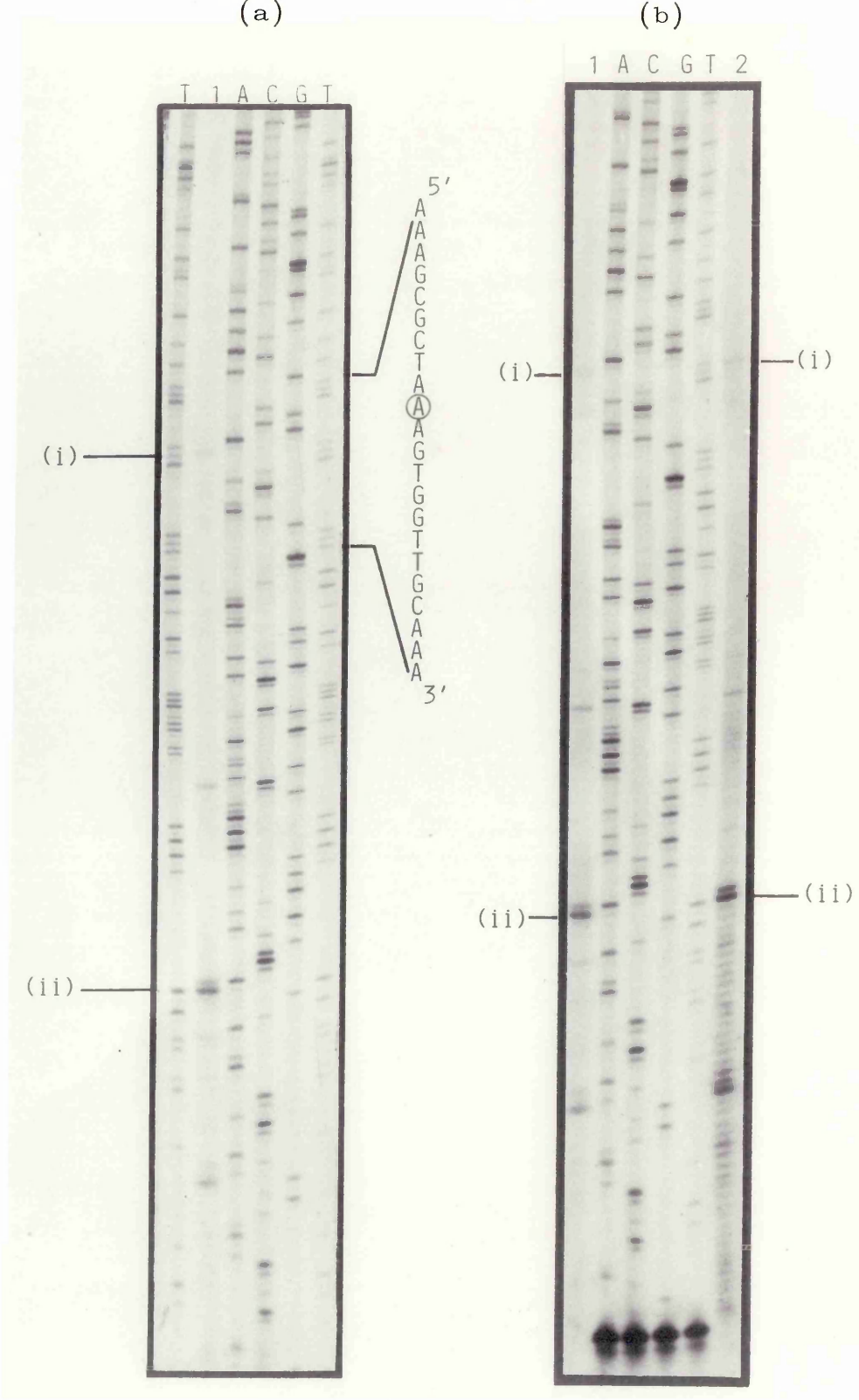


Figure 3.23: Primer extension transcript mapping of (a) pGM107 (track 1) and (b) pGM108 (tracks 1 & 2) encoded RNA's. The sequence around the transcript start site (encircled) was determined using the same primer. Run-off products (i) and (ii) are discussed in Section 3.8.2.

Figure 3.23(a)) maps the transcript start site at an A residue located 100 bp upstream of the translational start site. The location of the full length reverse run-off transcript band and the pattern of major premature termination species is identical for both pGM107- and pGM108-derived aroB mRNA's (Figure 3.23(b)). This evidence identifies the natural aroB transcription initiation site (positive 356, Figure 3.17) and precludes the possibility that expression of aroB in construct pGM107 may be from an artificially long transcript.

3.8.3 Identification of the aroB promoter

The DNA sequence immediately upstream of the aroB mRNA transcript start-site was searched by eye for promoter sequences. The consensus E.coli promoter sequence has two distinct elements (Hawley & McClure, 1983); a -10 region (or Pribnow box; Pribnow, 1975) specified by the sequence TAtAaT, and a -35 region characterised by TTGACa (where capital letters indicate highly conserved nucleotides). Other peripheral nucleotides contribute to a lesser extent to the overall consensus sequence (see Figure 3.24). The optimal separation between the two hexanucleotide sequences is typically 17 bp. Mulligan et al. (1985) have quantitatively characterised in vitro the effect of this spacer length on the activity of E.coli RNA polymerase binding. They found that a separation of 17 bp was optimal in terms of rapidity of RNA polymerase binding to the highly conserved hexanucleotides.

```
     .          .          .          .          .          .
240  CGTTGCTGCACGTTGAAACACCGCCGCGTGAAGTTCTGGAAGCGTTGGCCAATGAACGAA

     .          .    tcTTGACat   t  .  t  tg  TAtAaT .       +1
300  TCCGCTGTATGAAGAGATTGCCGACGTGACCATTCGTACTGATGATCAAAGCGCTAAAGT
                         -35                        -10         +--->
     .          .          .          .          .          .
360  GGTTGCAAACCAGATTATTCACATGCTGGAAAGCAACTAATTCTGGCTTTATATACACTC
     .          .          .          .          .          .
420  GTCTGCGGGTACAGTAATTAAGGTGGATGTCGCGTTATGGAGAGGATTGTCGTTACTCTC
                          RBS           MetGluArgIleValValThrLeu [8]
```

Figure 3.24: The sequence upstream of the aroB gene, identified by transcript mapping (Figure 3.23) as the aroB promoter. The consensus (Hawley & McClure, 1983) Pribnow box (-10) and -35 regions are indicated above the aroB promoter elements. The first eight residues of the DHQ synthase coding region are shown.

Immediately upstream of the aroB transcriptional start site (+1, Figure 3.23(a)) are sequences with considerable homology to the consensus E.coli -10 and -35 promoter elements. The aroB promoter -10 region (GATGAT) is separated by 17 bp, the optimal separation, from the -35 region (TTGCCG). Only three strongly conserved nucleotides differ from those observed for the standard E.coli promoter. The distance from the 3'-most nucleotide of the -10 region to the transcript start site is 10 bp, within the preferred 7-10 bp separation (Figure 3.24).

3.8.4 Possible aroB terminator

The 3'-flanking regions of the aroB coding sequence were searched for sequences capable of participating in transcription termination. Immediately 3' to the end of the aroB gene, and starting at position 1588, is a sequence (indicated by overlining in Figure 3.27) capable of forming a stem-loop structure (Figure 3.25). Such a structure is characteristic of a rho-independent terminator (Rosenberg & Court, 1979). A free-energy of formation (ΔG) was calculated by the rules specified by Tinoco et al. (1973), and the negative value obtained suggests that this stem-loop may form a stable secondary structure in vivo. This sequence may form part of an aroB terminator. (Figures 3.25 and 3.26).

Figure 3.27 summarises the location and organisation of the 5' and 3' sequence elements which influence (or may influence) expression of the aroB gene.

Figure 3.25 (facing).

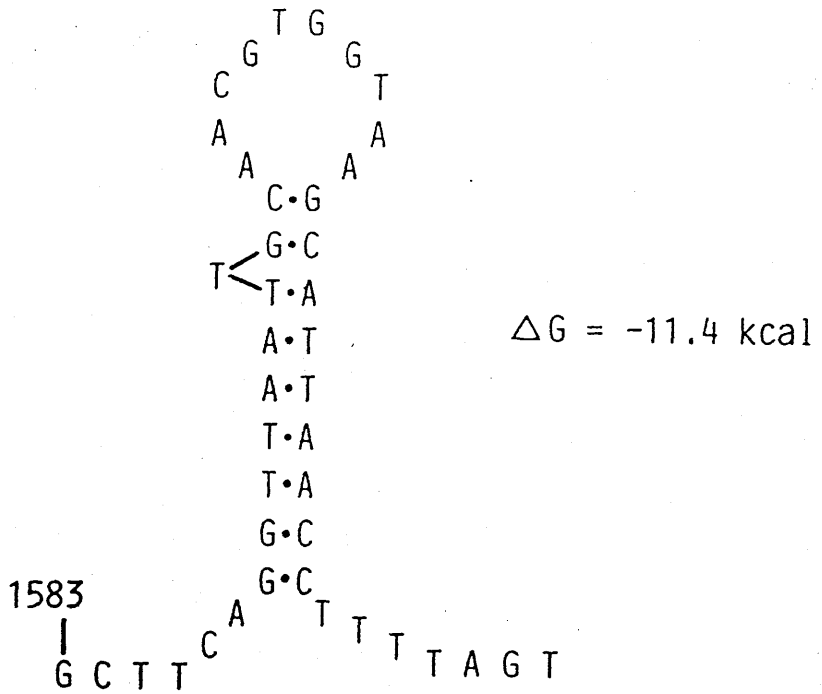
(Upper) An abbreviated aroB coding region showing the first 24 and last 10 amino acid residues. The inverted repeat sequence 3' to the aroB gene is overlined.

Figure 3.26 (facing).

(Lower) A potential stem-loop structure formed by the inverted repeat sequence described above. This may constitute an aroB terminator since it resembles the rho-independent structures described by Rosenberg & Court (1979). The free energy of formation was calculated by the rules of Tinoco et al. (1973).

```
    456
CGCGTTATGGAGAGGATTGTCGTTACTCTCGGGGAACGTAGTTACCCAATTACCATCGCA
      MetGluArgIleValValThrLeuGlyGluArgSerTyrProIleThrIleAla

TCTGGTTTGTTTAATGAA----------\ \----------CTTAACGCCATTGCCGAT
SerGlyLeuPheAsnGlu----------\ \----------LeuAsnAlaIleAlaAsp
                          aroB cod. seq
           1544                                    1580
TGTCAATCAGCGTAACAACAAGAAAGGTCAGGCCGCTTATCAAGCGTCTATTAGCTTCAG
CysGlnSerAlaEnd
                                                        1644
GTTAATTGCAACGTGGTAAGCATTAACCTTTTAGTGGGGTGTTAATGGTGAATTC
```



$\Delta G = -11.4$ kcal

Figure 3.27 (facing)

The *E.coli* *aroB* gene, encoding 3-dehydroquinate synthase. The transcript start site is marked (+ 1) and the consensus promoter sequences (Hawley & McClure, 1983) shown above the mapped *aroB* promoter.

```
                .         .         .         .         .         .
  60  CGATCGCGAAGAAAAGGTCATCAATGAGTTGACCGAGAAACAGGGTATTGTGCTGGCTAC
                .         .         .         .         .         .
 120  TGGCGGCGGCTCTGTGAAATCCCGTGAAACGCGTAACCGTCTTTCCGCTCGTGCGTTGTC
                .         .         .         .         .         .
 180  GTTTATCTTGAAACGACCATCGAAAAGCAACTTGCACGCACGCAGCGGATAAAAAACGCC
                .         .         .         .         .         .
 240  CGTTGCTGCACGTTGAAACACCGCCGCGTGAAGTTCTGGAAGCGTTGGCCAATGAACGAA

                .         .    tcTTGACat   t   .  t  tg  TAtAaT .       +1
 300  TCCGCTGTATGAAGAGATTGCCGACGTGACCATTCGTACTGATGATCAAAGCGCTAAAGT
                              -35                        -10        +--->
                .         .         .         .         .         .
 360  GGTTGCAAACCAGATTATTCACATGCTGGAAAGCAACTAATTCTGGCTTTATATACACTC
                .         .         .         .         .         .
 420  GTCTGCGGGTACAGTAATTAAGGTGGATGTCGCGTTATGGAGAGGATTGTCGTTACTCTC
                                  RBS            MetGluArgIleValValThrLeu [8]
                .         .         .         .         .         .
 480  GGGGAACGTAGTTACCCAATTACCATCGCATCTGGTTTGTTTAATGAACCAGCTTCATTC
      GlyGluArgSerTyrProIleThrIleAlaSerGlyLeuPheAsnGluProAlaSerPhe [28]
                .         .         .         .         .         .
 540  TTACCGCTGAAATCGGGCGAGCAGGTCATGTTGGTCACCAACGAAACCCTGGCTCCTCTG
      LeuProLeuLysSerGlyGluGlnValMetLeuValThrAsnGluThrLeuAlaProLeu [48]
                .         .         .         .         .         .
 600  TATCTCGATAAGGTCCGCGGCGTACTTGAACAGGCGGGTGTTAACGTCGATAGCGTTATC
      TyrLeuAspLysValArgGlyValLeuGluGlnAlaGlyValAsnValAspSerValIle [68]
                .         .         .         .         .         .
 660  CTCCCTGACGGCGAGCAGTATAAAAGCCTGGCTGTACTCGATACCGTCTTTACGGGGTTG
      LeuProAspGlyGluGlnTyrLysSerLeuAlaValLeuAspThrValPheThrAlaLeu [88]
                .         .         .         .         .         .
 720  TTACAAAAACCGCATGGTCGCGATACTACGCTGGTGGCGCTTGGCGGCGGCGTAGTGGGC
      LeuGlnLysProHisGlyArgAspThrThrLeuValAlaLeuGlyGlyGlyValValGly [108]
                .         .         .         .         .         .
 780  GATCTGACCGGCTTCGCGGCGGCGAGTTATCAGCGCGGTGTCCGTTTCATTCAAGTCCCG
      AspLeuThrGlyPheAlaAlaAlaSerTyrGlnArgGlyValArgPheIleGlnValPro [128]
                .         .         .         .         .         .
 840  ACGACGTTACTGTCGCAGGTCGATTCCTCCGTTGGCGGCAAAACTGCGGTCAACCATCCC
      ThrThrLeuLeuSerGlnValAspSerSerValGlyGlyLysThrAlaValAsnHisPro [148]
                .         .         .         .         .         .
 900  CTCGGTAAAAACATGATTGGCGCGTTCTACCAACCTGCTTCAGTGGTGGTGGATCTCGAC
      LeuGlyLysAsnMetIleGlyAlaPheTyrGlnProAlaSerValValValAspLeuAsp [168]
                .         .         .         .         .         .
 960  TGTCTGAAAACGCTTCCCCCGCGTGAGTTAGCGTCGGGGCTGGCAGAAGTCATCAAATAC
      CysLeuLysThrLeuProProArgGluLeuAlaSerGlyLeuAlaGluValIleLysTyr [188]
                .         .         .         .         .         .
1020  GGCATTATTCTTGACGGTGCGTTTTTTAACTGGCTGGAAGAGAATCTGGATGCGTTGTTG
      GlyIleIleLeuAspGlyAlaPhePheAsnTrpLeuGluGluAsnLeuAspAlaLeuLeu [208]
                .         .         .         .         .         .
1080  CGTCTGGACGGTCCGGCAATGGCGTACTGTATTCGCCGTTGTTGTGAACTGAAGGCAGAA
      ArgLeuAspGlyProAlaMetAlaTyrCysIleArgArgCysCysGluLeuLysAlaGlu [228]
                .         .         .         .         .         .
1140  GTTGTCGCCGCCGACGAGCGCGAAACCGGGTTACGTGCTTTACTGAATCTGGGACACACC
      ValValAlaAlaAspGluArgGluThrGlyLeuArgAlaLeuLeuAsnLeuGlyHisThr [248]
                .         .         .         .         .         .
1200  TTTGGTCATGCCATTGAAGCTGAAATGGGGTATGGCAATTGGTTACATGGTGAAGCGGTC
      PheGlyHisAlaIleGluAlaGluMetGlyTyrGlyAsnTrpLeuHisGlyGluAlaVal [268]
                .         .         .         .         .         .
1260  GCTGCGGGTATGGTGATGGCGGCGCGGACGTCGGAACGTCTCGGGCAGTTTAGTTCTGCC
      AlaAlaGlyMetValMetAlaAlaArgThrSerGluArgLeuGlyGlnPheSerSerAla [288]
                .         .         .         .         .         .
1320  GAAACGCAGCGTATTATAACCCTGCTCAAGCGGGCTGGGTTACCGGTCAATGGGCCGCGC
      GluThrGlnArgIleIleThrLeuLeuLysArgAlaGlyLeuProValAsnGlyProArg [308]
                .         .         .         .         .         .
1380  GAAATGTCCGCGCAGGCGTATTTACCGCATATGCTGCGTGACAAGAAAGTCCTTGCGGGA
      GluMetSerAlaGlnAlaTyrLeuProHisMetLeuArgAspLysLysValLeuAlaGly [328]
                .         .         .         .         .         .
1440  GAGATGCGCTTAATTCTTCCGTTGGCAATTGGTAAGAGTGAAGTTCGCAGCGGCGTTTCG
      GluMetArgLeuIleLeuProLeuAlaIleGlyLysSerGluValArgSerGlyValSer [348]
                .         .         .         .         .         .
1500  CACGAGCTTGTTCTTAACGCCATTGCCGATTGTCAATCAGCGTAACAACAAGAAAGGTCA
      HisGluLeuValLeuAsnAlaIleAlaAspCysGlnSerAlaEnd [362]
                .         .         .          ____ __  .       _____
1560  GGCCGCTTATCAAGCGTCTATTAGCTTCAGGTTAATTGCAACGTGGTAAGCATTAACCTT

                .         .         .
1620  TTAGTGGGGTGTTAATGGTGAATTC  1644
```

3.9 Amino acid homologies with other DHQ synthases

3.9.1 BESTFIT comparisons

Regions of amino acid homology between the DHQ synthase domains of the Saccharomyces cerevisiae (Duncan et al., 1986) and Aspergillus nidulans (Charles et al., 1986) arom multifunctional enzymes and the monofunctional E.coli DHQ synthase (Millar & Coggins, 1986; this study) were examined. The BESTFIT algorithm, as outlined in Section 2.18.3, was used to find the best region of similarity between any two of the given sequences. The percentage similarity was calculated in a strict and consistent manner being defined as the number of identities within the area recognised as homologous by BESTFIT.

In the case of DHQ synthase comparisons (and subsequent comparisons in Chapters 5 and 6) between monofunctional and multifunctional activities, an elevated level of homology could have been obtained by fixing the start point of one sequence (e.g. E.coli aroB). In this case the comparison would be optimised from residue 1 by insertion of small gaps to align the two sequences. The output of BESTFIT does not do this. BESTFIT simply finds the optimal alignment of the best region of similarity between two sequences. 'Optimal' is defined by maximising identities with respect to gaps - a calculated quality factor selects the best fit. Therefore the cost of consistency in comparing two separate BESTFIT outputs (e.g. E.coli/S.cerevisiae vs. E.coli/A.nidulans) is sometimes a slight decrease in the overall homologies

obtained. BESTFIT will always find a similarity between two sequences, and very often several such similarities. A degree of relevance or significance can only be assessed for these homologies when they are critically examined with the help of additional biochemical or genetic evidence.

A second consequence of using the BESTFIT output for comparisons is that areas of very low (but still significant) homology may be missed. The program optimises another set of non-significant identities by inserting gaps which do not acrue sufficient gap penalties to lower the quality factor below a positive selection threshold. Again in these cases (see Chapter 5) additional corroborating evidence (e.g. identifiable nucleotide binding consensus sequences within the given sequence) must be considered in order to select the correct match.

Areas of very high homology tend not to create such problems. Not unexpectedly, there is an inverse relationship between the number of gaps per unit length and the degree of identity. A corollary to this may be that if a highly homologous region does require a large gap insertion then the sequences flanking the gap must be very homologous by the selection criterion of BESTFIT. Otherwise the gap penalty incurred would have gained selection for a second, distinct set of identities.

In addition to the number of direct identities observed, a percentage similarity calculated by including the number of acceptable (conservative) replacements:- I-L-V, D-E, R-K,

T-S, was also computed (Dayhoff, 1978). A second class of acceptable change: those concerning amino acids that could be considered derived from another by a single base change in the corresponding codon, were not included.

3.9.2. Homology with the *Saccharomyces cerevisiae arom*

The BESTFIT homology obtained between the *E.coli* DHQ synthase amino acid sequence (362 amino acid residues) and the *S. cerevisiae arom* polypeptide sequence (1588 amino acid residues) (Duncan *et al.*, 1986) is shown in Figure 3.28. Homologies were found within the first 391 amino acids of the yeast sequence. This agrees with deletion mutation analysis (Duncan *et al.*, 1986) which identified the N-terminal region of the *S.cerevisiae arom* polypeptide as containing the DHQ synthase activity.

The first residues implicated as part of the BESTFIT match are residue 50 of the *E.coli* sequence and residue 55 of the yeast sequence. Overall there is 36% homology (122/336) between the two peptide sequences. If the number of conservative changes (23) are included, this figure rises to 43%. Significantly 65% (15/23) of these proposed conservative changes occur in regions of the two polypeptides already characterised by >50% identity. A notable feature of the match is the large gap at position 214 which must be inserted in the *E.coli* sequence for optimal alignment. Both before and after this insertion are regions of very high identity. Outwith this large insertion only two single-residue gaps, both the yeast sequence, are required for

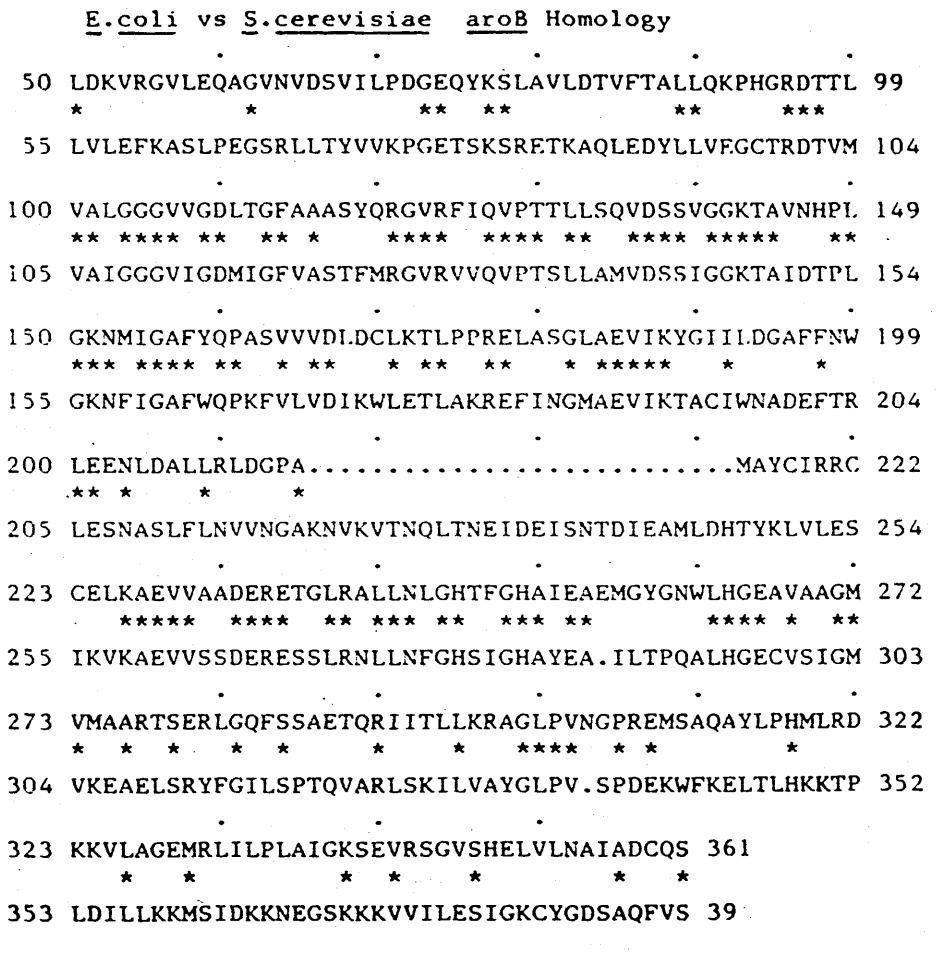


Figure 3.28: BESTFIT homology between the E.coli DHQ synthase sequence (upper strand) and the S.cerevisiae arom sequence (lower strand, Duncan et al., 1986a).

optimal alignment.

BESTFIT was unable to find any other match with such striking homology throughout the length (1588 amino acids) of the yeast arom polypeptide. It is clear therefore that the E.coli DHQ synthase sequence is related to a similar sequence located within the initial 400 amino acid residues of the S.cerevisiae arom polypeptide.

3.9.3. Homology with the Aspergillus nidulans arom polypeptide

A similar homology search was carried out with the E.coli DHQ synthase sequence and the A.nidulans arom polypeptide sequence (Charles et al., 1986). Figure 3.29 shows the result of a BESTFIT match between these two amino acid sequences. As in the comparison of Section 3.9.2 (E.coli vs. S.cerevisiae), a region of strong homology was found between the E.coli DHQ synthase sequence and the N-terminal region of the Aspergillus pentafunctional polypeptide sequence.

Within the limits of the BESTFIT comparison there are 124/331 identities (37%) occurring in the first 366 amino acid residues of the A.nidulans arom sequence. A consideration of conservative changes (18) elevates this degree of homology to 43%. Again a large insertion is required in the bacterial sequence to align the two polypeptides correctly. In the comparison of Section 3.9.2 the insertion occurred between residues 215/216 of the E.coli protein accommodating residues 220 to 247 of the S.cerevisiae protein. In this case the gap is situated at residues 208/209 of the prokaryotic polypeptide allowing accurate alignment of the extra 14

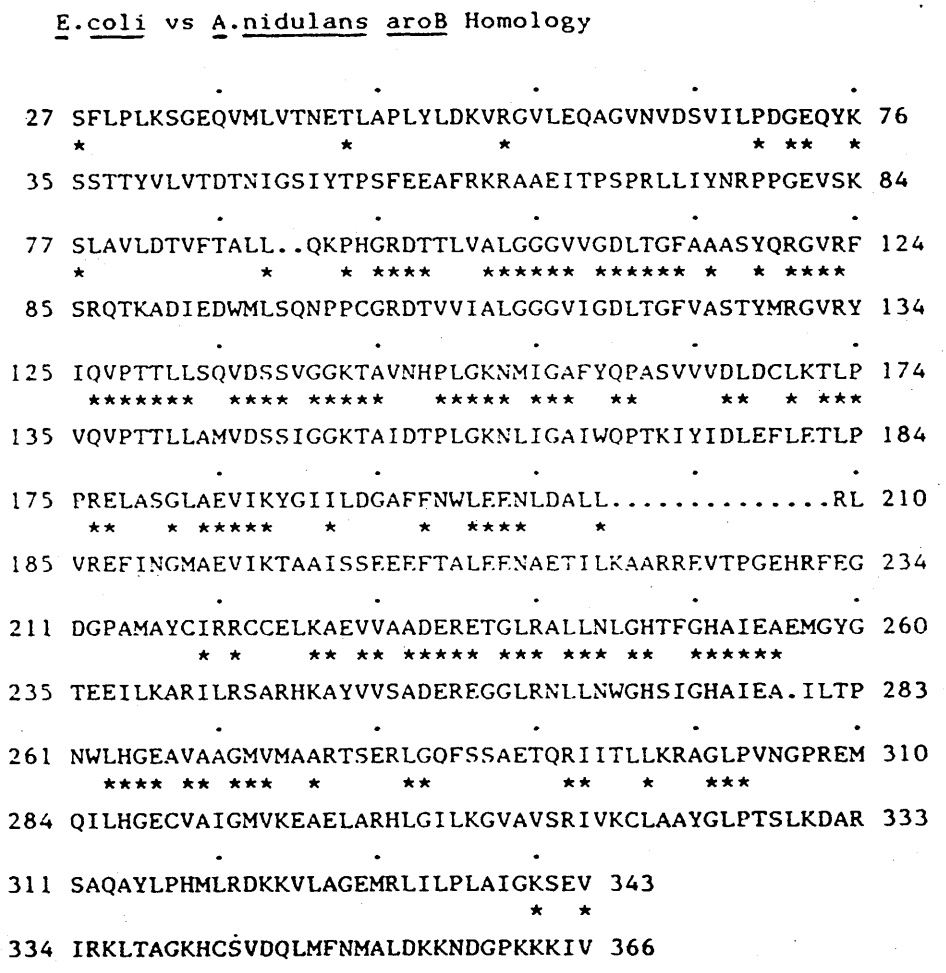


Figure 3.29: BESTFIT homologies between the E.coli DHQ synthase sequence (upper strand) and the A.nidulans arom sequence (lower strand, Charles et al., 1986).

residues present (219 — 233) in the A.nidulans sequence. Again this insertion is flanked by discrete regions of very high homology encompassing islands of acceptable conservative changes in a sea of consecutive identities. Only a single one-residue gap is required in the eukaryotic sequence to achieve this level of homology.

Interestingly the large gap required in the bacterial sequence (14) is almost half of the gap size required when the E.coli sequence is aligned with the S.cerevisiae arom sequence (Section 3.9.2).

3.9.4. S.cerevisiae/A.nidulans comparison

An apparent similarity between the initial 400 N-terminal residues of each of the eukaryotic arom polypeptides is inferred from the results presented in Sections 3.9.2 and 3.9.3. The exact degree of homology between the S.cerevisiae and A.nidulans arom proteins is shown in the BESTFIT comparison shown in Figure 3.30.

Both polypeptides are very similar. Overall between residues 14 to 390 of the S.cerevisiae and residues 15 to 382 of the A.nidulans, there are 197/367 (54%) direct identities. The observed homology increases to 60% when the 21 conservative changes are included. In some regions, there are runs as long as 75 residues with >90% identity (see residues 107 —182 of S.cerevisiae arom, Figure 3.30).

The difference in gap size required to align the E.coli DHQ synthase sequence with the S.cerevisiae (Section 3.9.2) and A. nidulans (Section 3.9.3) arom sequences, 27 residues

S.cerevisiae vs A.nidulans aroB Homology

```
 14 IIHVGYNIHDHLVETIIKHCPSSTYVICNDTN..LSKVPYYQQLVLEFKA 61
    **              *  * * ***   ***      *          *
 15 IIADFGLWRNYVAKDLISDCSSTTYVLVTDTNIGSIYTPSFEEAFRKRAA 64

 62 SLPEGSRLLTYVVKPGETSKSRETKAQLEDYLL..VEGCTRDTVMVAIGG 109
          *** *   *** **** ***  **  *     * ****  * **
 65 EITPSPRLLIYNRPPGEVSKSRQTKADIEDWMLSQNPPCGRDTVVIALGG 114

110 GVIGDMIGFVASTFMRGVRVVQVPTSLLAMVDSSIGGKTAIDTPLGKNFI 159
    *****  ****** ***** ***** ********************** *
115 GVIGDLTGFVASTYMRGVRYVQVPTTLLAMVDSSIGGKTAIDTPLGKNLI 164

160 GAFWQPKFVLVDIKWLETLAKREFINGMAEVIKTACIWNADEFTRLESNA 209
    ** ***     *   ****  ************** *    *** ** **
165 GAIWQPTKIYIDLEFLETLPVREFINGMAEVIKTAAISSEEEFTALEENA 214

210 SLFLNVVNGAKNVKVTNQLTNEIDEISNTDIEAMLDHTYKLVLESIKVKA 259
       *                              *          * *   **
215 ETIL.............KAARREVTPGEHRFEGTEEILKARILRSARHKA 251

260 EVVSSDERESSLRNLLNFGHSIGHAYEAILTPQALHGECVSIGMVKEAEL 309
     *** ****  ****** ******* ******* ******* ******* *********
252 YVVSADEREGGLRNLLNWGHSIGHAIEAILTPQILHGECVAIGMVKEAEL 301

310 SRYFGILSPTQVARLSKILVAYGLPVSPDEKWFKELTLHKKTPLDILLKK 359
     *  ***    * *  * * ***** *        **  *    * *
302 ARHLGILKGVAVSRIVKCLAAYGLPTSLKDARIRKLTAGKHCSVDQLMFN 351

360 MSIDKKNEGSKKKVVILESIGKCYGDSAQFV 390
    *  **** * *** * *  **  *   *  *
352 MALDKKNDGPKKKIVLLSAIGTPYETRASVV 382
```

Figure 3.30: BESTFIT homology between the S.cerevisiae (Duncan et al., 1986a; upper strand) and A. nidulans (Charles et al., 1986; lower strand) "aroB domains".

and 14 residues respectively, is again reflected in Figure 3.30. A gap of 13 residues in the A.nidulans sequence (interestingly 27 minus 14) ensures complete sequence alignment with the S.cerevisiae amino acid sequence. This consistency of gap requirement (and size) highlights two points. Firstly it acts as an indicator that the BESTFIT program is providing comparable data in all three cases. And secondly it precludes the possibility of a sequencing error resulting in the requirement for such a gap insertion. The data suggests that these observations reflect a fundamental and characteristic difference between all three sequences in this region which is therefore unlikely to be functionally important.

Coggins and Boocock (1986) have suggested that the two DHQ synthase domains of the Neurospora crassa arom polypeptide may play a role in maintaining the quaternary structure of the enzyme complex. A similar situation could also exist for both S.cerevisiae and A.nidulans arom complexes. Such a feature may reside within a non-catalytic portion of the protein structure and simply depend upon maintenance of suitable charged (or not) amino acid residues.

3.9.5. Significance of DHQ synthase homologies

One of the aims of this thesis, as outlined in Chapter One, was to examine the evolutionary relationship between the prokaryotic monofunctional shikimate pathway activities and the corresponding eukaryotic multifunctional activities. The BESTFIT comparisons detailed in the preceding sections can therefore be discussed in this wider context. This is

considered in Chapter 6 which draws together the results specifically presented in this Chapter and in Chapter 5 (shikimate kinase) in furtherance of this aim.

It is noteworthy how powerful and useful this approach is in identifying such homologies. When considered with the expanse of other biochemical and genetic data available (Chapters 1 and 6) it becomes obvious how significant the degree of relatedness between the prokaryotic and eukaryotic shikimate pathway activities really becomes.

3.9.6. Other *aroB* homologies

The *E.coli* DHQ synthase amino acid sequence was compared with sequences contained on the National Biomedical Research Foundation data base using the program WORDSEARCH (Section 2.18.3). No significant homologies were found. A comparison with the other *E.coli* shikimate pathway enzymes 3-dehydroquinase (Duncan *et al.*, 1986), shikimate dehydrogenase (Anton & Coggins, 1986), shikimate kinase (Millar *et al.*, 1986b; Defeyter and Pittard, 1986), EPSP synthase (Duncan *et al.*, 1984b) and chorismate synthase (White *et al.*, 1986) similarly failed to identify any extensive homologies.

Although differing in divalent cation requirements (Srinivasan *et al.*, 1963; Lambert *et al.*, 1985), a common feature of the DHQ synthase activities of *N.crassa* and *E.coli* is a catalytic requirement for NAD^+. The *E.coli* DHQ synthase sequence was searched by eye for any common structural motifs indicative of particular co-factor binding sites. It is

recognised that many nucleotide-binding proteins contain regions of conserved amino acid sequence (Wierenga and Hol, 1983; Walker *et al.*, 1982) involved in nucleotide binding. An example of such a motif (or 'fingerprint') has been described in detail for several NAD^+ and FAD linked enzymes (Hudson & Davidson, 1984).

The key structural feature implicated in this fingerprint model is a region of conserved amino acid sequence contained within a β-α-β secondary structure. Characteristically this conserved sequences contains 3 glycine residues arranged Gly-X-Gly-X-X-Gly, an invariant hydrophobic amino acid 4 residues preceding this sequence and an invariant negatively charged amino acid 18-20 residues after the last conserved glycine. In addition, several conserved hydrophobic residues forming the hydrophobic core of the unit are present within this 30-32 residue 'domain' (Wierenga & Hol, 1983).

This sequence has been shown to be conserved in several flavoproteins and dehydrogenases that bind FAD or NAD^+, including *E.coli* NADH dehydrogenases, human glutathione reductase, *E.coli* glutathione reductase, *E.coli* lipoamide dehydrogenase, *E.coli* chorismate mutase/prephenate dehydrogenase and dogfish lactate dehydrogenase (Hudson & Davidson, 1984 and references therein).

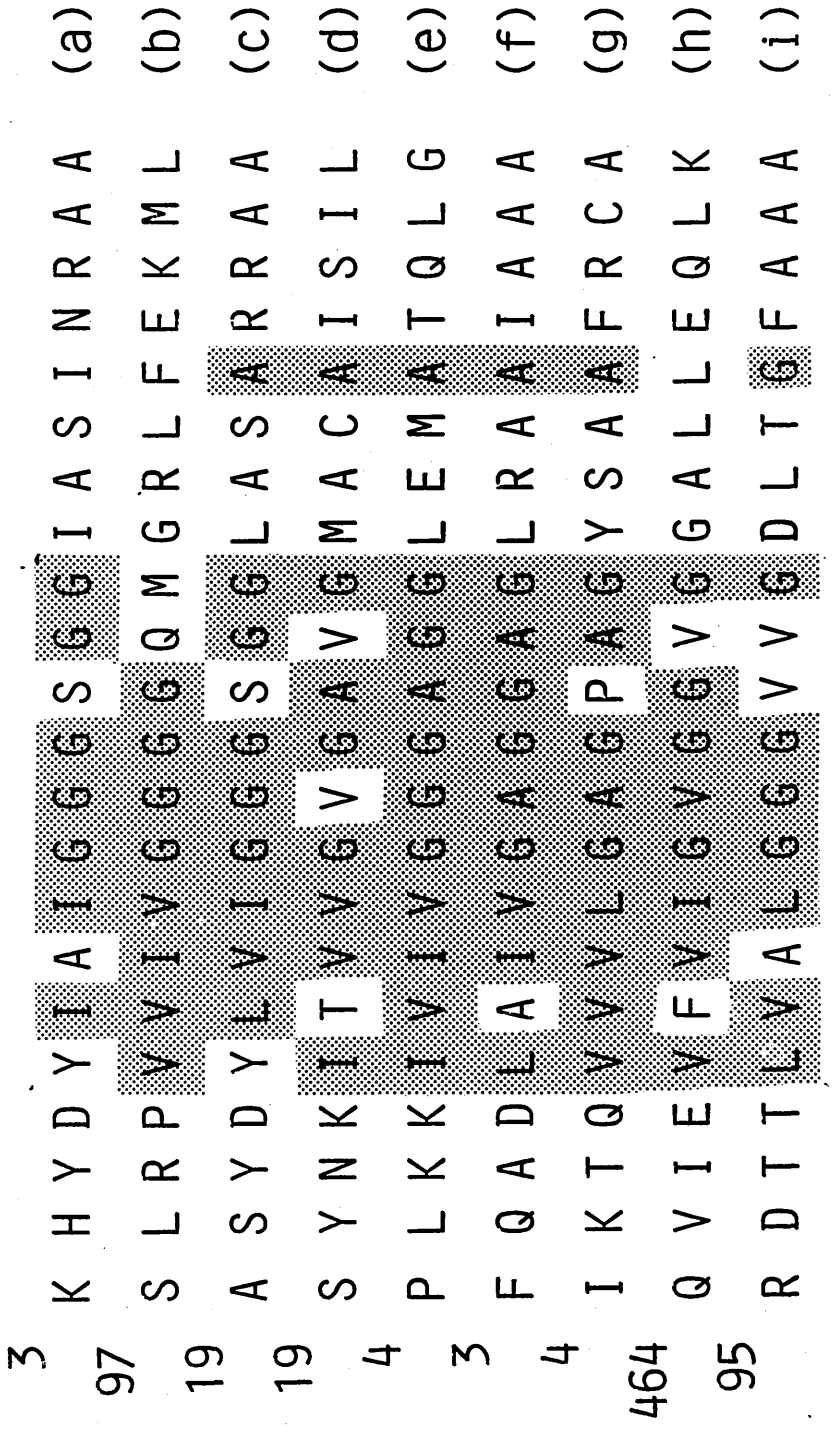
Although the *E.coli* DHQ synthase uses NAD^+ as a prosthetic group and not as a coenzyme *per se* (Maitra and Sprinson, 1978), a comparison of the enzyme sequence with the proposed motif reveals some interesting homologies. Residues 96-126 of

Figure 3.31 (facing):

Alignment of the adenine binding site of several flavoproteins and dehydrogenases with E.coli DHQ synthase (i) residues 95-116.

(a) E.coli glutathione reductase (Greer & Perham, 1986)

(b) - (h) are E.coli chorismate mutase/prephenate dehydrogenase; human glutathione reductase; lactate dehydrogenase (dogfish); E.coli NADH dehydrogenase; E.coli fumarate reductase; E.coli lipoamide dehydrogenase; and E.coli aspartokinase I/homoserine dehydrogenase I. (Hudson & Davidson, 1984 and references therein).


3 K H Y D Y I A I G G G S G G G I A S I N R A A A (a)
97 S L R P V V I V G G G G Q M G R L F E K M L (b)
19 A S Y D Y L V I G G G S G G L A S A R R A A (c)
19 S Y N K I T V V G G A V G G M A C A I S I L (d)
4 P L K K I V I V G G G A G G L E M A T Q L G (e)
3 F Q A D L A I V G A G G A G G L R A A I A A A (f)
4 I K T Q V V V L G A G P A G Y S A A F R C A (g)
464 Q V I E V F V I G V G G V G G A L L E Q L K (h)
95 R D T T L V A L G G G V V G D L T G F A A A (i)

the E.coli DHQ synthase sequences have a predicted β-α-β structure (Chou and Fasman, 1978). With the exception of the invariant negatively charged residue (the E.coli DHQ synthase sequence has a Gln residue), all of the conserved features of the proposed fingerprint model are present (Figure 3.31). Whether this sequence (or the observed variation from the consensus) has any functional significance is as yet unknown.

3.9.7. Future prospects

The determination of the primary structure of the E.coli DHQ synthase as described in this Chapter is the first step in unravelling a number of intellectually-challenging threads. The availability of milligram quantities of pure enzyme should allow X-ray crystallography to determine the 3-dimensional structure of the enzyme. The cloned gene should permit site-directed-mutagenesis to distinguish catalytically and structurally important residues. The exact cation requirement (Zn^{2+} or Co^{2+} or ?) can be directly examined. Finally it should be possible to resolve the catalytic mechanism of this unusual enzyme.

CHAPTER 4

CHORISMATE SYNTHASE FROM E.COLI K12: CLONING AND SEQUENCE ANALYSIS OF ITS GENE aroC

4.1 Introduction

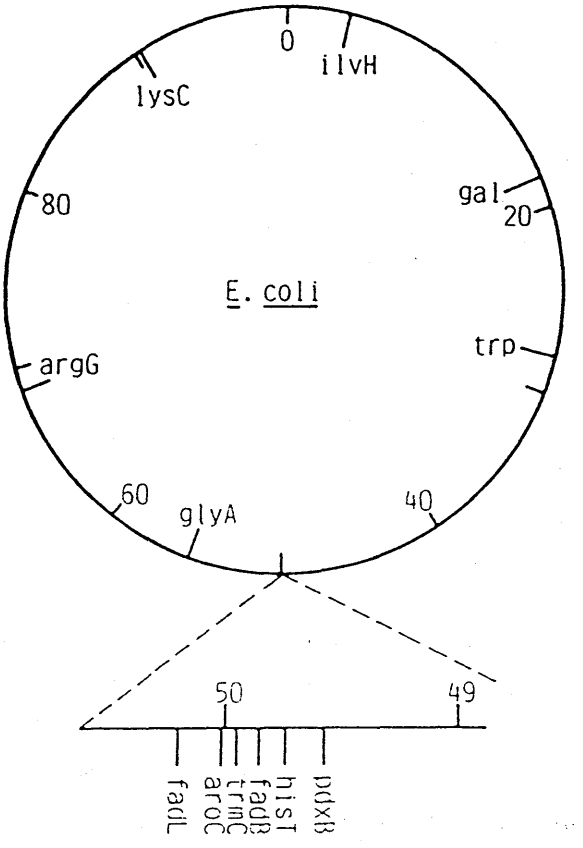
4.1.1 Previous work on the enzymology of chorismate synthase

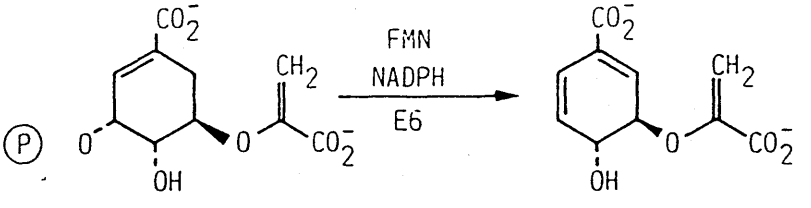
Chorismate synthase is by far the least well understood enzyme of the shikimate pathway (Figure 4.1). Purifications to homogeneity of chorismate synthase activities from *N.crassa* and *B.subtilis* have been reported (Welch *et al.*, 1974; Boocock, 1983; Hasan & Nester, 1978b). In contrast only a partial purification (approximately 4-fold), carried out under anaerobic assay conditions, has been reported for the *E.coli* enzyme, which by gel filtration criterion has a native M_r of 70-100,000 (Morell *et al.*, 1967).

Little is known structurally and mechanistically about this enigmatic enzyme except its sensitivity to molecular oxygen and absolute requirement for a reduced-flavin regenerating system (Morell *et al.*, 1967). In an attempt to redress this imbalance of factual information, this chapter reports the successful cloning and sequence analysis of the *E.coli* *aroC* gene which encodes chorismate synthase.

4.1.2 Location of the structural gene *aroC*

The *aroC* gene is known to map at around minute 50 (Figure 4.1) on the *E.coli* chromosome (Bachmann, 1983). The *aroC* locus has previously been used as a selectable marker to map an adjacent gene *trmC* (Hagervall & Björk, 1984) which encodes a tRNA(mnm^5s^2U) methyltransferase activity. Subsequent conjugational and transductional mapping experiments (Hagervall & Björk, 1984) established that the *aroC* and *trmC* genes





5-Enolpyruvyl-shikimate 3-phosphate

Chorismate

Figure 4.1: The chromosomal location of the E.coli aroC gene encoding chorismate synthase (E6).

were tightly linked (ca. 1.9 kbp). The aroC gene is located clockwise of trmC and fadL and the gene order at minute 50 is therefore hisT-fadB-trmC-aroC-fadL(prmB).

Previous work in this laboratory (I. Anton, unpublished work; Millar et al., 1986a) identified the Clarke & Carbon ColE1 plasmids pLC39-16, pLC26-33 and pLC33-1 (Clarke & Carbon, 1976) as potential aroC-carrying plasmids on the basis that they were known to carry the fadB marker. Only pLC33-1 was capable of complementing E.coli AB2849(aroC) (I. Anton, unpublished work).

4.2 Cloning the aroC gene

In cloning the aroC gene from a larger, lower copy-number plasmid pLC33-1 (ColE1), a similar direct approach was used to that described in Chapter 3 and Section 2.16 for the aroB gene. Complementation of the AroC$^-$ marker of E.coli AB2849 was used to identify positively those plasmids putatively carrying the intact aroC gene.

4.2.1 Identification of pLC33-1 as carrying aroC

I. Anton (unpublished work) had shown that when E.coli AB2849 (aroC) was transformed with the plasmid pLC33-1 (Clarke & Carbon, 1976) then the aromatic auxotroph was able to grow on unsupplemented media. Retransformation with a crude preparation of plasmid DNA (Section 2.11) confirmed that this transferable ability was a plasmid determinant.

Hagervall & Björk (1984) had identified a number of restriction sites in the locality of trmC and aroC on the TrmC$^+$ AroC$^+$ plasmids they had independently isolated from the Clarke &

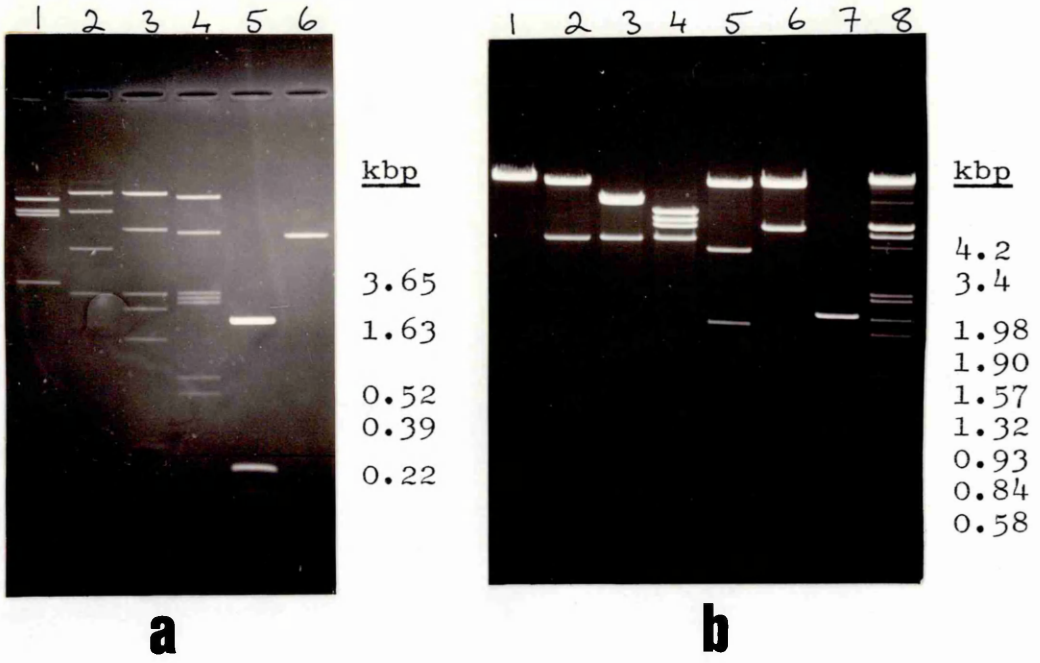
Figure 4.2 (facing)

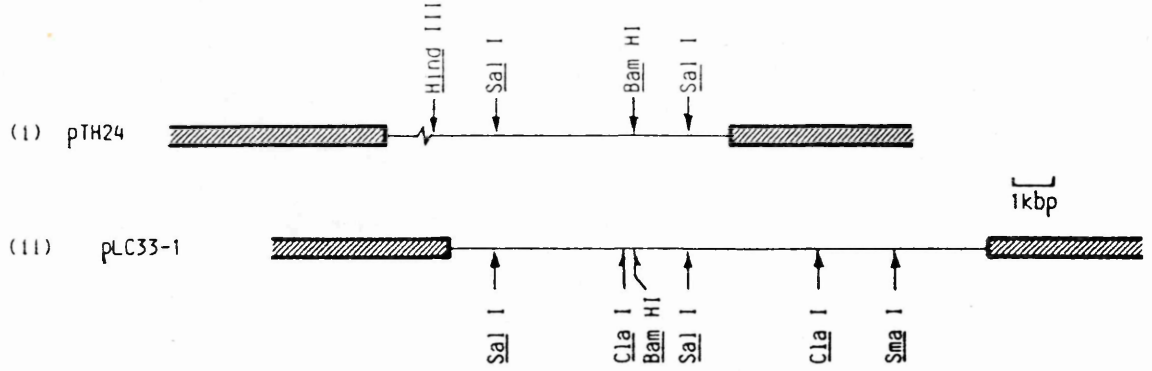
(a,b) 1% Agarose gel profiles of pLC33-1 restriction enzyme digestions. The lower figure summarises the pLC33-1 restriction mapping and comparison with the data of Hagervall & Björk (1984) is shown. The hatched and boxed areas represent vector DNA, *E.coli* genomic DNA is shown as a single line.

(a) Track 1 ClaI pLC33-1
Track 2 NruI pLC33-1
Track 3 PvuII pLC33-1
Track 4 HincII pLC33-1
Track 5 HinfI pAT153 (marker)
Track 6 EcoRI pAT153 (marker)

(b) Track 1 BamHI pLC33-1
Track 2 SmaI pLC33-1
Track 3 SmaI/BamHI pLC33-1
Track 4 SmaI/SalI pLC33-1
Track 5 SalI/BamHI pLC33-1
Track 6 SalI pLC33-1
Track 7 HinfI pAT153 (marker)
Track 8 EcoRI/HindIII λDNA (marker)

All digests were carried out as described in the text.


1 2 3 4 5 6
kbp
3.65
1.63
0.52
0.39
0.22
a
1 2 3 4 5 6 7 8
kbp
4.2
3.4
1.98
1.90
1.57
1.32
0.93
0.84
0.58
b


Hind III
Sal I
Bam HI
Sal I
(I) pTH24
1kbp
(II) pLC33-1
Sal I
Cla I
Bam HI
Sal I
Cla I
Sma I

Carbon gene bank (Clarke & Carbon, 1976). A comparison between the known restriction sites of the plasmid vector ColE1 (Dougan et al., 1978) and the restriction pattern of pLC33-1 (Figure 4.2(a)) indicated sites for ClaI, NruI, PvuII and HincII within the genomic DNA insert. In addition 2 SalI sites, defining a ca. 5kbp fragment of genomic insert DNA (gel (b), Figure 4.2), were identified.

Hagervall & Björk (1984) had tentatively assigned the aroC gene to a 5 kbp region of DNA bounded by SalI sites which contained an internal BamHI site. This mapping was on the basis of an estimated M_r for chorismate synthase (aroC gene product) of 70-100,000 (Morell et al., 1967) and not as defined by accurate sub-cloning. It was nevertheless clear that the genomic insert of pLC33-1 carried the aroC gene by complementation criterion, and displayed a pattern of restriction sites (Figure 4.2(b)) similar to those previously described for the area of genomic DNA specifying the trmC and aroC genes (Hagervall & Björk, 1984).

4.2.2 Construction of pGM601

Plasmid pLC33-1 was digested to completion with restriction enzyme ClaI and the products de-proteinised by phenol/chloroform extraction and ethanol precipitation (Section 2.13.1). The pattern of digestion products was examined by gel electrophoresis on a 1% agarose gel (Figure 4.2(a), gel(a) track 1).

This unfractionated mixture was ligated (Section 2.15) into ClaI-cut pAT153 (Twigg & Sherratt, 1980). The ligation mixture was used to transform $CaCl_2$-treated E.coli AB2849 (aroC) and following overnight growth at 37°C on LA + amp

(Section 2.16), transformants identified. Ampicillin resistant colonies were replica-plated on to unsupplemented media and Amp^rAroC^+ colonies selected after overnight growth at 37°C. Cloning into the ClaI site of pAT153 disrupts the tetracycline resistance gene therefore Amp^rAroC^+ phenotypes were also checked for Tet^s. The results are summarised in Table 4.1A.

A total of 5 colonies had $Amp^rAroC^+Tet^s$ phenotypes and were selected for further study, being designated AB2849/pGM601A —E. Mini-plasmid preparations (Section 2.11) were made from 10 ml LB + amp overnight cultures of each transformed strain. Digestion with restriction enzyme ClaI revealed that each culture contained a ClaI recombinant clone (Figure 4.3 gel(a)). A comparison between the digestion pattern of ClaI-cut pLC33-1 and ClaI-cut plasmid isolates indicated that each isolate at least had a 5.3 kbp ClaI fragment seen in the pLC33-1 digest (Figure 4.3(a)). Some of the isolates, pGM601 C, D, were obviously concatemers containing multiple cloned ClaI fragments of pLC33-1. Plasmid pGM601A (AB2849/pGM601A) was selected for further study and designated pGM601. Retransformation of E.coli AB2849 with this isolate and subsequent selection for $Amp^rAroC^+Tet^s$ confirmed the aroC marker as a plasmid determinant (data not shown).

The results of digestion of pGM601 with SalI and BamHI strongly suggested that this 5.3 kbp cloned ClaI fragment was derived from pLC33-1. The pattern of restriction sites within the genomic insert of pGM601 was identical to the corresponding ClaI region as shown in Figure 4.2(b). The

orientation of the cloned insert in pGM601 was such that the BamHI and SalI internal sites were located ca. 5.5 kbp and ca. 4.4 kbp respectively from their corresponding vector restriction sites. This spatial arrangement made it possible to delete these fragments and so map the aroC coding region more precisely.

4.2.3 Construction of pGM602

Plasmid pGM601 was digested with either SalI or BamHI and the products separated on a 1% LMT agarose gel (Section 2.13.2). A ca. 4.5 kbp SalI-derived band and a ca. 3.4 kbp BamHI-derived band were each excised and their respective DNA purified as described in Section 2.13.2. The two DNA preparations were recircularised (Section 2.15.2) by intramolecular ligation, and the resultant ligation mixtures used to transform $CaCl_2$-treated E.coli AB2849. Antibiotic (ampicillin resistant transformants were selected and tested for growth on unsupplemented minimal media. Prototrophs ($Amp^r AroC^+ Tet^s$) were only obtained from the SalI recircularised derivative of pGM601 (Table 4.1B). Not surprisingly, the BamHI deletion (of pGM601), which contained only 0.1 kbp of cloned genomic DNA was $aroC^-$. SalI-deleted pGM601 isolates were termed pGM602. Eight pGM602 (A — H) isolates were prepared and checked for plasmid DNA. All eight when digested with ClaI and SalI gave rise to a band of 1.65 kbp as expected (Figure 4.3).

(A)

Ligation	Colonies on LA + amp	Colonies on MM
ClaI-mix (pGM601)	41	5
ligation control	50	0
20 ng ccc. pAT153	ca. 200	0

(B)

Recircularisation	Colonies on LA + amp	Colonies on MM
SalI/pGM601 deletion	9	9
BamHI/pGM601 deletion	3	0
20ng ccc.pAT153	ca. 200	0

Table 4.1: Transformation and complementation of E. coli AB2849.

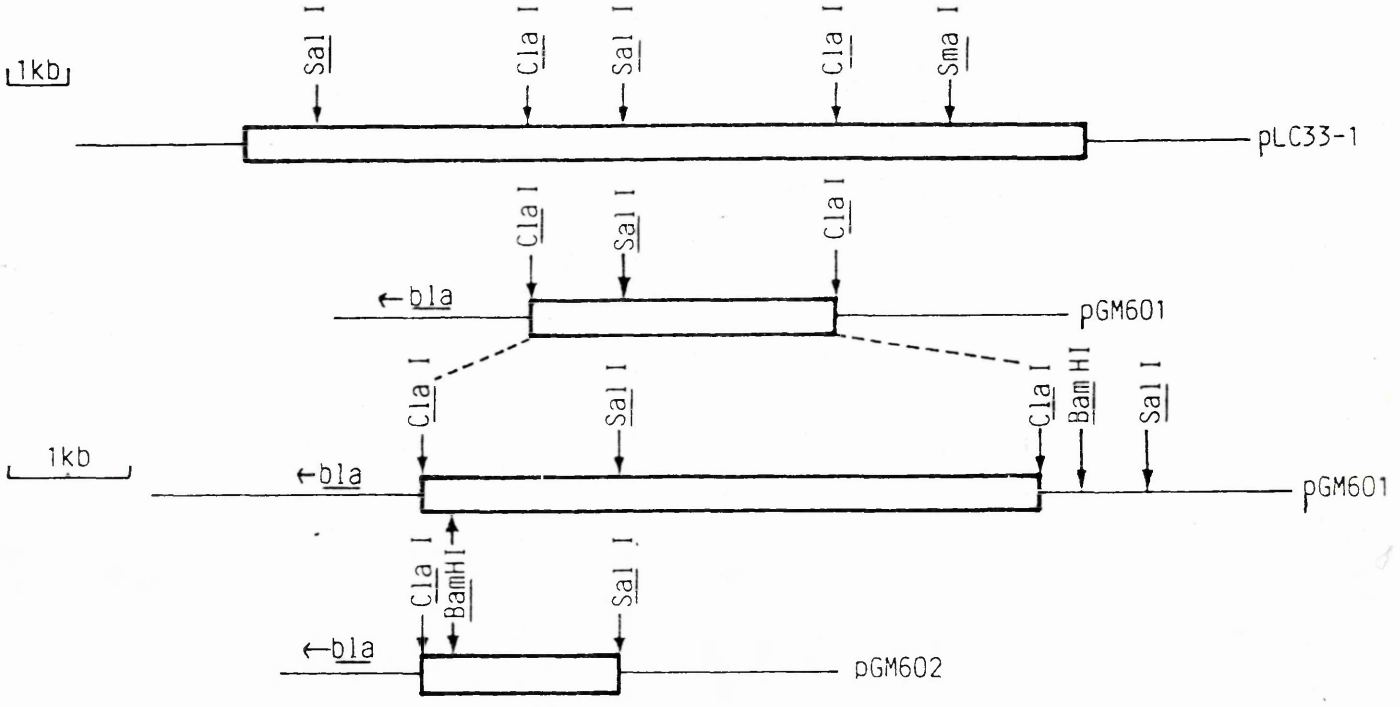
Figure 4.3 (facing)

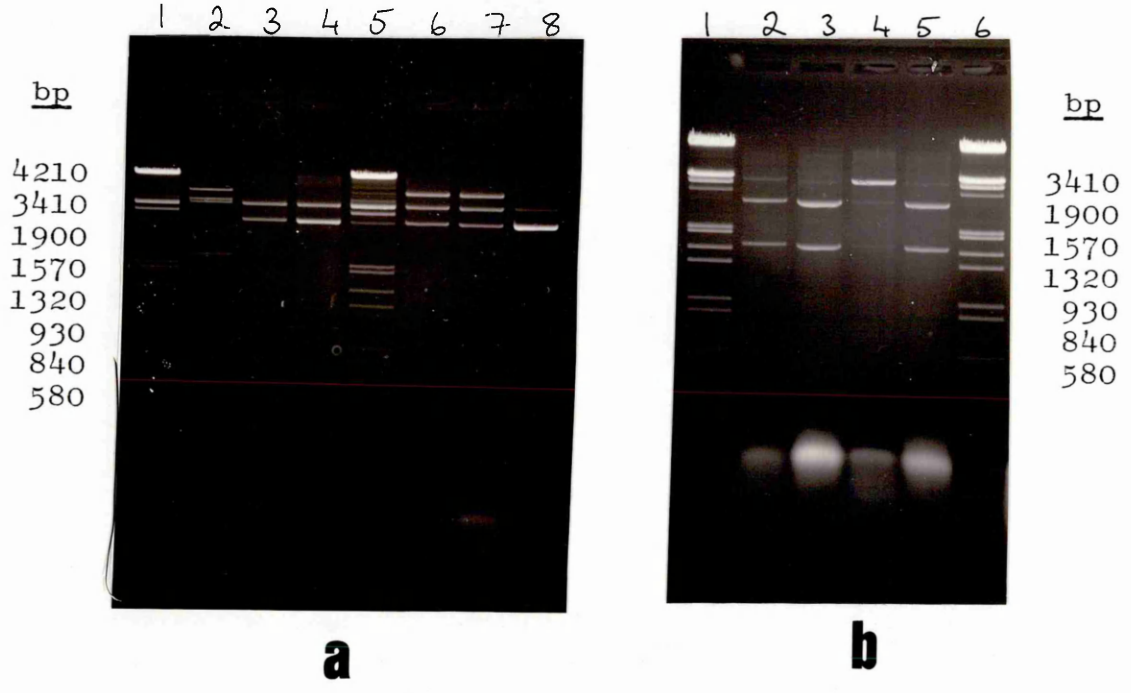
Sub-cloning strategy and relationship of *aroC*-plasmids pLC33-1, pGM601 and pGM602.

(Below) 1% Agarose gel profiles of restriction digests of rapid DNA preparations of (a) *E.coli* AB2849/pGM601 and (b) *E.coli* AB2849/pGM602.

(a)	Track 1 & 5	*Eco*RI/*Hind*III λDNA (marker)
	Track 2	*Cla*I pLC33-1
	Track 3	*Cla*I pGM601A
	Track 4	*Cla*I pGM601B
	Track 6	*Cla*I pGM601C
	Track 7	*Cla*I pGM601D
	Track 8	*Cla*I pGM601E
(b)	Track 1 & 6	*Eco*RI/*Hind*II λDNA (marker)
	Track 2	*Cla*I/*Sal*I pGM602A
	Track 3	*Cla*I/*Sal*I pGM602B
	Track 4	*Cla*I/*Sal*I pGM602C
	Track 5	*Cla*I/*Sal*I pGM602D

All digests were carried out as described in the text.


1kb
Sal I
Cla I
Sal I
Cla I
Sma I
pLC33-1
Cla I
Sal I
Cla I
←bla
pGM601
Cla I
Sal I
Cla I
BamHI
Sal I
1kb
←bla
pGM601
Cla I
BamHI
Sal I
←bla
pGM602


1 2 3 4 5 6 7 8
bp
4210
3410
1900
1570
1320
930
840
580
a
1 2 3 4 5 6
bp
3410
1900
1570
1320
930
840
580
b

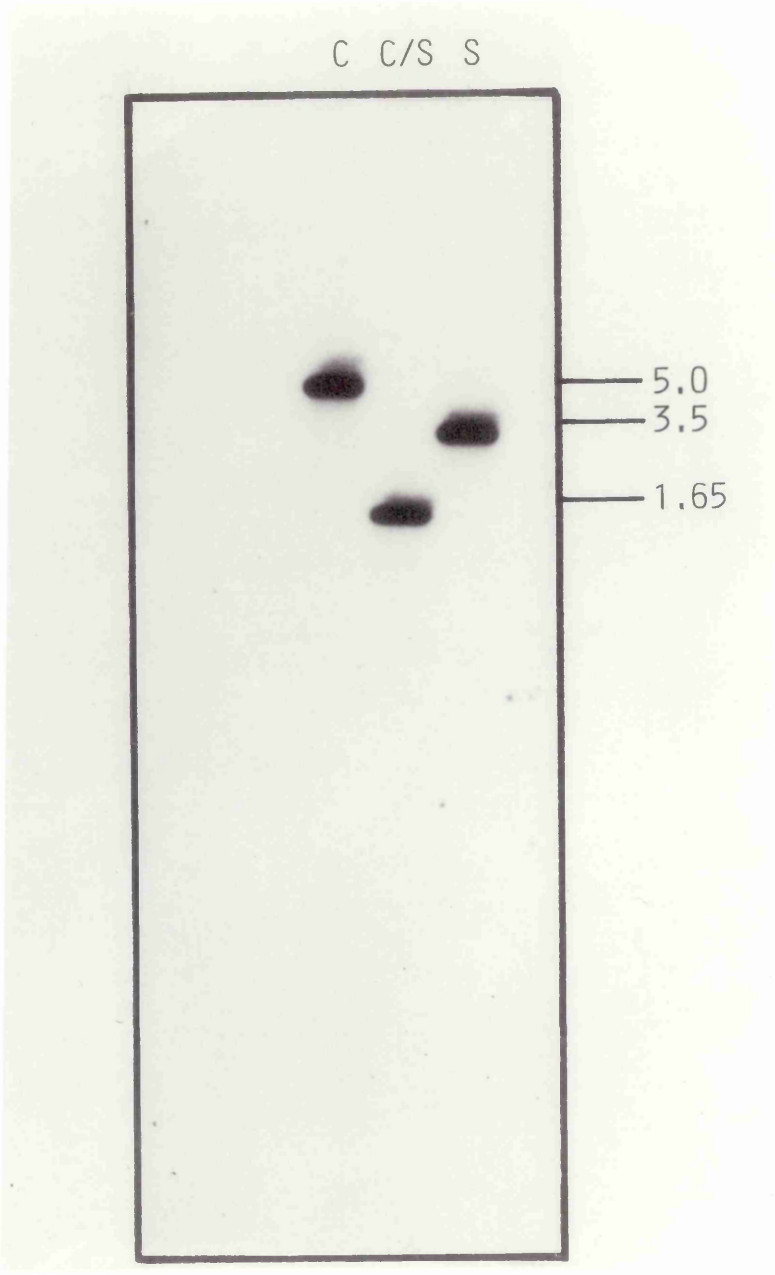


Figure 4.4: Southern hybridisation of the genomic aroC gene as described in Section 4.2.4. DNA was digested with either ClaI (C); SalI (S) or both (C/S). Molecular sizes (kbp) are shown.

The deletion of some 3.65 kbp of genomic material cloned in pGM601 (SalI-deletion) clearly did not affect the ability of the remaining cloned DNA to complement the nutritional deficiencies of (aroC) E.coli AB2849. The aroC structural gene must therefore be located within this 1.65 kbp fragment (Figure 4.3(b)) cloned in pGM602.

4.2.4 Genomic organisation of the aroC gene

Complementation by relief of auxotropy in this case is in a $recA^+$ background (E.coli AB2849). Similarly selection for $TrmC^+AroC^+$ (Hagervall & Björk, 1984) was in an $aroC^-$ $recA^+$ background. The possibility of a plasmid:chromosomal rearrangement always exists under these conditions, and so the genomic organisation of the aroC gene was investigated.

E.coli chromosomal (5 µg, high molecular weight) DNA was digested with either ClaI, SalI or ClaI + SalI and the products separated on a 0.8% agarose gel. The DNA was transferred to nitrocellulose and immobilised as detailed in Section 2.20.1. The chromosomal hybridisation pattern was investigated using the 1.65 kbp ClaI/SalI genomic insert of pGM602 (Figure 4.3(B)) as a probe. This was radiolabelled by nick-translation using ^{32}P α-dCTP to a specific activity of 2×10^8 dpm/µg and hybridised under high stringency conditions. All procedures are as outlined in Sections 2.20 and 2.21. The result is shown in Figure 4.4.

As expected the probe hybridises to a 1.65 kbp ClaI/SalI band (track C/S) and to a ca. 5.2 kbp ClaI band (track C). Surprisingly the probe finds not a 5 kbp SalI band as expected

but rather a 3.5 kbp SalI fragment (track S). The intensity of the hybridisation signal is equal for all three tracks (Figure 4.4).

The aroC gene (by complementation) is located within a 1.65 kbp ClaI/SalI fragment of genomic DNA as cloned in pGM602. 'Downstream' (to the right in Figure 4.3) is a ClaI site originally used in cloning the aroC gene in construct pGM601. This organisation is reflected on the chromosome (tracks C and C/S), Figure 4.4). The SalI site 'upstream' (to the left in Figure 4.3) appears to be 1.5 kbp distal to its location on the chromosome (track S, Figure 4.4).

Accurate restriction mapping of this region of the chromosome (Figure 4.2, Hagervall & Björk, 1984) independently confirms the location of the aroC gene on a 5 kbp SalI fragment. Clearly therefore either in the construction of pLC33-1 (Clarke & Carbon, 1976) or in the propagation of the plasmid in $recA^+$ strains some form of artificial construct has arisen. The former seems unlikely since the original ColE1 hybrid plasmids were made by the poly dA-dT tailing procedure and concatemers are normally not observed. The latter again seems improbably since pLC33-1 was isolated in this laboratory from a recA strain prior to mapping and sub-cloning in $RecA^+$ auxotrophs. One possibility is that the SalI site present 3.5 kbp from the ClaI site in the genome has been lost (by mutation?) at some stage of the initial isolation and/or propagation. This would explain the detection of a 3.5 kbp genomic SalI fragment by hybridisation (Figure 4.4) and a 5 kbp cloned SalI fragment as aroC-carrying regions.

4.2.5 Deletion analysis of pGM602

Deletion of ca. 1.4 kbp of genomic insert of pGM602 (BamHI-deletion Section 4.2.3) abolishes the plasmid's ability to complement aroC auxotrophs. This is not surprising (assuming the intact aroC gene is carried on pGM602) since this BamHI-deletion leaves only approximately 150 bp of genomic material cloned in the plasmid. To construct a series of more informative deletions pGM602 was accurately mapped with a number of restriction enzymes. Sites for EcoRV and NruI (in addition to the BamHI site) were located within the cloned genomic insert. The results of these deletion experiments are summarised in Figure 4.5.

Construct pGM603 was generated by removal of a ca. 450 bp NruI-SalI fragment of cloned E.coli DNA. Further deletion towards the ClaI end of the clone (1 kbp EcoRv-SalI deletion) gave plasmid pGM604. Complementation of E.coli AB2849 (aroC) was retained by the former but lost in the latter. This further defined the limit of DNA (E.coli) required for relief of auxotrophy of the aroC mutant, to a 1.2 kbp ClaI/NruI region of genomic DNA.

4.3 Overexpression of chorismate synthase from the cloned aroC gene

In parallel with the sub-cloning analysis of the aroC gene, work in this laboratory has concentrated on the development of a novel assay and purification procedure for the E.coli chorismate synthase (Millar et al., 1986a; White et al., 1986).

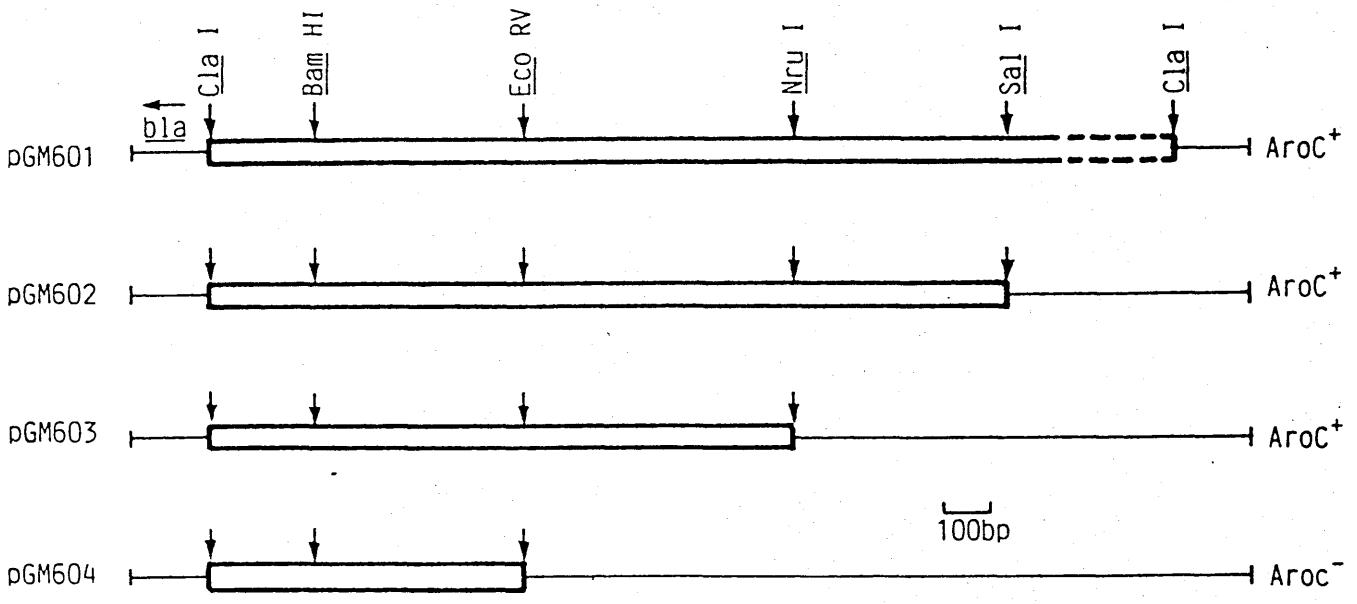


Figure 4.5: aroC deletion analysis. E.coli genomic DNA is boxed, plasmid vector DNA is represented by solid lines. Successive deletions employed in constructing pGM604 from pGM601 are shown and their resultant phenotypes when harboured in E.coli AB2849.

4.3.1 Overexpression in crude extracts

A rapid partial purification of chorismate synthase from both wild-type (*E.coli* K12) and overproducing cells (*E.coli* AB2849/pGM602) has been developed (P.J. White, unpublished work, Millar *et al.*, 1986a). Following FPLC separation, the specific activity of fractions obtained from pGM602-containing cell extracts was found to be ten-fold higher than similar fractions from wild-type cells. A noticeable feature in SDS PAGE analysis of the purified fractions obtained from *E.coli* AB2849/pGM602 is a protein band at M_r 40,000. This subunit molecular weight is consistent with the protein coding potential with the cloned genomic insert in pGM602 (1.65 kbp). Indeed the deletion analysis limit of 1.2 kbp for *aroC* complementation (pGM603) would theoretically fully encode a 40 kDa protein.

4.3.2 Expression from *tac-aroC* construct pGM605

The direction of transcription of *aroC* was examined by correlating the IPTG-induced elevation of *aroC* expression in *tac-aroC* constructs, with the orientation of the cloned insert.

The *Nru*I site identified in pGM603 (Figure 4.5) as essentially outwith the *aroC* coding region was used as a blunt cloning end. Plasmid pGM602 was digested with *Nru*I and *Hinc*II, and a fragment extending from the genomic *Nru*I site to the vector *Hinc*II site (1.73 kbp) was isolated and purified. This was ligated into *Sma*I cut pKK223/3 (see Chapter 3), and used to transform *E.coli* AB2849. Seventeen resultant amp^r colonies were tested for growth on unsupplemented

media. All were found to be $Amp^{r}AroC^{+}$, thirteen were selected for further study. The two possible orientations in which the NruI-HincII fragment could be cloned were resolved by the pattern of a BamHI digest of each of the putative tac-aroC plasmids. Clones with the NruI end nearest the tac promoter were characterised by a 705 bp BamHI fragment, conversely a 1030 bp BamHI fragment indicated a tac-proximal HincII site (see Figure 4.6B(i) & (ii)).

Mini-DNA preparations were made for each of the 13 tac constructs and each was digested with BamHI. Eight clones were tac-HincII proximal, four were internal concatemers and one was tac-NruI proximal. This latter plasmid was termed pGM605 and tested for lac inducible aroC overexpression.

E.coli AB2849/pGM605 cells were grown on a preparative scale following induction with IPTG (See Sections 2.23.1 and Chapter 3.7.2) and a crude extract prepared. Assay for chorismate synthase activity indicated that cells harbouring this plasmid were overproducing chorismate synthase some 400-fold relative to E.coli K12 level (P.J. White, unpublished work). SDS PAGE analysis of crude extracts (Figure 4.7) clearly shows a major protein band at ca. 40 kDa constituting some 5% of total soluble protein. A comparable result for tac-aroC in the opposite orientation (tac-HincII proximal) could not be obtained. Expression of aroC would appear to be from the NruI site towards the ClaI end of the genomic insert of pGM602.

Figure 4.6 (facing)

(Upper) 1% Agarose gel profile of putative tac-aroC constructs (a).

(Lower) The alternative possible orientations of the HincII/NruI fragment of pGM602 in expression vector pKK223/3 (see text).

(c) Track 1 BamHI tac-aroC (i)
Track 2 HinfI/EcoRI pAT153 (marker)
Track 3 BamHI tac-aroC (ii)
Track 4 BamHI tac-aroC (iii)

tac-aroC (i) has the insert in the correct orientation (706 bp fragment see (i) below).

All digests were carried out as described in the text.

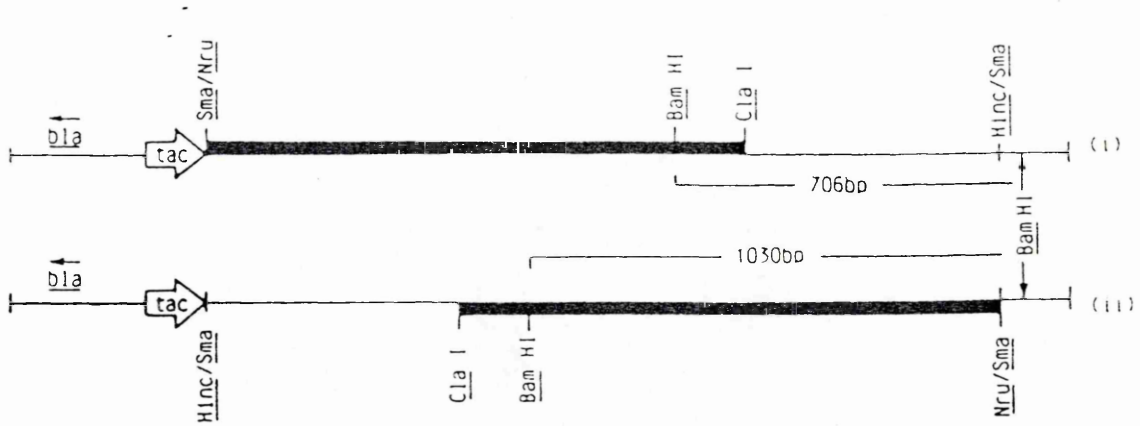
bp

1030
706

bp

1000
631
517
396
298
200

(a)



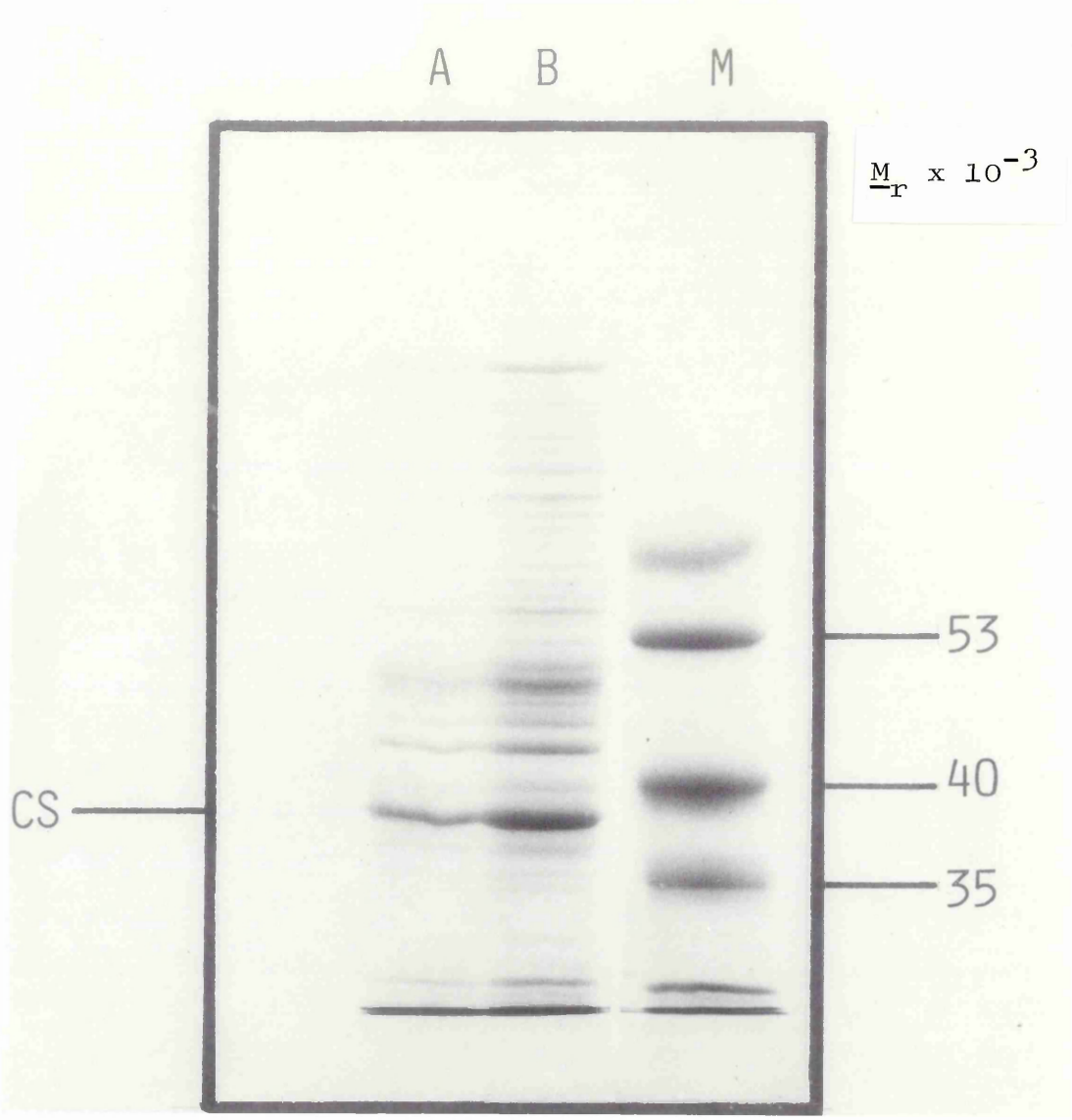


Figure 4.7: Coomassie stained 12.5% SDS PAGE analysis of crude extracts from E.coli AB2849/pGM605 (tac--aroC). Tracks A (20 µg) and B (40 µg) of protein were electrophoresed alongside marker proteins (M). The putative overexpressed chorismate synthase (CS) is indicated and discussed in Section 4.3.2.

4.3.3. *In vitro* coupled transcription-translation of pGM602

2.5 μg of pure plasmid pGM602 was substrate for the prokaryotic DNA-directed coupled transcription-translation system detailed in Section 2.26. Similarly protein products from 2.5 μg of pure pAT153 were expressed and analysed by SDS PAGE as an internal control. The results are shown in Figure 4.8.

A no DNA control showed no evidence of incorporation of L[^{35}S]-methionine in the autoradiograph (Figure 4.8). Both pGM602- and pAT153-directed expression show a number of protein products, separated by the 12.5% denaturing gel, with high levels of ^{35}S incorporation. Both plasmids (pGM602 and pAT153) characteristically show a major 30 kDa band corresponding to the *bla* gene product, the β-lactamase (*amp*^{r}, Figure 4.8). In addition pGM602 is expressing a protein of *ca.* 39 kDa subunit molecular weight (*aroC*, Figure 4.8).

4.4 DNA sequence analysis of the 1.65 kbp *Cla*I-*Sal*I insert of pGM602

4.4.1 Sequencing strategy

The presence of *Bam*HI, *Eco*RV and *Nru*I restriction sites within the 1.65 kbp *Cla*I/*Sal*I genomic insert of pGM602 (Figure 4.5) simplified the choice of strategy. Each of these restriction sites could be used as potential cloning sites from which to initiate and proliferate defined sequence data. An additional shotgun approach of cloning randomly generated *Hpa*II, *Taq*I and *Sau*3A sub-fragments of pGM602's cloned insert, was also performed in parallel to this definitive cloning/

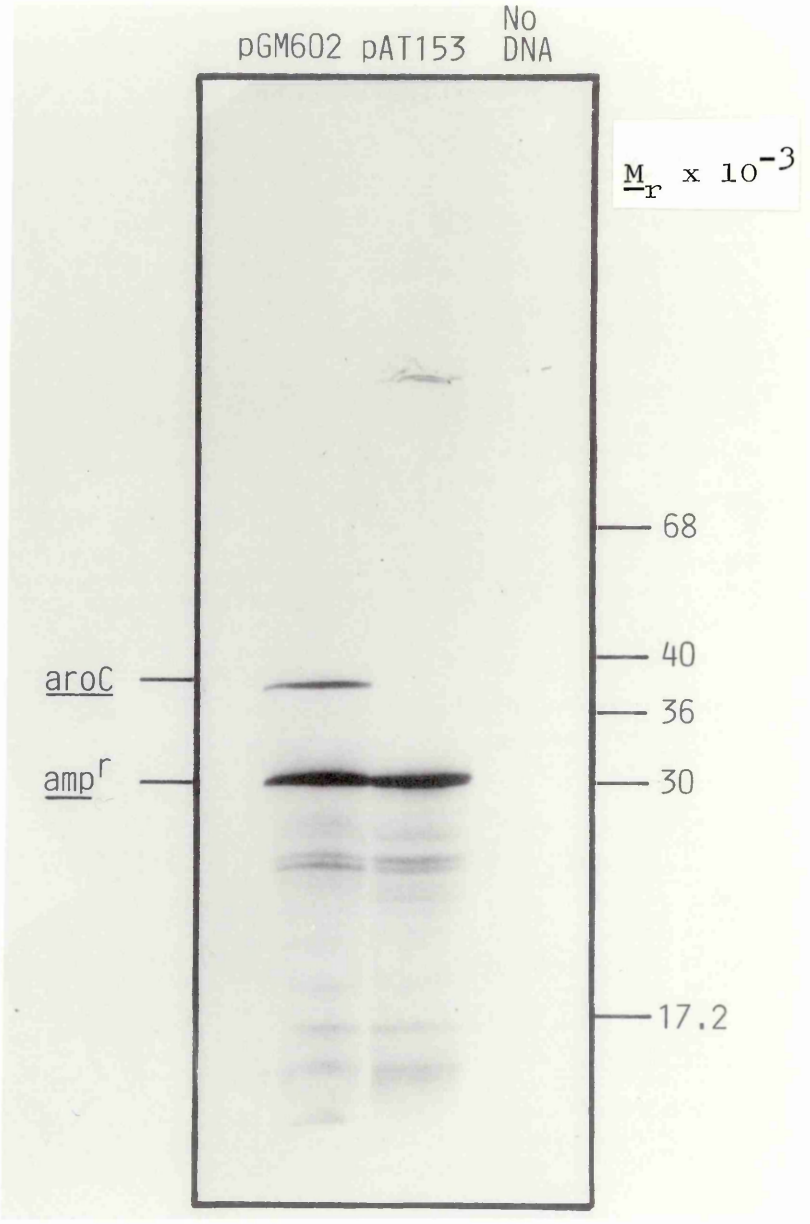


Figure 4.8: In vitro expression of pGM602 and pAT153 as described in Section 4.3.3. ^{35}S-Met incorporation into E.coli chorismate synthase (aroC) and vector-encoded β-lactamase (amp^r) is indicated. Molecular weights are shown ($\times 10^{-3}$)

sequencing. This was particularly apt since the ClaI end of pGM602 would also generate suitably cohesive ends enabling its cloning into AccI-prepared M13mp8RF (for HpaII and TaqI ligations).

4.4.2 Sub-cloning strategy

Plasmid pGM602 (1 µg) was digested with EcoRI and SalI and a ca. 1.65 kbp fragment purified from 1% LMT agarose by phenol/chloroform extraction (Section 2.13.2). Similarly 1 µg of plasmid was digested with BamHI and PstI and a ca. 950 bp fragment prepared. This fragment extended from the vector PstI through the ClaI cloning end to the genomic BamHI. The EcoRI/SalI DNA was ligated into suitably prepared M13mp8RF and the PstI/BamHI DNA fragment ligated to PstI/BamHI-cut M13mp9RF (Sections 2.17.2 and 2.17.3).

A ca. 500 bp BamHI/EcoRV fragment of the genomic insert of pGM602 (Figure 4.5) was prepared by double restriction enzyme digestion of pGM602 (1 µg) with the corresponding enzymes and the fragment isolated. This was cloned into BamHI/SmaI prepared M13mp8RF. In a directly analogous fashion, the EcoRV/NruI (550 bp) fragment of E.coli pGM602 insert DNA was cloned in SmaI-cleaved M13mp8RF. A ca. 750 bp DNA fragment generated by NruI digestion of pGM602 (1 µg), and extending from the genomic NruI site to the vector NruI site (within tet^r gene), was ligated to SmaI-cut M13mp8RF (Figure 4.5).

4.4.3 Distribution of HpaII, TaqI and Sau3A sites within the cloned insert of pGM602

A pool containing approximately 2 µg of the 1.65 kbp E.coli genomic insert in pGM602 was made by double restriction enzyme digests (ClaI/SalI) and preparative agarose gel electrophoresis (Sections 2.13.2). This pool was divided into three equal aliquots prior to digestion with one of HpaII, TaqI or Sau3A. A total of 50% of each secondary digest was purified (Section 2.13.1) for subsequent ligations and the remainder analysed by 2% agarose gel electrophoresis (Figure 4.9).

HpaII digestion produced a series of bands (6) within the 0.15 kbp to 0.35 kbp range (Figure 4.9) suggesting at least 5 internal HpaII sites. This distribution was ideal for sequence representation of the majority of the cloned insert. Digestion with TaqI and Sau3A was incomplete (Figure 4.9) but still provided useful data on the size distribution of these sites within the aroC clone. TaqI fragments of 250 bp and 390 bp were obvious with fainter bands at 410 bp and 630 bp corresponding to possible partial digests (Figure 4.9). The pattern of Sau3A digestion suggested two large fragments of 550 bp and 650 bp, a smaller band at approximately 170 bp and a number of smaller unseen bands inferred by the remaining non-allocated DNA.

All three unfractionated mixes (HpaII, TaqI and Sau3A) were ligated into either AccI-cut M13mp8RF or BamHI-cut M13mp8RF (Sau3A mix only). Following transformation of

Figure 4.9: 2% Agarose gel profile of secondary restriction digests of the 1.60 kbp ClaI/SalI insert of pGM602. Insert DNA was prepared as described in the text (Section 4.4.3) and digested with TaqI (Track 3), HpaII (Track 4) or Sau3A (Track 5). Marker sizes of EcoRI/HindIII λ (Track 1), EcoRI/HinfI pAT153 (Track 6) and HinfI pAT153 (Track 2) are shown in bp. All digests were carried out as described in the text.

$CaCl_2$-treated E.coli TG1 (Section 2.17.4) by these ligation mixes (and others described in Section 4.4.2 above) recombinant bacteriophage were identified. Conditions for single stranded template preparation, annealings and sequencing reactions were as described in Chapter Two. In total 171 individual single stranded templates were prepared and analysed by DNA sequencing.

4.4.4 1st round of DNA sequencing

(i) Construction of the BamHI/PstI and BamHI/EcoRV recombinant M13 clones described in Section 4.4.2 allowed sequence determination towards and away from the ClaI end of pGM602 respectively. The 223 bp separating the BamHI and ClaI sites contained 1 HpaII site 18 bp from the ClaI end.

Excluding the Sau3A site within the BamHI recognition sequence (5'GGATCC3'), there were an additional two Sau3A sites 15 bp and 31 bp from the BamHI site (Figure 4.10).

(ii) The ca. 1.65 kbp EcoRI/SalI M13 clone described above (Section 4.4.2) allowed sequence determination to extend for 210 bp from the SalI end of the clone towards the ClaI site. Two TaqI and two Sau3A sites were identified within this region. Perhaps more importantly a HpaII site at 204 bp from the SalI end was located (Figure 4.10).

(iii) Sequencing in both directions from the single NruI site identified some 350 bases in total. Determination of 250 bases running NruI $\longrightarrow$ SalI overlapped the SalI sequence intimated above (ii) and confirmed the NruI site at exactly 440 bp from the SalI cloning end (Figure 4.10). A HpaII site at 345 bp from the SalI end was added to the growing map.

Sequencing from the NruI site towards the ClaI identified an additional 100 bases including a TaqI site at position 572 (SalI is position 0).

(iv) Finally, sequencing from the EcoRV site towards SalI end identified over 200 base pairs with no internal HpaII, TaqI or Sau3A sites (Figure 4.10).

4.4.5 2nd round of DNA sequencing

Fifty eight recombinant M13 single stranded templates were prepared from the HpaII digest mix cloned in M13mp8RF (Section 4.4.3). Primary screening by A-track analysis (Section 2.17.10) revealed a minimum of 16 different classes of sequence information. Fourteen of these were derived from seven fragments sequenced in both directions. The remaining two classes were very small fragments (< 20 bp) one of which corresponded to 18 bp region between the ClaI end and the nearest HpaII site (Section 4.4.5(i)). Five of the seven fragments sequenced on both strands could be 'placed' from the known HpaII sites identified in Section 4.4.5 (Figure 4.10, 'HpaII').

From these data > 90% of the sequence within the 1.65 kbp insert of pGM602 could be assigned and confirmed on both strands (Figure 4.10).

4.4.6 3rd round of DNA sequencing

Initial screening of the 34 TaqI generated recombinant M13 clones by single track sequencing (Section 2.17.10) indicated 12 unique classes of sequence data. Six such clones

were completely overlapping opposite strands of three individual TaqI fragments. One of these pairs allowed unambiguous determination of the DNA sequence to within 50 bp of the SalI site. The remaining 50 bp was only sequenced (though several times on distinct clones) on one strand (Figure 4.10).

Of the remaining six types of TaqI sequences, four were non-overlapping opposite strands of two distinct TaqI fragments. The other two were sequences extending in opposite directions for a single TaqI site (Figure 4.10).

The determination of sequence towards the ClaI from TaqI sites at positions 572, 672, 747, 1010 and 1049 respectively (Figure 4.10, Figure 4.11) essentially completed the alignment of the fully sequenced HpaII clones (Section 4.4.5).

4.4.7 4th round of DNA sequencing

Final confirmation of some dubious base sequences was accomplished by sequence analysis of four Sau3A specified M13 clones (Figure 4.10, 'Sau3A'). The existence of internal Sau3A sites as predicted by restriction mapping (Section 4.4.3) and direct identification in sequence data (Section 4.4.4) was confirmed by cloning and sequencing from four of these sites. The data obtained exactly complemented other information obtained in Sections 4.4.4 to 4.4.6 (Figure 4.10).

4.4.8 Compilation of DNA sequence data

DNA sequences were entered as separate files into the PDP 11-34 computer using the programs detailed in Section 2.18.1. Sites used in cloning were the defined start and

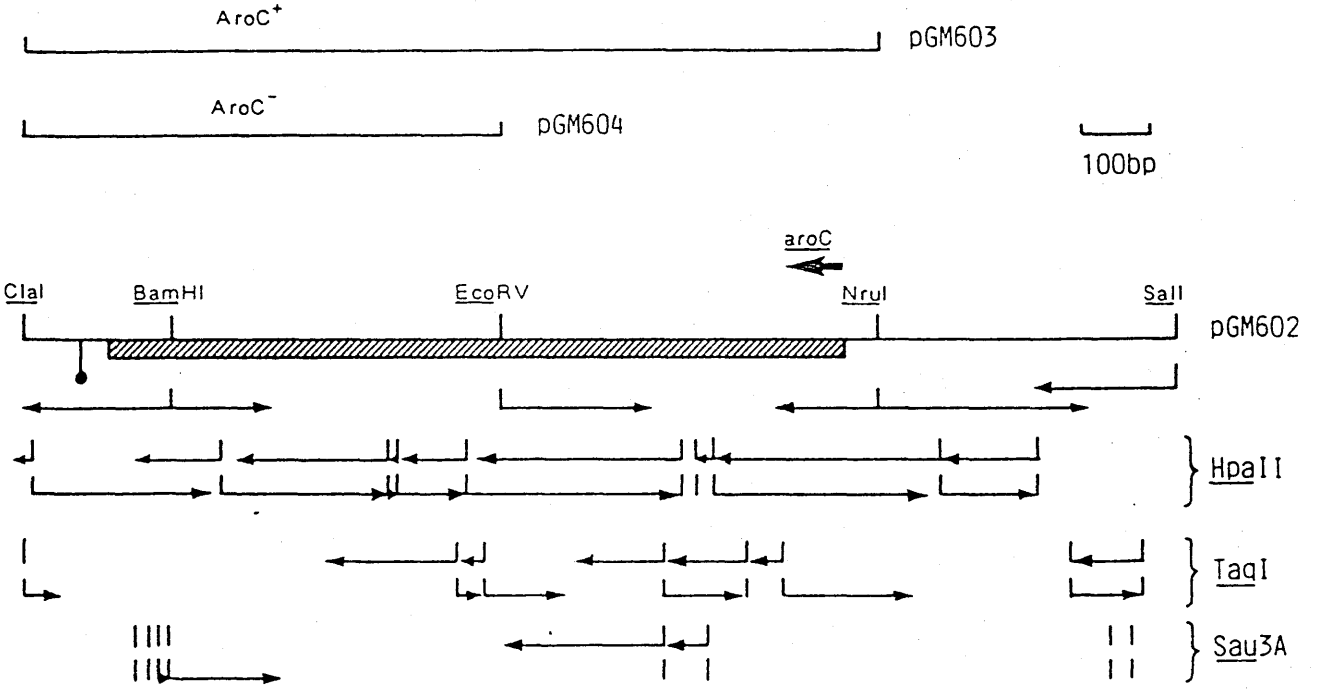


Figure 4.10: DNA sequencing strategy for the aroC gene (hatched area) within the cloned genomic insert of pGM602. The location of a potential terminator downstream of the gene is indicated as ( ).

finish for each file.

4.4.9 Complete DNA sequence of the 1.65 kbp genomic insert of pGM602

The complete double stranded sequence of the 1.65 kbp ClaI/SalI insert of pGM602 is shown in Figure 4.11. Inclusive of the two recognition sequences, SalI (position 3) and ClaI (position 1687), the total length is 1690 bp. Fully 97% of the sequence data was directly confirmed on its opposite strand. Only 50 bp at the SalI end was not sequenced on both strands. Each restriction site used either in sub-cloning or sequence determination was overlapped in reassembling the data (Figure 4.10).

The pattern of restriction sites for ClaI, SalI, BamHI, EcoRV, NruI are as predicted from the original restriction mapping data (Section 4.2).

4.5 Identification of the aroC structural gene

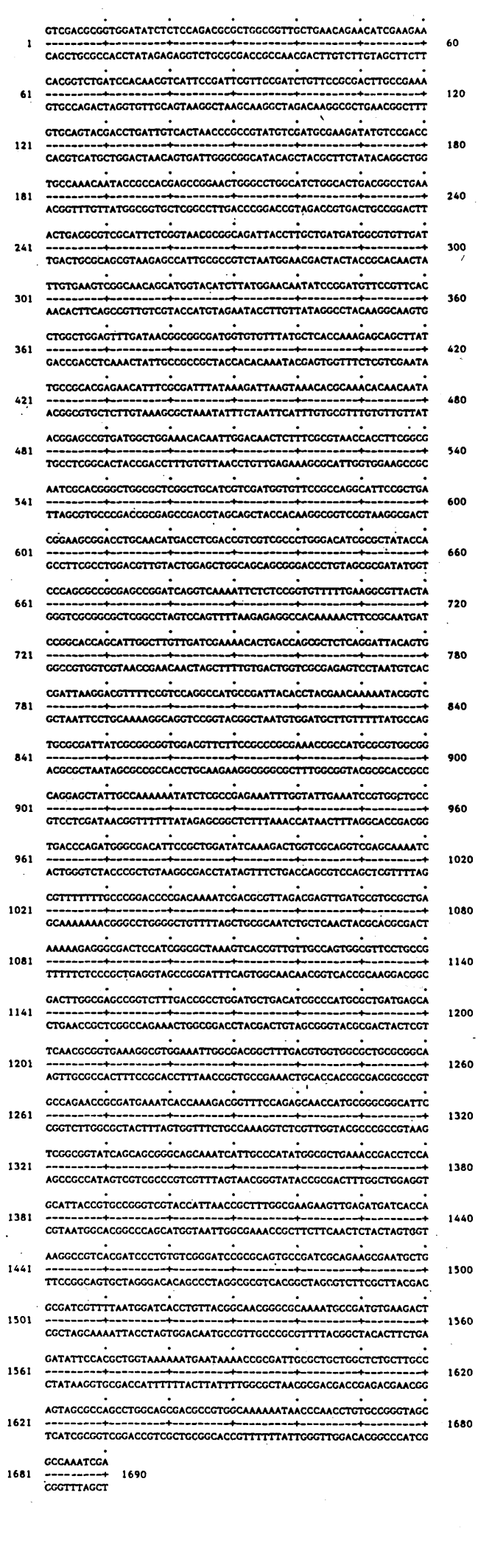
4.5.1 TRN TRP analysis of the DNA sequence data

Both strands of the 1.690 kbp ClaI/SalI DNA sequence were examined for potential open-reading frames (ORF) using the computer program TRN TRP (Section 2.18.2).

The 5'(SalI) $\longrightarrow$(ClaI)3' mRNA-like strand displayed a single large ORF extending from positions 462 to 1562 (Figure 4.12). When translated from the first initiation codon (ATG position 492), this was capable of encoding a 357 amino acid residue polypeptide of M_r 38,183 (Figure 4.14). Immediately preceding the putative start codon is the sequence GGAG (position 482) complementary to the 3'-end of E.coli

Figure 4.11 (facing)

The complete double-strand DNA sequence of the 1690 bp SalI (position 0)/ClaI (position 1690) insert of E.coli genomic DNA in pGM602. The sequencing strategy has been shown previously in Figure 4.10.



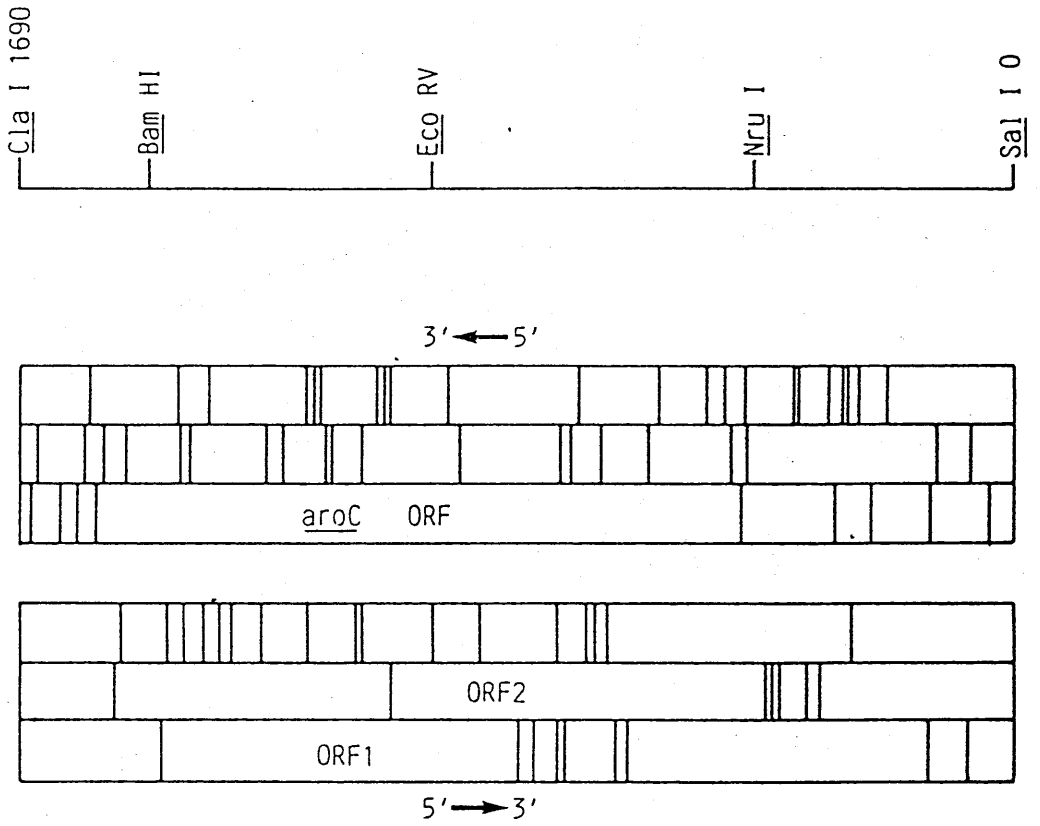


Figure 4.12: Protein coding regions within the sequenced ClaI/SalI insert of pGM602. Stop codons in all six reading frames are indicated and the positions of ORF's shown.

165 ribosomal RNA (Shine & Dalgarno, 1975).

The 5'(ClaI) → (SalI)3' strand exhibited a number of short ORF's the largest of which was 650 nucleotides long (Figure 4.12, ORFI). None of these were capable of encoding a polypeptide of molecular weight > 18-20 kDa. The pattern of plasmid encoded proteins in the in vitro expression of pGM602 (Section 4.3.3, Figure 4.8) suggests that the small ORF's (in particularly ORFI) observed in the opposite strand are not expressed. Bands at the predicted M_r 17.2 - 20,000 for these ORF's are not observed in pGM602-directed expression (Figure 4.8).

The position and orientation of this putative aroC coding sequence is entirely consistent with the data presented in Sections 4.2 and 4.3. In particular the deletion analysis of the NruI-SalI region (Section 4.2.5, Figure 4.5) which did not affect aroC complementation is rationalised in terms of removal of 5' upstream sequence. Moreover the use of the NruI site to clone aroC in the tac-aroC construct pGM605 (Section 4.3.2) means that the aroC coding region is brought very close to the vector tac promoter. This explains the very efficient and high levels of overexpression observed for pGM605 when fully induced by IPTG (Figure 4.7, Section 4.3.2).

4.5.2 TESTCODE analysis of the putative aroC gene

The TESTCODE program (described in Sections 2.18.3 and 3.6.8) was used to examine the statistical order in the nucleotide sequence suspected of being the aroC coding region.

The graphical output displaying the measure of the 'period three constraint' (Fickett, 1982) is shown in Figure 4.13(a) & (b).

Throughout the putative aroC gene (Figure 4.13(a)) the running value of the TESTCODE is consistently in panel A (> 95% probability of coding). The only 'trough' occurs at position 950 but overall the values obtained very strongly suggests that this is a genuine protein coding region.

A TESTCODE analysis of the opposite (complementary) strand sequence is shown in Figure 4.13(b). Surprisingly sequences corresponding to ORFI and ORF2 (Figure 4.12) both display high TESTCODE values. In particular ORFI appears just as likely to be protein coding as aroC ORF (Figure 4.12), as predicted by the TESTCODE program. It is possible that this antiparallel arrangement is fortuitous. Since there is an excess of RNY codons in a given gene (Shepherd, 1981), the same must hold for the same register in the complementary strand. Thus there must be an overall avoidance of pre-termination codons on both the sense and nonsense strand. If this hypothesis is correct then there should be a higher incidence of randomly distributed longer fortuitous ORFs on complementary strands than coding strands. Casino et al. (1981) have shown that the coding capacity for complementary strands exhibits such a preference for ORFs greater than 100 codons in length.

A cursory examination of the ORFI sequence indicates an ATG codon at position 284 on the opposite strand numerology

(the ORFI extends from positions 240 to 840). Immediately preceding this potential initiation codon is the hexamer AAGCGG, 6 bp upstream of the ATG, which could act as a potential ribosome binding site.

As yet no firm evidence other than this circumstantial TESTCODE prediction exists for expression of ORFI. Indeed evidence to the contrary suggests that plasmid pGM602 only expresses a 38-39 kDa protein from its genomic insert (Section 4.3.3, Figure 4.8). This presumably is the cloned *aroC* gene product chorismate synthase. Whether ORF1 is simply a fortuitous ORF or a real gene could be distinguished upon the detection (or not) of a functional mRNA. To examine this possibility a synthetic oligonucleotide (20-mer, sequence 5'ATG AAA TCA CCA AAG ACG GT3') complementary to residues 39-44 of the putative ORFI coding sequence was prepared. This was annealled to 10 µg of RNA prepared from *E.coli* AB2849/pGM602 as described in Section 2.19, and used as a primer for a reverse-transcriptase directed primer extension experiment. No reverse run-off products ($[\alpha\text{-}^{35}S]$ dATPαs incorporation into DNA) were detected on subsequent 6% polyacrylamide sequencing gels (Section 2.19.3).

It would appear therefore that ORFI is not expressed. However the possibility that the cloning procedure has disrupted the normal mode of ORFI expression cannot be discounted. There is no precedent for such structural anti-parallel overlapping genes in *E.coli*. Rak *et al.* (1982) have demonstrated expression of two proteins from overlapping and oppositely oriented genes on the transposable DNA insertion

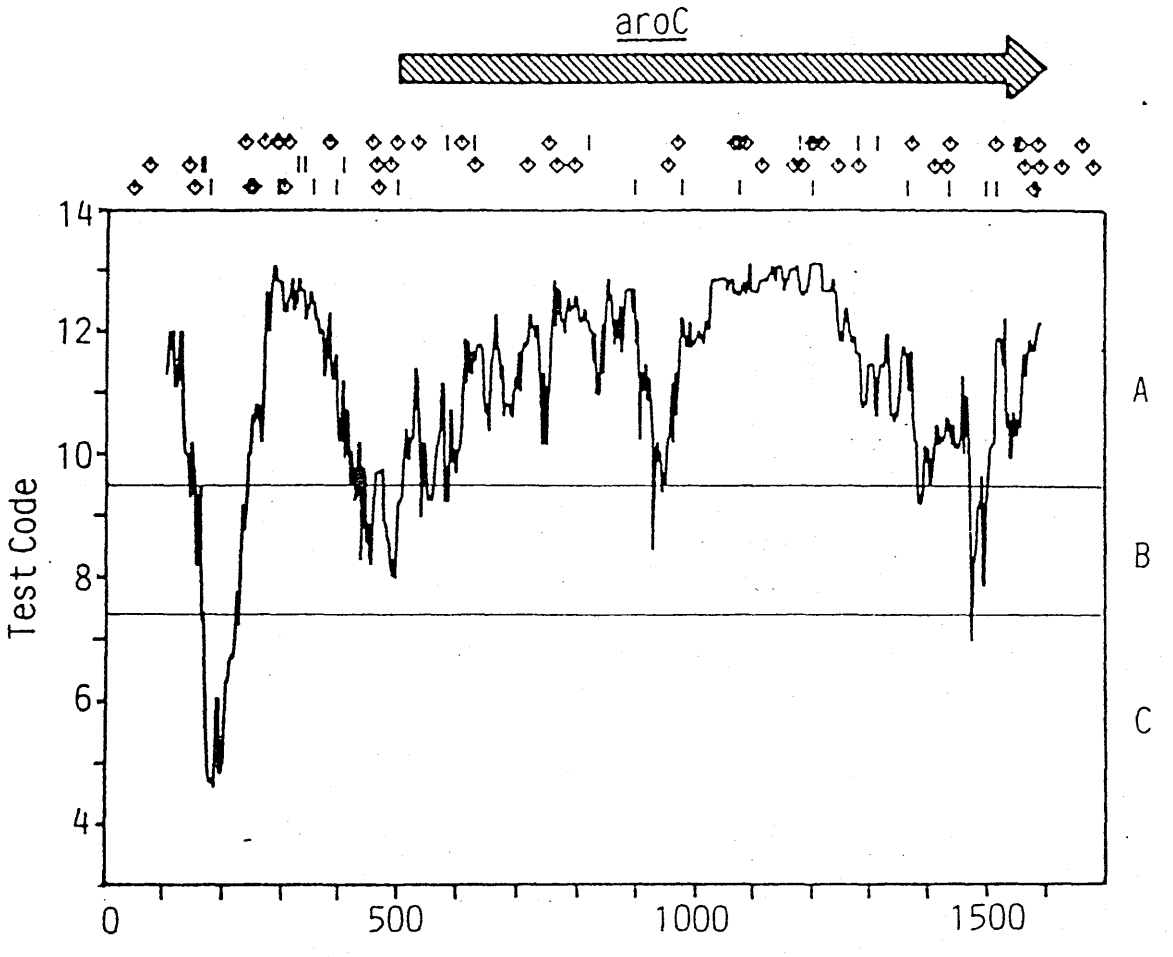


Figure 4.13(a): TESTCODE analysis of the aroC sense strand. All symbols and values are as previously described in Figure 3.15.

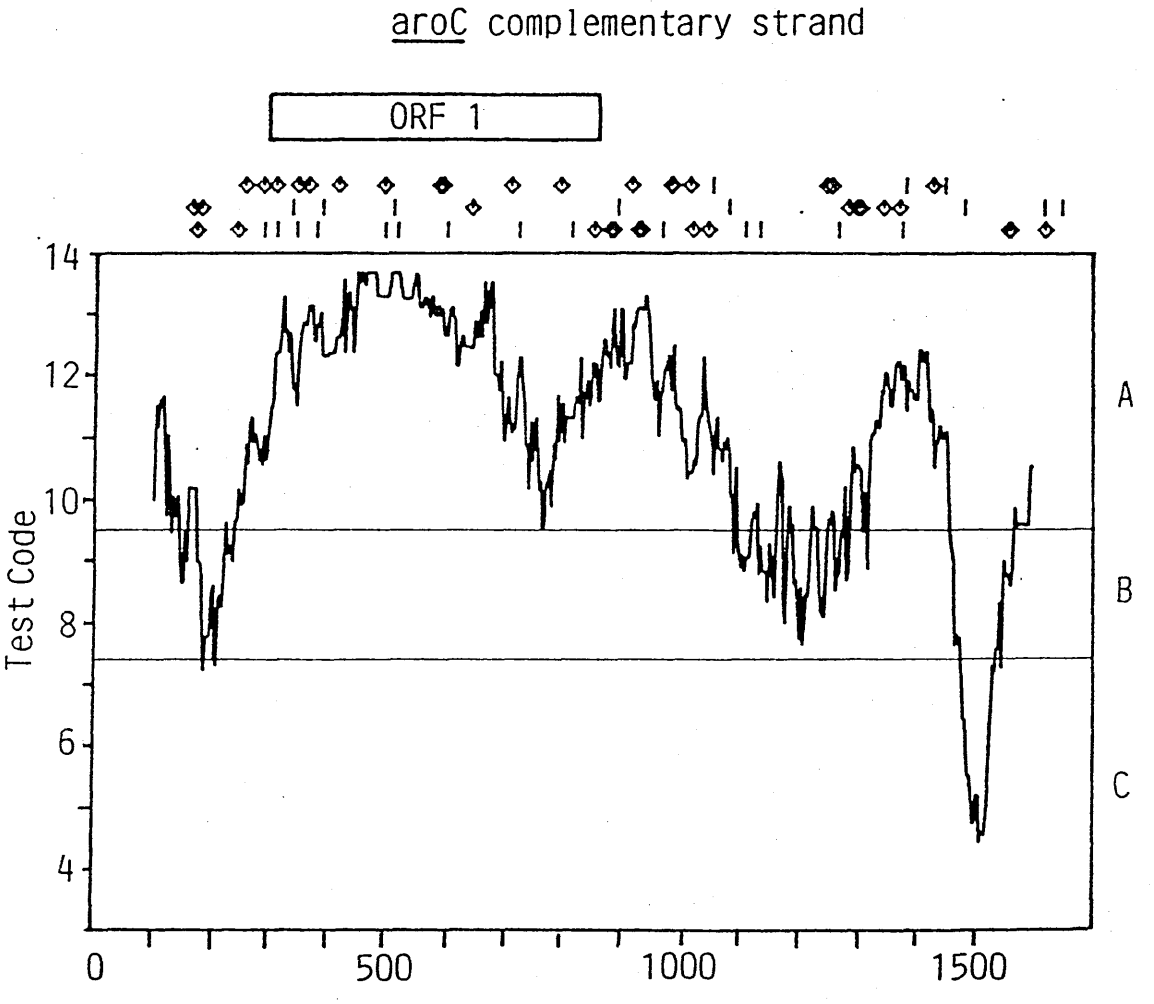


Figure 4.13(b): TESTCODE analysis of the aroC complementary (non-sense) strand. The position of ORF 1 is shown and is discussed in the text. All symbols and values are as specified in Figure 3.15.

sequence IS5. Furthermore the promoter for the smaller (located totally within the larger) IS5 ORF was only active in cell-free systems and not *in vivo*. Rak *et al*. (1982) have proposed that the smaller ORF may require its own or some other (regulatory) gene product in order to function. More detailed experimental analysis of ORFI, possible homologies with other proteins or forced expression as a fusion polypeptide must await.

4.5.3 Putative amino acid sequence of the *aroC* gene

The predicted amino acid sequence of the *E.coli* chorismate synthase, as deduced from the nucleotide sequence of its structural gene *aroC*, is shown in Figure 4.14. The predicted sub-unit M_r of 38,183 for the 357 amino acid residue protein is in agreement with the values observed in crude extract and *in vitro* expression analysis (Sections 4.3.1 and 4.3.3).

4.5.4 Codon utilisation of the *aroC* gene

(i) Background

The pattern of codon usage in *E.coli* is not random, being modulated by at least two distinct parameters (Gouy & Gautier, 1982; Grosjean & Fiers, 1982). In strongly expressed genes codons corresponding to minor tRNAs species are clearly avoided. Efficient translation is governed by the proper choice of degenerate codons promoting a codon:anticodon interaction of intermediate energy. Conversely codon usage in very weakly expressed genes or those required in low molar amounts (e.g. repressor genes) follows exactly the opposite rules. The pattern of codon utilisation in the *dnaG*

Figure 4.14 (facing)

The chorismate synthase (*aroC*) coding region. Note the position of the proposed ribosome binding site (RBS). The protein coding sequence was obtained by translating the primary DNA sequence data using the program TRN TRP.

```
   1  GTCGACGCGGTGGATATCTCTCCAGACGGGCTGGCGGTTGCTGAACAGAACATCGAAGA
  60  ACACGGTCTGATCCACAACGTCATTCCGATTCGTTCCGATCTGTTCCGCGACTTGCCGAA
 120  AGTGCAGTACGACCTGATTGTCACTAACCCGCCGTATGTCGATGCGAAGATATGTCCGAC
 180  CTGCCAAACAATACCGCCACGAGCCGGAACTGGGCCTGGCATCTGGCACTGACGGCCTGA
 240  AACTGACGGCGTCGCATTCTCGGTAACGCGGCAGATTACCTTGCTGATGATGGCGTGTTGA
 300  TTTGTGAAGTCGGCAACAGCATGGTACATCTTATGGAACAATATCCGGATGTTCCGTTCA
 360  CCTGGCTGGAGTTTGATAACGGCGGCGATGGTGTGTTTATGCTCACCAAAGAGCAGCTTA
 420  TTGCCGCACGAGAACATTTCGCGATTTATAAAGATTAAGTAAACACGCAAACACAACAAT
 480  AACGGAGCCGTGATGGCTGGAAACACAATTGGACAACTCTTTCGCGTAACCACCTTCGGC
          RBS      MetAlaGlyAsnThrIleGlyGlnLeuPheArgValThrThrPheGly   [16]
 540  GAATCGCACGGGCTGGCGCTCGGCTGCATCGTCGATGGTGTTCCGCCAGGCATTCCGCTG
      GluSerHisGlyLeuAlaLeuGlyCysIleValAspGlyValProProGlyIleProLeu   [36]
 600  ACGGAAGCGGACCTGCAACATGACCTCGACCGTCGTCGCCCTGGGACATCGCGCTATACC
      ThrGluAlaAspLeuGlnHisAspLeuAspArgArgArgProGlyThrSerArgTyrThr   [56]
 660  ACCCAGCGCCGCGAGCCGGATCAGGTCAAAATTCTCTCCGGTGTTTTTGAAGGCGTTACT
      ThrGlnArgArgGluProAspGlnValLysIleLeuSerGlyValPheGluGlyValThr   [76]
 720  ACCGGCACCAGCATTGGCTTGTTGATCGAAAACACTGACCAGCGCTCTCAGGATTACAGT
      ThrGlyThrSerIleGlyLeuLeuIleGluAsnThrAspGlnArgSerGlnAspTyrSer   [96]
 780  GCGATTAAGGACGTTTTCCGTCCAGGCCATGCCGATTACACCTACGAACAAAAATACGGT
      AlaIleLysAspValPheArgProGlyHisAlaAspTyrThrTyrGluGlnLysTyrGly   [116]
 840  CTGCGCGATTATCGCGGCGGTGGACGTTCTTCCGCCCGCGAAACCGCCATGCGCGTGGCG
      LeuArgAspTyrArgGlyGlyGlyArgSerSerAlaArgGluThrAlaMetArgValAla   [136]
 900  GCAGGAGCTATTGCCAAAAAATATCTCGCCGAGAAATTTGGTATTGAAATCCGTGGCTGC
      AlaGlyAlaIleAlaLysLysTyrLeuAlaGluLysPheGlyIleGluIleArgGlyCys   [156]
 960  CTGACCCAGATGGGCGACATTCCGCTGGATATCAAAGACTGGTCGCAGGTCGAGCAAAAT
      LeuThrGlnMetGlyAspIleProLeuAspIleLysAspTrpSerGlnValGluGlnAsn   [176]
1020  CCGTTTTTTTGCCCGGACCCCGACAAAATCGACGCGTTAGACGAGTTGATGCGTGCGCTG
      ProPhePheCysProAspProAspLysIleAspAlaLeuAspGluLeuMetArgAlaLeu   [196]
1080  AAAAAAGAGGGCGACTCCATCGGCGCTAAAGTCACCGTTGTTGCCAGTGGCGTTCCTGCC
      LysLysGluGlyAspSerIleGlyAlaLysValThrValValAlaSerGlyValProAla   [216]
1140  GGACTTGGCGAGCCGGTCTTTGACCGCCTGGATGCTGACATCGCCCATGCGCTGATGAGC
      GlyLeuGlyGluProValPheAspArgLeuAspAlaAspIleAlaHisAlaLeuMetSer   [236]
1200  ATCAACGCGGTGAAAGGCGTGGAAATTGGCGACGGCTTTGACGTGGTGGCGCTGCGCGGC
      IleAsnAlaValLysGlyValGluIleGlyAspGlyPheAspValValAlaLeuArgGly   [256]
1260  AGCCAGAACCGCGATGAAATCACCAAAGACGGTTTCCAGAGCAACCATGCGGGCGGCATT
      SerGlnAsnArgAspGluIleThrLysAspGlyPheGlnSerAsnHisAlaGlyGlyIle   [276]
1320  CTCGGCGGTATCAGCAGCGGGCAGCAAATCATTGCCCATATGGCGCTGAAACCGACCTCC
      LeuGlyGlyIleSerSerGlyGlnGlnIleIleAlaHisMetAlaLeuLysProThrSer   [296]
1380  AGCATTACCGTGCCGGGTCGTACCATTAACCGCTTTGGCGAAGAAGTTGAGATGATCACC
      SerIleThrValProGlyArgThrIleAsnArgPheGlyGluGluValGluMetIleThr   [316]
1440  AAAGGCCGTCACGATCCCTGTGTCGGGATCCGCGCAGTGCCGATCGCAGAAGCGAATGCT
      LysGlyArgHisAspProCysValGlyIleArgAlaValProIleAlaGluAlaAsnAla   [336]
1500  GGCGATCGTTTTAATGGATCACCTGTTACGGCAACGGGCGCAAAATGCCGATGTGAAGAC
      GlyAspArgPheAsnGlySerProValThrAlaThrGlyAlaLysCysArgCysGluAsp   [356]
1560  TGATATTCCACGCTGGTAAAAAATGAATAAAACCGCGATTGCGCTGCTGGCTCTGCTTGC
1620  CAGTAGCGCCAGCCTGGCAGCGACGCCGTGGCAAAAAATAACCCAACCTGTGCCGGGTAG
1680  CGCCAAATCGA  1690
```

gene and other regulatory genes of E.coli has been examined (Konigsberg & Godson, 1983). The dnaG gene contains an unusually large number of rare codons. This use of rare codons may form part of a mechanism to maintain low levels of expression of the dnaG protein. This hypothesis is supported by the observation that the E.coli repressor genes lacI, trpR and avaC also contain unusually high numbers of rare codons (Konigsberg & Godson, 1983).

The role of the tRNA population in regulating the rate of the translation process has been investigated (Ikemura, 1981; Grosjean & Fiers, 1982). The selection between degenerate codons recognised by iso-accepting tRNAs can be rationalised in terms of the frequency of these modulating codons. Codons AUA (Ile); CGG/AGA/AGG (Arg); CUA (Leu); and GGA/GGG (Gly) are strong candidates for this role representing very minor tRNAs species in E.coli.

(ii) The pattern of codon utilisation of the E.coli aroC gene is shown in Table 4.2. A comparison between the codon usages for E.coli aroA, aroB, aroD, aroE and aroL genes (Duncan et al., 1984b; Millar & Coggins, 1986, this study; Duncan et al., 1986b; Anton & Coggins, 1986; Millar et al., 1986b, this study, Defeyter & Pittard, 1986) is shown in Table 4.3.

A quantitative comparison of the frequency of modulating codons of the 6 aro genes is shown in Table 4.4. Grosjean & Fiers (1982) have examined the total number of each of the possible 64 codons used in 25 sequences corresponding to abundant E.coli proteins (S), and 18 sequences corresponding to weakly expressed E.coli proteins (W). The frequency of

	U	aroC	C	aroC	A	aroC	G	aroC	
U	Phe	9	Ser	2	Tyr	3	Cys	2	U
	Phe	3	Ser	4	Tyr	4	Cys	4	C
	Leu	1	Ser	1	ochre	0	opal	1	A
	Leu	3	Ser	3	amber	0	Trp	1	G
C	Leu	1	Pro	3	His	5	Arg	9	U
	Leu	6	Pro	2	His	2	Arg	15	C
	Leu	0	Pro	2	Gln	5	Arg	1	A
	Leu	11	Pro	10	Gln	9	Arg	0	G
A	Ile	13	Thr	2	Asn	3	Ser	2	U
	Ile	14	Thr	15	Asn	6	Ser	7	C
	Ile	0	Thr	2	Lys	15	Arg	0	A
	Met	7	Thr	3	Lys	1	Arg	0	G
G	Val	9	Ala	5	Asp	10	Gly	8	U
	Val	6	Ala	9	Asp	18	Gly	25	C
	Val	1	Ala	5	Gln	13	Gly	6	A
	Val	7	Ala	12	Gln	7	Gly	4	G

Table 4.2: aroC codon utilisation.

modulating codons in strongly expressed genes (S) was 0.02 (31/1516) and 0.17 (271/1612) for weakly (W) expressed genes.

Clearly the aroC gene falls into the weakly expressed category with a frequency of 0.09. It is interesting that of the six aro genes examined (Table 4.4) only the aroD gene (Duncan et al., 1986b) can be classed as strongly expressed. This however is based on the presence of only 1 modulating codon in 59 possible synonymous codons, and as such its validity is perhaps statistically bankrupt. The aroL gene encoding shikimate kinase II displays the highest incidence of rare codon frequency (0.24, Table 4.4). The possible significance of this is discussed further in Chapter Five.

This semi-quantitative assessment, based solely on incidence of rare codons within the aro gene's coding regions, is consistent with evidence regarding the constitutive expression of the aro genes (Tribe et al., 1976) and the observed wild-type levels of the E.coli shikimate pathway enzymes (Berlyn & Giles, 1967).

4.5.5 N-terminal amino acid sequence of purified chorismate synthase

E.coli chorismate synthase has been purified from an overproducing strain E.coli AB2849/pGM602. The sub-unit molecular weight, as determined by SDS PAGE, is 38-39,000 (White et al., 1986). The first thirty N-terminal amino acid residues have been determined by direct sequencing on a liquid phase amino acid sequencer (P.J. White, unpublished data). The results are shown in Figure 4.15.

		aro							aro							aro							aro						
	U	A	B	C	D	E	L	C	A	B	C	D	E	L	A	A	B	C	D	E	L	G	A	B	C	D	E	L	
	Phe	9	6	9	5	10	4	Ser	6	2	2	3	2	1	Tyr	9	5	3	5	5	2	Cys	2	5	2	0	2	1	U
U	Phe	9	4	3	3	4	1	Ser	5	3	4	5	4	1	Tyr	4	4	4	1	0	0	Cys	4	0	4	3	1	0	C
	Leu	10	10	1	1	4	2	Ser	2	3	1	0	2	2	Ochre	-	-	-	-	-	-	Opal	1	1	1	1	1	1	A
	Leu	5	6	3	1	8	2	Ser	2	5	3	2	2	1	Amber	-	-	-	-	-	-	Trp	2	2	1	2	2	1	G
	Leu	4	8	1	1	4	3	Pro	4	3	3	2	4	2	His	3	5	5	4	5	1	Arg	12	9	9	7	3	2	U
	Leu	3	8	6	5	5	1	Pro	3	2	2	0	2	1	His	5	2	2	2	3	1	Arg	7	8	15	2	5	4	C
C	Leu	0	0	0	0	1	1	Pro	2	2	2	2	4	3	Gln	5	4	5	2	3	4	Arg	0	0	1	0	1	1	A
	Leu	26	17	11	11	13	8	Pro	9	9	10	3	3	2	Gln	7	8	9	4	7	5	Arg	2	2	0	0	3	3	G
	Ile	17	12	13	9	9	5	Thr	7	3	2	3	4	1	Asn	9	5	3	2	9	4	Ser	1	4	2	2	5	0	U
	Ile	9	3	14	7	9	7	Thr	10	8	15	8	3	3	Asn	9	6	6	2	2	4	Ser	5	3	7	4	2	3	C
A	Ile	0	1	0	0	3	0	Thr	7	0	2	0	3	4	Lys	14	8	15	16	6	2	Arg	0	0	0	0	1	1	A
	Met	14	10	7	11	5	1	Thr	10	7	3	4	2	6	Lys	3	5	1	1	2	0	Arg	1	1	0	1	0	1	G
	Val	7	8	9	2	4	5	Ala	6	8	5	3	11	2	Asp	23	9	10	11	8	5	Gly	13	11	8	5	13	1	U
	Val	4	15	6	6	0	5	Ala	7	6	9	13	3	4	Asp	3	6	18	6	5	1	Gly	18	14	25	8	7	4	C
G	Val	4	3	1	4	4	2	Ala	15	5	5	6	6	3	Glu	16	16	13	11	11	5	Gly	2	2	6	0	4	3	A
	Val	9	6	7	5	5	3	Ala	18	20	12	7	10	9	Glu	6	8	7	7	6	1	Gly	4	7	4	0	1	2	G

Table 4.3: Codon utilisation of the aroA-E and aroL genes. Infrequently used (modulating) codons (see Table 4.4) are boxed.

Modulating Codons	aroC	aroB	aroL	aroD	aroE	aroA	s	w
CUA (Leu)	0	0	1	0	1	0	3	22
AUA (Ile)	0	1	0	0	3	0	2	27
CGA/G (Arg)	1	2	4	0	4	2	4	69
AGA/G (Arg)	0	1	2	1	1	1	4	45
GGA/G (Gly)	10	9	5	0	5	6	18	10
Total	11	13	12	1	14	9	31	271
Total possible	117	119	51	59	81	143	1516	1612
Frequency	0.09	0.11	0.24	0.017	0.17	0.06	0.02	0.17

Notes: 1. 's' and 'w' refer to the strongly and weakly expressed genes described by Grosjean & Fiers (1982).

Table 4.4: Frequency of modulating codons in the E.coli aro genes.

A comparison with the amino acid sequence predicted from the nucleotide sequence of the aroC gene (Figure 4.14; this study; White et al., 1986) indicates a direct match. This definitively identifies the aroC gene coding region by locating the translational initiation site.

4.6 Transcriptional regulatory features of the E.coli aroC gene.

4.6.1 aroC promoter

The DNA sequence upstream of the aroC gene translational initiation codon (position 492, Figure 4.14) was examined for potential transcriptional regulator features. A computer assisted search for the conserved hexanucleotide sequences corresponding to the '-35 region' TTGACa and '-10 region' (Pribnow box) TAtAaT failed to detect any exact matches. This finding is not surprising since the consensus sequences will, by definition, be rarely found as a perfectly conserved order. Two candidates each with limited homology to the conserved hexamers and each showing optimal 16-17 bp separation between the hexanucleotide sequences (Mulligan et al., 1986) were identified by eye search. The first exhibited the sequence ^{420}TTGCCG-17bp-ATTTAT^{448} preceding the triplet GAT at position 453. The second was comprised of ^{372}TTGATA-16bp-GTTTAT^{399} preceding the triplet CAC at position 404 (all numbering as in Figure 4.14).

RNA was prepared (Section 2.19.1) from E.coli AB2849/pGM602. A 20-mer oligonucleotide complementary to the sequence position 577-596 in Figure 4.14 (codons 29-35 of the aroC gene; sequence 5'-GCC AAT GCC TGG CGG AAC AC-3') was used in a

fMet - Ala - Gly - Asn - Thr - Ile -
Gly - Gln - Leu - Phe - Arg - Val -
Thr - Thr - Phe - Gly - Gln - Ser -
His - Gly - Leu - Ala - Leu - Gly -
Cys - Ile - Val - Asp - Gly - Val -
Pro - Pro - Gly - Ile - Pro - Leu

Figure 4.15: The N-terminal amino acid sequence of *E.coli* chorismate synthase.

primer extension synthesis experiment (Section 2.19.3) to attempt to identify the 5' end of the aroC mRNA. Despite repeated attempts a definitive result, as seen previously for analogous experiments with aroB and aroL (Figures 3.23 & 5.13), could not be obtained. It seems likely though that the aroC mRNA is originating between 50-150 bp upstream of the translational start site (data not shown). This area includes both potential promoter sequences identified by eye and detailed above.

4.6.2 Potential aroC terminator sequences

Immediately downstream of the aroC termination codon (position 163, Figure 4.14) occurs a sequence capable of forming the stem-loop structure shown in Figure 4.16. The free energy of formation, calculated by the method of Tinoco et al. (1977) suggests that on thermodynamic grounds this structure could exist in vivo. When considered with its position immediately adjacent to the 3' end of the gene, it seems likely that this structure forms part of a rho-independent terminator (Rosenberg & Court, 1979). However there is a noticeable absence of the conserved pyrimidine-rich sequence which normally follows such a stem-loop structure. The inverted repeat structure shown in Figure 4.16 is indicated by overlining in Figure 4.14.

4.7 Consideration of the aroC sequence

4.7.1 Homology with other shikimate pathway enzymes

In 1945 Horowitz proposed that one possible mode of evolution of biosynthetic pathways could entail the progressive

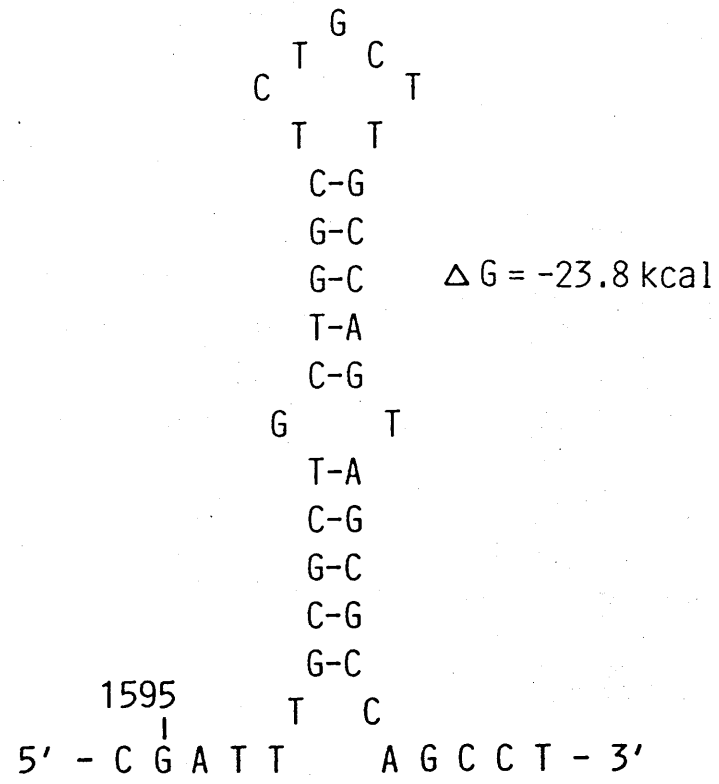
Figure 4.16(facing)

(Upper) An abbreviated *aroC* coding region showing the first 16 and last 20 amino acid residues. The inverted repeat sequence 3' to the *aroC* gene is overlined.

(Lower) A possible stem-loop structure formed by the inverted repeat sequence described above. This may constitute an *aroC* terminator (rho-independent, Rosenberg & Court, 1979). The free energy of formation was calculated by the rules of Tinoco *et al.* (1973).

```
481   AACGGAGCCGTGATGGCTGGAAACACAATTGGACAACTCTTTCGCGTAACCACCTTCGGC
                  MetAlaGlyAsnThrIleGlyGlnLeuPheArgValThrThrPheGly   [16]
              --------  aroC coding sequence  --------

1501  GGCGATCGTTTTAATGGATCACCTGTTACGGCAACGGGCGCAAAATGCCGATGTGAAGAC
      GlyAspArgPheAsnGlySerProValThrAlaThrGlyAlaLysCysArgCysGluAsp   [356]
                .         .         .         .____ ____      __
1561  TGATATTCCACGCTGGTAAAAAATGAATAAAACCGCGATTGCGCTGCTGGCTCTGCTTGC
      ___ _____.         .         .         .         .         .
1621  CAGTAGCGCCAGCCTGGCAGCGACGCCGTGGCAAAAAATAACCCAACCTGTGCCGGGTAG
                .
1681  CGCCAAATCGA  1691
```



building back of activities from the final metabolite of the pathway - retroevolution (Horowitz, 1945). This could involve gene duplication followed by subsequent mutation or, as already known for the chorismate mutase/prephenate dehydrogenase activity of the aromatic biosynthetic pathway, modification of a distinct effector site (Llewellyn *et al.*, 1980). It is clearly plausible that a biosynthetic pathway catalysing successive reactions involving alterations to a common structure (shikimic acid or chorismate perhaps) could have evolved in such a fashion.

For such a theory to be applicable to the shikimate pathway of *E.coli* at least two related phenomena would be expected. Firstly a progressively decreasing similarity in the three-dimensional structures of the pathway intermediates (chorismate —— DAHP) but a conserved interaction at their respective active sites. Secondly a measure of conserved amino acid homology within the 'core' structure of the pathway enzymes. This 'core' peptide sequence(s) would in effect recognise a common structural motif of the pathway intermediates and represent the pathway blueprint.

Without a detailed knowledge of the steric considerations at the active site of each of the pathway enzymes the first phenomenon is essentially for academic consumption only. The availability of the amino acid sequence for six shikimate pathway enzymes on the other hand makes the latter proposal directly testable.

The degree of homology between each of the six *aro* gene products was tested in a pair-wise fashion using the local homology search program BESTFIT (Section 2.18.3). Belfaiza *et al.* (1986) have recently shown that a similar approach can be used successfully to delineate modes of evolution. A comparison between the β-cystathionase (*metC* gene product) and the previous biosynthetic activity, cystathionine-γ-synthase (*metB*), reveals a degree of homology sufficiently strong to suggest that the structure of the MetB and MetC proteins evolved from a common ancestral gene (Belfaiza *et al.*, 1986).

The different lengths of the six *aro* enzymes presents a problem in trying to distinguish real cases of authentic relationships (possibly resulting from gene duplication and subsequent, limited divergence) and apparent homologies arising through spurious similarities.

The BESTFIT output for each of the paired shikimate pathway enzymes failed to detect any significant levels of conserved internal homologies. Taking the limits of each output as the first and last identities located, then the degree of identity rarely exceeded 10-15%. A consideration of conservative changes only slightly elevated this figure. On the contrary if the number of identities was calculated over the whole protein length then the figure dropped below 10%. At this level similarities are almost certainly fortuitous. A comparison between these levels of identity and those observed between multifunctional and monofunctional

arom activities (Chapter 3, Figure 3.28; Chapter 5, Figure 5.24; Chapter 6), clearly indicate that the shikimate pathway has not evolved by retro-evolution of a biosynthetic pathway. The hypothesis that the degree of relatedness between the prokaryotic and eukaryotic arom activities suggests a gene fusion/scission model is considered in Chapter 6.

CHAPTER 5

THE CLONING AND EXPRESSION OF THE AROL GENE FROM *ESCHERICHIA COLI* K12

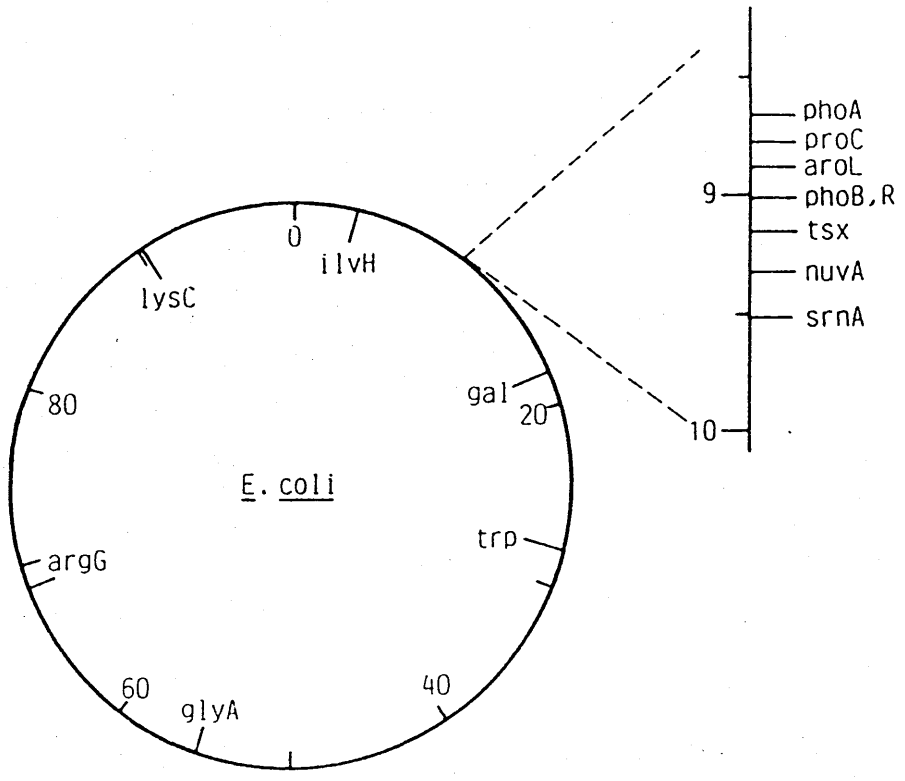
5.1 Background

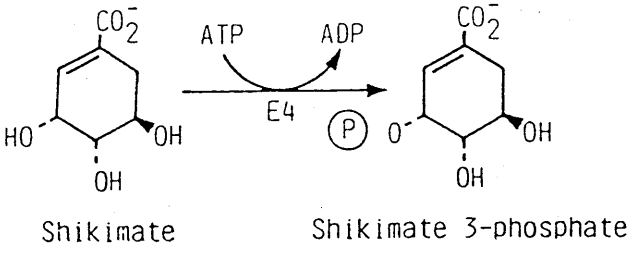
5.1.1 The existence of two shikimate kinase isoenzymes

Berlyn & Giles (1969) first demonstrated that E.coli contained two enzymatic activities capable of converting shikimate to shikimate 3-phosphate. Mutants auxotrophic for this particular step of the common pathway have never been isolated in E.coli (or S.typhimurium). This strongly suggests expression of separate genes from two independent chromosomal loci. Subsequently the two shikimate kinase activities from E.coli were resolved following ion-exchange chromatography, and shown to have similar molecular weights (of ca. 20,000) by gel filtration (Ely & Pittard, 1979; Section 1.3). The structural gene for the shikimate kinase II isoenzyme (the aroL gene) has been mapped to minute 9 of the E.coli chromosome (Figure 5.1), and has been shown to be 99% co-transducible with the adjacent proC gene (Ely & Pittard, 1979). The proC gene encodes the third enzyme in the proline biosynthetic pathway which catalyses the reduction of pyrroline-5-carboxylate to proline (Vogel & Davis, 1952).

5.1.2 The tyrR regulon

The gene tyrR is located at about minute 29 on the E.coli chromosome (Bachmann, 1983) and codes for a regulatory protein involved in transcriptional control of aromatic amino acid biosynthesis (Wallace & Pittard, 1969). The amino acids tyrosine, phenylalanine and tryptophan act in concert with the tyrR aporepressor (Table 5.1) to control expression of a number of genes coding for biosynthetic and





Shikimate Shikimate 3-phosphate

Figure 5.1: The chromosomal location of the E. coli aroL gene encoding shikimate kinase II (E4).

transport functions associated with aromatic amino acids (Brown & Somerville, 1971; Camakaris & Pittard, 1973; Camakaris & Pittard, 1982; Im, Davidson & Pittard, 1971; Whipp & Pittard, 1977). The tyrR gene encoding this auto-regulatory protein has recently been cloned into a multicopy plasmid (Cornish et al., 1982) and has been shown to code for a 63,000 dalton polypeptide. The aroL encoded shikimate kinase II is controlled by the tyrR repressor complexed with either tyrosine or tryptophan while the shikimate kinase I isoenzyme exhibits no such sensitivity to transcriptional regulation (Ely & Pittard, 1979; Figure 5.2).

5.2 Cloning strategy for the E.coli aroL gene

5.2.1 Selection by proC complementation (pMH423)

The existence of two shikimate kinase isoenzymes, and the associated lack of aromatic auxotrophs mutant for this step of the common pathway, invalidated the application of the approach used previously in this laboratory to clone the other shikimate pathway genes (Duncan & Coggins, 1983; Anton & Coggins, 1983; Millar et al., 1986a). Mutation in either of the shikimate kinase genes (only aroL has been mapped) would still produce an aromatic prototroph, therefore direct selection by complementation was not an available option.

The observed close linkage between aroL and proC (99% co-transducible, Ely & Pittard, 1979) suggested that selection for proC complementation could be used to indirectly clone the aroL gene. Plasmid pMH423, a 10 kbp EcoRI E.coli genomic

Table 5.1: The tyr regulon: genes controlled by tyrR

Operon	Gene product	Effector	Active repressor (inducer)
aroF tyrA	DAHP synthase (Tyr) Chorismate mutase prephenate dehydrogenase	Tyr (Phe at high concentration)	R-Tyr, R-Phe
aroG	DAHP synthase (Phe)	Phe + Trp	R_2-Phe-Trp, or R-Phe + R-Trp
aroL	Shikimate kinase II	Tyr/Trp	R-Tyr/R-Trp
tyrB	Tyrosine aminotransferase	Tyr	R-Tyr
aroP	Component of common aromatic transport	Tyr/Phe/Trp	R-Tyr/R-Phe/R-Tr
tyrP	Component of tyrosine-specific transport	Tyr Phe	R-Tyr R-Phe (inducer)
mtr	Component of tryptophan-specific transport	Trp Phe	R-Trp R-Phe (inducer)
tyrR	aporepressor	none, one, or more of Tyr, Phe, Trp	R and one or more or R-Tyr, R-Phe, R-Trp.

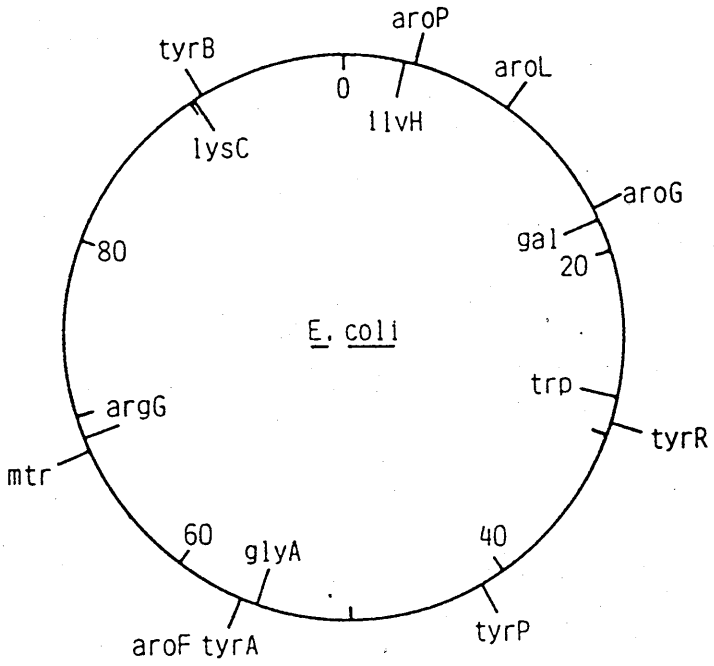


Figure 5.2: Chromosomal locations of the genes of the tyr regulon in E. coli.

On the outside of the chromosomal map are indicated those genes which are under the control of the pleitropic regulator tyrR.

On the inside of the circle are marker genes for reference.

fragment cloned in pAT153 (Figure 5.3), was a generous gift of Dr M.G. Hunter. This plasmid carried the intact proC gene by complementation criterion and was the starting material for the work described in this chapter.

5.2.2 Sub-cloning of the aroL gene, loss of ProC selection

Plasmid pMH423 was digested individually with a number of restriction enzymes and sites for PstI (2), BamHI (2), PvuII (2), HincII, Kpn I and SstII identified within the genomic insert (Figure 5.3). The two Pst I sites were of particular interest since the proC gene had been sequenced (Deutch et al., 1982) and shown to contain two PstI sites 800 bp apart. Fine restriction mapping showed that both PstI sites were located between the left-hand EcoRI[1] and BamHI[1] sites (see Figure 5.3). The two BamHI sites were separated by 2.7 kbp and the other mapped restriction sites were all located between these two sites (Figure 5.3). Close to the BamHI[2] site (within 50 bp) was a PvuII[2] site which itself was 1.2 kbp from another PvuII[1] site (Figure 5.3).

Plasmid pMH423 (2 μg) was digested with BamHI and the products separated on a 1% LMT agarose gel. Three bands were identified after ethidium bromide staining and low energy U.V. trans-illumination (Section 2.13.2). These migrated at 6.3, 4.3 and 2.7 kbp respectively. The three DNA bands corresponded to (i) 3 kbp-BamHI/EcoRI deleted pAT153 + ca. 3.3 kbp E.coli genomic DNA (6.3 kbp), (ii) BamHI[2] to vector

BamHI site (4.3 kbp) and (iii) BamHI[1] to BamHI[2] E.coli genomic DNA (2.7 kbp) as shown in Figure 5.3. All three bands were excised and their DNA purified. The 2.7 kbp and 4.3 kbp BamHI fragments were ligated into BamHI cut pAT153 while the 6.3 kbp BamHI fragment was recircularised with T4 DNA ligase (Sections 2.15.1 and 2.15.2). The three ligation mixes were used to transform $CaCl_2$-treated E.coli HB101 and the transformation mixes plated onto L-agar supplemented with 50 µg/ml ampicillin (Section 2.16). After overnight growth 50 amp^r colonies from each plate were replica plated onto L-agar supplemented with tetracycline (20 µg/ml). Several colonies of each with Amp^r Tet^s phenotypes (Table 5.2) were selected and plasmid DNA prepared (Section 2.11).

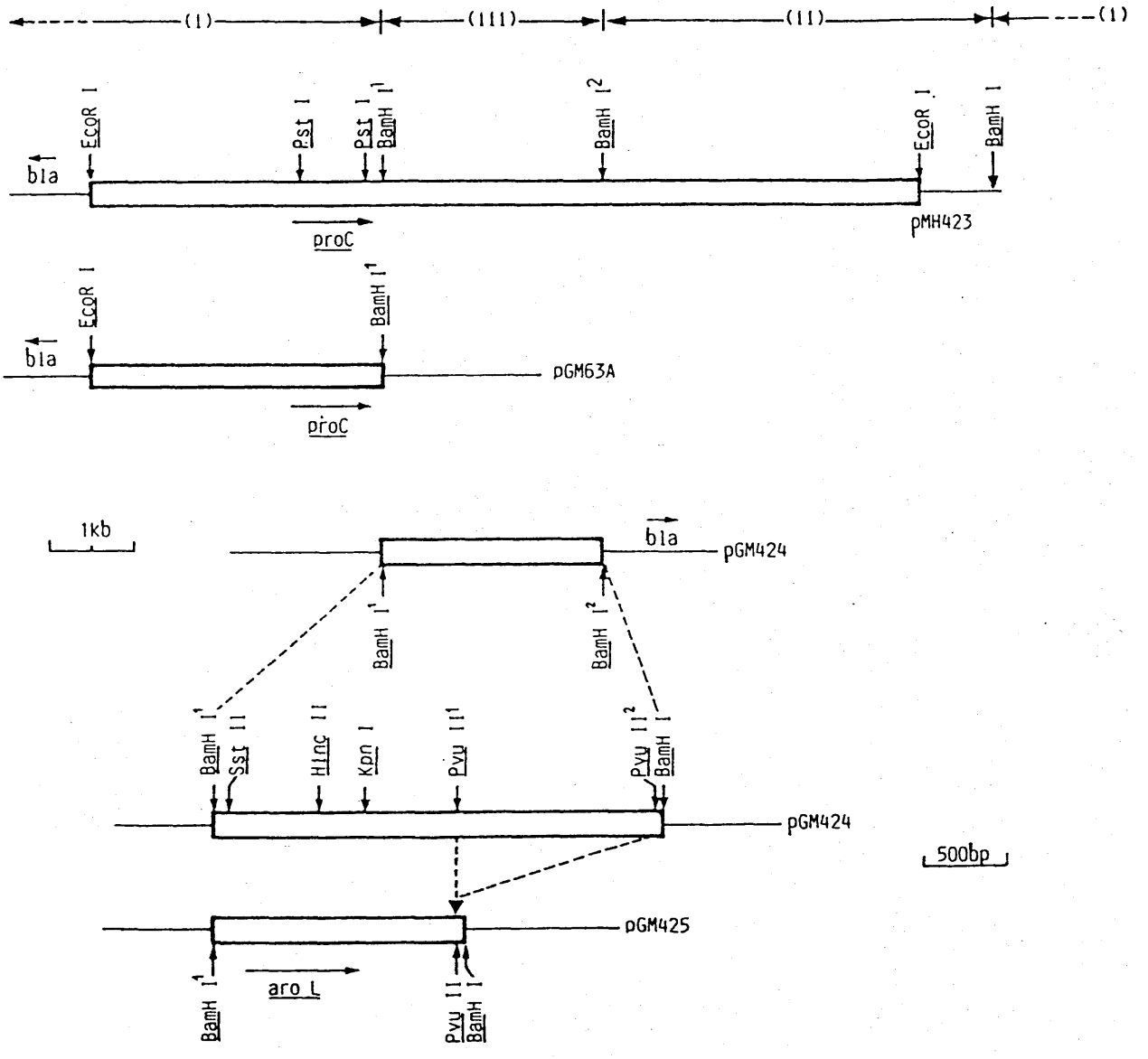
All three classes of recombinant plasmid contained the expected BamHI fragments, as verified by their restriction digest pattern (data not shown), and were each subsequently tested for the ability to complement E.coli HW0927 (proC). Only the 6.3 kbp BamHI (i) recircularised plasmid (plasmid pGM63A in Figure 5.3) complemented E.coli HW0927 (Table 5.2). No complementation was observed with the other two BamHI cloned fragments, pGM424 the 4.3 kbp (iii) fragment of pMH423 (Figure 5.3).

Plasmid pGM424 was further digested with PvuII and the products separated on a 1% LMT agarose gel. Deletion of the 1.2 kbp PvuII fragment from within the cloned insert of pGM424 resulted in a major band at 5.1 kbp. This band was

Ligation mix	Colonies on LA + amp	No. of 1st round colonies replicate plates (/50)	
		Colonies on LA + amp	Colonies on LA + Tet
2.7 kbp BamHI (iii) fragment/ pAT153	ca. 200	50	43
4.3 kbp BamHI (ii) fragment/ pAT153	ca. 250	50	29
6.3 kbp BamHI (i) fragment recircular-isation	ca. 200	50	8
ligation control	ca. 200	50	50
20 ng pAT153	ca. 300	50	50

Table 5.2: aroL subcloning.

Figure 5.3: Sub-cloning of the aroL gene. E.coli genomic DNA is shown by boxing, vector (plasmid) DNA is represented by solid lines. The positions of the proC and aroL are indicated. (i), (ii) & (iii) refer to the BamHI fragments discussed in the text. The deletion employed in creating pGM425 from pGM424 is shown as (▼). bla is the ampicillin resistance gene on the plasmid (β-lactamase).


(I)
(III)
(II)
(I)
EcoR I
Pst I
Pst I
BamH I 1
BamH I 2
EcoR I
BamH I
bla
proC
pMH423
EcoR I
BamH I 1
bla
pGM63A
proC
1kb
bla
pGM424
BamH I 1
BamH I 2
BamH I 1
Ssr II
Hinc II
Kpn I
Pvu II 1
Pvu II 2
BamH I
pGM424
500bp
pGM425
BamH I 1
aro L
Pvu II
BamH I

excised, the DNA phenol/chloroform purified (Section 2.13.2) and the DNA recircularised by ligation. This smaller derivative of pGM424, designated pGM425 was Amp^{r} Tet^{s} $ProC^{-}$ (Figure 5.3).

5.2.3 Levels of overexpression of the cloned shikimate kinase II

Crude extracts of 100 ml Minimal medium cultures of E.coli HB101, E.coli HB101/pMH423, E.coli HB101/pGM424, E.coli HB101/pGM425 and E.coli HB101/pGM63A were prepared (Section 2.6) and levels of shikimate kinase levels determined (Section 2.7.3).

An elevated level of shikimate kinase activity was observed in strains transformed with either plasmid pMG423 or pGM424 or pGM425. The coupled spectrophotometric assay for shikimate kinase measures the shikimate dependant oxidation of NADH (decreasing A_{340nm}; Section 2.7.3). The low levels at which this enzyme occurs naturally can be masked by the high endogenous NADH oxidase activity present in crude extracts even after high speed centrifugation. Quantitation of the exact level of overexpression (relative to untransformed E.coli HB101) depends therefore upon accurate measurement of the 'wild-type" (HB101) shikimate kinase level. To obtain a reliable wild-type (HB101) level of shikimate kinase activity and to compare the chromatographic behaviour of the wild-type and cloned activities the following experiment was performed.

A crude extract was prepared from 7 g (wet weight) cell paste of E.coli HB101/pMH423 as described in Section 2.23.3. The shikimate kinase level in crude extract was measured (0.35 μ/mg; protein estimation by method of Bradford (1976)) and 1.7 units of enzyme activity (ca. 5 mg protein) loaded onto a Pharmacia FPLC Mono Q anion exchange column. An increasing salt gradient was applied and eluted fractions assayed for shikimate kinase activity. The active fractions (27-29) were identified and the total recovery of activity calculated.

An identical procedure was used to work up a crude extract prepared from 7 g of untransformed E.coli HB101 cells (except no crude extract level of shikimate kinase level could be determined). Following chromatography on the same Mono Q column E.coli HB101 shikimate kinase activity was also found in fractions 27-29. This permitted a direct quantitation of the levels of overexpression (Table 5.3). Similarly crude extracts of the other plasmid-transformed strains were tested for shikimate kinase overexpression. The shikimate kinase activity always eluted in fractions 27-29. The results are summarised in Table 5.3.

The level of overexpression in pGM424 (2.7 kbp BamHI subclone of pMH423, see Figure 5.3) is double that of pMH433. More surprisingly pGM425, which has lost another 1.2 kbp of genomic DNA, shows a further doubling of enzyme activity. The overexpression in the case of pGM424 can be rationalised on the grounds of plasmid stability and copy-number factors.

Plasmid classification	*proC* complementation	shikimate kinase specific activity (units/mg protein) crude extract	FPLC	% activity recovered	shikimate kinase overexpression (fold)
None (wild type)	nd	nd	0.08	nd	1
pMH423	+	0.35	3.55	100	45
pGM63A	+	nd	nd	nd	nd
pGM424	-	1.36	8.3	75	103
pGM425	-	1.73	17.6	83	220

Table 5.3: Overexpression of shikimate kinase activity in plasmid transformed strains of *E. coli*.

A 2.7 kbp insert (pGM424) is potentially more stable than the larger 10 kbp insert of pMH423. The further increase in overexpression with pGM425 clearly establishes that the *aroL* gene is located within the 1.5 kbp *Bam*HI[1]/*Pvu*II[1] region of pGM424 and suggests that removal of the 1.2 kbp *Pvu*II fragment somehow either increases the stability of the *aroL* gene (or transcript) or promotes more efficient expression. Without additional evidence of copy-number level and mRNA half-life these alternatives cannot be distinguished.

Chromatography on this Mono-Q FPLC column also resolves the two shikimate kinase isoenzymes, which had previously been separated on a different kind of ion-exchange column by Ely & Pittard (1979). The putative shikimate kinase I activity consistently eluted in fractions 22-23 and was ca. 10% of the shikimate kinase II level in the untransformed *E.coli* HB101 preparation. This level of activity is consistent with the growth conditions under which the cells were prepared (growth on minimal medium). In the absence of the aromatic end products (tyrosine, tryptophan and phenylalanine) the *tyr*R-regulated shikimate kinase II becomes derepressed. Under these conditions shikimate kinase II can contribute as much as 90% of the total cellular shikimate kinase activity (Ely & Pittard, 1979).

The identical behaviour of the 'wild-type' (HB101) and the plasmid-encoded shikimate kinase II on these very high-resolution FPLC columns indicated that the overproduced

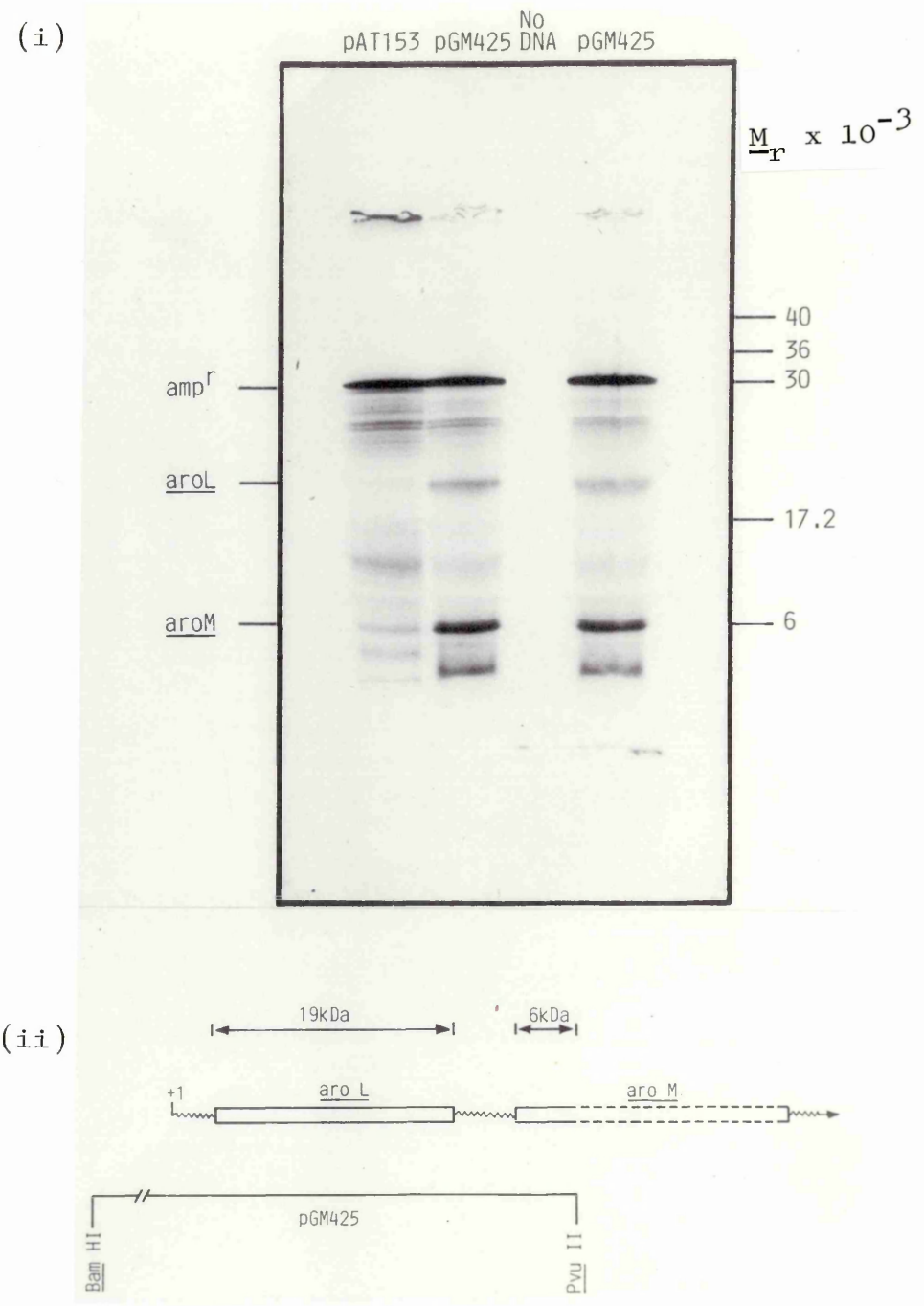


Figure 5.4: (i) In vitro expression of pGM425 and pAT153 as described in Section 5.2.4. ^{35}S-Met incorporation into shikimate kinase (aroL), vector encoded β-lactamase (amp^r) and the aroM gene product is indicated.

(ii) Extent of aroM coding potential in pGM425.

enzyme was either identical or at least very similar to the wild-type activity.

5.2.4 *In vitro* expression of the cloned shikimate kinase II gene

The results of *in vitro* DNA-directed coupled transcription/translation (Section 2.26) of 2.5 µg each of pAT153 and pGM425 are shown in Figure 5.4. Fractionation on a 15% SDS polyacrylamide gel resolved two main protein bands unique to pGM425-directed L- $[^{35}S]$ -methionine incorporation. The bands visualised on the subsequent autoradiogram (Figure 5.4) have M_r's of 19,000 and 6,000.and are designated *aroL* and *aroM* respectively. The nomenclature of the smaller protein band is after Defeyter & Pittard (1986) who have demonstrated that *aroL* is co-expressed on a single polycistronic mRNA with an as yet unidentified gene. The prefix *aro* implies aromatic biosynthetic/metabolic functionality and its assignation may prove premature.

The amount of incorporation of radionuclide into protein products is greater for *aroM* than *aroL* (Figure 5.4). This could imply that the smaller band is a premature termination product of the larger *aroL* band. Results presented in Section 5.3 suggest that the 6 kDa band is a distinct protein product and not a derivative of *aroL*. Indeed similar coupled transcription/translation of pGM424 demonstrates two major protein bands, encoded by the genomic insert, at 25 kDa and 19 kDa (M.G. Hunter unpublished work, cited in Millar *et al.*, 1986b). It is possible that in removing the 1.2 kbp *Pvu*II fragment to create pGM425 from pGM424 much

of the aroM coding region (at least 18-19 kDa) has been removed, leaving an N-terminal 6kDa truncated polypeptide. The region of aroM left in pGM425 after the PvuII deletion is approximately 7 kDa and contains 3 methionine residues (Defeyter & Pittard, 1986). The initial 7 kDa of aroL (67 amino acids) contains 2 methionine residues, and the only other methionine exists in the intact 19 kDa protein. It cannot therefore be unequivocably determined, on the grounds of ^{35}S-Met incorporation, whether the 6 kDa band observed in Figure 5.4 (aroM) is a fragment of aroL or a distinct protein product.

5.2.5 Genomic organisation of the aroL gene

The location in the genome of the E.coli aroL gene, predicted from subcloning and in vitro expression experiments indicates that it is situated on a 1.5 kbp BamHI/PvuII fragment, was directly examined by Southern hybridisation analysis. In addition, the observed linkage between aroL and another unidentified gene (aroM) within the 2.7 kbp BamHI region of the genome cloned in pGM424 (Defeyter & Pittard, 1986) was tested to preclude the possibility of an aberrant genome rearrangement having occurred during subcloning and propagation in recA$^+$ cells.

High molecular weight E.coli chromosomal DNA (5 µg) (generous gift of Mr S. Granger) was digested with either BamHI or BamHI and PvuII. The fragmented DNA was subjected

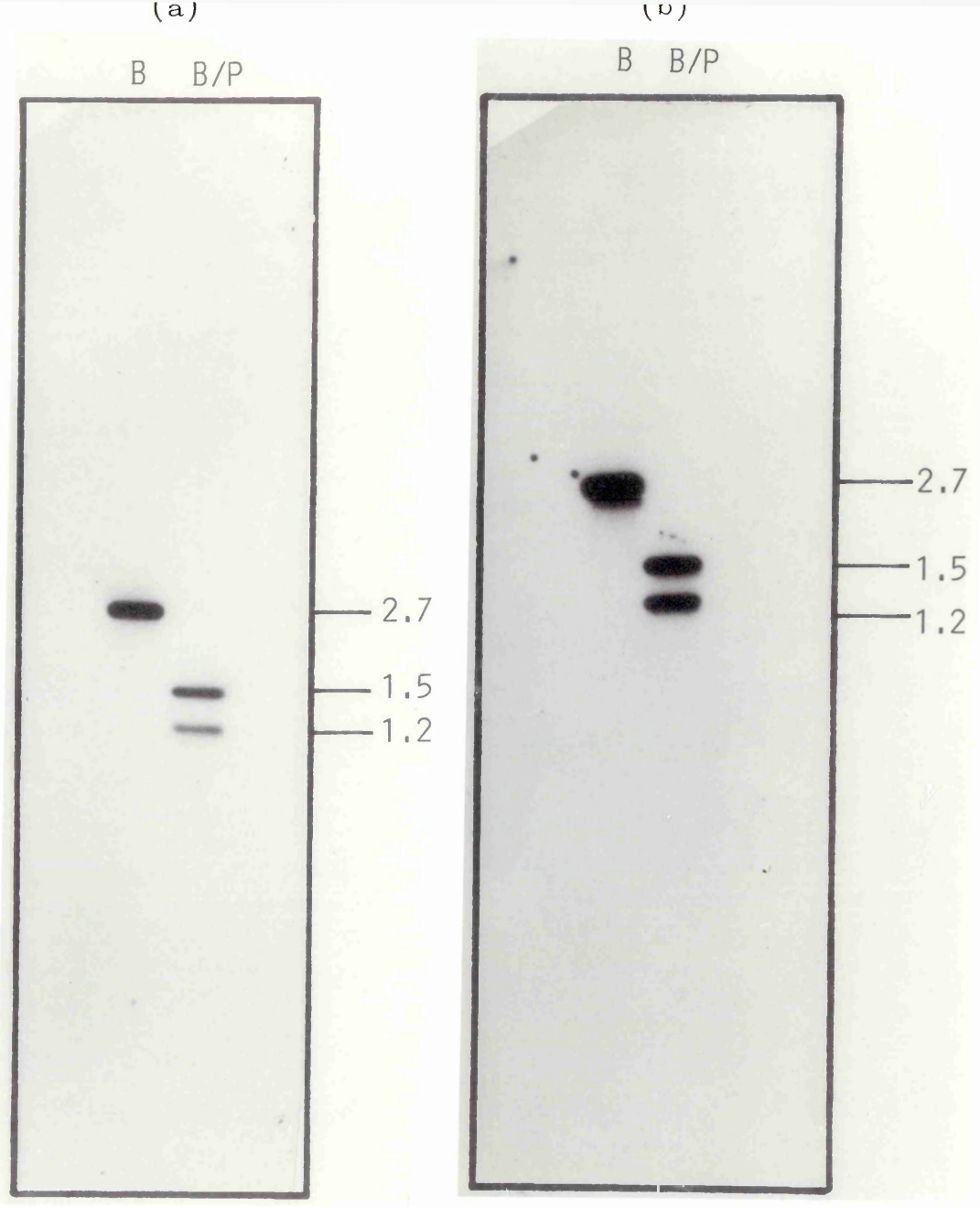


Figure 5.5: Southern hybridisation of the genomic aroL gene as described in Section 5.2.5. DNA was digested with BamHI (B) or BamHI and PvuII (B/P). Exposure was for (a) 4 hours or (b) overnight. Molecular sizes are indicated (kbp).

to electrophoresis on a 0.8% agarose gel and blot transferred to nitrocellulose (Section 2.20.1). The 2.7 kbp DNA insert of pGM424 was purified by excision of the appropriate band from LMT agarose (Section 2.13.2). The 2.7 kbp fragment was radiolabelled *in vitro* by nick-translation to a specific activity of 2.3×10^8 dpm/μg (Section 2.21) and used to probe the *aroL* gene. Prehybridisation, hybridisation and high stringency washing conditions were as described in Section 2.20. The result is shown in Figure 5.5.

The integrity of the 2.7 kbp BamHI fragment cloned in pGM424 appears to have been maintained from the chromosomal DNA to the plasmid. The 1.2 kbp PvuII region within the 2.7 kbp BamHI fragment, and its location close to the right-hand $BamHI^2$ (Figure 5.3) site, is also apparent on the chromosomal DNA. Any linkage (*aroL*/*aroM*, Defeyter & Pittard, 1986) within this region is therefore genuine and not due to rearrangement during cloning.

5.3 Sequence analysis of the *aroL* gene

5.3.1 Sequencing strategy, distribution of TaqI and Sau3A sites in pGM425

The *aroL* gene lies within a 1.5 kbp region of DNA bounded by BamHI and PvuII restriction sites. The direction of transcription of *aroL* is inferred from the results presented in Sections 5.2.3 and 5.2.4 and suggests that expression is from the $BamHI^1$ site (Figure 5.3) towards the PvuII site. The 1.5 kbp region of genomic DNA cloned within pGM425 was therefore chosen for sequence analysis. The pGM425 construct

was a BamHI clone (derivative of pGM424), with the right-hand BamHI[2] site very close to the PvuII hybrid site generated by the 1.2 kbp deletion (Section 5.2.2, Figure 5.3).

Plasmid pGM425 (2 μg) was digested with BamHI and subjected to electrophoresis on a 1% LMT agarose gel. The 1.5 kbp band was identified, excised and the DNA purified by phenol/chloroform extraction and ethanol precipitation (as in Section 2.13.2). The purified DNA was divided into three equal aliquots, one was maintained intact and the remaining two digested with either Sau3A or TaqI. Half of these secondary digestion products were purified by phenol/chloroform extraction and ethanol precipitation, while the remaining 50% was examined by 2% agarose gel electrophoresis. Digestion of the 1.5 kbp fragment with Sau3A was extensive (Figure 5.6(b)) resulting in a number of sub-fragments of 200 bp or less. The pattern of TaqI digestion was less extensive producing a number of larger, 300-500 bp bands.

The cohesive ends generated by BamHI (5'-GGATCC-3') and Sau3A (5'-GATC-3') digestion are the same. Therefore the Sau3A unfractionated secondary digestion mix of the 1.5 kbp BamHI DNA fragment of pGM425 was ligated into BamHI cut M13 mp8RF. Theoretically both the clone-defining BamHI ends and the internal Sau3A ends should be cloned. Sequence data representative of the entire 1.5 kbp region would then be obtainable from a large enough mixed population of recombinant bacteriophages. The TaqI mix was ligated into AccI cut M13 mp9RF and recombinant bacteriophage identified. The less widely

distributed TaqI sites would, hopefully, overlap and align their more frequent Sau3A counterparts. The intact 1.5 kbp BamHi fragment was also cloned into BamHI cut M13 mp8RF as sequence information from the resultant recombinant clones was required to identify the BamHI ends within the Sau3A mixed population. In total 36 Sau3A, 24 TaqI and 12 BamHI recombinant bacteriophage, constructed as described above, were identified and single stranded template prepared from each (Section 2.17).

5.3.2 1st round of DNA sequencing

A-track analysis (Section 2.17.10) of the twelve BamHI recombinant templates revealed only the expected two classes of DNA sequence data. Representatives of both classes were fully sequenced (Section 2.17) and the data analysed. One class had a PvuII site 45 bp from the BamHI cloning end and therefore was the right-hand $BamHI^2$ site (Figure 5.3). The other type of sequence data was, by elimination, from the $BamHI^1$ site and sequence information extending some 228 bp to a Sau3A site (passing a TaqI site at a position 201 bp from the $BamHI^1$ site) was obtained.

5.3.3 2nd round of DNA sequencing

Thirty six single stranded template preparations from Sau3A sub-cloning were examined by A-track analysis. Eighteen distinct types of sequence information were obtained and representatives of each class sequenced fully. Sixteen of these were quickly identified as eight pairs of complementary

sequences ranging from 228 bp long to as little as 30 bp (See Figure 5.6). The Sau3A site identified at position 228 bp from the BamHI[1] site (Section 5.3.2) was located and this sequence information verified by opposite strand sequencing. Two classes of sequence (160 bp) fragment contained an internal TaqI site at a position 140 bp from the cloned Sau3A end.

Although the order of sequence information could not be unequivocably assigned at this stage, the total data available amounted to some 1330 bp, over 1100 bp of which had been determined on both strands.

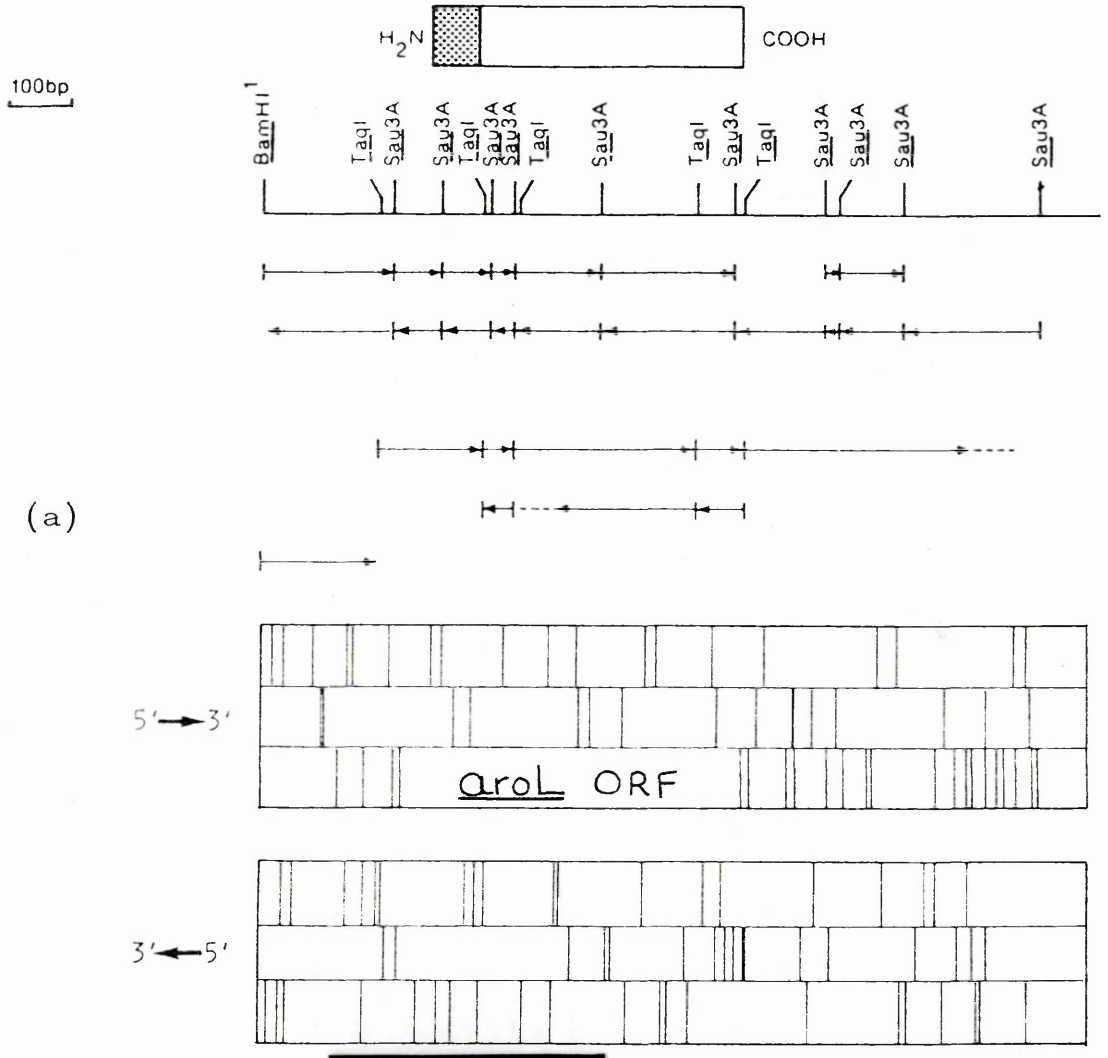
5.3.4 3rd round of DNA sequencing

Twenty four TaqI-derived recombinant single stranded templates were analysed by A-track sequencing and eight unique classes of DNA sequence data obtained. Six of the eight types of sequence information comprised three pairs of opposite strands. The remaining two DNA sequences included a short (170 bp) class and a large (sequence not fully determined) 450 bp class.

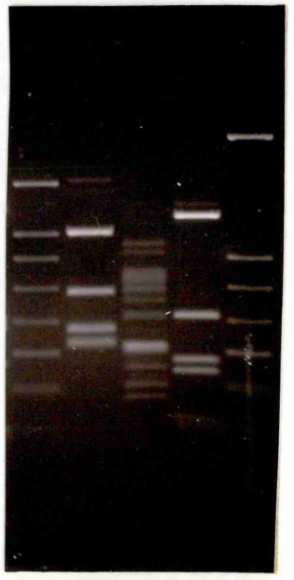
The information obtained in this round of sequencing was compared with the Sau3A (1330 bp) sequence data. The TaqI fragments overlapped the data already available from the Sau3A sequencing. This allowed the alignment of the ten 'blocks' of Sau3A sequences extending 1330 bp from the BamHI[1] site (Figure 5.6).

Figure 5.6 (facing)

(a) DNA sequencing strategy for the BamHI[1]/PvuII insert of pGM425. Above the restriction map, the shikimate kinase II (aroL) coding region is boxed and shaded indicating the extent of N-terminal amino acid sequencing. Stop codons in all possible (6) reading frames are indicated below. The direction and extent of each M13 clone sequenced is shown by an arrow.
(b) 2% Agarose gel profile of secondary restriction digests of the genomic insert of pGM425. Marker sizes from EcoRI/HinfI pAT153 (Track 1) and HinfI pAT153 (Track 5) are shown in bp. Secondary digests of insert DNA (prepared as described in Section 5.3.1) was with HpaII (Track 2), Sau3A (Track 3) or TaqI (Track 4). All digests were performed as described in the text.


H2N
COOH
100bp
BamHI
TaqI
Sau3A
(a)
5'→3'
aroL ORF
3'←5'

(b)


1631
1000
631
517
396
298
220
154
75
1 2 3 4 5

5.3.5 Compilation of sequence data

Each of the separate BamHI-, Sau3A- or TaqI-derived clones sequenced were entered as separate files into a DIGITAL PDP 11/34 computer. Where an internal concatemer was formed, each new sequence was assumed not to be a partial digestion product but treated as a separate file. Comparisons, alignments and overlaps were done essentially by eye and verified using the Staden programs (Section 2.18.1). The sequencing strategy is summarised in Figure 5.6.

The 1330 bp extending from the BamHI[1] site (Figure 5.7) was examined for convenient restriction sites. In addition to the known HincII (position 607), KpnI (position 940) and SstII (position 55) sites (Section 5.2.2, Figure 5.3), other sites at position 714 (NruI) and 314 (AvaI) were identified. Subsequent restriction mapping for these sites showed full correlation between predicted and experimentally observed data (gels (a) and (b) Figure 5.11).

5.3.6 The putative aroL coding region

The complete 1330 bp of DNA sequence data obtained are shown in Figure 5.7. Each nucleotide was verified by sequencing on the opposite strand and identifying the complementary base. All restriction sites used in the cloning were overlapped in reassembling the data.

Both strands were examined for potential open reading frames (ORF) using the program TRN TRP (Section 2.18.2). The positions of stop codons in all three reading frames of both strands are shown in Figure 5.6. A single ORF capable

Figure 5.7 (facing)

The double-strand DNA sequence of 1328 bp from the BamHI[1] site in pGM425 (Figure 5.3) towards the PvuII[1] site. The sequencing strategy has been shown previously in Figure 5.6.

1++A

```
     GATCCAGACCGGCGGACCAGATAGCCTTTCACAACGTGACCGCCAGGCCTTTGCCGCGGA
1    ---------+---------+---------+---------+---------+---------+  60
     CTAGGTCTGGCCGCCTGGTCTATCGGAAAGTGTTGCACTGGCGGTCCGGAAACGGCGCCT

     GCTGGAGAAGTGGTGGCTGGAAGTGCAACGTAGTCGTGGCTAAATGTAATTTATTATTTA
61   ---------+---------+---------+---------+---------+---------+  120
     CGACCTCTTCACCACCGACCTTCACGTTGCATCAGCACCGATTTACATTAAATAATAAAT

     CACTTCATTCTTGAATATTTATTGGTATAGTAAGGGGTGTATTGAGATTTTCACTTTAAG
121  ---------+---------+---------+---------+---------+---------+  180
     GTGAAGTAAGAACTTATAAATAACCATATCATTCCCCACATAACTCTAAAAGTGAAATTC

     TGGAATTTTTTCTTTACAATCGAAATTGTACTAGTTTGATGGTATGATCGCTATTCTCAT
181  ---------+---------+---------+---------+---------+---------+  240
     ACCTTAAAAAAGAAATGTTAGCTTTAACATGATCAAACTACCATACTAGCGATAAGAGTA

     GACACCGGCTTTCGCCGCATTGCGACCTATTGGGGAAAACCCACGATGACACAACCTCTT
241  ---------+---------+---------+---------+---------+---------+  300
     CTGTGGCCGAAAGCGGCGTAACGCTGGATAACCCCTTTTGGGTGCTACTGTGTTGGAGAA

     TTTCTGATCGGGCCTCGGGGCTGTGGTAAAACAACGGTCGGAATGGCCCTTGCCGATTCG
301  ---------+---------+---------+---------+---------+---------+  360
     AAAGACTAGCCCGGAGCCCCGACACCATTTTGTTGCCAGCCTTACCGGGAACGGCTAAGC

     CTTAACCGTCGGTTTGTCGATACCGATCAGTGGTTGCAATCACAGCTCAATATGACGGTC
361  ---------+---------+---------+---------+---------+---------+  420
     GAATTGGCAGCCAAACAGCTATGGCTAGTCACCAACGTTAGTGTCGAGTTATACTGCCAG

     GCGGAGATCGTCGAAAGGGAAGAGTGGGCGGGATTTCGCGCCAGAGAAACGGCGGCGCTG
421  ---------+---------+---------+---------+---------+---------+  480
     CGCCTCTAGCAGCTTTCCCTTCTCACCCGCCCTAAAGCGCGGTCTCTTTGCCGCCGCGAC

     GAAGCGGTAACTGCGCCATCCACCGTTATCGCTACAGGCGGCGGCATTATTCTGACGGAA
481  ---------+---------+---------+---------+---------+---------+  540
     CTTCGCCATTGACGCGGTAGGTGGCAATAGCGATGTCCGCCGCCGTAATAAGACTGCCTT

     TTTAATCGTCACTTCATGCAAAATAACGGGATCGTGGTTTATTTGTGTGCGCCAGTATCA
541  ---------+---------+---------+---------+---------+---------+  600
     AAATTAGCAGTGAAGTACGTTTTATTGCCCTAGCACCAAATAAACACACGCGGTCATAGT

     GTCCTGGTTAACCGACTGCAAGCTGCACCGGAAGAAGATTTACGGCCAACCTTAACGGGA
601  ---------+---------+---------+---------+---------+---------+  660
     CAGGACCAATTGGCTGACGTTCGACGTGGCCTTCTTCTAAATGCCGGTTGGAATTGCCCT

     AAACCGCTGAGCGAAGAAGTTCAGGAAGTGCTGGAAGAACGCGATGCGCTATATCGCGAA
661  ---------+---------+---------+---------+---------+---------+  720
     TTTGGCGACTCGCTTCTTCAAGTCCTTCACGACCTTCTTGCGCTACGCGATATAGCGCTT

     GTTGCGCATATTATCATCGACGCAACAAACGAACCCAGCCAGGTGATTTCTGAAATTCGC
721  ---------+---------+---------+---------+---------+---------+  780
     CAACGCGTATAATAGTAGCTGCGTTGTTTGCTTGGGTCGGTCCACTAAAGACTTTAAGCG

     AGCGCCCTGGCACAGACGATCAATTGTTGATTTTCGAGCGCCTATACTTAACGTTCATCC
781  ---------+---------+---------+---------+---------+---------+  840
     TCGCGGGACCGTGTCTGCTAGTTAACAACTAAAAGCTCGCGGATATGAATTGCAAGTAGG

     CGTGAAATAAGGAAGAACGATGCCAACGAAACCGCCTTATCCTCGTGAAGCATATATAGT
841  ---------+---------+---------+---------+---------+---------+  900
     GCACTTTATTCCTTCTTGCTACGGTTGCTTTGGCGGAATAGGAGCACTTCGTATATATCA

     GACGATTGAAAAAGGAAAGCCAGGACAGACGGTAACCTGGTACCAACTCAGAGCCGATCA
901  ---------+---------+---------+---------+---------+---------+  960
     CTGCTAACTTTTTCCTTTCGGTCCTGTCTGCCATTGGACCATGGTTGAGTCTCGGCTAGT

     TCCTAAACCAGACTCGTTGATCAGTGAACATCCGACCGCTCAGGAAGCGATGGATGCGAA
961  ---------+---------+---------+---------+---------+---------+  1020
     AGGATTTGGTCTGAGCAACTAGTCACTTGTAGGCTGGCGAGTCCTTCGCTACCTACGCTT

     AAAACGCTATGAGGACCCTGACAAAGAGTGACCGCATCAGACTGCTCGGAAGGGATTCTG
1021 ---------+---------+---------+---------+---------+---------+  1080
     TTTTGCGATACTCCTGGGACTGTTTCTCACTGGCGTAGTCTGACGAGCCTTCCCTAAGAC

     AGTGCCACTACAAGGGATCGCAACGACGCACTCATTGTTCATCCACCTTACTTTTCCTTT
1081 ---------+---------+---------+---------+---------+---------+  1140
     TCACGGTGATGTTCCCTAGCGTTGCTGCGTGAGTAACAAGTAGGTGGAATGAAAAGGAAA

     CGTCGTTAATTACCGGGCAAGTGTGAAGCACCATGCTGACATTACTTCTCGTACAAATGA
1141 ---------+---------+---------+---------+---------+---------+  1200
     GCAGCAATTAATGGCCCGTTCACACTTCGTGGTACGACTGTAATGAAGAGCATGTTTACT

     CAAAAAGCGTAGCAGCAGACGTGCACGGCATAATGTAAAGATTCCAAATGATTCCAGTAA
1201 ---------+---------+---------+---------+---------+---------+  1260
     GTTTTTCGCATCGTCGTCTGCACGTGCCGTATTACATTTCTAAGGTTTACTAAGGTCATT

     TGGATTTGTTATTGTTTAATATTCTAATTATTAGAAAAACATGAATTATGAAAAATGTGA
1261 ---------+---------+---------+---------+---------+---------+  1320
     ACCTAAACAATAACAAATTATAAGATTAATAATCTTTTTGTACTTAATACTTTTTACACT

     CGCAGATC
1321 --------  1328
     GCGTCTAG
```

of encoding a polypeptide of ca. 20 kDa sub-unit molecular weight was identified from positions 219 to 810 running from the BamHI[1] site (position 0) towards the PvuII[1] site. When translated from the first methionine (ATG) codon at position 286 the sequence predicted a polypeptide chain of M_r 19,068.

The area preceding the proposed translational initiation codon was examined for possible ribosome binding sites. Nine bases upstream of the position 286 ATG codon is the sequence 5'GGGGAAAA3' which could act as a weak ribosome binding site. Homology between the 3' 16S rRNA sequence (Shine & Dalgarno, 1975) and this proposed sequence is, however, poor, and confined to the sequence GGA (nucleotides 274-276).

5.3.7 TESTCODE analysis of the aroL coding region

The likelihood of this region being genuinely protein coding was examined using the TESTCODE algorithm (Section 2.18.3). The statistical basis for this quantitative assessment of protein coding probability has been discussed previously (Sections 2.18.3; 3.6.9).

Throughout the proposed aroL coding region the TESTCODE graphical output (Figure 5.8(a)) consistently displays a >95% probability of protein coding (panel A). Immediately preceding and following the putative shikimate kinase II gene, the measure of 'period three constraint' (Fickett, 1982) suggests a non-coding region (panel B, C; Figure 5.8(a)).

Conversely, an examination of the opposite strand (Figure 5.8(b)) displays neither large ORF's nor statistically likely coding regions. This would therefore appear to be the aroL complementary strand.

5.3.8 Further statistical examination of the aroL coding region

Staden (1984) described a similar algorithm for examining the 'preferential base frequency' of a nucleic acid sequence as a measure of protein coding potential. Rather than summing the statistical output (as done by Fickett's TESTCODE and therefore independent of reading frame), the ANALYSEQ (Staden, 1984) program examines each reading frame separately. The value of preferential base frequency (pbf) is computed for each reading frame (Section 2.18.4), the highest identified and a dot placed along the meridian of that reading frame. A potential protein coding region is identified by (i) lack of stop codons, (ii) a peak in the pbf and (iii) an undisturbed meridian line. Such an output is shown for the aroL coding region in Figure 5.9.

In addition potential ribosome binding sites are searched for by homology to sequences complementary to the 3' 16s rRNA sequence (Shine & Dalgarno, 1975) and marked in each frame by a vertical line. Similarly potential E.coli promoter sequences, both -35 and -10 elements (Rosenberg & Court, 1979; Hawley & McClure, 1983) are identified (panels A and B, Figure 5.9) by vertical lines. In both cases the height of the vertical 'score' is proportional to increasing homology with the conserved translational or transcriptional control sequences.

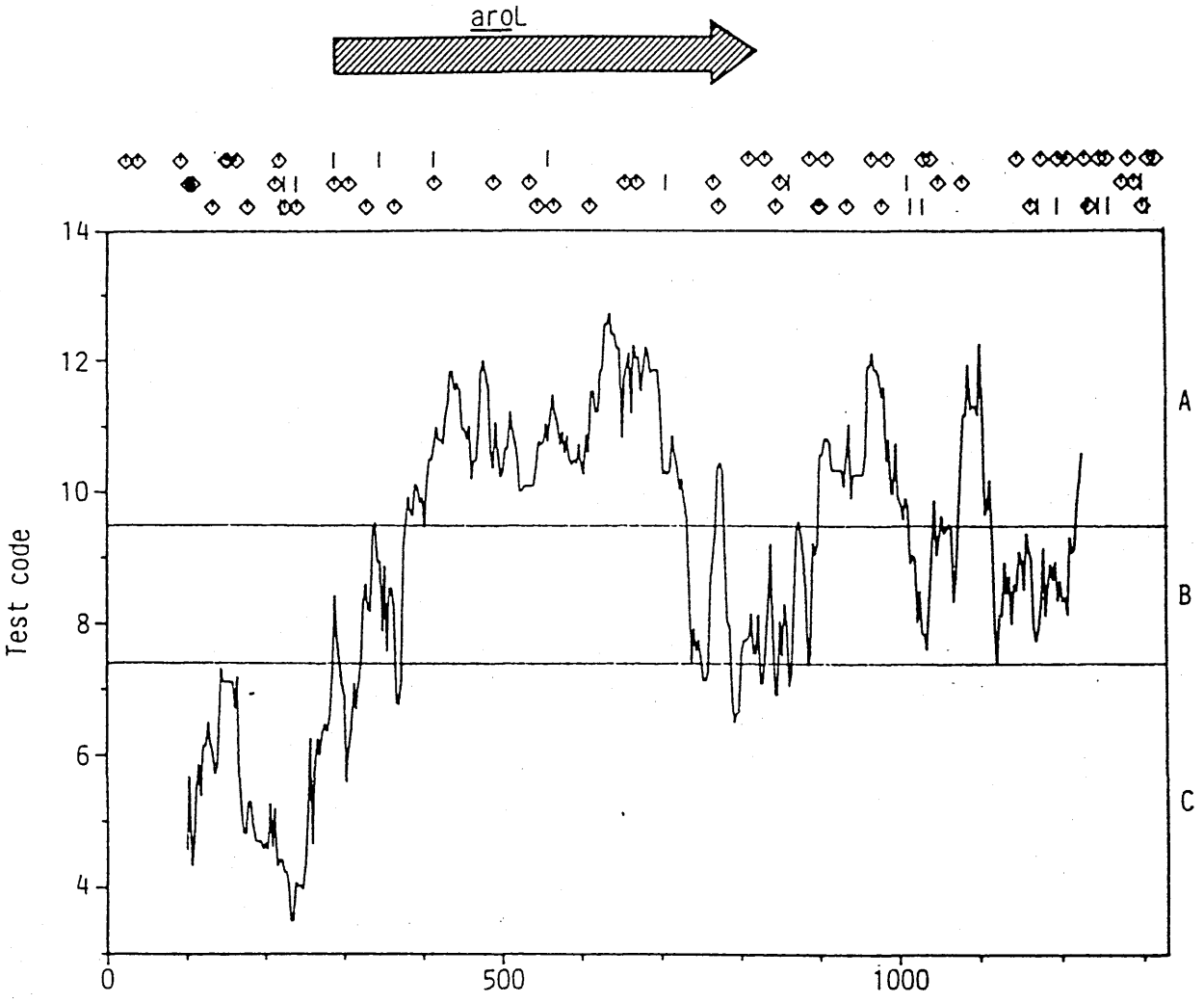


Figure 5.8(a): TESTCODE analysis of the aroL sense strand. All symbols and values are as described previously (in Figure 3.15).

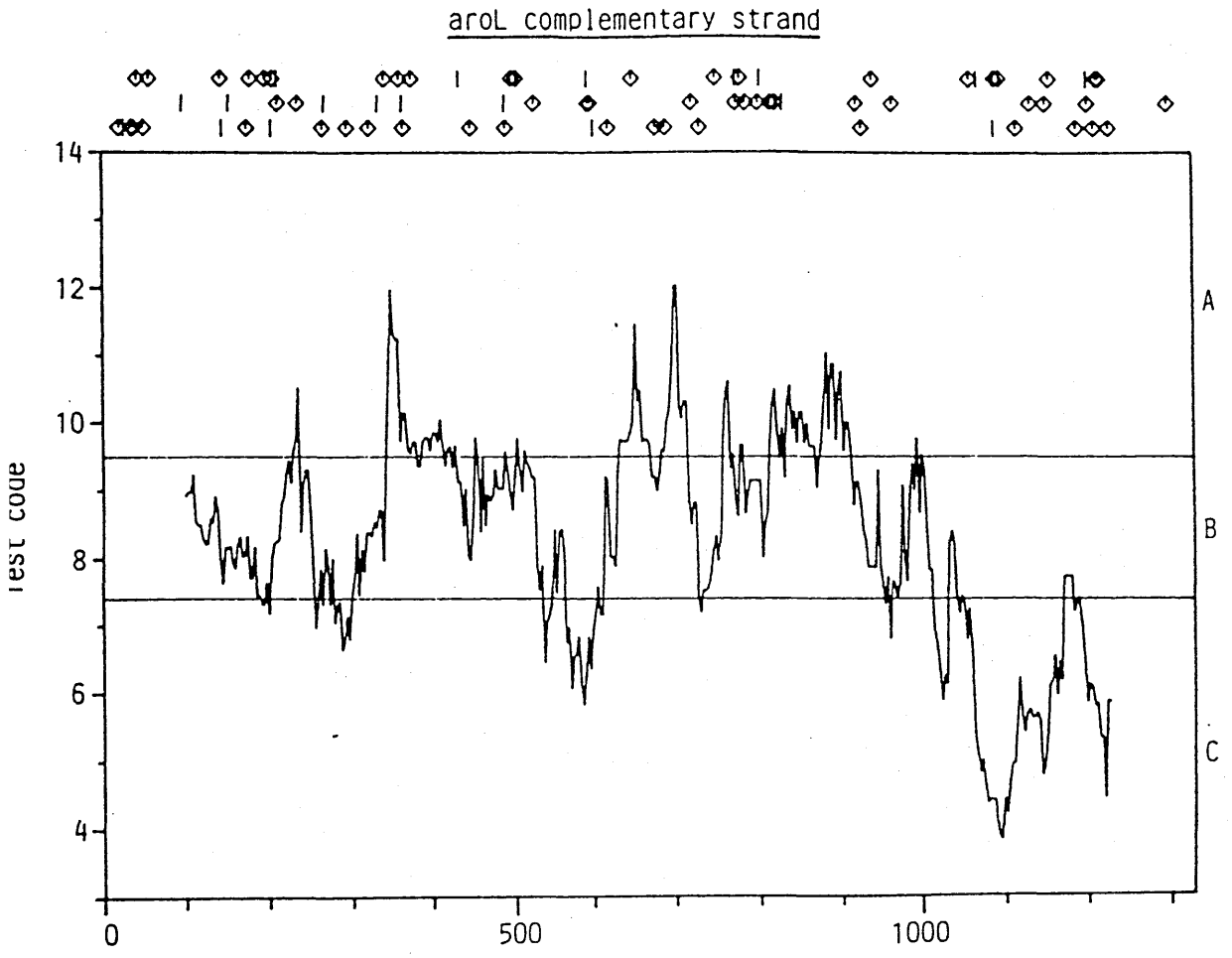


Figure 5.8(b): TESTCODE analysis of the aroL complementary (non-sense) strand. All symbols and values are as described previously (in Figure 3.15).

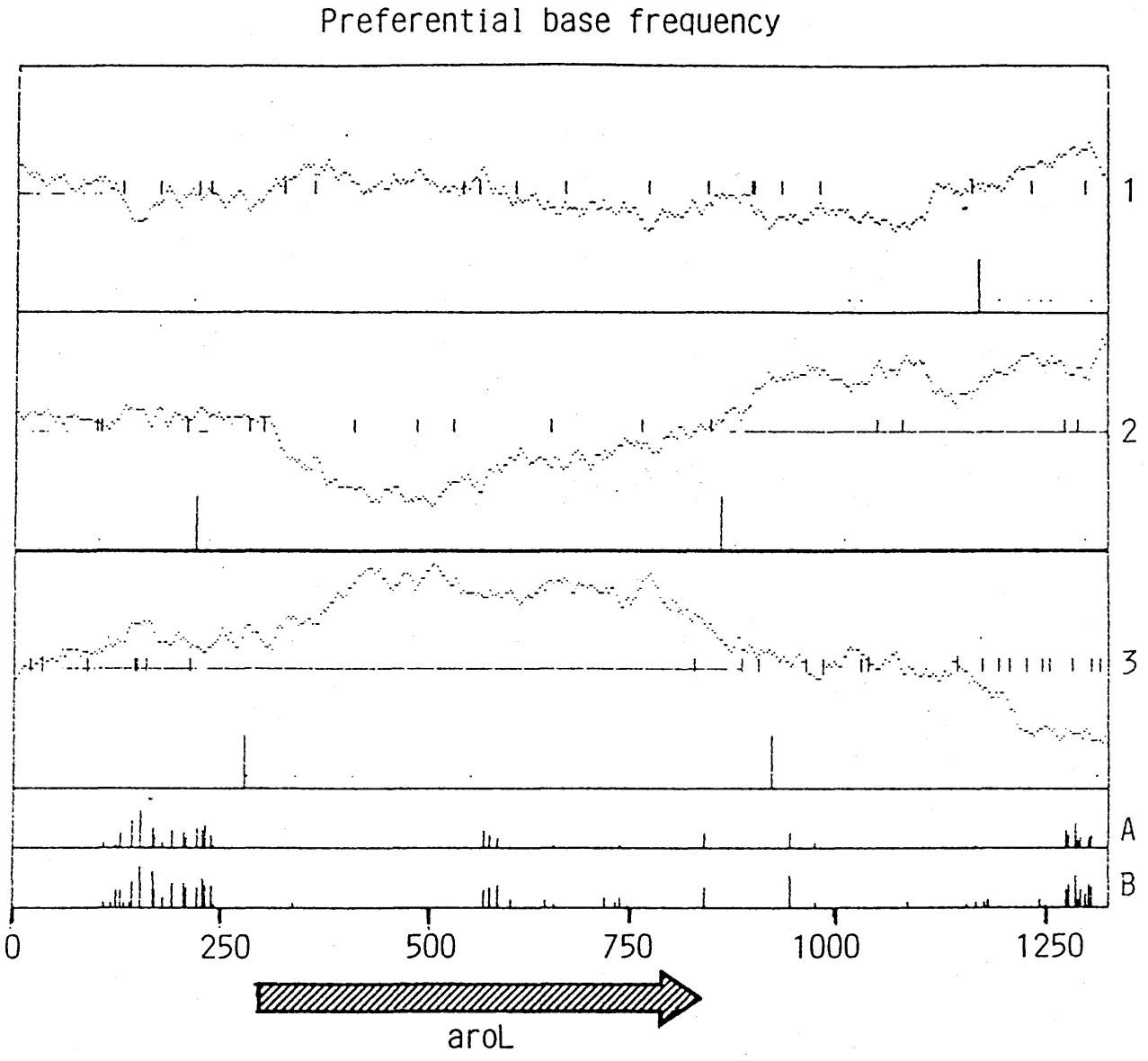


Figure 5.9: ANALYSEQ analysis of the aroL sense strand. Panels 1, 2 and 3 represent the three possible reading frames, vertical lines along the meridian are stop codons. Vertical lines along the base are potential Shine-Dalgarno sites. Panels A and B map potential consensus promoter sequences throughout the sequence length (abscissa).

The aroL gene is clearly identifiable in Figure 5.9. Immediately preceding the proposed translational start at position 286 is a predicted ribosome binding site. Also heavily concentrated around positions 140-200, and peaking at position 160, are sequences recognisable as E.coli promoter elements. The validity of these (TESTCODE and ANALYSEQ) computer predictions are examined experimentally in Sections 5.4 and 5.5.

5.3.9 Codon utilisation of the proposed aroL gene

The pattern of codon utilisation for the putative aroL gene (Table 5.4) further confirms this as a likely coding region. The bias against usage of so-called modulating codons (Gouy & Gautier, 1982; Grosjean & Fiers, 1982) is apparent for aroL and is discussed in greater detail in Section 4.5.4. The aroL gene would appear, on the frequency of modulating codon usage criterion, to be a weakly expressed E.coli gene, which is consistent with the low levels of shikimate kinase found in wild-type E.coli cells.

5.4 Definitive location of the aroL coding region

5.4.1 The N-terminal amino acid sequence of shikimate kinase II

In parallel with the molecular biological studies on the E.coli aroL gene, work in this laboratory has also led to the development of a reproducible purification scheme for the aroL-encoded shikimate kinase. A three-column purification protocol for shikimate kinase II from an over-producing strain (E.coli HW87/pMH423) has been developed by Dr A. Lewendon and is described in detail in Millar et al.

	U	aroL	C	aroL	A	aroL	G	aroL	
U	Phe	4	Ser	1	Tyr	2	Cys	1	U
	Phe	1	Ser	1	Tyr	0	Cys	0	C
	Leu	2	Ser	2	ochre	0	opal	1	A
	Leu	2	Ser	1	amber	0	Trp	1	G
C	Leu	3	Pro	2	His	1	Arg	2	U
	Leu	1	Pro	1	His	1	Arg	4	C
	Leu	1	Pro	3	Gln	4	Arg	1	A
	Leu	8	Pro	2	Gln	5	Arg	3	G
A	Ile	5	Thr	1	Asn	4	Ser	0	U
	Ile	7	Thr	3	Asn	4	Ser	3	C
	Ile	0	Thr	4	Lys	2	Arg	1	A
	Met	1	Thr	6	Lys	0	Arg	1	G
G	Val	5	Ala	2	Asp	5	Gly	1	U
	Val	5	Ala	4	Asp	1	Gly	4	C
	Val	2	Ala	3	Gln	5	Gly	3	A
	Val	3	Ala	9	Gln	1	Gly	2	G

Table 5.4: aroL codon utilisation.

(1986b). An overall yield of 35% was obtained and a purification of 80-fold resulted in homogeneous shikimate kinase as judged by polyacrylamide-gel electrophoresis in the presence of SDS. The apparent sub-unit M_r of the purified *E.coli* shikimate kinase II was estimated to be 20,000 under these conditions.

The availability of purified enzyme allowed the determination of the N-terminal amino acid sequence on a liquid phase sequencer. This work is summarised in Figure 5.10 and was carried out by Dr A. Lewendon (Millar *et al.*, 1986b). The first 24 N-terminal residues were unambiguously identified and agreed exactly with the nucleotide-predicted protein sequence. The single exception to this was that the fMet translational initiation residue was missing, presumably lost post-translationally (Figure 5.10). Nevertheless this information confirmed the *aroL* gene coding region as predicted in Section 5.3.

5.4.2 Amino acid composition of shikimate kinase II

The amino acid composition of an acid hydrolysate of shikimate kinase was determined (Dr A. Lewendon, Millar *et al.*, 1986b). Comparison with the values predicted from the deduced nucleotide sequence and the experimentally-obtained values are shown in Table 5.5. The overall agreement between the two is excellent. There can be no doubt that the polypeptide encoded by nucleotides 289 to 810 (in Figure 5.7) is the 18,937 (-fMet) molecular weight shikimate kinase II. The *aroL* coding region, including 5' and 3' flanking sequences, is shown in Figure 5.12.

Residue no.	Amino acid identified (as phenylthio-hydantoin)	Yield (nmol)
1	Thr	2.3
2	Gln	13.0
3	Pro	8.5
4	Leu	11.5
5	Phe	12.3
6	Leu	11.5
7	Ile	10.6
8	Gly	10.0
9	Pro	8.0
10	Arg	12.2
11	Gly	7.6
12	Cys	5.2
13	Gly	7.8
14	Lys	56.
15	Thr	2.9
16	Thr	3.3
17	Val	7.1
18	Gly	6.7
19	Met	6.9
20	Ala	6.3
21	Leu	6.7
22	Ala	6.2
23	Asp	6.3
24	Ser	0.5

Figure 5.10: N-Terminal amino acid sequence of E.coli shikimate kinase II.

The sequence was determined, as described in Materials and Methods section, on 20 nmol of protein. The repetitive yield from residues 1-24, by least-squares regression analysis, was 97% (correlation coefficient 0.83).

E.coli shikimate kinase (aroL)

Residue	Relative amino acid composition (Leu = 17)	Predicted amino acid composition
Cys [c]	3.2	3
Asx	13.6	14
Met [a]	3.3	4
Thr [b]	13.2	14
Ser [b]	7.3	8
Glx	27.1	26
Pro	8.8	8
Gly	10.2	10
Ala	18.3	18
Val	14.5	15
Ile	11.6	12
Leu	17	17
Tyr	2.1	2
Phe	5.5	5
His	2.1	2
Lys	2.4	2
Arg	12.5	12
Trp	Nd	2

Table 5.5: Amino acid composition of E.coli shikimate kinase compared with that predicted from the aroL gene sequence

a. Determined as methionine sulphone
b. Experimental values at t = 0
c. Determined as cysteic acid.

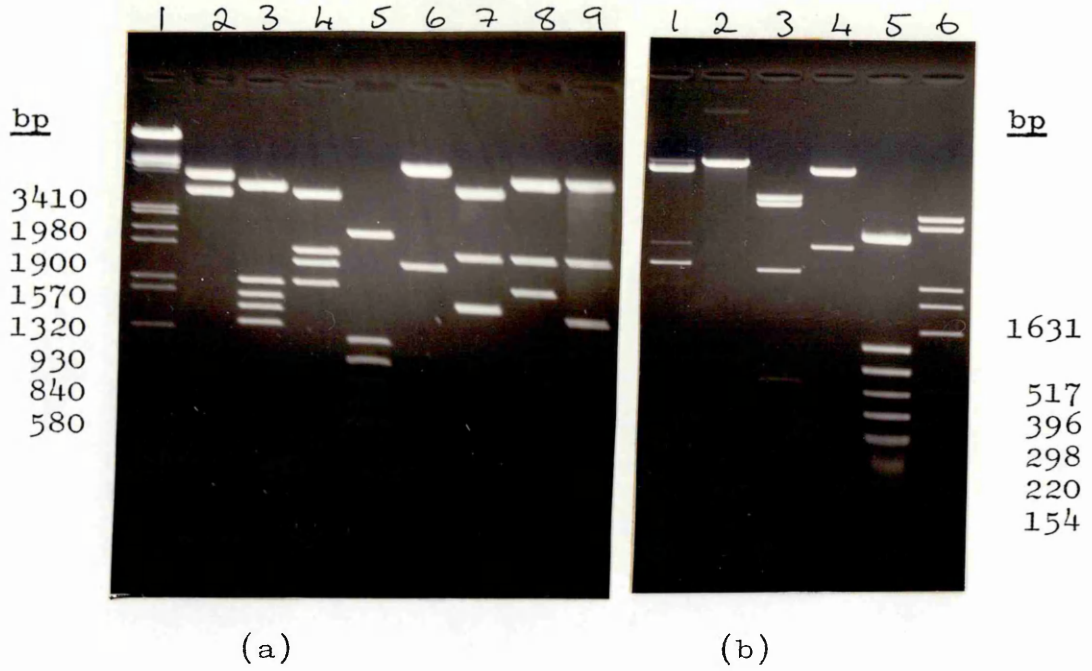


Figure 5.11: (a,b) 1% Agarose gel profiles identifying NruI and AvaI sites predicted from the DNA sequencing data. Marker fragment sizes are shown in bp, for EcoRI/HindIII λ (gel(a) Track 1) and HinfI pAT153 (gel (a) Track 5; gel (b) Track 5).

(a) Track	2	BamHI	(b) Track	1	SstII/AvaI
	3	BamHI/NruI		2	SstII
	4	HindIII/NruI		3	BamHI/AvaI
	6	PvuII/NruI		4	AvaI
	7	PvuII/SstII		6	AvaI/HincII
	8	SstII/NruI			
	9	KpnI/NruI			

All digests (other than markers) were of pGM424 DNA.

Figure 5.12 (facing)

The *E.coli* shikimate kinase coding region (*aroL* gene encoded). Note the initiation methionine lost post-translationally ([*]) and the position of the proposed ribosome binding site (RBS). The protein coding sequence was obtained using the program TRN TRP on the primary DNA sequence data.

```
                .         .         .         .         .         .
   1  GATCCAGACCGGCGGACCAGATAGCCTTTCACAACGTGACCGCCAGGCCTTTGCCGCGGA
                .         .         .         .         .         .
  61  GCTGGAGAAGTGGTGGCTGGAAGTGCAACGTAGTCGTGGCTAAATGTAATTTATTATTTA
                .         .         .         .         .         .
 121  CACTTCATTCTTGAATATTTATTGGTATAGTAAGGGGTGTATTGAGATTTTCACTTTAAG
                .         .         .         .         .         .
 181  TGGAATTTTTTCTTTACAATCGAAATTGTACTAGTTTGATGGTATGATCGCTATTCTCAT
                .         .         .         .         .         .
 241  GACACCGGCTTTCGCCGCATTGCGACCTATTGGGGAAAACCCACGATGACACAACCTCTT
                                        RBS          [*]ThrGlnProLeu    [5]
                .         .         .         .         .         .
 301  TTTCTGATCGGGCCTCGGGGCTGTGGTAAAACAACGGTCGGAATGGCCCTTGCCGATTCG
      PheLeuIleGlyProArgGlyCysGlyLysThrThrValGlyMetAlaLeuAlaAspSer    [25]
                .         .         .         .         .         .
 361  CTTAACCGTCGGTTTGTCGATACCGATCAGTGGTTGCAATCACAGCTCAATATGACGGTC
      LeuAsnArgArgPheValAspThrAspGlnTrpLeuGlnSerGlnLeuAsnMetThrVal    [45]
                .         .         .         .         .         .
 421  GCGGAGATCGTCGAAAGGGAAGAGTGGGCGGGATTTCGCGCCAGAGAAACGGCGGCGCTG
      AlaGluIleValGluArgGluGluTrpAlaGlyPheArgAlaArgGluThrAlaAlaLeu    [65]
                .         .         .         .         .         .
 481  GAAGCGGTAACTGCGCCATCCACCGTTATCGCTACAGGCGGCGGCATTATTCTGACGGAA
      GluAlaValThrAlaProSerThrValIleAlaThrGlyGlyGlyIleIleLeuThrGlu    [85]
                .         .         .         .         .         .
 541  TTTAATCGTCACTTCATGCAAAATAACGGGATCGTGGTTTATTTGTGTGCGCCAGTATCA
      PheAsnArgHisPheMetGlnAsnAsnGlyIleValValTyrLeuCysAlaProValSer   [105]
                .         .         .         .         .         .
 601  GTCCTGGTTAACCGACTGCAAGCTGCACCGGAAGAAGATTTACGGCCAACCTTAACGGGA
      ValLeuValAsnArgLeuGlnAlaAlaProGluGluAspLeuArgProThrLeuThrGly   [125]
                .         .         .         .         .         .
 661  AAACCGCTGAGCGAAGAAGTTCAGGAAGTGCTGGAAGAACGCGATGCGCTATATCGCGAA
      LysProLeuSerGluGluValGlnGluValLeuGluGluArgAspAlaLeuTyrArgGlu   [145]
                .         .         .         .         .         .
 721  GTTGCGCATATTATCATCGACGCAACAAACGAACCCAGCCAGGTGATTTCTGAAATTCGC
      ValAlaHisIleIleIleAspAlaThrAsnGluProSerGlnValIleSerGluIleArg   [165]
                .         .         .         .         .         .
 781  AGCGCCCTGGCACAGACGATCAATTGTTGATTTTCGAGCGCCTATACTTAACGTTCATCC
      SerAlaLeuAlaGlnThrIleAsnCysEnd [173]
                .         .         .         .         .         .
 841  CGTGAAATAAGGAAGAACGATGCCAACGAAACCGCCTTATCCTCGTGAAGCATATATAGT
                .         .         .         .         .         .
 901  GACGATTGAAAAAGGAAAGCCAGGACAGACGGTAACCTGGTACCAACTCAGAGCCGATCA
                .         .         .         .         .         .
 961  TCCTAAACCAGACTCGTTGATCAGTGAACATCCGACCGCTCAGGAAGCGATGGATGCGAA
                .         .         .         .         .         .
1021  AAAACGCTATGAGGACCCTGACAAAGAGTGACCGCATCAGACTGCTCGGAAGGGATTCTG
                .         .         .         .         .         .
1081  AGTGCCACTACAAGGGATCGCAACGACGCACTCATTGTTCATCCACCTTACTTTTCCTTT
                .         .         .         .         .         .
1141  CGTCGTTAATTACCGGGCAAGTGTGAAGCACCATGCTGACATTACTTCTCGTACAAATGA
                .         .         .         .         .         .
1201  CAAAAAGCGTAGCAGCAGACGTGCACGGCATAATGTAAAGATTCCAAATGATTCCAGTAA
                .         .         .         .         .         .
1261  TGGATTTGTTATTGTTTAATATTCTAATTATTAGAAAAACATGAATTATGAAAAATGTGA

1321  CGCAGATC  1328
```

5.5 Transcript mapping of the aroL gene

5.5.1 Preparation of RNA

Total cellular RNA was prepared from an exponentially growing culture (50 ml, L broth + amp) of E.coli HB101/pGM425. The method of RNA isolation followed that of Aiba et al. (1981) with the modifications described in Section 2.19.1. The RNA preparation was examined by agarose gel electrophoresis and ethidium bromide staining. The pattern observed has been shown previously (Chapter 3, Figure 3.23).

5.5.2 Primer extension analysis

A synthetic oligonucleotide primer, 25 nucleotides long (sequence 5'-TAC CAC AGC CCC GAG GCC CGA TCA G-OH-3') was a gift of Dr M.G. Hunter. This oligonucleotide is complementary to nucleotides 18-43 in the aroL gene coding sequence. It was annealed to RNA prepared from E.coli HB101/pGM425 (Section 5.5.1), and primer extension synthesis carried out and analysed as described in Chapter Two (Section 2.19). The results are shown in Figure 5.13.

The length of the oligonucleotide-primed reverse run-off (primer extension product) was measured by comparison with a M13mp8 sequence ladder (Figure 5.13). The 5' end of the aroL transcript was identified as an A residue 125 bp upstream of the initiation methionine codon (ATG). Regions of potential secondary structure at the 5' end of the mRNA resulted in a number of premature-termination bands (see I and II, Figure 5.13). The largest (168 bases), most intensely radio-labelled primer extension product (A, Figure

5.13) maps accurately at position 161.

Analysis of the nucleotide sequence immediately upstream of the transcriptional start site revealed a -10 and a -35 region (16 bp apart) with considerable homology to the consensus sequence for E.coli promoter sequences (Hawley & McClure, 1983). The -10 region contains all 4 highly conserved nucleotides (5' TA-A-T 3') together with three of the five less strongly conserved elements. The -35 region has the sequence 5' ACTTCATTC 3' which lacks only two of the strongly conserved nucleotides 5' TTGAC- 3' (Hawley & McClure, 1983). The transcript start itself is located as the centre nucleotide in the triplet 5' TAT 3'. Typically the transcriptional start site for E.coli genes is an A flanked 5' and 3' by C and T respectively. The aroL promoter sequences are shown in Figure 5.14.

5.5.4 A potential aroL operator?

In addition to the full length (168b) reverse run-off product (A), two other faint primer extension bands (I & II) can be seen in Figure 5.13. These correspond to positions 98 and 102 bp upstream of the translational start. Both are located in a region spanning some 20 bp (87-105 bp upstream of the ATG) which is very A/T rich and has a potential imperfect 18 bp two-fold rotational symmetry. This may account for the occurrence of premature reverse primed transcript terminations at locations I and II (Figure 5.13). This region (L3) is the third of a triplet of 18 bp imperfect repeats (L1 and L2 are the others) found in the vicinity of

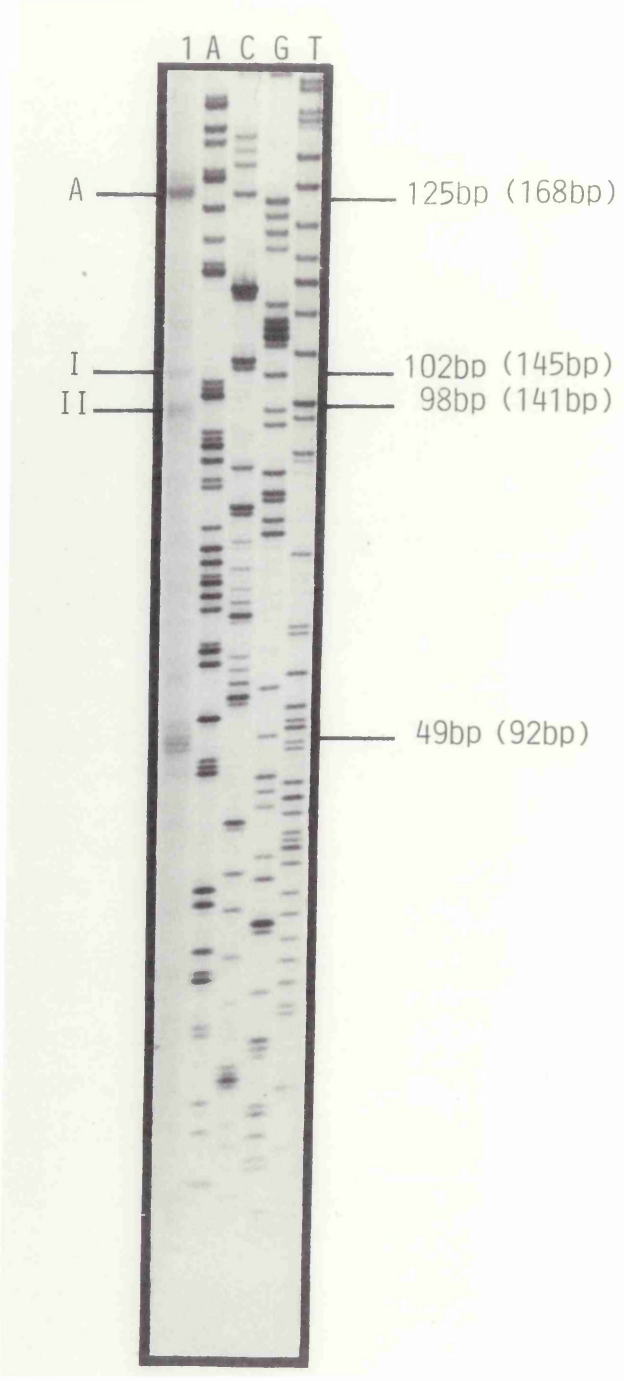


Figure 5.13: Primer extension transcript mapping of pGM425 encoded RNA (Section 5.5.2) is shown in track 1. M13mp8 (tracks A, C, G, T) was sequenced as a size marker and is shown in (bp). The full length (A) and premature termination products (I and II) are discussed in the text; and their distances upstream of the AUG codon are shown in bp.

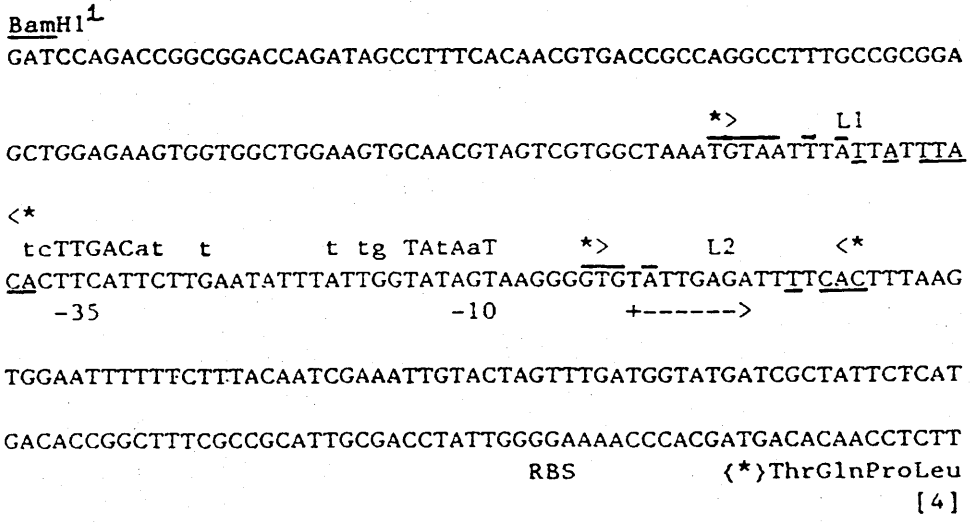


Figure 5.14: The region upstream of the aroL gene identified by transcript mapping (Figure 5.13) as the aroL promoter. Note the regions of dyad symmetry L1 and L2 which straddle the aroL -35 and -10 promoter sequences. The BamHI[1] site (Figure 5.3) is shown for reference.

transcript start site (Figure 5.15(a)).

The E.coli aroF and aroG genes are repressed by the TyrR gene product (Section 5.1.2) complexed with tyrosine and phenylalanine (Camakaris & Pittard, 1983). Garner & Herrmann (1985) have isolated constitutive aroF mutants with lesions in the regulatory region of the aroF gene. The mutations were found in two 18 bp imperfect repeats whose axis of symmetry are separated by 51 bp. These sequences have been proposed as the aroF operator and the site of tyrR repressor binding (Garner & Herrmann, 1985). The two arms of the repressor binding site ($aroF_{O1}$ and $aroF_{O2}$) are situated 61 and 113 bp, respectively, upstream of the transcriptional initiation site (Figure 5.15) and the $aroF_{O1}$ imperfect repeat overlaps the -35 region of the RNA polymerase binding site.

The axes of the 18 bp imperfect repeats (aroL1, L2, L3, Figure 5.15(a)) found close to the aroL promoter occur at -28, +6 and +30 respectively. The aroL1 and aroL2 'arms' are separated (axially) by 52 bp, a situation very similar to the aroF operator sequences. The aroL1 sequence overlaps the 5' end of the -35 region of the aroL promoter. The aroL2 imperfect repeat overlaps the transcriptional start site. The aroF and aroL operator sequences share several conserved nucleotides as can be seen in Figure 5.15(b).

By analogy to the mutationally defined aroF operator we propose that these sequences (aroL1 and aroL2) are involved in binding the tyrR regulator protein. The location of the transcript start site and -35 region of the promoter within these possible regulatory sequences strongly supports

this hypothesis. The occurrence of a third (aroL3) sequence may be fortuitous or may impart a potential for selectivity in the operator to differentially bind the tyrR repressor complexed with either tyrosine or tryptophan. The three dimensional structure of the trpR repressor (regulating the aroH gene) has recently been solved at 2.6 Å resolution (Schevitz et al., 1985). Binding of the L-tryptophan co-repressor to the aporepressor induces conformational changes which stabilize the functionally critical orientation of the active repressor responsible for operator-specific contacts. It is possible that analogous conformational changes occur within the tyrR repressor depending upon the nature of the ligand bound, and are manifest as alternate operator-specific contacts. This may explain the presence of three regions (L1, L2 and L3) of two-fold rotational symmetry in the aroL promoter region.

All three aroL 18 bp imperfect repeats exhibit some homology with the consensus prokaryotic DNA-binding protein recognition sequence $TGTGTN_{(6-10)}ACACA$ (Gicquel-Sanzey & Cossart, 1982).

5.5.5 Derepression of aroL in a tyrR mutant

The presence of tyrR-regulated sequences upstream of aroL was confirmed by comparison of the specific activities of shikimate kinase II (from pGM424) in a $tyrR^+$ and $tyrR^-$ background. Duplicate minimal medium cultures (50 ml) of E.coli HB101/pGM424 and E.coli HW1045 (tyrR)/pGM424 were each grown to exponential phase and crude extracts were

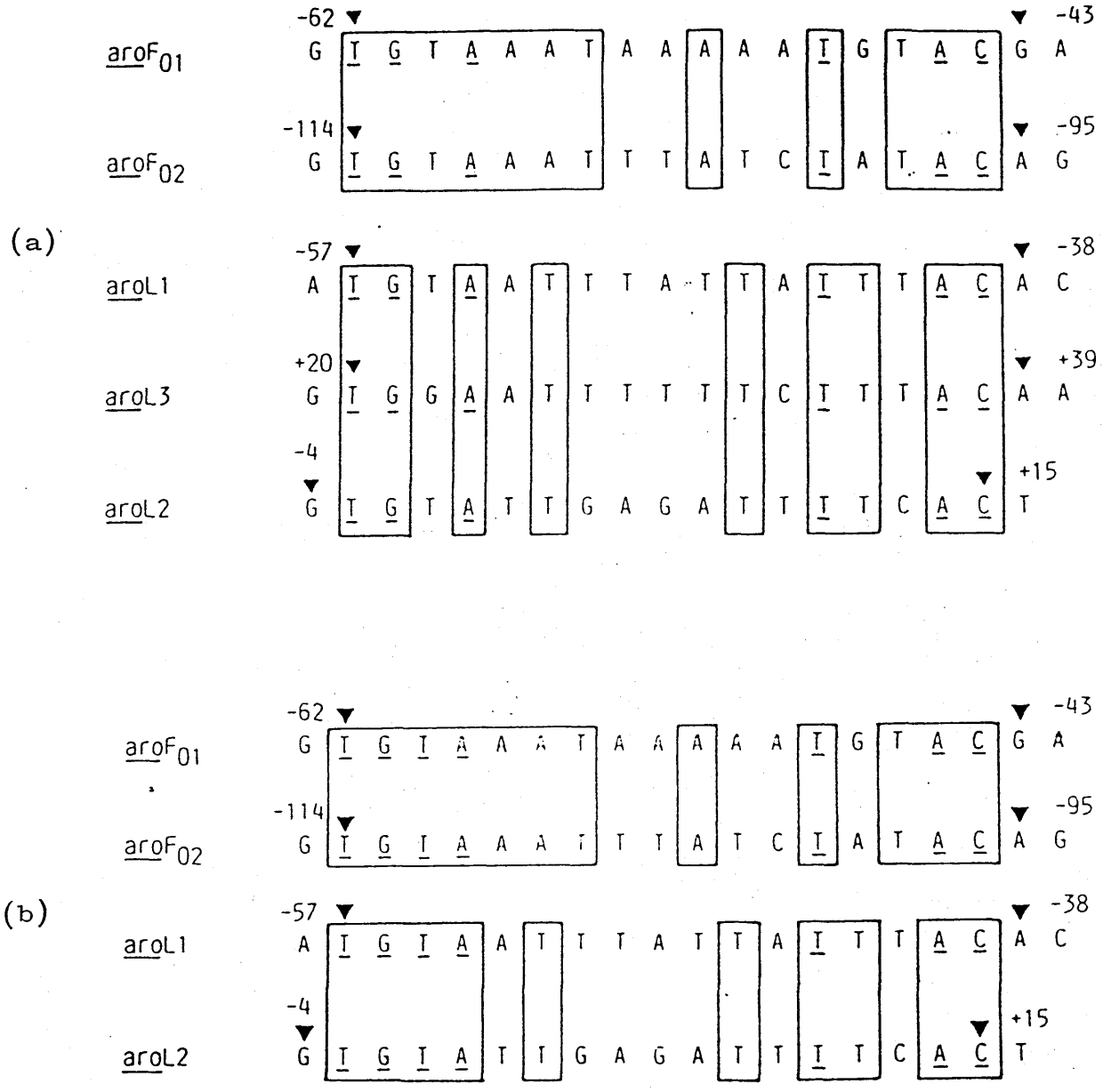


Figure 5.15: (a) Potential inverted repeat operator sequences upstream of the aroL gene, compared with the aroF operator (Garner & Herrmann, 1985).

(b) 52 bp-separated inverted repeats in aroF and aroL, common nucleotides are underlined.

All numbering is w.r.t. the respective transcript start sites.

prepared (Section 2.6). The specific activity of pGM424-overexpressed shikimate kinase II was 50% higher in a *tyrR* background (1.35 u/mg) than in the 'wild-type' (HB101) strain (0.9 u/mg). Although comparison in this instance is between two non-isogenic strains, an internal control (3-dehydroquinase level) indicated that at least one other shikimate pathway enzyme activity level was directly comparable.

De-repression of the chromosomal *aroL* allele in un-supplemented media can result in a 5-10 fold increase in shikimate kinase activity (Ely & Pittard, 1979). The limited de-repression in pGM424-encoded shikimate kinase levels (*tyrR*$^-$) suggests that in a wild-type cell harbouring pGM424 the *tryR* repressor is being titrated out by the large numbers of *aroL* operators. This would allow considerable escape synthesis of shikimate kinase. Under normal (*tyrR*$^+$) conditions most operators would be free of repressor and hence any subsequent de-repression is confined to a sub-population of plasmid molecules. This hypothesis, supported by the experimental data and by a similar phenomenon reported by Defeyter & Pittard (1986), has one major flaw. *TyrR* is autoregulatory and therefore titration by a large number of plasmid encoded operators would result in increased *tyrR* expression. However the nature of the *tyrR* operator has yet to be reported, and one could postulate that the *tyrR* repressor may have a higher affinity for its own operator. A *tyrR* operator-specific dissociation constant significantly lower than for any other operator in its regulation might ensure maximal sensitivity

in the regulation of its own expression. Under these conditions the limited de-repression of plasmid encoded operators would occur as observed for aroL. Despite the limited de-repression, the results do suggest that the aroL operator is intact in pGM424.

5.5.6 Organisation of the aroL operator

Both the argR (Cunin et al., 1983) and the lexA (Little et al., 1981; Brent & Ptashne, 1981) repressors recognise multiple DNA binding sites similar to those proposed for tyrR binding to aroF (Garner & Herrmann, 1985) and aroL (Millar et al., 1986b; Defeyter & Pittard, 1986). In each case the basic operator sequence is a 18 bp stretch of DNA which exhibits an imperfect two-fold rotational symmetry. Gicquel-Sanzey & Cossart (1982) have recently compared homologies between several different prokaryotic DNA-binding regulatory proteins and their respective sites of action. The observation that gene-regulatory proteins are often multimeric suggests sub-unit symmetrical binding with such regions of dyad symmetry as a general type of DNA-regulatory protein interaction (Gicquel-Sanzey & Cossart, 1982).

It has also been proposed that the separation distance between multiple palindromic sequences within an operator sequence can be directly correlated with the molecular weight of the repressor (Garner & Herrmann, 1985). The tyrR product at 63 kDa (Cornish et al., 1982) recognises sequences some 51 bp apart (Garner & Herrmann, 1985) while the ArgR product at 12 kDa (Cunin, 1983) binds to palindromes 3 bp apart

Figure 5.16 (facing)

The E.coli aroL gene, encoding shikimate kinase. The positions of the potential operator sequences L1 and L2 are shown (* → ←*). The transcript start site is marked (+1) and the consensus (Hawley & McClure, 1983) promoter sequences shown above the mapped aroL promoter.

```
GATCCAGACCGGCGGACCAGATAGCCTTTCACAACGTGACCGCCAGGCCTTTGCCGCGGA

                                                    *⟶    L1
GCTGGAGAAGTGGTGGCTGGAAGTGCAACGTAGTCGTGGCTAAATGTAATTTATTATTTA

⟵*
  tcTTGACat  t       t tg TAtAaT      *⟶    L2    ⟵*
CACTTCATTCTTGAATATTTATTGGTATAGTAAGGGGTGTATTGAGATTTTCACTTTAAG
   -35                        -10        +1

                                          +-----⟶

TGGAATTTTTTCTTTACAATCGAAATTGTACTAGTTTGATGGTATGATCGCTATTCTCAT

GACACCGGCTTTCGCCGCATTGCGACCTATTGGGGAAAACCCACGATGACACAACCTCTT
                                    RBS          (*)ThrGlnProLeu
                                                              [4]
TTTCTGATCGGGCCTCGGGGCTGTGGTAAAACAACGGTCGGAATGGCCCTTGCCGATTCG
PheLeuIleGlyProArgGlyCysGlyLysThrThrValGlyMetAlaLeuAlaAspSer
                                                             [24]
CTTAACCGTCGGTTTGTCGATACCGATCAGTGGTTGCAATCACAGCTCAATATGACGGTC
LeuAsnArgArgPheValAspThrAspGlnTrpLeuGlnSerGlnLeuAsnMetThrVal
                                                             [44]
GCGGAGATCGTCGAAAGGGAAGAGTGGGCGGGATTTCGCGCCAGAGAAACGGCGGCGCTG
AlaGluIleValGluArgGluGluTrpAlaGlyPheArgAlaArgGluThrAlaAlaLeu
                                                             [64]
GAAGCGGTAACTGCGCCATCCACCGTTATCGCTACAGGCGGCGGCATTATTCTGACGGAA
GluAlaValThrAlaProSerThrValIleAlaThrGlyGlyGlyIleIleLeuThrGlu
                                                             [84]
TTTAATCGTCACTTCATGCAAAATAACGGGATCGTGGTTTATTTGTGTGCGCCAGTATCA
PheAsnArgHisPheMetGlnAsnAsnGlyIleValValTyrLeuCysAlaProValSer
                                                            [104]
GTCCTGGTTAACCGACTGCAAGCTGCACCGGAAGAAGATTTACGGCCAACCTTAACGGGA
ValLeuValAsnArgLeuGlnAlaAlaProGluGluAspLeuArgProThrLeuThrGly
                                                            [124]
AAACCGCTGAGCGAAGAAGTTCAGGAAGTGCTGGAAGAACGCGATGCGCTATATCGCGAA
LysProLeuSerGluGluValGlnGluValLeuGluGluArgAspAlaLeuTyrArgGlu
                                                            [144]
GTTGCGCATATTATCATCGACGCAACAAACGAACCCAGCCAGGTGATTTCTGAAATTCGC
ValAlaHisIleIleIleAspAlaThrAsnGluProSerGlnValIleSerGluIleArg
                                                            [164]
AGCGCCCTGGCACAGACGATCAATTGTTGA
SerAlaLeuAlaGlnThrIleAsnCysEnd
                            [173]
```

(Cunin et al., 1983). The validity of this observation and the possibility of co-operative effects in repressor binding to the dual sites in the operator sequences has yet to be substantiated.

5.6 Increased overexpression of aroL-encoded shikimate kinase II

5.6.1 Strategy

The purification scheme for shikimate kinase II from the (cq. 45-fold) overexpressing strain E.coli HW87/pMH423 (Section 5.4) yielded 95 units of activity (1.3 mg of pure protein, Table 5.6(b)) from 16 g of cells. In order to obtain much more enzyme for detailed mechanistic and structural studies the aroL gene was placed under the control of the very strong tac promoter (Section 3.2, Figure 3.3). This was achieved by cloning a mutationally-altered aroL gene into the expression vector pKK223/3.

5.6.2 Site-directed mutagenesis of the aroL ribosome binding site

The proposed aroL ribosome binding site, 5' GGGAAA 3', Figure 5.16, was mutationally altered to a BamHI site by site-directed mutagenesis (M.G. Hunter, unpublished work). A synthetic oligonucleotide with a three base mis-match (Figure 5.17(a)) was used to create the new BamHI site. Mutants were selected by screening for the appearance of the additional BamHI site and confirmed by direct DNA sequencing (M.G. Hunter, unpublished work). The altered sequence still contained a limited homology with the consensus ribosome binding site sequence (Shine & Dalgarno, 1975).

5.6.3 tac-aroL construct pGM450

1 μg of recombinant M13 mp8RF containing the mutated aroL RBS on a 2.7 kbp BamHI insert (gift of M.G. Hunter) was digested with BamHI and subjected to electrophoresis on a 1% LMT agarose gel. A 2.5 kbp fragment and a 250 bp fragment were observed, rather than an intact 2.7 kbp fragment (pGM424 insert, Figure 5.3), and the 2.5 kbp fragment excised and its DNA purified. The 2.5 kbp fragment was ligated into BamHI cut pAT153, transformed into competent E.coli HB101 and recombinant Amp^r Tet^s colonies identified. One such recombinant was designated pGM429 and an internal PvuII deletion (1.2 kbp deletion) made to it using the same procedure as that used to generate pGM425 from pGM424 (Section 5.2.2, Figure 5.3). This new construct, pGM430, was essentially identical to pGM425 except its BamHI insert was 1.3 kbp and not 1.5 kbp as in pGM425 (see Figure 5.17(b)).

The 1.3 kbp BamHI insert of pGM430 was excised from a 1% LMT agarose gel and ligated to BamHI cut pIH223/3 (gift of Dr I.S. Hunter), before transformation into E.coli HW1111 ($lacI^q$). This tac expression vector is a derivative of pKK223/3 (Section 3.4.1, Figure 3.3) which has lost its vector BamHI site but retains the more useful polylinker BamHI site. Several Amp^r colonies were tested for 1.3 kbp BamHI inserts and one plasmid (pGM450) was selected as the tac-aroL construct.

The intermediate cloning into pAT153 (to generate pGM429) was required since both pIH223/3 and M13mp8RF have

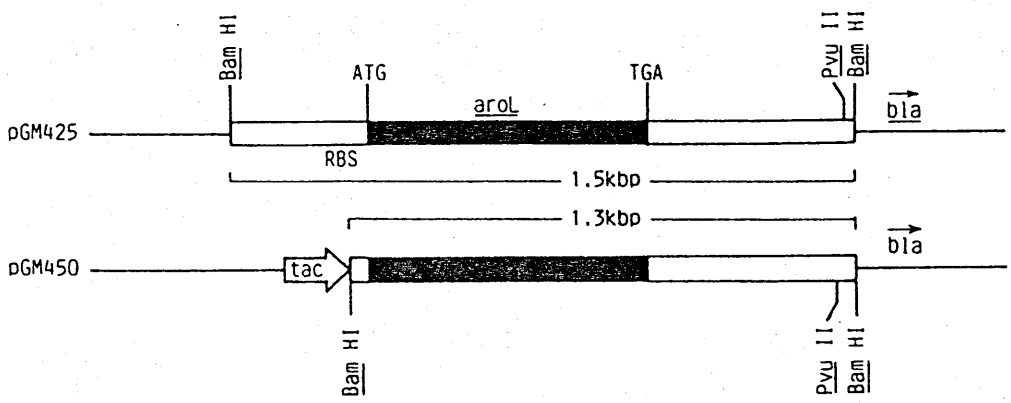
Figure 5.17 (facing)

(a) Mutagenesis used in the creation of a new BamHI site (see B*, (b) over) in mp8:mutRBS aroL (M.G. Hunter, unpublished).

(b) Summary of the construction of tac-aroL plasmid pGM450 as described in the text (Section 5.6.3).

(c) 1% Agarose gel profiles of restriction digests of pGM429 and pGM430.

gel (i) Track 1 EcoRI/HindIII λ (marker)
2 BamHI/PvuII pGM424
3 BamHI/PvuII pGM429

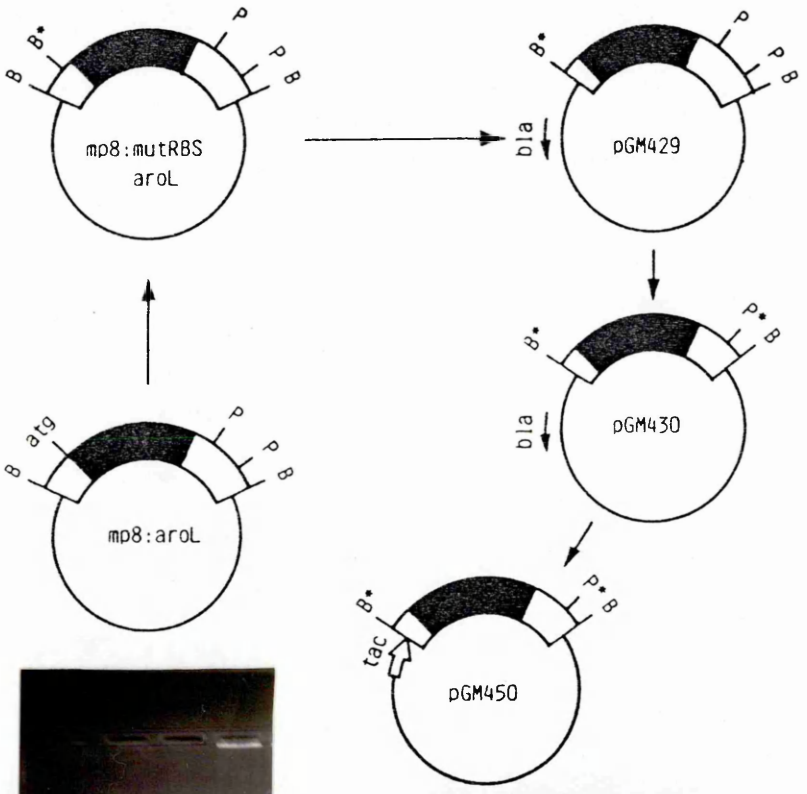


gel (ii) Track 1 EcoRI/HindIII λ (markers)
2 BamHI pGM425
3 BamHI pGM430
4 BamHI pGM430
5 HinfI pAT153 (markers)

(a)

```
260                RBS          (*)---->
5'TTGCGACCTATTGGGGAAAACCCACGATGAC 3'  WT.SEQ
3'AACGCTGGATAACCCCTTTTGGGTGCTACTG 5'
                    ** *
       acctattggggGGaTcccacgatgac      oligo.seq
                   BamHI
5'ttgcgacctattggggGGATCCcacgatgac 3' MUTANT.SEQ
```

(b)

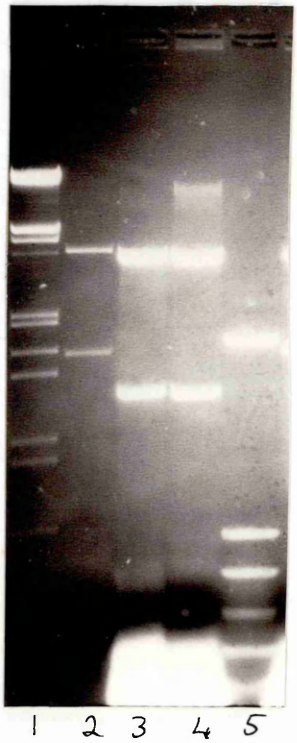


(c)

bp
1980
1900
1570
1320
930
840
580

1 2 3

(i)



bp
1631
1570
1320
930
840
580
396
220

1 2 3 4 5

(ii)

vector PvuII sites. As a result these could not be used during the deletion of the internal 1.2 kbp PvuII region from pGM429. Plasmid pGM450 (tac-aroL) construction is summarised in Figure 5.17(b).

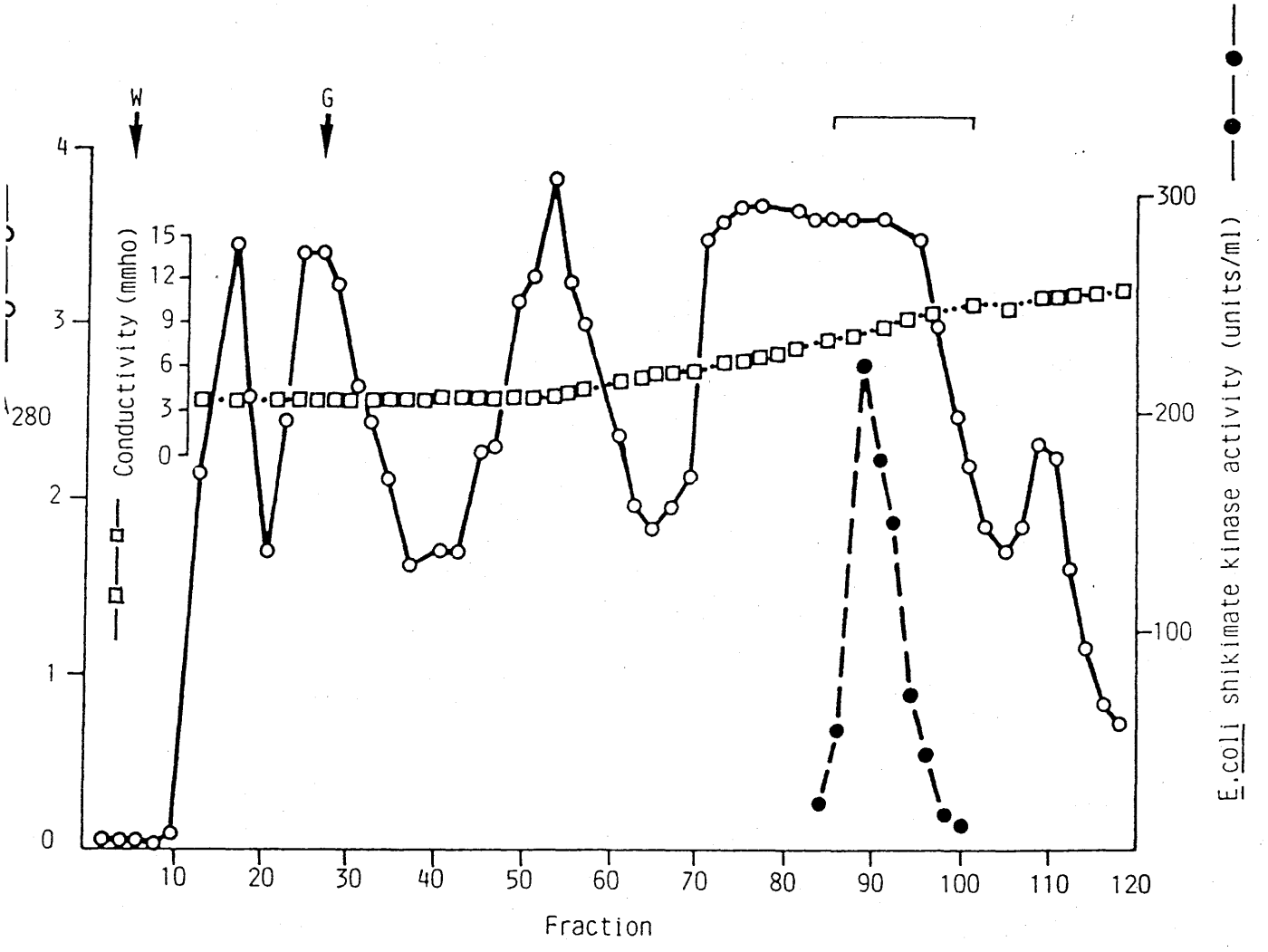
5.6.4 Purification of shikimate kinase II from a tac-aroL strain

(i) 8 x 500 mls Lbroth + amp. (pre-warmed to 37°C) were each inoculated with 15 mls of an overnight culture (Lbroth + amp.) of E.coli HW1111/pGM450. After growth, with vigorous shaking, for 1 hour at 37°C, IPTG was added to a final concentration of 0.5 mM. The cells were grown for a further 9 hours before harvesting (Section 2.23.1) and storage at -20°C as a cell paste.

(ii) 17 g (wet weight) of E.coli HW1111/pGM450 were thawed slowly, broken by two passages through a French Pressure cell and extracted into 30 mls of 50 mM-Tris/HCl buffer, pH 7.5 containing 50 mM-KCl, 5 mM - $MgCl_2$ and 0.4 mM DTT (Buffer A, Section 2.23.4). The supernatant of a post-DNase treated high speed centrifugation was termed the crude extract (Section 2.23.4). The enzyme preparation scheme is outlined in Section 2.23.4 and is based on the method developed in this laboratory by Dr A. Lewendon (Millar et al., 1986b).

The total shikimate kinase activity in crude extract was 23,200 units. This was 100-fold greater than the level observed in a crude extract prepared from 16 g of E.coli HW87/pMM423 (Millar et al., 1986b; Table 5.6(b)). The latter overproducing strain was overexpressing shikimate kinase ca. 45-fold relative to the wild-type level (Section

Figure 5.18: DEAE-Sephacel Chromatography of E.coli shikimate kinase. Step 2 of purification detailed in Section 2.23.4 (Table 5.6B).



Crude extract, dialysed in Buffer A, containing 23,220 units was loaded on to the column and washed (w) with 350 ml of Buffer A (Section 2.23.4). A linear gradient (600 ml) of 50-300mM-KCl in Buffer A was applied (G). Flow rate after loading was 60 ml/hr and 14 ml fractions were collected. The pooled fractions are indicated.

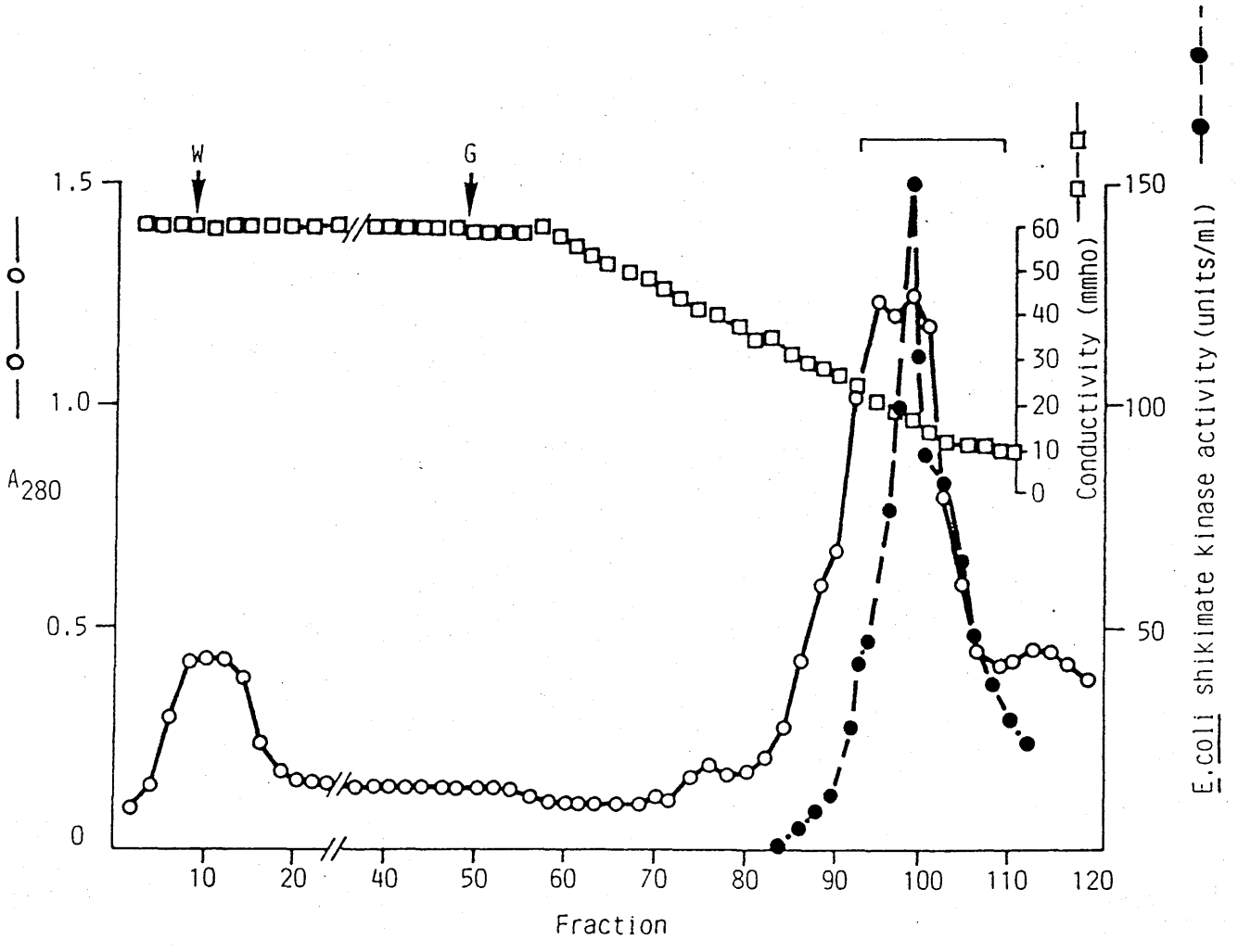
5.2.3). On this basis the tac-aroL construct pGM450, under fully inducing conditions, is overexpressing shikimate kinase II some 5,000 fold. SDS PAGE analysis of the crude extract (Figure 5.21) reveals a major protein band at ca. 19,000 molecular weight comprising some 2-3% of the total soluble protein.

The crude extract was subjected to ion-exchange chromatography on DEAE-Sephacel and shikimate kinase eluted with an increasing (50-300 mM KCl) salt gradient. Fractions 84-100 were pooled (96 ml) and over 14,000 units of shikimate kinase recovered (Figure 5.18). The supernatant of a subsequent 30% Ammonium Sulphate fractionation (99 mls, 11,600 U) was further enriched for shikimate kinase by chromatography on Phenyl-Sepharose.

Shikimate kinase remained bound to the Phenyl-Sepharose column until the Ammonium Sulphate concentration was almost zero (Figure 5.19). Active fractions (92-108) were pooled (72.5 ml) and 9,200 units recovered.

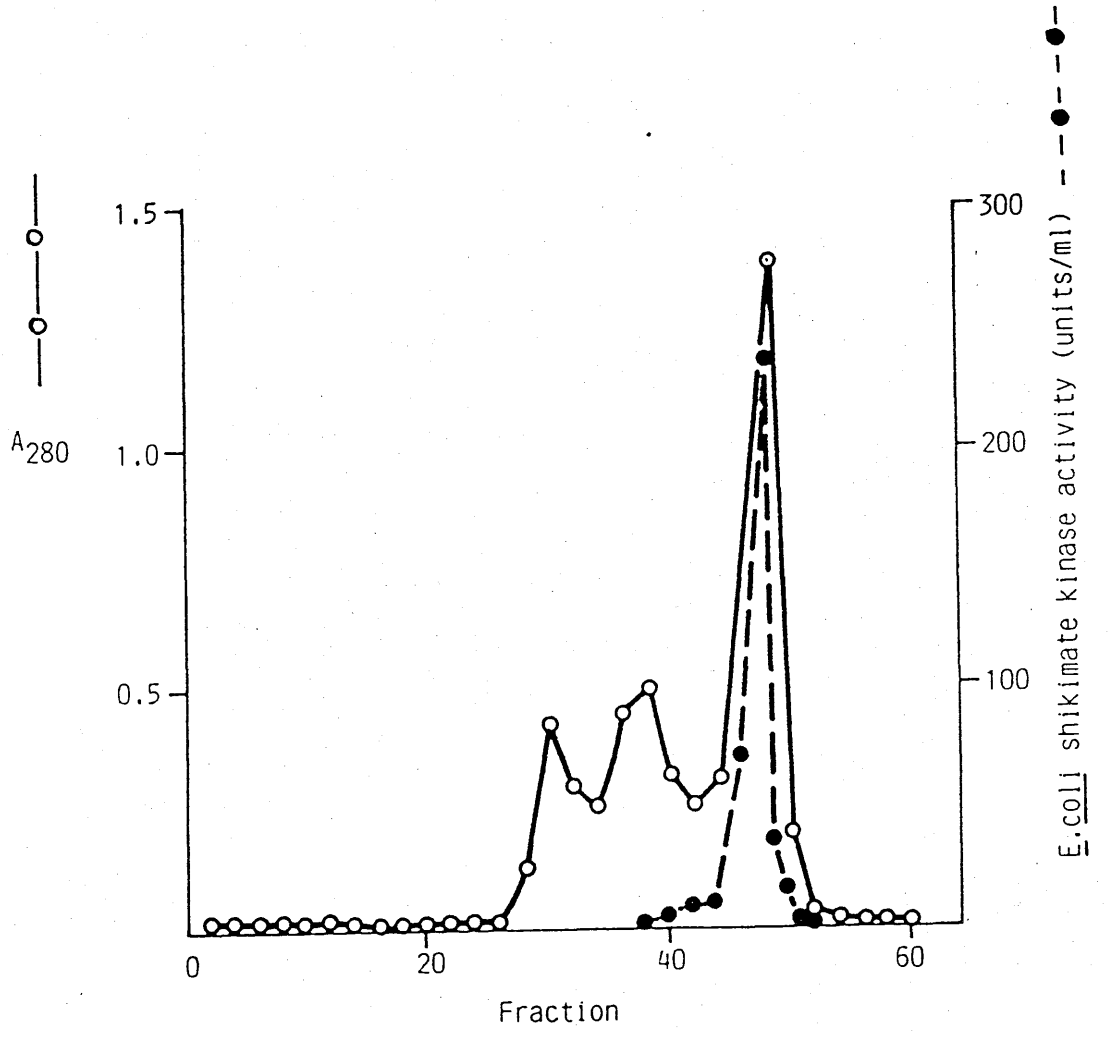
At this stage an abortive attempt to bind the enzyme to phosphocellulose resulted in the loss of some 2,000 units (22% loss) of shikimate kinase activity. The remainder of the active preparation was concentrated by vacuum-dialysis and dialysed into Buffer A containing glycerol (10%) before further purification by gel-filtration chromatography. The sample was split into two 10 ml aliquots and each separately fractionated on a Sephacryl S-200 gel filtration column (Figure 5.20(a) & (b)).

Figure 5.19: Phenyl-Sepharose Chromatography of E.coli shikimate kinase. Step 3 of purification outlined in Section 2.23.4 (Table 5.6B).



Fractions pooled from DEAE-Sephacel chromatography were treated with Ammonium Sulphate (Section 2.23.4) and 11,614 units in 99 ml of Buffer B loaded on to the column. Following an extensive Buffer B overnight wash (w), the column was treated with a linear gradient of 1.2M - 0M $(NH_4)_2SO_4$ in Buffer B (G). Flow rate 20 ml/hr, 10 ml fractions collected. Pooled fractions are indicated.

Figure 5.20a. Gel Filtration on Sephacryl-S200 of E.coli shikimate kinase. Step 4 of purification outlined in Section 2.23.4 (Table 5.6B).



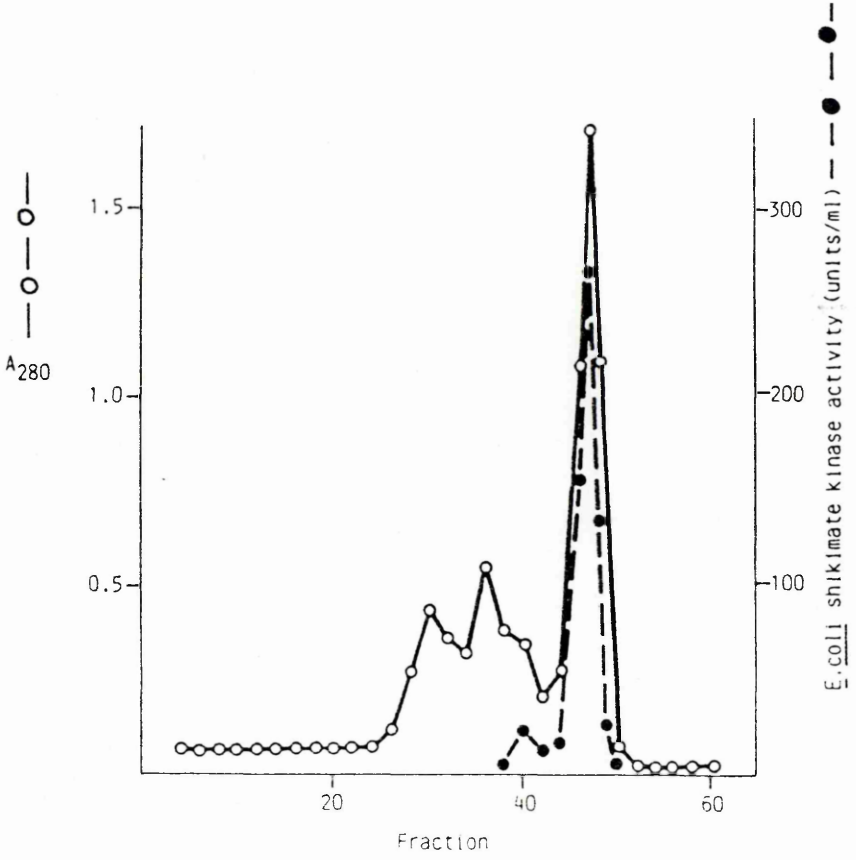
ca. 6,000 units of shikimate kinase activity was dialysed into Buffer A containing 10% (v/v) glycerol. This dialysed sample (20 ml) was divided into two aliquots and each loaded on to a S200 gel filtration column. Protein was eluted with Buffer A; flow rate 10 ml/hr and 4 ml fractions collected.

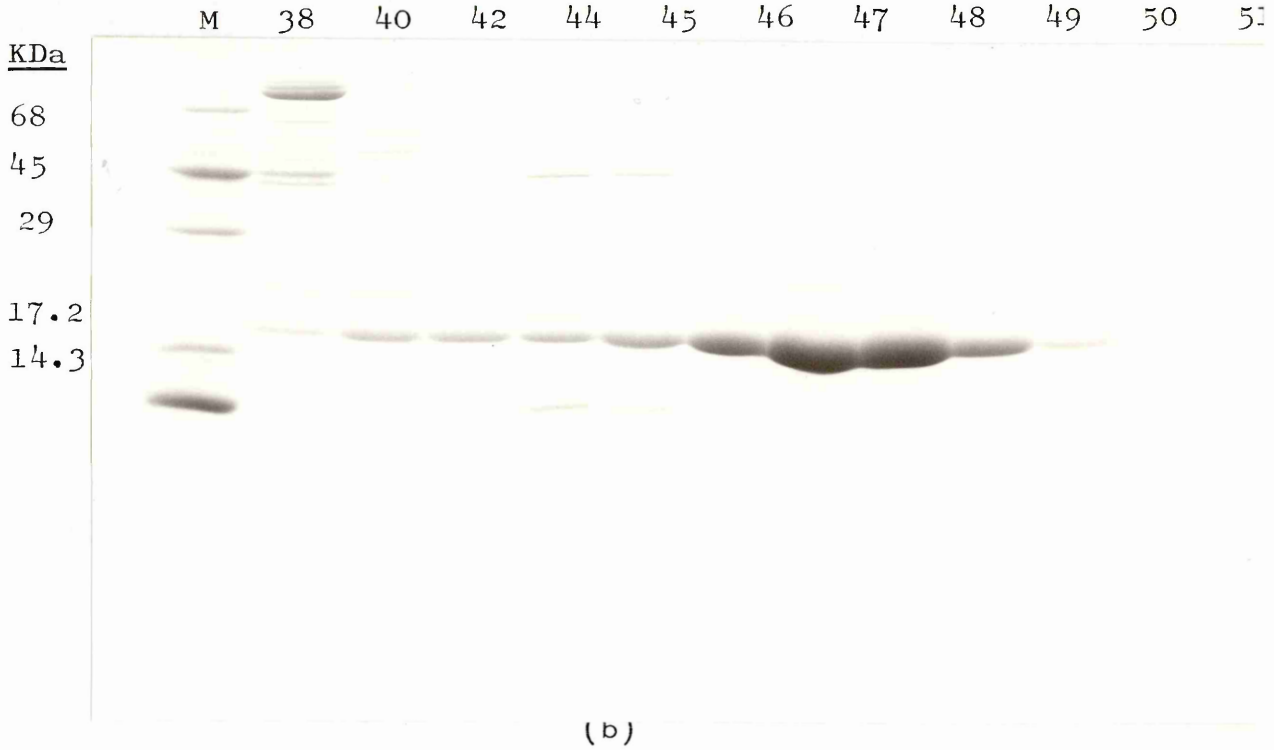
Figure 5.20b (facing).

(a) Sephacryl-S200 chromatography of the second aliquot of *E.coli* shikimate kinase preparation carried out as described in Figure 5.20a (Section 2.23.4; Table 5.6B).

(b) 12.5% SDS PAGE profile of the active fractions from the gel filtration column shown above. Protein is stained with Coomassie Blue. Sizes of marker proteins (M) are shown.

(a)


A280
1.5
1.0
0.5
300
200
100
E.coli shikimate kinase activity (units/ml)
20
40
60
Fraction


M
38
40
42
44
45
46
47
48
49
50
KDa
68
45
29
17.2
14.3

(b)

Fractions 46-48 of each gel filtration run (Figure 5.20(b)) were pooled and a combined total of 5,100 units recovered. The shikimate kinase II preparation was essentially homogeneous as judged by SDS PAGE (Figure 5.21). Further attempts to concentrate the enzyme sample by vacuum dialysis and mini-column ion exchange chromatography (5 ml DEAE-sephacel) resulted in heavy losses of activity (2,500 U lost). A final yield of 11% (2,645 Eu) was obtained overall representing about 50% of the pure enzyme preparation obtained in the gel-filtration pool. By omitting the abortive phosphocellulose and later DEAE-sephacel columns, this yield could easily be tripled. SDS PAGE analysis of the successive stages of the purification protocol is shown in Figure 5.21. Long term storage of the purified shikimate kinase II at -20°C in Buffer A (Section 2.23.4) containing 50% (v/v) glycerol resulted in a 60% loss of activity over a 6 month period.

A comparison between this purification scheme and that reported in Millar *et al.* (1986b) is shown in Table 5.6. Clearly this overexpressing strain (*E.coli* HW11/pGM450) could produce the hundred milligram quantities of pure shikimate kinase II which could be required for detailed n.m.r. or X-ray crystallographic structural analysis.

5.6.5 Shikimate kinase II is monomeric

The sub-unit M_r of *E.coli* shikimate kinase II was estimated to be 20,000 by SDS PAGE (Figure 5.21). The native M_r was estimated by gel filtration on a Superose 12

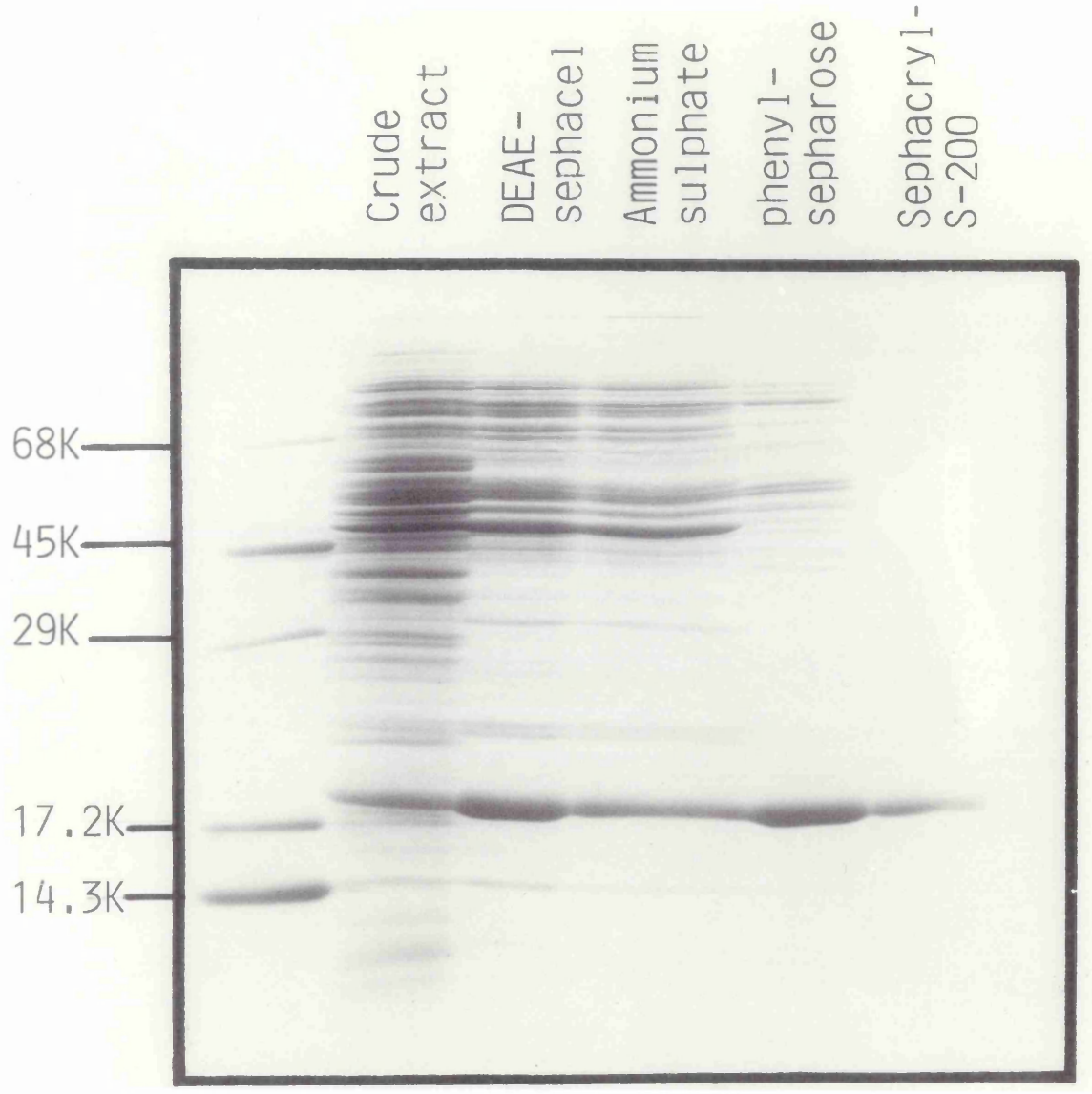
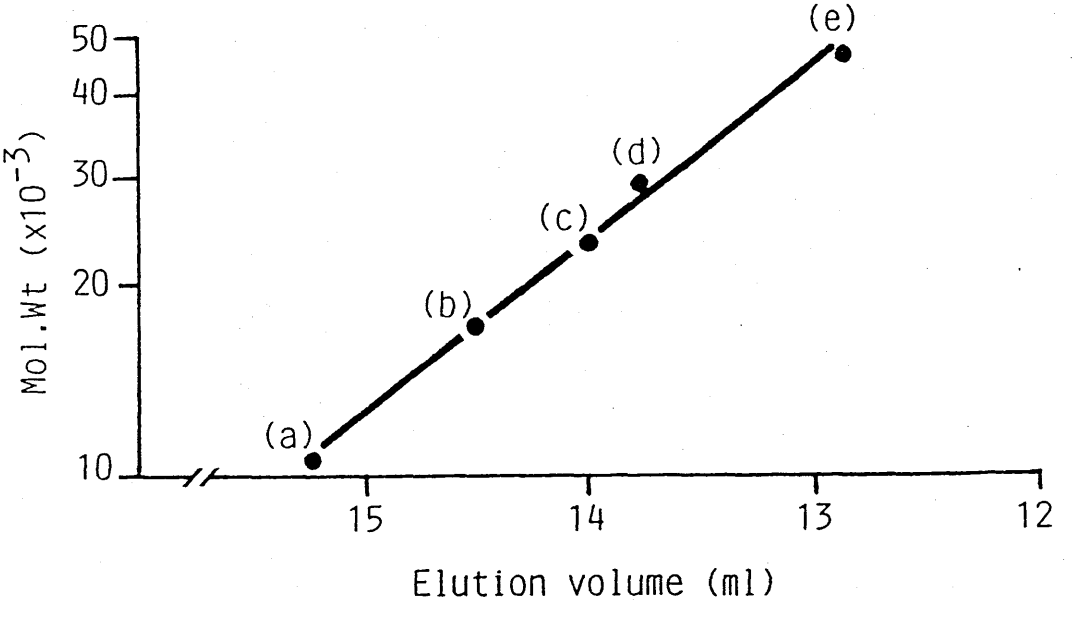


Figure 5.21: Coomassie stained 15% SDS PAGE of each step in the (pGM450) overexpressed shikimate kinase purification. Crude extract (125 µg); DEAE-Sephacel (75 µg); Ammonium Sulphate (75 µg); Phenyl-Sepharose (35 µg) and Sephacryl S-200 (10 µg) samples were analysed (Section 5.6.4). Protein size markers are shown in kDa (K).

Figure 5.22: Superose 12 FPLC separation of shikimate kinase. Purified shikimate kinase (10 units) was applied and chromatography carried out as described in Section 2.23.5. The elution volumes of the standard and unknown (shikimate kinase) proteins are shown below.



(a) horse heart cytochrome C M_r 12,800

(b) sperm whale myoglobin M_r 17,200

(c) E.coli shikimate kinase

(d) E.coli DHQ synthase M_r 38,900

(e) chicken ovalbumin M_r 45,000

Table 5.6(A & B) facing.

Comparison of purification schemes for *E.coli* shikimate kinase purified from (A) *E.coli* HW87/pMH423 (Millar *et al.*, 1986b) and (B) *E.coli* HW1111/pGM450 (Section 2.23.4). Identical assay procedures were performed in both cases.

A. The purification of shikimate kinase from E.coli HW87/pMH423

Step		Vol. (ml)	Protein (mg/ml)	Activity (units/ml)	Total activity (units)	Yield (%)	Specific activity (units/mg)	Purification (fold)
1	Crude extract	48	6.8	5.66	272	100	0.83	1
2	DEAE-Sephacel	47	0.7	5.99	281	103	8.56	10.3
3	Phenyl-Sepharose	50	0.14	3.80	190	70	27.1	31.6
4	Sephacryl S200	10.5	0.13	9.04	95	35	69.5	80.7

B. Purification of E.coli shikimate kinase from 17g of E.coli HW1111/pGM450

Stage	Volume (ml)	Activity (units/ml)	Total activity (units)	Protein (mg/ml)	Specific activity (units/mg)	Yield (%)	Purification (fold)
Crude extract	35	663.5	23,220	45	14.7	100	1
DEAE-sephacel	96	153	14,688	5.8	26.4	63	1.8
Ammonium sulphate	99	117	11,614	5.0	23.4	50	1.6
Phenyl sepharose	127	72.5	9,207	1.36	53.3	39	3.6
Sephacryl S-200	4.4	601.2	2,645	9.1	66.0	11	4.5

column as 20-22,000 (Figure 5.22). It would appear that shikimate kinase II is monomeric. Interestingly four of the five monofunctional E.coli activities which correspond to the arom multifunctional polypeptide, are monomeric. This is discussed in greater detail in Chapter Six.

5.7 Shikimate kinase II: homologies

5.7.1 Homologies with other kinases/ATPases

In Bacillus subtilis there is a single shikimate kinase, which is a component of a trifunctional multienzyme complex. This complex contains a bifunctional polypeptide carrying DAHP synthase and chorismate mutase activities as well as a monofunctional shikimate kinase polypeptide (Nakatsukara & Nester, 1972). The kinase component polypeptide, which is active only in the complex, has been purified to homogeneity and has a M_r of 10,000 (Huang et al., 1975). The amino acid composition of the B. subtilis shikimate kinase sub-unit is shown in Table 5.7.

The regulatory properties of the B.subtilis shikimate kinase suggest that it may be a key allosteric step in the shikimate pathway (Huang et al., 1975; Nasser et al., 1969). A direct comparison between the E.coli and B.subtilis shikimate kinase sub-unit amino acid compositions (Tables 5.5 and 5.7) is a dubious exercise because of the very different sub-unit molecular weights (19,000 and 10,000 respectively). In order to partially alleviate this problem only the first 87 N-terminal residues of E.coli shikimate kinase II (cq. 10,000 molecular weight) were compared with the B.subtilis amino

acid composition (Table 5.7). Such a comparison is still highly contentious but justifiable in as much as the B.subtilis composition is the only other prokaryotic shikimate kinase data available.

Not surprisingly the two compositions are very different but the choice of the N-terminus region of the E.coli sequence for comparison was not random. Evidence to be presented (Section 5.7.1) suggests that the adenine nucleotide binding domain of E.coli shikimate kinase II is near the N-terminus of the polypeptide. In this context therefore it is interesting to note the very different lysine and threonine contents between the E.coli composition and its B.subtilis counterpart (Table 5.7).

Walker et al. (1982) and Finch & Emmerson (1984) have identified two highly conserved regions (A and B) in kinases and other ATP/ADP-requiring enzymes that are believed to contribute to the nucleotide-binding fold. By a combination of n.m.r. and X-ray diffraction analysis Fry et al. (1986) have proposed that the MgATP binding site of adenylate kinase is located near three protein segments that are homologous in sequence in a number of enzymes that utilize MgATP. Adenylate kinase is the only protein among those sharing these extensive sequence homologies for which the binding site of the nucleotide substrate has been characterised. In phosphofructokinase (Evans et al., 1981) where an X-ray structure also exists, only one of the three homologous segments is found (segment B, after Walker et al., 1982).

Amino acid	Amino acid composition of the N-terminal 87 residues of *E.coli* shikimate kinase II[1]	Amino acid composition of *B.subtilis* shikimate kinase subunit[2]
Asx	3	8
Thr	10	1
Ser	3	nd
Glx	11	14
Pro	3	3
Gly	8	6
Ala	10	8
Cys	1	1
Val	6	10
Met	3	2
Ile	5	4
Leu	8	8
Tyr	0	2
Phe	4	3
His	0	2
Lys	1	8
Arg	6	3
Trp	2	1

[1]deduced from the nucleotide sequence reported here.

[2]Huang *et al.*, 1975.

Table 5.7: The amino acid composition of N-terminal 87 residues of *E.coli* shikimate kinase II compared with the amino acid composition of the *B.subtilis* shikimate kinase subunit.

Consensus sequences have been deduced for sequences A and B from several such proteins, and E.coli shikimate kinase II exhibits at least one of these structural motifs (Figure 5.23). Sequence A (Walker et al., 1982) is characterised by the consensus G-X_4-G-K-T-X_6-I/V; shikimate kinase contains the sequence G-P-R-G-C-G-K-T-T G-M-A-L-A at residues 8-22. The functional role for the sequence A in adenylate kinase (Fry et al., 1986) has been proposed as a glycine-rich flexible loop that is terminated by the conserved lysine residue and which interacts with the α-phosphoryl group of MgATP. Secondary-structure predictions, using the method of Chou & Fasman (1978), suggest that for E.coli shikimate kinase II this homologous A segment exhibits this loop structure containing the conserved lysine residue.

The occurrence of this binding fold close to the N-terminus is not unique to porcine adenylate kinase or E.coli shikimate kinase; in bacteriophage T_4 polynucleotide kinase an almost identical situation is observed (Midgley & Murray, 1985). The highly conserved nature of this 'GKT domain' (Segment A) in a number of ATP and GTP utilising enzymes suggests an equivalent mechanistic role for this conserved region in all these proteins (Figure 5.23).

The second conserved region (Segment B) in kinases exhibits the following consensus sequence R/K-X_3-G-X_3-L-(hydrophobic)$_4$ - followed by an aspartic acid residue (Walker et al., 1982). A homologous region in the E.coli shikimate kinase II sequence cannot be unambiguously identified although

Protein	Residues	Sequences
adenylate kinase	8-29	S K I I F V V **G** G P G S **G K G** T Q C E K I **V**
recA protein	59-80	G R I V E I Y **G** P E S S **G K T** T L T L Q V **I**
rho protein	172-193	G Q R G L I V **A** P P K A **G K T** M L L Q N I **A**
DNA B protein	223-244	S D L I I V A **A** R P S M **G K T** T F A M N L **V**
EF-Tu	11-32	H V N V G T I **G** H V D H **G K T** T L T A A I **T**
DNA helicase II	22-43	R S N L L V L **A** G A G S **G K T** R V L V H R **I**
shikimate kinase	1-22	T Q P L F L I **G** P R G C **G K T** T V G M A L **A**

Figure 5.23: Homologies between adenine-nucleotide binding proteins and E.coli shikimate kinase II, strongly conserved regions are boxed as discussed in the text. Note EF-Tu is a guanine-nucleotide binding protein. The source of the protein sequences is as detailed in Millar et al., 1986b.

one possible location is between residues 138 (arginine) and 152 (aspartic acid). As observed for phosphofructokinase (Segment B) and p21 GTPase (Segments A & B) it is very often the case that only a sub-set of the three homologous regions (Fry et al., 1986) are identifiable on primary structural data.

5.7.2 Homology with S.cerevisiae and A.nidulans arom sequences

Homology between the E.coli shikimate kinase II sequence and the S.cerevisiae (Duncan et al., 1986) and A.nidulans (Charles et al., 1986) arom sequences were examined using the BESTFIT program (Section 2.18.3). The overall homology between E.coli and S.cerevisiae or A.nidulans was relatively low (22% and 20% respectively) but notable in two respects.

Firstly the region of both fungal sequences identified as homologous with the bacterial shikimate kinase sequence was the same. Residues 872-1022 (Figure 5.24(b)) of the A.nidulans and residues 894-1040 (Figure 5.24(a)) of the S.cerevisiae sequences matched the E.coli sequence (starting in each case from residue 8).

Secondly the majority of identities were clustered in a region corresponding to the N-terminus of the prokaryotic sequence. Significantly, 17/32 and 22/30 identities, between E.coli/S.cerevisiae and E.coli/A.nidulans respectively, were located within the first 50 residues of the match. Closer examination of this sequence reveals it is the region identified in E.coli shikimate kinase II (Section 5.7.1; Figure 5.23) as the A segment involved in nucleotide binding (GKT domain). All of the conserved residues of the consensus

(A) sequence G-X_4-G-K-T-X_6-I/V are retained in both S.cerevisiae and A.nidulans arom polypeptides. The single exception to this is the conservative substitution of serine for threonine in the A.nidulans sequence (Figure 5.24(b)).

A BESTFIT comparison between these two fungal sequences, within this proposed shikimate kinase domain, shows extensive homology (44% identity; Figure 5.24(c)). An examination of other highly conserved regions between the two fungal proteins failed to identify any of the remaining proposed segments associated with MgATP binding (Walker et al., 1982; Fry et al., 1986).

5.7.3 The aroL gene: relationship to shikimate kinase I

Gel filtration of partially purified E.coli shikimate kinase I and II isoenzymes yields a single peak of activity at ca. 20,000 molecular weight (Ely & Pittard, 1979). Only the aroL encoded kinase isoenzyme is regulated by the pleiotropic effector tyrR. The possibility exists that a gene duplication, followed by subsequent divergence to produce the constitutive shikimate kinase gene was responsible for the dual activity in E.coli.

E.coli high molecular weight DNA (5 µg) was digested with BamHI or BamHI and PvuII and (following agarose gel electrophoresis) blot transferred to nitrocellulose paper (Section 2.20.1). A nick-translated (Section 2.21) 2.7 kbp BamHI insert of pGM424 (specific activity 2.3 x 10^8 dpm/µg) was used to probe the immobilised chromosomal DNA. Hybridisation

Figure 5.24 (facing).

BESTFIT output of:

(a) E.coli vs. S.cerevisiae aroL homology.

(b) E.coli vs. A.nidulans aroL homology.

(c) S.cerevisiae vs. A.nidulans aroL homology.

The numbering in each case refers to the peptide sequences of the respective proteins.

E.coli vs S.cerevisiae aroL Homology

(a)

```
  8 IGPRGCGKTTVGMALADSLNRRFVDTD.QWLQSQLNMTVAEIVEREEWAG 56
    ** *  ****     *  *    ** *    *   *  *   *    *
894 IGMRAAGKTTISKWCASALGYKLVDLDELFEQQHNNQSVKQFVVENGWEK 943

 57 FRARETAALEAVTAPSTVIATGGGIILTEFNRHFMQNNGIVVYLCAPVSV 106
    **  **     *
944 FREEETRIFKEVIQNYGDDGYVFSTGGGIVESAESRKALKDFASSGGYVL 993

107 LVNRLQAAPEEDLRPTLTGKPLSEEVQEVLEERDALYREVAHIIIDA 153
       *        *          **  **   *   * *       *
994 HLHRDIEETIVFLQSDPSRPAYVEEIREVWNRREGWYKECSNFSFFA 1040
```

E.coli vs A.nidulans aroL Homology

(b)

```
  8 IGPRGCGKTTVGMALADSLNRRFVDTDQWLQSQLNMTVAEIVEREEWAGF 57
    ** ** ** * *      *** *** *  *     **   *     * **
872 IGMRGAGKSTAGNWVSKALNRPFVDLDTELETVEGMTIPDIIKTRGWQGF 921

 58 RARETAALEAVTAPSTVIATGGGIILTEFNRHFMQNNGIVVYLCAPVSVL 107
    *  *   *                                       *  *
922 RNAELEILKRTLKERSRGYVFACGGGVVEMPEARKLLTDYHKTKGNVLLL 971

108 VNRLQAAPEEDLRPTLTGKPLSEEVQEVLEERDALYREVAHIIIDATNEP 157
                                    *                 *
972 MRDIKKIMDFLSIDKSRPAYVEDMMGVWLRRKPWFQECSNIQYYSRDASP 1021

158 S 158
    *
1022 S 1022
```

S.cerevisiae vs A.nidulans aroL Homology

(c)

```
893 IIGMRAAGKTTISKWCASALGYKLVDLDELFEQQHNNQSVKQFVVENGWE 942
    ***** *** *   *   **    ****   *                **
871 IIGMRGAGKSTAGNWVSKALNRPFVDLDTELETVEGM.TIPDIIKTRGWQ 919

943 KFREEETRIFKEVIQNYGDDGYVFSTGGGIVESAESRKALKDFASSGGYV 992
     **  *  * *          ****  *** **  * ** * *     * *
920 GFRNAELEILKRTLKERSR.GYVFACGGGVVEMPEARKLLTDYHKTKGNV 968

993 LHLHRDIEETIVFLQSDPSRPAYVEEIREVWNRREGWYKECSNFSFFAPH 1042
    * * ***     **  * *******    ** **  *  ****
969 LLLMRDIKKIMDFLSIDKSRPAYVEDMMGVWLRRKPWFQECSNIQYYSRD 1018

1043 CS 1044
     *
1019 AS 1020
```

and washing conditions were at low stringency (Section 2.20.5) designed to detect divergents of 25-30% mismatch. At this level of stringency authentic hybrids would not be masked by totally aspecific hybrids.

A short time exposure (30 minutes, Figure 5.25(a)) resolved the expected pattern of hybridisation. As previously seen (Figure 5.5), the chromosomal organisation of the BamHI (2.7 kbp) region of DNA encoded aroL is preserved on plasmid pGM424. A 3 hour exposure of the same filter (Figure 5.25(b)) failed to detect any other strongly hybridising bands. This time limit of exposure (3 hours) appeared to be the threshold beyond which high background smearing (see track M, Figure 5.25(b)) made interpretation unreliable.

It is possible that the shikimate kinase I gene is the product of a gene duplication of aroL (or vice versa) which retained the gross overall structure (2.7 kbp BamHI region with internal PvuII sites). Subsequent mutations within this region, but without affecting the restriction sites, could produce a structurally distinct protein and account for the separability of the two isoenzymes on ion-exchange chromatography (Ely & Pittard, 1979). Similarly, mutations in the operator sequences could account for the differential modes of regulation of the two genes. If this hypothesis were true, the second shikimate kinase gene (shikimate kinase I) could also be cloned on a 2.7 kbp BamHI fragment. In an aroF$^-$ aroG$^-$ background, mutations in aroL dramatically affect growth on unsupplemented (lacking aromatic end products)

medium (Defeyter & Pittard, 1986). Growth characteristics on this medium can therefore be used to screen for the AroL$^+$ (normal growth) and aroL$^-$ (poor growth) phenotypes. Defeyter & Pittard (1986) have reported only isolating aroL-specific 2.7 kbp BamHI recombinant E.coli clones capable of conferring elevated growth rates on a aroG aroF aroL strain. Unless there is a selection against non-aroL shikimate kinase genes under such 'complementation' conditions, it is difficult to avoid the conclusion that the shikimate kinase I gene is not a gene duplication of aroL. Such a hypothesis will only be confirmed once the kinetic, mechanistic and structural properties of both isoenzymes are elucidated and their respective gene sequences compared.

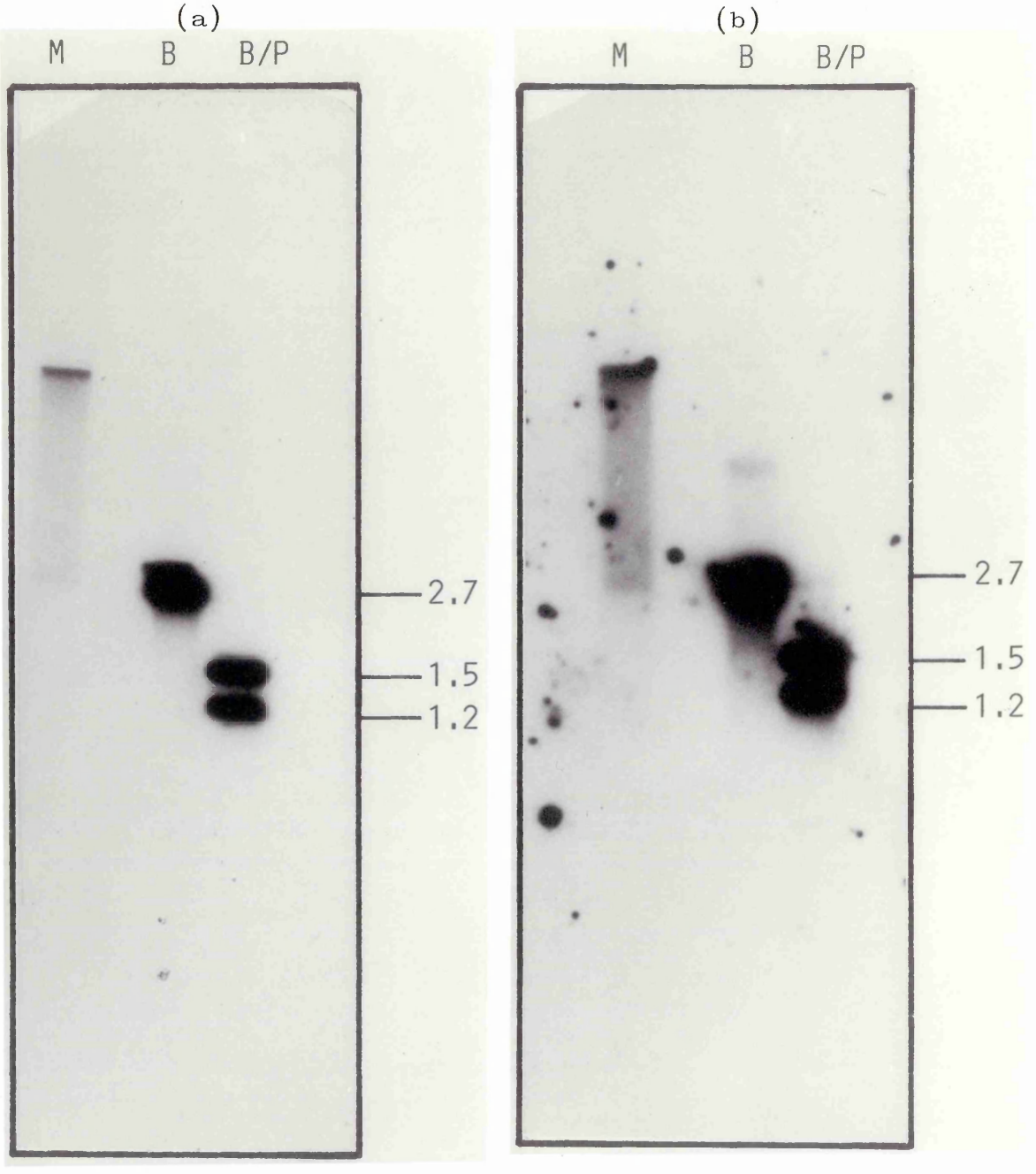


Figure 5.25: Southern hybridisation (low stringency) of the E.coli aroL gene. DNA was digested with BamHI (B) or BamHI and PvuII (B/P). Size marker DNA (M) was probed also. Both 30 minute (a) and 3 hour (b) exposures are shown. Molecular sizes are shown in kbp. The hybridisation pattern is discussed in Section 5.7.3.

CHAPTER 6

THE AROM COMPLEX: A MODEL FOR THE EVOLUTION OF MULTI-FUNCTIONAL PROTEINS - A DISCUSSION

6.1 Introduction

6.1.1 An overview

A unifying theme and one of the primary aims of this thesis, as outlined in Chapter One, was to use the sequences of the *E.coli* *aroL* and *aroB* genes as analytical tools. The elucidation of these two gene sequences completed the required set of five bacterial shikimate (*arom*) pathway gene sequences needed for comparative purposes, the others being *aroA* (Duncan *et al.*, 1984b), *aroD* (Duncan *et al.*, 1986b) and *aroE* (Anton & Coggins, 1986). The complete primary structure of each of these five central, monofunctional *E.coli* shikimate pathway enzymes had therefore been established in this laboratory. The availability of the *A.nidulans* (Charles *et al.*, 1986) and *S.cerevisiae* (Duncan *et al.*, 1986a) gene sequences (*AROM* and *ARO1* respectively), which encode the pentafunctional *arom* complexes in these fungi, meant that possible relationships between the monofunctional prokaryotic activities and their corresponding multifunctional eukaryotic counterparts could be directly investigated. In this respect, the *E.coli* 3-dehydroquinate synthase (*aroB*) and shikimate kinase II (*aroL*) amino acid sequences became more than passive sequence data.

Although lacking a corresponding *multifunctional* activity, the determination of the *E.coli* chorismate synthase (*aroC*) sequence was equally as important. The availability of this sequence allowed detailed examination of the *E.coli* shikimate pathway for characteristics indicative of a retroevolutionary

origin. This particular aspect was discussed in Chapter Four but will be reaffirmed as part of an overall assessment.

The structures of the preceding chapters concerning aroB (Chapter 3), aroC (Chapter 4) and aroL (Chapter 5) were conceived in order that they should "dovetail" into this final discussion. It is the sincere hope of this author that the reader finds this to be the case. In complementing the preceding results chapters, this discussion expands upon the more wide ranging implications of the homologies observed between the E.coli aro genes and their fungal homologues. Fuelled by the substantial corroborative evidence, see below, speculation is permitted where perhaps it was restrained before. Hypotheses regarding possible evolutionary relationships between the monofunctional and multifunctional systems will be expanded by considering the mode of expression of the E.coli aro genes (and possible constraints this may put on our thinking on evolutionary routes). In essence this chapter stands not apart from the results presented previously (Chapters 3-5), but rather exists as an equally integral component, to be considered in assessing the significance of the results.

6.1.2 Expansion of homology discussions - rationale

Any discussion regarding the level of relatedness and respective origins of the mono- and multifunctional shikimate pathway activities must satisfy at least one fundamental criterion. It is vital to establish that the activities being compared are related, as a prerequisite of evolutionary

speculation. This seemingly simplistic statement does in fact form the basis of valid discrimination between divergent and convergent routes to any given peptide sequence (and structure). It is necessary to go further than to simply highlight limited sequence homology between any two (or three) sequences, in presenting a balanced assessment of their respective levels of similarity. Corroborative evidence, both factual and circumstantial, must be considered to validate any presupposed identity. In this respect the plethora of genetic, biochemical and kinetic data (some already outlined in Chapter One), in addition to the striking sequence homologies to be discussed, exists to substantiate the case in point. Having achieved this necessary qualification, discussion of the alternative hypotheses for the evolutionary origin of the different pathway organisations can follow.

6.2 Shikimate pathway sequence comparisons

6.2.1 Bacterial: fungal homologies

BESTFIT comparisons of the E.coli DHQ synthase (aroB) and shikimate kinase II (aroL) sequences, in Chapters 3 and 5 respectively, had established that each was homologous with a discrete region of the S.cerevisiae and A.nidulans arom polypeptides. Furthermore the sequences of both fungal multifunctional proteins in these DHQ synthase and shikimate kinase 'domains' were highly conserved (Figures 3.30 and 5.24(c)). The working model of the arom multifunctional enzyme is that of a mosaic structure of five independently folded domains. This is based on the genetic (Giles et al.,

1967a; Rines et al., 1969) and biochemical (Boocock, 1983; Smith & Coggins, 1983; Coggins & Boocock, 1986) data discussed in Chapter One. This hypothesis is supported therefore by the identification of discrete DHQ synthase and shikimate kinase domains and can be substantiated by similarly examining aroA, aroD and aroE homology with the fungal sequences. Such evidence is shown in Figures 6.1 - 6.3.

Comparison of each of the monofunctional E.coli enzymes with the S.cerevisiae and A.nidulans arom proteins established, by primary homology criteria, that five sub-regions of the multifunctional enzyme exist (Figure 6.1). Each of the bacterial sequences is homologous to a discrete, non-overlapping segment of the fungal protein (Duncan et al., 1986a).

The order of bacterial homologies along the fungal polypeptide is DHQ synthase (aroB): EPSP synthase (aroA): shikimate kinase (aroL): dehydroquinase (aroD): shikimate dehydrogenase (aroE). This order is the same as that predicted for N.crassa arom (Giles et al., 1967a) and A.nidulans (Roberts, 1969) but is inconsistent with the incomplete genetic mapping data of de Leeuw (1967) for the S.cerevisiae ARO1 locus.

Examination of the homologies (Figure 6.1) highlights some interesting features. Within the DHQ synthase domain there are two sub-domains comprising two large regions of very highly conserved amino acid residues. These sub-domains are separated by a small, less well conserved region which requires, for optimal alignment with the yeast sequence, the insertion of a gap in both E.coli and A.nidulans sequences.

The lack of significant homology within this region (between the two fungal sequences) and its absence altogether in E.coli suggests that it plays a non-catalytic, structural role perhaps in connecting the two sub-domains. Greer & Perham (1986) have noticed a similar feature when aligning the two internal sub-domains of the E.coli and human glutathione reductase sequences.

Within the first of the two sub-domains is the proposed DHQ synthase nucleotide binding site. The conserved residues within the E.coli DHQ synthase sequence (residues 95-116) implicated in NAD^+ binding (Figure 3.31) are also highly conserved in both fungal sequences (Figure 6.1; S.cerevisiae residues 105-113). Overall the homology between the two fungal sequences is higher than either pairwise comparison of fungal:bacterial sequences. The level of direct identity between the E.coli and S.cerevisiae DHQ synthase domains at 36% (Duncan et al., 1986a), is still significant.

The five domains within the arom enzyme are separated by four linker regions, the first of which occurs between the DHQ synthase and EPSP synthase domains. In this connector, an eleven residue peptide sequence between the last residue of the aroB-encoded sequence and the initiation methionine of the aroA-encoded sequence, still retains some fungal: fungal homology (Figure 6.1; Figure 6.3(a)).

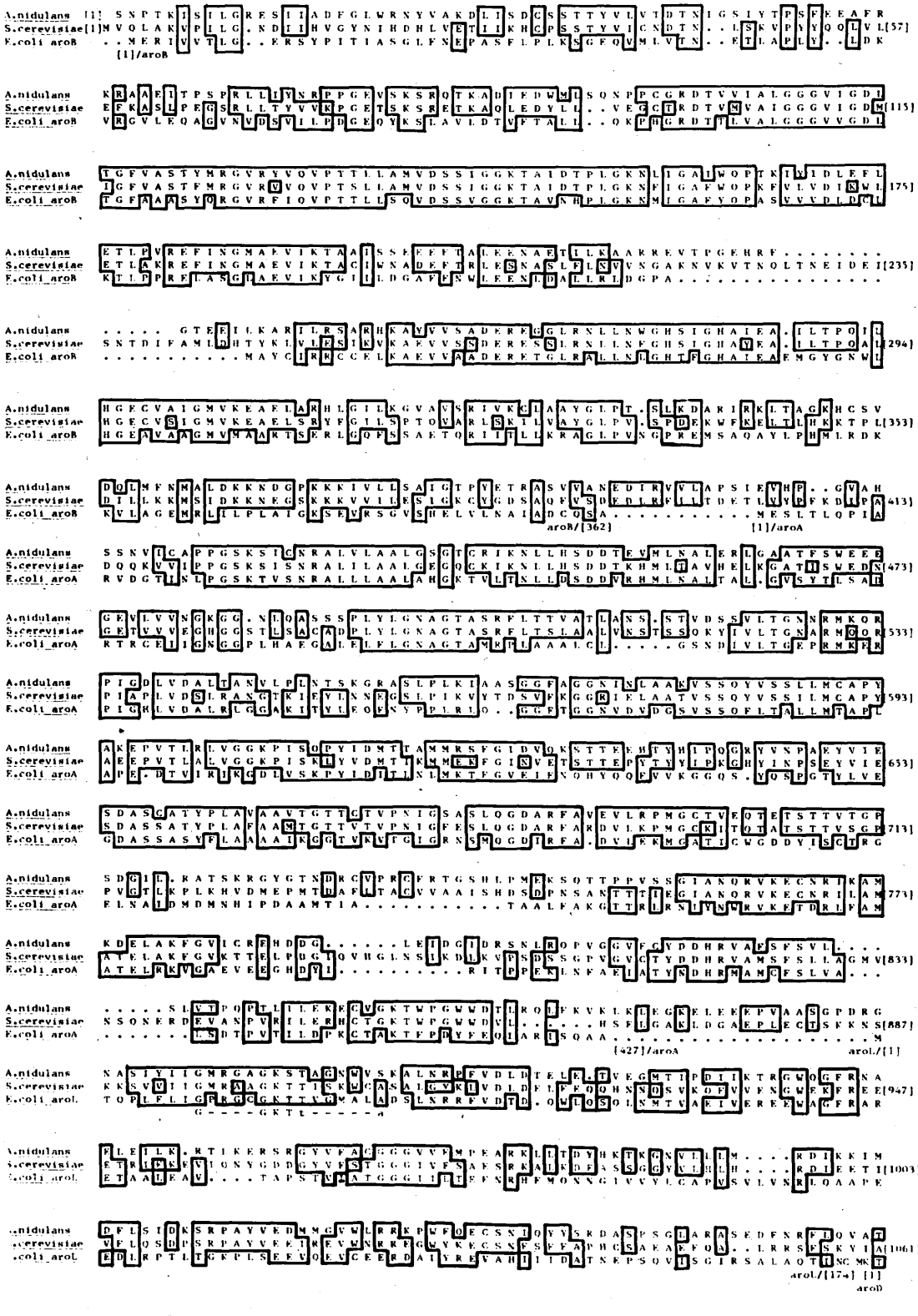
The EPSP synthase domain of the S.cerevisiae arom enzyme is 38% homologous (by identity) with the E.coli sequence (Duncan et al., 1986a). Large areas of extended homology exist throughout the sequence although identification of

sub-domains is less easy. The connector region between the aroA and aroL (shikimate kinase) domains consists of a 20 amino acid residue region which also retains some inter-fungal homologies (Figure 6.1; Figure 6.3(b)).

The remaining three domains, shikimate kinase, dehydroquinase and shikimate dehydrogenase (aroL, aroD and aroE; Figure 6.1) all exhibit lower levels (23-25%) of bacterial: fungal homologies although interfungal homology (identity) remains significant (Duncan et al., 1986a). Similarly the connector regions, particularly the aroL/aroD boundary, are less well defined (Figure 6.1; Figure 6.3).

The nucleotide sequences of the aroB gene 3' flanking sequences and the aroA gene 5' flanking sequences (Duncan et al., 1984b) were examined and compared with the corresponding connector region (aroB/aroA) in the S.cerevisiae nucleotide sequence (Duncan et al., 1986a). Similarly the nucleotide sequences around the other S.cerevisiae connectors were compared with the corresponding bacterial flanking sequences. Zalkin & Yanofsky (1982) have demonstrated using a similar approach that the yeast tryptophan synthase bifunctional enzyme has arisen by fusion of E.coli-like genes. Specifically, nucleotide sequence comparison of a portion of the yeast TRP5 connector region with the 3' end of the E.coli trpC and trpC-trpB intercistronic region demonstrates the mutational events which presumably occurred during or after gene fusion.

It is not unreasonable to suggest that examination of the connector regions within the arom enzyme may provide a clue to identify the DNA sequence(s) associated with the


A.nidulans
S.cerevisiae
E.coli aroB
[1]/aroB
aroB/[362]
[1]/aroA
[427]/aroA
aroL/[1]
aroL/[174]
aroD

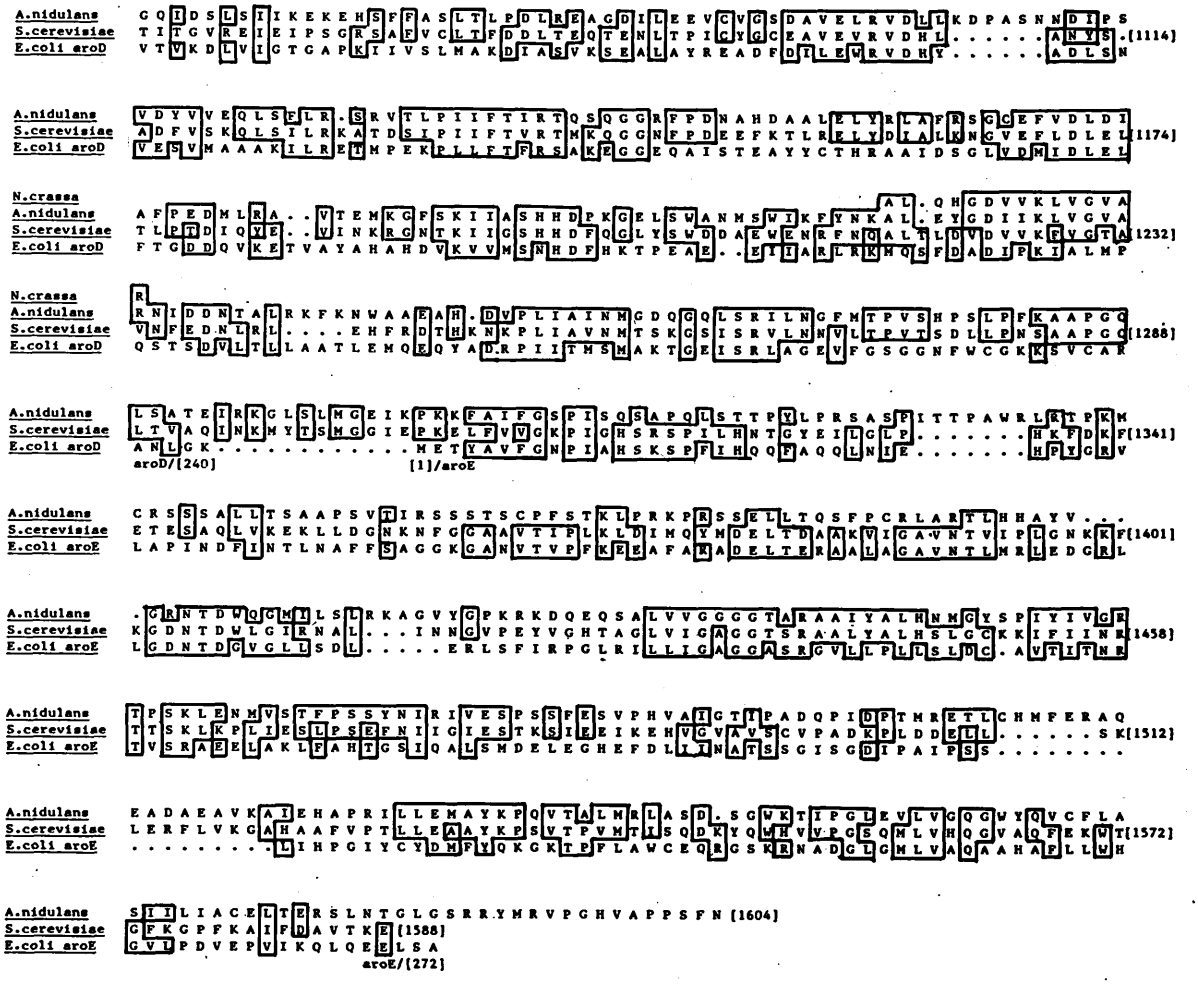


Figure 6.1: Homologies between the A.nidulans and S.cerevisiae arom sequences and E.coli aroB, aroA, aroL, aroD and aroE encoded shikimate pathway enzymes. The gapping introduced to maximise homology is adapted from Duncan et al. (1986a). Direct identities and conservative changes (I/L/V; D/E; S/T; F/W/Y & R/K) are boxed. The kinase ATP-binding site consensus sequence and the N.crassa dehydroquinase active site peptide sequence are indicated.

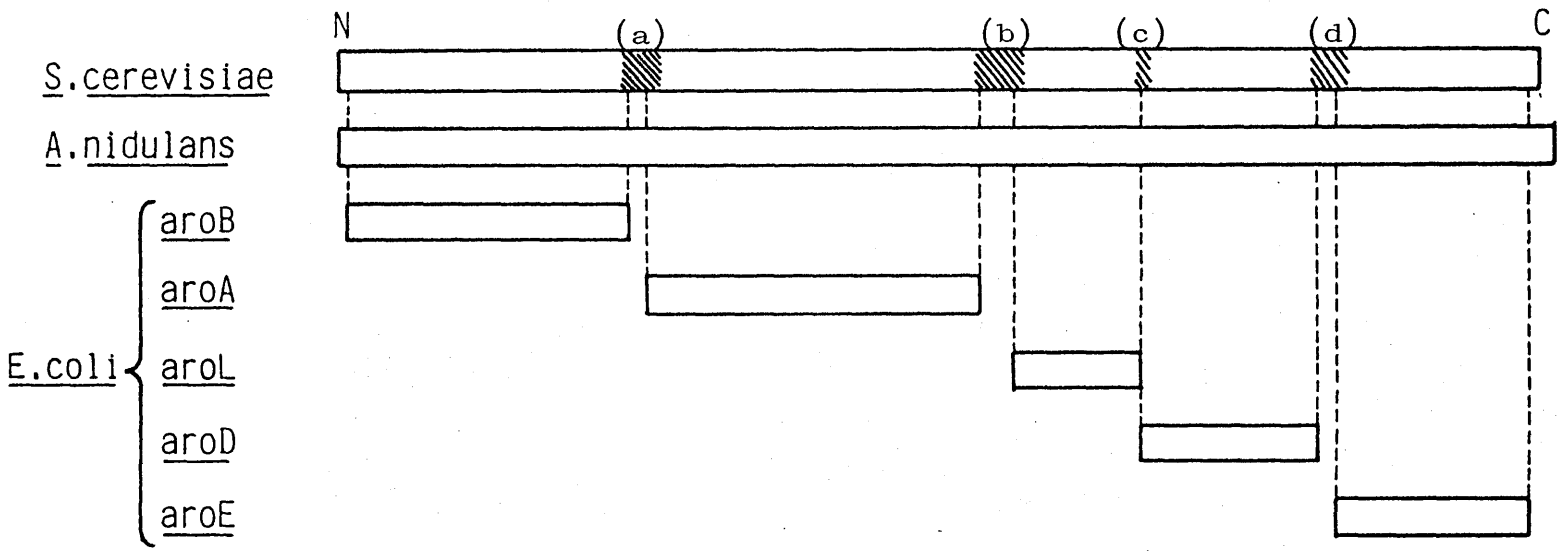


Figure 6.2: Diagramatic representation of the homologies shown in Figure 6.1, between the S.cerevisiae and A.nidulans arom sequences and the monofunctional E.coli enzymes. The connector regions of the yeast enzyme (see Figure 6.3(a)-(d)) are shaded.

genetic events which led to gene fusion/scission. Unfortunately there is no discernable nucleotide homology between the bacterial and fungal sequences. This of course does not exclude the possibility that these fungal sequences could be derived from transposable element(s) which may have been responsible for the transposition of the bacterial genes to (or from) the single fungal locus. Alternatively the nucleotide (and therefore peptide) sequence(s) at the domain junctions could be catalytically, and to some extent structurally, inert. As such, regardless of the evolutionary direction (fusion or scission), these sequences may be able to mutate freely. It is perhaps therefore not surprising that there is no visible homology between the nucleotide sequences of the connectors shown in Figure 6.3 and the corresponding bacterial DNA flanking sequences.

A closer examination of the four proposed linked regions is shown in Figure 6.3. For clarity each linker is defined as the region of arom sequence containing the inter-domain boundary and located between the next immediately adjacent fully conserved residues (in all three sequences). Since the connectors do not seem to be the remnants of ancestral nucleotide sequences (bacterial) involved in genetic transposition then a specific non-catalytic role must be assumed. However, only the aroD/aroE connector (Figure 6.3(d)) has fungal connectors of identical size. Also all four connectors show only a partial degree of interfungal and bacterial: fungal homology (Figure 6.1). Zalkin et al. (1984) have noted a similar situation in comparing the connector regions

of the bifunctional TRP3 and trp-1 gene products of S.cerevisiae and N.crassa respectively. They have proposed that the primary structure of the connector, and to a certain degree the length of the connector, are of secondary importance and may not be conserved in homologous multi-functional polypeptides (Zalkin et al., 1984).

It seems likely that the connector regions within both fungal arom polypeptides represent 'stuffer' regions which enjoy a certain degree of freedom both in size and composition. The observed conservation of some residues in the aroB/aroA, aroA/aroL and aroD/aroE connectors (Figure 6.3) may suggest that these are essential residues for linker function or more likely that these (the multifunctional enzymes) are the more ancient form of structural organisation. This is based on the assumption that the A.nidulans and S.cerevisiae are not derived from each other but rather from a common fungal ancestor. In this respect the connector regions may provide a valuable piece of evidence in favour of gene scission.

Alternatively the absence of nucleotide homology between the bacterial flanking sequences and the fungal connector DNA sequences could indicate a very ancient fusion event. The subsequent divergence of the non-essential connector DNA sequences could now be manifest in only limited fungal connector primary sequence homology and the observed absence of bacterial flanking sequence motifs.

Similarly although the aroB and aroA homologies (Figure 6.1) are high and connectors therefore well defined, the same is not true for the aroL, aroD and aroE homologies. The

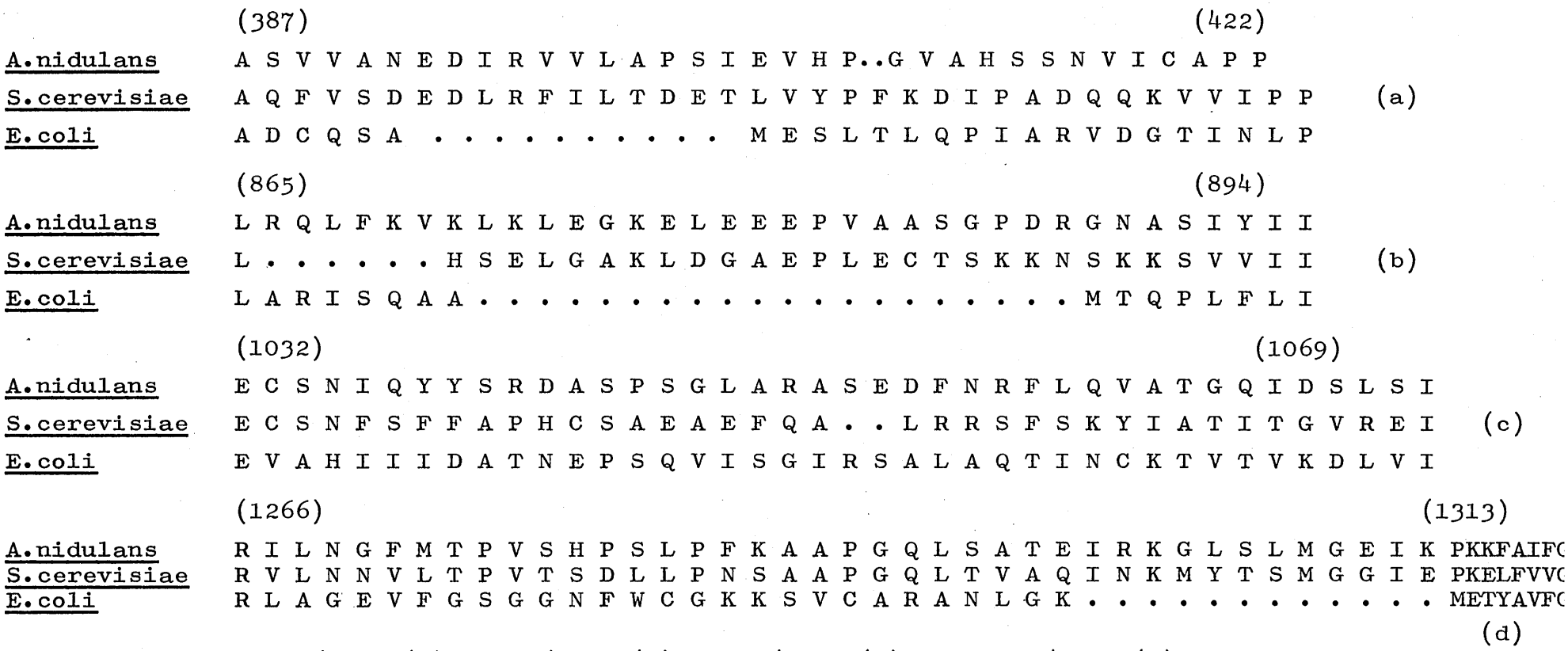


Figure 6.3: The aroB/aroA (a); aroA/aroL (b); aroL/aroD (c) and aroD/aroE (d) connector regions of A. nidulans and S.cerevisiae arom polypeptides. All numbering is w.r.t. the S.cerevisiae sequence (see Figure 6.1).

location of the connector sequences in these cases is therefore somewhat subjective.

The homology between the two fungal enzymes is consistently high throughout the polypeptide sequences (Figure 6.1). The level of identity between the bacterial enzymes and the five subregions of the fungal proteins varies from ca. 35-40% for DHQ synthase (*aroB*) and EPSP synthase (*aroA*), to as low as 20% for the dehydroquinase (*aroD*) activity. At this decreased degree of similarity, despite the fact that the BESTFIT homology program uniquely identifies these dehydroquinase sequences as related, it could be argued that such "homology" is fortuitous. In such a case, the possibility of independent (functionally convergent) evolution cannot be ruled out.

6.2.2 Corroborative evidence

Duncan *et al.* (1986a) have shown that the individual, catalytically active domains of the *S.cerevisiae arom* polypeptide can be defined both genetically and biochemically. Deletion and insertion analyses of the cloned *S.cerevisiae* ARO1 gene has produced a set of subclones specifying altered yeast *arom* polypeptides. The ability of the yeast ARO1 gene to complement mutants in four *E.coli aro* genes is affected by either deletion at, or insertion into various restriction sites within the ARO1 gene (Duncan *et al.*, 1986a).

This work is conceptually analogous to the limited proteolysis of the *N.crassa arom* multifunctional enzyme described in Chapter One. In both cases truncated (or altered) polypeptides are produced which still retain the ability to

catalyse one or more of the multifunctional enzyme activities. When considered with the chemical modification experiments of Smith & Coggins (1983), which established that the E2 and E3 active sites were spatially distinct in the N.crassa arom enzyme, this work extends the identification of individual catalytic domains on the arom polypeptide to four of the five enzymatic functions.

Although the level of homology between the dehydroquinase domain of the S.cerevisiae/A.nidulans arom sequences (Duncan et al., 1986a; Charles et al., 1986) and the E.coli aroD-encoded enzyme is low (Figure 6.1), one striking feature does exist. The sequence around the active site lysine residue of the dehydroquinase domain of the N.crassa arom enzyme has been determined on a pentadecapeptide (Chaudhuri & Coggins, 1986). A comparison with the two fungal arom sequences and the E.coli dehydroquinase sequence identifies a similar sequence in all three (Figure 6.1). This sequence occurs in a region of the S.cerevisiae arom enzyme identified by homology, and confirmed by deletion and complementation criterion (Duncan et al., 1986a), as being the dehydroquinase domain.

Similarly the region of the S.cerevisiae and A.nidulans arom sequences identified as being homologous to the aroL-encoded shikimate kinase (Chapter Five) contains a conserved feature. The ATP-binding site sequences (Walker et al., 1982; see Chapter Five) recognised within the E.coli shikimate kinase peptide sequence, are strongly conserved in both fungal enzymes (Figure 6.1). The unambiguous identification of such a 'kinase-specific' sequence completes the assignation

of all five *arom* domains. Each of the domains of the *S.cerevisiae* *A.nidulans* or *N.crassa* *arom* multifunctional enzymes can be identified by sequence homology with the *E.coli* monofunctional enzymes, and/or by their containing a unique feature allowing biochemical or genetic discrimination.

It is perhaps prudent at this stage to step back and to reassess the question of relatedness. The catalytic activities of the monofunctional *E.coli* enzymes and the multifunctional *arom* enzymes of fungi are identical. Their substrates, co-factors and sensitivity to inhibitors are equally alike. It is perhaps therefore not surprising that an element of sequence homology should exist in comparing the bacterial and fungal polypeptides. It could be argued that the identity observed (Figure 6.1) represents nothing more than functional convergence towards a peptide sequence specifying, in each case, the necessary three dimensional structure required for efficient catalysis. Whilst a detailed comparative kinetic study of the monofunctional and multifunctional activities is lacking, some evidence tends to suggest that differences do exist. Specifically, Duncan *et al.* (1984a) showed that the *N.crassa* and *E.coli* EPSP synthase activities show a 5-10 fold difference in their binding constants for substrates and product. Similarly the DHQ synthase activity of *N.crassa* is Zn^{2+} dependent (Lambert *et al.*, 1985) whereas the *exact* cation requirement for the *E.coli* enzyme is as yet unascertained. These small, though not insignificant, differences could betray *parallel* evolutionary routes for the bacterial and fungal enzymes.

This arguement is embarassed not only by the lack of experimental data, but also by the facts existing to the contrary. The levels of direct identity for the DHQ synthase and EPSP synthase domains are so high (Figure 6.1) as to represent indisputably related sequences. Also within the regions of low homology, the presence of conserved features, as discussed above, argues for divergence outwith essential structural regions as being the real source of dissimilarity. This argument can be extended to enzymes whose primary structures are largely different due to such extensive divergence but who retain a 'core' structural element specifying the necessary catalytic residues. Such an example is the serine proteases which exhibit differences in substrate binding (and primary structure) but retain a fundamental catalytic machinery indicative of a common ancestral origin. A lack of primary structure homology therefore is not always what it may seem. Doolittle (1981) has pointed out that resemblances between seemingly unrelated proteins is more definitive when based on conserved tertiary structural features rather than primary sequence homologies alone.

The most suggestive evidence for the hypotheses that the fungal and bacterial arom activities are derived from a common ancestor, regardless of the evolutionary direction, is their size. As can be seen in Table 6.1, the sum molecular weight of the five E.coli common pathway activities corresponding to the fungal arom protein, is approximately 160 kDa. The N.crassa arom is 165 kDa (Lumsden & Coggins, 1977) whilst the

S.cerevisiae and A.nidulans enzymes are both 175 kDa (Duncan et al., 1986a; Charles et al., 1986). It seems extremely unlikely that sequence convergence would have produced five separate peptide sequences whose sum molecular weight differed by less than 5% from that of an independently evolving, but catalytically identical pentafunctional protein.

Ferritti et al. (1986) have applied a similar argument to the evolutionary origin of the Streptococcus faecalis bifunctional antibiotic-modifying enzyme encoded by plasmid pIP800. The acetyl- and phosphotransferase activities (aminoglycoside) catalysed on this 56 kDa bifunctional enzyme are found as two ca. 30 kDa monofunctional activities in other bacteria. The gene specifying the bifunctional S. faecalis enzyme can be sub-cloned and the two activities expressed (functionally) separately. These observations have prompted speculation that the bifunctional resistance factor in S.faecalis arose by gene fusion (Ferretti et al., 1986).

On balance it seems likely that the monofunctional and multifunctional shikimate pathway activities of bacteria and fungi are functionally related. The early genetic mapping data of the N.crassa, S.cerevisiae and A.nidulans arom genes (Giles et al., 1967a; Rines et al., 1969; De Leeuw, 1967; Roberts, 1969; see Chapter One) has been validated by subsequent cloning and DNA sequencing work. The presence of five separate domains on the arom protein has been established by chemical modification and limited proteolysis experimental

Pathway step	Activity	Subunit M_r	Length (amino acids)	Quaternary Structure
1	3-dehydroquinate synthase (*aroB*)	38,880	362	monomer
2	3-dehydroquinase (*aroD*)	26,377	240	dimer
3	shikimate dehydrogenase (aroE)	29,380	272	monomer
4	shikimate kinase (*aroL*)	18,937	173	monomer
5	EPSP synthase (*aroA*)	46,112	427	monomer
		159,689	1475	

Table 6.1: Sum of the DNA-derived subunit molecular weights of the five *E. coli* enzymes which correspond to the activities of the fungal *arom* multifunctional protein.

data (Section 1.9). Finally the undisputed relatedness of the five monofunctional E.coli enzymes with the individual domains of the yeast and A.nidulans arom multifunctional enzymes has been established by primary sequence comparisons (Figures 6.1 and 6.2).

Whether the bacterial enzymes fused to produce the multifunctional arom enzyme or vice versa is the topic of the remainder of this discussion.

6.3 Evolution of the arom multifunctional enzyme

6.3.1 Scission or fusion?

The hypothesis that the shikimate pathway arose by gene duplication and subsequent divergence has been discussed previously in Chapter Four. No discernable homologies were found in comparing, in a pair-wise fashion, the six E.coli activities leading from DAHP to chorismate (E1 $\longrightarrow$ E6; Figure 1.1). Conversely, the observation has been made that the five bacterial sequences (E1 $\longrightarrow$ E5) are homologous to separate domains of the fungal polypeptide (Section 6.2.1; Duncan et al., 1986a). Neither the monofunctional nor multifunctional shikimate pathway activities are therefore the products of prior gene duplication events.

Similarly the wealth of evidence, both genetic and biochemical, dismisses the possibility that the two extremes of structural organisation evolved independently. This then leaves two main contenders for the mechanism by which the different types of organisation arose during the process of evolution. Either the monofunctional bacterial enzymes (genes)

fused to form the multifunctional fungal arom-type protein or the pentafunctional enzyme represents the most ancient form which subsequently split in the bacteria.

6.3.2 Gene fusion

The gene fusion hypothesis is automatically the most attractive of the two alternatives. One can readily envisage a melting together of genes for monofunctional enzymes, the mutational removal of termination codons and the production of a multifunctional protein with separately folded but structurally interacting domains. The demonstrable isolation of active proteolytic fragments of multifunctional enzymes is strong evidence for this evolutionary route. The problem of transposing the widely scattered monofunctional genes to a single locus could be compensated for by a positive selection pressure working on a system which co-ordinately expresses several genes of a given biosynthetic pathway.

A possible corollary is provided by the case of fungal tryptophan synthase. This example is often cited as a classical case of gene fusion of the α and β subunits encoded by separate bacterial genes, in an operon with the order β-α (Zalkin & Yanofsky, 1982). All multifunctional tryptophan synthases examined to date have resulted from a α-β fusion, the inverse of the bacterial chromosomal order. It could be argued that if the fungal system was the more ancient then why when gene scission took place was the chromosomal order not retained, but rather inverted?

Similarly, parsimony prescribes that it is easier to evolve a complex protein from pre-existing functional units than to rely on mutation and selection to create an intricate, large and multi-catalytic enzyme. Also if a multifunctional protein was to be split then each of the monofunctional enzymes (genes) would have to obtain the necessary regulatory sequences to ensure its expression. In the case of E.coli these sequences would include promoter elements and ribosome binding site(s), accurately positioned upstream of the structural gene.

6.3.3 Gene scission

The case for gene scission is less strong and depends largely on argument by analogy. The splitting of the complex arom gene could have occurred in response to selection pressure during streamlining of the rapidly replicating bacterial genome. Many hypotheses have been put forward regarding possible advantages for the organisation of five aromatic biosynthetic enzymes on a single multifunctional enzyme (see Chapter One, Section 1.9). None have been experimentally proven. Only the obvious stoichiometric and co-ordinate expression of pathway activities exists as an example of possible positive selective pressure. It is conceivable that scission of the 'genes' from within the complex locus, and their transposition into more advantageous genomic positions, could provide a directly selectable evolutionary pressure. Certainly the arrangement of both aroA (Duncan

& Coggins, 1986) and aroL (Defeyter & Pittard, 1986) in operons suggests the possible existence of an alternate expression selection pressure other than co-ordinate production of enzyme activities. A corollary to this statement is that the structural organisation of these two aro genes in operons in E.coli argues against gene fusion. Although, in the case of the serC-aroA operon, the loss of serine and aromatic biosynthetic linkage would not be detrimental for fungi who= do not produce the iron chelator enterochelin (see Section 1.8.6).

Another line of evidence for a scission mechanism depends upon recent hypotheses on the origins of eukaryotic and prokaryotic cells. Senapathy (1986) has compared codon distribution statistics in genes and proposed that prokaryotic genes could have evolved from primitive unicellular eukaryotes (c.f. yeast) by intron loss.

This hypothesis is based on the observation that the upper limit of reading-frame length in a random nucleotide sequence is longer for prokaryotes than eukaryotes (Senapathy, 1986). As a result, the only way eukaryotic cells could evolve long protein coding sequences would be to splice short segments together. Therefore prokaryotic gene sequences, using this model, could not have evolved independently from primordial DNA sequences but rather evolved from primitive unicellular eukaryotic genes by los ing introns.

This observation unfortunately provides the alternative hypothesis that the primordial DNA sequences contained monofunctional, intron-punctuated aromatic genes that evolved into intronless multifunctional genes in yeast and A.nidulans,

but into separate intronless genes in other primitive eukaryotes from which the bacteria subsequently evolved.

In support of such a prokaryotic lineage, Pace *et al.* (1986) have examined phylogenetic relationships between the three major kingdoms using 16S-like rRNA sequence comparisons. They concluded that the eukaryotic nuclear line of descent did not originate from within either bacterial lines (archaebacteria, or eubacteria).

6.3.4 The origin of the *arom* enzyme

A cursory examination of the fungal and bacterial shikimate pathway enzyme sequences (Figure 6.1) does not immediately reveal clues which might aid in identifying their respective origins. One notable feature though is the total lack of introns in both *S.cerevisiae* and *A.nidulans* sequences. This is not a unique feature since, for example, the triosephosphate isomerase genes from *E.coli*, *S.cerevisiae* and *S.pombe* lack introns whereas those from human and maize have many introns (McKnight *et al.*, 1986). Therefore the absence of introns from both fungal *arom* sequences cannot be regarded as evidence *per se* for a bacterial (gene fusion) origin.

One clue may come from the observation that in many plants the shikimate pathway enzymes are separable, with the exception of the E2 and E3 activities which occur as a bifunctional enzyme (Section 1.5.2). The *Petunia hybrida* EPSP synthase gene has many introns (Shah *et al.*, 1986) but remains 48% homologous with the bacterial enzyme (G. Kishore, personal communication). It could be that investigation of

the gene sequences of the aromatic genes in species like Petunia or pea, which have both separable and multifunctional activities co-existing, may reveal which way evolution is work. Is the bifunctional E2/E3 a partial fusion event which has remained fixed in the population? Or could it be the last remnant of an almost complete scission event? The occurrence (or not) of introns in both the separable and bifunctional genes of the same plant species may point the way to the true evolutionary origin.

One important aspect which has not been discussed yet is the quaternary structural organisation of the shikimate pathway enzymes. As detailed in Table 6.1, four of the five E.coli enzymes, with the exception of dehydroquinase, are monomeric. The N.crassa (Lumsden & Coggins, 1977, 1978) and, almost certainly, the S.cerevisiae arom enzymes are homodimers. Having established that the enzymatic activities of both types of organisation (bacterial and fungal) are related, then this may provide suggestive evidence for gene fusion. Four of the bacterial monofunctional enzymes are stable and catalytically active in a monomeric form. Their organisation as a multifunctional, homodimeric enzyme in fungi could have occurred as a result of a selectable advantage gained by the dimerisation of multifunctional subunits. This dimerisation could have resulted from the dehydroquinase domain retaining its bacterial 'need' to exist as a dimer. In this model the remaining four activities are simply passengers, but once organised as a dimer may confer some

overall advantage.

At first site this hypothesis may seem at odds with the complementation data of Giles et al (N.crassa arom) (1967a,b) which demonstrated domain specific mutations affecting catalytic activity through quaternary structural disruption. Also the stability of the N.crassa arom even under mild proteolysis conditions (Smith & Coggins, 1983; Boocock, 1983) again suggests extensive domain interactions. These sub-unit 'contacts' could have evolved after the initial dehydroquinase-directed dimerisation.

It is difficult to turn this argument around in favour of gene scission. Most proteins do not exist as complex quaternary structures unless they have to for catalytic reasons. If fungal dimerisation is a prerequisite for catalytic activity, then after gene scission why are the bacterial enzymes nearly all monomeric?

6.4 Conclusions

It was Kirschner & Bisswanger (1976) who noted in surveying the occurrence of multifunctional enzymes, that this class of macromolecules were predominantly found in the pathways of amino acid biosynthesis. The differential structural organisations of the pre-chorismate aromatic biosynthetic enzymes of bacteria and fungi, exist to substantiate and to expand the significance of this observation.

The aims of this thesis, as outlined in Chapter One and restated at the beginning of this discussion, were to determine sufficient structural information of a number of

E.coli shikimate pathway enzymes. Data that would complement the wealth of factual information already available (see Sections 1.3 - 1.8) and allow detailed structural, functional and evolutionary comparison between the monofunctional and multifunctional aromatic biosynthetic systems. The results presented in Chapters 3-5 provided that data and the discussion so far has considered such a comparison. This final conclusion will attempt to objectively assess the merit of such an exercise, and to put into perspective the information obtained.

Until recently structural comparisons of multifunctional proteins had relied heavily upon the classic work of Yanofsky and others on the fungal TRP3 and TRP5 gene products. Comparison with the separable bacterial genes/enzymes encoding the same biosynthetic activities allowed interpretation and speculation on a host of relevant factors. Evolutionary origin, structural interactions between otherwise non-covalently attached domains, kinetic and/or functional advantage to be gained by multi-functional organisation. These and other considerations had been applied only to bifunctional enzymes. The work described here forms part of a group effort which has asked the same questions of a pentafunctional enzyme.

The question of evolutionary origin (fusion or scission) posed earlier is in itself only of secondary importance. It could be argued that it is optimistic, *in extremis*, to hope to gain definitive information from one set of five bacterial sequences and two fungal pentafunctional sequences. The difficulties which I have outlined in separating such alternatives perhaps not only confirms this but also highlights

a major flaw in attempting such comparisons. Too little is known about evolutionary direction between prokaryotes, eukaryotes and archaebacteria to apply any real central dogma to this system.

Notwithstanding, the identification of five separate domains within the fungal arom sequence (Figures 6.1; 6.3) has widespread significance. The homologies observed define, by implication, those residues required for efficient catalysis. Dissimilarity between the fungal and bacterial sequences should prove to be within regions which, under no direct selection pressure, are free to mutate. This is of course with the proviso that primary structure dissimilarity does not necessarily mean tertiary structural differences. A lack of identity does not always mean a lack of homology. For this reason the comparisons shown in Figure 6.1 include not only direct identities but also amino acid changes considered to be 'conservative' in nature.

The identification of the separate domains immediately defines the inter-domain (or connector) regions. These in turn identify regions within the multifunctional sequence required for structural integrity and maintenance of inter-domain separation (Figure 6.3). Analysis of and artificial changes to these connectors could, in the future, reveal much about the tertiary (and quaternary) stability of the arom complex and some factors which may govern it.

Similarly the comparison between the bacterial and fungal sequences has revealed the presence of two sub-domains within the DHQ synthase region. Until now this was totally

unsuspected, and the location within one of these sub-domains of the region identified previously (Section 3.9.6; Millar & Coggins, 1986) as the NAD^+ binding site, supports this hypothesis. The second DHQ synthase sub-domain may contain sequences involved in DHQ binding. Whether the hypothetical common structural motif within each of the shikimate pathway enzymes discussed in Chapter Four could bear any resemblance to this second sub-domain remains unresolved. At first sight the identification of five non-overlapping domains (identified by direct homology) and a common structural element may seem mutually exclusive. However if the latter evolved convergently within each of the five domains and displays no obvious primary structure signature then the two could co-exist. Three-dimensional structural analysis and comparison of each of the five domains could reveal such a conserved feature.

The acid test of the proposal that five domains exist would be the excision and expression of each individually. For a number of reasons this may not be trivial, especially for 'C-terminus-proximal' activities. However such work is under way and could provide much information not only on individual expression but also on factors as diverse as protein folding (and regulation thereof) and quaternary stability interactions between the various domains.

What has been achieved already is definitive evidence which confirms a readily assumed but little substantiated hypothesis. The pentafunctional arom complex of S.cerevisiae, A.nidulans and (probably) N.crassa is a mosaic of functional

and structural domains. These independently folded protein domains are arranged almost as 'beads on a thread'. The simplicity of this statement should not disguise the significance of the observation.

REFERENCES

Ahmad, S. & Jensen, R.A. (1986).
The evolutionary history of two bifunctional proteins that emerged in the purple bacteria.
Trends in Biochemical Science, 11, 108-112.

Ahmed, S.I. & Giles, N.H. (1969).
Organisation of enzymes of the common aromatic synthetic pathway: Evidence for aggregation in fungi.
Journal of Bacteriology, 99, 231-237.

Aiba, H., Adhya, S. & de Crombrugghe, B. (1981).
Two functional gal promoters in intact E.coli cells.
Journal of Biological Chemistry, 256, 11905-11910.

Amhrein, N., Schab, J. & Steinrucken, H.C. (1980).
The mode of action of the herbicide glyphosate.
Nuturwiss, 67, 356-357.

Anton, I.A. & Coggins, J.R. (1983).
Subcloning of the Escherichia coli shikimate dehydrogenase gene (aroE) from the transducing bacteriophage λspc1.
Biochemical Society Transactions, 12, 275-276.

Anton, I.A. & Coggins, J.R. (1986).
The complete amino acid sequence of the E.coli aroE gene.
Biochemical Journal, in the press.

Bachmann, B. (1983).
Linkage map of Escherichia coli K12, edition 7.
Microbiological Reviews, 44, 180-230.

Belfaiza, J., Parsot, C., Martel, A., Bouthier de la Tour, C., Margarita, D., Cohen, G. & Sait-Girons, I. (1986).
Evolution in biosynthetic pathways: Two enzymes catalysing consecutive steps in methionine biosynthesis. originate from a common ancestor and possess a similar regulatory region.
Proceedings of the National Academy of Sciences (USA), 83, 867-871.

Berylyn, M.B., Ahmed, S.I. & Giles, N.H. (1970).
Organisation of polyaromatic biosynthetic enzymes in a variety of photosynthetic organisms.
Journal of Bacteriology, 104, 768-774.

Berlyn, M.B. & Giles, N.H. (1969).
Organisation of enzymes in the polyaromatic synthetic pathway: separability in bacteria.
Journal of Bacteriology, 99, 222-230.

Biggin, M.D., Gibson, T.J. & Hong, G.F. (1983).
Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination.
Proceedings of the National Academy of Sciences (U.S.A. 80, 3963-3965.

Birnboim, H.C. & Doly, J. (1979).
A rapid alkaline extraction procedure for screening recombinant plasmid DNA.
Nucleic Acids Research, 7, 1513-1523.

Bisswanger, H. & Schimcke-Ott, E. (1980).
Multifunctional Proteins. New York: John Wiley & Sons.

Block, K. & Vance, D. (1977).
Control mechanisms in the synthesis of saturated fatty acids.
Annual Review of Biochemistry, 46, 263-298.

Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.C., Heynecker, H.L., Boyer, H.W., Crosa, J.H. & Falkow, S. (1977).
Construction and characterisation of new cloning vehicles. II. A multi-purpose cloning system.
Gene, 2, 95-113.

Boocock, M.R. (1983).
Ph.D. Thesis. University of Glasgow.

Boudet, A.M. & Lécusson, R. (1974).
Studies on the association of 5-Dehydroquinate hydro-lyase and shikimate: $NADP^+$-oxidoreductase in higher plants.
Planta, 119, 71-79.

Bowen, J.R. & Kosuge, T. (1979).
In vivo activity, purification and characterisation of shikimate kinase from Sorghum.
Plant Physiology, 64, 382-386.

Bradford, M.M. (1976).
A rapid and sensitive method for the quantitation of microgram quantities of protein utilising the principle of protein-dye binding.
Analytical Biochemistry, 72, 248-254.

Brent, R. & Ptashne, M. (1981).
Mechanism of action of the lexA gene product.
Proceedings of the National Academy of Sciences (USA), 78, 4204-4208.

Brosius, J., Ullrich, A., Raker, M.A., Gray, A., Dull, T.J., Guttell, R.R. & Noller, H.F. (1981a).
Construction and fine mapping of recombinant plasmids containing the rrnB ribosomal RNA operon of E.coli.
Plasmid, 6, 112-118.

Brosius, J., Dull, T., Sleeter, D.D. & Noller, H.F. (1981b).
Gene organisation and primary structure of a ribosomal RNA operon from E.coli.
Journal of Molecular Biology, 148, 107-128.

Brown, K.D. (1968).
Regulation of aromatic amino acid biosynthesis in Escherichia coli K12.
Genetics, 60, 31-48.

Brown, K.D. & Somerville, R.L. (1971).
Repression of aromatic amino acid biosynthesis in Escherichia coli K12.
Journal of Bacteriology, 108, 386-399.

Bruni, C.B., Carlomagno, M.S., Formisano, S. & Paolella, Cr. (198
Primary and secondary structural homologies between the His4 gene product of Saccharomyces cerevisiae and the hisIE and hisD gene products of Escherichia coli and Salmonella typhimurium.
Molecular and General Genetics, 203, 389-396.

Burgoyne, L., Case, M.E. & Giles, N.H. (1969).
Purification and properties of the aromatic (arom) synthetic enzyme aggregate of Neurospora crassa.
Biochimica et Biophysica Acta, 191, 452-462.

Camakaris, H. & Pittard, J. (1973).
Regulation of tyrosine and phenylanine biosynthesis in Escherichia coli K12: properties of the tyrR gene product.
Journal of Bacteriology, 115, 1135-1144.

Camakaris, H. & Pittard. (1983).
Tyrosine biosynthesis. In: Amino acids: Biosynthesis and genetic regulation, p.339-350. London: Addison-Wesley.

Carter, P.E., Dunbar, B. & Fothergill, J.E. (1983).
The serine proteinase chain of human complement component C1s.
Biochemical Journal, 215, 565-571.

Casino, A., Cipollaro, M., Guerrni, A.M., Mastrocinque, G., Spena, A. & Scarlato, V. (1981).
Coding capacity of complementary DNA strands.
Nucleic Acids Research, 9, 1499-1518.

Catcheside, D.E.A., Storer, P.J. & Klein, B. (1985).
Cloning of the ARO cluster gene of Neurospora crassa and its expression in Escherichia coli.
Molecular and General Genetics, 199, 446-451.

Chaconas, G. & van de Sande, J.H. (1980).
5'-^{32}P Labelling of RNA and DNA restriction fragments.
Methods in Enzymology, 65, 75-85.

Charles, I.G., Keyte, J.W., Brammar, W.J. & Hawkins, A.R. (1985).
Nucleotide sequence encoding the biosynthetic dehydroquinase function of the pentafunctional AROM locus of Aspergillus nidulans.
Nucleic Acids Research, 13, 8119-8128.

Charles, I.G., Keyte, J.W., Brammar, W.J., Smith, M. & Hawkins, A.R. (1986).
The isolation and nucleotide sequence of the complex AROM locus of Aspergillus nidulans.
Nucleic Acids Research, 14, 2201-2213.

Chaudhuri, S. & Coggins, J.R. (1982).
Catabolic dehydroquinase from N.crassa.
Neurospora Newsletter, 29, 12-13.

Chaudhuri, S. & Coggins, J.R. (1985).
The purification of shikimate dehydrogenase from Escherichia coli.
Biochemical Journal, 226, 217-223.

Chaudhuri, S., Lambert, J.L., McColl, L.A. & Coggins, J.R. (1986).
Purification and characterisation of 3-dehydroquinase from Escherichia coli.
Biochemical Journal, 239,

Chiariotti, L., Alifano, P., Carlomagno, M.S. & Bruni, C.B. (1986)
Nucleotide sequence of the Escherichia coli hisD gene and of the Escherichia coli and Salmonella typhimurium hisIE region.
Molecular and General Genetics, 203, 382-388.

Chou, P.Y. & Fasman, G.D. (1978).
Empirical predictions of protein conformation.
Annual Review of Biochemistry, 47, 251-277.

Clark, J.E. & Ljungdahl, L.G. (1982).
Purification of 5,10-Methylenetetrahydrofolate cyclohydrolase from Clostridium formicoaceticum.
Journal of Biological Chemistry, 257, 3833-3836.

Clarke, L. & Carbon, J. (1976).
A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire E.coli genome.
Cell, 9, 91-99.

Coggins, J.R. (1982).
The arom pentafunctional enzyme.
Neurospora Newsletter, 29, 8.

Coggins, J.R. & Boocock, M.R. (1986).
The arom multifunctional enzyme. In: Multidomain Proteins - Structure and Evolution (eds. Hardie, D.G. & Coggins, J.R.). Elsevier Science Publishers. pp.259-282.

Coggins, J.R., Boocock, M.R., Campbell, M.S., Chaudhuri, S., Lambert, J.M., Lewendon, A., Mousdale, D.M. & Smith, D.D.S. (1985).
Functional domains involved in aromatic amino acid biosynthesis.
Biochemical Society Transactions, 13, 299-303.

Coggins, J.R. & Hardie, G.D. (1986).
The domain as the fundamental unit of protein structure and evolution. In: Multidomain Proteins: Structure and Evolution (eds. Hardie, D.G. & Coggins, J.R.). Elsevier Science Publishers. pp.1-5.

Coggins, J.R., Smith, D.D.S., Chaudhuri, S., Coia, A.A., Lambert, J.M. & Nimmo, G.A. (1981).
The arom multifunctional enzyme of Neurospora crassa: Structural and functional studies on a pentafunctional polypeptide.
Biochemical Society Transactions, 9, 44P.

Comai, L., Facciotti, D., Hiatt, W.R., Thompson, G., Rose, R.E. & Stalker, D.M. (1985).
Expression in plants of a mutant aroA gene from Salmonella typhimurium confers tolerance to glyphosate
Nature, 317, 741-744.

Comai, L., Sen, L.C., Stalker, D.M. (1983).
An altered aroA gene product confers resistance to the herbicide Glyphosate.
Science, 221, 370-371.

Cornish, E.C., Davidson, B.E. & Pittard, J. (1982).
Cloning and characterisation of Escherichia coli K12 regulator gene tyrR.
Journal of Bacteriology, 152, 1276-1279.

Crawford, I.P. (1975).
Regulation of enzyme synthesis in the tryptophan pathway of Acinetobacter calcoaceticus.
Annual Review of Bacteriology, 39, 87-120.

Crawford, I.P. (1980).
Gene fusions in the tryptophan pathway: Tryptophan synthase and Phosphoribosyl-anthranilate isomerase: Indoleglycerol-phosphate synthase. In: Multifunctional Proteins, eds. Bisswanger, H. & Schmincke-Ott, E. Ch.5, p.151-173.

Crawford, I.P., Nichols, B.P. & Yanofsky, C. (1980). Nucleotide sequence of the trpB gene in E.coli and S. typhimurium. Journal of Molecular Biology, 142, 489-502.

Crawford, I.P. & Stauffer, G.V. (1980). Regulation of tryptophan biosynthesis. Annual Review of Biochemistry, 49, 163-195.

Cunin, R., Eckhardt, T., Piette, J., Boyen, A., Pierard, A., & Glansdorff, N. (1983). Molecular basis for modulated regulation of gene expression in the arginine regulon of E.coli K12. Nucleic Acids Research 11, 5007-5019.

Dagert, M. & Ehrlich, S.D. (1979). Prolonged incubation in calcium chloride improves the competence of Escherichia coli cells. Gene, 6, 23-28.

Dansette, P. & Azerad, R. (1974). Stereospecificity of hydrogen transfer catalysed by NADPH-dehydroshikimate reductase of E.coli. Biochimie, 56, 751-755.

Davis, B.D. (1951). Aromatic biosynthesis. I. The role of shikimic acid. Journal of Biological Chemistry, 191, 315-325.

Davis, W.D. & Davidson, B.E. (1982). The nucleotide sequence of aroG, the gene for 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase (phe) in E.coli K12. Nucleic Acids Research, 10, 4045-4058.

Dayhoff, M.M. (1978). Atlas of Protein Sequence and Structure (National Biomedical Research Foundation, Washington, D.C., U.S.A.), pp36.

De Boer, H.A., Comstock, L.J. & Vasser, M. (1983). The tac promoters, A functional hybrid derived from the trp and lac promoters. Proceedings of the National Academy of Sciences (USA), 80, 21-25.

Defeyter, R.C. & Pittard, J. (1986). Genetic and molecular analysis of aroL, the gene for shikimate kinase II in Escherichia coli K12. Journal of Bacteriology, 165, 226-232.

de Leeuw, A. (1967). Gene-enzyme relationships in polyaromatic auxotrophic mutants in Saccharomyces cerevisiae. Genetics, 56, 554-555.

Demerec, M. (1964).
Clustering of functionally related genes in Salmonella typhimurium.
Proceedings of the National Academy of Science (USA), 51, 1057-1060.

Dev, I.K. & Harvey, R.J. (1978).
A complex of N^5,N^{10}-Methylenetetrahydrofolate dehydrogenase and N^5, N^{10}-Methylenetetrahydrofolate cyclohydrolase from Escherichia coli.
Journal of Biological Chemistry, 253, 4245-4253.

Devereux, J., Haeberli, P. & Smithies, O. (1984).
A comprehensive set of sequence analysis programs for the VAX.
Nucleic Acids Research, 12, 387-395.

DeVries, J.K. & Zubay, G. (1967).
DNA directed peptide synthesis, II. The synthesis of the α-fragment of the enzyme β-galactosidase.
Proceedings of the National Academy of Sciences (USA), 57, 1010.

Donahue, T.F., Farabaugh, P.J. & Fink, G.R. (1982).
The nucleotide sequence of the HIS4 region of yeast.
Gene, 18, 47-59.

Doolittle, R.F. (1981).
Similar amino acid sequences: Chance or common ancestry?
Science, 214, 149-159.

Dougan, G., Saul, M., Warren, G. & Sherratt, D. (1978).
A functional map of plasmid ColE1.
Molecular and General Genetics, 158, 325-327.

Doy, C.H. & Brown, K.D. (1965).
Control of aromatic biosynthesis: The multiplicity of 7-phospho-2-oxo-3-deoxy-D-arabino-heptonate D-erythrose-4-phosphate-lyase (pyruvate phosphorylating) in Escherichia coli.
Biochimica et Biophysica Acta, 104, 377-389.

Duncan, K. & Coggins, J.R. (1983).
Subcloning of the Escherichia coli genes aroA (5-enolpyruvylshikimate 3-phosphate synthase) and aroB (3-dehydroquinate synthase).
Biochemical Society Transactions, 12, 274-275.

Duncan, K., Lewendon, A. & Coggins, J.R. (1984a).
The purification of 5-enolpyruvylshikimate 3-phosphate synthase from an overproducing strain of Escherichia coli.
FEBS Letters, 165, 121-127.

Duncan, K., Lewendon, A. & Coggins, J.R. (1984b).
The complete amino acid sequence of Escherichia coli 5-enolpyruvyl shikimate 3-phosphate synthase.
FEBS Letters, 170, 59-63.

Duncan, K. & Coggins, J.R. (1986).
The serC-aroA operon of Escherichia coli.
Biochemical Journal, 234, 49-57.

Duncan, K., Dacey, S.A., Edwards, R.M. & Coggins, J.R. (1986a).
Homologies between the arom pentafunctional enzyme of Saccharomyces cerevisiae and the corresponding monofunctional Escherichia coli enzymes.
Biochemical Journal, in the press.

Duncan, K., Chaudhuri, S., Campbell, M.S. & Coggins, J.R. (1986b).
The overexpression and complete amino acid sequence of Escherichia coli 3-dehydroquinase.
Biochemical Journal, 238, 475-483.

Ely, B. & Pittard, J. (1979).
Aromatic amino acid biosynthesis: Regulation of shikimate kinase in E.coli K12.
Journal of Bacteriology, 138, 933-943.

Evans, P.R., Farrants, G.W. & Hudson, P.J. (1981).
Phosphofructokinase: structure and control.
Philosophical Transactions of the Royal Society London B, 293, 53-62.

Ferrara, P., Duchange, N., Zakin, M.M. & Cohen, G.H. (1984).
Internal homologies in the two aspartokinase-homoserine dehydrogenases of Escherichia coli K12.
Proceedings of the National Academy of Sciences (USA), 81, 3019-3023.

Ferretti, J.J., Gilmore, K.S. & Courvalin, P. (1986).
Nucleotide sequence analysis of the gene specifying the bifunctional 6'-Aminoglycoside Acetyltransferase 2"-Aminoglycoside Phosphotransferase enzyme in Streptococcus faecalis and identification and cloning of gene regions specifying the two activities.
Journal of Bacteriology, 167, 631-638.

Fickett, J.W. (1982).
Recognition of protein coding regions in DNA sequences.
Nucleic Acids Research, 10, 5303-5318.

Finch, P.W. & Emmerson, P.T. (1984).
The nucleotide sequence of the uvrD gene of E.coli.
Nucleic Acids Research, 12, 5789-5798.

Fotheringham, I.G., Dacey, S.A., Taylor, P.P., Smith, T.J., Hunter, M.G., Finlay, M.E., Primrose, S.B., Parker, D.M. & Edwards, R.M. (1986).
The cloning and sequence analysis of the aspC and tyrB genes from Escherichia coli K12.
Biochemical Journal, 234, 593-604.

Frost, J., Bender, J., Kadonaga, J.T. & Knowles, J.R. (1984).
Dhydroquinate synthase from E.coli: Purification, cloning and the construction of overproducers of the enzyme.
Biochemistry, 23, 4470-4475.

Fry, D.C., Kuby, S.A. & Mildvan, A.S. (1986).
ATP-binding site of adenylate kinase: mechanistic implications of its homology with ras-encoded p21, F_1-ATPase and other nucleotide binding proteins.
Proceedings of the National Academy of Sciences (USA), 83, 907-911.

Gaertner, F.H. (1972).
Purification of two multienzyme complexes in aromatic/tryptophan pathway of Neurospora crassa.
Archives of Biochemistry and Biophysics, 151, 277-284.

Gaertner, F.H. & Cole, K.W. (1977).
A cluster gene: evidence for one gene, one polypeptide, five enzymes.
Biochemical and Biophysical Research Communications, 75, 259-264.

Garner, C. & Herrmann, K.M. (1983).
Biosynthesis of phenylalanine. In: Amino acids: Biosynthesis and genetic regulation, eds. Herrmann, K.M. & Somerville, R.L. Ch. 18, p.323-338. London: Addison-Wesley.

Garner, C. & Herrmann, K.M. (1985).
Operator mutations of the Escherichia coli aroF gene.
The Journal of Biological Chemistry, 260, 3820-3825.

Gibson, F. & Pittard, J. (1968).
Pathways of biosynthesis of aromatic amino acids and vitamins and their control in microorganisms.
Bacteriological Review, 32, 465-492.

Gicquel-Sanzey, B. & Cossart, P. (1982).
Homologies between different prokaryotic DNA-binding regulating proteins and their sites of action.
EMBO Journal, 1, 591-595.

Gilbert, W., Marchionni, M. & McKnight, G. (1986).
On the antiquity of introns.
Cell, 46, 151-154.

Giles, N.H. (1978).
The organisation, function and evolution of gene clusters in Eukaryotes.
American Naturalist, 112, 641-657.

Giles, N.H., Case, M.E., Partridge, C.W.H. & Ahmed, S.I. (1967a).
A gene cluster in Neurospora crassa coding for an aggregate of five aromatic synthetic enzymes.
Proceedings of the National Academy of Sciences (USA), 58, 1453-1460.

Giles, N.H., Partridge, C.W.H., Ahmed, S.I. & Case, M.E. (1967b).
The occurrence of two dehydroquinases in Neurospora crassa, one constitutive and one inducible.
Proceedings of the National Academy of Sciences (USA), 58, 1930-1937.

Gollub, E., Zalkin, H. & Sprinson, D.B. (1967).
Correlation of genes and enzymes, and studies on regulation of the aromatic pathway in Salmonella.
Journal of Biological Chemistry, 242, 5323-5328.

Gouy, M. & Gautier, C. (1982).
Codon usage in bacteria: correlation with gene expressivity.
Nucleic Acids Research, 10, 7055-7074.

Gowrishankar, J. & Pittard, J. (1982).
Construction from Mu dl (lac Ap^r) lysogens of lambda bacteriophage bearing promoter-lac fusions: isolation of λppheA-lac.
Journal of Bacteriology, 150, 1122-1129.

Greer, S. & Perham, R.N. (1986).
Glutathione reductase from Escherichia coli: Cloning and sequence analysis of the gene and relationship to other Flavoprotein Disulphide Oxidoreductases.
Biochemistry, 25, 2736-2742.

Grosjean, H. & Fiers, W. (1982).
Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes.
Gene, 18, 199-209.

Gross, S.R. & Fein, A. (1960).
Linkage and function in Neurospora.
Genetics, 45, 885-904.

Hagervall, T.G. & Björk, G.R. (1984).
Genetic mapping and cloning of the gene (trmC) responsible for the synthesis of tRNA (mnm^5s^2U) methyltransferase in Escherichia coli K12.
Molecular and General Genetics, 196, 201-207.

Hardie, G.D. & McCarthy, A.D. (1986).
Evolution of fatty acid synthase systems by gene fusion. In: Multidomain Proteins - Structure and Evolution (eds. Hardie, D.G. & Coggins, J.R.).
Elsevier Science Publishers. pp.229-258.

Hasan, N. & Nester, E.W. (1978a).
Dehydroquinate synthase in Bacillus subtilis.
Journal of Biological Chemistry, 253, 4999-5004.

Hasan, N. & Nester, E.W. (1978b).
Purification and properties of chorismate synthase from Bacillus subtilis.
Journal of Biological Chemistry, 253, 4993-4998.

Haslam, E. (1974).
The Shikimate Pathway.
London: Butterworths.

Hawley, D.K. & McClure, W.R. (1983).
Compilation and analysis of Escherichia coli promoter DNA sequences.
Nucleic Acids Research, 11, 2237-2255.

Henner, D.J. & Hoch, J.A. (1980).
The Bacillus subtilis chromosome.
Microbiological Reviews, 44, 57-82.

Herrmann, K.M. (1983).
The common aromatic biosynthetic pathway. In: Amino acids: biosynthesis and genetic regulation, ed. Herrmann, K.M. & Somerville, R.L. Ch. 17, p.301-322. London: Addison-Wesley.

Herrmann, K.M. & Somerville, R.L. (1983).
Amino Acids: biosynthesis and genetic regulation.
London: Addison Wesley.

Hirs, C.H.W. (1967).
Performic acid oxidation.
Methods in Enzymology, 11, 197-199.

Hoiseth, S.K. & Stocker, B.A.D. (1985).
Genes aroA and serC of Salmonella typhimurium, constitute an operon.
Journal of Bacteriology, 163, 355-361.

Holmes, D.S. & Quigley, M. (1981).
A rapid boiling method for the preparation of bacterial plasmids.
Analytical Biochemistry, 114, 193-197.

Horowitz, N.H. (1945).
cited in Horowitz, N.H. (1965).
'Evolving genes and proteins'
(eds. Bryson, V. & Vogel, H.J.) pp.15-24. London, Academic Press.

Huang, L., Montoya, A.L. & Nester, E.W (1975).
Purification and characterisation of shikimate kinase enzyme activity in Bacillus subtilis.
Journal of Biological Chemistry, 250, 7675-7681.

Hudson, G.S. & Davidson, B.E. (1984).
Nucleotide sequence and transcription of the Phenylalanine and Tyrosine operons of Escherichia coli K12.
Journal of Molecular Biology, 180, 1023-1051.

Huiet, L. (1984).
Molecular analysis of the Neurospora qa-1 regulatory region indicates that two interacting genes control qa gene expression.
Proceedings of the National Academy of Sciences (USA),

Im, S.W.K., Davidson, H. & Pittard, J. (1971).
Phenylalanine and tyrosine biosynthesis in Escherichia coli K12: mutants derepressed for 3-deoxy-D-arabino-heptulosonic acid 7-phosphate synthetase (phe), 3-deoxy-D-arabinoheptulosonic acid 7-phosphate synthetase (tyr), chorismate mutase T-prephenate dehydrogenase and transaminase A.
Journal of Bacteriology, 108, 400-409.

Jensen, R.A. & Pierson, D.L. (1975).
Evolutionary implications of different types of microbial enzymology for L-tyrosine biosynthesis.
Nature, 254, 667-671.

Kania, J. & Müller-Hill, B. (1980).
in: Multifunctional Proteins.
(Bisswanger, H. & Schmincke-Ott, E., eds.)
pp.31-48, Wiley, New York.

Katinka, M., Cossart, P., Sibilli, L., Saint-Girons, I., Chahignac, M.A., Le Bras, G., Cohen, G.N. & Yaniv, M. (1980).
Nucleotide sequence of the thrA gene of Escherichia coli.
Proceedings of the National Academy of Sciences (USA), 77, 5730-5733.

Kelly, R.B., Cozzarelli, N.R., Deutscher, M.P., Lehman, I.R. & Kornberg, A. (1970).
Enzymatic synthesis of deoxyribonucleic acid XXXII. Replication of duplex deoxyribonucleic acid by polymerase at a single strand break.
Journal of Biological Chemistry, 245, 39-45.

Kinghorn, J.R. & Hawkins, A.R. (1982).
Cloning and expression in E.coli K12 of the biosynthetic dehydroquinase function of the arom cluster gene from the Eukaryote, Aspergillus nidulans.
Molecular and General Genetics, 186, 145-152.

Kinghorn, J.R., Schweizer, M., Giles, N.H. & Kushner, S.R. (1981)
The cloning and analysis of the aroD gene of E.coli K12.
Gene, 14, 73-80.

Kirschner, K. & Bisswanger, H. (1976).
Multifunctional proteins.
Annual Review of Biochemistry, 45, 143-166.

Königsberg, W. & Godson, N.G. (1983).
Evidence for the use of rare codons in the dnaG gene and other regulatory genes of Escherichia coli.
Proceedings of the National Academy of Science (USA), 80, 687-691.

Koshiba, T. (1979).
Organisation of enzymes of the shikimate pathway of Phaseolus mungo seedlings.
Plant and Cellular Physiology, 20, 667-670.

Laemmli, U.K. (1970).
Cleavage of structural proteins during the assembly of the head of bacteriophage T4.
Nature, 227, 680-685.

Lambert, J.M., Boocock, M.R. & Coggins, J.R. (1985).
The 3-dehydroquinate synthase activity of the penta-functional arom enzyme complex of Neurospora crassa is zinc-dependent.
Biochemical Journal, 226, 817-829.

Larimer, F.W., Morse, C.C., Beck, A.K., Cole, K.W. & Gaertner, F.H. (1983).
Isolation of the ARO1 cluster gene of Saccharomyces cerevisiae.
Molecular and Cellular Biology, 3, 1609-1614.

Levin, J.G. & Sprinson, D.B. (1964).
The enzymatic formation and isolation of 5-enolpyryvyl-shikimate 3-phosphate.
Journal of Biological Chemistry, 239, 1142-1150.

Lewendon, A. & Coggins, J.R. (1983).
Purification of 5-enolpyruvylshikimate 3-phosphate synthase from Escherichia coli.
Biochemical Journal, 213, 187-191.

Little, J.W., Mount, D.W. & Yanisch-Perron, C.R. (1981).
Purified lexA protein is a repressor of the recA and lexA genes.
Proceedings of the National Academy of Sciences (USA), 78, 4199-4203.

Llewellyn, D.J., Daday, A. & Smith, G.D. (1980).
Evidence for an artifically evolved bifunctional 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase-Chorismate mutase in Bacillus subtilis.
Journal of Biological Chemistry, 255, 2077-2084.

Lumsden, J. & Coggins, J.R. (1977).
The subunit structure of the arom multienzyme complex of Neurospora crassa. A possible pentafunctional polypeptide chain.
Biochemical Journal, 161, 599-607.

Lumsden, J. & Coggins, J.R. (1978).
The subunit structure of the arom multienzyme complex of Neurospora crassa. Evidence from peptide maps for the identity of the subunits.
Biochemical Journal, 169, 441-444.

McCandliss, R.J., Poling, M.J. & Herrmann, K.M. (1978).
3-deoxy-D-arabino-heptulosonate 7-phosphate synthase: Purification and molecular characterisation of the phenylalanine-sensitive isoenzyme from Escherichia coli.
Journal of Biological Chemistry, 253, 4259-4265.

McCarthy, A.D. & Hardie, D.G. (1984).
Fatty acid synthase - an example of protein evolution by gene fusion.
Trends in Biochemical Sciences, 9, 60-63.

McKenzie, K.Q. & Jones, E.W. (1977).
Mutants of the formyltetrahydrofolate interconversion pathway of Saccharomyces cerevisiae.
Genetics, 86, 85-102.

McKnight, G.L., O'Hara, P.J. & Parker, M.L. (1986).
Nucleotide sequence of the triosephosphate isomerase gene from Aspergillus nidulans: implications for a differential loss of introns.
Cell, 46, 143-147.

Maitra, U.S. & Sprinson, D.B. (1978).
5-dehydro-3-deoxy-D-arabino-heptulosonic acid 7-phosphate: An intermediate in the 3-dehydroquinate synthase reaction.
Journal of Biological Chemistry, 253, 5426-5430.

Maniatis, T., Fritsch, E.F. & Sambrook, J. (1982).
Molecular Cloning: A Laboratory Manual. New York: Cold Spring Harbour Publications.

Matchett, W.H. (1974).
Indole channeling by tryptophan synthase of Neurospora.
Journal of Biological Chemistry, 249, 4041-4049.

Matchett, W.H. & De Moss, J.A. (1975).
The subunit structure of Tryptophan synthase from Neurospora crassa.
Journal of Biological Chemistry, 250, 2941-2946.

Mattern, I.A. & Pittard, J. (1971).
Regulation of tyrosine biosynthesis in Escherichia coli K12: isolation and characterisation of operator mutants.
Journal of Bacteriology, 107, 8-15.

Messing, J. (1983).
New M13 vectors for cloning.
Methods in Enzymology, 101, 20-78.

Messing, J. & Vieira, J. (1982).
A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments.
Gene, 19, 269-276.

Midgley, C.A. & Murray, N.E. (1985).
T4 polynucleotide kinase; cloning of the gene (pseT) and amplification of its product.
EMBO Journal, 4, 2695-2703.

Millar, G. & Coggins, J.R. (1986).
The complete amino acid sequence of 3-dehydroquinate synthase of Escherichia coli K12.
FEBS Letters, 200, 11-17.

Millar, G., Anton, I.A., Mousdale, D.M., White, P.J. & Coggins, J.R. (1986a).
Cloning and overexpression of Escherichia coli aroC gene encoding the enzyme chorismate synthase.
Biochemical Society Transactions, 14, 262-263.

Millar, G., Lewendon, A., Hunter, M.G. & Coggins, J.R. (1986b).
The cloning and overexpression of the aroL gene from Escherichia coli K12.
Biochemical Journal, 237, 427-437.

Miozzari, G.F. & Yanofsky, C. (1979).
Gene fusion during the evolution of the tryptophan operon in Enterobacteriaceae.
Nature, 277, 486-489.

Morell, H., Clark, M.J., Knowles, P.F. & Sprinson, D.B. (1967)
The enzymatic synthesis of chorismic and prephenic acids from 5-enolpyruvylshikimate 3-phosphate.
Journal of Biological Chemistry, 242, 82-90.

Morell, H. & Sprinson, D.B. (1968).
Shikimate kinase isoenzymes in Salmonella typhimurium.
Journal of Biological Chemistry, 243, 676-677.

Mousdale, D.M. & Coggins, J.R. (1984).
Purification and properties of 5-enolpyruvylshikimate 3-phosphate synthase from seedlings of Pisum sativum L.
Planta, 160, 78-83.

Mousdale, D.M. & Coggins, J.R. (1985).
Subcellular localisation of the common shikimate-pathway enzymes in Pisum sativum L.
Planta, 163, 241-249.

Mousdale, D.M., Campbell, M.S. & Coggins, J.R. (1986).
Purification and characterisation of bifunctional dehydroquinate-shikimate: NADP oxireductase from Pea Seedlings.
Planta, in the press.

Mulligan, M.E., Brosius, J. & McClure, W.R. (1986).
Characterisation in vitro of the effect of spacer length on the activity of Escherichia coli RNA polymerase at the TAC promoter.
Journal of Biological Chemistry, 260, 3529-3538.

Nakanishi, N. & Yamamoto, M. (1984).
Analysis of the structure and transcription of the aro3 cluster gene in Schizosaccharomyces pombe.
Molecular and General Genetics, 195, 164-169.

Nakatsukasa, W.M. & Nester, E.W. (1972).
Regulation of aromatic amino acid biosynthesis in Bacillus subtilis 168.
Journal of Biological Chemistry, 247, 5972-5979.

Nasser, D., Henderson, G. & Nester, E.W. (1969).
Regulated enzymes of aromatic amino acid synthesis: control, isozymic nature and aggregation in B.subtilis and B.licheniformis.
Journal of Bacteriology, 98, 44-50.

Nichols, B.P. & Yanofsky, C. (1979).
Nucleotide sequence of trpA of Salmonella typhimurium and Escherichia coli: An evolutionary comparison.
Proceedings of the National Academy of Sciences (USA), 76, 5244-5248.

Nimmo, G.A. & Coggins, J.R. (1981).
The purification and properties of the tryptophan-sensitive 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase from Neurospora crassa.
Biochemical Journal, 197, 427-436.

Pace, N.R., Olsen, G.J. & Woese, C.R. (1986).
Ribosomal RNA phylogeny and the primary lines of evolutionary descent.
Cell, 45, 325-326.

Patel, V.B. & Giles, N.H. (1979).
Purification of the AROM multi-enzyme aggregate from Euglena gracilis.
Biochimica et Biophysica Acta, 567, 24-34.

Paukert, J.L., Williams, G.R. & Rabinowitz, J.C. (1977).
Formyl-methenyl-methylenetetrahydrofolate synthase: Correlation of enzymic activities with limited proteolytic degredation from yeast.
Biochemical and Biophysical Research Communications, 77, 147-154.

Pittard, J. & Wallace, B.J. (1966). Distribution and function of genes concerned with aromatic biosynthesis in Escherichia coli.
Journal of Bacteriology, 91, 1494-1508.

Platt, T. (1978).
Regulation of gene expression in the tryptophan operon of Escherichia coli. In: The Operon. Eds. Miller, J.H. & Reznikoff, p.263-302.
New York: Cold Spring Harbour Publications.

Polley, L.D. (1978).
Purification and characterisation of 3-Dehydroquinate hydrolyase and shikimate oxidoreductase. Evidence for a bifunctional enzyme.
Biochimica et Biophysica Acta, 526, 259-266.

Pribnow, D. (1975).
Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter.
Proceedings of the National Academy of Sciences (USA), 72, 784-788.

Rak, B., Lusky, M. & Hable, M. (1982).
Expression of two proteins from overlapping and oppositely oriented genes on transposable DNA insertion element IS5.
Nature, 297, 124-128.

Rines, H.W., Case, M.E. & Giles, N.H. (1969).
Mutants in the arom gene cluster of Neurospora crassa specific for biosynthetic dehydroquinase.
Genetics, 61, 789-800.

Roberts, C.F. (1969).
Isolation of multiple aromatic amino-acid mutants in A. nidulans.
Aspergillus Newsletter, 10, 18-21.

Rosenberg, M. & Court, D. (1979).
Regulatory sequences involved in the promotion and termination of RNA transcription.
Annual Review of Genetics, 13, 19-53.

Sanderson, K.E. & Roth, J.R. (1983).
Linkage map of Salmonella typhimurium.
Microbiological Reviews, 47, 410-553.

Sanger, F. (1981).
Determination of nucleotide sequences in DNA.
Science, 214, 1205-1210.

Schevitz, R.W., Otwinowski, Z., Joachimiak, A., Lawson, C.L. & Sigler, P.B. (1986).
The three dimensional structure of the trp repressor.
Nature, 317, 782-786.

Schminke-Ott, E. & Bisswanger, H. (1980).
In: Multifunctional Proteins.
(Bisswanger, H. & Schmincke-Ott, E., eds.), pp.1-29, Wiley, New York.

Schweizer, E., Werkmeister, K. & Jain, M.K. (1978).
Fatty acid biosynthesis in yeasts.
Molecular and Cellular Biochemistry, 21, 95-106.

Schweizer, M., Roberts, L.M., Höltke, H-J., Takabayashi, K., Höllerer, E., Hoffmann, B., Muller, G., Köttig, H. & Schweizer, E. (1986).
The pentafunctional FAS1 gene of yeast: its nucleotide sequence and order of catalytic domains.
Molecular and General Genetics, 203, 479-486.

Senapathy, S. (1986).
Origin of eukaryotic introns: A hypothesis, based on codon distribution statistics in genes, and its implications.
Proceedings of the National Academy of Sciences (USA), 83, 2133-2137.

Shah, D.M., Horsch, R.B., Klee, H.J., Kishore, G.M., Winter, J.A., Tumer, N.E., Hironaka, C.M., Sanders, P.R., Gasser, C.S., Aykent, S., Siegel, N.R., Rogers, S.G. & Fraley, R.T. (1986).
Engineering herbicide tolerance in transgenic plants.
Science, 233, 478-481.

Shepherd, J.C.W. (1981).
Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification.
Proceedings of the National Academy of Sciences (USA), 78, 1596-1600.

Shine, J. & Dalgarno, L. (1975).
Determinant of cistron specificity in bacterial ribosomes.
Nature, 254, 34-38.

Shultz, J., Hermodson, M.A., Garner, C.C. & Herrmann, K.M. (1984)
The nucleotide sequence of the aroF gene of Escherichia coli and the amino acid sequence of the encoded protein, the tyrosine-sensitive 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase.
Journal of Biological Chemistry, 259, 9655-9661.

Silbert, D.F., Jorgensen, S.E. & Lin, E.C.C. (1962).
Repression of transaminase A by tyrosine in Escherichia coli.
Biochimica et Biophysica Acta. 73, 232-240.

Smith, D.D.S. (1980).
Ph.D. Thesis. University of Glasgow.

Smith, D.D.S. & Coggins, J.R. (1983).
Isolation of a bifunctional domain from the penta-functional arom enzyme complex of Neurospora crassa.
Biochemical Journal, 213, 405-415.

Smith, M.A., Gerrie, L.M., Dunbar, B. & Fothergill, J.E. (1982).
Primary structure of bovine complement activation fragment C4a, the third anaphylatoxin.
Biochemical Journal, 207, 253-260.

Srinivasan, P.R., Rothschild, J. & Sprinson, D.B. (1963).
The enzymic conversion of 3-deoxy-D-arabino-heptulosonat 7-phosphate to 5-Dehydroquinate.
Journal of Biological Chemistry, 238, 3176-3182.

Staden, R. (1978).
Further procedures for sequence analysis by computer.
Nucleic Acids Research, 5, 1013-1015.

Staden, R. (1980).
A new method for the storage and manipulation of DNA gel reading data.
Nucleic Acids Research, 8, 3673-3694.

Staden, R. (1984).
Computer methods to locate signals in nucleic acid sequences.
Nucleic Acids Research, 12, 505-519.

Stalker, D.M., Hiatt, W.R. & Comai, L. (1985).
A single amino acid substitution in the enzyme 5-enolpyruvylshikimate-3-phosphate synthase confers resistance to the herbicide glyphosate.
The Journal of Biological Chemistry, 260, 4724-4728.

Steinrucken, H.C. & Amrhein, N. (1980).
The herbicide glyphosate is a potent inhibitor of 5-enolpyruvylshikimate 3-phosphate synthase.
Biochemical and Biophysical Research Communications, 94, 1207-1212.

Strauss, A. (1979).
The genetic fine structure of the complex locus aro3 involved in early aromatic amino acid biosynthesis in Schizosaccharomyces pombe.
Molecular and General Genetics, 172, 233-241.

Sugimoto, S. & Shiio, I. (1980).
Purification and properties of bifunctional 3-deoxy-D-arabinoheptulosonate-7-phosphate/chorismate mutase component A from Brevibacterium flavum.
Journal of Biochemistry, 87, 881-890.

Takeda, Y., Nishimura, A., Nishimura, Y., Yamada, M., Yasuda, S., Suzuki, H. & Hirota, Y. (1981).
Synthetic ColE1 plasmids carrying genes for penicillin-binding proteins of E.coli.
Plasmid, 6, 86-98.

Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. & Gralla, J. (1973).
Improved estimation of secondary structures in ribonucleic acids.
Nature New Biology, 246, 40-41.

Tribe, D.E., Camakaris, H. & Pittard, J. (1976).
Constitutive and repressible enzymes of the common pathway of aromatic biosynthesis in E.coli K12: Regulation of enzyme synthesis at different growth rates
Journal of Bacteriology, 127, 1085-1097.

Twigg, A.J. & Sherratt, D. (1980).
Trans-complementable copy-number mutants of plasmid ColE1.
Nature, 283, 216-218.

Vapnek, D., Hautala, J.A., Jacobson, J.W., Giles, N.H. & Kushner, S.R. (1977).
Expression in E.coli K12 of the structural gene for catabolic dehydroquinase of N.crassa.
Proceedings of the National Academy of Sciences (USA), 74, 3508-3512.

Vogel, H.J. & Davis, B.D. (1952).
Glutamic semialdehyde and pyroline-5-carboxylic acid, intermediates in the biosynthesis of proline.
Journal of American Chemical Society, 74, 109-112.

Walker, J.E., Saraste, M., Runswick, M.J. & Gray, M.J. (1982).
Distantly related sequences in the α and β subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold.
EMBO Journal, 1, 945-951.

Wallace, B.J. & Pittard, J. (1967).
Genetic and biochemical analysis of the isoenzymes concerned in the first reaction of aromatic biosynthesis in Escherichia coli.
Journal of Bacteriology, 93, 237-244.

Wallace, B.J. & Pittard, J. (1969).
Regulator gene controlling enzymes concerned in tyrosine biosynthesis in Escherichia coli.
Journal of Bacteriology, 97, 1234-1241.

Warburg, R.J., Mahler, I., Tipper, D.J. & Halvorson, H.O. (1984).
Cloning of the Bacillus subtilis 168 aroC gene encoding dehydroquinase.
Gene, 32, 57-66.

Weiss, U. & Edwards, J.M. (1980).
The Biosynthesis of Aromatic Compounds.
New York: John Wiley & Sons.

Welch, G.R., Cole, K.W. & Gaertner, F.H. (1974).
Chorismate synthase of Neurospora crassa: a flavoprotein
Archives of Biochimistry and Biophysics, 165, 505-518.

Whipp, M.J. & Pittard, J. (1977).
Regulation of aromatic amino acid transport systems in Escherichia coli K12.
Journal of Bacteriology, 132, 453-461.

White, P.J., Millar, G. & Coggins, J.R. (1986).
Cloning and expression of the aroC gene from Escherichia coli K12: purification and characterisation of the gene product chorismate synthase.
Biochemical Journal, submitted.

Wierenga, R.K. & Hol, W.G.J. (1983).
Predicted nucleotide-binding properties of p21 protein and its cancer-associated variant.
Nature, 302, 842-844.

Yamamoto, E. (1980).
Purification and metal requirements of 3-dehydroquinate synthase from Phaseolus mungo seedlings.
Phytochemistry, 19, 779-781.

Yanofsky, C. (1981).
Attenuation in the control of expression of bacterial operons.
Nature, 289, 751-758.

Yourno, J., Kohno, T. & Roth, J.R. (1970).
Enzyme evolution: generation of a bifunctional enzyme by fusion of adjacent genes.
Nature, 228, 820-824.

Zakin, M.M., Duchange, N., Ferrara, P. & Cohen, G.H. (1983).
Nucleotide sequence of the metL gene of Escherichia coli.
Journal of Biological Chemistry, 258, 3028-3031.

Zalkin, H. (1980).
Anthranilate synthase: Relationships between bifunctional and monofunctional enzymes. In:
Multifunctional Proteins, ed. Bisswanger, H. & Schmincke Ott, E. Ch. 4, p.123-149. New York: John Wiley & Sons.

Zalkin, H. & Yanofsky, C. (1982).
Yeast gene TRP5: Structure, function, regulation.
Journal of Biological Chemistry, 257, 1491-1500.

Zalkin, H., Paluh, J.L., van Cleemput, M., Moye, W.S. & Yanofsky, C. (1984).
Nucleotide sequence of Saccharomyces cerevisiae genes TRP2 and TRP3 encoding bifunctional Anthranilate synthase: Indole-3-glycerol phosphate synthase.
Journal of Biological Chemistry, 259, 3985-3992.

Zurawski, G., Brown, K., Killingly, D. & Yanofsky, C. (1978).
Nucleotide sequence of the leader region of the phenylalanine operon of Escherichia coli.
Proceedings of the National Academy of Sciences (USA), 75, 4271-4275.

Zurawski, G., Gunsalus, R.P., Brown, K.D. & Yanofsky, C. (1981). Structure and regulation of aroH, the structural gene for tryptophan-repressible 3-deoxy-D-arabino-heptulosonate 7-phosphate of Escherichia coli. Journal of Molecular Biology, 145, 47-73.

Bolivar, F. & Backman, K. (1979) Plasmids of Escherichia coli as cloning vectors Methods in Enzymology, 68, 245-258.

Messing, J., Crea, R. & Seeburg, P.H. (1981) A method for shotgun sequencing DNA. Nucleic Acids Research, 9, 309-321.