STUDIES ON THE SHIKIMATE DEHYDROGENASE GENE OF *ESCHERICHIA COLI*

by

Ian Alexander Anton

Thesis submitted for the degree of Doctor of Philosophy in the Faculty of Science, University of Glasgow.

November 1985

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

ACKNOWLEDGEMENTS

I would like to thank:

John de Banzie, Michael McGarvey and John Subak-Sharpe for persuading me to do it; Robb Krumlauf and George Birnie for lessons in cloning DNA; Ian Hamilton and Charles Fewson for decreasing my ignorance about bacteria; Richard Hayward and David Meek for providing *E.coli* NO1267 together with much useful advice and information (and David for telling me about λ *spc*1 in the first place); Nat Harvey and Alec Brown for the prompt and skilful construction and repair of apparatus; Bryan Dunbar, John Fothergill, and Linda Fothergill for much assistance with the N-terminal amino acid sequencing; Jim Jardine for operating the amino acid analyser; Bob Eason for a great deal of advice and help with computing, and for setting up such a useful computing system within the department; Andrew Coulson for introducing us to the WISGEN package and a lot of help in using it; the University of Glasgow and the SERC for financial support.

I am very grateful to Lorna Peedle for typing this thesis with so much patience and skill.

I would like to thank everyone on D-floor throughout my stay for their help, encouragement, good humour and friendship. Special thanks are owed to Ken, Martin, and Ann.

I am indebted to my supervisor John Coggins for his cheerful and enthusiastic support throughout this project and for his advice, gentle criticism, and patience.

And Elizabeth, for many things.

ABBREVIATIONS

As well as those given in the Biochemical Journal's "Instructions to Authors" the following abbreviations were used:

A.a.	amino acid
ABF	anaerobic bacterial ferredoxin
ABR	anaerobic bacterial rubredoxin
A_x	absorbance at x nm measured with a 1 cm path
βME	2-mercaptoethanol
Da	daltons
DAHP	3-deoxy-D-arabino-heptulosonate-7-phosphate
DHQ	3-dehydroquinate
DHS	3-dehydroshikimate
DS	double stranded
DTT	dithiothreitol
E0	DAHP synthase
E1	DHQ synthase
E2	dehydroquinase
E3	shikimate dehydrogenase
E4	shikimate kinase
E5	EPSP synthase
E6	chorismate synthase
EPSP	5-enolpyruvylshikimate-3-phosphate
IPTG	isopropyl-β-D-thio-galactopyranoside
L Amp	L agar + ampicillin
L Tet	L agar + tetracycline

M Amp	M agar + ampicillin
MM	minimal medium
m.w.	molecular weight
NBT	nitro-blue tetrazolium
O/N	overnight (approximately 5 pm to 9 am)
ORF	open reading frame
PABA	p-aminobenzoate
PAGE	polyacrylamide gel electrophoresis
PEG	polyethylene glycol
PHBA	p-hydroxybenzoate
phe	L-phenylalanine
PMS	phenazine methosulphate
PMSF	phenylmethanesulphonyl fluoride
RF	replicative form
s.a.	specific activity
SA	shikimate
S-D	Shine-Dalgarno
SDS	sodium dodecyl sulphate
shik3P	shikimate-3-phosphate
SS	single stranded
trp	L-tryptophan
tyr	L-tyrosine
u	units of enzyme activity
X-gal	5-bromo-4-chloro-3-indolyl-β-galactoside

CONTENTS

LIST OF FIGURES

SUMMARY

Aromatic compounds are made via the shikimate pathway. The N.crassa pentafunctional arom enzyme has five shikimate pathway activities on one polypeptide whereas in E.coli all seven activities are separate enzymes. It has been hypothesised that arom arose by the fusion of genes for monofunctional enzymes. To test this proposal requires comparison of the sequences of arom and its monofunctional counterparts. Towards this future goal the author set out to sequence the E.coli aroE gene encoding shikimate dehydrogenase ("E3").

AroE was cloned from the previously isolated transducing phage λspc1 by selection in an aroE$^-$ auxotroph. The E3 overexpressed by strains carrying these clones is identical to wild-type E3 by native and SDS PAGE. A protocol was developed which permits the renaturation of E3 after SDS PAGE.

A 1.82 kbp region of DNA containing aroE was sequenced on both strands by the M13/dideoxy method. The open reading frame (ORF) corresponding to aroE was initially identified by size and by further subcloning.

A high level of E3 overproduction was obtained by placing the aroE gene in an expression vector. This gave E3 specific activities more than 300 times higher than in wild-type cells.

Overproduced E3 was purified to homogeneity using a previously developed method. 20g (wet weight) of cells yielded 10 mg of E3.

The E3 amino acid sequence deduced from the DNA sequence was confirmed by N-terminal amino acid sequencing and amino acid analysis of the overproduced E3.

Within the 1.82 kbp of DNA sequenced on both strands, together with the adjacent 0.6 kbp sequenced on only one strand, two large ORF's were found in addition to aroE ("UPSORF 1" and "UPSORF 2"). Biased codon utilisation and Fickett analysis hinted that UPSORF's 1 and 2 might be genes, as might be a truncated ORF ("UPSORF 3?") at the end of the single strand sequence. It remains to be seen if these UPSORF's encode proteins.

Strong indirect evidence that UPSORF 2 encodes a protein comes from its predicted amino acid sequence (180 a.a.'s) which contains four internal homologous sub-regions, suggesting internal gene duplication. In particular, each sub-region has two pairs of cysteine residues. Preliminary sequence comparisons indicate the possibility of very weak homologies between UPSORF 2 and some iron-sulphur proteins.

Truncation of sequences $\geqslant$1.4 kbp and $\geqslant$0.8 kbp upstream from the 5' end of aroE may be the cause of the observed 4-5 fold and 20-25 fold reductions (respectively) in E3 specific activity with some of the smaller aroE subclones (relative to the larger subclones). This may be indirect evidence that aroE is part of a "mixed" operon, with UPSORF's 1, 2, and 3, analogous to that found for aroA.

CHAPTER 1 INTRODUCTION

PART A - GENERAL INTRODUCTION

1.0 The problem

The reader, like the author and all other animals, cannot make aromatic compounds *de novo*. The work described in this thesis was not provoked by envy of plants, bacteria, and other microorganisms - all of which are equipped for the biosynthesis of aromatics - but rather by the striking differences between these organisms in the exercise of their synthetic abilities.

All aromatic compounds are made from the common precursor chorismate (reviewed by Haslam, 1974; Weiss and Edwards, 1980; Herrmann, 1983). The route to chorismate, which is known as the shikimate pathway or common (aromatic) pathway, is the same in all species so far studied and is shown in Figure 1.1. It is in the organisation of the enzyme activities which catalyse the seven universal reactions of the pathway that one sees great differences between species.

In the bacterium *Escherichia coli* (*E.coli*) all seven common pathway activities are separate enzymes (Berlyn and Giles, 1969) encoded by widely scattered genes (Pittard and Wallace, 1966). In contrast, in the fungus *Neurospora crassa* (*N.crassa*) five of the seven common pathway activities occur as a multifunctional protein - the *arom* multifunctional enzyme - consisting of two identical pentafunctional poly-

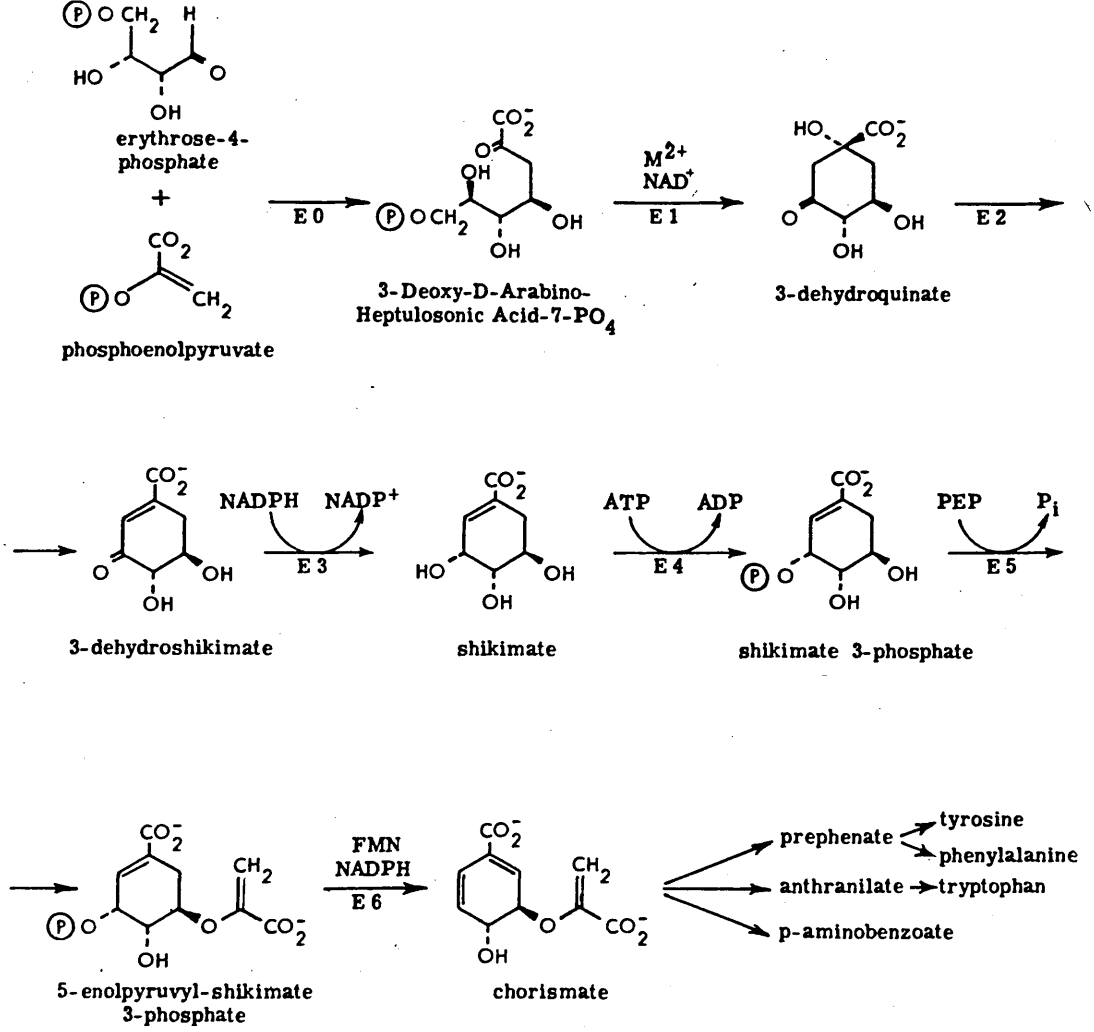


Figure 1.1 The shikimate pathway

Abbreviations used:

E0	DAHP synthase	EC 4.1.2.15
E1	DHQ synthase	EC 4.6.1.3
E2	dehydroquinase	EC 4.2.1.10
E3	shikimate dehydrogenase	EC 1.1.1.25
E4	shikimate kinase	EC 2.7.1.71
E5	EPSP synthase	EC 2.5.1.19
E6	chorismate synthase	EC 4.6.1.4
DAHP	3-deoxy-D-arabino-heptulosonate-7-phosphate	
DHQ	3-dehydroquinate	
DHS	3-dehydroshikimate	
Shik3P	shikimate-3-phosphate	
EPSP	5-enolpyruvylshikimate-3-phosphate	

peptide subunits encoded by a single gene (Lumsden and Coggins, 1977, 1978; Gaertner and Cole, 1977; Giles *et al.*, 1967a). This system thus provides a stark example of a phenomenon which is not well understood. For instance, why should a particular pathway involve multifunctional proteins in some species but not in others? Within one species, why are some functions handled by multifunctional proteins and not others? Such questions, and ramifications thereof, provided the main stimulus to the work described here. The shikimate pathway is also interesting in itself both for pure and applied reasons.

1.1 Format of the introduction

There follows a brief survey of various facets of the shikimate pathway. Attention is given to the organisation of the pathway in different species. The known properties of individual enzymes are summarised and progress made in their purification, and in cloning the corresponding genes, is described.

The latter part of the general introduction considers multifunctional proteins and the many questions they raise. A more detailed description of the *arom* multifunctional enzyme will be given here.

Part B specifically introduces the work described in this thesis.

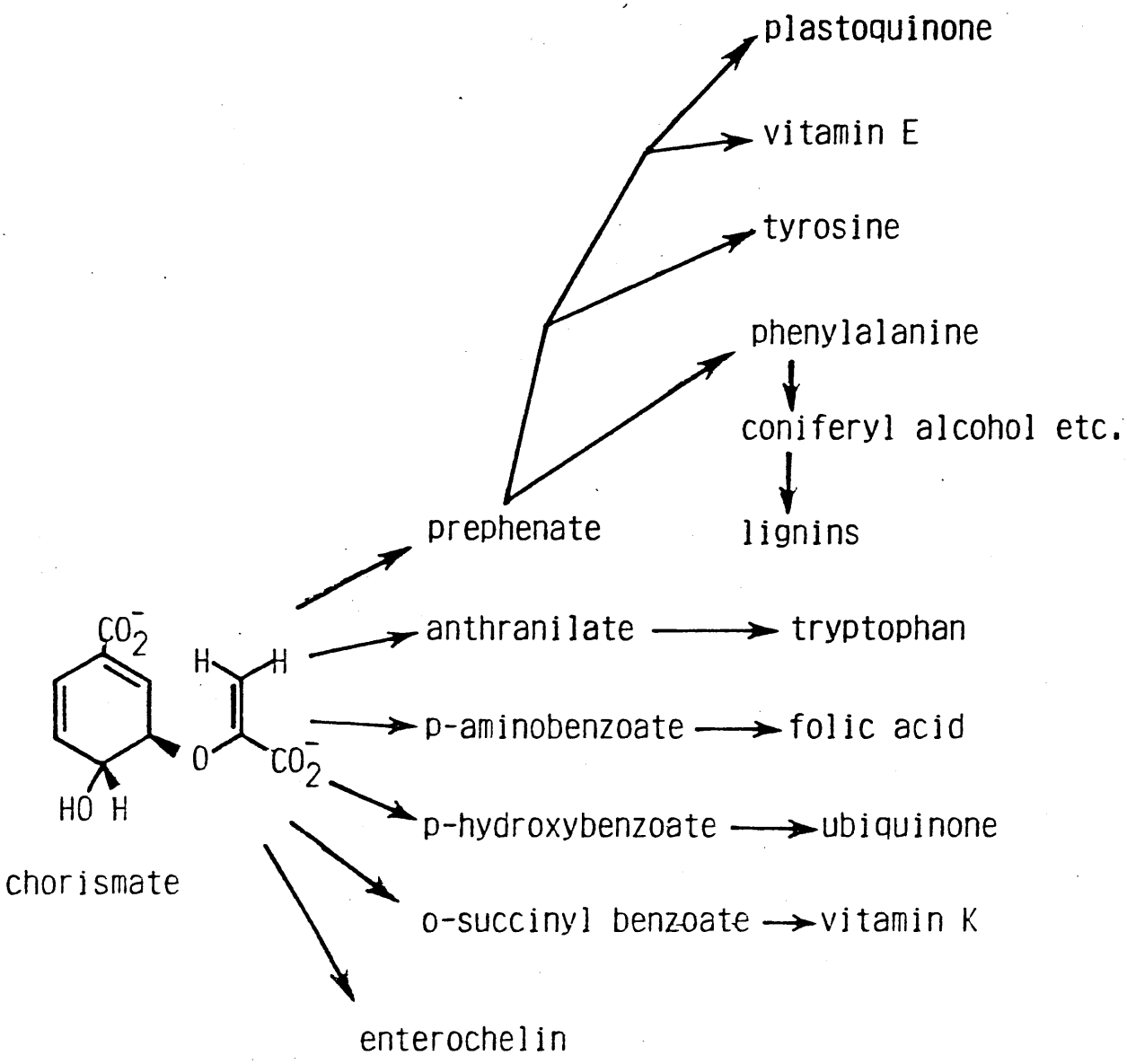


Figure 1.2 The utilisation of chorismate

1.2 Elucidation of the pathway

The deduction of the sequence of intermediates in the shikimate pathway was a long and intricate process in which key roles were played by the groups of B.D. Davis and D.B. Sprinson. Some of the decisive events are described below.

Davis (1951), using the then newly invented technique of penicillin enrichment, was the first to isolate auxotrophic mutants of E.coli which required supplementation with more than one aromatic end-product for growth. He demonstrated the central role of shikimate in the pathway. In a similar way, 3-dehydroquinate (DHQ) and 3-dehydroshikimate (DHS) were shown to be earlier precursors of shikimate (Davis and Weiss, 1953).

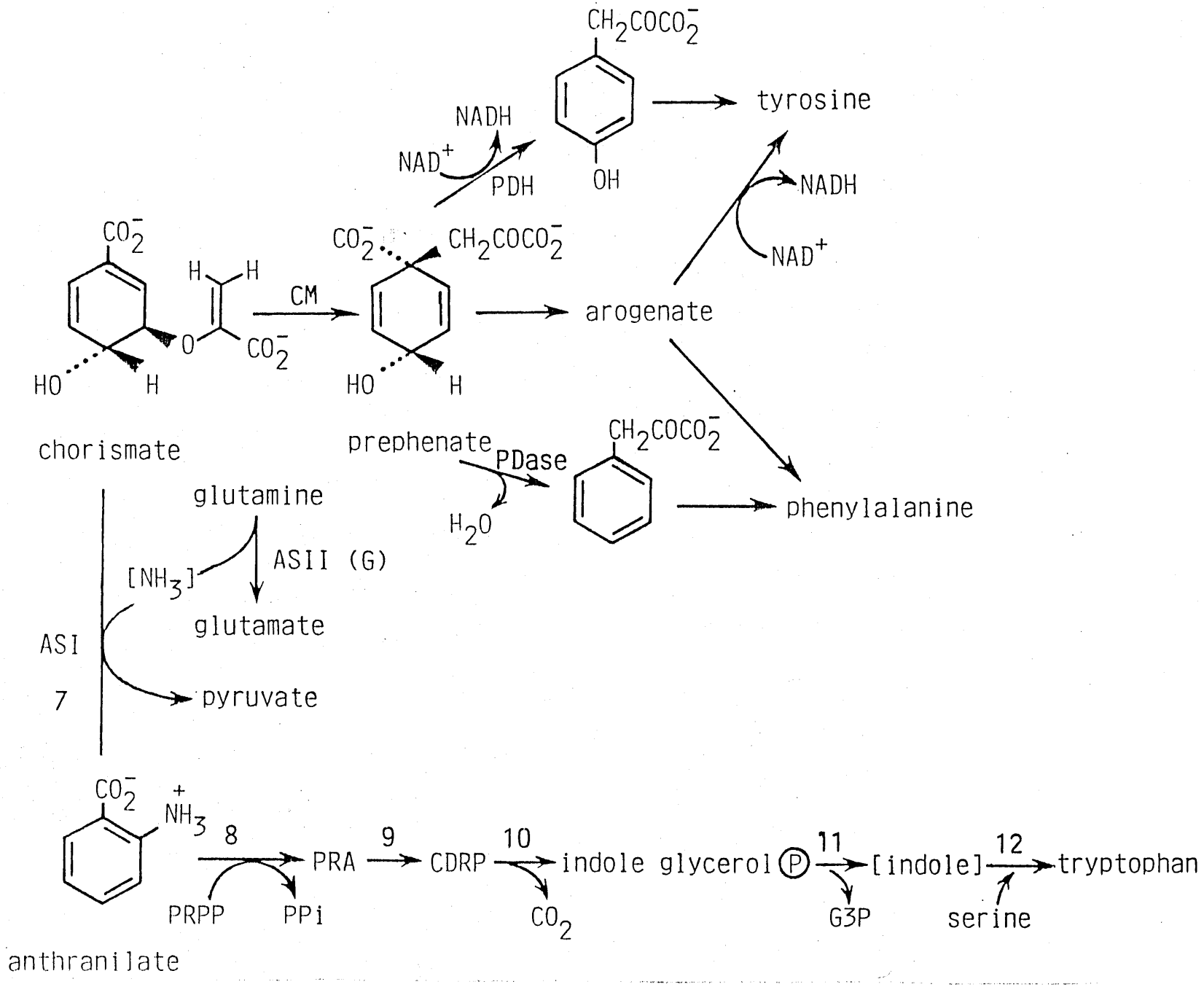
The enzymatic formation of 5-enolpyruvylshikimate-3-phosphate (EPSP) from shikimate-3-phosphate was demonstrated by Levin and Sprinson (1964) who also showed that there was a common pathway intermediate after EPSP and guessed correctly that it was chorismate (Levin and Sprinson, 1964; and references therein).

1.3 Utilisation of chorismate

Many pathways diverge from chorismate to give the end-products of aromatic biosynthesis and these are shown in Figure 1.2. There may be some minor end-products which are as yet undiscovered. The major end-products are the three aromatic amino acids phenylalanine, tyrosine, and tryptophan and the pathways to these are shown in Figure 1.3 (Umbarger, 1978). Phenylalanine is the precursor of lignins which are

Figure 1.3 Biosynthesis of phenylalanine, tyrosine, and tryptophan from chorismate

Enzymes	CM	chorismate mutase
	PDH	prephenate dehydrogenase
	PDase	prephenate dehydratase
	ASI(7)	anthranilate synthase, catalytic subunit
	ASII(G)	anthranilate synthase, glutaminase subunit
	8	anthranilate phosphibosyl transferase
	9	phosphoribosyl anthranilate isomerase
	10	indole glycerol phosphate synthase
	11	tryptophan synthase, step A
	12	tryptophan synthase, step B
Intermediates	PRA	phosphoribosyl anthranilate
	CDRP	(O-carboxyphenylamino)-1-deoxyribulose-5-phosphate
	PRPP	phosphoribosyl pyrophosphate
	G3P	glyceraldehyde-3-phosphate


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

some of the major structural polymers of wood tissue - one of the more abundant materials in the biosphere.

1.4 Organisation of the shikimate pathway in fungi

1.4.1 The arom multifunctional enzyme of N.crassa

1.4.1A Early genetic studies in N.crassa

Mutants of N.crassa which require all three aromatic amino acids as growth factors are known as arom mutants. They are defective in one or more enzyme activities of the common pathway and can be grouped into different classes by, for example, enzyme assays or by genetic complementation tests in heterokaryons.

Gross and Fein (1960) reported that arom mutations in at least three distinct complementation groups all mapped in a cluster on linkage group II of N.crassa. This work was extended by Giles et al. (1967a). Five distinct complementation groups of arom mutations all mapped in a very tight cluster at the "arom region". Each of these complementation groups is associated with the loss of one of the activities for steps 2 through 6 of the shikimate pathway (activities E1, E2, E3, E4 and E5; see Figure 1.1). Fine structure genetic mapping revealed that the five classes of simple mutations map in five discrete non-overlapping subregions of the arom cluster. As well as mutations associated with the loss of only one

activity other mutations were found causing the loss of two or more activities. These pleiotropic mutations map within the arom region and their complementation map shows a clear polarity (hence, "polarity mutants") implying that the arom cluster is transcribed as a single unit.

It was also discovered that the five activities which mapped to the arom locus all cosedimented rapidly during sucrose density gradient centrifugation, suggesting that the five activities were associated in a multienzyme complex (Giles et al., 1967a).

1.4.1B Purification of the N.crassa arom complex to homogeneity

Early attempts at purification did not immediately conflict with the idea of a multienzyme complex but gave confusing results (Burgoyne et al., 1969). Only with the advent of a new purification procedure, which took full precautions against endogenous proteases, did it become clear that the arom complex consisted of two identical 165 kDa pentafunctional subunits and was thus a multifunctional enzyme (Lumsden and Coggins, 1977, 1978; Gaertner and Cole, 1977).

Although the genetic data of Giles et al. (1967a) were originally interpreted in terms of separate genes these results are very useful when reinterpreted in terms of a multifunctional polypeptide. These data will be considered again in the section on arom structure.

1.4.2 Separability of the shikimate pathway enzymes in other fungi

Ahmed and Giles (1969) investigated the separability of the shikimate pathway enzymes in the following species of fungi:

(Phycomycetes)	Rhizopus stolonifer
	Phycomyces nitens
	Absidia glauca
(Ascomycetes, like N.crassa)	Aspergillus nidulans (A.nidulans)
(Basidiomycetes)	Coprinus lagopus
	Ustilago maydis

The activities which cosediment in extracts of N.crassa - E1, E2, E3, E4, and E5 - also sediment together on sucrose density gradients in extracts of all these six species of fungi. The sedimentation coefficients are all similar and comparable with that observed in the case of N.crassa.

At least three of the same five "arom" activities (E1, E2, and E3) in the budding yeast Saccharomyces cerevisiae (S.cerevisiae) also cosediment in sucrose gradients, with a sedimentation coefficient like that of the N.crassa arom complex (De Leeuw, 1967).

None of the above "aggregates" have been purified to homogeneity so it is not known whether they constitute penta-functional enzymes, as in N.crassa, or multienzyme complexes. However, other evidence (discussed below) suggests that at least in budding yeast the complex is a multifunctional protein.

1.4.3 Genetic organisation of the shikimate pathway in other fungi and cloning of arom loci

1.4.3A A.nidulans

The molecular cloning of a subset of the A.nidulans arom locus (which is very similar, if not identical, to that in N.crassa) has been reported (Kinghorn and Hawkins, 1982). Their recombinant plasmid was isolated by selecting for relief of auxotrophy in an aroD$^-$ E.coli strain defective for dehydroquinase (E2) activity, an approach which has been successful for a number of lower eukaryotic genes. The interpretation of this work is complicated by the existence in N.crassa and A.nidulans of a second, inducible, catabolic dehydroquinase activity (see Section 1.8.4B) which is heat-stable unlike the constitutive arom dehydroquinase activity. The gene for the N.crassa catabolic dehydroquinase (qa-2) was originally cloned by relief of auxotrophy in an aroD$^-$ E.coli mutant (Vapnek et al., 1977) so that the selection procedure used by Kinghorn and Hawkins (1982) could lead to the isolation not only of an arom clone but also a catabolic E2 clone. Unfortunately the criteria used to identify the particular A.nidulans recombinant obtained were both "negative" - the E2 activity produced was not heat-stable and did not react with antibodies which recognise A.nidulans catabolic dehydroquinase. Since the location of the cloned gene relative to the insert is unknown, and since slight truncation of a catabolic E2 gene could lead to loss of heat stability and loss of the relevant epitopes, there is

slight doubt whether part of the A.nidulans arom locus has actually been cloned. However, from the restriction map their clone definitely does not contain the E.coli aroD$^+$ gene (Kinghorn et al., 1981).

1.4.3B N.crassa

A preliminary report has been published of the cloning of that portion of the N.crassa arom gene which specifies the biosynthetic dehydroquinase activity (Catcheside and Storer, 1984).

1.4.3C S.cerevisiae

In the budding yeast S.cerevisiae a complex locus specifying shikimate pathway activities was identified by De Leeuw (1967). He found four complementation groups in a preliminary study, each corresponding to a deficiency in one of the common pathway activities E1, E3, E4, and E5. Mutations in all four complementation groups are closely linked in a cluster, the "arom-1 locus", as in N.crassa. The finding of various classes of polarity mutant which affect two, three, or four activities, and which can be arranged in a unique linear order, is also very reminiscent of the situation in N.crassa. Furthermore, completely noncomplementing mutants were isolated and found to lack E2 activity in addition to the other four activities.

From the above evidence and the cosedimentation data it appears likely that S.cerevisiae might possess a pentafunctional

arom polypeptide. Further support for this idea comes from the cloning of the S.cerevisiae arom-1 locus, also called the "ARO1 cluster gene" (Larimer et al., 1983). The ARO1 clone was isolated selectively by its ability to complement an appropriate yeast mutant. The insert fragment from the original recombinant plasmid was subcloned to yield a 6.2 kbp fragment which still retained full complementation activity. Since the N.crassa arom polypeptide contains about 1500 amino acid residues (Lumsden and Coggins, 1978) it is rather unlikely that the 6.2 kbp fragment contains five separate genes. Moreover, when present as an episome in yeast cells the cloned locus results not only in the overexpression of all five activities but also of an unidentified polypeptide identical in size (by SDS PAGE) to N.crassa arom.

The 6.2 kbp fragment carrying ARO1 is present being sequenced in this laboratory (K. Duncan, unpublished work) and the results are eagerly awaited.

1.4.3D Schizosaccharomyces pombe

In the fission yeast Schizosaccharomyces pombe (S.pombe), which is only distantly related to the budding yeast S.cerevisiae, there is a complex locus aro3 containing five genetically defined subregions A-E, in that linear order (Strauss, 1979). These subregions all apparently affect shikimate pathway enzymes, from the growth factor requirements of aro3 alleles, but a specific biochemical defect has only been assigned to mutations in three of them - A: $E1^-$, C: $E4^-$, E: $E3^-$. From

the polarity of the complementation patterns characterising the nonsense aro3 alleles it was inferred that the direction of transcription of the aro3 locus, whether it encoded a multifunctional protein or a polycistronic mRNA for several proteins, is from subregion A to E. This implies that the order and spacing of the three assigned activities within the aro3 locus is the same as that in the N.crassa arom locus (5'-E1, E5, E4, E2, E3-3'; see Section 1.10) but this does not provide evidence that aro3 encodes a multifunctional polypeptide. However, the aro3 locus clearly resembles the N.crassa arom locus very closely at the genetic level.

Most, if not all, of the aro3 locus of S.pombe has been isolated as a set of overlapping cloned DNA fragments by Nakanishi and Yamamoto (1984). They also demonstrated that E2 is encoded by aro3 and probably by subregion D, thus extending the analogy with N.crassa arom. They discovered that the whole locus is represented by a single 4.5 kb mRNA transcript. Although it seems likely that aro3 encodes a pentafunctional polypeptide the aggregation state of the shikimate pathway enzymes has never been scrutinised in this species. Also, it has yet to be established that E5 is encoded by aro3.

1.5 Organisation of the shikimate pathway in bacteria

1.5.1 Separability of the shikimate pathway enzymes in bacteria

The five common pathway activities E1-E5 sediment independently during sucrose density gradient centrifugation

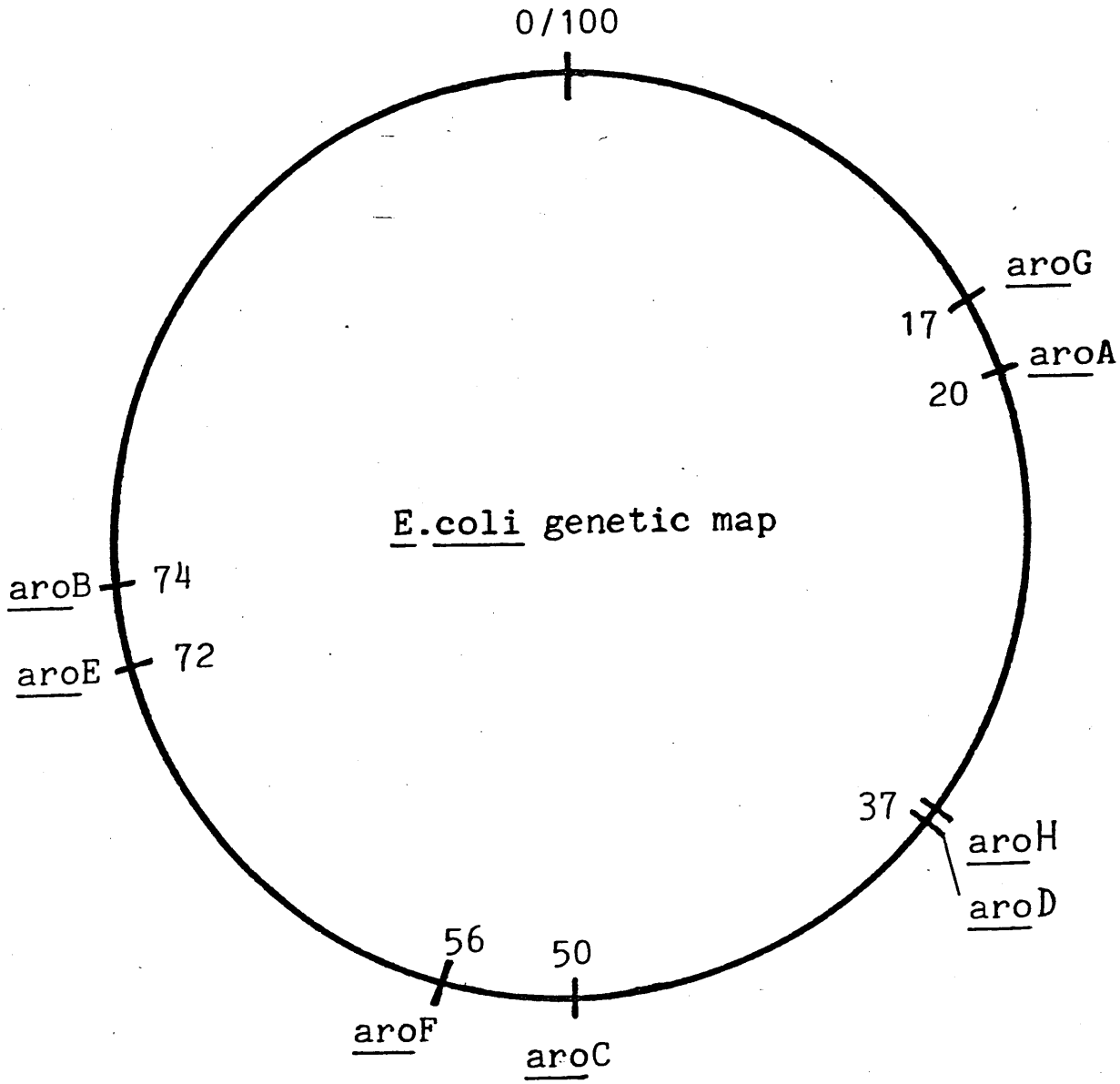
of extracts from the following six species of bacteria: E.coli, Salmonella typhimurium (S.typhimurium), Aerobacter aerogenes, Bacillus subtilis (B.subtilis), Pseudomonas aeruginosa, and Streptomyces coelicolor (Berlyn and Giles, 1969). Two peaks of shikimate kinase activity were identified in E.coli and S.typhimurium.

Although it is clear that these bacteria do not contain multifunctional arom enzymes (or subsets thereof) or stable multienzyme complexes it would perhaps be premature to discard the possibility that some of them might contain weakly bound multienzyme complexes that do not survive normal extraction procedures (Fulton, 1982).

1.5.2 Genetic organisation of the shikimate pathway in E.coli

The common pathway enzymes in E.coli are encoded by widely scattered genes (Pittard and Wallace, 1966; Bachmann, 1983). The distribution of the genes is shown in Figure 1.4 and a very similar distribution is seen in S.typhimurium (Gollub et al., 1967). There are no operons which contain more than one shikimate pathway gene.

The failure to isolate multiple aromatic auxotrophs of E.coli defective for shikimate kinase, together with the finding of two distinct shikimate kinase activities in E.coli (Berlyn and Giles, 1969), suggested that there are (at least) two E4 genes in this species. This appears to be the case and ingenious procedures eventually allowed the isolation of



Gene	Map position (min)	Enzyme encoded	
aroF	56	EO(Tyr)	DAHP synthase (Tyr sensitive)
aroG	17	EO(Phe)	DAHP synthase (Phe sensitive)
aroH	37	EO(Trp)	DAHP synthase (Trp sensitive)
aroB	74	E1	DHQ synthase
aroD	37	E2	dehydroquinase
aroE	72	E3	shikimate dehydrogenase
aroA	20	E5	EPSP synthase
aroC	50	E6	chorismate synthase

Figure 1.4 Positions of genes for shikimate pathway enzymes on the E.coli chromosome

strains with mutations defining aroL, the structural gene for the isozyme E4II. The aroL gene maps at about 9 minutes on the E.coli chromosome (Ely and Pittard, 1979). The location of the gene for the E4I isozyme is a mystery.

Although at least one gene is known for every shikimate pathway enzyme in E.coli there is another gene, aroI, defined by a temperature-sensitive mutant which requires phenylalanine, tyrosine, and tryptophan for growth at 42°C (Gibson and Pittard, 1968). The aroI gene maps at 84 minutes on the E.coli chromosome and its function is unknown. M. Boocock has suggested (personal communication) that it might encode a diaphorase subunit for the chorismate synthase activity (see Section 1.8.8).

The purification and properties of individual E.coli shikimate pathway enzymes and the cloning of their genes will be considered in Section 1.8.

1.6 Organisation of the shikimate pathway in photosynthetic organisms

1.6.1 Separability of the shikimate pathway enzymes in photosynthetic organisms

Berlyn et al. (1970) studied whether the enzyme activities E1-E5, from the photosynthetic organisms listed below, could be separated on sucrose density gradients:

Anabaena variabilis (prokaryotic blue-green algae)

Chlamydomonas reinhardi (unicellular green flagellate)

Euglena gracilis (unicellular green flagellate)

Physcomitrella patens (P.patens; moss - a lower plant)

Nicotiana tabacum (higher plant)

In Anabaena all five activities were separable, as in other prokaryotes. In Euglena all five activities cosediment as a large aggregate. In the other three species E2 and E3 cosediment but E1, E4, and E5 are quite separate.

The complex of the five activities in Euglena was subsequently purified 2000-fold (Patel and Giles, 1979). Its native m.w. appears to be slightly less than that of N.crassa arom and it is still not known whether this is a multifunctional enzyme or a multi-enzyme complex or a combination of both.

Boudet and Lecussan (1974) examined the separability of E2 and E3 (they did not look at other activities) in a variety of higher plants including Zea mays, Pisum sativum, Phaseolus vulgaris, and Triticum sativum. They used a range of enzyme purification techniques and in all cases E2 and E3 copurified.

Koshiba (1979) showed that in Phaseolus mungo (mung beans) E1, E4, and E5 are all separable by ion-exchange chromatography and gel filtration but E2 and E3 are not.

1.6.2 Purification of multifunctional shikimate pathway enzymes from plants

It appears that in many plants E2 and E3 are associated in some way. The manner of this association was first established in the moss P.patens. From this lower plant Polley (1978) purified to homogeneity a monomeric bifunctional polypeptide having both E2 and E3 activities which copurified in constant ratio. The molecular weight of this bifunctional

protein is 48 kDa as estimated by SDS PAGE.

In the higher plant P.sativum (peas) the E2, E3 complex has been purified to homogeneity and is a monomeric bifunctional enzyme of m.w. 60 kDa (M.S. Campbell, unpublished work).

The purification and properties of various monofunctional plant enzymes from the shikimate pathway will be described in Section 1.8.

1.6.3 Genetic organisation of the shikimate pathway in plants

Almost nothing is known about this aspect. In the Triticinae the structural genes for the two isozymic forms of E3 activity are located on the nuclear chromosomes (Koebner and Shepherd, 1982) thus dashing hopes that they might have been conveniently "packaged" in the chloroplast genome to the advantage of gene cloners.

1.7 Organisation of the five "arom" activities in fungi, bacteria, and photosynthetic organisms - a summary

Only a small sample of species from the three groups - fungi, bacteria, and photosynthetic organisms - has been looked at, and only in a few cases has there been detailed characterisation of some of the shikimate pathway enzymes. However, bearing these points in mind, there is a clear trend in the results. The five activities found in the N.crassa arom enzyme are also associated in other fungi and at least in S.cerevisiae there is strong circumstantial evidence for an arom-like enzyme. In bacteria and blue-green algae the

same five activities are all separate enzymes, whereas in plants and Chlamydomonas reinhardi - but not in Euglena which in this resembles the fungi more closely - only E2 and E3 are associated, probably as a bifunctional enzyme as has been demonstrated in two cases.

1.8 A survey of the seven shikimate pathway catalytic activities

1.8.1 Preliminary remarks

This section considers the individual reactions of the shikimate pathway. Work on the purification of monofunctional shikimate pathway enzymes and the cloning of their genes will also be described here. Attention will be drawn to the general similarities between particular activities whether they occur in mono- or multifunctional enzymes, a feature of relevance to later discussion, and the practical importance of studies on the pathway will become apparent.

1.8.2 DAHP synthase (E0) and the regulation of the common pathway

This area has been well reviewed by Herrmann (1983). In E.coli carbon flow through the common pathway is regulated at the first step but chorismate, the immediate end-product, is not a feedback inhibitor in this species. E.coli possesses three DAHP synthase (E0) isozymes each sensitive to feedback inhibition (and repression) by one of the end-products of the

major terminal pathways - phenylalanine, tyrosine and tryptophan (Doy and Brown, 1965). These isozymes are designated EO(Phe), EO(Tyr), and EO(Trp) and have all been purified to homogeneity (Herrmann, 1983). E1, E2, E3, E4I, E5, and E6 are all constitutively expressed and they are not regulated at the protein level by any of the three aromatic amino acids, nor by chorismate or DAHP (Tribe *et al.*, 1976). The *aroL* gene, encoding the second shikimate kinase isozyme E4II, possibly represents a second point of control since its expression is affected by the *tyrR* locus (Ely and Pittard, 1979).

The *aroF* gene for EO(Tyr) has been cloned (Zurawski *et al.*, 1978) and sequenced (Shultz *et al.*, 1984). Davies and Davidson (1982) cloned and sequenced the *aroG* gene for EO(Phe). Similarly, the EO(Trp) gene, *aroH*, has been cloned and sequenced (Zurawski *et al.*, 1981).

In addition to being regulated by feedback inhibition at the enzyme level the DAHP synthase activities in *E.coli* are also regulated at the transcriptional level by feedback repression (Umbarger, 1978; Herrmann, 1983). The expression of *aroG* is controlled by the levels of phenylalanine and tryptophan which act via the *tyrR* gene product. The expression of the "tyr" operon, which includes *aroF*, is regulated by the levels of tyrosine (and possibly of phenylalanine), also via the *tyrR* protein. The level of tryptophan controls the transcription of *aroH* via the *trp* repressor. Attenuation

has been ruled out as an additional form of control for aroG and aroH but not completely for aroF.

From ^{13}C nuclear magnetic resonance studies on whole E.coli cells it appears that feedback inhibition is quantitatively the major regulatory mechanism for the shikimate pathway (Ogino et al., 1982).

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

B.subtilis has only a single DAHP synthase. It has been purified to homogeneity from strain 168 in which EO and chorismate mutase form a bifunctional enzyme which in turn is part of a complex that includes noncovalently bound shikimate kinase (Herrmann, 1983). The EO is insensitive to inhibition by aromatic amino acids but is inhibited by chorismate and prephenate. There is evidence that the bifunctional EO-chorismate mutase arose from a monofunctional EO during mutagenesis of the wild-type Marburg strain of B.subtilis (Llewellyn et al., 1980). It appears that in strain 168 the wild-type gene for chorismate mutase has been lost and its function replaced inefficiently by a mutated prephenate allosteric binding site of the original EO.

1.8.3 DHQ synthase (E1)

The *E.coli aroB* gene, which encodes DHQ synthase, has been cloned (Duncan and Coggins, 1983). Subsequently, other workers placed *aroB* in the expression vector pKK223-3 (see Section 4.2.2) giving a strain which overproduces DHQ synthase 1000-fold, thus facilitating the purification of this enzyme to homogeneity (Frost *et al.*, 1984). 50% pure E1 was also obtained from wild-type *E.coli* and the major band on SDS PAGE of this material corresponded exactly to the pure over-expressed E1 (Frost *et al.*, 1984). The specific activity of the overproduced E1 is twelve times greater than that observed previously for "pure" *E.coli* E1 by Maitra and Sprinson (1978). Frost *et al.* (1984) found E1 to be a monomer of 40 kDa whereas Maitra and Sprinson (1978) estimated the subunit m.w. to be 57 kDa. These serious discrepancies must cast considerable doubt on the earlier work of Maitra and Sprinson. However, the observations (Maitra and Sprinson, 1978) that *E.coli* E1 requires Co^{2+} and catalytic NAD^{+} for full activity are not in dispute.

The *aroB* gene has been sequenced in this laboratory and confirmation obtained by N-terminal amino acid sequencing and determination of the amino acid composition of the over-produced polypeptide (G. Millar, unpublished results). The sequence predicts a m.w. of 39 kDa for the E1 protein.

DHQ synthase has been purified to homogeneity from *B.subtilis* where it is part of a multienzyme complex with

chorismate synthase and the associated flavin reductase (Hasan and Nester, 1978b; see Section 1.8.8). The m.w. of the E1 subunit is 17 kDa and, like the E.coli enzyme, it requires a divalent transition metal cation (Co^{2+} or Mn^{2+}) and catalytic quantities of NAD^+.

Yamamoto (1980) purified the E1 from mung beans to homogeneity. It too requires NAD^+ and a divalent transition metal cation.

The DHQ synthase activity of the N.crassa arom multifunctional enzyme requires catalytic NAD^+ and Zn^{2+} for activity thus resembling the monofunctional E1's described above (Lambert et al., 1985).

1.8.4 Dehydroquinase (E2)

1.8.4A Biosynthetic dehydroquinases

The monofunctional E.coli dehydroquinase has been purified to homogeneity from wild-type K12 cells (S. Chaudhuri and J.R. Coggins, submitted to Biochem. J.). It is a dimer and the subunit m.w. deduced from the DNA sequence (see below) is 28 kDa. The enzyme can be inhibited by treatment with borohydride in the presence of substrate which strongly suggests that a Schiff's base intermediate is formed between an active site lysine residue and the carbonyl group of the substrate. This is known to be a feature of two other dehydroquinases described below.

The E.coli aroD gene which encodes E2 was originally cloned by Kinghorn et al. (1981). It was later subcloned and sequenced by Duncan (1985). The sequence has been confirmed by purification of the overproduced polypeptide (S. Chaudhuri, unpublished work) and determination of the amino acid composition and the N-terminal amino acid sequence (M.S. Campbell, unpublished results).

Little is known about the E2 activities of the purified bifunctional plant E2, E3 enzymes.

Like the E.coli E2, the dehydroquinase activity of N.crassa arom can be specifically inhibited by treatment with sodium borohydride in the presence of substrate, again implying a Schiff's base intermediate (Smith and Coggins, 1983). This "substrate trapping" has been exploited to radiolabel (using tritiated borohydride), isolate, and sequence an active site peptide which contains the essential lysine residue of arom E2 (S. Chaudhuri and J.R. Coggins, submitted to Biochem. J.).

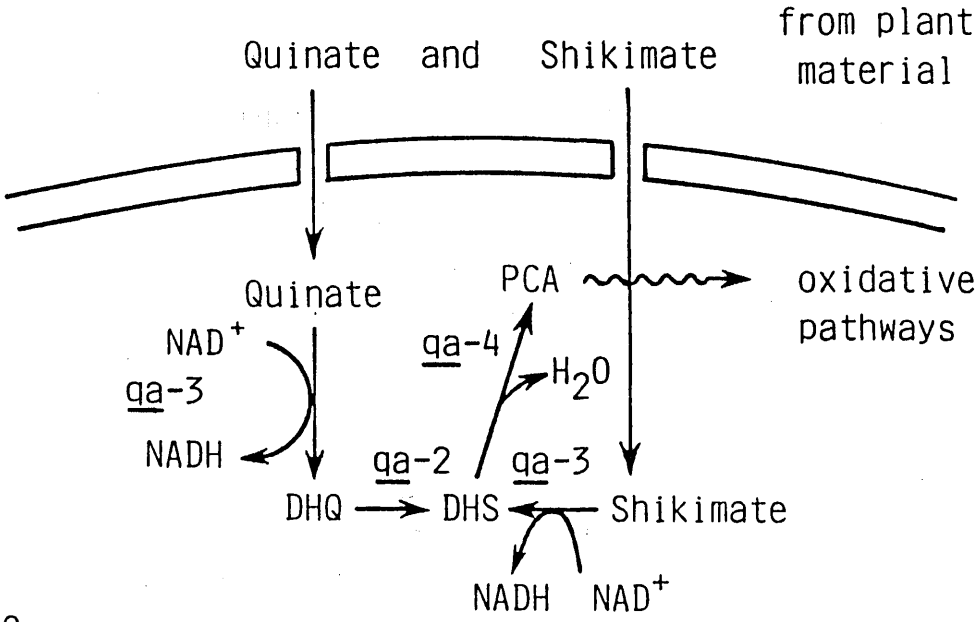
1.8.4B Catabolic dehydroquinases

The arom E2 is not the only dehydroquinase activity in N.crassa. There is also an inducible, catabolic dehydroquinase - the qa-2 gene product - which forms part of a degradative pathway for the utilisation of quinate (Giles et al., 1967b; Chaleff, 1974). N.crassa can grow on the plant product quinate as sole carbon source and shikimate can be used similarly. This catabolic pathway is shown in Figure 1.5. The enzymes

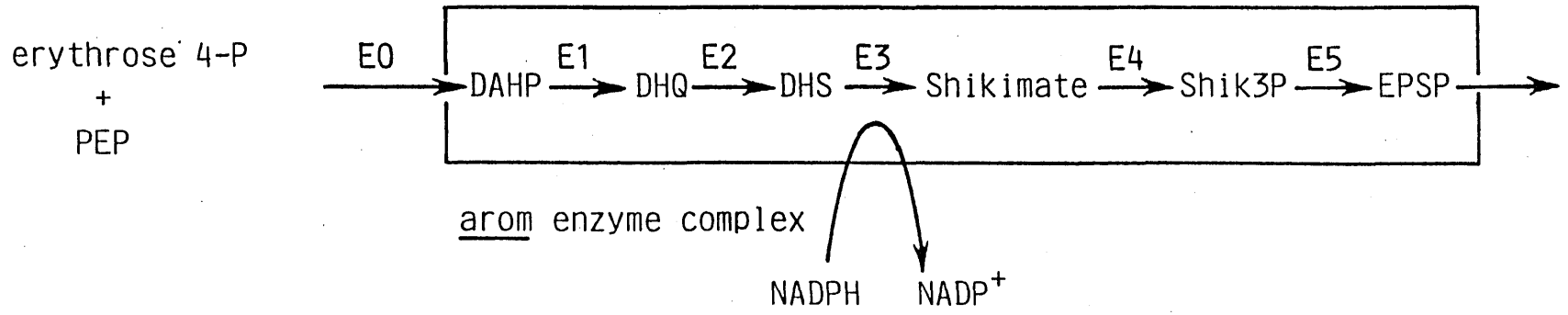
Figure 1.5 The catabolic quinate pathway and the biosynthetic shikimate pathway in *N. crassa*

Intermediates:	DHQ	dehydroquinate
	DHS	dehydroshikimate
	PCA	protocatechuate
Genes and enzymes:	*qa*-2	catabolic dehydroquinase
	qa-3	catabolic quinate/shikimate dehydrogenase
	qa-4	dehydroshikimate dehydratase

Catabolic quinate pathway



Biosynthetic shikimate pathway



of the pathway are all separate monofunctional activities. However, the separate genes all lie together in the "qa cluster" and they are coordinately expressed as monocistronic transcripts. Induction by substrates is mediated by the qa-1 gene products. The whole qa cluster has been cloned (Schweizer et al., 1981) and this system is now an important model for transcriptional regulation in eukaryotes (e.g. Huiet, 1984).

The catabolic dehydroquinase has been purified to homogeneity (S. Chaudhuri and J.R. Coggins, submitted to Biochem. J; Chaudhuri and Coggins, 1981). Early attempts at purification yielded material which was proteolytically degraded (Hautala et al., 1975). The enzyme appears to be a dodecamer of 20 kDa subunits and, like other E2's, is inhibited by borohydride in the presence of substrate (Chaudhuri and Coggins, 1981). The amino acid sequence has been revealed by the sequence of the qa-2 gene (Alton et al., 1982; M.E. Case, personal communication) and there is no readily detectable homology at the protein level between the E.coli aroD sequence and the N.crassa qa-2 sequence (Duncan, 1985).

In corn seedlings there are two E2 activities, one associated with a shikimate dehydrogenase and the other with a quinate dehydrogenase (Graziana et al., 1980). It is possible that these are both bifunctional enzymes and that the second E2 plays a catabolic role.

1.8.5 Shikimate dehydrogenase (E3)

1.8.5A E.coli shikimate dehydrogenase

Shikimate dehydrogenase (dehydroshikimate reductase; shikimate:$NADP^+$ oxidoreductase; E3) catalyses the NADPH-linked reduction of 3-dehydroshikimate. This reaction was first detected in partially purified extracts of E.coli (Yaniv and Gilvarg, 1955). The K'_{eq} in the direction of shikimate is about 28 at pH 7.0. Quinate is not a significant substrate for E.coli E3 nor can NAD^+ be substituted as cofactor. E.coli E3 transfers the hydrogen atom from the A side of the reduced nicotinamide ring (Dansette and Azerad, 1974).

The only monofunctional biosynthetic shikimate dehydrogenase which has been purified to homogeneity is that from E.coli (Chaudhuri and Coggins, 1985). The purification factor required was 20,000 fold. It is a monomer of m.w. 29 kDa. There are very few examples known of monomeric $NAD(P)^+$-linked dehydrogenases. $NADP^+$-linked dihydrofolate reductase from a variety of species is monomeric (Volz et al., 1982) as is the catabolic NAD^+-linked quinate/shikimate dehydrogenase of N.crassa (Barea and Giles, 1978; see below).

The E.coli aroE gene, which encodes E3, has been cloned and sequenced (Anton and Coggins, 1983; this study) and the DNA sequence has been confirmed by analysis of the overproduced polypeptide (this study).

1.8.5B The catabolic quinate/shikimate dehydrogenase of N.crassa

The inducible, catabolic quinate/shikimate dehydrogenase of N.crassa has been purified to homogeneity (Barea and Giles, 1978; see Figure 1.5). It is the product of the qa-3 gene (Chaleff, 1974) and is a monomer. Its activities as a shikimate dehydrogenase and as a quinate dehydrogenase are of similar magnitude, and it has a preference for NAD^+ over $NADP^+$ as cofactor, unlike all known biosynthetic E3's. There is almost certainly only a single active site from inhibition studies (Barea and Giles, 1978) and from genetic evidence (Chaleff, 1974). The qa-3 gene has been sequenced and predicts a m.w. of 35 kDa for the gene product (Alton et al., 1982; M.E. Case, personal communication). The question of possible homologies between qa-3 and aroE will be discussed in Chapter Five.

1.8.5C Plant shikimate dehydrogenases

The association between E2 and E3 in higher plants has been described in Section 1.6.1 and the purification to homogeneity of two bifunctional plant E2, E3 enzymes was outlined in Section 1.6.2.

Balinsky et al. (1971) and Dowsett et al. (1971) carried out detailed kinetic studies on the partially purified E3 from peas. It appears to have an ordered sequential mechanism in which $NADP^+$ or NADPH binds first to the enzyme. The enzyme from peas, like that from E.coli, shows A-stereospecificity

for hydrogen transfer (Davies *et al.*, 1972). There have been unsuccessful attempts to use analogues of dehydroshikimate as herbicides (Baillie *et al.*, 1972).

1.8.6 Shikimate kinase (E4)

The existence of at least two isozymes of *E.coli* E4 was discussed in Section 1.5.2. Both forms appear to have native molecular weights of about 20 kDa (Ely and Pittard, 1979).

The *aroL* gene encoding E4II has recently been cloned and sequenced. The overexpressed polypeptide has been purified to homogeneity and the sequence confirmed by determination of the amino acid composition and the N-terminal amino acid sequence (G. Millar, A. Lewendon, M. Hunter, and J.R. Coggins, submitted to Biochem. J.). The sequence of E4II contains the faint homologies found between other ATP utilising enzymes by Walker *et al.* (1982).

1.8.7 EPSP synthase (E5)

E.coli E5 was the first monofunctional EPSP synthase to be purified to homogeneity (Lewendon and Coggins, 1983). It is a monomeric enzyme and the m.w. deduced from the sequence (see below) is 46 kDa. The monofunctional E5 from pea seedlings has also been purified to homogeneity and it too is a monomer with an estimated m.w. of 50 kDa (Mousdale and Coggins, 1984).

The E.coli aroA gene encoding E5 has been cloned and homogeneous enzyme can now be obtained easily in milligram quantities (Duncan and Coggins, 1983; Duncan et al., 1984a). The aroA gene has been sequenced and the sequence confirmed by analysis of the overproduced polypeptide (Duncan et al., 1984b). The aroA gene is part of an operon which also includes the gene, serC, for phosphoserine aminotransferase. This is the first mixed operon discovered in E.coli (Duncan, 1985; see Section 5.8 for further discussion of this topic).

In plants, E5 is the target of the enormously successful and relatively benign herbicide glyphosate (N-phosphonomethyl-glycine) which is a competitive inhibitor with respect to phosphoenolpyruvate. The first indications that E5 might be the site of action of this compound came from the finding that the EPSP synthases of Aerobacter aerogenes (Amrhein et al., 1980) and of N.crassa (Boocock and Coggins, 1983) were highly sensitive to inhibition by glyphosate. That this was actually the site of action in vivo was confirmed, at least in bacteria, by the finding that either a mutated aroA gene (Comai et al., 1983) or overproduction of E5 (Duncan et al., 1984a) could make bacteria much less sensitive to glyphosate. The E5 from pea seedlings is an order of magnitude more sensitive to glyphosate than the E5 from E.coli or from N.crassa (Duncan et al., 1984a).

1.8.8 Chorismate synthase (E6)

Chorismate synthase catalyses what is probably the least understood step of the shikimate pathway. In particular, the requirement for reduced flavin nucleotides is mysterious.

E.coli E6 has been partially purified and some of its requirements for activity have been studied (Morell *et al.*, 1967). It was found to be very sensitive to O_2 and was most readily assayed under an H_2 or N_2 atmosphere in the presence of a reduced flavin adenine dinucleotide regenerating system.

The gene *aroC* encodes at least the catalytic subunit of *E.coli* chorismate synthase and is situated at about 50 minutes on the genetic map. Recombinant plasmids have been identified, from the Clarke and Carbon (1979) *E.coli* genomic library (see Part B of the introduction), which can complement *E.coli* *aroC*$^-$ mutants. This has been done both by direct selection from the entire library (Hagervall and Bjork, 1984), and by testing of a recombinant plasmid, pLC33-1, known to carry the *fabB* gene, a flanking marker of *aroC* (I. Anton, unpublished work). More definitive identification of the complementing gene has come from the subcloning of complementing fragments of pLC 33-1 into the high copy number plasmid vector pAT153 and the demonstration that *E.coli* strains carrying these subclones overexpress E6 20-40 fold (Millar *et al.*, 1985). Clearly much work remains to be done on the *E.coli* chorismate synthase.

E6 from B.subtilis has been purified to homogeneity (Hasan and Nester, 1978b). It is part of a complex which includes E1 (see Section 1.8.3) and a flavin reductase ("diaphorase") subunit required for E6 activity. The subunit m.w.'s of the three components of the complex are 13 kDa (flavin reductase), 24 kDa (E6), and 17 kDa (E1). The flavin reductase has also been purified to homogeneity in a dissociated form (Hasan and Nester, 1978a). The flavin reductase is specific for NADPH and requires M^{2+} and either FMN or FAD for maximal rates of NADPH oxidation.

Preliminary work suggests that the N.crassa chorismate synthase can now be purified to homogeneity (M. Boocock and P.J. White, unpublished results).

1.9 Multifunctional proteins

1.9.1 Definition and distribution

Kirschner and Bisswanger (1976) defined multifunctional proteins as those proteins in which more than one distinct biochemical function is located on a single polypeptide chain. The field has been reviewed (Kirschner and Bisswanger, 1976; Bisswanger and Schmincke-Ott, 1980). Enzymes with broad substrate specificity are excluded from the definition. Multifunctional enzymes must clearly be distinguished from multienzyme complexes, like pyruvate dehydrogenase, which contain several kinds of subunit each carrying out different functions. Fatty acid synthase in fungi is both a multifunctional protein and a multienzyme complex since it comprises

two different kinds of multifunctional subunits (McCarthy and Hardie, 1984). Non-catalytic functions such as binding functions (but not conventional allosteric effector binding) are usually considered acceptable as autonomous activities (Schmincke-Ott and Bisswanger, 1980), however, these are really semantic problems which stem from the need to impose simplifying subdivisions on a continuous spectrum of structure and function.

Multifunctional proteins are found in all classes of organism and are involved in very diverse cellular processes. In E.coli, DNA polymerase I and aspartokinase I-homoserine dehydrogenase I are two contrasting examples of multifunctional enzymes, and in mammalian cells fatty acid synthase and immunoglobulins are multifunctional (Schmincke-Ott and Bisswanger, 1980). The simian virus 40 large tumour antigen is multifunctional (Rigby and Lane, 1983). The N.crassa arom enzyme has already been described. Despite this considerable diversity it is clear that the known multifunctional enzymes are particularly common in the amino acid biosynthetic pathways of prokaryotes and fungi.

Various generalisations can be made about the reactions catalysed by multifunctional enzymes. All known naturally occurring multifunctional enzymes catalyse reactions which are functionally related. Many catalyse two or more consecutive steps in a biosynthetic pathway. However, there are some multifunctional enzymes which catalyse nonconsecutive steps from the same pathway, for example, the HIS4 gene

product of S.cerevisiae is a trifunctional enzyme which catalyses the second, third, and tenth steps in histidine biosynthesis (Donahue et al., 1982). In E.coli the thrA gene encodes a bifunctional enzyme (aspartokinase I-homoserine dehydrogenase I) which catalyses the first and third steps in threonine biosynthesis (Katinka et al., 1980).

Enzyme activities which are found in a multifunctional polypeptide in one species are often organised in a different way in unrelated species: the same activities may be carried by independent polypeptides, or a number of distinct subunits may associate to form a multienzyme complex. However, related organisms usually have comparable types of organisation. The structural organisation of the enzymes in the shikimate and tryptophan pathways of a range of organisms is shown in Figure 1.6. The evolution of multifunctional proteins will be considered in Section 1.11.

1.9.2 The structure of multifunctional proteins

1.9.2A Structural domains within proteins

Structural "domains" within proteins were first recognised by crystallographers in immunoglobulin molecules and are a very common feature in enzymes and proteins. The term refers to a spatially separate, compact unit within a protein (Rossmann and Argos, 1981; Phillips et al., 1983). The structure of a domain is usually reminiscent of a small,

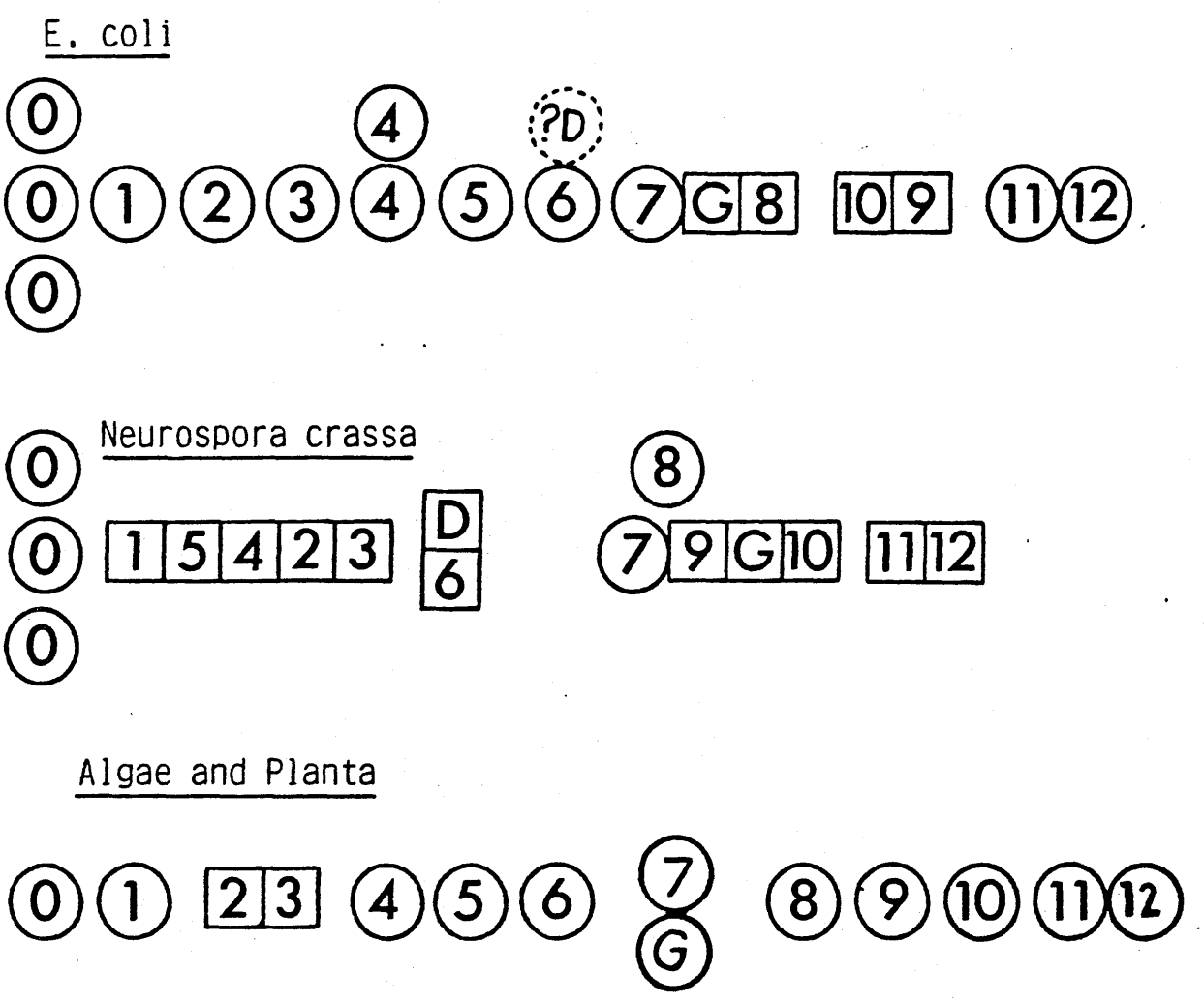


Figure 1.6 Structural organisation of enzymes in the shikimate pathways and tryptophan pathways of a range of organisms

The enzymes are numbered as in Figures 1.1 and 1.3. Rectangles represent multifunctional polypeptides while circles represent monofunctional polypeptides. Joined circles represent multienzyme complexes. After Boocock (1983).

compact protein and is often described as being a tight "glob". A few of the attributes commonly held by domains are listed below (Rossmann and Argos, 1981):

1. Domains often have a specific function (for example, the nucleotide binding domain in dehydrogenases).
2. Active sites often lie between domains.
3. Domains are made up of one, or a very small number of, continuous stretches of a polypeptide chain.
4. Limited proteolysis will often separate the component domains of a protein.

There are several basic possibilities for the structure of multifunctional enzymes. Firstly, one could have an active site catalysing more than one reaction. In most cases this is very implausible from what is understood of enzyme function, and there is usually some evidence available which suggests that the active sites of a particular multifunctional enzyme are separate (see below).

Secondly, one could hypothesise that one domain of a protein could carry two or more active sites, or at least contribute towards two or more active sites. This model implies that it would be almost impossible to isolate functional fragments by limited proteolysis (see below). A further corollary is that it is very unlikely that - within one domain - those parts of the primary structure directly involved in a particular activity would come from just one, or a few, contiguous stretches of the sequence.

Thirdly, it has been proposed that each active site or binding site of a multifunctional protein is located within a discrete structural "domain" of the folded polypeptide, thus envisaging multifunctional proteins as monofunctional proteins strung together (Kirschner and Bisswanger, 1976). This hypothesis is often referred to as the "mosaic" model or the "beads on a string" model. Note that Kirschner and Bisswanger's usage of the term "domain" is slightly different from that introduced above: allowance must be made for one of their "domains" to include, if necessary, two adjacent interacting domains (as defined earlier) as in a dehydrogenase, for example. However, their basic proposal is clear. The mosaic model predicts that it may be feasible to separate the component activities as different fragments of the multifunctional protein. It also strongly suggests that the amino acid sequences directly responsible for a particular activity should be mainly contiguous, as has been observed in known domains.

The available evidence concerning the structure of multifunctional proteins is very far from being definitive and is considered in the next section. However, much of it favours the mosaic model.

1.9.2B Evidence for the mosaic model of multifunctional proteins

Some of the experimental approaches described here will

be illustrated for the N.crassa arom enzyme in Section 1.10.

The ideal technique for probing the structure of multifunctional proteins, but one which requires a large investment in time and effort, is high resolution X-ray crystallography. Unfortunately very few relevant structures have yet been solved using this method. The only multifunctional protein whose whole structure is known at atomic resolution is a single type of immunoglobulin G molecule (Alberts et al., 1983). As well as the function represented by the antigen binding sites there are also the functions associated with the F_c region of IgG. These include binding to specific receptors on placental and phagocytic cells as well as binding to and activating complement. The IgG molecule has a clear domain structure and, as predicted from internal amino acid sequence homologies, all the domains have very similar folding patterns. This is one of the most obvious examples of evolution by gene duplication. The different functions are associated with different domains.

The crystal structure of the Klenow fragment of E.coli DNA polymerase I has been determined very recently (Ollis et al., 1985). This proteolytic fragment has the DNA polymerase and the 3'$\longrightarrow$5' exonuclease functions but not the 5'$\longrightarrow$3' exonuclease activity. Unfortunately the relationships between structure and function here are not yet sufficiently understood for many conclusions to be drawn. However, there are two distinct domains the smaller of which binds dTMP and the larger of which has a very deep crevice

apparently well suited for binding B-DNA. Each domain is built from a contiguous stretch of amino acid sequence.

Hopefully the detailed structures of more multifunctional proteins will be determined soon. In the absence of X-ray crystallographic data various indirect approaches have been used to study the structure of multifunctional proteins. Some of these are outlined below.

In a handful of cases (to be described in Section 1.11.2) sequence homologies have been demonstrated between a multifunctional enzyme and the corresponding monofunctional enzymes from a different species. Since the monofunctional enzymes are independent folding units these homologies favour the idea of discrete functional domains within multifunctional enzymes.

Chemical modification has been used to show that the different reactions of a multifunctional enzyme are catalysed at independent active sites. It is often possible to inactivate specifically one function of a multifunctional enzyme without harming the other activities. This will be illustrated for arom in Section 1.10.2.

Genetic analysis is a powerful tool in the study of multifunctional enzymes from prokaryotes and lower eukaryotes. At a coarse level the isolation of mutants each deficient in only one activity of a multifunctional enzyme is evidence for the independence of the active sites. Fine structure mapping of mutations in the genes for a variety of multifunctional enzymes has been carried out, for example in the

td locus of *N.crassa* which encodes a bifunctional tryptophan synthase (Bonner *et al.*, 1965) and in the *HIS4* gene of *S.cerevisiae* which encodes a trifunctional enzyme involved in histidine biosynthesis (Donahue *et al.*, 1982). In these cases, as for *N.crassa arom* (see Section 1.4.1A), mutations inactivating particular activities usually map in single, non-overlapping regions of the gene. This supports the mosaic model.

Limited proteolysis is one of the most useful techniques for demonstrating the existence of structural sub-regions corresponding to the component activities of a multifunctional enzyme. The linking regions between domains are known to be vulnerable to proteases and it is unlikely that internal fragments of a single domain would maintain a stable conformation. The Klenow fragment of *E.coli* DNA polymerase I is a classic example of the use of limited proteolysis to give functional subsets of multifunctional proteins. The application of this method to the *N.crassa arom* enzyme will be described in Section 1.10.3. Limited proteolysis has been used to study many multifunctional proteins including vertebrate fatty acid synthase (McCarthy and Hardie, 1984) and aspartokinase I-homoserine dehydrogenase I of *E.coli* (the bifunctional *thrA* gene product). In the latter case one can isolate a C-terminal fragment having only homoserine dehydrogenase activity (Katinka *et al.*, 1980).

Functional subsets of multifunctional proteins have also been isolated by exploiting chain termination (nonsense) mutants. A thrA nonsense mutant in E.coli, for instance, has permitted the isolation of an N-terminal fragment having only aspartokinase activity.

1.10 The arom multifunctional enzyme of N.crassa

1.10.1 Genetic studies

The results of the genetic analysis of the N.crassa arom locus have already been summarised (see Section 1.4.1A; Giles et al., 1967a; Rines et al., 1969). When considered in terms of a single pentafunctional polypeptide these results support the hypothesis that each activity resides on an independent structural unit. The five classes of mutants in which only a single enzyme activity is lost are most simply interpreted as being due to missense mutations. They demonstrate that the active sites of arom are probably autonomous. The fine structure genetic map shows that mutations affecting individual activities map in unique non-overlapping regions which is the pattern the mosaic model predicts. The polarity mutants, at least some of which are suppressible by nonsense suppressors and are therefore probably chain termination mutants, also imply that there are discrete sub-regions within the polypeptide which are responsible for individual functions. Furthermore, the

polarity data suggest that the order of the activities along the polypeptide chain is (starting from the amino terminus): E1, E5, E4, E2, E3 (see Figure 1.7).

Giles *et al.* (1967a) observed interallelic complementation between mutants affecting E3 and also between mutants affecting E1. *Arom* is a dimer of identical subunits (Lumsden and Coggins, 1977; 1978) and the interallelic complementation suggests that at least the E1 and E3 regions of each subunit interact with the homologous regions of the other subunit.

1.10.2 Chemical modification studies

In addition to the genetic data, there is substantial evidence from *in vitro* modification of the purified *arom* enzyme that all five active sites are spatially distinct.

The E1 activity can be specifically destroyed by removal of the essential Zn atom with chelating agents (Lambert *et al.*, 1985). The E2 activity alone can be inactivated by treatment with sodium borohydride in the presence of dehydroquinate (see Section 1.8.4A; Smith and Coggins, 1983). Treatment of *arom* with formaldehyde plus sodium borohydride inactivates both E2 and E3, however, only E3 is protected from inactivation by shikimate (J. Lumsden and J.R. Coggins, unpublished results). E5 can be uniquely inactivated in a variety of ways: by oxidation, by the competitive inhibitor glyphosate, or by an endogenous *N.crassa* protease (Boocock, 1983; Boocock and Coggins, 1983). E4 can be almost uniquely destroyed by very mild treatment with trypsin or subtilisin (Smith and Coggins, 1983; Boocock, 1983; see below).

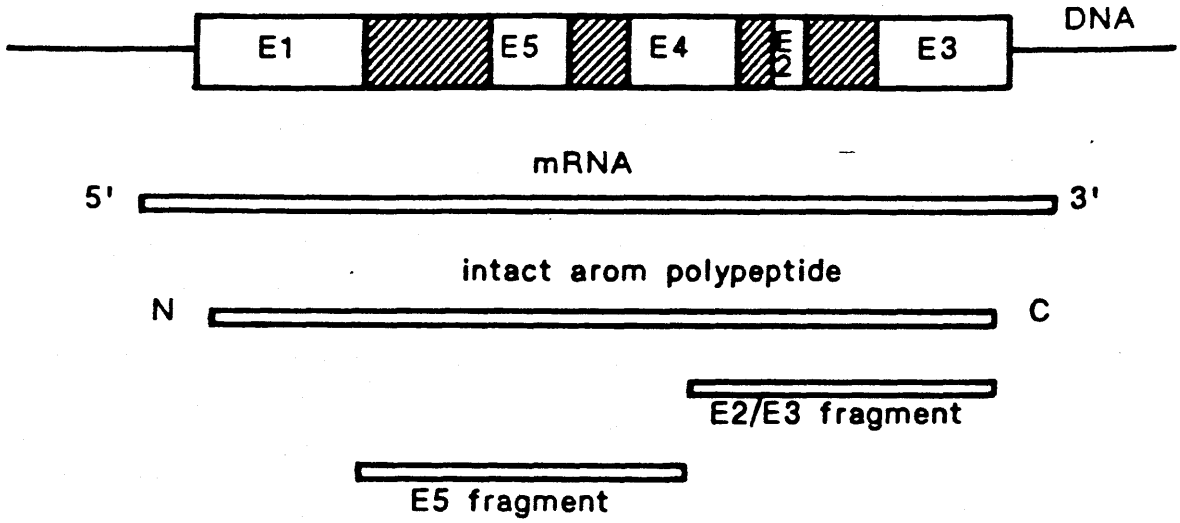


Figure 1.7 The arom multifunctional enzyme of N.crassa

1.10.3 Limited proteolysis studies

The susceptibility of the arom enzyme to cleavage by endogenous N.crassa proteases was one of the major problems in the purification of this protein. Such susceptibility is a general problem in the study of multifunctional proteins. However, the knowledge that the arom polypeptide could be cut in several places while still retaining all five activities encouraged deliberate attempts to fragment the intact protein into functional subsets.

The subunit m.w. of the arom enzyme is 165 kDa. Mild treatment of intact arom with trypsin or subtilisin gives a rapid initial cleavage resulting in two fragments of 110 kDa and 68 kDa on SDS PAGE (Smith and Coggins, 1983). E4 activity is destroyed in parallel with the disappearance of the 165 kDa arom polypeptide but the other four activities are much more robust. Under native conditions the 110 and 68 kDa fragments remain noncovalently bound in a complex indistinguishable on native gels from intact arom enzyme. Smith and Coggins (1983) showed that the 68 kDa fragment carries the active site lysine of E2 and can be renatured and stained for E3 activity after treatment with 8M urea. It would appear that the 68 kDa fragment represents a bifunctional subset of the arom polypeptide. From the genetic data it is almost certainly a C-terminal fragment (see Figure 1.7). It is interesting that at least the E3 activity can refold following denaturation in urea since this implies that no information other than that contained within the 68 kDa fragment is required for

this activity. Subsequent work, involving arom fragments generated by V8 protease and renaturation after SDS PAGE, has lowered the minimum size of polypeptide fragment required for recovery of E3 activity to 44 kDa (M. Boocock, unpublished results). Further trimming may well be possible.

Smith and Coggins (1983) suggested that the 110 kDa fragment (presumptively complementary to the 68 kDa fragment) carries E1 and E5 activities on the basis of the genetic data. It has now been shown that a 74 kDa fragment, derived from the 110 kDa fragment by prolonged treatment with trypsin and chymotrypsin, carries E5 activity (Boocock, 1983; Coggins et al., 1985). This fragment has been isolated preparatively. E1 activity is lost during the proteolysis procedure used here and the 68 kDa fragment is trimmed to 63 kDa.

Although much work remains to be done, the results already obtained from limited proteolysis of the arom enzyme support the mosaic model.

1.11 The evolution of multifunctional proteins

Questions about the evolution of multifunctional proteins can be divided into those concerned with "why?" they evolved and those directed at "how?" they evolved. As will shortly be described, the work for this thesis was mainly aimed at helping to answer a question of the second type.

1.11.1 Why have multifunctional proteins evolved?

The first possibility to consider is that there are selective advantages to be gained by having some functions organised in multifunctional proteins. Many hypothetical advantages have been put forward but none are yet generally accepted (Schmincke-Ott and Bisswanger, 1980). Some of these proposals are given below.

Multifunctional proteins may allow greater economy in the use of genetic information: one may obtain the expression of n activities using only one promoter, terminator, and ribosome binding site. However, this assumes that the kinetic properties of the different activities are sufficiently similar for the production of equimolar amounts of the different active sites to be appropriate.

The physical strength of a covalent connection between two functions may be useful in rare cases.

It has often been suggested that multifunctional enzymes might allow more efficient catalysis due to "channelling" of the product of one active site to the next active site. In a related argument Giles *et al.* (1967a) suggested that the pentafunctional *N.crassa arom* enzyme might be necessary to segregate the shikimate pathway from the competing catabolic quinate pathway. However, this does not explain why E4 and E5 are included in *arom*. Also, *S.cerevisiae* apparently lacks any catabolic quinate pathway. In general, channelling cannot explain cases where a multifunctional enzyme catalyses non-sequential steps in a pathway unless there are other

subunits, in a multienzyme complex, which catalyse the intervening steps. No evidence for substrate channelling has yet been found with the N.crassa arom enzyme (G.A. Nimmo, J.M. Lambert, M.R. Boocock, J.R. Coggins, unpublished results). There is clearly a need for further work in this area. This will be facilitated by the cloning and overexpression of the genes for the relevant enzymes.

The above description of the possible merits of multifunctional proteins, in comparison with monofunctional ones, has rather ignored a third "contestant" - the multienzyme complex. It is quite conceivable that a particular selective advantage could equally well be gained by two different routes. It may be that in certain cases the covalent attachment of functions allows topologies which are not readily evolved as multienzyme complexes. Evolutionary processes are constrained by their functional environment (Gould, 1984). For example, the symmetry relationships in two different homo-oligomers may put barriers in the way of the evolution from these subunits of a multienzyme complex with a particular arrangement of active sites.

A second major possibility is that some multifunctional proteins may confer no selective advantage and are found today only because their genes (after arising by chance) became fixed in the population and are difficult to eliminate. It is much easier for two genes to be fused (see below) by a deletion event than it is for the process to be reversed, whereupon all the control sequences which were lost during

the original deletion event must somehow be restored.

It has been argued (Schmincke-Ott and Bisswanger, 1980) that the relatedness of the functions in all known multifunctional proteins rules out the possibility that multifunctional proteins arose, say, by accidental gene fusions followed by neutral fixation. This argument is not impregnable. It may be that accidental fusions between unrelated genes do occur but that these are more likely to be detrimental, for example, due to the disruption of regulatory processes. Furthermore, there may be some accidental fusions of unrelated genes that have simply not yet been detected. For instance, a bifunctional enzyme with two known activities could conceivably have an unknown, unrelated "Cinderella" activity which goes undetected because the appropriate assay is never performed.

1.11.2 How have multifunctional proteins evolved?

It is generally assumed that multifunctional proteins have evolved from monofunctional proteins. There are three plausible mechanisms for how this might occur:

1. There might be adaptation of an existing monofunctional enzyme to catalyse a second reaction (Llewellyn *et al.*, 1980).
2. There might be internal duplication within the gene for a monofunctional enzyme and subsequent adaptation of one active site for catalysis of a second reaction (Schmincke-Ott and Bisswanger, 1980).

3. There might be fusion of the genes for two monofunctional enzymes (Bonner *et al.*, 1965).

It has been proposed (see Section 1.8.2) that the bifunctional enzyme chorismate mutase/DAHP synthase of *B.subtilis* strain 168 arose through adaptation of a regulatory prephenate binding site in an existing DAHP synthase (Llewellyn *et al.*, 1980). However, this is an atypical case given the special nature of the chorismate mutase reaction. Also, the simple adaptation hypothesis is embarrassed by the fact that most multifunctional proteins tend to have "large" subunit molecular weights.

The evolution of immunoglobulins provides a clear example of gene duplication and adaptation (see Section 1.9.2B). It is not known how general this mechanism might be. Internal, repeating sequence homologies within a multifunctional protein would strongly imply a gene duplication mechanism for the evolution of that protein.

There is good evidence that several multifunctional enzymes evolved by a gene fusion mechanism. This evidence is based on the finding of sequence homologies between a multifunctional protein and its monofunctional counterparts from a different species (see Figure 1.8).

In *E.coli* the *trpD* gene is part of the *trp* operon and encodes a bifunctional enzyme having both the glutaminase (anthranilate synthase II; ASII) and the anthranilate phosphoribosyl transferase (PRT) activities of the tryptophan pathway (Miozzari and Yanofsky, 1979; see Figure 1.3). Genetic analysis and limited proteolysis show that the two

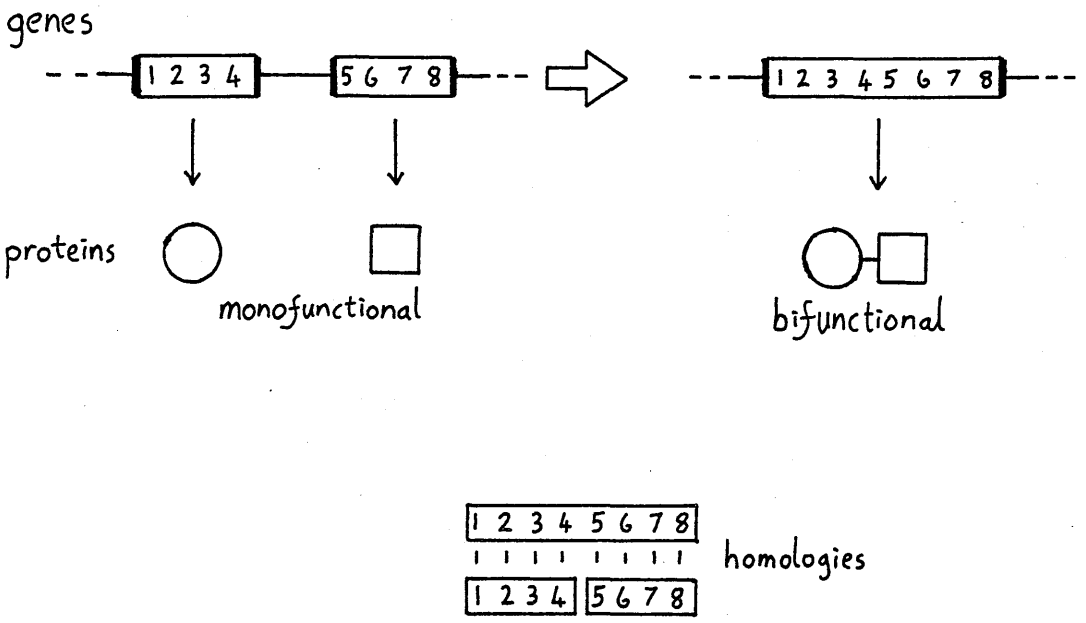


Figure 1.8 Evolution of multifunctional proteins by gene fusion

activities are associated with different segments of the polypeptide chain, ASII being N-terminal. The E.coli trpD gene is often written trp(G)D. In contrast, in the related enteric bacterium Serratia marcescens (S.marcescens) the same two activities occur as separate proteins. In this species the trpD gene encodes only the PRT activity and the operon contains an extra gene, trpG, which encodes ASII and lies immediately 5' of trpD, thus: ....-trpG-trpD-... Miozzari and Yanofsky (1979) compared the sequence of E.coli trpD with that of the S.marcescens trpG-trpD region. They found very extensive homology between the S. marcescens trpG gene and the N-terminal part of E.coli trpD and between the S.marcescens trpD gene and the C-terminal part of E.coli trpD. The intercistronic region in S.marcescens was only 16 nucleotides long. It appears that there has been a gene fusion event in the E.coli operon, although one cannot entirely rule out the possibility of an ancestral fusion and subsequent separation in S.marcescens.

In the yeast S.cerevisiae the TRP5 gene encodes a bifunctional tryptophan synthase whereas in E.coli tryptophan synthase is a multienzyme complex consisting of two kinds of subunits encoded by the trpA and trpB genes. Sequence comparisons show clear homologies (at the protein level) between the N-terminal part of TRP5 and trpA and between the C-terminal part of TRP5 and trpB, thus implying that TRP5 arose by fusion of two genes for monofunctional subunits (Zalkin and Yanofsky, 1982). Similarly, there are homologies

between the bifunctional S.cerevisiae TRP3 gene product (ASII/indole glycerol phosphate synthase) and the appropriate regions of the E.coli trpD and trpC (bifunctional; indole glycerol phosphate synthase/phosphoribosyl anthranilate isomerase) gene products, suggesting that TRP3 evolved by gene fusion (Zalkin et al., 1984).

Fragmentary sequence homologies have been detected between the multifunctional mammalian fatty acid synthase and the corresponding monofunctional bacterial enzymes perhaps showing that the former arose by gene fusion (McCarthy and Hardie, 1984).

It is not known yet how general the gene fusion mechanism might be. Both gene duplication and gene fusion would predict the occurrence of independent functional domains encoded by contiguous genetic regions.

1.11.3 Gene fusion and the N.crassa arom multifunctional enzyme

It is not known how the N.crassa arom protein evolved. However, there are two very weak lines of evidence which point to gene fusion being the most plausible hypothesis. Firstly, there are the similarities between the various enzyme activities of arom and the corresponding monofunctional activities of E.coli. These were outlined in Section 1.8. Secondly, there is the arom subunit molecular weight which is very close to the sum of the corresponding E.coli subunit molecular weights. This is illustrated in Table 1.1. The bifunctional E2/E3

Table 1.1 Subunit molecular weights of the E.coli enzymes corresponding to the five arom activities

E.coli activity	Subunit m.w.(kDa)	Reference
E1	39	G. Millar, unpublished
E2	28	Duncan (1985)
E3	29	This study
E4	20	Ely and Pittard (1979)
E5	46	Duncan et al. (1984b)
	Total:162	

NOTE: N.crassa arom subunit m.w. is 165 kDa by SDS PAGE.

from peas has a subunit m.w. of 60 kDa (see Section 1.6.2) which also agrees well with the hypothesis that it arose by the fusion of two E.coli-like monofunctional activities (28 + 29 kDa). Although the bifunctional P.patens E2/E3 appears to have a subunit m.w. too small (48 kDa) to be simply related to the E.coli E2 and E3 polypeptides it should be remembered that an E2 with a lower subunit m.w. (20 kDa) has been characterised: the catabolic E2 of N.crassa (20 + 29 kDa = 49 kDa). The 68 kDa proteolytic fragment of the N.crassa arom enzyme, which carries the E3 activity and at least the active site lysine of E2, can be trimmed further to 63 kDa (see Section 1.10.3), but it remains to be determined whether this fragment carries all sequences required for E2 activity.

1.11.4 This project

One of the long term aims of this laboratory is to find out how the arom multifunctional enzyme evolved. The working hypothesis is that this protein evolved by the fusion of genes for monofunctional activities. To test this idea it is planned to compare the amino acid sequence of an arom multifunctional enzyme with the amino acid sequences of the corresponding monofunctional enzymes from E.coli. The finding of homologies between each of the E.coli enzymes and the relevant parts of arom would be diagnostic of gene fusion. As part of this long term project I describe here the cloning

and sequencing of the E.coli aroE gene encoding shikimate dehydrogenase (E3). This route to obtaining sequence information also allows the overexpression of the cloned gene and the purification of relatively large amounts of the encoded protein for detailed study. The technical approach used to clone aroE is introduced in the next section.

PART B - INTRODUCTION TO THIS PROJECT

1.12 General approach

The method chosen to clone the E.coli shikimate dehydrogenase gene, aroE, exploited an Aro^- auxotrophic mutant of E.coli which is specifically defective in the aroE gene. This strain is E.coli AB2834 (Pittard and Wallace, 1966). It lacks E3 activity and will not grow on a minimal medium (lacking aromatics) which would support the growth of wild-type E.coli. A cloned, wild-type, aroE gene should be able to complement the auxotrophic mutation of AB2834 cells, thus allowing growth on minimal medium and providing a direct selection for the presence of this gene. This complementation approach is often called cloning by "relief of auxotrophy". One would "shotgun" fragments of the E.coli genome into E.coli AB2834 using a suitable plasmid vector and then plate the transformed cells on minimal medium. Any colonies which appeared - the "putative positives" - would be studied to see if they contained a recombinant plasmid carrying the aroE gene.

Cloning E.coli genes by complementation is a widely applicable technique (Clarke and Carbon, 1979) since a large number of wild-type E.coli genes can confer a selectable phenotype on the appropriate, and available, mutant. The expanding literature testifies to the widespread awareness of this fact.

1.13 Shortcuts in cloning E.coli genes by complementation

The logistics of cloning by relief of auxotrophy can be simplified by starting from a known small subset of the E.coli genome which is thought to harbor the desired gene. Such a simplification was possible in this case. There are many transducing phages available that carry defined fragments of the E.coli genome. The aroE gene is located at about 72 minutes on the E.coli linkage map (see Figure 1.4). While studying E.coli ribosomal protein genes which map very close to aroE Nomura and his coworkers isolated the defective transducing phage λspc1 (Jaskunas et al., 1975a; see Figure 1.9). This phage carries bacterial sequences from E.coli K12 and was thought to carry aroE at its extreme left end. It seemed worthwhile to see if any restriction fragments of λspc1 DNA were able to complement the aroE auxotrophic mutant E.coli AB2834 (see Chapter 3).

It should be noted that the E.coli genomic library constructed by Clarke and Carbon (1979), using the plasmid vector ColE1, has allowed shortcuts in cloning a number of

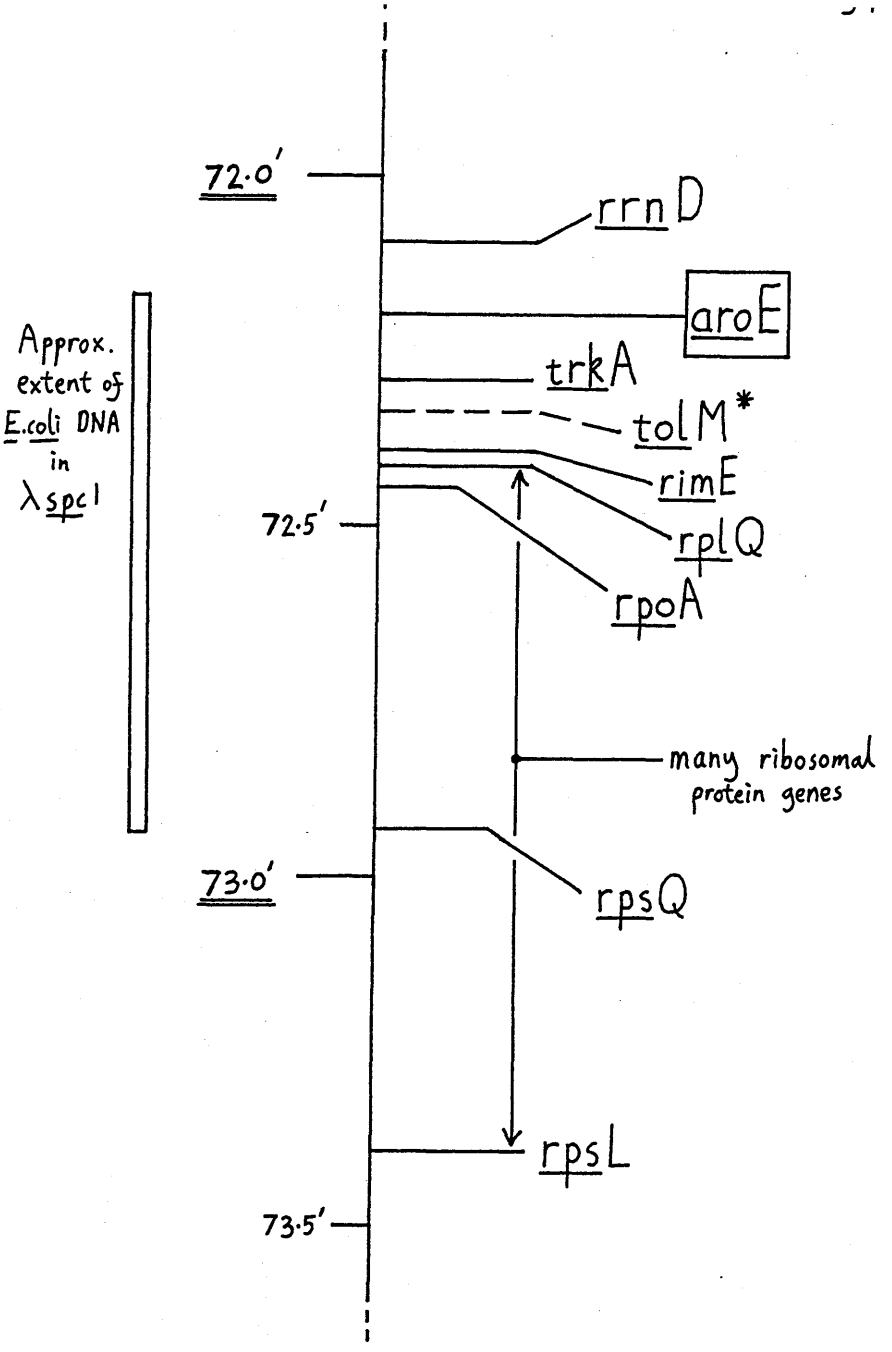


Figure 1.9 The region of the E.coli genetic map around aroE

From Bachmann (1983). The approximate extent of the E.coli DNA carried in the transducing phage λspc1 is shown. The asterisk indicates that the position of tolM with respect to nearby markers is uncertain. rrnD represents a ribosomal and transfer RNA operon. The trkA gene product is involved in the transport of potassium.

genes. Clarke and Carbon initially screened their library by testing clones for the ability to complement a wide variety of E.coli mutants. This permitted the tentative assignment of particular genes, and thus regions of the linkage map, to particular recombinant plasmids. They were, however, very aware of the dangers of mistaking suppression for complementation. They identified two recombinant plasmids in their library as being able to complement an aroE auxotrophic mutant: pLC16-1 and pLC12-24. These are not considered further. An important and useful extension of the work of Clarke and Carbon is the "gene-protein index" for E.coli K12 (Neidhardt et al., 1983).

CHAPTER 2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Chemicals

General chemicals were obtained from many different manufacturers and were of analytical reagent grade or the best available quality.

All solutions were made using glass-distilled water, except for those solutions used during the annealing and synthesis steps of DNA sequencing where deionised, distilled water was used.

Fine chemicals and radiochemicals. 3-dehydroquinate (ammonium salt) and 3-dehydroshikimic acid were the generous gifts of S. Chaudhuri.

Shikimic acid was obtained from the Aldrich Chemical Co., Gillingham, Dorset, U.K.

$NADP^+$ ($diNa^+$ salt), NADH (grade II, Na^+ salt), and PEP (K^+ salt) were obtained from Boehringer Corp., Lewes, East Sussex, U.K.

ATP(A-5394, essentially vanadium free), ampicillin, tetracycline, streptomycin sulphate, chloramphenicol, ethidium bromide, Coomassie Brilliant Blue G-250, PMSF, benzamidine, and DTT were obtained from Sigma Chemical Co., Poole, Dorset, U.K.

2-mercaptoethanol, 99% formic acid (Analar), 30% hydrogen peroxide (Analar), concentrated hydrochloric acid (Aristar), xylene cyanol, bromophenol blue, nitro-blue tetrazolium, phenazine methosulphate, polyethylene glycol 6000, Amberlite MB3, CsCl, L-proline, L-leucine, L-phenylalanine, L-tyrosine, L-tryptophan, p-aminobenzoic acid, p-hydroxybenzoic acid, acrylamide, bis-acrylamide, N,N,N',N'-tetramethylene diamine, SDS, and enzyme grade ammonium sulphate (especially low in heavy metals) were obtained from BDH Chemicals, Poole, Dorset, U.K.

Thiamine hydrochloride (vitamin B_1) was obtained from Fluka, Fluorochem Ltd., Glossop, Derbyshire, U.K.

Agarose and low melting point agarose (for DNA gels), urea, IPTG, and X-gal were obtained from BRL, Gibco Ltd., P.O. Box 35, Paisley, U.K.

Bactotryptone, yeast extract, and "Bactoagar" were obtained from Difco, Detroit, U.S.A.

Oxoid nutrient broth powder and Oxoid No.1 agar were obtained from Oxoid Ltd., London SE1, U.K.

Deoxy- and dideoxy nucleotides (dATP, dTTP, dCTP, dGTP, ddATP, ddTTP, ddCTP, ddGTP) and $[\alpha\text{-}^{35}S]$ dATPαS for chain termination sequencing were obtained from Amersham International plc, Amersham, U.K. The nucleotides were part of a kit (see below).

Prior to use in nucleic acid manipulations <u>phenol</u> was redistilled and stored in small aliquots at -20°C. Prior to use in manipulations involving protein <u>urea</u> was recrystallised from ethanol.

2.1.2 Chromatographic media

Sephadex G-75 (superfine grade), DEAE-Sephacel, Sephacryl S-200 (superfine grade), and 2'5' ADP-Sepharose 4B were all obtained from Pharmacia, Milton Keynes, Bucks, U.K.

2.1.3 Enzymes and proteins

The following were obtained from Boehringer Corp., Lewes, East Sussex, U.K.: catalase (EC 1.11.1.6) from beef liver, glutamate dehydrogenase (EC 1.4.1.3) from beef liver, glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) from rabbit muscle, aldolase (EC 4.1.2.13) from rabbit muscle, carbonic anhydrase (EC 4.2.1.1) from bovine erythrocytes, pyruvate kinase (EC 2.7.1.40) from rabbit muscle/lactate dehydrogenase (EC 1.1.1.27) from rabbit muscle (mixed suspension for use as coupling enzymes), and alkaline phosphatase (EC 3.1.3.1) from calf intestine (Grade 1). Prior to use in DNA cloning experiments the calf intestinal alkaline phosphatase was further purified (to remove residual nucleases) by gel filtration on Sephadex G-75 (superfine) as described by Efstratiadis et al. (1977).

Deoxyribonuclease I (EC 3.1.4.5) from bovine pancreas, bovine serum albumin, proteinase K, RNase A, and lysozyme were obtained from Sigma Chemical Co., Poole, Dorset, U.K.

Gelatin was obtained from Serva, Uniscience Ltd., Cambridge, U.K.

All restriction endonucleases were obtained from BRL, Gibco Ltd., P.O. Box 35, Paisley, U.K., as were T4 DNA ligase and nuclease-free bovine serum albumin.

Klenow fragment of E.coli DNA polymerase I was obtained from Amersham International plc, Amersham, U.K. as part of a kit (see below) which also included the M13 oligonucleotide primer.

2.1.4 Bacterial strains and episomes

The bacterial strains used are shown in Table 2.1. Plasmid vector pAT153 (Twigg and Sherratt, 1980) was the gift of R. Krumlauf. The plasmid expression vector pKK223-3 (J. Brosins, unpublished work) was the gift of J.R. Knowles. Further supplies of these two plasmids were obtained by exploiting their ability to replicate (see below). The sequencing phage vectors M13mp8 and M13 mp9 (Messing and Vieira, 1982) were obtained from Amersham International plc, Amersham, U.K. as part of a kit (see below).

2.2 General biochemical methods

pH measurements were made with a Radiometer pH probe calibrated at room temperature.

Conductivity measurements were made with a Radiometer conductivity meter type CDM2e (Radiometer, Copenhagen, Denmark).

Protein estimations were done by the method of Bradford (1976), with bovine serum albumin as the standard.

Table 2.1 Bacterial strains used

Strain	Genotype	Origin/Reference
E.coli K12	wild-type ATCC 14948 F^-, λ lysogenic	American Type Culture Collection (Rockville, Maryland, U.S.A.)
E.coli AB2834	*aro*E-353, *mal*-352, *tsx*-352, F^-, λ^R, λ^- *sup*E-42	CGSC (*E.coli* Genetic Stock Centre, Dept. of Human Genetics, Yale University, New Haven, U.S.A.) Pittard and Wallace (1966)
E.coli NO1267	str^R, double lysogen of λ*dspc*1 and λ *cI857S7*, *thi*	Jaskunas *et al.* (1975a)
E.coli HB101	*pro*$^-$ *leu*$^-$ *thi*$^-$ *lacY*$^-$ *hsdR*$^-$ *endA*$^-$ *recA*$^-$ *rpsL*20 *ara*-14 *galK*2 *xyl*-5 *mtl*-1 *sup*E44	Bolivar and Backman (1979)
E.coli JM101	Δ *lac pro*, *thi*, *sup*E, F' *traD*36, *proAB*, *lacI*^{q}*Z*ΔM15	Messing and Vieira (1982)
E.coli TG1	This is an *hsd*Δ5 (EcoK r^-m^-) version of JM101	T. Gibson (unpublished)
E.coli Ymel	F^+, *sup*F, λ^s	Gift of D.W. Meek

DNA was determined spectrophotometrically at 260nm (Maniatis et al., 1982). An A_{260} value of 1.0 (1 cm path) corresponds to 50 µg/ml of double-stranded DNA. This method is only applicable to pure preparations of DNA.

Absorbance (A_x) values refer throughout to absorbances measured in a cuvette having a 1 cm path length, at x nm.

Dialysis membranes were boiled for at least 15 min in 1mM Na_2EDTA and then rinsed thoroughly with d.w. prior to use.

2.3 General microbiological techniques

Many of these techniques are described in Maniatis et al. (1982).

Purification of strains. Before use all new strains were streaked on plates for single colonies, usually twice in succession. Plate tests were used to check genetic markers (see below).

Preservation of strains. For short term use plates with bacterial colonies were stored (sealed) at 4°C. For short or medium term use O/N cultures in Nutrient Broth or L Broth remained viable at 4°C up to about six months. For long term storage two different methods were employed (usually both): stabs and storage at -20°C in medium containing 15% (v/v, final concentration) glycerol (Maniatis et al., 1982). The latter is preferable for strains containing plasmids and anyway gives better long term viability (several years).

Replica plating was carried out using a "mushroom" apparatus and velvets.

Growth of bacteria. Solid and liquid media for the growth of bacteria are described in the next section. The growth of bacteria in liquid cultures was followed at 650 nm using a Gilford-Unicam model 252 spectrophotometer. The linear range is from A_{650} = 0 - 0.6 (determined experimentally using a culture of E.coli HB101 grown to stationary phase on L broth and diluted). The A_{650} of denser cultures was obtained after dilution of a sample in the appropriate medium. Unless otherwise stated all bacteria were grown at 37°C. Drug selection was always used (except during the expression phase of transformations - see below) on strains containing a plasmid with a selectable drug resistance marker. Inocula for moderate to large cultures were usually taken from a 10 ml culture in Nutrient Broth or L Broth which had been grown O/N without shaking after inoculation from a stock culture. The inoculum volume was usually 1-5% of the volume of the new medium.

Harvesting of bacteria for biochemical purposes was done by centrifugation, usually in a rotor at 4°C in an MSE High Speed 18 refrigerated centrifuge. Larger volumes were spun in the 6 x 300 ml angle rotor at 6 krpm (7000 g) for 5-10 min, while smaller volumes were spun in the 8 x 50 ml angle rotor at 6 krpm (5400 g) for 5-10 min.

2.4 Solid and liquid media for the growth of bacteria

Unless otherwise stated all media (and supplements) were sterilised by autoclaving.

2.4.1 Rich media

L Broth contains, per litre of d.w.: 10 g Bactotryptone
5 g yeast extract
10 g NaCl

After autoclaving add 5 ml sterile 20% (w/v) D-glucose per litre (0.1% final), except when the medium is to be used for preparing bacteria for titration of λ phage whereupon the glucose is omitted and 0.2% (w/v, final) maltose is added instead.

L agar: add 15 g Bactoagar per litre of L Broth before autoclaving. Supplement with D-glucose (as above) before pouring, or maltose (as above) for bottom agar to be used for λ phage titration. Top agar for λ phage titration is the same as bottom agar but contains only 7 g/l Bactoagar.

Nutrient Broth contains per litre of d.w.: 13 g Oxoid No.1 nutrient broth powder.

Stab medium: add 6 g/l Bactoagar to Nutrient Broth before autoclaving.

2xTY medium contains per litre of d.w.: 16 g Bactotryptone
10 g yeast extract
5 g NaCl

H agar contains per litre of d.w.: 10 g Bactotryptone
8 g NaCl
12 g Bactoagar

H top agar: as for H agar but only 8 g/l Bactoagar.

2.4.2 Defined minimal media

All defined minimal media used were based on one version of M9 salts medium. This contains, per litre of d.w.:

1 g	NH_4Cl
0.13 g	$MgSO_4 \cdot 7H_2O$
3 g	KH_2PO_4
6 g	Na_2HPO_4

After autoclaving, add vitamin B_1 (thiamine hydrochloride; final concentration 2 μg/ml), $CaCl_2$ (final concentration 0.1mM), and D-glucose (final concentration 2 g/l). The $CaCl_2$ and D-glucose are autoclaved together while the vitamin B_1 is filter sterilised. So that the same minimal medium would suffice for E.coli HB101 (which requires leucine and proline) or K12 the above salts were supplemented prior to autoclaving with L-leucine (100 mg/l final) and L-proline (150 mg/l final) to give "minimal medium (MM)". However, for the growth of the host strains for M13 sequencing proline (and leucine) are omitted because these strains contain an unstable F' episome carrying the pro^+ marker in a pro^- background (see Table 2.1) and one selects for maintenance of the episome on a true minimal medium.

For solid defined media the salts must be autoclaved separately from the agar to avoid precipitation problems. Hence, the M9 salts are made up as a 2 times concentrated stock solution and autoclaved separately before being mixed

with the agar solution. For solid defined media Oxoid No.1 agar was used, rather than Bactoagar, and at 12.5 g/l (final). After mixing the two halves and cooling to 55°C the vitamin B_1, $CaCl_2$ and glucose are added in the same concentrations as for minimal medium and the plates are poured (M agar).

For plate tests on aromatic auxotrophs the above solid defined medium was supplemented with various compounds to the final concentrations given below. For the growth of shikimate pathway aromatic auxotrophs plates were supplemented with the five end-products: phe (80 mg/l), tyr (80 mg/l), trp (40 mg/l), PABA (320 µg/l), and PHBA (320 µg/l). For testing auxotrophs with subsets of the above five aromatic compounds PABA and PHBA were used at the same concentrations but phe, tyr, and trp were used at 20 mg/l. All these end-products can be added to the medium prior to autoclaving. As described in Chapter 3, supplements of the common pathway intermediates DHQ, DHS, or shikimic acid were always given along with phe (20 mg/l) + tyr (20 mg/l), and were each used at 100 mg/l. These intermediates were made up in concentrated stock solutions, filter sterilised, and the appropriate volume (of the order of 0.1 ml) spread on the M agar + phe + tyr plates a few hours before use.

2.4.3 Drug supplements

Ampicillin (Amp) was used at a final concentration of 100 µg/ml. A stock solution of 5 mg/ml was filter sterilised

and stored at -20°C. Hot agar was cooled to 55°C before ampicillin was added. L agar + ampicillin = L Amp plates. M agar + ampicillin = M Amp plates. Plates containing ampicillin could be kept for at least 4 weeks if stored at 4°C.

Tetracycline (Tet) was used at a final concentration of 15 μg/ml. A stock solution of 10 mg/ml in absolute ethanol was stored in the dark at -20°C. Hot agar was cooled to 55°C before tetracycline was added. L agar + tetracycline = L Tet plates, which were used the same day that they were made.

Streptomycin (Str) was used at a final concentration of 25 μg/ml. A stock solution of 20 mg/ml was filter sterilised and stored at -20°C.

Chloramphenicol for plasmid amplification was used at a final concentration of 150 μg/ml. A stock solution of 34 mg/ml in absolute ethanol was stored at -20°C.

2.5 Crude E.coli cell extracts (for analytical purposes)

For assaying E3 and E2 in crude E.coli extracts: grow up 2 x 50 ml bacterial cultures in L Broth (supplemented, if appropriate, with ampicillin) each in a 250 ml conical flask on an orbital shaker at 37°C. After they reach the desired A_{650} cool on ice, pool, and harvest the bacteria by centrifugation at 4°C. Resuspend pellets in sonication buffer I:

0.2M KCl

0.2M KH_2PO_4 (pH to 7.0 with KOH)

2mM $MgCl_2$

1mM βME

Spin down. Resuspend pellet in 3.5 ml sonication buffer I. Break cells by sonication: 3-5 x 30 s bursts, each separated by a 30 s interval for cooling, using a Dawe soniprobe type 7532A at 80 W (output setting 2) with the sample vial in a brass cooling block threaded on the probe and immersed in iced water. Spin the disrupted cell suspension at 4°C in a precooled rotor for 2 h with 199,000 g at the maximum radius (10 x 10 ml MSE aluminium angle rotor (119) in an MSE Prepspin 50 centrifuge). The supernatant, which forms the crude extract, is removed and kept on ice. E3 and E2 activities are stable in this extract for at least 36 h but assays were always performed as soon as possible.

In a preliminary experiment in which extracts were assayed for E4(and E3) the procedure was as above but with the following modifications: grow the 2 x 50 ml cultures in minimal medium. Use sonication buffer II which does not contain P_i:

0.2M Tris-HCl pH 7.5

0.2M KCl

5mM $MgCl_2$

0.4mM DTT

(A. Lewendon, unpublished work). Since E.coli E4 activity is thought to be unstable 10 mg/ml bovine serum albumin was included in the final cell suspension (prior to sonication) for a duplicate extract. For the same reason the duration of the high speed spin was reduced to 1 h and assays were done at once.

2.6 Enzyme assays

Continuous spectrophotometric assays were done at 25°C in masked semi-micro quartz cuvettes (1 cm path, 1 ml). The instrument used throughout was a Gilford-Unicam model 252 spectrophotometer with a slave chart recorder. One unit of activity is defined as the amount of enzyme required to catalyse the conversion of 1 μmole of substrate per minute under the defined conditions.

Dehydroquinase (E2) was assayed spectrophotometrically at 234nm (ε = 12,000 $M^{-1}cm^{-1}$). The assay mixture contained (final concentrations) 0.2mM DHQ, 100mM KP_i pH 7.0. DHQ was used to initiate the assay of crude E.coli cell extracts. The KP_i was prepared as a two times concentrated stock solution (200mM KH_2PO_4/KOH pH 7.0).

Shikimate dehydrogenase (E3) was assayed in the reverse direction by monitoring the reduction of $NADP^+$ at 340nm (ε = 6,200$M^{-1}cm^{-1}$). The assay mixture contained (final concentrations) 4mM shikimic acid, 2mM $NADP^+$, 100mM Tris-HCl pH 9.0. When assaying crude E.coli extracts the assay was initiated with shikimic acid after recording any blank rate (usually very small).

Shikimate kinase (E4) was assayed spectrophotometrically at 340nm (ϵ = 6,200$M^{-1}cm^{-1}$) by coupling the release of ADP to the pyruvate kinase and lactate dehydrogenase reactions (G.A. Nimmo, unpublished work). Assays contained 1mM shikimic acid, 2.5mM ATP (neutralised with KOH), 1mM PEP (neutralised with KOH), 0.1mM NADH, pyruvate kinase (3U/ml) and lactate dehydrogenase (2.5U/ml) in 50mM KCl, 2.5mM $MgCl_2$, 50mM triethanolamine.HCl/KOH pH 7.0. The buffer was made up as a two-fold concentrated stock solution. The assay of crude E.coli extracts was initiated with shikimic acid after recording the rather rapid blank rate.

2.7 Polyacrylamide gel electrophoresis of proteins

2.7.1 SDS PAGE

Polyacrylamide gel electrophoresis in the presence of sodium dodecyl sulphate was done by the method of Laemmli (1970). Separation gels were polymerised from solutions containing 375mM Tris-HCl pH 8.8, the stated concentration (w/v) of acrylamide (at an acrylamide:bis-acrylamide ratio of 30:0.8), 0.033% (v/v) N,N,N',N',-tetramethylene diamine, 0.05% (w/v) ammonium persulphate and 0.1% (w/v) SDS. Stacking gels were polymerised from solutions containing 125mM Tris-HCl pH 6.8, 3% (w/v) acrylamide (30:0.8), 0.067% (v/v) N,N,N',N'-tetramethylene diamine, 0.1% (w/v) ammonium persulphate and 0.1% (w/v) SDS. Laemmli gels were poured in a slab format

(20 cm x 15 cm x 1.2 mm: Raven Scientific Ltd., Haverhill, Suffolk, U.K.). Samples for SDS PAGE were adjusted to at least 1% SDS, at least 2% (v/v)βME, at least 10% (v/v) glycerol, and a trace of bromophenol blue and heated to 100°C for 2-5 minutes. SDS PAGE was done at room temperature. The well buffer stock solution is 30 g/l Tris base, 144 g/l glycine and is diluted 10-fold and supplemented with 0.1% (w/v) SDS for SDS PAGE.

The protocol used for renaturing proteins after SDS PAGE is fully described in Section 3.4.4.

The molecular weight markers used in SDS PAGE experiments were: bovine serum albumin (68 kDa), catalase (60 kDa), glutamate dehydrogenase (53 kDa), aldolase (40 kDa), glyceraldehyde-3-phosphate dehydrogenase (36 kDa), and carbonic anhydrase (29 kDa). See Weber and Osborn (1969).

2.7.2 Non-denaturing gel electrophoresis

This was done by the method of Davis (1964) but using slab gels rather than tube gels. Gels were polymerised as described for Laemmli separation gels but without any SDS and with 0.5M Tris-HCl pH 8.8. The gels are poured in the same slab gel apparatus described above for SDS PAGE (after cleaning!). The well buffer is a 25-fold dilution of the stock Laemmli well buffer (without SDS) supplemented with 0.1mM DTT. Native gels were pre-electrophoresed (40mA) and were run in the cold room (4°C) at 25 mA. Samples for non-denaturing gels were incubated with 1-50mM DTT on ice for 1 h prior to loading (with glycerol).

2.7.3 Staining of polyacrylamide protein gels

Gels were stained for protein with 0.1% (w/v) Coomassie Brilliant Blue G-250 in 10% (v/v) acetic acid, 50% (v/v) methanol and destained at room temperature in 10% (v/v) acetic acid, 10% (v/v) methanol.

Non-denaturing gels, and renatured SDS gels, were stained for E3 activity using the nitro-blue tetrazolium (NBT) dye-linked method, as adapted by Lumsden and Coggins (1977). Gels were first soaked in 100mM Tris-HCl pH 8.8-9.0 for 1/2-1 h to remove DTT. The (freshly made up) staining mixture consisted of 0.5mM $NADP^+$, 1mM shikimic acid, 0.62 mg/ml NBT, 300mM Tris-HCl pH 9.0, to which was added 0.0125 volumes of 0.5 mg/ml phenazine methosulphate. The gel in the staining mixture (together with control gel in mixture minus shikimate) was incubated at room temperature in the dark until bands developed (minutes-hours) whereupon the gel is transferred to d.w. and photographed.

Most protein gels were photographed by the author on a light box with a Polaroid CU-5 camera and type 665 +ve/-ve film. A red filter was used to photograph Coomassie stained gels by this method. The gel in Figure 4.7 was photographed by the Medical Illustration Unit of Glasgow University. In all figures the direction of electrophoresis is from top to bottom unless otherwise stated.

2.8 Digestion of DNA with restriction enzymes

2.8.1 Buffers

All restriction enzyme digests were carried out, essentially as described in Maniatis et al. (1982), in one of four buffers:

"Low salt" - 10mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 1mM DTT

"Medium salt" - 50mM NaCl, 10mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 1mM DTT

"High salt" - 100mM NaCl, 50mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 1mM DTT

"SmaI" - 20mM KCl, 10mM Tris-HCl pH 8.0, 10mM $MgCl_2$, 1mM DTT

These were made up as 5x stocks. The glass-ware, plastic-ware, d.w., stock NaCl, Tris-HCl, KCl, and $MgCl_2$ solutions used were all first autoclaved. The stock DTT solution was made up in autoclaved d.w. The 5x stocks were stored at -20°C in small aliquots. In general, all manipulations involving DNA were done using autoclaved materials and implements wherever possible.

2.8.2 Conditions

Analytical digests were commonly done in a 20 µl volume. The appropriate buffer and temperature were selected using Maniatis et al. (1982) or the restriction enzyme supplier's data sheets. Nuclease-free bovine serum albumin was added to all restriction digests to a final concentration of 100 µg/ml.

A typical analytical digest would consist of:

4 µl 5x buffer

2 µl 1 mg/ml nuclease-free bsa

0.1- 1 µg DNA

restriction enzyme suspension

(autoclaved) d.w. to 20 µl

Most of the restriction enzymes can be used in more than one of the general buffers thus facilitating double digests. However, where no common buffer was usable the lower salt digestion was done first and then the digest conditions adjusted. For double digests involving SmaI the DNA was purified after the first digest (see Section 2.12).

2.8.3 Restriction mapping

This was done mainly by using a combination of single and double digests and logic. Sequential digests of fragments cut out of low melting point agarose gels were also used occasionally. These were done exactly as described in Maniatis *et al.* (1982). Fortuitous partial digestions were often very useful. The gel systems used to separate and size the products of mapping digests are considered in the next section.

2.9 Gel electrophoresis of DNA

DNA was most commonly run in horizontal submerged agarose slab gels essentially as described in Maniatis *et al.* (1982). The dimensions of the agarose gels were either

11 (long) x 13 x 0.7 cm or 22 x 13 x 0.7 cm. The Tris-borate (TBE) buffer system was used here, as for all DNA gels. The 10x stock contains, per litre:

108 g Tris base

55 g boric acid

9.3 g Na_2 EDTA.$2H_2O$

For analysing very small DNA restriction fragments polyacrylamide slab gels were used, with the Tris-borate buffer system, as in Maniatis et al. (1982). The slab gel apparatus used was the same as that described above for protein gels. (Polyacrylamide sequencing gels are considered in Section 2.14).

Samples for both agarose and polyacrylamide gels were loaded after adding 0.1 - 0.2 volumes of "sample loading buffer": 50% (w/v) sucrose, 0.25% (w/v) bromophenol blue.

After electrophoresis, DNA gels were stained in 1 µg/ml ethidium bromide (in d.w.), destained briefly in d.w., examined on a long-wave U.V. transilluminator (U.V. Products Inc.), and photographed (if desired) through a red filter using a Polaroid CU-5 camera with type 665 +ve/-ve film. In all figures, the direction of electrophoresis is from top to bottom unless otherwise stated.

The sizes of restriction fragments were determined by running marker fragments of known size alongside the unknown fragments. The most commonly used sets of marker fragments were phage λ cI857S7 cut with HindIII, λ cut with HindIII

+ EcoRI, and plasmid pAT153 cut with HinfI. The sizes of these sets are given in the appropriate figure legends. The known sequence of pAT153 also allowed precise tailoring of markers for sizing particular fragments.

Isolation of restriction fragments from agarose gels was initially done by the method of Dretzen et al. (1981) in a few early experiments. However, this method was soon abandoned in favour of the more convenient low melting point agarose technique which was performed as described in Maniatis et al. (1982) with the following modifications:

(i) the diluted molten agarose is cooled from 65°C to 37°C (not room temperature) prior to extraction with phenol.

(ii) 3 phenol extractions were performed (rather than just 1) to reduce contamination of the purified DNA with agarose.

2.10 Isolation of plasmid DNA

2.10.1 Large scale isolation of pure plasmid DNA

The large scale isolation of plasmid DNA, in pure form, was done using a scaled up version of the Birnboim and Doly (1979) alkaline/SDS procedure followed by CsCl/ethidium bromide isopycnic density gradient centrifugation.

The isolation was usually done from a 500 ml culture (grown in a 2000 ml conical flask on an orbital shaker). In some early preparations chloramphenicol amplification was

used but it was generally found more satisfactory simply to grow the culture to stationary phase. The cell pellet from 500 ml of culture is resuspended in 20 ml of lysis solution and thereafter the volumes follow the proportions used in the Birnboim and Doly (1979) "mini-prep". Following the first precipitation (with isopropanol) the pellet of nucleic acids is resuspended in 10mM Tris-HCl pH 8.0, 2mM Na_2 EDTA (TE) and subjected to CsCl/ethidium bromide isopycnic density gradient centrifugation as described in Maniatis et al. (1982). The lower of the two bands in the gradient, which contains supercoiled plasmid DNA, is harvested by suction. Ethidium bromide is removed by 5-6 extractions with isopropanol (saturated with TE, saturated with CsCl). The CsCl is removed by dialysis against TE. Purified plasmid DNA was routinely stored at 4°C in TE containing a tiny drop of $CHCl_3$. For long term storage (years) the plasmid DNA was kept at 4°C in the dark in 1.5M NaCl, 10mM Tris-HCl pH 8.0, 2mM Na_2 EDTA plus a tiny drop of $CHCl_3$.

2.10.2 Rapid small scale isolation of plasmids

It is very useful to have a method for quickly isolating a small, impure but restrictable sample of a recombinant plasmid from a small O/N culture. This permits many clones to be screened for the presence of a recombinant with the desired insert. The method of Holmes and Quigley (1981) was used for this purpose. This gives DNA which cuts readily with restriction enzymes and can also be used successfully

to transform bacteria. These mini-preps are stable when stored at -20°C.

2.11 Isolation of phage λ spc1 DNA

2.11.1 Isolation of phage λ spc1

E.coli strain NO1267 was used to prepare phage λspc1. This strain (see Table 2.1) is a double lysogen of the defective λspc1 and the helper λcI857S7. The helper phage strain carries the standard temperature sensitive allele of the cI immunity repressor gene and is also defective in the lysis function, having an amber mutation in the S gene. The practical implications of this are that:

(i) NO1267 can be grown at 30°C as a stable double lysogen.

(ii) Both types of prophage can be induced to enter the lytic cycle by heating the cells to 42°C.

(iii) One must lyse the cells.

(iv) One obtains, initially, a mixture of the two kinds of phage from which λspc1 must be separated prior to the isolation of λspc1 DNA.

The method used for the isolation of λspc1 phage is a modified version of that given in Maniatis et al. (1982), based on advice from D.W. Meek (personal communication): Inoculate 25 ml Nutrient Broth + streptomycin with a single colony of E.coli NO1267 and grow O/N at 30°C. Inoculate 2 x 500 ml Nutrient Broth (in 2000 ml conical flasks; prewarmed to 30°C)

each with 10 ml of fresh O/N. Incubate at 30°C on an orbital shaker with vigorous aeration until A_{650} = 0.45. Induce by placing the flasks in a 42°C shaking waterbath for 35 minutes, then incubate at 39°C with vigorous orbital shaking for 2.75 h. Harvest the cells by centrifugation at 4°C and resuspend in 20 ml SM phage buffer: 100mM NaCl, 8mM $MgSO_4$, 50mM Tris-HCl pH 7.5, 0.01% (w/v) gelatin. Add 0.4 ml chloroform and shake vigorously at 37°C for 10 minutes to lyse the cells. Add deoxyribonuclease I (10 µg/ml final) and shake for 10 minutes at 37°C. Pellet cellular debris (4°C, 10,000 g, 10 min). The supernatant is the crude phage suspension. The phage are then partially purified by centrifugation on a CsCl step density gradient: 3 steps - densities of 1.7, 1.5, and 1.3 g/ml - with the CsCl dissolved in SM phage buffer. Layer 1 ml of each step in two 14 ml polycarbonate tubes for the MSE 6 x 14 ml Ti swing-out rotor (MSE Prepspin 50 centrifuge) or equivalent, then layer 9 ml of crude phage suspension in each tube. Spin for 2 h at 33 krpm, 20°C. The phage band lies at the interface between the 1.3 and 1.5 g/ml steps and is harvested by suction from above. It should have a strong bluish-white opalescence. The λ*spc*1 transducing phage is then separated from the helper phage (and both further purified) by a shallow CsCl equilibrium density gradient: adjust the phage suspension from the step gradient to a density of 1.5 g/ml using CsCl and SM phage buffer. Spin at 27 krpm, 20°C for 18-24 h in an MSE 10 x 10 ml aluminium angle rotor (119; MSE Prepspin 50 centrifuge). Let

the rotor coast to a halt without braking. Harvest both bands individually - the top band is λspc1. Dialyse the phage extensively against SM phage buffer to remove the CsCl.

2.11.2 Determination of phage titre

Preparation of plating bacteria. Inoculate 50 ml of L Broth (with 0.2% (w/v) maltose rather than D-glucose) with a single colony of E.coli Ymel (see Table 2.1) and grow for 10 h in a 250 ml conical flask on an orbital shaker. Pellet the cells and resuspend in sterile 0.01M $MgSO_4$. Store at 4°C. This plating bacteria suspension can be used for up to 3 weeks but better results are obtained if it is used fresh.

Plating the phage. Make serial 10-fold dilutions of phage stock in SM phage buffer (see above). Add 0.1 ml of plating bacteria suspension to 0.1 ml of each dilution to be assayed. Incubate 37°C, 10-20 min. Add 3 ml of λ phage top agar at a temperature of 45°C, mix quickly but gently, and pour on to a plate of bottom agar. Incubate plates at 37°C O/N. Count plaques.

2.11.3 Isolation of phage λ DNA

This was done, for both λspc1 and λcI857S7, by the method given in Maniatis et al. (1982).

2.12 Construction of recombinant plasmids ("cloning")

2.12.1 General points

Methods for restriction enzyme digestion of DNA have already been described (see Section 2.8). As mentioned briefly there, it is necessary when manipulating small quantities of DNA (especially when the condition of the "ends" is important) to avoid contamination with deoxyribonucleases. Thus, all Eppendorf tubes and plastic pipette tips should be autoclaved before use. Likewise, all solutions should be autoclaved if possible or, where there are labile components, made up with autoclaved d.w. All the methods considered below can be found, in some form, in Maniatis et al. (1982). An accessible Eppendorf centrifuge is essential.

"TE" = 10mM Tris-HCl pH 8.0, 2mM Na_2 EDTA.

"Phenol:chloroform" refers to a 24:24:1 mixture of (respectively) phenol:chloroform:isoamyl alcohol which has been saturated with TE. "Chloroform" refers to chloroform which has been saturated with d.w.

2.12.2 Preparation of vector and insert DNA for cloning

After restriction enzyme digestion the DNA must be purified prior to ligation: adjust the volume of the completed digest to 150 µl with TE. Extract once with an equal volume of phenol:chloroform (unless otherwise stated retain the aqueous phase). Back extract the phenol:chloroform with 80 µl TE and combine the aqueous phases. Extract with an

equal volume of phenol:chloroform followed by three extractions with chloroform. Adjust to 0.25M sodium acetate with an autoclaved 4M stock solution, pH 7.4. Precipitate the DNA: add 2.15 volumes of absolute ethanol and either freeze in dry-ice/methanol (10 min) or leave at -20°C for 1 h or longer. Spin in the Eppendorf centrifuge for 10-15 min (room temperature). Remove the supernatant carefully - the desired pellet is often "invisible". Dry the pellet very briefly under vacuum. Resuspend in a suitable volume of TE and store at -20°C until required. It is usually wise to run a small sample of the purified DNA on an agarose gel to check that the DNA has not been lost. Note that DNA purified from low melting point agarose gels is ready for use in ligations.

2.12.3 Dephosphorylation of DNA

In some experiments the cut vector was dephosphorylated to prevent self-ligation. Calf intestinal alkaline phosphatase was used (see Section 2.1.3) and the method employed was that given in Maniatis *et al.* (1982) with the changes shown below. 1 unit of calf intestinal alkaline phosphatase (CIP) activity is defined as the amount of enzyme which releases 1 μmole of p-nitrophenol from p-nitrophenyl phosphate per minute at 37°C in 0.1M glycine/KOH pH 10.5, 1mM $MgCl_2$, 0.1mM $ZnCl_2$. Vector DNA to be dephosphorylated was purified after restriction enzyme digestion as described above. Spermidine was omitted from the dephosphorylation buffer.

After the heat step the DNA was again purified as described above.

2.12.4 Ligations

These were performed, using T4 DNA ligase, in a final volume of 20 µl. The 10x ligase buffer used was 660mM Tris-HCl pH 7.6, 66mM $MgCl_2$ and a ligation contained:

vector DNA

insert DNA

d.w. to 20 µl

2 µl 10x ligase buffer

2 µl 5mM ATP

2 µl 0.1M DTT

T4 DNA ligase

Before adding ligase heat the mixture to 65°C for 1 minute. Incubate the completed mixture 2-3.5 h at 14°C followed by 1.5 h at 22°C. Then transform the appropriate host bacteria. With phosphorylated vector DNA it is better to use a molar excess of insert DNA if possible (say 3-fold). However, with dephosphorylated vector it is better to use a molar excess of vector DNA if possible (say 3-fold) since this reduces the number of recombinants with multiple inserts.

2.13 Transformation of bacteria with plasmids

The method of Dagert and Ehrlich (1979) was used. Prior to plating on media containing ampicillin or tetracycline a standard expression period of 90 min was allowed.

2.14 M13/dideoxy sequencing of DNA

2.14.1 Technical procedures

A detailed summary of the M13/dideoxy sequencing method is given in Section 3.5.1. The detailed procedures were carried out as described in Amersham's "M13 Cloning and Sequencing Handbook" (1983). Where a choice of methods and materials is possible those actually used are described below. The basic materials for M13 cloning and sequencing were obtained in the form of kits from Amersham as described in Section 2.1. The M13 Cloning Kit N.4501 was used together with the M13 Sequencing Kit N.4502.

When preparing SS M13 template DNA an additional (optional) step was used immediately before the ethanol precipitation step, namely one extraction with 200 µl of chloroform to remove residual polyethylene glycol 6000 since the latter gives rise to artefactual bands on the sequencing gels. When a particular template was found to give artefactual bands either a different isolate of the same clone was sequenced or the original preparation was re-extracted with chloroform.

The radioactive label used was $[\alpha\text{-}^{35}S]$ dATPαS.

The 6% polyacrylamide sequencing gels used were 20 x 40 x 0.04 cm and were not poured using the buffer gradient system. Different loadings of the same sequencing reactions were used to obtain different running times and thus maximise the length of sequence that could be determined from each template. Usually (but see below) one loading was run for about 1.75 h

and another for about 4.5 h either on the same or separate gels. In early experiments gels were run at 25-28 mA without any preheating. A number of troublesome "compressions" of the band spacing occurred as a result. Later all gels were preheated (30 min, 33 mA) before the samples were loaded, and then run at 30-32 mA. This cured all "compression" problems.

After electrophoresis the gels were fixed and then dried on Whatman 3MM paper using a Bio-Rad Model 1125 gel dryer before autoradiography (16-48 h) using Fuji RX film.

In one case the technique of clone "turn-around" was used (see Chapter 3). The DS replicative form was prepared from SS template DNA using the *in vitro* method.

For extended sequencing up to 400 bases the concentration of ddNTP was reduced to 0.75 of that normally used and the period of electrophoresis was extended to 5.5 h.

2.14.2 Compilation of the sequence

At every stage between reading the sequencing gel autoradiographs and the final print-out from the computer a system of multiple checks was used in an effort to prevent errors. However, by far the best line of defence against mistakes was the sequencing of both strands.

The primary data from the sequencing gels were recorded on specially designed sheets. While reading the autoradiographs attention was paid to the following possibilities:

(i) weak bands (in the upper reaches of a gel) especially the first C of a doublet or an A at the end of a run of A's

(ii) artefactual bands

(iii) compressions, especially in the vicinity of C/G rich regions.

A ruler was used very frequently to check the spacings between bands by comparison with adjacent sets of tracks. It was often useful to measure the spacing between bands well above and below a suspect region and to compare the total number of bands within that wide region with the expected total.

Overlapping sequences and complementary strands were quite easily identified by visual inspection of the auto-radiographs.

All doubtful regions were so indicated on the primary data sheets. It was always possible to deduce the sequence of a region, using both strands, even in areas of bad compression. However, as a precaution, where there was ambiguity in the same vicinity on both strands this region was sequenced again on both strands using gels run at a higher current (hence higher temperature) until the sequence could be read unambiguously on both strands. Even where there was only ambiguity on one strand and the other strand was unambiguous both strands were sequenced again (except in cases where the original ambiguity was clearly a faint artefactual band with no trace of a complementary band in the other strand, and no anomalies in the spacing). The original "consensus" sequence proved correct in all cases.

When a fully overlapping sequence had been obtained the consensus sequence for each strand was entered into a Digital PDP 11-34 computer using part of the program BATIN (Staden, 1980). The sequence of each strand was entered as segments (each into a separate file). A print-out of each file was subsequently checked against the original consensus sheets. As the main check on the accuracy of the computer files the files containing complementary strands were compared using the program TTEM (R. Eason, unpublished). A perfect match was found in all cases. Each file was entered beginning and ending with half a restriction site. This greatly facilitated the final merger of files (in the correct order) to construct the complete sequences of both strands of the HindIII-ClaI fragment. The total number of bases in the files for each whole strand were equal, and equal to the number expected from the sum of the component files.

Although all sequences had been checked visually for double inserts the program CUTSIT was used to locate all restriction sites in the final sequence. All known sites were found but no unexpected ones.

2.14.3 Analysis of the sequence data

Except for the DOTPLOT program (see below) all programs were run on a Digital PDP 11-34 computer.

The program TRNTRP (Staden, 1978) was used to translate DNA sequences in any desired reading frame. The relative positions of stop codons in all six frames were plotted manually.

The program HAIRPN (Staden, 1978) was used in a limited search for palindromic structures. Only those structures with "loop" sizes $\geqslant 3$ and $\leqslant 10$ bases would have been found by this search and, furthermore, only those where the symmetry related elements immediately flanking the central "loop" were $\geqslant 4$ bp long. A print-out of the positions of palindromic structures meeting these conditions was obtained. Each one was examined with a view to extending it.

The DNA sequence was examined for speculative promoters in two ways. Firstly, the region upstream from aroE was checked manually using a print-out of the sequence and a double (sliding) card system to accommodate the variable spacing between the "-35" region and the Pribnow box. The assumption was made that the final T in the Pribnow box would be conserved (see Section 5.9.1). This assumption is valid for 108 of the 112 well-established promoters listed by Hawley and McClure (1983). Only promoters acting in the BamHI to HindIII direction were searched for. Each T upstream from aroE was checked to see if the sequences further upstream loosely fitted the consensus E.coli promoter sequence (see Section 5.9.1).

Secondly, a SEARCH program was used to look for sequences, in the ClaI-HindIII sequence, differing from the Pribnow box consensus by only a single base and for sequences differing by no more than one base from the first five bases (the most conserved) of the "-35" region consensus sequence. All "finds" were examined manually to see if anything approximating

the other region/box was present.

The amino acid sequences of *E.coli* shikimate dehydrogenase and *N.crassa* catabolic quinate/shikimate dehydrogenase were compared using the DOTPLOT program from the WISGEN software package (Devereux *et al.*, 1984). This package was devised by the Wisconsin Genetics Computer Group and was available on the Edinburgh Regional Computing Centre's VAX 11/750, from where DOTPLOT was run with the assistance of A. Coulson.

2.15 Large scale purification of overproduced E.coli E3 from pIA321//AB2834

2.15.1 Growth of cells

7 x 50 ml L Broth plus ampicillin, in 250 ml conical flasks, were each inoculated with 2.5 ml of an O/N culture of pIA321//AB2834 (grown in L Broth + ampicillin) and incubated on an orbital shaker for 2 h at 37°C. Then 13 x 500 ml prewarmed L Broth + ampicillin, in 2000 ml conical flasks were each inoculated with 25 ml from the 50 ml cultures and shaken at 37°C for 5.25 h. 1.25 h after this inoculation the *lac* inducer IPTG was added to each 500 ml culture to a final concentration of 5×10^{-4}M. After 5.25 h of growth the cultures had reached stationary phase (A_{650} = 3.7 - 3.8) and were removed to ice prior to harvesting by centrifugation (MSE 6 x 750 ml rotor, 10°C, 15 min, 5 krpm). The cells were stored at -20°C until required.

2.15.2 Purification of E3

This was carried out almost exactly as described in Chaudhuri and Coggins (1985) with a few minor changes which are described below. The method is summarised in Chapter 4 and further details are given in the figure legends.

Buffers were prepared from a stock of 1M Tris-HCl pH 7.5 without any adjustment to the pH after dilution.

Buffer 1 (extraction): 100mM Tris-HCl pH 7.5, 1mM Na_2 EDTA, 0.4mM DTT, 1.2mM PMSF.

Buffer 2 (post-ammonium sulphate dialysis, equilibration and washing of DEAE-Sephacel column): 50mM Tris-HCl pH 7.5, 50mM KCl, 0.4mM DTT, 1.2mM PMSF. The salt gradient was based on this buffer.

Buffer 3 (post-gradient dialysis, equilibration of 2 ml mini-DEAE-Sephacel column for concentrating enzyme): 50mM Tris-HCl pH 7.5, 0.4mM DTT, 1.2mM PMSF. This is twice the concentration of Tris-HCl used by Chaudhuri and Coggins.

Buffer 4 (elution from 2 ml DEAE-Sephacel column) = Buffer 3 containing 1M KCl.

Buffer 5 (equilibration and running of Sephacryl S-200 superfine column): 500mM KCl, 50mM Tris-HCl pH 7.5, 0.4mM DTT.

Buffer 6 (post-Sephacryl dialysis and equilibration of ADP-Sepharose column): 25mM Tris-HCl, 0.4mM DTT.

Buffer 7 (elution from ADP-Sepharose column) = Buffer 6 containing 1mM $NADP^+$.

Buffer 8 (first post-ADP-Sepharose dialysis, to remove $NADP^+$): 50mM Tris-HCl pH 7.5, 0.4mM DTT, 1mM benzamidine.

Buffer 9 (storage dialysis buffer): 50% (v/v) glycerol, 50mM KCl, 50mM Tris-HCl pH 7.5, 0.4mM DTT, 1mM benzamidine. The storage buffer used by Chaudhuri and Coggins (1985) did not contain any KCl.

PMSF (in ethanol) and DTT were added just before use.

Due to the limited capacity of ADP-Sepharose and the large quantity of E3 present a 20 ml bed volume was used for this column rather than the 5 ml used by Chaudhuri and Coggins.

2.16 Protein sequencing

2.16.1 Reduction and carboxymethylation

This was carried out essentially as described by Lumsden and Coggins (1978). All glass-ware used was acid-washed.

100 nanomoles (3 mg) of pure E.coli E3 was dialysed exhaustively (3 days, 7 changes) against 0.5% (w/v) ammonium bicarbonate. The dialysed material was freeze-dried, re-suspended in 4 ml d.w. and freeze-dried again. The dry E3 was dissolved in 2 ml of 0.1M Tris-HCl pH 8.2, 8M urea (recrystallised), 2mM DTT and incubated in the dark, under N_2, at room temperature for 1 h. The solution was then made 15mM in iodoacetic acid (from a fresh 300mM stock solution, pH adjusted to 7.5 with NaOH) and incubated in the dark for a further 1 h. The reaction was terminated by the addition of

DTT to a final concentration of 30mM. The solution was then dialysed exhaustively against 0.5% ammonium bicarbonate. The dialysed material was freeze-dried, taken up in d.w., and freeze-dried twice more.

2.16.2 Automated N-terminal amino acid sequencing

This was done in collaboration with J.E. Fothergill, L.A. Fothergill, and B. Dunbar at the SERC funded protein sequencing facility at Aberdeen University. The determination of the N-terminal amino acid sequence was carried out using a Beckman Model 890C automatic liquid phase sequencer operated by B. Dunbar as described in Smith *et al.* (1982). Thiazolinone samples were collected into tubes containing 0.2 ml of freshly prepared 1M HCl containing 1% (v/v) ethanethiol, thus readily permitting the detection of serine and threonine residues. The phenylthiohydantoin samples were analysed by high pressure liquid chromatography on a Waters 5 μm "Resolve" C_{18} reverse phase column with a pH 5.00 Na acetate-acetonitrile buffer system (Carter *et al.*, 1983). S-carboxymethylcysteine was used as an internal standard.

2.17 Amino acid analysis of E.coli E3

2.17.1 General points

The starting material was *E.coli* E3 purified to homogeneity from the overproducing strain pIA321//AB2834 (see Section

2.15 and Chapter 4). The E3 was almost homogeneous by the criterion of SDS PAGE after the penultimate stage of the purification. The last step (affinity chromatography on ADP-Sepharose) is a powerful one (giving more than a 40-fold purification during the purification of E3 from wild-type E.coli) thus one might expect the purified, overproduced E3 to have an unusually low percentage of residual contamination with other proteins. This was possibly an important factor in the good agreement between the observed amino acid composition and that predicted from the gene sequence.

Great care was taken at all stages of the analysis to prevent contamination with extraneous proteins or amino acids. All glass-ware used was soaked overnight in concentrated nitric acid before being rinsed exhaustively with tap water followed by d.w., drained and then dried in an oven. The clean glass-ware was stored in a new seal-top plastic container which had been thoroughly scrubbed. Two pairs of disposable plastic gloves (from a dedicated box) were worn at all times and changed regularly. Plastic pipette tips were taken from a dedicated, unpunctured bag normally kept sealed.

2.17.2 Performic acid oxidation

30 nanomoles (0.9 mg) of pure E.coli E3 was dialysed exhaustively against 0.5% (w/v) ammonium bicarbonate. DL-nor-leucine was then added as an internal standard and the material was freeze-dried, redissolved in a few ml of

d.w., and lyophilised again.

Performic acid oxidation was then carried out (to convert the cysteine and methionine residues to derivatives stable under the subsequent acid hydrolysis conditions) by the method of Hirs (1967). The oxidation was terminated by dilution with d.w. followed by lyophilisation. After redissolving in d.w. and further lyophilisation the "oxidised" protein was dissolved in 200 µl 99% formic acid, diluted up to 2 ml with d.w., divided equally between 4 pyrex test-tubes (avoiding contact with the sides) and freeze-dried once more prior to acid hydrolysis.

2.17.3 Acid hydrolysis

Each of the above tubes received 500 µl of 6M HCl (Aristar) containing 0.17% (v/v) 2-mercaptoethanol, without any of the acid touching the sides of the upper part of the tube. The tubes were then sealed under vacuum (with the assistance of J.R. Coggins) and placed in a 105°C heating block. One tube was removed after 23.9, 47.8, 72.1, and 94.9 hours. Tubes were stored at -20°C until all were ready, whereupon they were opened and desiccated over concentrated H_2SO_4 and NaOH pellets, under vacuum. After resuspension in d.w. and further desiccation as before the samples were left O/N under simple vacuum prior to resuspension in 125 µl and analysis.

2.17.4 Analysis and data processing

The protein hydrolysates were analysed on an LKB Model 4400 amino acid analyser (with integrator) operated by J. Jardine. Data processing was performed by the author except for integration of the proline peaks from the original chart recorder traces which was done by J. Jardine.

Standards were run immediately before and after every set of 3 or 4 experimental runs. Although there was not usually much difference an average value of the nmol/area figure was calculated for each standard amino acid (in all cases).

For a given loading, the 95 hour timepoint values were all about 3.3 times lower than those for the 24, 48, and 72 hour timepoints. However, the DL-nor-leucine values were, incomprehensibly, very similar in all four cases. For this reason the 95 hour timepoint data were discarded.

For each of the other three timepoints two 5 µl loadings, one 10 µl loading, and one 25 µl loading were run. Data from the 25 µl loadings were not used except for cysteic acid (see below) because many of the peaks were overloaded and merged with adjacent peaks.

Considerable problems were encountered with the integrator. Peaks were often misassigned by the machine so print-outs were examined by eye and the appropriate calculations performed manually (in all cases) using correctly assigned areas for each peak. More seriously, for unknown reasons the integrator sometimes made errors in deciding which points to integrate

between and with what baseline. Thus, the integration markers were examined for every peak in every print-out (including the standard runs) and data from peaks with obvious integration errors were discarded. This was done before calculating the predicted amino acid composition from the gene sequence and before calculating the experimental composition.

Except for threonine, serine, valine and isoleucine (see Chapter 4) the values, in nmol, for the three different runs with each of the three different timepoints (minus integrator error values) were simply averaged (but see below). Each average was based on the following number of values: asx(9), methionine sulphone(9), glx(5), gly(9), ala(9), leu(9), tyr(8), phe(9), his(9), lys(9), arg(7). The values, in nmol, for each amino acid within a given run were summed and from this figure the values for each a.a. were normalised to give a total of 100 nmol of amino acids per run. This allows one to average a different number of values for particular amino acids, without distorting the true experimental ratio, where two different sizes of loading have been used.

With the cysteic acid peak the integrator made an error in every run (including standard runs), starting the integration from the bottom of a sharp, shallow valley of roughly constant size which always preceded the main peak. This error was clearly of least significance for values calculated from the largest (25 µl) loadings. Fortunately the cysteic acid peak

is well isolated so merging of peaks is not a problem. Also, with the 25 µl loading the experimental peaks were about the same size as the standard peaks thus effectively removing the error. However, since the 25 µl loadings (for all three timepoints) always tended to give slightly higher values for all the other amino acids the value for cysteic acid, calculated from the average of the 24, 48, and 72 hour 25 µl runs, was normalised using the values for arginine.

CHAPTER 3 SUBCLONING AND DNA SEQUENCING OF THE E.COLI AROE GENE

3.1 Preliminaries

3.1.1 Isolation of phage λspc1 DNA

λspc1 was prepared from E.coli strain NO1267 (Jaskunas et al., 1975a). This strain is a double lysogen of the defective λspc1 and of the helper λcI857S7. In Chapter Two there is a detailed description of the isolation of the two kinds of phage, of their separation, and of the extraction of their DNA. λspc1 is known to form the upper band in the CsCl equilibrium density gradient which is used to separate the two types of phage (D.W. Meek, personal communication). The two phage bands in the equilibrium gradient were of approximately equal intensity. The infective titre of the λspc1 band was much lower than that of the helper band, as expected:-

upper band: 1.1×10^{11} plaque-forming units per ml.

lower band: 3.3×10^{12} plaque-forming units per ml.

Thus, it appears that the defective λspc1 phage were slightly contaminated with approximately 3% of helper phage - a negligible amount. The identity of the two phage DNAs was confirmed by analysis of restriction fragment patterns. A restriction map of λspc1 was available (R. Hayward, personal communication). It had been compiled from the work of many people in different laboratories (C. Ma, D.W. Meek, and R. Hayward, unpublished results; Jaskunas

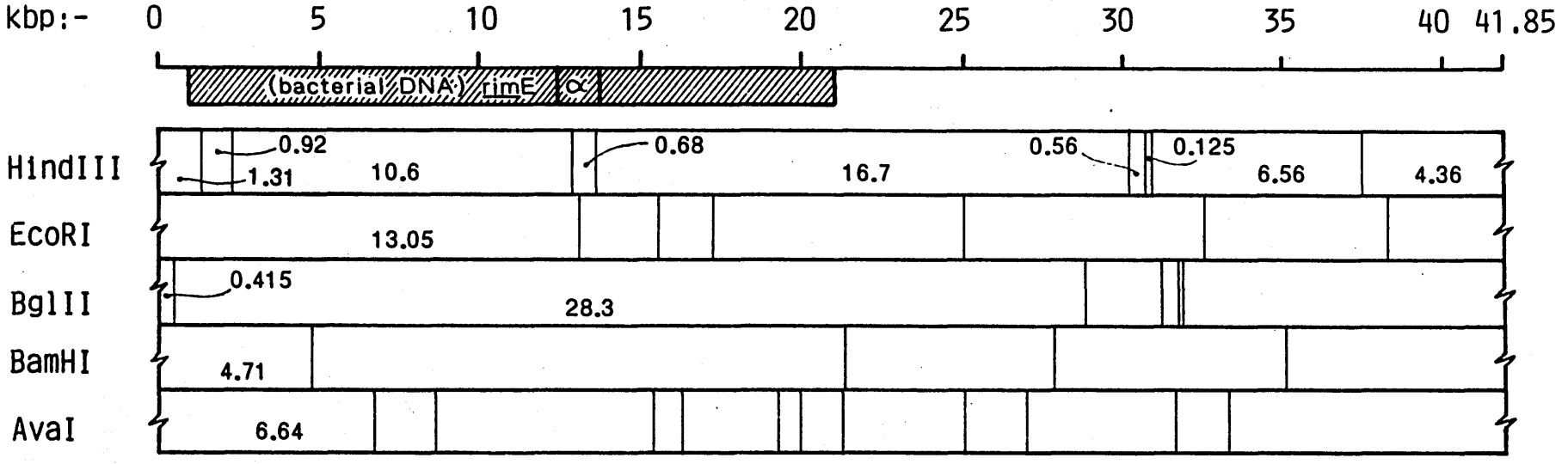


Figure 3.1 λspc1 restriction sites relevant to the initial subcloning of aroE

The restriction sites (indicated by vertical lines) are taken from the map compiled by R. Hayward and his collaborators (see text). As described in Chapter One λspc1 was thought to carry aroE at its left hand end. α denotes the location of the rpoA gene (RNA polymerase α subunit). Restriction fragment sizes are in kbp.

et al., 1975b). Figure 3.1 shows the sites relevant to the initial subcloning of aroE. Single digests of the presumptive λspc1 DNA with the restriction enzymes HindIII, EcoRI, SalI, and BglII gave DNA fragments of the sizes expected from the map (data not shown). Digestion of the presumptive λcI857S7 DNA with the same enzymes gave patterns identical to those of the standard λcI857S7 DNA that was used to make the restriction fragment size markers. These latter patterns are readily distinguishable from those of λspc1.

3.1.2 E.coli AB2834

We tested our stock of E.coli AB2834 to confirm that it was still an aroE$^-$ auxotrophic mutant. In plate tests the strain behaved as expected (see Table 3.1). Minimal medium must be supplemented with all three aromatic amino acids plus p-aminobenzoate and p-hydroxybenzoate to allow E.coli AB2834 to grow well. This implies a block in the common pathway. The failure of AB2834 to respond to a DHS supplement together with its good response to a supplement of shikimic acid strongly suggest that the defect is in the aroE gene. These common pathway intermediate supplements were given together with phenylalanine and tyrosine. DHQ, DHS, and shikimic acid alone can be growth factors for suitable mutants. However, some mutants respond poorly to these compounds if they are given alone but very well if the intermediate is given along with tyrosine and phenylalinine (Davis and Weiss, 1953; Ely

Table 3.1 Growth of E.coli strains AB2834 and K12 on solid media

Supplement	Response of AB2834	Response of K12
none	-	+++
trp	-	"
phe	-	"
tyr	-	"
trp + PABA + PHBA	-	"
phe + tyr	-	"
phe + tyr + trp	-	"
phe + tyr + trp + PABA	+	"
phe + tyr + trp + PHBA	+/-	"
phe + tyr + trp + PABA + PHBA	+++	"
phe + tyr + DHQ	-	"
phe + tyr + DHS	-	"
phe + tyr + shikimic acid	+++	"

NOTES: 1. In all cases the supplements were added to minimal medium.

2. Details of the various media are given in Chapter Two.

3. Key: +++ = wild-type growth
++ = moderate growth
+ = slight growth
+/- = barely detectable growth
- = no growth detectable

4. E.coli K12 is taken as wild-type.

and Pittard, 1979). For an *aroE*$^-$ mutant the following scheme was elucidated: the strain accumulated DHS heavily and this competitively inhibited utilisation of shikimic acid. With increasing inhibition the various end products of aromatic synthesis were eliminated in a fixed order, tyrosine and phenylalanine being the first to be lost (Davis and Weiss, 1953). As a control for the stability of DHS in plates an E2$^-$ *aroD* auxotrophic strain was tested. This strain, *E.coli* AB2848, responded well to a DHS supplement which confirmed the stability of DHS in plates.

As a further check crude cell extracts were made from *E.coli* strains AB2834, AB2848, and K12. Each of these extracts was assayed for E3 activity. The results are shown in Table 3.2. There was no detectable E3 activity in AB2834 in contrast to the activities of the AB2848 and K12 control strains. Thus it was concluded that our stock of AB2834 was almost certainly a genuine *aroE*$^-$ mutant.

3.2 Initial subcloning of aroE from λspc1

3.2.1 Isolation of the first putative positives

The available restriction map of λ*spc*1 (Figure 3.1) showed a BglII site at the extreme left end, outside the region of transduced bacterial DNA. The next BglII site to the right was over 28 kbp away. The distance from the left end to the first EcoRI site was 13 kbp and this site was known to be in the RNA polymerase/ribosomal proteins operon (Jaskunas *et al.*, 1975b). This operon lies to the

Table 3.2 E3 specific activities of crude cell extracts of E.coli strains AB2834, AB2848 and K12

Strain	Specific activity of E3 in crude extracts (units/mg)
K12	0.05
AB2848	0.132
AB2834	none detectable

NOTES:
1. The limit of detection was less than 0.0006 u/mg
2. Mixing experiments (e.g. K12 extract + AB2834 extract) were not done but later work involving the activity staining of SDS gels, after renaturation, effectively rules out the possibility of some inhibitory species in the AB2834 extracts.

right of aroE when the E.coli linkage map is orientated with the phage. Thus, the 12.6 kbp BglII-EcoRI fragment from the left end of λspc1 was expected to include the aroE gene. This fragment was cloned in a suitable plasmid vector so that its ability to relieve the auxotrophy of E.coli AB2834 could be tested.

The multicopy plasmid vector pAT153 (Figure 3.2) was used for this experiment (Twigg and Sherratt, 1980; the details of the construction of pAT153 given in the paper are now known to be slightly in error - Old and Primrose, 1981). The copy number of pAT153 is 1.5 to 3 times higher than that of pBR322 so that higher levels of plasmid specified gene products are obtained. The pAT153 used was prepared from a strain (E.coli AB2829) with a functional EcoK host modification system because AB2834 has a functional EcoK host restriction system and the pAT153 sequence contains an EcoK site which if left unmethylated would result in about a 20-fold reduction in the efficiency of transformation of AB2834 by pAT153.

There are no BglII sites in pAT153. However, BglII sticky ends can be ligated with BamHI sticky ends (though the resulting hybrid site is cleavable by neither of the two enzymes). λspc1 DNA was digested with BglII and then with EcoRI, whereas pAT153 was digested with BamHI and then with EcoRI. After purification the DNA from the two digests was ligated and used to transform competent AB2834 cells. (This is, in effect, a "mini-shotgun" experiment).

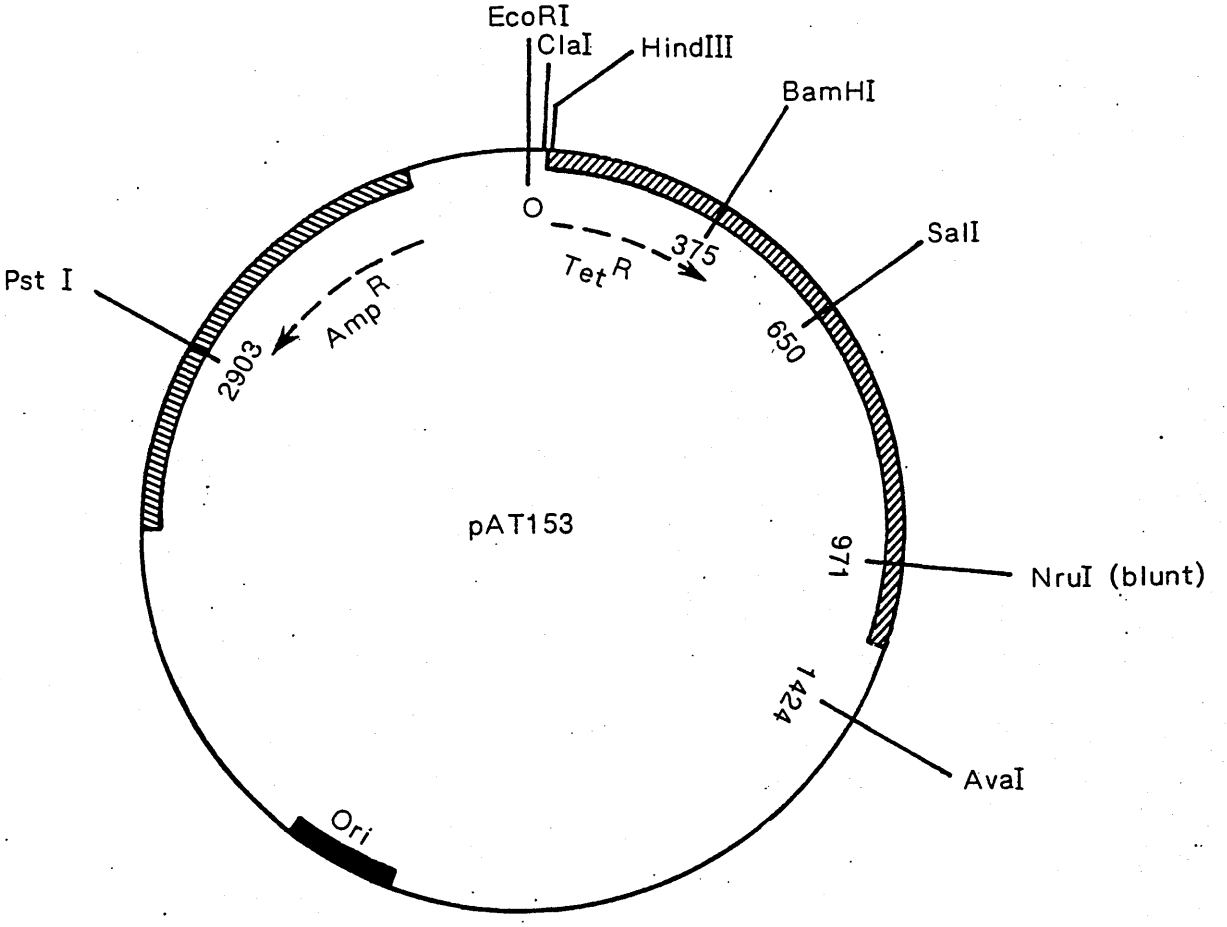


Figure 3.2 Plasmid vector pAT153

The origin of DNA replication is marked "Ori". The two drug resistance markers Amp^R and Tet^R, specifying resistance to ampicillin and tetracycline respectively, are hatched and the direction of transcription is shown for each. Some useful restriction sites are marked and the cleavage coordinates are shown. The total length of pAT153 is 3657 base-pairs: pAT153 is derived in a known way from pBR322 (Twigg and Sherratt, 1980) and the complete DNA sequence of the latter plasmid is known (Sutcliffe, 1979).

The cells were plated onto L agar + ampicillin to select for transformed cells. After overnight growth about 970 colonies were replica plated onto minimal agar. 27 colony replicas grew well on the minimal plates whereas the remainder showed no signs of growth. These 27 colonies constituted the putative positives. As a control 3×10^8 AB2834 cells were plated directly on minimal agar after $CaCl_2$ treatment and transformation with pAT153. No colonies appeared on these control plates which implied that the reversion rate of AB2834 is rather low. The replica plating step was used, rather than direct plating onto minimal medium, to ensure that any cloned aroE genes would have had sufficient time to be expressed in the cell. This was later shown to be an unnecessary precaution: Aro^+ AB2834 cells carrying aroE recombinants can be recovered after plating directly on minimal medium, provided the cells are suspended in rich medium.

3.2.2 Ways in which false positives could arise

There are a number of ways in which one could be misled by false positives:-

(a) The most obvious danger is AB2834 cells reverting to prototrophy. As mentioned above this is a very rare event. Also, since the initial selection is for ampicillin resistance, and since only a tiny proportion of the cells are actually transformed by any kind of plasmid, the chances of getting an Amp^R, Aro^+ revertant are very remote. However, it is still wise to check that the Aro^+ phenotype of a putative

positive is due to a plasmid. This can be done by preparing plasmid on a small scale from several positives and back transforming into AB2834. One would then expect that all Amp^R colonies would also be Aro^+.

(b) AB2834 carries the wild-type recA gene. If a recombinant plasmid carried an inactive version of the aroE gene it is conceivable that complementation might arise from recombination between the nonfunctional plasmid-borne allele and the (different) nonfunctional AB2834 allele. In this case it is very unlikely that all cells which became Amp^R, in a back transformation experiment, would also become Aro^+.

(c) It is possible that the plasmid in a putative positive does not carry the desired gene but rather a gene which suppresses the mutant phenotype of the host cell (e.g. a suppressor tRNA gene could suppress a nonsense mutation in a gene). This possibility can be effectively ruled out, where one is using a high copy number vector, if one can show that cells transformed with the recombinant plasmid express much higher levels of the gene product than are found in normal cells.

3.2.3 Evidence that some of the recombinant plasmids carry aroE

Plasmids isolated from putative positives 4L and 8L and a control sample of pAT153 were used to transform AB2834. The cells were initially plated on L agar + ampicillin and colonies were then replica plated onto both

Table 3.3 Back transformation of AB2834

Plasmid	No. of colonies		
	LAmp	LTet(replicas)	MAmp (replicas)
+ve 4L	69	0	69
+ve 8L	(39	not done	39
	(52	0	52
pAT153	410	410	0

L agar + tetracycline and onto minimal agar + ampicillin. The results are shown in Table 3.3. The two putative positives, 4L and 8L, both yielded plasmids that conferred the Aro^+ phenotype on all transformed cells. Positive 4L was then studied in more detail.

The sensitivity to tetracycline suggested that the plasmids were recombinant. Restriction analysis of the plasmid from 4L showed that it contained the expected 12.6 kbp BglII-EcoRI insert. This recombinant was designated pIA306.

When subcloning from λ*spc*1 by the "mini-shotgun" approach one must be alert to the following possible complications:

(a) cloning a partial digest product of λ*spc*1 and/or cloning into a partial digest of the vector.

(b) Cloning more than one fragment into the one recombinant plasmid.

(c) Cloning a hybrid fragment of λ*spc*1 formed by ligation of the λ cohesive ends. Hence, when considering the possible complete (and partial) restriction digest products of λ*spc*1 it is essential to consider both the linear and the circular forms of the phage.

The shikimate dehydrogenase specific activities of crude cell extracts were determined for the following *E.coli* strains: K12, AB2834, AB2834 transformed with pIA306 (=pIA306//AB2834), HB101, and pIA306//HB101. Dehydroquinase specific activities were also determined as

an internal control. The results are shown in Table 3.4. Cells carrying pIA306 show much higher levels of shikimate dehydrogenase than normal. In contrast, dehydroquinase specific activities are fairly constant between comparable strains. It could be argued that our K12 strain has an abnormally low level of E3 but this criticism is not applicable to the comparison between HB101 and pIA306// HB101. The cultures used to make the latter two crude extracts were taken at almost identical A_{650} values and they grew at almost equal rates. The final A_{650} values were not determined for the other three cultures but there were no great differences between them as judged by eye. (Factors that can influence the expression of a gene carried by a particular multicopy plasmid are discussed in Section 3.3.5 below). It was concluded that E3 was being overexpressed due to the presence of *aroE* on the multicopy recombinant plasmid pIA306. Hence, *aroE* must be present on λ*spc*1 between the BglII and EcoRI sites previously described. Later results demonstrate unequivocally that *aroE* has been cloned but the findings described above justified proceeding further.

Table 3.4 Overexpression of E3 in strains carrying pIA306

E.coli strain	Crude extract E3 specific activity (units/mg)	Crude extract E2 specific activity (units/mg)	Final A_{650} of culture used
AB2834	0	0.032	not determined
K12	0.086	0.032	not determined
pIA306//AB2834	1.7	0.034	not determined
HB101	0.12	0.018	0.82
pIA306//HB101	1.3	0.018	0.78

NOTES: 1. $\frac{\text{E3 s.a.(pIA306//HB101) - E3 s.a.(HB101)}}{\text{E3 s.a.(HB101)}} = 9.8$

2. $\frac{\text{E3 s.a.(pIA306//AB2834)}}{\text{E3 s.a.K12}} = 19.8$

3. All crude extracts were made from cells grown in L Broth supplemented, in the case of strains containing pIA306, with ampicillin to maintain selection (see Chapter Two for details).

3.3 Further subcloning of aroE

3.3.1 Isolation of plasmids pIA305, 307, and 304

Further subcloning experiments were performed to locate the aroE gene more precisely (see Figure 3.3 and Table 3.5). Initially the same approach was used as for pIA306 namely a "mini-shotgun" of a particular λspc1 digest into AB2834 followed by restriction analysis of plasmids isolated from Aro^+ clones. In this way plasmids pIA305, 307, and 304 were obtained. The 2.47 kbp HindIII-BamHI fragment was the only region of λspc1 DNA common to all these recombinants thus implying that aroE was located on this fragment.

3.3.2 Construction of pIA303

The 2.47 kbp common denominator fragment was now cloned specifically. A two stage isolation procedure was used since a HindIII + BamHI digest of λspc1 would yield several fragments of about 2.5 kbp. λspc1 DNA was digested with BamHI and the fragments separated by electrophoresis in a 0.6% agarose gel. The 4.7 kbp fragment (see Figure 3.1) was isolated from the gel by the method of Dretzen et al. (1981) and digested with HindIII. The products of this second digest were separated by electrophoresis in a 1% agarose gel and the 2.47 kbp fragment was isolated and then ligated with cut (HindIII + BamHI) and phosphatased pAT153 vector. The ligation mix was used to transform E.coli HB101 and the cells were plated on L Amp. 9 Amp^R

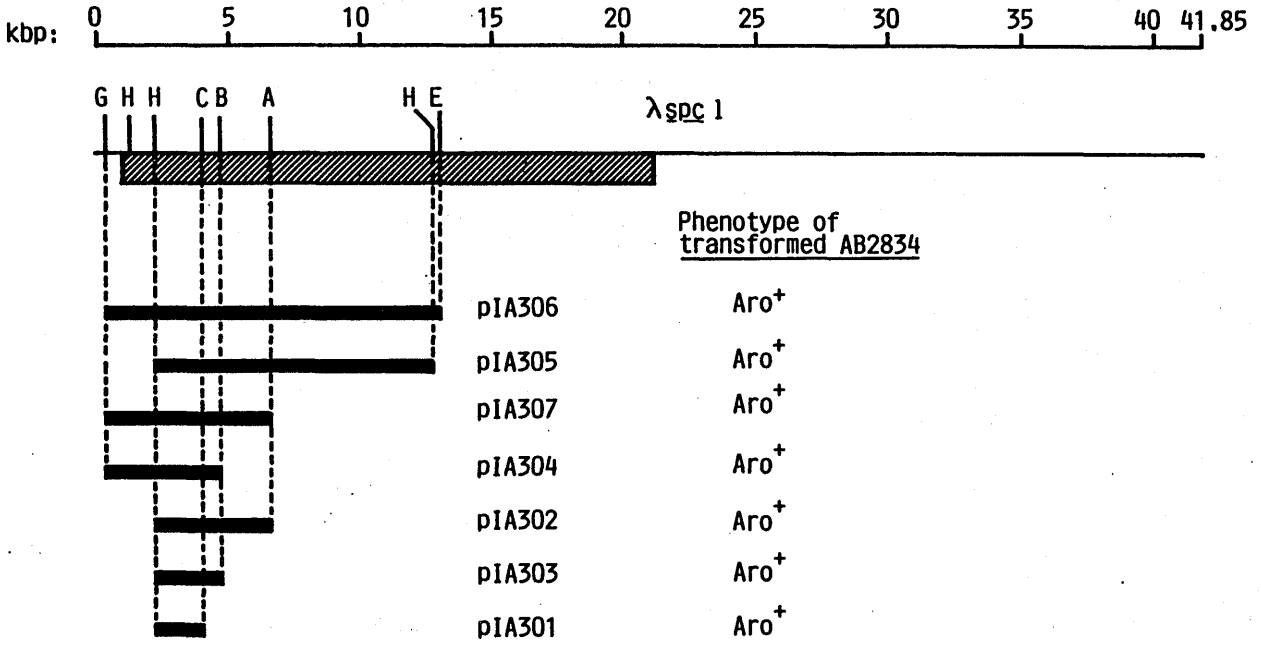


Figure 3.3 Subcloning of aroE from λspc1

The solid horizontal bars represent the fragments cloned in the subclones shown. Key to restriction sites: G = BglII, H = HindIII, C = ClaI, B = BamHI, A = AvaI, E = EcoRI.

Table 3.5 AroE subclones

Recombinant plasmid	Fragment cloned	pAT153 vector digest
pIA306	12.6 kbp BglII-EcoRI from λspc1	BamHI + EcoRI
pIA305	10.6 kbp HindIII from λspc1	HindIII
pIA307	6.2 kbp BglII-AvaI from λspc1	BamHI + AvaI
pIA304	4.3 kbp BglII-BamHI from λspc1	BamHI
pIA302	4.4 kbp HindIII-AvaI from pIA307 (gel)	HindIII + AvaI
pIA303	2.47 kbp HindIII-BamHI from λspc1 (gel)	HindIII + BamHI
pIA301	1.82 kbp HindIII-ClaI from pIA307 (gel)	ClaI + HindIII

NOTES:
1. Fragments in the lower section were cloned specifically after isolation from agarose gels whereas those in the upper section were cloned by the "mini-shotgun" approach.
2. Since the AvaI recognition sequence is degenerate (CPyCGPuG) the isolation of pIA307 was fortuitous.
3. The structures of these subclones are shown in Figure 5.8.

colonies were tested and shown to contain the desired recombinant by restriction analysis. This recombinant was called pIA303. Six of these pIA303 plasmid isolates were shown to confer the Aro$^+$ phenotype on AB2834. These results confirmed the suggestion from earlier subcloning experiments that *aroE* is located on the 2.47 kbp HindIII-BamHI fragment.

3.3.3 Further restriction mapping of the pIA307 insert

The available compilation of λ*spc*1 restriction sites did not include any sites within the pIA303 insert region nor any sites closely flanking this region. Further mapping of restriction sites in the insert of pIA307 was carried out. It was hoped that new sites might allow the position of *aroE* to be better defined. The results are shown in Figure 3.4 while Figures 3.5, 3.6, and 3.7 show examples of the data used to construct the new restriction map.

3.3.4 Construction of pIA301 and pIA302

The newly mapped ClaI site in the 2.47 kbp HindIII-BamHI region was used to construct pIA301. This contains the 1.82 kbp HindIII-ClaI fragment cloned into HindIII + ClaI cut, and phosphatased, pAT153. In addition, the 4.4 kbp HindIII-AvaI fragment was cloned into HindIII + AvaI cut, and phosphatased, pAT153 to give pIA302. In both cases the desired insert fragments were isolated from

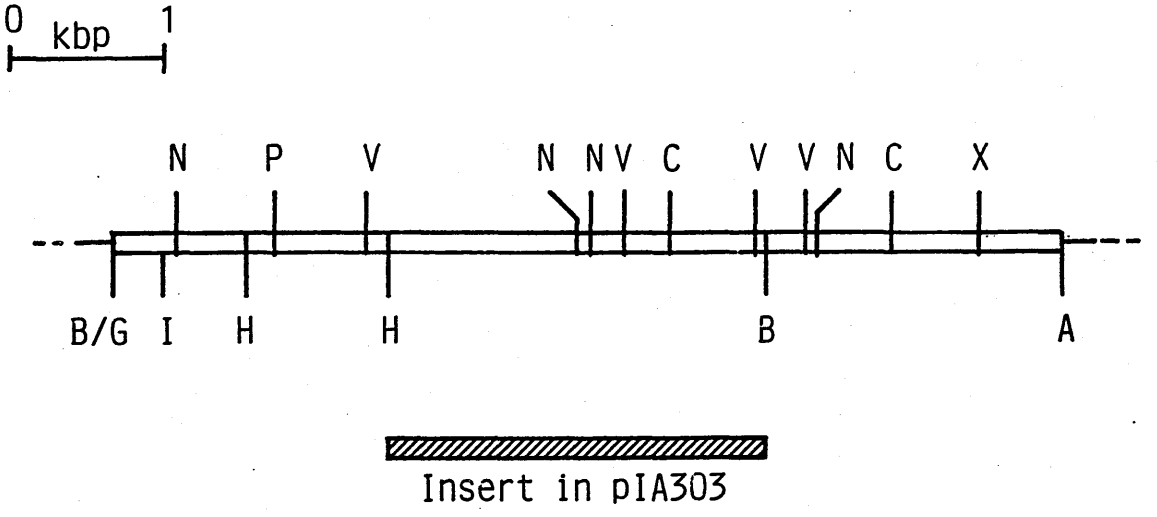


Figure 3.4 Restriction sites in the insert of pIA307

Key to restriction sites: B/G = BamHI-BglII hybrid site, H = HindIII, B = BamHI, C = ClaI, A = AvaI, P = PstI, N = HincII, V = PvuII, X = XbaI, I = HpaI.

Sites labelled underneath the bar are from the previously described R. Hayward compilation. As well as the sites in Figure 3.1 this compilation also gave sites for SalI, HpaI, KpnI, SmaI, and XhoI. However, of these only HpaI has a site in the region under consideration. Sites labelled above the bar were identified during this project. 2 AccI sites and an SstI site have not been mapped. There are no sites for SstII.

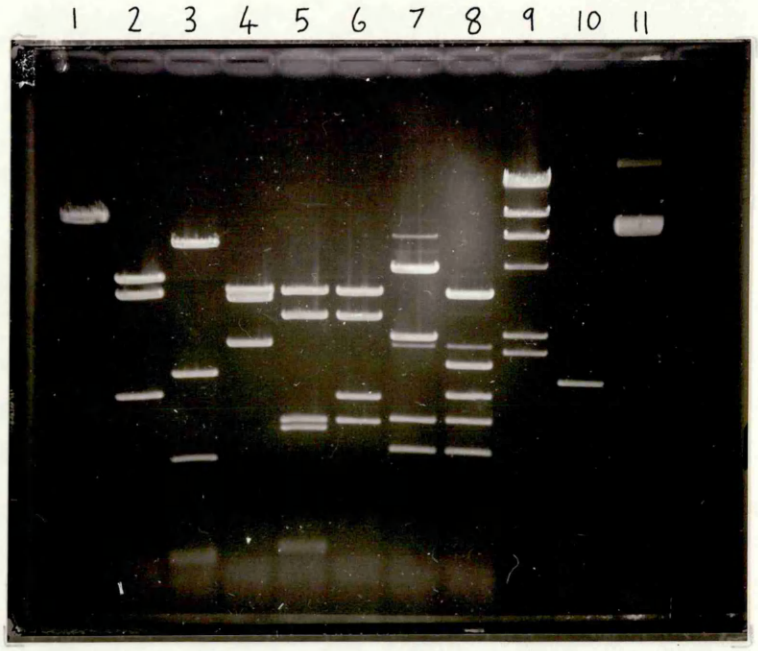


Figure 3.5 Agarose gel analysis of restriction digests of pIA307

Lane: 1. EcoRI cut pIA307
2. ClaI cut pIA307
3. PvuII cut pIA307
4. PstI + BamHI cut pIA307
5. HincII + BamHI cut pIA307
6. HincII cut pIA307
7. AvaI + HindIII cut pIA307 (incomplete digest)
8. HindIII + ClaI cut pIA307 (incomplete digest)
9. λcI857S7 cut with HindIII (size markers)
10. pAT153 cut with HinfI (size markers)
11. pIA307 (undigested)

In each case 0.8μg of pIA307 was digested and subsequently loaded on the gel. A 1% agarose gel was used. DNA fragments were stained with ethidium bromide and the gel photographed with a Polaroid camera on a U.V. transilluminator. The original photograph shows the fainter low m.w. bands better than the copy above.

λ/HindIII fragments (kbp): 23.6, 9.64, 6.64, 4.34, 2.26, 1.98, 0.56, 0.14 .

pAT153/HinfI fragments (kbp): 1.631, 0.517, 0.396, 0.298, 0.221 and 0.220, 0.154 and 0.145, 0.075 .

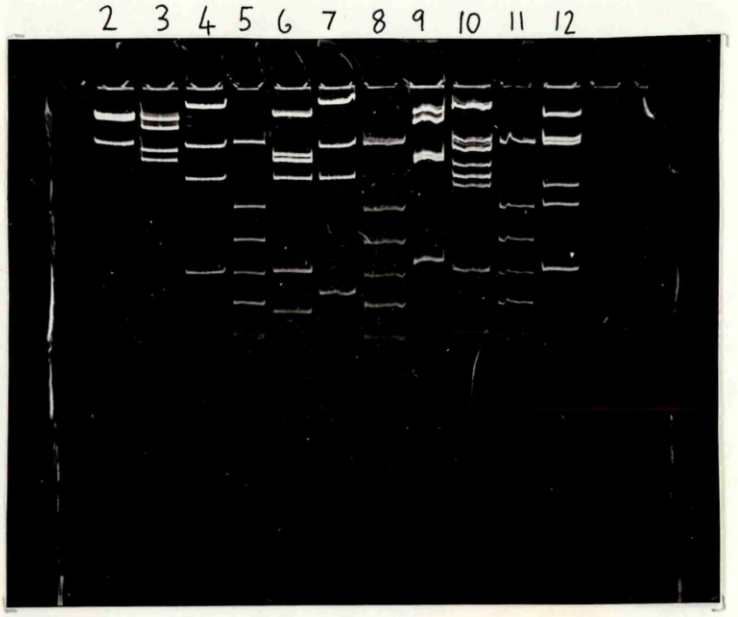


Figure 3.6 Polyacrylamide gel analysis of restriction digests of pIA307

Electrophoresis was through an 8% polyacrylamide gel which was subsequently stained with ethidium bromide.

Lane: 2. ClaI cut pIA307
3. HincII cut pIA307
4. PvuII cut pIA307
5. pAT153 cut with HinfI (size markers)
6. HincII + PvuII cut pIA307
7. BamHI + PvuII cut pIA307
8. as for 5.
9. BamHI + HincII cut pIA307
10. HindIII + PvuII cut pIA307
11. as for 5
12. ClaI + PvuII cut pIA307

pAT153/HinfI markers (kbp): 1.631, 0.517, 0.396, 0.298, 0.221 and 0.220, 0.154, 0.145, 0.075

This gel was run to look for small fragments that would have escaped detection on agarose gels (e.g. bottom of lane 3).

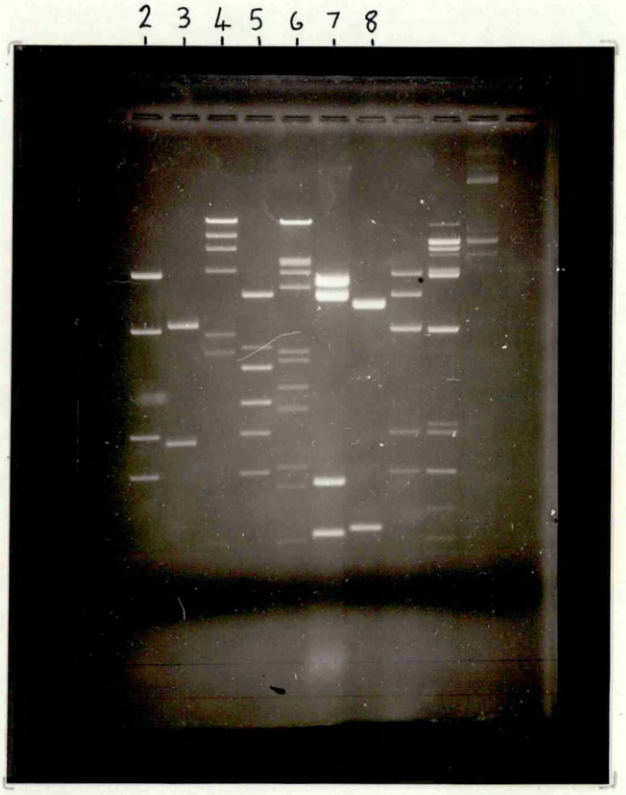


Figure 3.7 Agarose gel analysis of restriction digests of pIA307 - precise sizing of particular fragments.

1.5% agarose gel.

The relevant marker fragments are underlined. Sizes are in kbp.

Lane: 2. BamHI + HindIII cut pIA307 - for sizing of the 2.47 kbp fragment (second from top)
3. pAT153/BamHI + PstI - markers (2.532, 1.125)
4. λcI857S7/HindIII - markers (23.6, 9.64, 6.64, 4.34, 2.26, 1.98, 0.56, 0.14)
5. HindIII + ClaI cut pIA307 - for sizing of the 1.82 kbp fragment (second complete digest product from the top)
6. λcI857S7/HindIII + EcoRI - markers (21.7, 5.24, 5.05, 4.21, 3.41, 1.98, 1.90, 1.57, 1.32, 0.93, 0.84, 0.58, 0.14)
7. ClaI + BamHI cut pIA307 - for sizing of the 0.60 kbp fragment (fourth from top)
8. pAT153/HindIII + SalI - markers (3.036, 0.621)

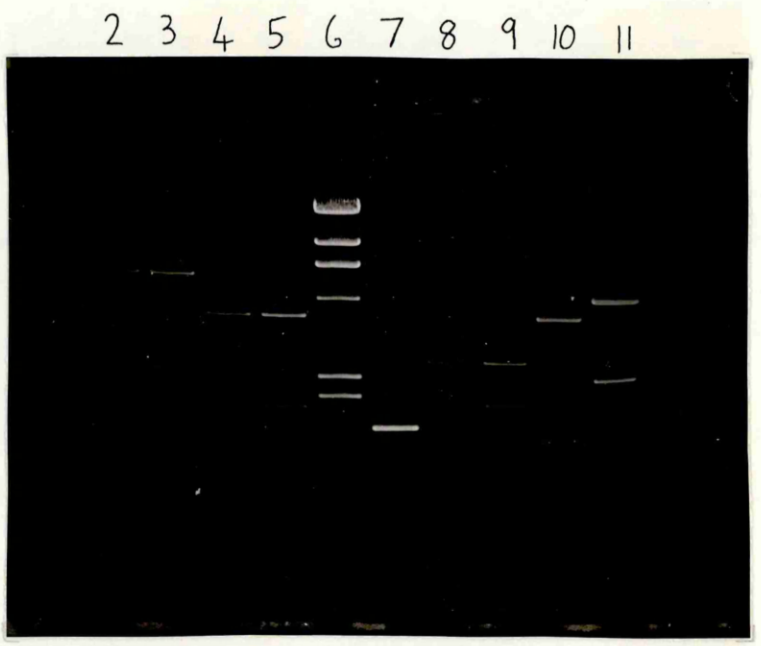


Figure 3.8 Agarose gel analysis of restriction digests of pIA301 and pIA302

Electrophoresis was through a 1% agarose gel which was subsequently stained with ethidium bromide.

Lane:
2. ClaI cut pIA301 (isolate 3/1/+1)
3. " " " " " 3/1/+3)
4. HindIII + ClaI cut pIA301 (isolate 3/1/+1)
5. " " " " " " 3/1/+3)
6. λcI857S7/HindIII markers
7. pAT153/HinfI markers
8. HincII cut pIA301 (isolate 3/1/+1)
9. " " " " " 3/1/+3)
11. AvaI + HindIII cut pIA302

λ/HindIII markers (kbp) : 23·6, 9·64, 6·64, 4·34, 2·26, 1·98, 0·56, 0·14

pAT153/HinfI markers (kbp) : 1·631, 0·517, 0·396, 0·298, 0·221 and 0·220, 0·154, 0·145, 0·075

digests of pIA307 by electrophoresis in a 1% low melting point agarose gel. This method of isolating bands from agarose gels proved much more convenient than that of Dretzen *et al*. The desired band is simply cut out of the low m.p. gel, the agarose melted and extracted with phenol, and the DNA purified (see Chapter Two for further details). Figure 3.8 shows pIA301 and pIA302 restriction digests analysed on a 1% agarose gel. Both these subclones confer the Aro^+ phenotype on AB2834. Hence, it appeared that *aroE* was contained within the 1.82 kbp HindIII-ClaI region.

3.3.5 E3 specific activities in strains carrying particular subclones

In an experiment like that described in Section 3.2.3 the E3 specific activities of crude cell extracts were determined for the following *E.coli* strains:- K12, AB2834, pIA307//AB2834 and pIA301//AB2834. Three independent isolates of pIA301 were tested since the cloned insert DNA had been exposed to the combination of ethidium bromide and long-wave U.V. during the isolation procedure and this slightly increases the remote chances of picking up a mutant DNA fragment. E2 specific activities were also determined, again as an internal control. The results are shown in Table 3.6. The cultures grew at similar rates (data not shown), after being inoculated to identical initial turbidities, and their final A_{650} values covered a narrow range (1.17 - 1.31). The E2 specific activities were all very similar and were comparable with those in

Table 3.6 E3 specific activities in crude cell extracts of AB2834 carrying pIA301 and pIA307

	E.coli strain	Crude extract E3 specific activity (units/mg)	Crude extract E2 specific activity (units/mg)	Final A_{650} of culture used
(a)	K12	0.092	0.026	1.24
(b)	AB2834	0	0.026	1.26
(c)	pIA307//AB2834	1.7	0.029	1.17
(d)	pIA301(1)//AB2834	0.078	0.030	1.30
(e)	pIA301(2)//AB2834	0.089	0.031	1.31
(f)	pIA301(3)//AB2834	0.076	0.030	1.24

NOTES: 1. All crude extracts were made from cells grown in L broth, supplemented with ampicillin in the case of strains carrying subclones, as in Section 3.2.3.

2. $$\frac{\text{E3 s.a.(pIA307//AB2834)}}{\text{E3 s.a. K12}} = \frac{1.7}{0.092} = 18.5$$

3. Three independent isolates of pIA301 were tested.

Section 3.2.3. The specific activity of E3 in pIA307// AB2834 was the same as that observed previously for pIA306//AB2834. However, although all three isolates of pIA301 confer the Aro$^+$ phenotype on AB2834 the E3 specific activity in strains containing pIA301 was, in all three cases, slightly less than the K12 wild-type value and was thus about 21 times lower than that obtained with the larger subclone. It seemed possible that the small insert of pIA301 did not span all the sequences required for the proper expression of aroE. Experiments analogous to the one described above were done using the other subclones. The aim was to define the minimum region of DNA required for maximal aroE expression.

Results for three independent isolates of pIA303, in HB101, are shown in Table 3.7. The cultures were grown from different initial turbidities, hence the much greater spread of the final A_{650} values: (a), (c), and (d) are roughly comparable whereas (b) is about two times higher. Note that (b) also shows higher E2 and E3 specific activities. pIA303 clearly gives overexpression of E3 but about three times less than that originally observed with pIA306.

Table 3.8 shows the results for another experiment with HB101 as the host for various subclones. Row (g) again shows that a higher final A_{650} correlates with higher E2 and E3 specific activities.

Table 3.9 contains data from an experiment in which AB2834 was the host for a variety of subclones. Row (c)

Table 3.7 E3 specific activities in crude cell extracts of HB101 carrying independent isolates of pIA303

	E.coli strain	Crude extract E3 specific activity (units/mg)	Crude extract E2 specific activity (units/mg)	Final A_{650} of culture used
(a)	HB101	0.078	0.011	0.62
(b)	pIA303(8)//HB101	0.50	0.019	1.3
(c)	pIA303(20)//HB101	0.31	0.010	0.58
(d)	pIA303(28)//HB101	0.40	0.013	0.79

NOTES: 1. All crude extracts were made from cells grown in L broth, supplemented with ampicillin in the case of strains carrying subclones.

2. $$\frac{\text{E3 s.a. pIA303(20)//HB101} - \text{E3 s.a. HB101}}{\text{E3 s.a. HB101}} = 3.0$$

3. Three independent isolates of pIA303 were tested: (8), (20), and (28).

Table 3.8 E3 specific activities in crude cell extracts of HB101 carrying various aroE subclones

	E.coli strain	Crude extract E3 specific activity (units/mg)	Crude extract E2 specific activity (units/mg)	Final A_{650} of culture used
(a)	HB101	0.12	0.018	0.82
(b)	pIA306//HB101	1.3	0.018	0.78
(c)	pIA307//HB101	1.3	0.019	0.85
(d)	pIA304//HB101	0.62	0.017	0.72
(e)	pIA305//HB101	1.0	0.020	0.76
(f)	pIA303(8)//HB101	0.40	0.017	0.90
(g)	pIA303(8)//HB101	0.61	0.027	2.6

NOTES: 1. Values in the upper section (rows (a) and (b)) are from Section 3.2.3.

2. All crude extracts were made from cells grown in L broth, supplemented with ampicillin in the case of strains carrying subclones.

3. Samples of the final cultures were taken for the small scale preparation of plasmid. This was done to check that no gross deletions had occurred. None were detected.

Table 3.9 E3 specific activities in crude cell extracts of AB2834 carrying various aroE subclones

	E.coli strain	Crude extract E3 specific activity (units/mg)	Crude extract E2 specific activity (units/mg)	Final A_{650} of culture used
(a)	AB2834	0	0.032	1.2
(b)	pIA306//AB2834	2.1	0.036	1.2
(c)	pIA306//AB2834	5.4	0.052	2.9
(d)	pIA307//AB2834	1.6	0.033	1.2
(e)	pIA304//AB2834	0.48	0.037	1.3
(f)	pIA305//AB2834	1.7	0.036	1.2
(g)	pIA303(8)//AB2834	0.41	0.038	1.2
(h)	pIA307//AB2834	2.5	0.036	1.0
(i)	pIA302(2)//AB2834	1.9	0.023	1.0
(j)	pIA302(6)//AB2834	1.8	0.019	0.92

NOTES: 1. The values in the lower section are from a separate experiment (rows (h), (i) and (j)).

2. All crude extracts were made from cells grown in L broth supplemented with ampicillin in the case of strains carrying subclones.

3. Two independent isolates of pIA302 were tested: (2) and (6).

Table 3.10. Summary of experiments on the relationship between the E3 specific activity in crude extracts of a strain and the subclone it carries

Left — Insert (not to scale) — Right	Subclone	A	B	
BglII HindIII ClaI BamHI AvaI HindIII EcoRI				
⊢——————————⊣	pIA306	1	1	)
⊢——————⊣	pIA307	0.8	1	)
⊢——————⊣	pIA305	0.8	0.7	) Class I
⊢——————⊣	pIA302	0.6*	-	)
⊢——————⊣	pIA304	0.2	0.4	)
⊢——————⊣	pIA303	0.2	0.2, 0.4	) Class II
⊢——⊣	pIA301	0.04	-	) Class III

NOTES: 1. Column A shows the relative values (to 1 significant figure) of E3 specific activities in crude cell extracts made from approximately equivalent cultures of AB2834 carrying the subclone shown. Column B does likewise for HB101 after subtraction of the contribution of the host. The highest value has arbitrarily been given the value of 1.

2. (*) The relative value for pIA302//AB2834 may be too low since the E2 s.a. is rather low in both cases (see Table 3.9).

3. The structures of these subclones are shown in Figure 5.8.

once more demonstrates that E2 and E3 specific activities are elevated when the cultures are grown to higher cell densities.

The main results of this group of experiments are summarised in Table 3.10.

3.3.6 Preliminary interpretation of the specific activity results

Before considering the differences between the various subclones two general conclusions should be noted. Firstly, if two otherwise identical cultures are grown to different final A_{650} values then the culture with the higher A_{650} will show higher E3 and E2 specific activities. Hence, one should only compare the specific activities of different cultures if they have been grown to roughly similar A_{650} values. Secondly, overexpression of E3 is less marked in HB101 than in AB2834. The reason for this is unknown.

Plasmids pIA306, 307, 305, and probably 302 can be classed as giving "high" E3 specific activities - class I. Plasmids pIA304 and 303 - class II - give E3 levels which are rather lower than those of class I. pIA301 - class III - shows much lower levels than class I. These results must be interpreted cautiously because many factors may be involved.

The size of the insert and/or its location in the vector might influence gene expression. For example, the size of the insert might affect plasmid replication and thus the copy number. However, the size of the insert in pIA302 (4.4 kbp, class I) is almost the same as that

of the insert in pIA304 (4.3 kbp, class II). This makes insert size less plausible as a way of explaining the differences between the subclones. None of the subclones have lost vector fragments which are known to affect plasmid replication or partition.

Particular insert sequences, or hybrid vector-insert sequences created at the ligation sites, could influence plasmid functions. It is known that strong constitutive promoters often cannot be cloned in the absence of a downstream transcriptional termination signal (Gentz *et al.*, 1981) presumably because high levels of transcription interfere with plasmid replication. If the insert of, say, pIA301 spans *aroE* and also a highly transcribed gene but does not include the latter's terminator then this might cause some plasmid instability and thus effectively reduce the copy number.

It is unlikely that interactions between the transcription of the plasmid Tet^R gene and transcription of *aroE* could alone explain the data. In both pIA307 (class I) and pIA304 (class II), for example, the *aroE* gene must have the same position and orientation relative to Tet^R transcription.

A simple hypothesis is that class II and III insert fragments do not span all the sequences required for full expression of the *aroE* gene. Sequences required for maximal *aroE* expression are unlikely to be to the left of the HindIII site which forms one end of the inserts in pIA305,

302, 303, and 301 since both pIA305 and pIA302 are in class I. However, one could imagine that part of the aroE promoter straddles the BamHI site at one end of the pIA304 and pIA303 (class II) inserts. Cleavage here might leave only a partially functional sequence. Cleavage at the ClaI site when cloning the pIA301 (class III) insert would give total loss of the sequence. However, this hypothesis is not without difficulties. For instance, one must postulate a messenger RNA which has an unusually long leader sequence if cleavage at the ClaI site is not to remove the ribosome binding site (and N-terminal portions of E3). The ClaI site is 0.6 kbp from the BamHI site (see Figure 3.4). Most prokaryotic mRNAs have 5' untranslated leader sequences of less than 200 bases (reviewed by Kozak, 1983). At this stage in the work it seemed sensible to postpone further consideration of the specific activity results until after the relevant region of DNA had been sequenced. Further discussion will be found in Chapter Five.

Before starting to sequence the aroE gene and the surrounding DNA it was desirable to check that the E3 produced in the various subclone carrying strains was similar to that produced in E.coli K12. These comparisons are described in the next section. There was the possibility - not considered above - that the E3 polypeptide had a partially redundant C-terminal portion which was slightly truncated in pIA303 and 304 and drastically truncated in pIA301.

3.4 Analysis by PAGE of E3 from different subclones and from E.coli K12

3.4.1 Background

The experiments described in this section exploit a specific gel stain for shikimate dehydrogenase activity - the nitro-blue tetrazolium (NBT) dye-linked method, as adapted by Lumsden and Coggins (1977). The technique is very sensitive. When some crude extract of wild-type E.coli K12 is run on a polyacrylamide gel under nondenaturing conditions it is easy to detect the band of E3 activity. This seemed a suitable way of comparing E3 in wild-type E.coli K12 with that produced in the various subclone carrying strains.

3.4.2 Analysis by nondenaturing PAGE of E3 from different strains

Crude cell extracts of E.coli K12 and of a selection of subclone containing strains were made in the same way as for the determination of specific activities. Samples of these extracts were run on nondenaturing polyacrylamide slab gels. To allow precise comparisons between lanes the salt concentration, glycerol concentration, and final volume were kept constant between different samples on a particular gel. For each strain duplicate samples were loaded on opposite halves of the gel. After electrophoresis one half was stained normally for E3 activity. The other half was stained in a mixture lacking the specific substrate

shikimic acid. This served as a control against nonspecific reduction of the NBT. The results obtained using crude cell extracts of E.coli strains AB2834, K12, pIA301//AB2834 (class III), pIA303//AB2834 (class II), and pIA307//AB2834 (class I) are shown in Figure 3.9.

The minus shikimic acid control half shows no bands but does have three very faint diffuse regions which stretch horizontally across the upper part of the complete gel. The nature of these is unknown. In the plus shikimic acid half AB2834 gives only a very faint activity band and even this could simply be "splashover" from the adjacent K12 lane where there is a very prominent band.

The three subclone containing strains each give only one prominent band and the intensities of these increase in the expected order. These three bands, together with the K12 band, all have the same mobility. This supports the view that the (putatively) cloned aroE gene produces an E3 at least superficially identical to that found in E.coli K12. The mobility of a protein in this gel system is a function of both its net charge and of the Stokes radius. Thus it is conceivable, but unlikely, that two bands of identical mobilities represent different proteins whose Stokes radii and charges combine to give identical mobilities. (However, a single experiment at a different gel concentration (8%) also showed all bands to have the same mobility). The faint bands above the main activity bands are probably due to aggregates that have become linked by intermolecular disulphide bonds which survive,

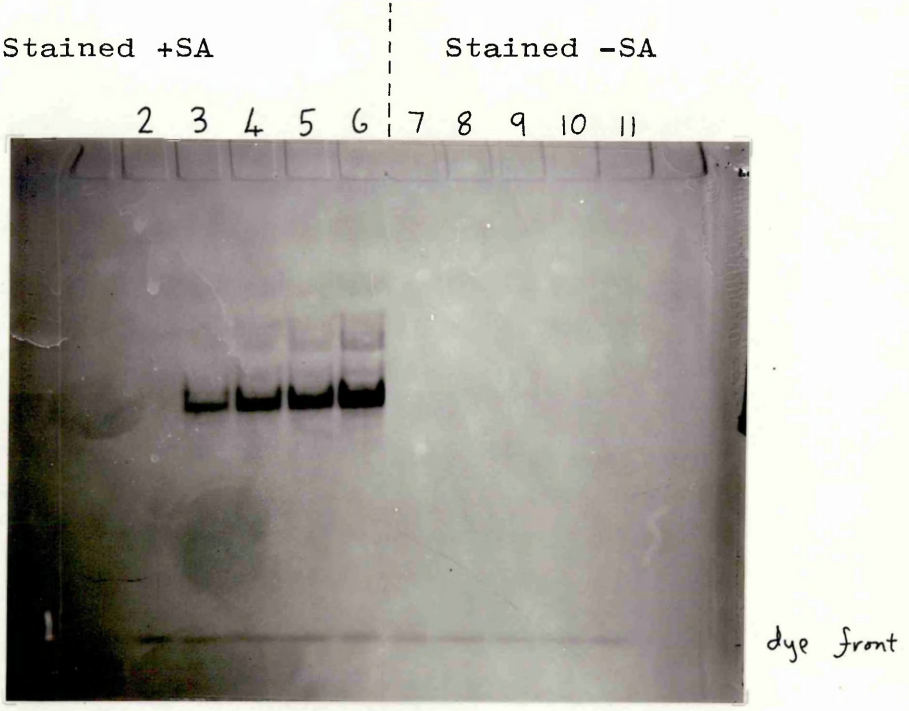
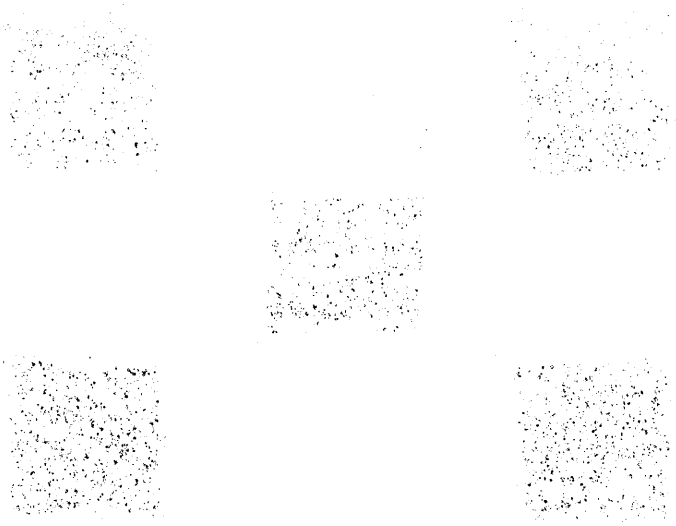


Figure 3.9 Comparison of E3 in K12 with that in subclone carrying strains.

Nondenaturing PAGE was carried out in a 10% slab gel. The two halves were stained for E3 activity as shown above. Identical amounts of protein were loaded in each lane (105 μg) and the samples were treated with 50mM DTT for 1 hour at 0°C prior to loading. The samples were of crude cell extracts of particular strains:

Lanes 2,11. *E.coli* AB2834
3,10. *E.coli* K12
4,9 *E.coli* pIA301//AB2834
5,8 *E.coli* pIA303//AB2834
6,7 *E.coli* pIA307//AB2834

Note that the crude extract of pIA301//AB2834 was made from a culture with a higher final A_{650} than that of the K12 culture and, as expected, had a slightly higher than usual relative specific activity.



as some do, the treatment with DTT prior to loading. The faint bands underneath might be the result of proteolysis or of intramolecular disulphide linkages.

3.4.3 Recovery of enzyme activity after treatment with SDS-introduction

More evidence that the E3 from subclone carrying strains is outwardly indistinguishable from that in E.coli K12 came from experiments in which samples of crude cell extracts were subjected to SDS PAGE by the method of Laemmli (1970). After treatment designed to renature enzymes these gels were stained for E3 activity. It proved possible to detect specific bands. This work was prompted by diverse reports in the literature which describe the successful renaturation of proteins from SDS-protein complexes. Weber and Kuter (1971) incubated a variety of enzymes with 2-mercaptoethanol and 1% SDS for 1 h at room temperature. This is rather different from the pre-loading treatment used for SDS PAGE samples. Nonetheless, they showed that the SDS could be removed in the presence of 6M urea (using an ion-exchange resin) and that various enzymes could subsequently be renatured from the urea solution. Rosenthal and Lacks (1977) demonstrated that certain extra- and intracellular nucleases could be re-natured in gels, after SDS PAGE, by prolonged washing in 40mM Tris-HCl pH 7.6, 2mM $MgCl_2$, 0.02% sodium azide. No urea was used at any stage. They assert that the pore

size of a 10% separating gel should allow SDS micelles to diffuse out and they emphasise that the purest available SDS should be used since some contaminants appear to bind very tightly to the proteins. Dottin et al. (1979) showed that alcohol dehydrogenase (Drosophila) and lactate dehydrogenase (Fundulus), amongst other enzymes, could be renatured in gels (and activity stained) after SDS PAGE followed by isoelectric focussing in the presence of 7M urea and 2% NP40. Early work on protein blotting provides further examples of proteins, such as the lac repressor, which can be renatured (using the criterion of restored function) after SDS PAGE (Bowen et al., 1980): SDS is removed from the gel by washing firstly in 4M urea, 1% Triton X-100, 50mM NaCl, 2mM Na_2 EDTA, 0.1mM DTT, 10mM Tris-HCl pH 7.0, and secondly in the same buffer minus Triton X-100. (The urea itself is removed during the later blotting stage). It was also known that a tryptic fragment (subunit m.w. = 68kDa) of N.crassa arom could regain E3 activity after denaturing PAGE in the presence of 8M urea (Smith and Coggins, 1983).

Thus, it seemed that it might be possible to renature E.coli (and N.crassa) E3 in gels after SDS PAGE and to stain the resulting activity bands specifically. As well as permitting further comparisons of the E3 in various strains it was also hoped that the renaturation approach would yield a preliminary estimate of the E.coli E3 subunit m.w. which was unknown at this stage in the work since the enzyme had not yet been purified to homogeneity.

3.4.4 Recovery of E3 activity after SDS PAGE - results and discussion

A protocol for renaturing E3 after SDS PAGE was designed on the basis of the work cited above. It was initially tested using purified *N.crassa arom* which was the generous gift of M.R. Boocock (*arom* preparation "A21"; Boocock, 1983). The first protocol used was as follows:

Buffer A = 50mM sodium phosphate pH 7·0, 1mM DTT.

1. Samples were boiled for 2 minutes exactly in $\geqslant$ 1% SDS, $\geqslant$ 2% 2-mercaptoethanol, prior to loading. More prolonged boiling should be avoided due to the potential risk of *trans-cis* rotation of peptide bonds. After electrophoresis:
2. Soak the gel with gentle agitation at room temperature in Buffer A + 8M urea (freshly recrystallised), 1mM glycinamide, 0.1% Triton X-100: use 150 ml over about 4 h.
3. Continue for another 3 h using 100 ml of the same buffer but without Triton X-100.
4. Soak in a total of 500 ml of Buffer A (2 changes) for 12 h .
5. Remove DTT prior to staining by soaking for 1 h in a total of 500 ml of 100mM Tris-HCl pH 9.0 (two changes).
6. Stain for activity (at least 1 h).

This protocol was successful in that one could see activity bands for A21 *arom* but these were rather faint. It also permitted the detection of a faint activity band for a crude extract of pIA307//AB2834. However, M.R. Boocock

greatly improved the protocol by introducing the following modifications (Boocock, 1983):

(a) Include 0.01% 2-mercaptoethanol, 0.1mM EDTA in the well buffer used for electrophoresis.

(b) Do not use urea; proceed as follows:

Soak the gel with gentle agitation at room temperature in Buffer A + 0.1% Triton X-100, 0.1mM EDTA - use a total of 1 l in several changes over 7-12 h .

(c) Continue for about 12 h using a further 1 l of the same buffer but without Triton X-100.

(d) As for 5. above.

(e) Stain for activity.

This improved version was used in all the experiments described below. Figure 3.10 shows the E3 activity bands detected after SDS PAGE (and renaturation) of crude cell extracts of E.coli K12, AB2834, pIA301//AB2834, and pIA307//AB2834. A similar experiment is shown in Figure 3.11. The fainter details of the photographs were much more easily seen when the original gels were viewed on a light box. The following reproducible features should be noted:

1. All activity bands are dependent on shikimic acid (or at least on something in the shikimic acid bottle).
2. AB2834 shows no bands.
3. K12, pIA301//AB2834, and pIA307//AB2834 each show (i) an activity band of m.w. about 29kDa (all three bands having identical mobilities) and (ii) an

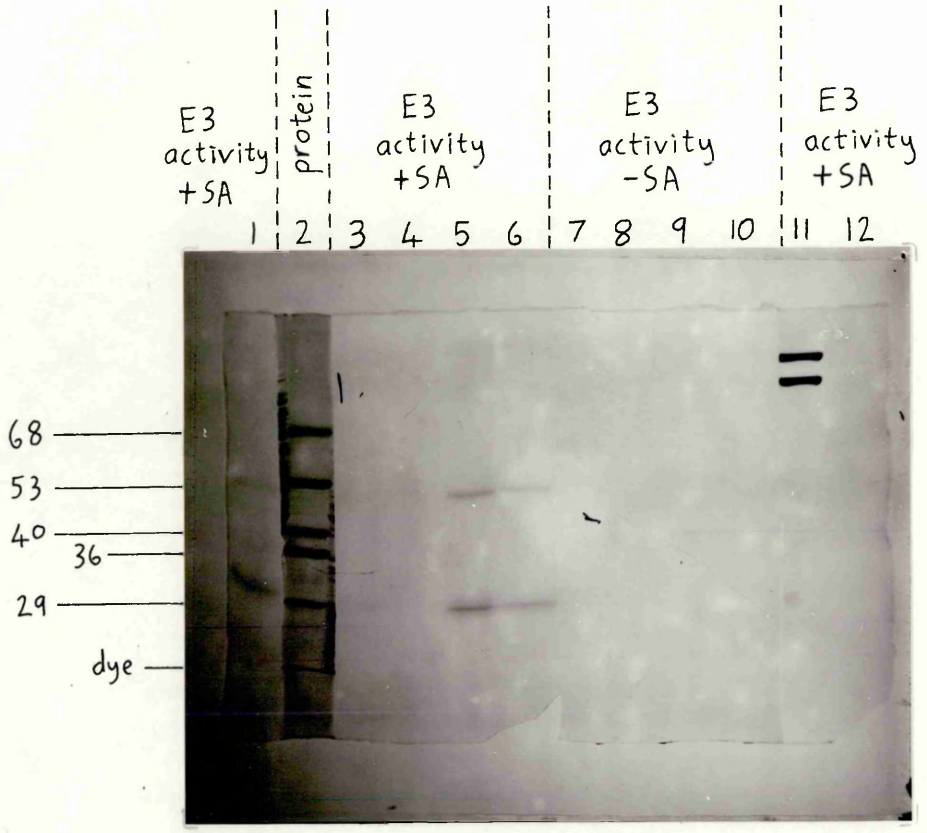


Figure 3.10 Detection of E3 activity bands in crude cell extracts after SDS PAGE - (1)

Crude extracts of various E.coli strains were run, plus a sample of pure A21 arom (some of which is proteolytically nicked, hence the doublet) as an internal control for the efficiency of renaturation. Gel (10%) was cut into sections prior to staining as shown above each section.

Lane: 1. pIA301//AB2834 (96 µg total protein)
2. molecular weight markers (see Chapter Two)
3. K12 (68 µg total protein)
4. AB2834 (84 µg total protein)
5. pIA307//AB2834 (60 µg total protein)
6. " "
7. K12
8. AB2834
9. pIA307//AB2834
10. " "
11. A21 arom
12. AB2834

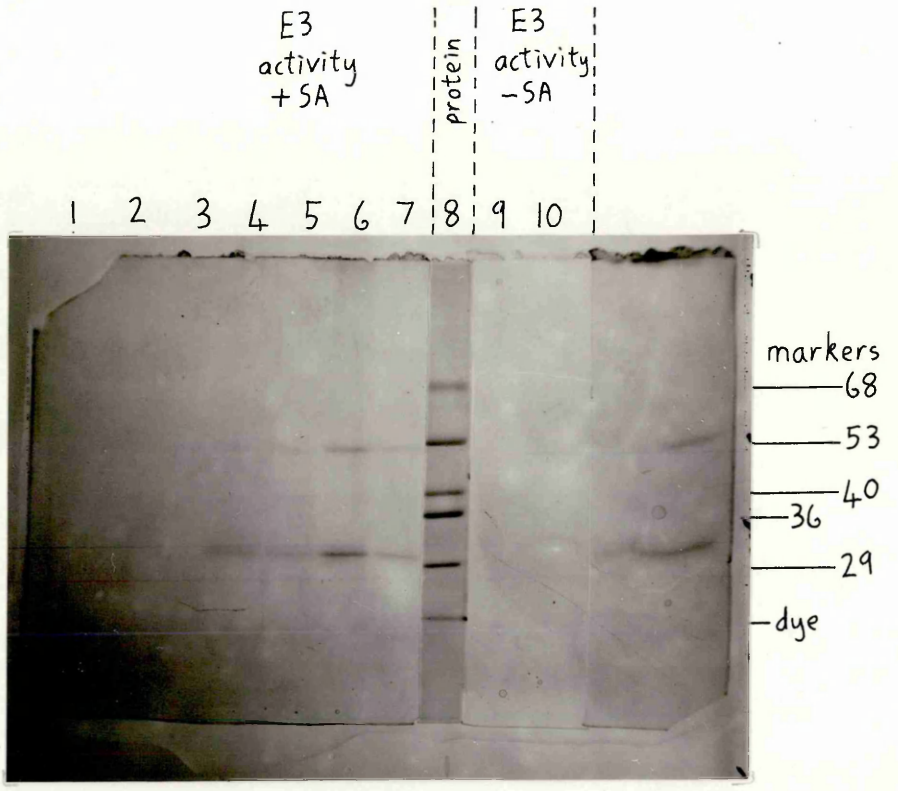


Figure 3.11 Detection of E3 activity bands in crude cell extracts after SDS PAGE - (2)

10% gel. Lane:
1. AB2834
2. AB2834
3. blank to prevent splashover into 1. & 2.
4. pIA301//AB2834
5. K12 + pIA301//AB2834
6. pIA307//AB2834
7. pIA302//AB2834
8. molecular weight markers (see Chapter Two)
9. pIA307/AB2834
10. pIA302//AB2834

activity band of m.w. about 49 kDa (all three bands having identical mobilities).

4. The upper band is always fainter than the lower band.
5. Both the upper and lower bands are much more intense in the pIA307//AB2834 lanes than in the K12 or pIA301//AB2834 lanes.

Leaving aside momentarily the surprising presence of two bands, these experiments again fail to reveal any differences, other than of quantity, between the E3 from K12 and that from a large or the smallest subclone. This implies that pIA301 must carry most (if not all) of the aroE coding sequence. Hence, the initial DNA sequencing work (see Section 3.5) concentrated on the 1.82 kbp HindIII-ClaI fragment which forms the insert of pIA301.

Returning to the finding of two activity bands in place of the anticipated single band, the following conclusions can be drawn. The points listed above suggest that both bands are, in a loose sense, "encoded" by even the smallest subclone. The question then arises of whether the smaller m.w. band is, in some way, a subset of the larger. The data presented do not permit a definite answer. However, for a number of reasons it seems unlikely that there are two separate genes giving independent polypeptides with E3 activity. Firstly, the maximum coding capacity of the pIA301 insert is 607 amino acids and thus is unlikely to code for more than 120 Da x 607 = 73kDa worth of polypeptide(s). "29" kDa + "49" kDa = "78" kDa which would

be a tight squeeze. Secondly, AB2834 lanes lack both bands. There are two obvious ways in which the smaller band could be a subset of the larger band. The smaller could be a proteolytic fragment of the larger. Alternatively, oxidation could lead to disulphide crosslinking of the smaller species to another protein: there are (rare) precedents for such linkages surviving boiling in SDS/ 2-mercaptoethanol, for example, between the β and β' subunits of E.coli RNA polymerase (J.R. Coggins, unpublished results).

It should be emphasised that the relative intensity of the two bands is not a safe guide to the relative abundance or activity of the two species since their respective efficiencies of renaturation might be quite different. Hence, the presence of only a single major band on nondenaturing gels is not necessarily at variance with the SDS PAGE results.

Since the problem of the two bands could not be simply resolved no conclusions could then be drawn about the E.coli E3 subunit m.w.(s). Further discussion will be found after the DNA sequencing work has been described (Chapter Five).

3.5 DNA sequencing of aroE

3.5.1 Introduction

3.5.1A Choice of sequencing method

All rapid DNA sequencing methods exploit polyacrylamide gel systems which can resolve single-stranded DNA fragments, of up to several hundred nucleotides, which differ in length by a single nucleotide. Also, the available methods (with the exception of the recently developed "genomic sequencing" technique of Church and Gilbert, 1984) all feature nested sets of radiolabelled DNA fragments having one end in common. The fragments are made in four groups and the members of each group all end randomly at a particular type of base.

Two major DNA sequencing techniques are currently in use. The method of Maxam and Gilbert (1977) relies on base-specific chemical cleavage of DNA fragments radiolabelled at one common end. The method of Sanger et al. (1977) involves copying the DNA to be sequenced, using the Klenow fragment of DNA polymerase I starting from a uniquely hybridising primer (see Figure 3.13) which provides the common end; base-specific endpoints are obtained by using specific chain-terminating dideoxy nucleotide analogues.

The dideoxy method was chosen for this project mainly because it can be used in a shotgun fashion when combined with the phage M13 cloning system developed by J. Messing and his coworkers (Messing, 1983). This obviates the

need for a highly detailed restriction map of the DNA to be sequenced, a normal requirement when using the Maxam and Gilbert method. Overall the M13/dideoxy approach is probably the most efficient one available, especially for sequencing many kilobases of DNA as was the eventual aim in this laboratory.

3.5.1B M13/dideoxy sequencing

A single-stranded (SS) DNA template is an essential prerequisite for sequencing by the dideoxy method. This requirement is easily met by exploiting the life cycle of the single-stranded filamentous phage M13. The key stages in M13/dideoxy sequencing are set out below and are summarised in Figure 3.12. Further details are given in Chapter Two - the procedures used were essentially those given in Amersham's "M13 Cloning and Sequencing Handbook" (Amersham, 1983).

Stage 1 - M13 cloning:

(a) Ligate the double-stranded (DS) DNA fragment(s) to be sequenced with cut DS M13 replicative form (RF).

(b) Transform competent E.coli host cells with ligation mix. In this project E.coli JM101 was used initially but later E.coli TG1 was substituted since the latter, unlike JM101, is EcoK restriction minus. (The recombinant plasmid which provided the source material for sequencing was isolated from the EcoK modification minus strain E.coli HB101).

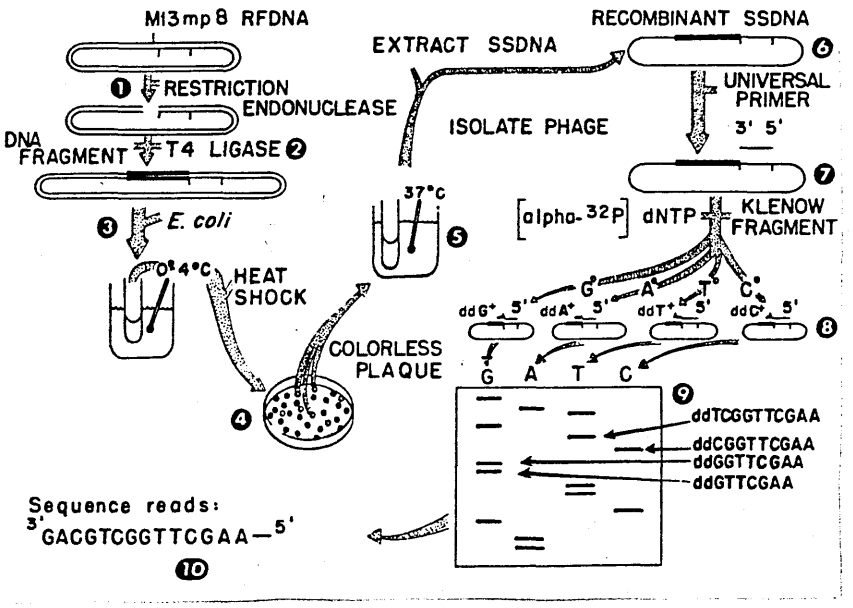


Figure 3.12 M13/dideoxy sequencing
(After Messing, 1983)

There are a wide variety of M13mp vectors and E.coli host cells (see text). α-^{35}S dNTPαS can be used as the label (see text).

(c) Plate out the transformed cells in the presence of X-gal and IPTG and grow up plaques. Blue plaques are almost always nonrecombinant although very occasionally a particular insert will still allow complementation of the lacZ M15 deletion (Close et al., 1983). Colourless plaques are usually, but not always, recombinant.

Stage 2 - Preparation of SS templates:

(a) Pick and grow up individual colourless plaques, in 1.5 ml cultures.

(b) Remove cells from the culture and precipitate the phage from the supernatant with PEG.

(c) Extract the SS template DNA from the phage and purify it.

(d) Run samples of the templates on a 1% agarose gel to check that DNA is present in each case and to see if any of the templates are contaminated with material from a different plaque.

Stage 3 - Sequencing reactions:

(a) Anneal each template with the "universal" primer.

(b) Carry out the four different sequencing reactions with each annealed template/primer as summarised below:

"A"-template/primer + Klenow + ddATP/dNTP's + label (see below).

"C"- " + " + ddCTP/ " + " .

"G"- " + " + ddGTP/ " + " .

"T"- " + " + ddTTP/ " + " .

(c) Chase the reaction with a mixture of all four dNTP's.

Stage 4 - Electrophoresis and autoradiography:

(a) Add formamide (+ dye mix) to each reaction and heat in order to denature the DNA.

(b) Electrophorese samples on a thin polyacrylamide gel at high temperature (use two separate loadings so as to obtain a long run and a short run).

(c) Fix the gel and dry it down.

(d) Autoradiograph the gel.

Stage 5 - Data processing:

This is discussed further on and in Chapters Two and Five.

The radioactive label used during this project was deoxyadenosine 5'-(α- [^{35}S] thio)triphosphate (Biggin et al., 1983). The main advantage of the ^{35}S analogue over [α-^{32}P] dATP is that the bands on the autoradiographs of the sequencing gels are much sharper due to the lower energy of the electrons emitted during ^{35}S decay. It is therefore possible to read the sequence further up the gel. Also, even above where the sequence can be read reliably it is usually still possible to count the bands (especially T's and to a lesser extent A's) and to estimate the spacing between them. This "unreadable" pattern of bands can often be used to match up two different clones even if they are complementary strands. Other advantages of using ^{35}S for sequencing include its logistically more convenient half-life and the lower radiation dose received by the user. The only disadvantage is a minor one: prior to autoradiography one must dry down the gel as the radiation is less penetrating than from ^{32}P. Although not essential drying the gels also improves the resolution obtained with ^{32}P.

3.5.1C Potential pitfalls in DNA sequencing

It is important to sequence across every restriction site that is used to make fragments for sequencing because closely adjacent restriction sites of the same kind may have been mistaken for single sites during restriction mapping.

A particular template insert may represent an artefactual ligation of two (or more) fragments of DNA which are not normally contiguous. When fragments for sequencing are generated by digestion with a restriction enzyme that cuts frequently, as in this project, it is straightforward to examine the sequence for internal sites of the enzyme used initially. Internal sites should always be assumed to be artefactual and not the result of partial digestion.

It is dangerous to sequence only one strand of a region of DNA due to the risk of missing a base. Such a frameshift error would have a catastrophic effect on the amino acid sequence deduced from a gene sequence. Normally the spacing between bands declines gradually up the gel and this regularity is useful when reading the sequence, for example, in prompting one to look for a faint band where the spacing implies that one should exist. However, certain stretches of sequence, especially G/C rich ones, can give rise to the "compression effect" (also known as "pile-up") where the spacing between successive bands is reduced abruptly, sometimes to zero (Fiddes, 1976; Sanger et al., 1977). This effect, which is a general problem with all rapid DNA sequencing techniques, is due to the DNA forming

base-paired loops during electrophoresis despite the denaturing conditions (urea and high temperature) that are normally used. The spacing of bands above a compression is often expanded. Usually compressions are obvious but there are insidious cases which can escape notice (see Figure 3.20). Fortunately the position of a compression is generally shifted by a few bases when the complementary strand is sequenced. Thus, the whole of both strands should be sequenced, as was done in this project.

3.5.1D Outline of the sequencing strategy used in this project

This section provides a brief overview of the DNA sequencing work. The initial aim was to obtain the protein coding sequence of the *aroE* gene. Evidence presented in Sections 3.3 and 3.4 suggested that the desired region of DNA was probably contained within the 1.82 kbp HindIII-ClaI fragment which forms the insert of pIA301. Hence, the objective became the sequence of this 1.82 kbp fragment. All material for sequencing came from a class I plasmid, pIA307, which had been prepared on a large scale.

At the start defined restriction fragments from the region of interest were isolated and then cloned specifically using suitably cut M13 mp 8/9 vectors. This work exploited available restriction sites from the map given in Figure 3.4 and generated islands of sequence whose locations were known.

In the second round of sequencing the 1.82 kbp fragment was digested with the frequently cutting restriction enzyme HpaII. The resulting small fragments were cloned into M13 in a shotgun fashion. Templates were prepared from many plaques and screened by "T-tracking" (see Section 3.5.2) to eliminate redundant clones. Unique clones, which include those with complementary strands, were then sequenced. Complementary strands were readily identified by visual inspection of the gel autoradiographs as were matches to the regions of sequence obtained from the first round.

Another set of small fragments, overlapping those generated by HpaII, was then required so as to allow the whole sequence to be pieced together. The third round of sequencing was therefore analogous to the second except that the 1.82 kbp fragment was digested with a different restriction enzyme - TaqI.

Due largely to the unfavourable distribution of TaqI sites a fourth stage was required to complete the sequence of the 1.82 kbp fragment. A more selective cloning strategy was used in this final stage to span a gap in the sequence.

3.5.2 First round of M13/dideoxy sequencing

The sequencing vectors M13mp8 and M13mp9 were used throughout this project (Messing and Vieira, 1982). Each contains a "polylinker" which facilitates the insertion of a wide variety of restriction fragments (see Figure 3.13 and Table 3.11). Moreover, the orientation of the

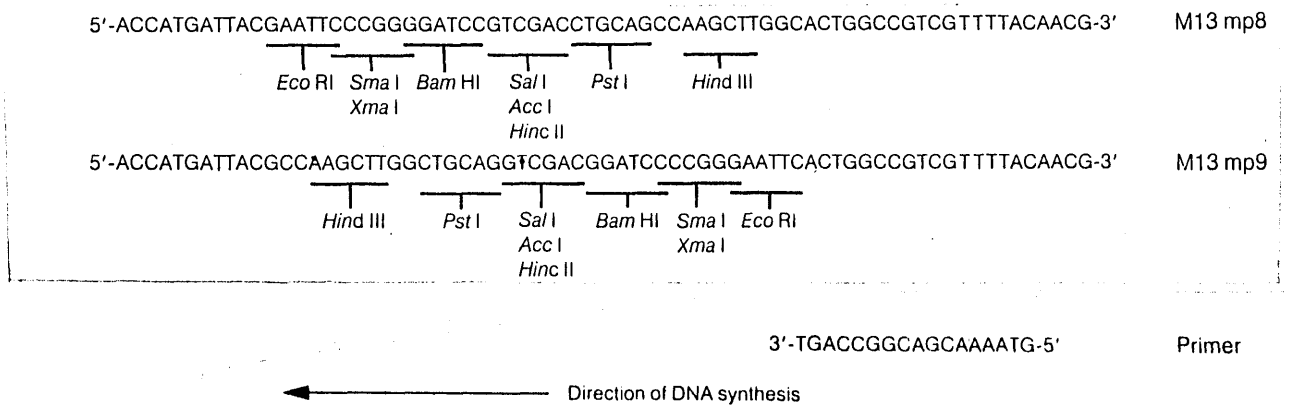


Figure 3.13 The cloning sites of M13 sequencing vectors mp8 and mp9.

The sequence of the 17 base primer used in this project is shown. Note that the order of the restriction sites is reversed in mp8 compared to mp9 thus allowing either strand of a double digest DNA fragment to be cloned at will (Messing and Vieira, 1982).

Table 3.11 The versatility of the mp8/9 cloning sites

Vector site	Some restriction enzymes giving ends compatible with the vector site
AccI GT'CGAC	ClaI AT'CGAT HpaII C'CGG TaqI T'CGA
BamHI G'GATCC	Sau3A 'GATC
SmaI CCC'GGG)) HincII GTC'GAC) (blunt ends)	HaeIII GG'CC AluI AG'CT HincII GTPy'PuAC PvuII CAG'CTG

NOTES:
1. All the restriction enzymes listed in the right half of the table were used during the DNA sequencing work.
2. The recognition sequences are written 5' to 3'. The site of cleavage is indicated thus: '.

polylinker in mp8, relative to the primer hybridisation site, is the opposite of that in mp9. It is therefore possible to select either strand of a double digest restriction fragment for cloning and sequencing since one can choose the orientation of the insert.

Defined restriction fragments, of known location, were cloned individually for the first round of sequencing of the 1.82 kbp HindIII-ClaI region. Particular fragments were purified from low m.p. agarose gels, after electrophoresis of the appropriate restriction digests of pIA307. The isolated fragments were either cloned forthwith, or, in some cases, cleaved further. For the latter the M13 vectors were cut so as to accept only the desired secondary digest product (in the correct orientation). For example, to clone the 0.6 kbp ClaI-BamHI fragment (which flanks the 1.82 kbp HindIII-ClaI fragment), in both orientations with respect to the primer, the previously isolated 2.47 kbp HindIII-BamHI fragment was digested with ClaI. The resulting mixture of fragments was ligated both with AccI + BamHI cleaved M13mp8 and, separately, with AccI + BamHI cut M13mp9. The templates derived from these two ligations carried complementary strands of the ClaI-BamHI fragment. Thus one could sequence into this fragment from opposite ends. There is no danger of ambiguity provided that the vector-insert junction is as expected.

When, as here, purified single digest fragments are cloned it is necessary to prepare templates from many plaques to have a good chance of obtaining at least one

Table 3.12 First round of M13 cloning - defined fragments

Template name	Insert fragment	Vector
3-4	0.6 kbp ClaI-BamHI	AccI + BamHI cut mp8
4-2	"	AccI + BamHI cut mp9
13-2	0.84 kbp PvuII	SmaI cut mp8
17-4	1.08 kbp HincII - BamHI	HincII + BamHI cut mp8
18-8	1.27 kbp HindIII-HincII	HincII + HindIII cut mp9
16-3	1.6 kbp HindIII-PvuII	"
15-5	1.7 kbp PvuII	SmaI cut mp8
21-3	0.090 kbp HincII	"

NOTES: 1. The insert fragments are shown in Figure 3.14.
2. The template name is given for the one which was sequenced from a particular class.

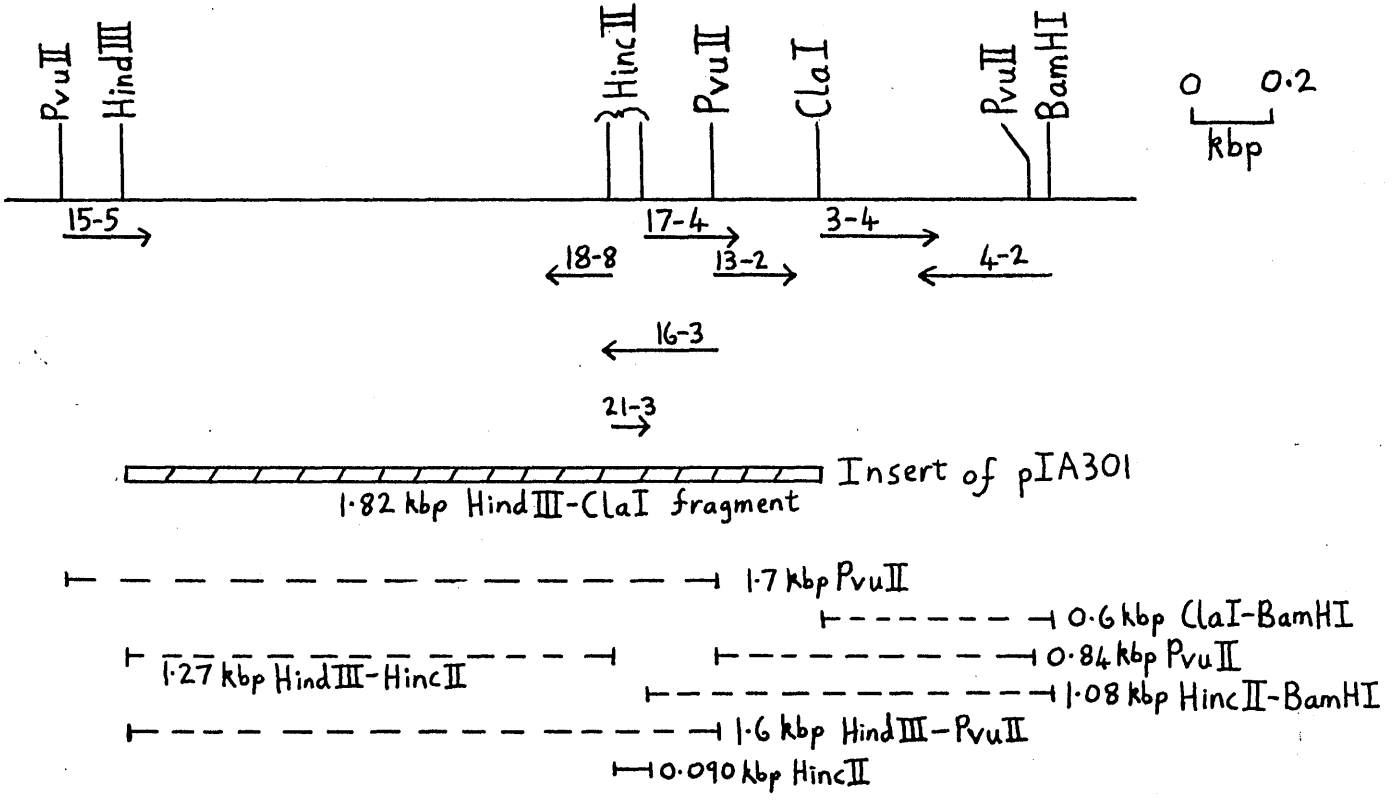


Figure 3.14 First round of M13/dideoxy sequencing.

The horizontal arrows indicate the direction and extent of sequence determinations. Dotted lines represent the fragments cloned in M13 (see Table 3.12). The 1.82 kbp HindIII-ClaI insert fragment of pIA301 is shown.

representative of both possible orientations (e.g. six plaques give a 1 in 64 chance of not getting both orientations). Thus, it is helpful to screen the templates, prior to sequencing, by "T-tracking" in which only the "T" sequencing reaction is performed and the products run on a gel. Clones with the insert in different orientations can be selected by inspection of the auto-radiographs.

Table 3.12 describes the first round M13 clones and Figure 3.14 shows the runs of sequence obtained using these clones. E.coli JM101 was used as the host strain for this stage.

3.5.3 Second round of M13/dideoxy sequencing

It was planned to digest the 1.82 kbp HindIII-ClaI region with a restriction enzyme which cuts frequently (one with a tetranucleotide recognition sequence). The resulting fragments would then be cloned in M13 using a shotgun approach. The 1.82 kbp fragment was purified from a low m.p. agarose gel after electrophoresis of a HindIII + ClaI digest of pIA307 (see Figure 3.15). A sample of the fragment was digested with HpaII (see Table 3.11) and analysed by agarose gel electrophoresis. Figure 3.15 shows the pattern of fragments obtained; one of the seven visible fragments is inconveniently large (0.68 kbp) but, overall, digestion by HpaII offered a useful way of obtaining pieces of the 1.82 kbp region for sequencing.

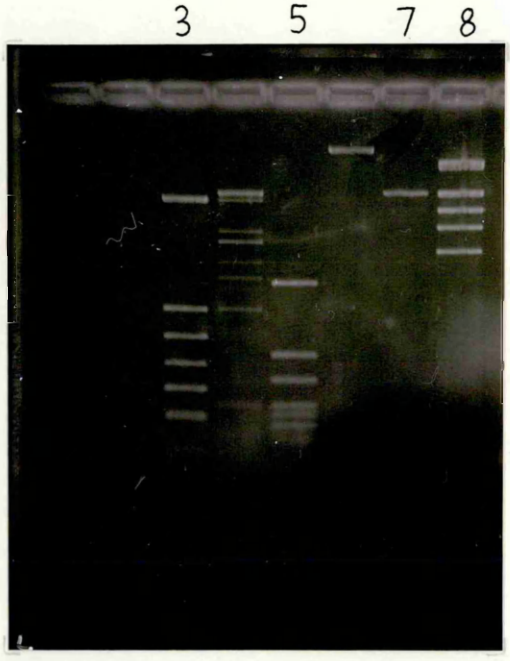


Figure 3.15 Digestion of the 1.82 kbp HindIII-ClaI fragment with HpaII

Lane: 3. pAT153 cut with HinfI, markers (kbp):- 1.631, 0.517, 0.396, 0.298, 0.221 and 0.220, 0.154 and 0.145, 0.075.
5. 1.82 kbp HindIII-ClaI fragment cut with HpaII. Estimated sizes of the digestion products (kbp):- 0.68, 0.33, 0.24, 0.18, 0.15, 0.13, 0.092 (total = 1.80 kbp).
7. Purified 1.82 kbp fragment.
8. pIA307 cut with HindIII + ClaI.

Electrophoresis was through a 2% agarose gel which was subsequently stained with ethidium bromide.

A complete HpaII digest of the 1.82 kbp fragment was cloned randomly ("shotgunned") into AccI cut M13mp9. E.coli TG1 (r^-m^-) was used as the host strain from here on. It was expected that the HpaII-ClaI fragment would be cloned (Table 3.11) but that the HindIII-HpaII fragment would be lost. Templates were prepared from 60 colourless plaques so as to give a reasonable probability of getting all possible strands. 57 templates were analysed by T-tracking. Systematic examination of the T-track autoradiographs revealed, unexpectedly, 15 different classes of clone. Representatives of all the classes were sequenced.

From the sequencing autoradiographs five pairs of complementary strands were identified by inspection and all of these pairs could be localised by comparison with first round sequences. Another strand could be matched to a first round sequence (18-8) and was clearly one strand of the large 0.68 kbp fragment. (The complementary strand was identified very tentatively by the "unreadable" pattern of residues).

One of the three remaining classes contained a tiny HpaII-ClaI fragment with three bases between the two sites. This fragment, which formed part of a larger insert, was presumed to be the ClaI end.

The two sequences that were still unaccounted for remain so to this day. A computer search program failed to locate segments of these last two sequences in the final contiguous sequence of the 1.82 kbp fragment. They presumably originate from incomplete digestion of the

pIA307 during the initial isolation of the 1.82 kbp fragment, causing the latter to be slightly contaminated with partial digest product(s).

The contribution of the second round HpaII clones to the overall sequencing strategy is shown in Figure 3.16 which is drawn partly with the benefit of hindsight. Figure 3.17 shows an example of a sequencing gel autoradiograph.

3.5.4 Third round of M13/dideoxy sequencing

Digestion of the 1.82 kbp fragment with TaqI gave fragments of sizes 0.98, 0.53, 0.18 and 0.12 kbp. This enzyme was thus unsuitable for generating a comprehensive set of small fragments, overlapping the HpaII fragments, for sequencing. However, three TaqI sites had been found (with their potential use in mind) during the sequencing of the HpaII clones. All three lay in the half nearest the ClaI end. Their distribution implied fragments of the sizes observed and was such as to allow completion of the ClaI half of the sequence. For this reason the TaqI fragments were cloned using AccI cut M13mp9 even though it was realised that the largest fragment (which, from 15-5 first round sequence data, extended to the HindIII site) would not be cloned and would anyway have left a large gap in the sequence.

Templates were prepared, sorted by T-tracking and representatives of the useful individual classes were sequenced, all in a manner analogous to that described

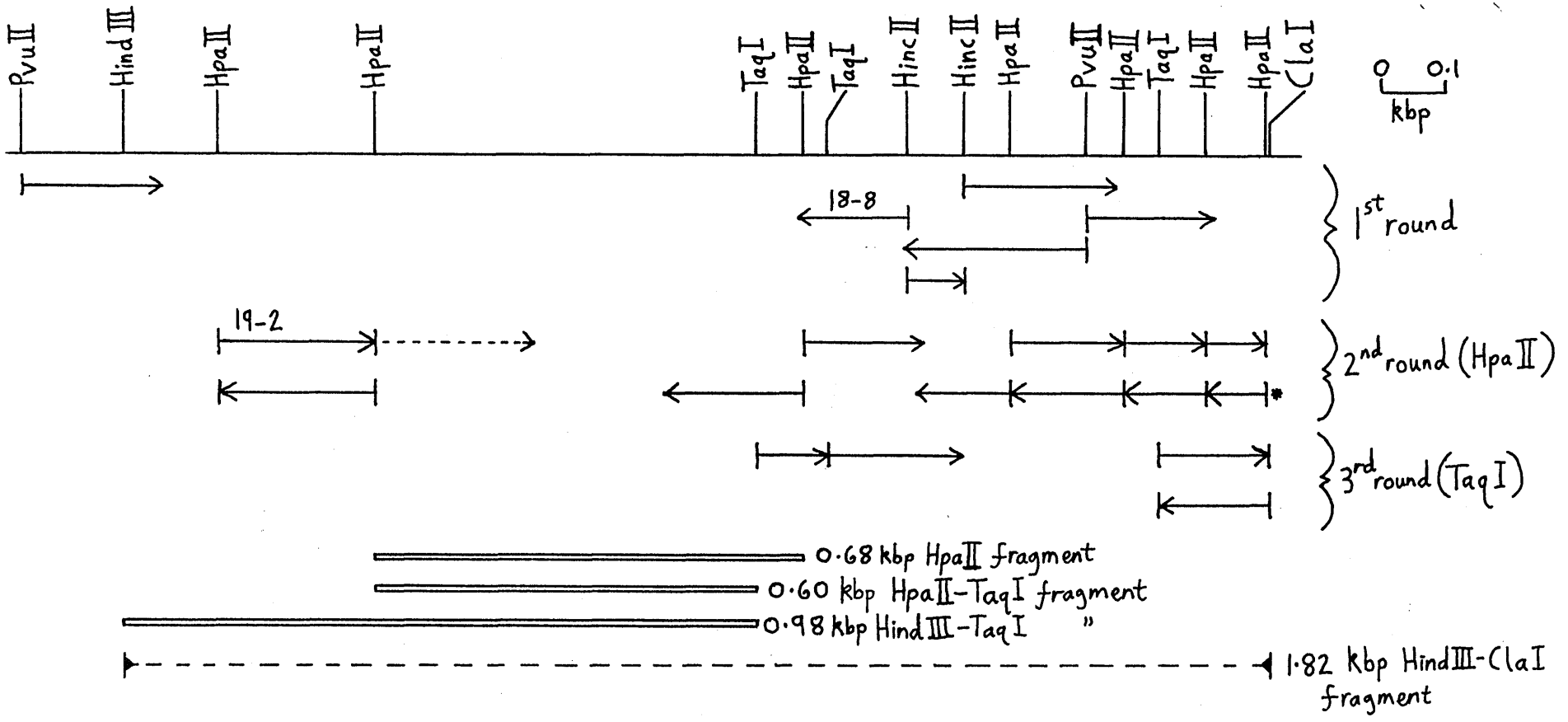


Figure 3.16 Second and third rounds of M13/dideoxy sequencing.

The horizontal arrows indicate the direction and extent of sequence determinations. The broken arrow indicates a second round sequence whose position was only established definitively after the fourth round. *indicates the tiny ClaI-HpaII fragment.

Figure 3.17 An example of a sequencing gel autoradiograph.


22-14
22-21
22-25
19-17
A C G T A C G T A C G T A C G T

for the HpaII clones. The contribution of these TaqI clones to the overall sequencing strategy is shown in Figure 3.16. The presence of a HpaII site three bases from the ClaI site was confirmed.

3.5.5. Fourth round of M13/dideoxy sequencing

After the third round the right hand half of the 1.82 kbp fragment's sequence was essentially complete. It was possible to sequence rightwards from the HindIII site, past the start of 19-2 (see Figure 3.16) by "turning-around" (see below) the insert in the first round clone 18-8. Also, a HaeIII site had been found just after the start of 19-2 and so the complementary strand back to the HindIII site could be obtained by cloning the small HindIII-HaeIII fragment in the correct orientation. Figure 3.16 shows clearly the crucial problem that remained: how to span the major gap in the left hand half of the sequence. Sequencing leftwards from the TaqI site end of the large 0.98 kbp HindIII-TaqI fragment would clearly gain a little ground. A potentially useful Sau3A site was found very near the end of the 19-2 sequence.

To investigate the extent of useful Sau3A, AluI, and HaeIII sites in the vicinity of the gap these enzymes were used to digest the 0.68 kbp HpaII and the 0.98 kbp HindIII-TaqI fragments (see Figure 3.16). These two fragments were obtained by electrophoresing suitable digests of purified 1.82 kbp fragment on a low m.p. agarose gel. The desired bands were cut out of the gel in the

usual way. However, the subsequent analytical digests were performed in molten low m.p. agarose at 37°C without further purification of the DNA (see Chapter Two). HaeIII cleaved off the expected 0.15 kbp fragment from the HindIII end of the 0.98 kbp material but made no other detectable cuts in this nor in the 0.68 kbp HpaII fragment. AluI cut the 0.68 kbp HpaII fragment into 2 pieces of 0.36 and 0.30 kbp. This would prove a very useful site. No additional AluI sites were detected in the 0.98 kbp fragment. Sau3A digestion of the 0.98 kbp fragment revealed four subfragments of 0.33, 0.25, 0.21, and 0.12 kbp. These Sau3A sites would also prove very useful in completing the sequence.

It proved possible and convenient to cut both the 0.60 kbp HpaII-TaqI (see Figure 3.16) and the 0.98 kbp HindIII-TaqI fragments directly from gels of the corresponding digests of pIA307. These two fragments provided most of the raw material for the fourth round of M13 cloning work which is summarised in Table 3.13.

A HaeIII digest of the 0.98 kbp fragment was ligated with HindIII + HincII cut M13mp9. Only the sought after small HindIII-HaeIII fragment was expected to be cloned (and in the correct orientation).

The first round clone 18-8 insert (the 1.27 kbp HindIII-HincII fragment cloned in HindIII + HincII cut M13mp9) was "turned-around" (see Chapter Two) by synthesising a replicative form <u>in vitro</u> and cutting out the insert from the now double-stranded molecule by an EcoRI + HindIII

Table 3.13 Fourth round of M13 cloning

Ligation	Insert	Vector
24	HindIII-EcoRI insert from first round clone 18-8	HindIII + EcoRI cut mp8
25	0.60 kbp HpaII-TaqI	AccI cut mp9
26	AluI cut 0.60 kbp HpaII-TaqI	AccI + SmaI cut mp9
27	Sau3A cut 0.98 kbp HindIII-TaqI	BamHI cut mp8
28	"	BamHI + AccI cut mp9
29	HaeIII cut 0.98 kbp HindIII-TaqI	HindIII + HincII cut mp9

digest. The products of this digestion were separated by electrophoresis in a low m.p. agarose gel and the insert band was cut out, purified, and ligated with HindIII + EcoRI cut M13mp8, thus placing the HindIII site of the insert adjacent to the primer.

The 0.60 kbp HpaII-TaqI fragment was ligated into AccI cut M13mp9. One of the possible orientations permits sequencing in from the TaqI end.

The 0.60 kbp HpaII-TaqI fragment was digested with AluI and ligated with AccI + SmaI cut M13mp9. Both of the expected inserts would be orientated to permit sequencing in from the AluI ends (i.e. outwards, in both directions, from the middle of the 0.60 kbp fragment).

A Sau3A digest of the 0.98 kbp fragment was ligated with BamHI cut M13mp8, to clone Sau3A fragments, and also with BamHI + AccI cut M13mp9 to clone the Sau3A-TaqI fragment in the correct orientation for sequencing in the direction of the TaqI site.

Templates derived from these ligations were screened, where necessary, by T-tracking and then sequenced. The runs of sequence obtained from these fourth round clones are shown in Figure 3.18, and they proved sufficient to complete the sequence.

3.5.6 The overall sequencing strategy

Figure 3.19 shows the overall strategy used to sequence the 1.82 kbp HindIII-ClaI fragment, and the sequence itself is shown in Figure 3.21. The complete sequences of both

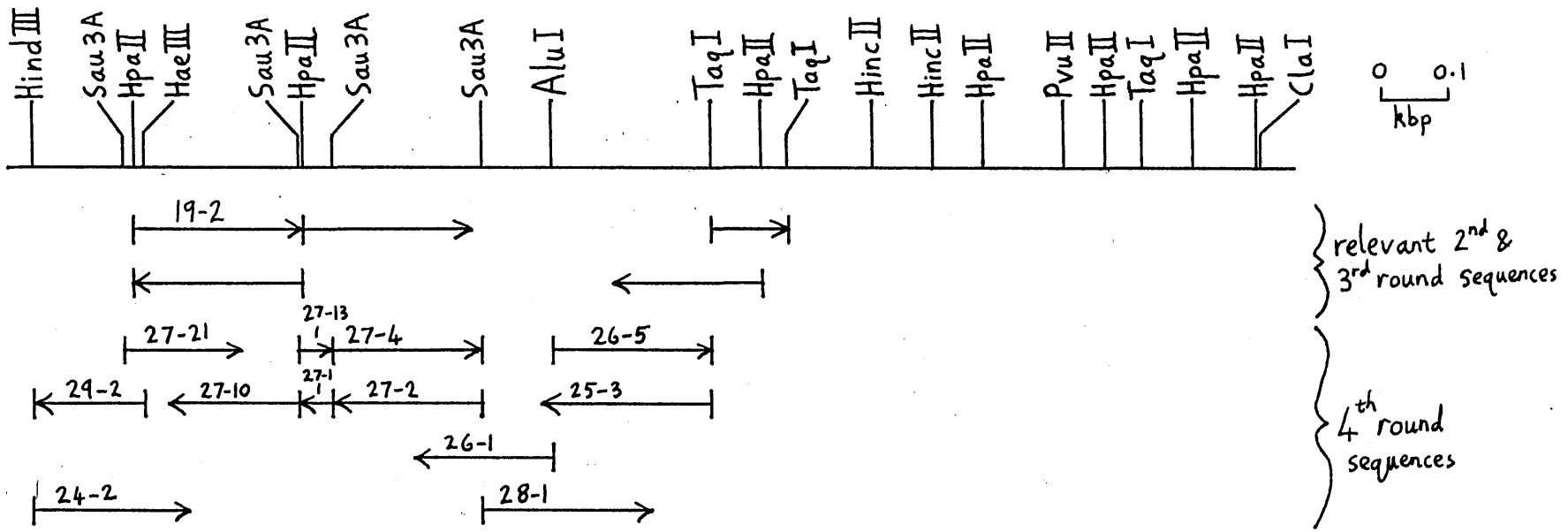


Figure 3.18. Fourth round of M13/dideoxy sequencing

The horizontal arrows indicate the direction and extent of sequence determinations. Only the relevant Sau3A, HaeIII, and AluI sites are shown. The first number above each of the fourth round sequence runs indicates which ligation that run originated from (see Table 3.13).

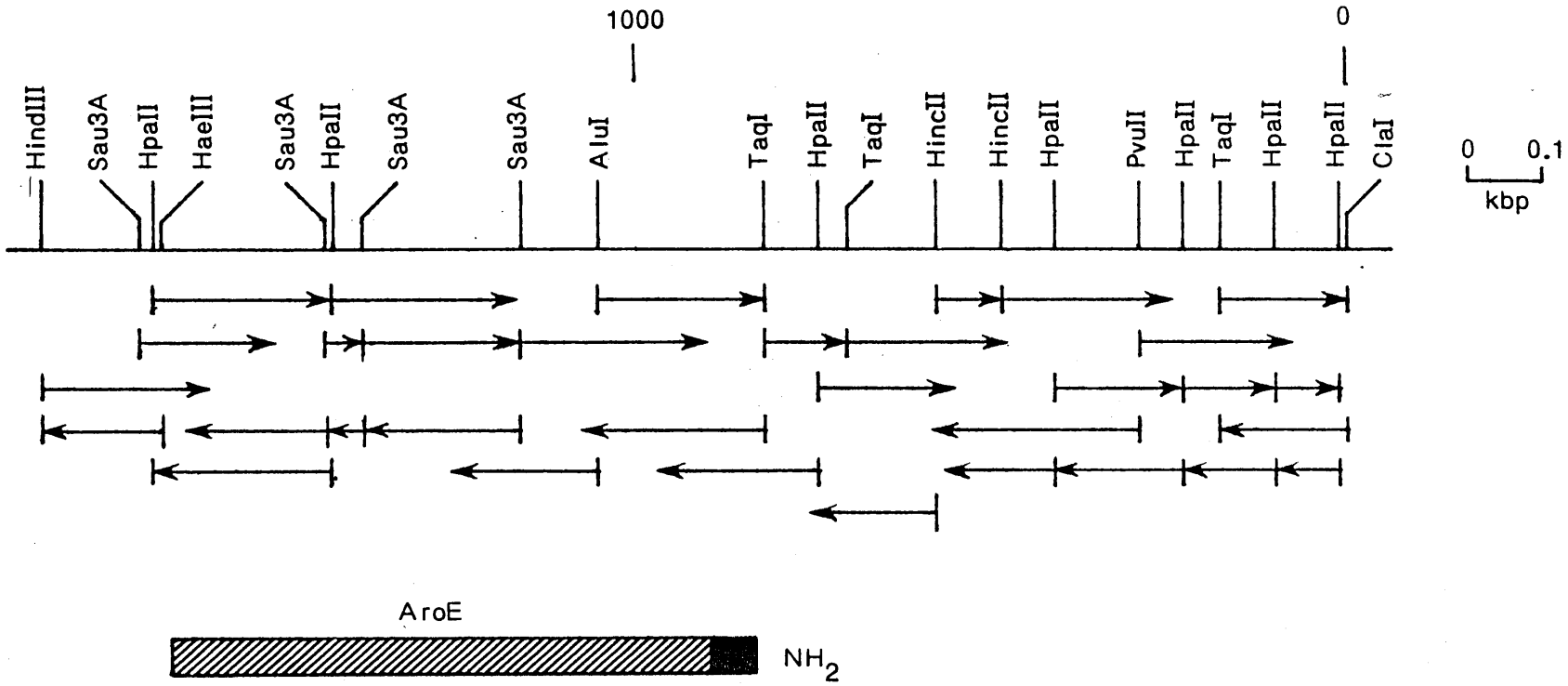


Figure 3.19 Overall DNA sequencing strategy

As previously, the horizontal arrows indicate the direction and extent of sequence determinations. Only the relevant Sau3A, HaeIII, and AluI sites are shown. At the top "1000" and "0" indicate bp from the ClaI site.

The box labelled AroE represents the aroE coding sequence and is based on work described in Chapter Four. The extent of N-terminal amino acid sequence determination is indicated by the solid shading.

strands were determined and all the restriction sites used were overlapped.

The weakest overlap is that between the sequence of the HpaII clone 19-2 and the sequence of the small Sau3A clone 27-13 (see Figure 3.18). The sequence at the end of the 19-2 run reads (5' to 3')...GATCGCCG(G), the last base being inferred from the fact that the insert in 19-2 is a HpaII fragment. Only a single base separates the GATC Sau3A site from the CCGG HpaII site and the effective overlap is thus five bases. However, error could only arise from a very unlikely possibility: the existence of another unsuspected pair of Sau3A and HpaII sites, separated by a G residue, and lying within about 50 bases of the known pair. A greater separation would make it almost certain that the hypothetical HpaII fragment would have been detected amongst the products of the analytical HpaII digest of the 1.82 kbp fragment shown in Figure 3.15. Even a separation of 50 bases would have led to a marked discrepancy between the observed - 1.82 kbp - and predicted sizes of the HindIII-ClaI fragment. The sequence shown in Figure 3.21 contains 1,826 bases.

An experiment was done to confirm the sequence of this region, despite the low probability of being misled by the short overlap. Clone 24-2 (see Figure 3.18 and T bles 3.13 and 3.12) permits the 1.27 kbp HindIII-HincII fragment to be sequenced from the HindIII end. This clone was resequenced using conditions biased towards being able

Figure 3.20 An example of a "compression"

In panel (iii) the sequence around the arrow can be read unambiguously as ...ATGCCCTGA... Panel (i) shows the complementary strand of the same region; here the three G residues corresponding to the three C's in panel (iii) are compressed, making it impossible to count the number of G's (although measurements of the spacing of widely flanking bands and comparison with adjacent, uncompressed sequences - not shown - allow an estimate of three to be made). Panel (ii) shows the same strand as in panel (i) but from a gel run at a higher temperature (hence the "fuzziness" of the bands) to check the conclusions from the complementary strand. Here it seems more plausible that there are indeed three G residues although clearly the conditions were not yet fully denaturing.

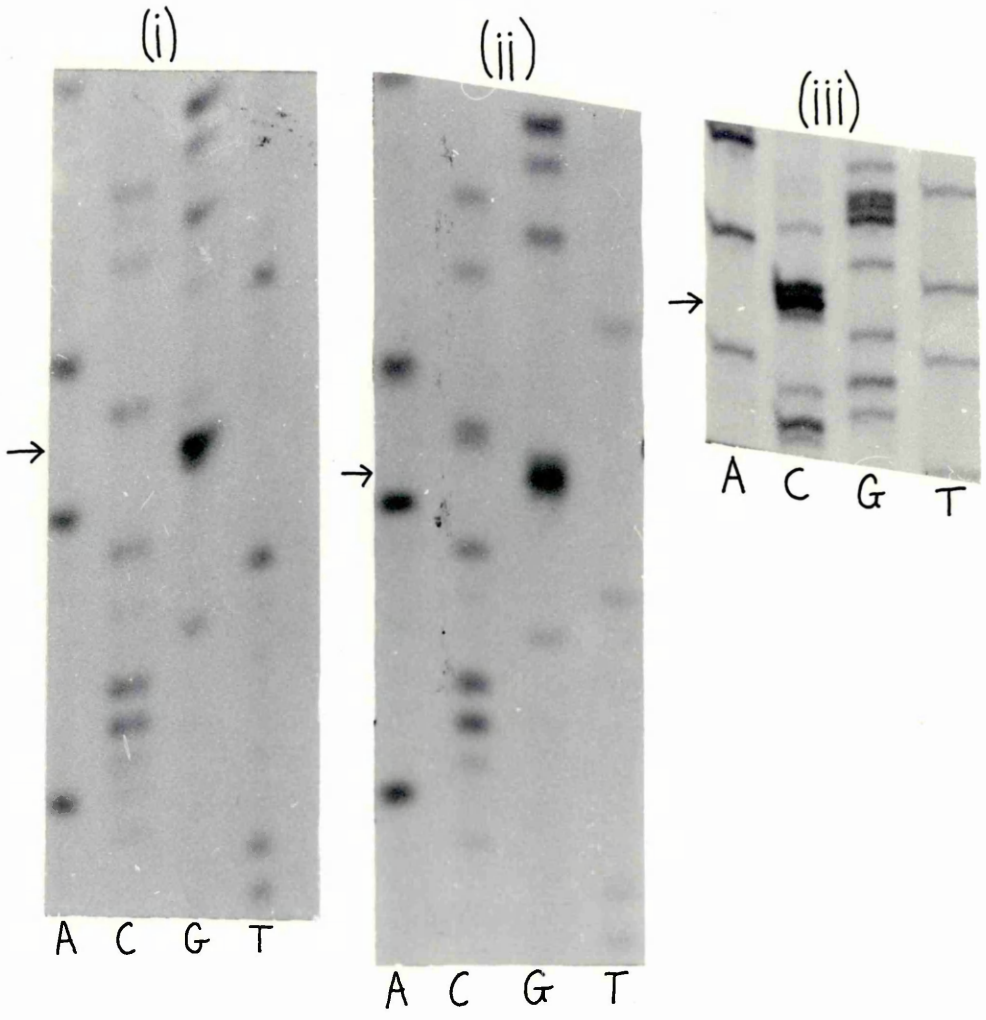

(i)
(ii)
(iii)
A C G T
A C G T
A C G T

Figure 3.21 Sequence of the 1.82 kbp HindIII-ClaI fragment

The sequence of one strand of the HindIII-ClaI fragment is shown. It is written 5' to 3' (as usual) starting from the ClaI end. Thus it begins with the 4 bases of a ClaI site which remain after ClaI digestion and ends with the first base of a HindIII site.

```
      10         20         30         40         50         60         70         80         90         100        110        120
CGATAAACCG GACGAAACAG CAATTACATG CCCCCAATGT CGGACGGGCC ATCTGGTCCA GCGCCGCTCC CGTTATGGCA AAACATTTCA CTCTTGTGAT CGCTACCCGG AGTGTCAATT

      130        140        150        160        170        180        190        200        210        220        230        240
TGCCATTAAC TTCAAACCCA TAGCTGGAGA ATGCCCTGAG TGTCATTATC CGCTACTCAT CGAAAAGAAA ACCGCGCAGG GTGTAAAACA CTTTTGTGCC AGTAAACAAT GTGGAAAGCC

      250        260        270        280        290        300        310        320        330        340        350        360
GGTTTCGGCG GAATAATAAC GTGAATAATA ACCTGCAAAG AGACGCTATC GCAGCTGCGA TAGATGTTCT CAATGAAGAA CGTGTCATCG CCTATCCAAC GGAAGCCGTT TTCGGTGTTG

      370        380        390        400        410        420        430        440        450        460        470        480
GGTGCGATCC TGATAGCGAA ACAGCAGTGA TGCGACTGTT GGAGTTAAAA CAGCGTCCGG TTGATAAGGG GCTGATTTTA ATCGCAGCAA ATTACGAGCA GCTTAAACCC TATATTGATG

      490        500        510        520        530        540        550        560        570        580        590        600
ACACCATGTT GACTGACGTG CAGCGTGAAA CCATTTTTTC CCGCTGGCCA GGTCCTGTCA CCTTTGTCTT TCCCGCGCCT GCGACAACAC CGCGCTGGTT GACGGGCCGC TTTGATTCGC

      610        620        630        640        650        660        670        680        690        700        710        720
TTGCTGTACG AGTCACCGAC CATCCGTTGG TGGTTGCTTT GTGCCAGGCT TATGGTAAAC CGCTGGTTTC TACCAGTGCC AACTTGAGTG GATTGCCACG TTGTCGAACA GTAGACGAAG

      730        740        750        760        770        780        790        800        810        820        830        840
TTCGCGCACA ATTTGGCGCG GCGTTCCCGG TTGTGCCTGG TGAAACGGGG GGGCGTTTAA ATCCTTCAGA AATCCGCGAT GCCCTGACGG GTGAACTGTT TCGACAGGGG TAACATAATG

      850        860        870        880        890        900        910        920        930        940        950        960
GAAACCTATG CTGTTTTTGG TAATCCGATA GCCCACAGCA AATCGCCATT CATTCATCAG CAATTTGCTC AGCAACTGAA TATTGAACAT CCCTATGGGC GCGTGTTGGC ACCCATCAAT

      970        980        990        1000       1010       1020       1030       1040       1050       1060       1070       1080
GATTTCATCA ACACACTGAA CGCTTTCTTT AGTGCTGGTG GTAAAGGTGC GAATGTGACG GTGCCTTTTA AAGAAGAGGC TTTTGCCAGA GCGGATGAGC TTACTGAACG GGCAGCGTTG

      1090       1100       1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
GCTGGTGCTG TTAATACCCT CATGCGGTTA GAAGATGGAC GCCTGCTGGG TGACAATACC GATGGTGTAG GCTTGTTAAG CGATCTGGAA CGTCTGTCTT TTATCCGCCC TGGTTTACGT

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300       1310       1320
ATTCTGCTTA TCGGCGCTGG TGGAGCATCT CGCGGCGTAC TACTGCCACT CCTTTCCCTG GACTGTGCGG TGACAATAAC TAATCGGACG GTATCCCGCG CGGAAGAGTT GGCTAAATTG

      1330       1340       1350       1360       1370       1380       1390       1400       1410       1420       1430       1440
TTTGCGCACA CTGGCAGTAT TCAGGCGTTG AGTATGGACG AACTGGAAGG TCATGAGTTT GATCTCATTA TTAATGCAAC ATCCAGTGGC ATCAGTGGTG ATATTCCGGC GATCCCGTCA

      1450       1460       1470       1480       1490       1500       1510       1520       1530       1540       1550       1560
TCGCTCATTC ATCCAGGCAT TTATTGCTAT GACATGTTCT ATCAGAAAGG AAAAACTCCT TTTCTGGCAT GGTGTGAGCA GCGAGGCTCA AAGCGTAATG CTGATGGTTT AGGAATGCTG

      1570       1580       1590       1600       1610       1620       1630       1640       1650       1660       1670       1680
GTGGCACAGG CGGCTCATGC CTTTCTTCTC TGGCACGGTG TTCTGCCTGA CGTAGAACCA GTTATAAAGC AATTGCAGGA GGAATTGTCC GCGTGAATCA GGCCATCCAG TTTCCGGACA

      1690       1700       1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
GGGAAGAGTG ATCTTACCCA GCAATAGTGG ACACGCGGCT AAGTGAGTAA ACTCTCAGTC AGAGGTGACT CACATGACAA AAACAGTATC AACCAGTAAA AAACCCCGTA AACAGCATTC

      1810       1820
GCCTGAATTT CGCAGTGAAG CCCTGA
```

to read the sequence at long distances from the primer site (up to 400 bases; the ratio of dideoxynucleotides to deoxynucleotides is reduced and the period of electrophoresis is extended - see Chapter Two for details). This extended 24-2 sequence matched the 19-2 sequence as far as the HpaII site and continued (with slight uncertainty at those positions underlined):-

.... -A-A-T-A-T-C-A-C-C-A-(?)-T-G-A-T- ....

This matches the sequence of 27-13 after the HpaII site:-

.... -A-A-T-A-T-C-A-C-C-A-C-T-G-A-T- ....

Hence, the extended 24-2 sequence provides a much more reliable overlap.

3.6 Conclusion

The cloning and sequencing of a region of DNA able to complement an *E.coli aroE*$^-$ mutant has been described. Only after further work described in the next chapter was it certain that the *aroE* structural gene lay within this region.

CHAPTER 4 LOCALISATION OF THE AROE CODING SEQUENCE AND CONFIRMATION OF THE DNA SEQUENCE BY ANALYSIS OF THE OVERPRODUCED POLYPEPTIDE

4.1 Localisation of the aroE coding sequence

4.1.1 Open reading frames in the 1.82 kbp HindIII-ClaI sequence

As a first step towards locating the aroE coding region, within the sequence of the 1.82 kbp HindIII-ClaI fragment, the computer program TRNTRP (Staden, 1978) was used to translate both strands of the sequence in all three possible reading frames. For each of the six frames the positions of stop codons were plotted (see Figure 4.1). The two largest open reading frames (ORF's) are both found in the ClaI(5')-HindIII(3') strand.

The second largest ORF has 624 bases between stop codons, 615 bases from the start of the first plausible initiation codon (GTG), and 441 bases from the start of the first methionine codon (ATG). Even the 615 base stretch is unlikely to encode a polypeptide larger than about 23 kDa, taking 110 Da as the average m.w. of an amino acid residue.

The largest ORF has 846 bases between stop codons, 843 bases from the first plausible initiation codon, and 816 bases from the first ATG. From the first methionine this ORF would encode a polypeptide of about 30 kDa. At about this time E3 was purified to homogeneity from E.coli K12 in this laboratory (Chaudhuri and Coggins, 1985)

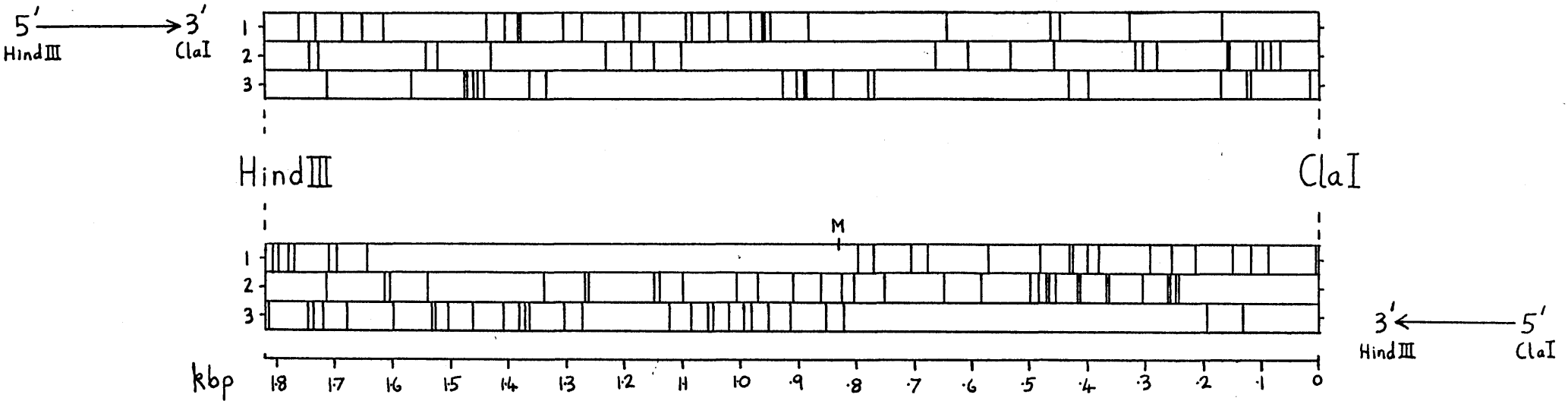


Figure 4.1 Open reading frames in the 1.82 kbp HindIII-ClaI sequence.

The sequence of each strand was translated in all three possible reading frames using the computer program TRNTRP. Unadorned vertical lines indicate the positions of stop codons. "M" shows the position of the first methionine in the largest open reading frame.

and the subunit m.w. observed on SDS Laemmli gels was 30 kDa. Hence, it appeared that the largest ORF was much more likely to contain the aroE coding sequence than the second largest one.

4.1.2 Further subcloning of aroE using pAT153

Further subcloning experiments were performed to confirm that only the largest ORF could contain the aroE gene. The 1.27 kbp HindIII-HincII fragment (see Figure 4.2) carries the whole of the largest ORF but includes only the latter 250 bases of the second largest ORF. This fragment was cloned by ligating HindIII + NruI cut, and phosphatased, pAT153 with a HincII digest of the purified 2.47 kbp HindIII-BamHI fragment of pIA307, transforming E.coli AB2834 (aroE$^-$), and screening Amp^R transformants for the Aro$^+$ phenotype in the way described previously. Five of the six independent isolates were found to contain the desired insert - this construct was called pIA309. (The sixth isolate contained the very small 0.090 kbp HincII fragment in addition to the 1.27 kbp HindIII-HincII fragment). Back transformation confirmed that pIA309 can complement E.coli AB2834. This rules out the 624 base ORF as a possible location for the aroE gene, leaving the largest ORF as the only candidate. How this might relate to the specific activity results discussed in Section 3.3.6 will be considered in Chapter Five.

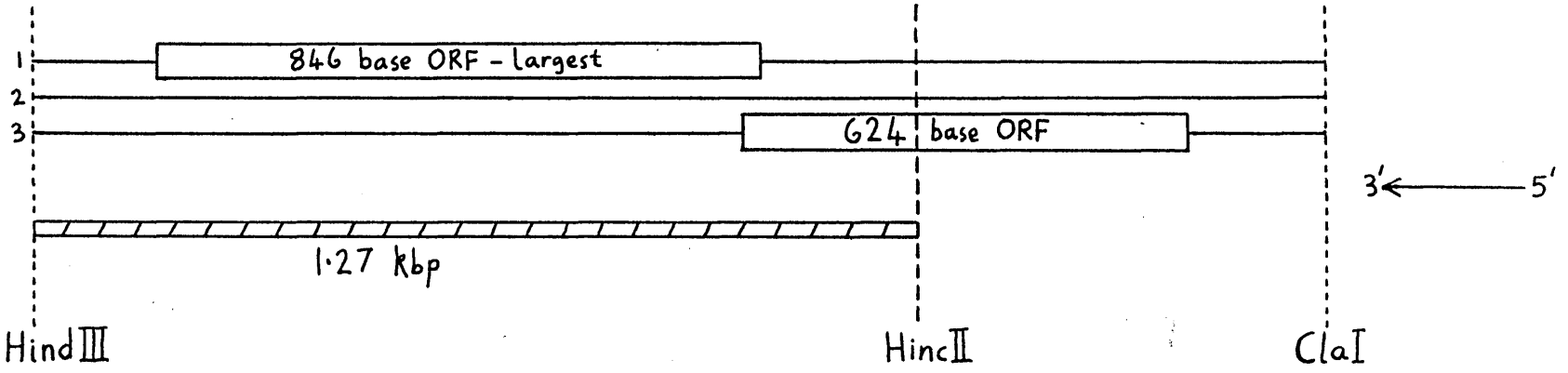


Figure 4.2 Relationship of the 1.27 kbp HindIII-HincII fragment to the two largest ORF's.

The two largest ORF's in the 1.82 kbp HindIII-ClaI sequence are shown schematically. The 1.27 kbp HindIII-HincII fragment, which forms the insert of pIA309, spans the whole of the largest ORF but includes only the latter 250 bases of the second largest.

4.1.3 The need for analysis of the E3 polypeptide

That only the largest ORF could contain the *aroE* coding sequence is not conclusive evidence that it actually does. As mentioned in Section 3.2.2 there was the possibility (remote, given the observed overexpression) of being misled by a suppressor gene. Also, assuming that the largest ORF did represent *aroE*, there was the need to locate the translational initiation codon. This need could be met by determination of the N-terminal amino acid sequence of purified overproduced E3 and this approach also offered a way of showing definitively that the ORF sequence encodes E3. It was also decided to determine the amino acid composition of the purified overproduced E3 since this would confirm the overall sequence.

4.2 Construction of a strain which greatly overproduces *E.coli* E3

4.2.1. Justification

The procedure developed by Chaudhuri and Coggins (1985) for the purification of E3 from wild-type *E.coli* K12 yields about 0.013 mg of homogeneous polypeptide from 20g (wet weight) of cells. Thus, it would have been tedious to purify enough wild-type E3 for determination of the N-terminal amino acid sequence (using a conventional liquid phase sequencer) and for determination of the amino acid composition. The levels of E3 overexpression obtainable with subclones, such as pIA307, would have

overcome this difficulty. However, the wish to raise antibodies against *E.coli* E3, and especially the long term aim of obtaining crystals suitable for high resolution X-ray crystallography, all prompted consideration of ways of obtaining rather higher levels of E3 overexpression than could be gained from the copy number of the vector pAT153.

4.2.2 Expression vector pKK223-3

In seeking higher levels of *E.coli* E3 overexpression the approach adopted was to place the putative coding sequence downstream, in the correct orientation, from a powerful promoter. The expression vector pKK223-3 was used for this purpose (J. Brosius, unpublished results; this vector is available commercially from Pharmacia P-L Biochemicals Inc. but for this study it was obtained through the generosity of J.R. Knowles). This vector contains the strong *trp-lac* hybrid *tac*(I) promoter (De Boer *et al.*, 1983) which has the "-35" region of the *E.coli trp* promoter and the "-10" region of the *E.coli lac*UV5 promoter; the *lac*UV5 segment also contains the *lac* operator and the Shine-Dalgarno sequence although the latter is irrelevant here. As shown in Figure 4.3 there is an M13mp8 polylinker situated downstream from the *tac* promoter, thus facilitating the positioning of genes behind the promoter. The polylinker is followed by a DNA segment containing the strong *rrn*B ribosomal RNA transcription

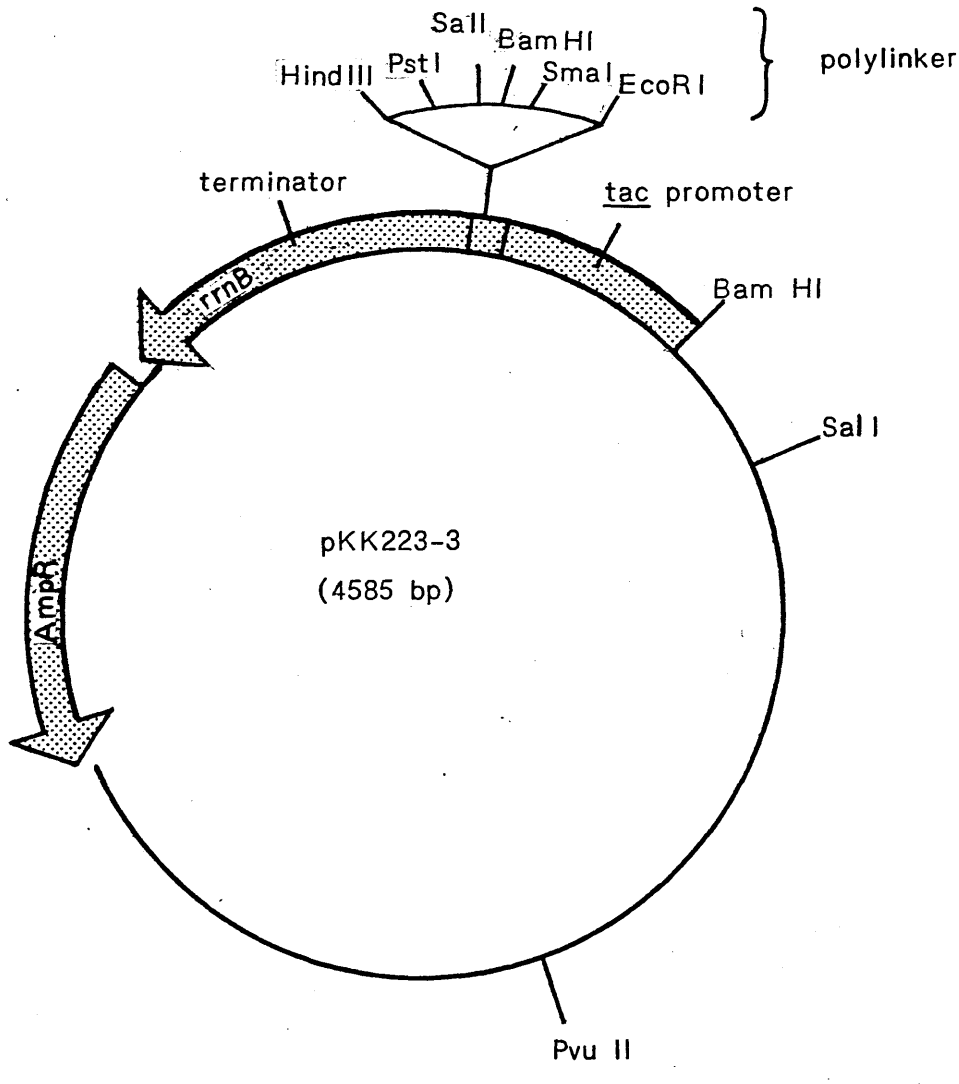


Figure 4.3 Expression vector pKK223-3. This vector carries the Amp^R selectable marker.

terminators. The position of these terminators stabilises the plasmid (Gentz *et al.*, 1981), the remainder of which consists of pBR322 sequences. The *tac* promoter is controllable by the *lac* repressor and can be induced by IPTG. However, transcription from the *tac* promoter of pKK223-3 is only partially inhibited in uninduced cells carrying the wild-type *lacI* gene due to the multiple copies of the plasmid titrating out the *lac* repressor.

pKK223-3 has been used by Frost *et al.* (1984) to overexpress the *E.coli aroB* gene which encodes dehydroquinate synthase. They obtained about one thousand times the normal wild-type level of enzyme.

4.2.3 Construction of pIA321

The 1.27 kbp HindIII-HincII fragment (see Figure 4.2) was cloned into pKK223-3 so that it would be transcribed in the HincII to HindIII direction, the correct orientation for overexpressing the product of the largest ORF. A HincII digest of the purified 2.47 kbp HindIII-BamHI fragment of pIA307 was ligated with SmaI + HindIII cut pKK223-3, *E.coli* AB2834 was transformed, and Amp^R transformants were tested for the Aro^+ phenotype. Six independent isolates were examined by restriction analysis and found to be identical. None contained the very small 0.090 kbp HincII fragment. This construct was called pIA321 (see Figure 4.4).

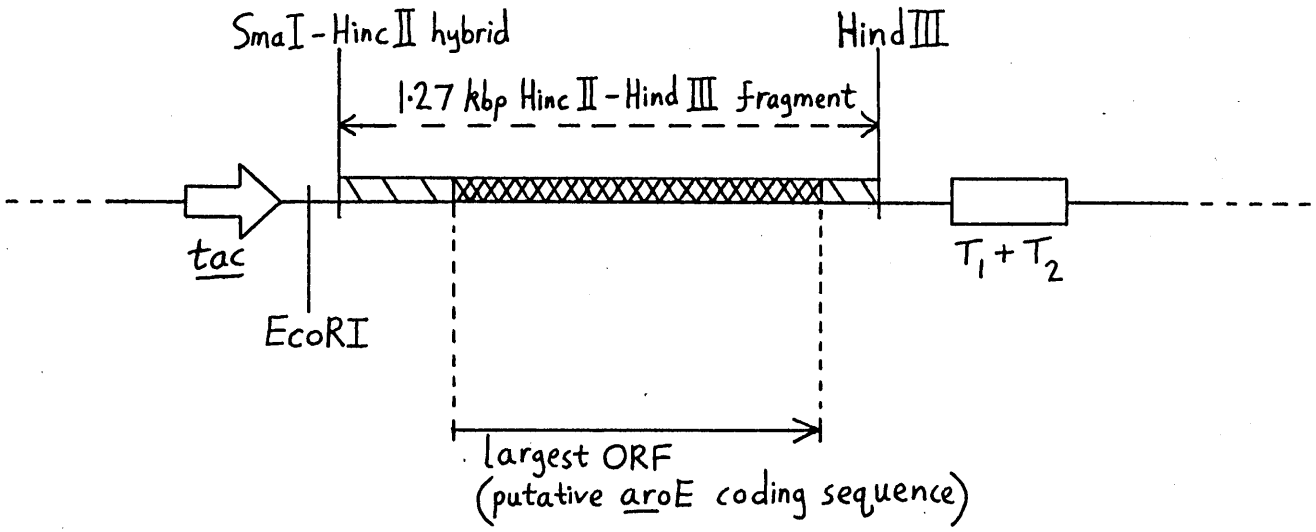


Figure 4.4 pIA321.

The structure of pIA321 is shown schematically and not to scale. The 1.27 kbp HincII-HindIII fragment, which contains the putative aroE coding sequence, is cloned between the SmaI and HindIII sites of the pKK223-3 polylinker as shown. The terminators (T_1 + T_2) are also shown.

4.2.4 Specific activity of E3 in pIA321//AB2834

Table 4.1 summarises the E3 specific activity values observed in crude cell extracts of pIA321//AB2834 grown to various final cell densities in L Broth medium. IPTG gives a worthwhile improvement in the specific activity (increases of 37% and 48% in the two cases shown in Table 4.1). From earlier results the correlation between higher final A_{650} values and higher E3 specific activities is not unexpected. However, the specific activity of pIA321//AB2834(1) grown to A_{650} = 3.08 shows a useful three fold increase over that of pIA306//AB2834 grown to A_{650} = 2.9, the most similar A_{650} available from previously described results (see Table 3.9). Thus, pIA321 is significantly better than any previous subclone.

When pIA321//AB2834 was grown to higher final A_{650} values than had been tried previously ($A_{650} \doteq 4$ or 4.7, where the culture is almost stationary) the E3 specific activity, 30-33 u/mg, was almost double that obtained at A_{650} = 3. Although no *E.coli* K12 cultures were grown to such high cell densities it is worth noting that a specific activity of 33 u/mg is three hundred times greater than the E3 specific activity in crude extracts of *E.coli* K12 grown to a final A_{650} of 3.1 (0.11 u/mg). This is not as good as the one thousand fold factor obtained by Frost *et al.* (1984), but there are many possible explanations for this. For example, it is conceivable that there is a weak transcriptional terminator between the HincII site and the start of the ORF.

Table 4.1 E3 specific activities in crude cell extracts of E.coli AB2834 carrying pIA321

Strain and growth conditions	Final A_{650} of culture	E3 specific activity (U/mg) in crude cell extract
pIA321//AB2834(1)	1.65	7.3
" " " + IPTG	1.63	10.0
" " " + IPTG	3.08	16.4
" " " + IPTG	4.15	31.0
" " " + IPTG	4.75	33.0
pIA321//AB2834(2)	1.62	6.4
" " " + IPTG	1.58	9.5
" " " + IPTG	3.95	30.0
" " " + IPTG	4.7	33.0

NOTES:
1. All crude extracts were made from cells grown in L Broth supplemented with ampicillin.
2. The inducer IPTG was used to obtain maximal expression from the *tac* promoter of pIA321. It was added to a final concentration of 5×10^{-4}M $1\frac{1}{2}$ hr after inoculation of the cultures (see Chapter Two).
3. Two independent isolates of pIA321 were used: (1) & (2).

The specific activity of E.coli E3 purified to homogeneity is 1200 U/mg (Chaudhuri and Coggins, 1985). Thus, E3 constitutes about 3% of the total protein in small scale crude cell extracts of pIA321//AB2834 grown almost to stationary phase.

4.3 Large scale purification of overproduced E.coli E3 from pIA321//AB2834

4.3.1 Objectives and approach

The need to determine the N-terminal amino acid sequence of E3 was discussed in Section 4.1.3 as were the advantages of determining the amino acid composition of the protein. For the accomplishment of these two tasks a total of about 5 mg of homogeneous E3 was required. The experiments described in Section 4.2.4, together with the previous work on the purification of E3 from E.coli K12 (Chaudhuri and Coggins, 1985), suggested that the required amount of pure E3 could be obtained easily from 20g (wet weight) of cells of the overproducer strain pIA321//AB2834.

The method developed by Chaudhuri and Coggins (1985) for the isolation of pure E3 from E.coli K12 is summarised below:

1. Cell breakage
2. $(NH_4)_2SO_4$ fractionation
3. Ion-exchange chromatography on DEAE-Sephacel
4. Gel filtration on Sephacryl S-200

5. Affinity chromatography on ADP-Sepharose
6. Ion-exchange chromatography on a Pharmacia Mono Q FPLC column.

The degree of purification afforded by this procedure (approximately 20,000 fold) is greater than that required when an overproducing strain such as pIA321//AB2834 is used as the starting material. The procedure used was therefore a simplified version in which the final step was omitted.

4.3.2 Growth of cells

The medium used was L Broth plus ampicillin, as for the specific activity experiments. IPTG was added as an inducer. For small scale analytical crude extracts cells were always grown in 50ml cultures in 250ml conical flasks on an orbital shaker. For the production of pIA321//AB2834 cells on a larger scale 13 x 500ml cultures were grown up in 2000ml conical flasks on an orbital shaker. Perhaps due to inferior aeration the growth of these large scale cultures levelled off at a somewhat lower A_{650} than expected. The cultures were eventually taken for harvesting at A_{650} = 3.7-3.8. A total of about 40g (wet weight) of cell paste was obtained from the 6.5 l of culture.

4.3.3 Purification of E3 from pIA321//AB2834

Further details are given in Chapter Two. The starting material was 20g (wet weight) of pIA321//AB2834 cells and the purification of overproduced E3 is summarised in Table 4.2.

Table 4.2 Purification scheme for shikimate dehydrogenase from pIA321 transformed E. coli AB2834

Step	Vol (ml)	Concn. of Protein (mg/ml)	Total Protein (mg)	Activity (units/ml)	Total Activity (units)	Specific Activity (units/mg)	Purification (fold)	Yie (%)
1. Crude extract	75	20.8	1560	414	31100	19.9	1	100
2. 30-55% Satn. $(NH_4)_2SO_4$	36	24	864	807	29100	33.6	1.7	94
3. DEAE-Sephacel	33	1.9	63	539	17800	284	14.3	57
4. Sephacryl S-200	15.5	1.05	16.3	670	10400	638	32.1	33
5. ADP-Sepharose	14.6	0.72	10.5	619	9040	860	43.2	29

NOTES: 1. The results presented are for a purification starting from 20g (wet weight) of cells.
2. E3 was, as usual, assayed in the reverse direction at 25°C in 100mM Tris-HCl pH 9.0 buffer (by following the reduction of $NADP^+$ at 340 nm). Chaudhuri and Coggins (1985) used 100mM Na_2CO_3 pH 10.6 but an otherwise identical assay - their values were multiplied by the empirically determined factor of 1.09, to make them directly comparable with values determined here, before being quoted in the text.

The cells were broken by two passages through a French pressure cell. This was followed by deoxyribonuclease I treatment and centrifugation at 30,000 x g for 30 min: the supernatant represented the crude extract. The specific activity given for the crude extract in Table 4.2 is not directly comparable with those given in Table 4.1 because all small scale crude cell extracts were prepared in a different way - cells were disrupted by sonication which was followed by a high speed spin at 199,000 x g for 2 h : the supernatant formed the crude extract. Chaudhuri and Coggins (1985) obtained a specific activity of 0.059 units/mg for their crude extract of E.coli K12 (grown on minimal medium) compared with 19.9 units/mg for the pIA321//AB2834 crude extract here, a 337-fold difference.

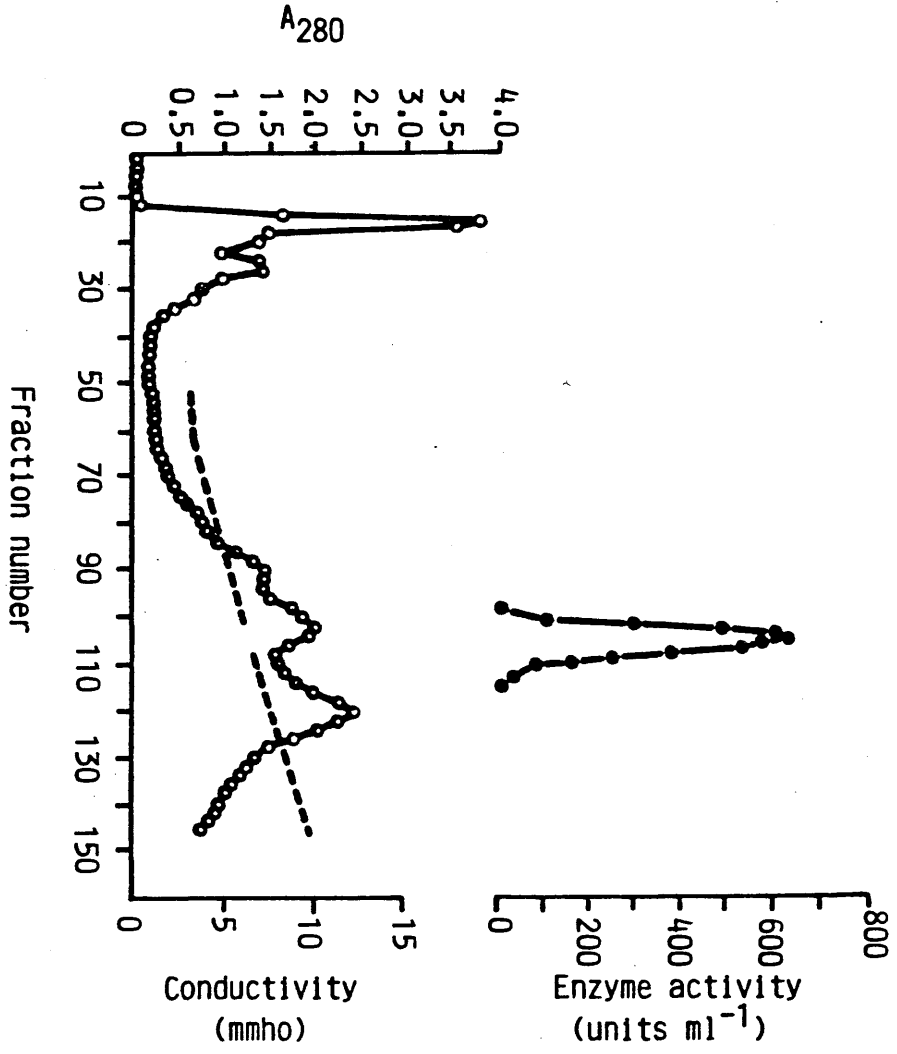
After $(NH_4)_2SO_4$ fractionation of the crude extract the E3 activity was found in the 30-55%-saturation fraction.

The third step was ion-exchange chromatography on DEAE-Sephacel (Figure 4.5). The peak fractions that were pooled from the DEAE-Sephacel column contained a total of only 60% of the activity originally loaded - this deliberate paring of the yield was done in the hope of minimising the number of chromatographic steps required to obtain pure overproduced E3. The pool, which contained 17,800 units at a specific activity of 284 units/mg, was then concentrated prior to the fourth step. (An accident during the concentration procedure resulted in only about 83% of the material being carried forward to the next step, although the author's loss was the cold room floor's gain.) In Table 4.2 the yield after the fourth step (Sephacryl S-200)

Too faint

Figure 4.5 Chromatography of overproduced E.coli E3 on DEAE-Sephacel (step 3 of the purification scheme).

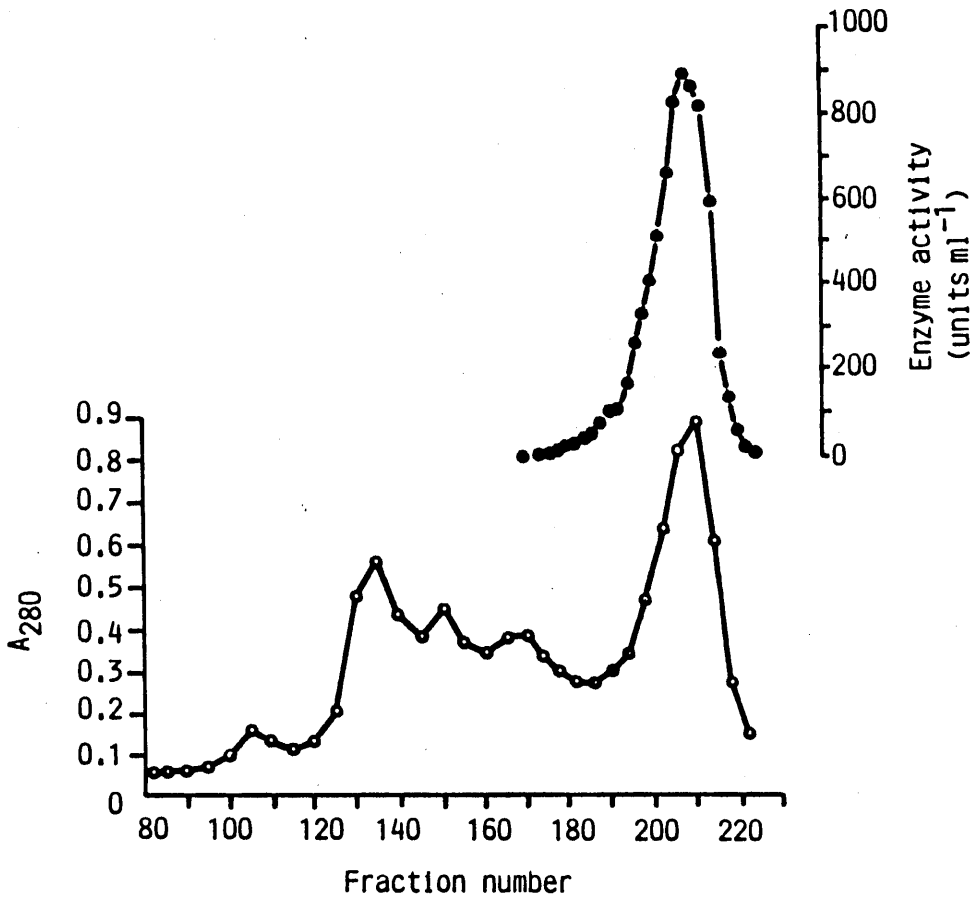
Enzyme from step 2 (29,100 units in 36ml) was loaded onto a column of DEAE-Sephacel (16.5 cm x 2.1 cm) equilibrated in buffer 2, as described in Chapter Two. The column was washed with this buffer until the A_{280} of the eluate reached a constant value of about 0.2. The column was then eluted with an 800ml linear gradient of KCl (50-350mM) in buffer 2. The flow rate was 36 ml/h and 5.5ml fractions were collected. O, A_{280}; ● , E3 activity (units/ml); -----, conductivity (mmho).


A_{280}
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
10
30
50
70
90
110
130
150
Fraction number
0
5
10
15
Conductivity
(mmho)
0
200
400
600
800
Enzyme activity
(units ml^{-1})

189 facing

Figure 4.6 Chromatography of overproduced E.coli E3 on Sephacryl S-200 (step 4 of the purification scheme).

Enzyme from step 3 (17,800 units in 33 ml) was concentrated (to 2.7 ml) and applied to a column of Sephacryl S-200 (85 cm x 2.1 cm) equilibrated in buffer 5 (see Chapter Two). The enzyme was eluted with the same buffer. The flow rate was 4 ml/h and 1 ml fractions were collected. O, A_{280}; ●, E3 activity (units/ml).


1000
800
600
400
200
0
Enzyme activity
(units ml^{-1})
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A_{280}
80
100
120
140
160
180
200
220
Fraction number

is therefore slightly lower than it would otherwise have been.

The fourth step was gel filtration on Sephacryl S-200 (Figure 4.6). The E3 activity elutes late in a peak of activity which corresponds clearly to a large peak of protein. This step in the purification is particularly effective because the enzyme is monomeric (Chaudhuri and Coggins, 1985). Fractions 200-215 inclusive, those having more than 400 units/ml, were pooled. This pool contained 10,400 units (71% of the 14,800 units loaded) at a specific activity of 638 units/mg.

The fifth and final step was affinity chromatography on ADP-Sepharose (elution profile not shown). Due to the limited capacity of this chromatographic medium a 20ml bed volume was used rather than the 5 ml specified in Chaudhuri and Coggins (1985). The pool from the fourth step was loaded onto the column which was then washed to remove unbound material. The E3 activity was then eluted from the column with 1mM $NADP^+$ in buffer 6 (see Chapter Two). The final pool of activity contained 9,040 units at a specific activity of 860 units/mg. The total amount of protein was 10.5mg. This material was homogeneous by the criterion of SDS PAGE, as shown in Figure 4.7 - in the original gel a barely detectable contaminant band (of greater mobility than E3) was visible in the lane loaded with 25µg of E3. Polyacrylamide gel electrophoresis under native conditions, after incubation of the sample in 50mM DTT at 0^oC for 1 h prior to loading, revealed two barely detectable minor bands when 5µg of E3 was loaded (gel not shown).

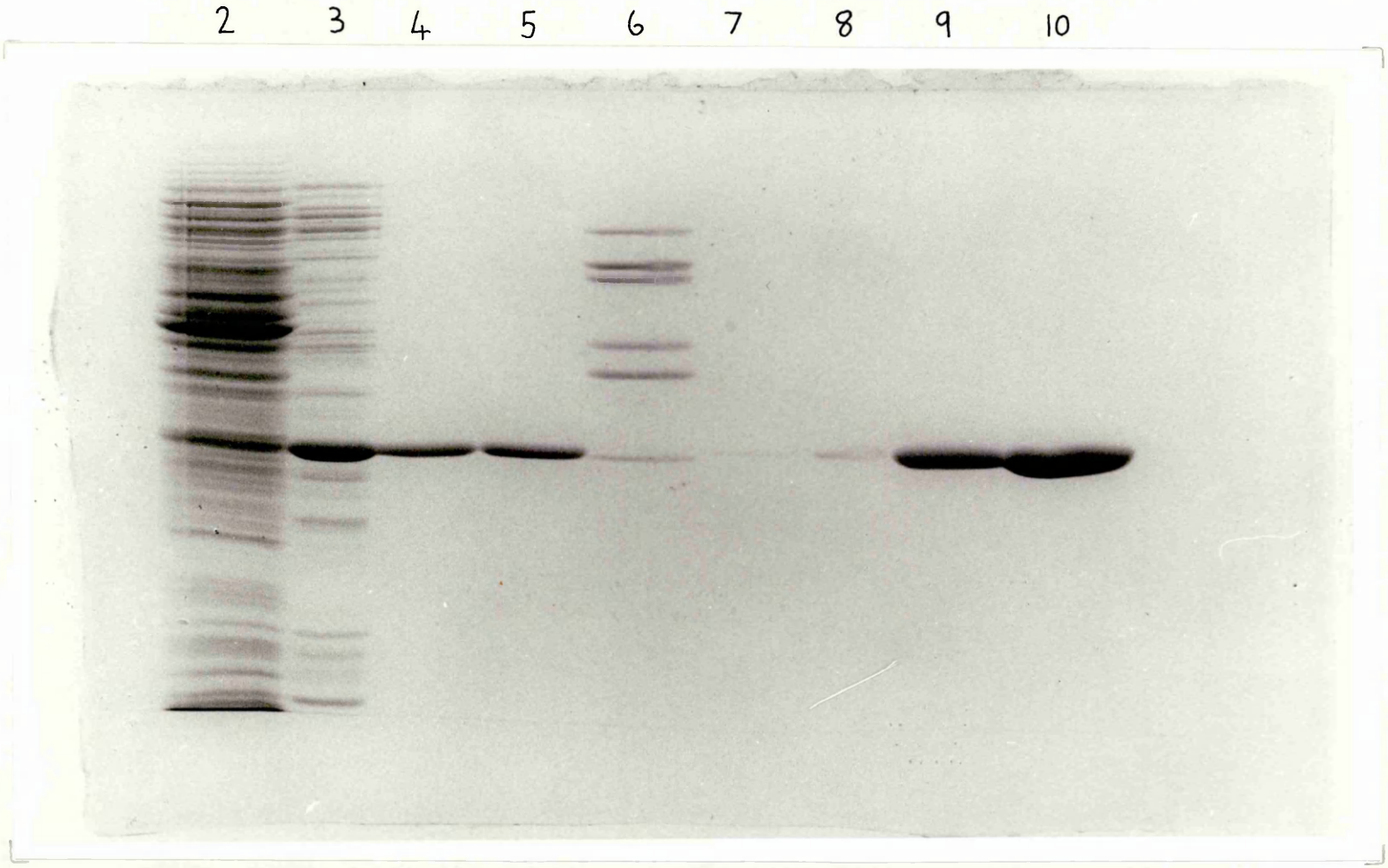
Facing

Figure 4.7 The purification of overproduced E.coli E3.

Samples were subjected to SDS PAGE in a 12.5% gel. After electrophoresis the gel was stained for protein.

Lane: 2. (N.B. not from the purification) sample of a small scale crude extract of pIA321//AB2834 (grown to A_{650} = 4.1), 100 µg total protein in sample.
3. post-DEAE-Sephacel pool, 15 µg total protein
4. post-Sephacryl S-200 pool, 6.7 µg total protein
5. post-ADP-Sepharose final pool, 5 µg
6. molecular weight markers (see Chapter Two; BSA, catalase, GluDH, aldolase, GAPDH, carbonic anhydrase).
7. 0.43 µg pure E.coli K12 E3
8. 1 µg, final pool
9. 10µg, final pool
10. 25µg, final pool

Lanes 3,4, and 5 were loaded with equal amounts of E3 activity. Pure E.coli K12 E3 was the generous gift of S. Chaudhuri.


2
3
4
5
6
7
8
9
10

Overall, the overproduced E3 had to be purified about 43-fold to give homogeneous material. This compares with 20,000-fold for *E.coli* K12 E3 (*op.cit.*). The final yield of overproduced E3 was 29%. One may speculate that it might be possible to obtain homogeneous overproduced E3 using only the first four steps by sacrificing yield. This could be useful for obtaining gram quantities of E3 for crystallisation experiments since otherwise the cost of the ADP-Sepharose might be limiting.

4.3.4 Specific activity of overproduced E.coli E3

The overproduced E3 appears to be effectively homogeneous yet its specific activity of 860 units/mg is significantly lower than the value of 1200 units/mg (see footnotes to Table 4.2) obtained by Chaudhuri and Coggins (*op.cit.*) for the specific activity of pure *E.coli* K12 E3. In both cases protein concentrations were estimated by the method of Bradford (1976) using bovine serum albumin as the standard. However, the extremely small amounts of K12 E3 required the use of the "micro" assay rather than the standard assay, and discouraged repetitions. These difficulties, and the fact that traces of glassware detergent can cause anomalously high A_{595} values with the Bradford system, might be responsible for the higher K12 value (a trace of detergent in the buffer control "blank" tube could generate an artefactually low protein concentration hence higher specific activity). The specific activity of the overproduced E3 was calculated using the average of

four activity determinations and the average of two protein determinations done with two different buffer controls; volumes were measured either using Hamilton microsyringes or Gilson Pipetmen whose calibration had been checked by weighing; removal of samples at $4^{o}C$ was done after precooling the Pipetmen.

Clearly such precautions are to no avail if, despite the contrary evidence from two different gel systems, the overproduced E3 is significantly impure. Further evidence is available which argues against contamination. During the later N-terminal sequencing of the overproduced E3 there was no trace of heterogeneity. It could be argued that the putative contaminant(s) did not have an accessible N-terminus but the later amino acid analysis of the over-produced E3 gave a composition which matched very closely the composition predicted from the DNA sequence. All these different strands of evidence make it implausible that the overproduced E3 is sufficiently impure to explain the discrepancy in the specific activity values.

4.3.5 Comparison of purified E3 from E.coli K12 with that from pIA321//AB2834

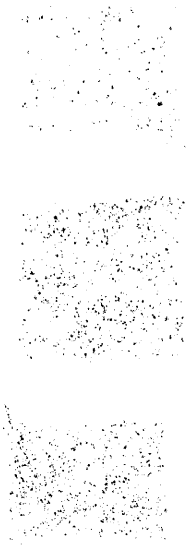
The results obtained from gel analysis of small scale crude extracts (Section 3.4) suggested that there would be no detectable differences between purified E3 from E.coli K12 and that from pIA321//AB2834. Further data tend to confirm this view. At a gross level both display similar

Figure 4.8 Comparison of purified K12 E3 with purified pIA321//AB2834 E3 by nondenaturing PAGE

Samples were subjected to nondenaturing PAGE in a 10% gel after incubation for 1 h at 0°C in the presence of 50mM DTT. The gel was stained for E3 activity.

Lane: 1. 0.4 µg pure E.coli K12 E3
2. 0.4 µg pure pIA321//AB2834 E3

The salt concentration, glycerol concentration, and final volume were the same in both samples. Pure E.coli K12 E3 was the generous gift of S. Chaudhuri.



chromatographic behaviour on DEAE-Sephacel and on Sephacryl S-200 during purification. Lanes 7 and 8 in Figure 4.7 show that the two purified enzymes have identical mobilities during SDS PAGE. In addition, Figure 4.8 shows that they also have identical mobilities during PAGE under non-denaturing conditions.

It would perhaps be worthwhile to compare the over-produced and K12 versions by one-dimensional peptide mapping, and by their kinetic parameters, as in Duncan *et al.* (1984a).

4.4 Determination of the N-terminal amino acid sequence of the overproduced E.coli E3

This work was done in collaboration with J.E. Fothergill, L.A. Fothergill, and B. Dunbar in the SERC funded protein sequencing facility at Aberdeen University. The determination of the N-terminal amino acid sequence was carried out using a Beckman Model 890C automatic liquid phase sequencer. Details of the methodology are given in Chapter Two. The resulting sequence is shown in Figure 4.9. Table 4.3 shows the yield (in nmol) obtained for each residue and Figures 4.10a and 4.10b present the same data graphically. The identification of the first thirty amino acids was unambiguous. Residues 31 and 35 were almost certainly histidine and arginine respectively but there was an element of doubt due to the HPLC retention times being shorter than usual. Residue 36 could not be identified. Residue 40 was definitely isoleucine even though no yield could be calculated.

1 10

Met - Glu - Thr - Tyr - Ala - Val - Phe - Gly - Asn - Pro -

11 20

Ile - Ala - His - Ser - Lys - Ser - Pro - Phe - Ile - His -

21 30

Gln - Gln - Phe - Ala - Gln - Gln - Leu - Asn - Ile - Glu -

31 40

(—) - Pro - Tyr - Gly - (—) - (—) - Leu - Ala - Pro - Ile -
(His) (Arg)

41 50

Asn - (—) - Phe - Ile - Asn - (—) - Leu - Asn - Ala - Phe -
(Asp)

51 60

Phe - (—) - Ala - Gly - Gly - Lys - Gly - Ala - Asn - Val -

Figure 4.9 The N-terminal amino acid sequence of E.coli shikimate dehydrogenase

The repetitive yield from residues 1-60 was 94% by regression analysis (correlation coefficient 0.96) and the initial amount of protein sequencing was 53 nmol. The gaps are discussed in the text.

Table 4.3 Yield obtained of each residue during N-terminal amino acid sequencing of overproduced E.coli E3.

RESIDUE NO.	AMINO ACID	YIELD(nmol)
1	M	30.5
2	E	46.0
3	T	11.5
4	Y	31.5
5	A	34.7
6	V	42.5
7	F	35.1
8	G	31.8
9	N	33.7
10	P	22.8
11	I	31.8
12	A	23.8
13	H	14.5
14	S	4.1
15	K	28.3
16	S	3.9
17	P	18.4
18	F	24.3
19	I	24.1
20	H	10.6
21	Q	17.2
22	Q	25.9
23	F	13.7
24	A	14.1
25	Q	13.9
26	Q	24.6
27	L	18.2
28	N	10.75
29	I	12.8
30	E	12.5
31	—	00
32	P	7.2
33	Y	8.8
34	G	8.4
35	—	00
36	—	00
37	L	11.8
38	A	5.3
39	P	5.9
40	I	00
41	N	7.0
42	—	00
43	F	4.5
44	I	00
45	N	4.2
46	—	00
47	L	3.2
48	N	2.8
49	A	2.7
50	F	5.2
51	F	5.0
52	—	00
53	A	2.6
54	G	2.0
55	G	2.0
56	K	1.2
57	G	0.859
58	A	1.5
59	N	1.1
60	V	1.4

Figure 4.10a Yield obtained of each residue during N-terminal amino acid sequencing of overproduced *E.coli* E3; residues 1-30.

RESIDUES 1 TO 30
GRAPH OF YIELD(NMOLES) V. RESIDUE NO.
MAX.VAL. Y AXIS.....50

M E T Y A V F G N P I A H S K S P F I H Q Q F A Q Q L N I E

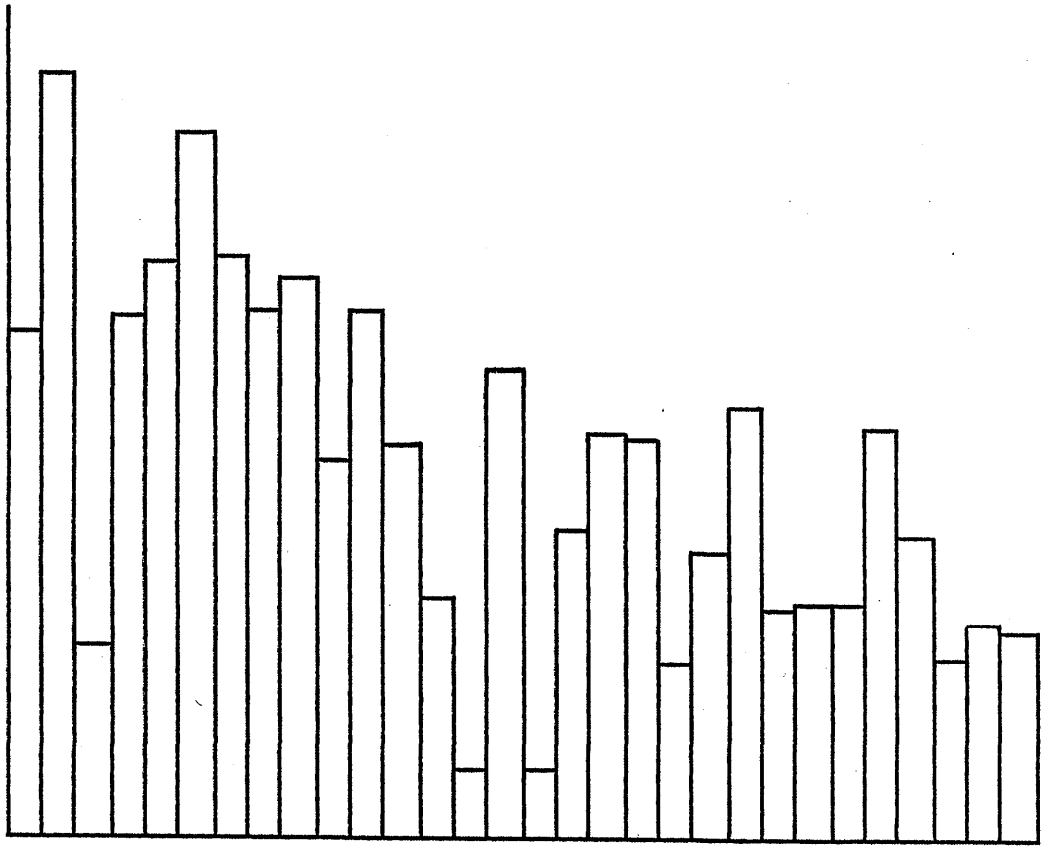
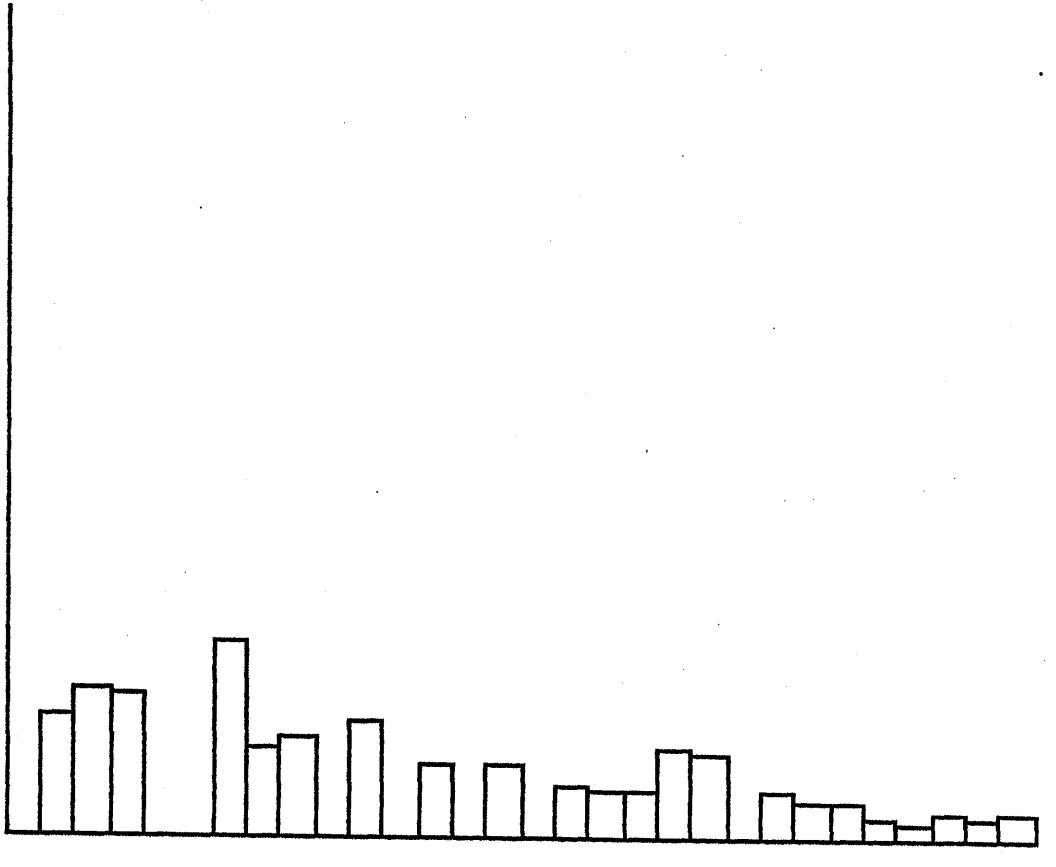


Figure 4.10b Yield obtained of each residue during N-terminal amino acid sequencing of overproduced *E.coli* E3; residues 31-60.

RESIDUES 31 TO 60
GRAPH OF YIELD(NMOLES) V. RESIDUE NO.
MAX.VAL. Y AXIS.....50

P Y G L A P I N F I N L N A F F A G G K G A N V



Residue 42 was probably aspartic acid but the preceding asparagine residue made it impossible to rule out another asparagine residue at position 42. Residue 44 was undoubtedly isoleucine even though no yield could be calculated. Residues 46 and 52 could not be identified.

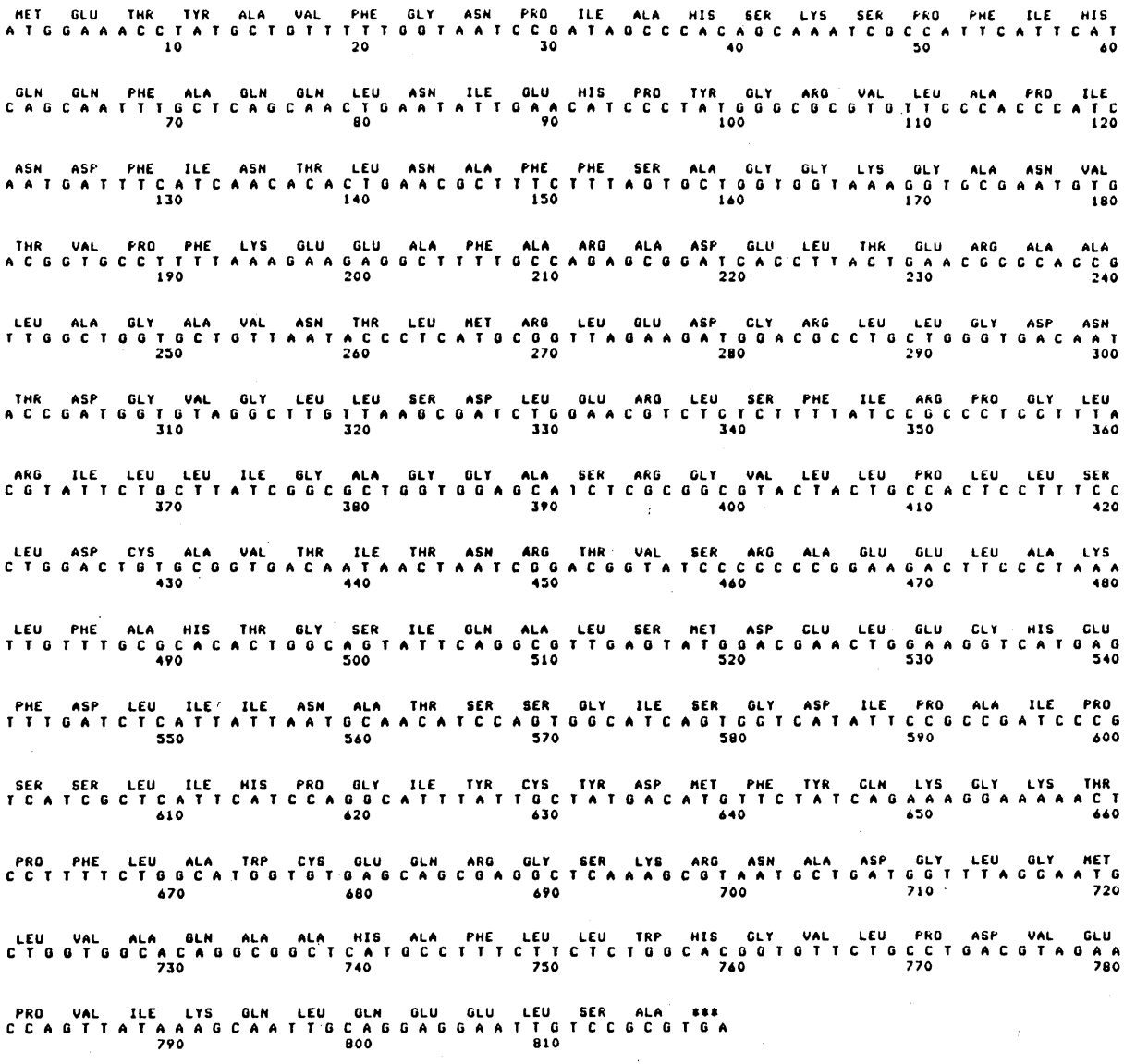
4.5 Definitive identification of the E.coli shikimate dehydrogenase coding sequence

The N-terminal amino acid sequence of overproduced E.coli E3 was compared with the amino acid sequence predicted for the longest open reading frame in the DNA sequence of the 1.82 kbp HindIII-ClaI fragment (see Sections 4.1.1 and 4.1.2), starting from the first methionine residue in the ORF. A perfect match was found (compare Figures 4.9 and 4.11), thus definitively locating the coding sequence for shikimate dehydrogenase. Figure 4.11 shows the complete nucleotide sequence of the E.coli aroE gene and the corresponding amino acid sequence of E.coli shikimate dehydrogenase.

From the DNA sequence the residues which could not be identified during the N-terminal amino acid sequencing - residues 36, 46, and 52 - are valine, threonine, and serine respectively. Failure to detect threonine and serine after many cycles is not surprising since the anilinothiazolinone derivatives of threonine and serine are relatively unstable. However, it remains unclear why valine was not detected at

Figure 4.11 The complete nucleotide sequence of the *E.coli* *aroE* gene and the corresponding amino acid sequence of *E.coli* shikimate dehydrogenase.

Nucleotides are numbered in the 5' to 3' direction beginning with the first base of the ATG triplet encoding the N-terminal methionine.


MET GLU THR TYR ALA VAL PHE GLY ASN PRO ILE ALA HIS SER LYS SER PRO PHE ILE HIS
ATGGAAACCTATGCTGTTTTTGGTAATCCGATAGCCCACAGCAAATCGCCATTCATTCAT
10 20 30 40 50 60
GLN GLN PHE ALA GLN GLN LEU ASN ILE GLU HIS PRO TYR GLY ARG VAL LEU ALA PRO ILE
CAGCAATTTGCTCAGCAACTGAATATTGAACATCCCTATGGGCGCGTGTTCCCACCCATC
70 80 90 100 110 120
ASN ASP PHE ILE ASN THR LEU ASN ALA PHE PHE SER ALA GLY GLY LYS GLY ALA ASN VAL
AATGATTTCATCAACACACTGAACGCTTTCTTTAGTGCTGGTGGTAAAGGTGCGAATGTG
130 140 150 160 170 180
THR VAL PRO PHE LYS GLU GLU ALA PHE ALA ARG ALA ASP GLU LEU THR GLU ARG ALA ALA
ACGGTGCCTTTTAAAGAAGAGGCTTTTGCCAGAGCGGATCACCTTACTGAACGCGCCACCG
190 200 210 220 230 240
LEU ALA GLY ALA VAL ASN THR LEU MET ARG LEU GLU ASP GLY ARG LEU LEU GLY ASP ASN
TTGGCTGGTGCTGTTAATACCCTCATGCGGTTAGAAGATGGACGCCTGCTGGGTGACAAT
250 260 270 280 290 300
THR ASP GLY VAL GLY LEU LEU SER ASP LEU GLU ARG LEU SER PHE ILE ARG PRO GLY LEU
ACCGATGGTGTAGGCTTGTTAAGCGATCTGGAACGTCTCTCTTTTATCCGCCCTCCTTTA
310 320 330 340 350 360
ARG ILE LEU LEU ILE GLY ALA GLY GLY ALA SER ARG GLY VAL LEU LEU PRO LEU LEU SER
CGTATTCTGCTTATCGGCGCTGGTGGAGCATCTCGCGGCGTACTACTGCCACTCCTTTCC
370 380 390 400 410 420
LEU ASP CYS ALA VAL THR ILE THR ASN ARG THR VAL SER ARG ALA GLU GLU LEU ALA LYS
CTGGACTGTGCGGTGACAATAACTAATCGGACGGTATCCCGCCGCGAAGACTTCCCTAAA
430 440 450 460 470 480
LEU PHE ALA HIS THR GLY SER ILE GLN ALA LEU SER MET ASP GLU LEU GLU GLY HIS GLU
TTGTTTGCGCACACTGGCAGTATTCAGGCGTTGAGTATGGACGAACTGGAAGGTCATGAG
490 500 510 520 530 540
PHE ASP LEU ILE ILE ASN ALA THR SER SER GLY ILE SER GLY ASP ILE PRO ALA ILE PRO
TTTGATCTCATTATTAATGCAACATCCAGTGGCATCAGTGGTCATATTCCGGCCGATCCCG
550 560 570 580 590 600
SER SER LEU ILE HIS PRO GLY ILE TYR CYS TYR ASP MET PHE TYR GLN LYS GLY LYS THR
TCATCGCTCATTCATCCAGGCATTTATTGCTATGACATGTTCTATCAGAAAGGAAAAACT
610 620 630 640 650 660
PRO PHE LEU ALA TRP CYS GLU GLN ARG GLY SER LYS ARG ASN ALA ASP GLY LEU GLY MET
CCTTTTCTGGCATGGTGTGAGCAGCGAGGCTCAAAGCGTAATGCTGATGGTTTACGAATG
670 680 690 700 710 720
LEU VAL ALA GLN ALA ALA HIS ALA PHE LEU LEU TRP HIS GLY VAL LEU PRO ASP VAL GLU
CTGGTGGCACAGGCGGCTCATGCCTTTCTTCTCTGGCACGGTGTTCTGCCTGACGTAGAA
730 740 750 760 770 780
PRO VAL ILE LYS GLN LEU GLN GLU GLU LEU SER ALA ***
CCAGTTATAAAGCAATTGCAGGAGGAATTGTCCGCGTGA
790 800 810

at residue 36. The identities of amino acid residues 31, 35, and 42, which had been slightly in doubt after the protein sequencing, were confirmed by the DNA sequence.

The relationship of the E3 coding sequence to the restriction map of the 1.82 kbp fragment may be seen in Figure 3.19.

4.6 Determination of the amino acid composition of the overproduced E.coli E3

As a final check on the complete aroE sequence the amino acid composition of overproduced E.coli E3 was determined so that it could be compared with the amino acid composition predicted by the DNA sequence.

Purified E.coli E3 was oxidised by performic acid and samples were analysed after hydrolysis with 6M-HCl at 105°C for 24, 48, and 72 h, as described in Chapter Two. The amino acid composition obtained is shown in Table 4.4 alongside the predicted composition. Tryptophan is destroyed by the procedure used. Since asparagine and glutamine are hydrolysed quantitatively to aspartic and glutamic acid respectively the abbreviations Asx and Glx have been used in Table 4.4 to indicate the ambiguous origins of the aspartic and glutamic acid residues. Some serine and threonine is destroyed in a time dependent manner during hydrolysis so for these two amino acids the data from the three timepoints were extrapolated to zero time.

Table 4.4 The amino acid composition of overproduced E.coli E3 compared with the amino acid composition predicted from the E.coli aroE gene sequence

Amino acid	Relative amino acid composition based on Lys = 8 residues	Amino acid composition predicted from the DNA sequence
Cys[a]	3.00	3
Asx	23.5	24
Met[b]	5.22	5
Thr[c]	12.3	12
Ser[c]	17.4	17
Glx	27.9	27
Pro	12.8	13
Gly	26.1	25
Ala	30.8	30
Val[d]	13.3	13
Ile[d]	17.6	18
Leu	34.5	35
Trp	nd	2
Tyr	5.06	5
Phe	14.5	14
His	7.91	8
Lys	(8)	8
Arg	12.9	13

[a] Determined as cysteic acid

[b] Determined as methionine sulphone

[c] Experimental values were extrapolated to zero time

[d] Values were calculated using only data from the 72 h timepoint.

Valine and isoleucine are often released more slowly than other amino acids due to steric hindrance of hydrolysis by the β-branched sidechains. A plot of the data for the three timepoints showed that the amounts of valine and isoleucine had reached a plateau by 72 h, hence only data from the 72 h timepoint were used. Performic acid oxidation of cysteine and methionine, to cysteic acid and methionine sulphone respectively, allowed quantitative determination of these two amino acids.

The relative amino acid composition was calculated using the assumption that there are 8 lysine residues in the polypeptide. (Each of glycine, alanine, leucine, lysine and arginine was tried - using the number of residues predicted by the DNA sequence - and lysine gave the closest fit to the predicted total number of amino acids - 272). The overall agreement between the observed and predicted values in Table 4.4 is good. It is thus very unlikely that there is a frameshift error anywhere in the *aroE* sequence.

4.7 Conclusion

The match between the DNA sequence of the longest ORF in the 1.82 kbp HindIII-ClaI fragment and the N-terminal amino acid sequence of purified *E.coli* E3 formally locates *aroE*, the gene for shikimate dehydrogenase. The accuracy of the *aroE* sequence is confirmed by the finding that the amino acid composition of purified *E.coli* E3 fits very

closely with that predicted from the DNA sequence.

The location of the aroE gene so far from the ClaI site reopened the issue of the wide variation in E3 specific activities in strains carrying different aroE subclones. Whether the sequence data might shed light on this and other questions is considered in the next chapter.

CHAPTER 5 ANALYSIS OF THE SEQUENCE DATA

5.1 Subunit molecular weight of E.coli E3 and the "2 bands" problem

The subunit molecular weight for E3 predicted by the sequence is 29,380 daltons. This agrees well with the size of 30 kDa obtained by SDS PAGE of the purified enzyme (Chaudhuri and Coggins, 1985; this study).

In Section 3.4.4 it was shown that E3 activity could be recovered after SDS PAGE. However, SDS PAGE of *E.coli* crude cell extracts unexpectedly gave two bands with E3 activity: one of about 29 kDa, which from the sequence data can now be said to represent uncorrupted E3, and another of about 49 kDa. The reproducible features of these two bands can be summarised as follows:-

(i) both are shikimate-dependent

(ii) neither is detectable in *E.coli* AB2834

(iii) both are present in pIA301//AB2834 extracts

(iv) both are increased in intensity in extracts of strains overproducing E3.

These observations strongly suggest that both bands are dependent for their existence on the region of the genome represented by the 1.82 kbp HindIII-ClaI fragment. There is no need, therefore, to invoke the idea of some other dehydrogenase nonspecifically catalysing the E3 reaction. The sequence data for the 1.82 kbp fragment show no ORF capable of encoding a 49 kDa polypeptide.

It seems likely, as mentioned in Chapter Three, that the species giving rise to the 49 kDa band contains, in some way, the E3 polypeptide. This matter was not pursued further experimentally so the origin of the 49 kDa band remains unclear. Application of the renaturation protocol to an SDS Laemmli gel of homogeneous K12 E3 gives a single 30 kDa activity band (S. Chaudhuri, unpublished results).

In Chapter Three it was hypothesised that the 49 kDa band might be an artefact due to the formation of refractory disulphide linkages between E3 and some other protein. One can speculate further in this direction though the ratio of data to ideas is unfavourable. The existence of only one putatively artefactual band need not imply that there would have to be only one particular polypeptide to which E3 crosslinks since possibly only one partnership out of many would survive the sample preparation procedure used for SDS PAGE. However, it is perhaps more plausible that the hypothetical crosslinking would occur to some species with which the E3 polypeptide was already associated, conceivably another E3 polypeptide. E3 purified to homogeneity from E.coli K12 is a monomer (Chaudhuri and Coggins, 1985) but it has been noted (Fulton, 1982) that the protein concentrations found in cytoplasm are so high that non-ideal effects become very important. Thus observations made at low protein concentrations may not accurately reflect the in vivo situation. The 49 kDa molecular weight of the putatively crosslinked species is not at variance with the idea of a crosslinked E3 dimer since the apparent m.w. would depend on the positions of the crosslinked groups

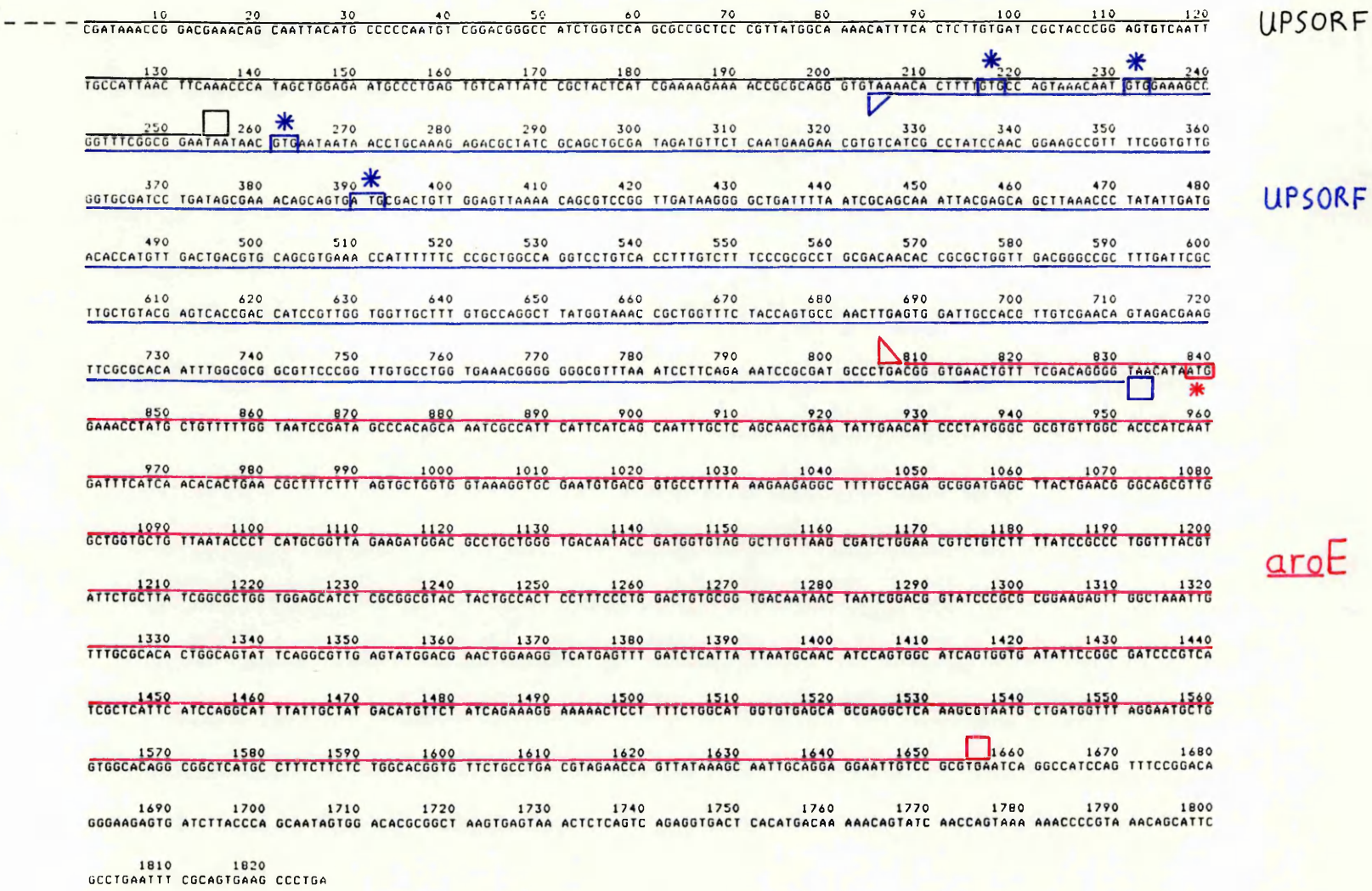
within the polypeptide chains.

The sequence data suggest an alternative hypothesis to explain the 49 kDa band. Examination of the open reading frames in the 1.82 kbp HindIII-ClaI sequence (see Figures 4.1, 4.2, and 5.1) shows that upstream of the aroE gene and in the same orientation there is a substantial ORF (624 bases between stop codons) whose termination codon is only 4 bases before the initiation codon of aroE. This ORF immediately upstream of aroE will be referred to as UPSORF 1. It is possible, as discussed later, that UPSORF 1 is a functional gene and that it is part of an operon which includes the aroE gene. If this is the case, and there is as yet no direct evidence to support it, then the 49 kDa band could arise from ribosomes occasionally frameshifting during translation of the C-terminal part of UPSORF 1 thus resulting in readthrough into the aroE gene and the production of a fusion protein. Precedents for this situation are known (Grosjean and Fiers, 1982; Atkins et al., 1979). Perhaps some feature of the sequence at the end of UPSORF 1 predisposes ribosomes to shift frame sometimes rather than to terminate translation.

Various experimental approaches could be used to pursue this problem. The readthrough hypothesis would predict that an extract of pIA309//AB2834 should not give a 49 kDa band since in pIA309 UPSORF 1 is drastically truncated (see Figure 4.2). Also, it should be possible to lyse cells directly in SDS PAGE sample dissolving buffer thus making oxidation artefacts unlikely.

Figure 5.1 Large ORF's upstream of aroE in the 1.82 kbp HindIII-ClaI sequence

The sequence of one strand of the HindIII-ClaI fragment is shown. It is written 5' to 3' starting from the ClaI end. The stop codon which precedes an ORF is marked by a triangle whereas the stop codon which ends an ORF is shown by a square. The extent of the aroE ORF is indicated by the red line, and the known initiation codon is boxed and asterisked. The extent of UPSORF 1 is indicated by the blue line and some of the hypothetical initiation codons are boxed and asterisked. UPSORF 2 (see text) straddles the ClaI site and is marked by the black line.


UPSORF 2
UPSORF 1
aroE

The possibility that UPSORF 1 is a gene, in an operon containing aroE, is relevant to the discussion of another problem: the wide variation in E3 specific activities obtained for strains carrying different subclones of aroE (Sections 3.3.5 and 3.3.6). This problem will be discussed later. In the next sections preliminary sequence data are presented for the 0.6 kbp ClaI-BamHI region (see Figure 3.14), thus extending the available sequence further upstream from aroE. This new information permits the identification of another large ORF whose amino acid sequence has interesting features.

5.2 Locations of large open reading frames upstream of aroE

5.2.1 Large ORF's in the 1.82 kbp HindIII-ClaI sequence

The stop codons in all six reading frames of the HindIII-ClaI sequence were shown in Figure 4.1. The precise position and extent within the sequence of each of the two largest ORF's (aroE and UPSORF 1) is shown in Figure 5.1. Upstream from, and slightly overlapping, UPSORF 1 is a segment of ORF which continues uninterrupted to the ClaI site: this segment was found to join in frame with a segment of ORF in the sequence of the ClaI-BamHI region (see below) giving a large ORF which will be referred to as UPSORF 2.

5.2.2 Sequence of the 0.6 kbp ClaI-BamHI fragment on one strand

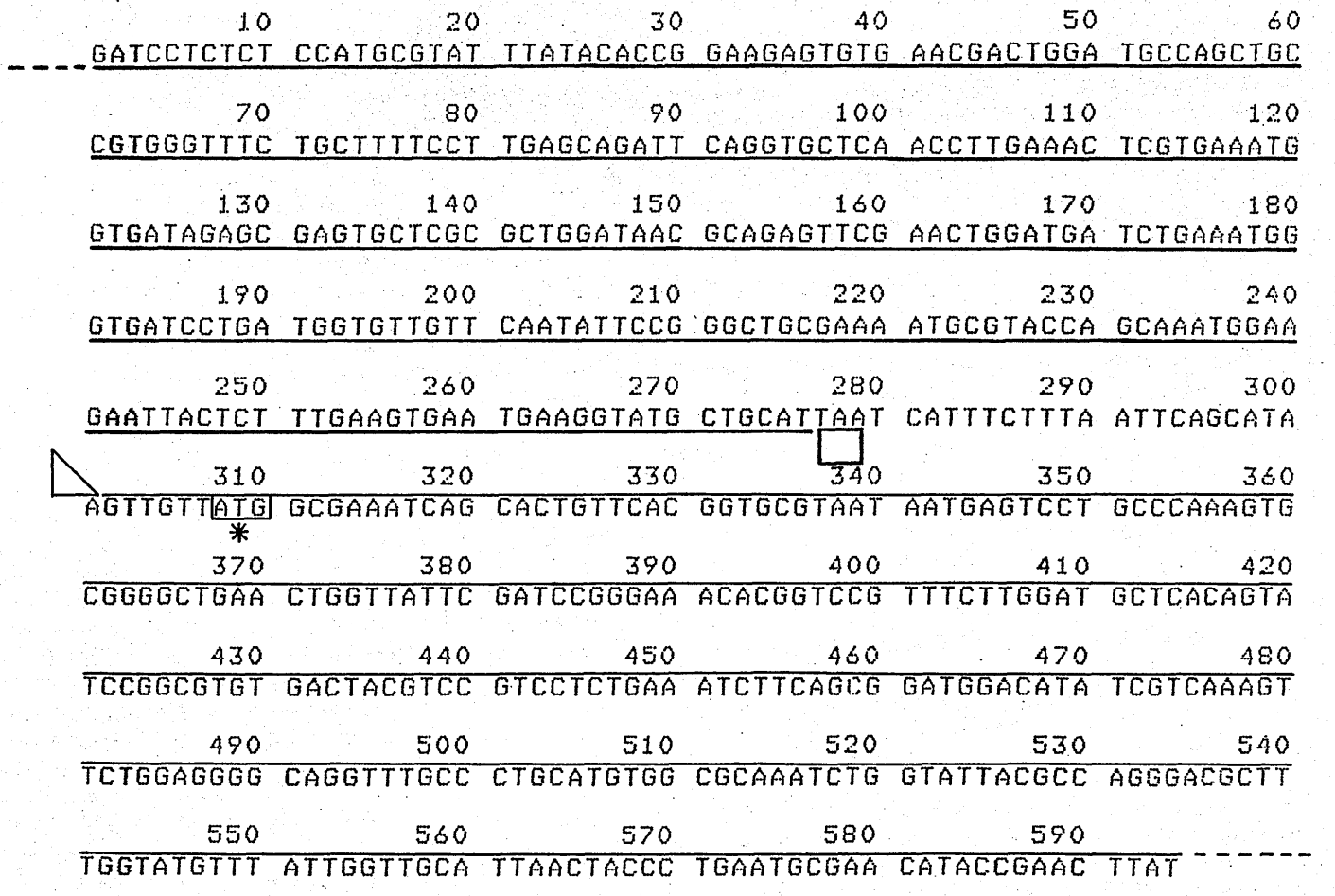
The HindIII-ClaI fragment was sequenced on both strands. The abutting ClaI-BamHI piece was sequenced mainly on one strand during the first round of M13 sequencing (see Figure 3.14 and Table 3.12) by sequencing in from either end of the fragment. The use of altered ratios of dideoxy- to deoxynucleotides and extended periods of electrophoresis were required to obtain runs of sequence long enough to overlap. An overlap of 49 base-pairs was obtained. The sequence is shown in Figure 5.2. It must be emphasised that this sequence should be regarded as preliminary since being mainly on one strand it may harbor insidious errors, as discussed in Section 3.5.1C. However, it is worth noting that the overlapping sequences agree perfectly and also that, very unusually, there are no ambiguities or doubtful regions in the two runs of single strand sequence.

The overlap across the ClaI site was obtained by extended reading of the 13-2 template (see Figure 3.14 and Table 3.12). This permitted the tentative ClaI-BamHI sequence to be fused with the ClaI-HindIII sequence thus extending the distance upstream from *aroE* for which DNA sequence was available.

The stop codons in all six reading frames of the ClaI-BamHI sequence were plotted as an extension to Figure 4.1. This enlarged ORF diagram for the whole HindIII-BamHI region is shown in Figure 5.3. As mentioned above in Section 5.2.1

Figure 5.2 Preliminary sequence of the 0.6 kbp ClaI-BamHI fragment and the locations of large ORF's within the sequence

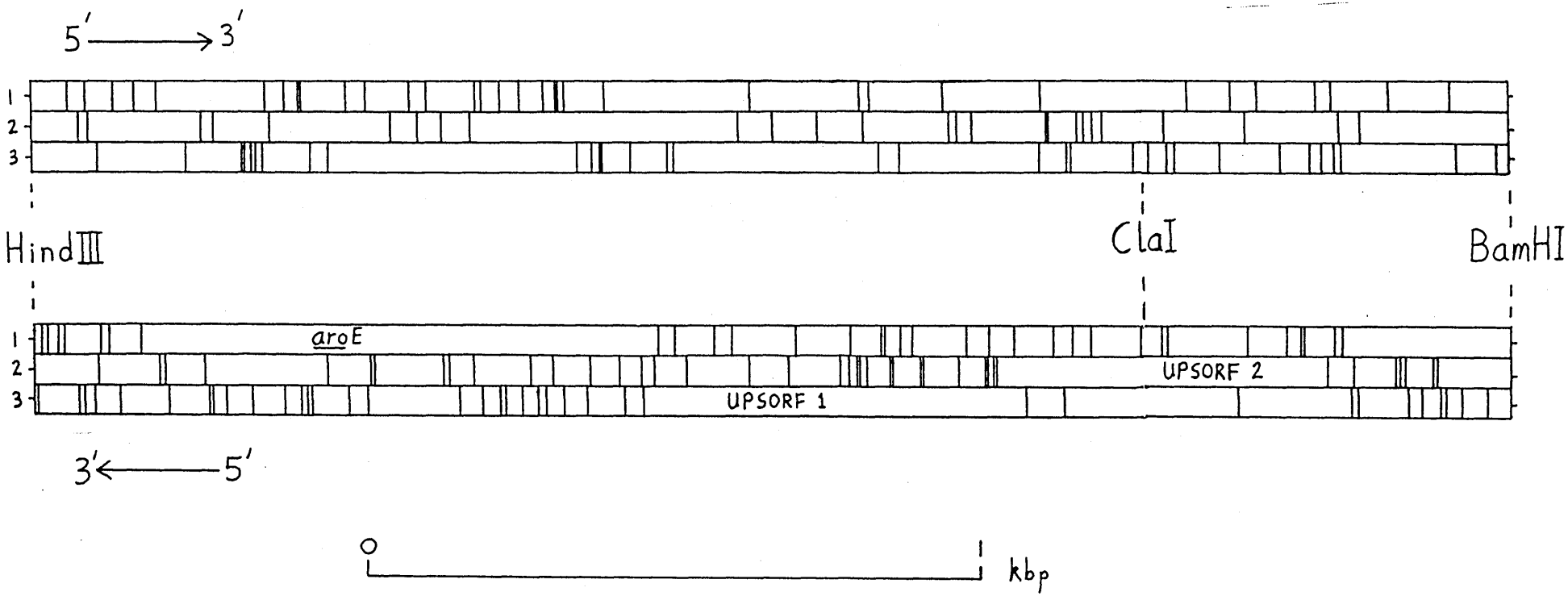
The preliminary sequence of one strand of the ClaI-BamHI fragment is shown. It is written 5' to 3' starting from the BamHI end. The stop codon which precedes an ORF is marked by a triangle whereas the stop codon which ends an ORF is shown by a square. The extent of UPSORF 2, which continues past the ClaI site (see Figure 5.1), is indicated by the black line and the first hypothetical initiation codon is boxed and asterisked. A segment of ORF which continues uninterrupted to the BamHI site is shown by the blue line.


10 20 30 40 50 60
GATCCTCTCT CCATGCGTAT TTATACACCG GAAGAGTGTG AACGACTGGA TGCCAGCTGC
70 80 90 100 110 120
CGTGGGTTTC TGCTTTTCCT TGAGCAGATT CAGGTGCTCA ACCTTGAAAC TCGTGAAATG
130 140 150 160 170 180
GTGATAGAGC GAGTGCTCGC GCTGGATAAC GCAGAGTTCG AACTGGATGA TCTGAAATGG
190 200 210 220 230 240
GTGATCCTGA TGGTGTTGTT CAATATTCCG GGCTGCGAAA ATGCGTACCA GCAAATGGAA
250 260 270 280 290 300
GAATTACTCT TTGAAGTGAA TGAAGGTATG CTGCATTAAT CATTTCTTTA ATTCAGCATA
310 320 330 340 350 360
AGTTGTTATG GCGAAAATCAG CACTGTTCAC GGTGCGTAAT AATGAGTCCT GCCCAAAGTG
370 380 390 400 410 420
CGGGGCTGAA CTGGTTATTC GATCCGGGAA ACACGGTCCG TTTCTTGGAT GCTCACAGTA
430 440 450 460 470 480
TCCGGCGTGT GACTACGTCC GTCCTCTGAA ATCTTCAGCG GATGGACATA TCGTCAAAGT
490 500 510 520 530 540
TCTGGAGGGG CAGGTTTGCC CTGCATGTGG CGCAAATCTG GTATTACGCC AGGGACGCTT
550 560 570 580 590
TGGTATGTTT ATTGGTTGCA TTAACTACCC TGAATGCGAA CATACCGAAC TTAT

UPSORF 2

Figure 5.3 Open reading frames in the HindIII-BamHI sequence

The sequence of each strand was translated in all three possible reading frames using the computer program TRNTRP. Unadorned vertical lines indicate the positions of stop codons. The stop codons in the ClaI-BamHI region are based on preliminary sequence data.


5′ ⟶ 3′
1
2
3
HindIII
ClaI
BamHI
aroE
UPSORF 2
UPSORF 1
3′ ⟵ 5′
kbp

there is a large open reading frame, UPSORF 2, which straddles the ClaI site.

5.2.3 The amino acid sequences of UPSORF 1 and UPSORF 2

Table 5.1 gives the rough sizes of the polypeptides which would be expected if UPSORF 1 and UPSORF 2 are expressed genes. The amino acid sequences predicted from the DNA sequences are shown in Figures 5.4 and 5.6. Figure 5.5 sketches the arrangement of large ORF's and some putative initiation codons in the HindIII-BamHI region.

5.3 Are UPSORF 1 and UPSORF 2 genes?

It is not known whether UPSORF's 1 and 2 are expressed as polypeptides. Further work is required to test this idea, however, considerable indirect evidence points to UPSORF 1 and UPSORF 2 being genes. This evidence is outlined below, prior to detailed discussion in the following sections:

1. The codon utilisation within the two ORF's shows the bias found in genes.
2. For UPSORF 2, the striking pattern in the deduced amino acid sequence makes it very likely that this ORF encodes a protein.
3. The E3 specific activities in strains carrying different *aroE* subclones suggest that the level of *aroE* expression is affected by sequences far upstream from *aroE*. One plausible explanation is that UPSORF's 1 and 2 are genes in an operon which includes *aroE*.

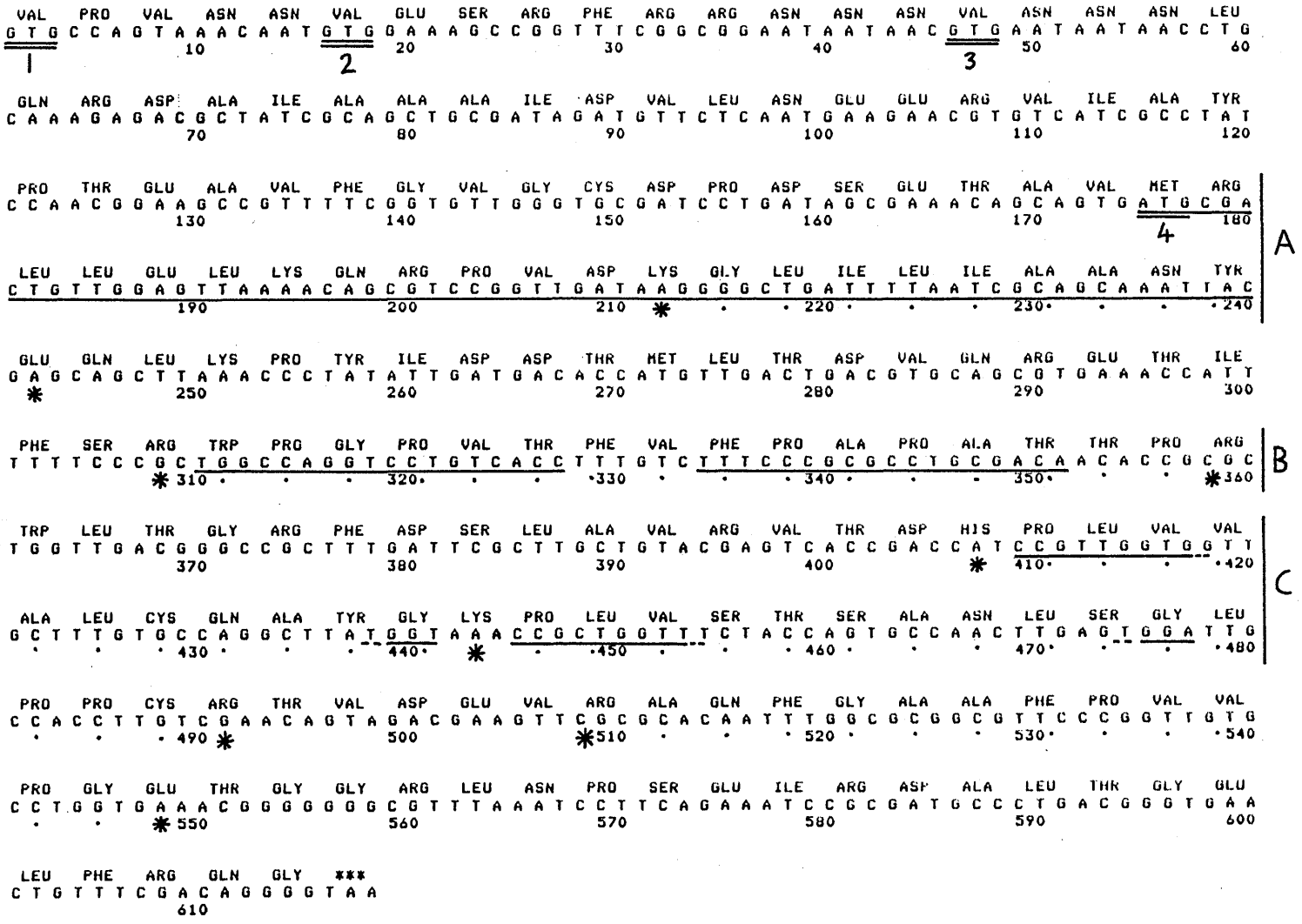
Table 5.1 Polypeptide sizes predicted from the sequences of UPSORF 1 and UPSORF 2

	Number of amino acids	Estimated m.w.(kDa)
(a) UPSORF 1		
Initiation codon: 1st GTG	205	23
2nd "	200	22
3rd "	190	21
1st ATG	147	16
(b) UPSORF 2		
Initiation codon: 1st ATG	180	20

NOTE: m.w.'s estimated on the assumption of an average m.w. per amino acid residue of 110 Da.

Figure 5.4 Amino acid sequence of UPSORF 1

The DNA sequence of UPSORF 1 was translated, starting from the first GTG codon. Four possible initiation codons are labelled 1-4. Other features of possible significance (to be discussed in Section 5.6.1) are also marked. Long runs of neutral/hydrophobic amino acids are indicated by dots and the flanking charged amino acids are asterisked.


VAL PRO VAL ASN ASN VAL GLU SER ARG PHE ARG ARG ASN ASN ASN VAL ASN ASN ASN LEU
GTG CCA GTA AAC AAT GTG GAA AGC CGG TTT CGG CGG AAT AAT AAC GTG AAT AAT AAC CTG
1 2 3
GLN ARG ASP ALA ILE ALA ALA ALA ILE ASP VAL LEU ASN GLU GLU ARG VAL ILE ALA TYR
CAA AGA GAC GCT ATC GCA GCT GCG ATA GAT GTT CTC AAT GAA GAA CGT GTC ATC GCC TAT
PRO THR GLU ALA VAL PHE GLY VAL GLY CYS ASP PRO ASP SER GLU THR ALA VAL MET ARG
CCA ACG GAA GCC GTT TTC GGT GTT GGG TGC GAT CCT GAT AGC GAA ACA GCA GTG ATG CGA
4
LEU LEU GLU LEU LYS GLN ARG PRO VAL ASP LYS GLY LEU ILE LEU ILE ALA ALA ASN TYR
CTG TTG GAG TTA AAA CAG CGT CCG GTT GAT AAG GGG CTG ATT TTA ATC GCA GCA AAT TAC
A
GLU GLN LEU LYS PRO TYR ILE ASP ASP THR MET LEU THR ASP VAL GLN ARG GLU THR ILE
GAG CAG CTT AAA CCC TAT ATT GAT GAC ACC ATG TTG ACT GAC GTG CAG CGT GAA ACC ATT
PHE SER ARG TRP PRO GLY PRO VAL THR PHE VAL PHE PRO ALA PRO ALA THR THR PRO ARG
TTT TCC CGC TGG CCA GGT CCT GTC ACC TTT GTC TTT CCG GCG CCT GCG ACA ACA CCG CGC
B
TRP LEU THR GLY ARG PHE ASP SER LEU ALA VAL ARG VAL THR ASP HIS PRO LEU VAL VAL
TGG TTG ACG GGC CGC TTT GAT TCG CTT GCT GTA CGA GTC ACC GAC CAT CCG TTG GTG GTT
ALA LEU CYS GLN ALA TYR GLY LYS PRO LEU VAL SER THR SER ALA ASN LEU SER GLY LEU
GCT TTG TGC CAG GCT TAT GGT AAA CCG CTG GTT TCT ACC AGT GCC AAC TTG AGT GGA TTG
C
PRO PRO CYS ARG THR VAL ASP GLU VAL ARG ALA GLN PHE GLY ALA ALA PHE PRO VAL VAL
CCA CCT TGT CGA ACA GTA GAC GAA GTT CGC GCA CAA TTT GGC GCG GCG TTC CCG GTT GTG
PRO GLY GLU THR GLY GLY ARG LEU ASN PRO SER GLU ILE ARG ASP ALA LEU THR GLY GLU
CCT GGT GAA ACG GGG GGG CGT TTA AAT CCT TCA GAA ATC CGC GAT GCC CTG ACG GGG GAA
LEU PHE ARG GLN GLY ***
CTG TTT CGA CAG GGG TAA
610

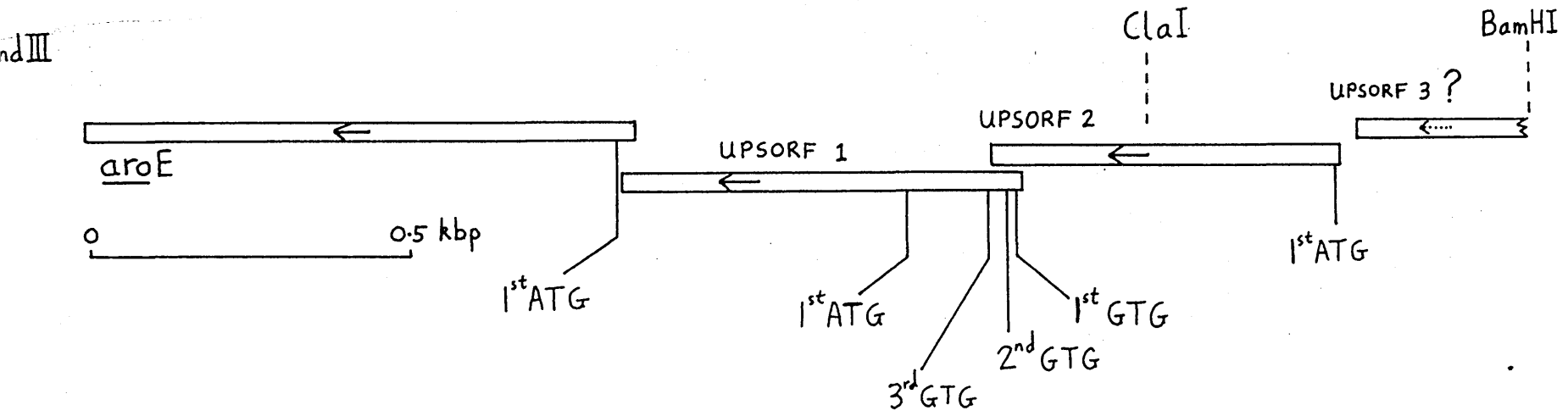


Figure 5.5 Arrangement of large ORF's and putative initiation codons in the HindIII-BamHI region - a summary

The ORF's are represented by the horizontal boxes. The initiation codon for aroE has been determined experimentally (see Chapter Four) and the positions of some putative initiation codons are shown for UPSORF 1 and UPSORF 2. It is not known whether the segment of ORF which extends from the BamHI site to near the start of UPSORF 2 - labelled "UPSORF 3?" - continues beyond the BamHI site for a significant distance. The horizontal arrows within the boxes show the 5' to 3' direction.

4. The pattern of putative transcriptional and translational control sequences within the HindIII-BamHI sequence is consistent with the operon hypothesis mentioned above, although this is tenuous evidence at best.

There are thus two separate but interwoven questions to be considered. Are UPSORF's 1 and 2 genes and, if so, do they form an operon which includes aroE? There is also a short segment of ORF, labelled "UPSORF 3?" in Figure 5.5, which is truncated by the BamHI site and which may constitute a third putative gene (see below).

5.4 Patterns of codon utilisation in aroE, UPSORF 1, and UPSORF 2

5.4.1 Background

Codon utilisation in E.coli genes is not random (Gouy and Gautier, 1982; Grosjean and Fiers, 1982). All genes, except for those which are only expressed at extremely low levels, tend to use codons corresponding to major species of tRNA's in the cell and the degree of this bias is greater in genes for highly expressed proteins. Another component of the observed bias is that in genes for very abundant proteins the choice between synonymous codons NNU and NNC is such as to favour intermediate values of codon-anticodon binding energy but for weakly expressed genes this rule does not hold. The expected biases for moderately/weakly

expressed E.coli genes, such as the shikimate pathway genes, are shown in Table 5.4.

5.4.2 Codon utilisation in aroE and comparison with some other common pathway E.coli aro genes

The codon usage of aroE is shown in Table 5.2. Alongside it are the codon usages for aroA and aroD (Duncan et al., 1984b; Duncan, 1985). In Table 5.4 the observed biases are compared with the expected biases. The patterns for the three aro genes are all very similar and are in good agreement with the expected pattern.

5.4.3 Codon utilisation in UPSORF 1 and UPSORF 2

The codon usages of UPSORF 1 and UPSORF 2 are shown in Table 5.3 alongside the codon usage for aroE. The codon utilisation of UPSORF 1 was calculated on the assumption that the translational initiation site is the third GTG codon which is the first plausible initiation codon after the end of UPSORF 2; the pattern is the same if one starts from the first ATG. Note that the polypeptide sequence predicted for UPSORF 2 contains 16 cysteine residues out of a total of 180 amino acids. In Table 5.4 the observed biases are compared with those expected for moderately/ weakly expressed E.coli genes. For both UPSORF 1 and UPSORF 2 the agreement is reasonably good, although poorer than in the case of the aro genes. However, this less satisfactory agreement will be at least partly due to UPSORF's 1 and 2 being smaller than the aro genes and thus their codon usage patterns being more vulnerable to statistical fluctuations.

Table 5.2 Codon utilisation in aroE, aroA, and aroD.

		aroE	aroA	aroD
ARG	CGA	1	0	0
	CGC	5	7	2
	CGG	3	2	0
	CGU	3	12	7
	AGA	1	0	0
	AGG	0	1	1
LEU	CUA	1	0	0
	CUC	5	3	5
	CUG	13	26	11
	CUU	4	4	1
	UUA	4	10	1
	UUG	8	5	1
SER	UCA	2	2	0
	UCC	4	5	5
	UCG	2	2	2
	UCU	2	6	3
	AGC	2	5	4
	AGU	5	1	2
THR	ACA	3	7	0
	ACC	3	10	8
	ACG	2	10	4
	ACU	4	7	3
PRO	CCA	4	2	2
	CCC	2	3	0
	CCG	3	9	3
	CCU	4	4	2
ALA	GCA	6	15	6
	GCC	3	7	13
	GCG	10	18	7
	GCU	11	6	3
GLY	GGA	4	2	0
	GGC	7	18	8
	GGG	1	4	0
	GGU	13	13	5

		aroE	aroA	aroD
VAL	GUA	4	4	4
	GUC	0	4	6
	GUG	5	9	5
	GUU	4	7	2
LYS	AAA	6	14	16
	AAG	2	3	1
ASN	AAC	2	9	2
	AAU	9	9	2
GLN	CAA	3	5	2
	CAG	7	7	4
HIS	CAC	3	5	2
	CAU	5	3	4
GLU	GAA	11	16	11
	GAG	6	6	7
ASP	GAC	5	3	6
	GAU	8	23	11
TYR	UAC	0	4	1
	UAU	5	9	5
CYS	UGC	1	4	3
	UGU	2	2	0
PHE	UUC	4	4	3
	UUU	10	9	5
ILE	AUA	3	0	0
	AUC	6	9	7
	AUU	9	17	9
MET	AUG	5	14	11
TRP	UGG	2	2	2
Stop	UAA			
Stop	UAG			
Stop	UGA	1	1	1

Table 5.3 Codon utilisation in aroE, UPSORF 1, and UPSORF 2

		aroE	UPSORF 1	UPSORF 2
ARG	CGA	1	4	1
	CGC	5	5	5
	CGG	3	0	1
	CGU	3	4	3
	AGA	1	1	0
	AGG	0	0	0
LEU	CUA	1	0	1
	CUC	5	1	1
	CUG	13	6	6
	CUU	4	2	2
	UUA	4	3	1
	UUG	8	7	0
SER	UCA	2	1	3
	UCC	4	1	3
	UCG	2	1	1
	UCU	2	1	2
	AGC	2	1	0
	AGU	5	2	1
THR	ACA	3	4	3
	ACC	3	5	2
	ACG	2	4	2
	ACU	4	1	0
PRO	CCA	4	3	1
	CCC	2	2	2
	CCG	3	5	6
	CCU	4	6	4
ALA	GCA	6	5	4
	GCC	3	4	2
	GCG	10	5	5
	GCU	11	5	2
GLY	GGA	4	1	5
	GGC	7	2	3
	GGG	1	5	3
	GGU	13	5	4

		aroE	UPSORF 1	UPSORF 2
VAL	GUA	4	2	2
	GUC	0	4	3
	GUG	5	5	1
	GUU	4	8	4
LYS	AAA	6	3	10
	AAG	2	1	3
ASN	AAC	2	2	2
	AAU	9	5	3
GLN	CAA	3	2	3
	CAG	7	5	5
HIS	CAC	3	0	3
	CAU	5	1	4
GLU	GAA	11	9	8
	GAG	6	2	4
ASP	GAC	5	5	2
	GAU	8	7	3
TYR	UAC	0	1	3
	UAU	5	3	3
CYS	UGC	1	2	8
	UGU	2	1	8
PHE	UUC	4	2	2
	UUU	10	6	6
ILE	AUA	3	1	1
	AUC	6	4	3
	AUU	9	3	5
MET	AUG	5	2	2
TRP	UGG	2	2	0
stop	UAA		1	1
stop	UAG			
stop	UGA	1		

N.B. For UPSORF 1 codon utilisation was calculated starting from the 3rd GTG (190 a.a.'s). 180 a.a.'s in UPSORF 2.

Table 5.4 Expected biases in codon usage for moderately/weakly expressed E.coli genes and comparison with observed biases in aroE, aroA, aroD, UPSORF 1, and UPSORF 2

	Expected biases	aroE	aroA	aroD	UPSORF 1	UPSORF 2
*LEU	CUG most favoured	+	+	+	+/-	+
	CUA least favoured	+	+	+	+	+/-
*LYS	AAA > AAG	+	+	+	+	+
*ILE	AUU > AUC >> AUA	+	+	+	+/-	+
*GLN	CAG > CAA	+	+	+	+	+
*GLU	GAA > GAG	+	+	+	+	+
ARG	CGC & CGU favoured	+/-	+	+	+/-	+
GLY	GGC & GGU favoured	+	+	+	+/-	-
ASP	GAU > GAC	+	+	+	+	+
TYR	UAU > UAC	+	+	+	+	+/-
PHE	UUU > UUC (often)	+	+	+	+	+

NOTE: 1. The most clear cut expected biases are asterisked.

2. Good agreement with the expected bias is indicated by "+", marginal agreement by "+/-", and disagreement by "-".

If the data for UPSORF's 1 and 2 are pooled then the match with the expected biases is as good as for aroE. Hence, the codon utilisation data support the hypothesis that both UPSORF 1 and UPSORF 2 are genes.

Although the small truncated "UPSORF 3?" (see Figure 5.5) contains only 92 known codons its codon usage is heavily biased in the manner expected for a gene. Only two amino acids occur more than 6 times, leucine and glutamic acid. Of the 16 leucine codons 7 are CUG and none are CUA. Of the 14 glutamate codons 10 are GAA. This suggests that UPSORF 3 may also be a gene.

5.5 Analysis of UPSORF's 1, 2, and 3 by the method of Fickett

5.5.1 Background

Fickett (1982) devised the "TESTCODE" procedure to distinguish protein coding regions from fortuitous ORF's in DNA sequences. This test depends mainly on finding statistical order in the base sequence (together with a small contribution from the base composition). Statistical order in protein coding regions is due to non-random codon utilisation. One consequence of biased codon utilisation is that the number of bases separating (e.g.) T's is much more likely to be 2,5,8...(2 + 3n) than it is to be 3n or 1 + 3n. Although Fickett's method does not provide a line of evidence independent from that adduced in the previous section it does provide a more quantitative assessment of

a particular ORF. It is applicable to stretches of DNA sequence greater than 200 bp.

The Fickett method was originally tested on 400,000 bases of sequence. It gave "no opinion" in 20% of cases and was wrong in 5% of cases (although the latter include cases with an assigned "probability of coding" as low as 0.84).

5.5.2 Results

TESTCODE gave a "probability of coding" of 0.98 for the whole of UPSORF 1 and predicted that this region encodes a protein. However, starting from the first ATG in UPSORF 1 the verdict was "no opinion" ("probability of coding" = 0.77).

For UPSORF 2 the verdict was "no opinion" ("probability of coding" = 0.58).

The most categorical result was obtained for the truncated UPSORF 3 where TESTCODE gave a "probability of coding" of 1.00 (TESTCODE indicator value of 1.25). It thus seems likely that UPSORF 3 is a gene.

5.6 Preliminary examination of the deduced amino acid sequences of UPSORF 1 and UPSORF 2

5.6.1 UPSORF 1

On the uncertain assumption that UPSORF 1 encodes a protein the deduced amino acid sequence was examined by eye

for any prominent features. This must be regarded as a highly tentative exercise.

One possible initiation site for translation is the first ATG(Met) codon. The a.a. sequence following the first Met shows some similarity to an E.coli signal sequence (Watson, 1984). This feature is labelled "A" in Figure 5.4. Many E.coli polypeptides with signal sequences begin (in their de novo form) either with Met Lys... or else have one or more lysines and/or arginines within the first six amino acids. Also, in most cases the run of uncharged/hydrophobic amino acids characteristic of a signal sequence is preceded by a lysine or arginine residue. Both these features are found in UPSORF I. However, the run of uncharged/hydrophobic amino acids in UPSORF 1's putative signal sequence (9 residues) is unusually short, and the run's distance from the initiation codon is rather long, compared to most known precedents thus casting doubt on this idea.

Much of the latter part of UPSORF 1 consists of relatively long runs of uncharged/hydrophobic amino acids. These are marked on Figure 5.4. However, the overall polarity of the hypothetical polypeptide (as defined by Capaldi and Vanderkooi, 1972), calculated from the first ATG, is 40%, the same as for E3.

There are some interesting sequence patterns within the neutral/hydrophobic runs but these may be completely fortuitous. The first (marked "B" in Figure 5.4) is as follows:

...Arg Trp Pro Gly Pro Val Thr Phe Val Phe Pro Ala Pro Ala Thr Thr Pro Arg ...

in which a hexapeptide sequence is closely followed by a very similar hexapeptide sequence. The second (marked "C" in Figure 5.4) consists of two neutral/hydrophobic runs, separated by a lysine residue, which both begin Pro Leu Val.

The small size of the putative UPSORF 1 polypeptide (16-23 kDa) led to the suggestion that UPSORF 1 might be the missing gene for shikimate kinase (E4) I (see Chapter One). The estimated molecular weight of E4I is 20 kDa (Ely and Pittard, 1979). To test whether there was any E4 gene(s) in the vicinity of aroE, crude extracts were made of E.coli AB2834 carrying the largest aroE subclone pIA306 (and of pIA307//AB2834) and assayed for E4 activity as described in Chapter Two. No sign of E4 overexpression was seen. However, this preliminary experiment must be interpreted with caution since both forms of E4 are known to be unstable. There is no trace in UPSORF 1 of the weak homologies between ATP using enzymes that were found by Walker et al. (1982).

5.6.2 UPSORF 2

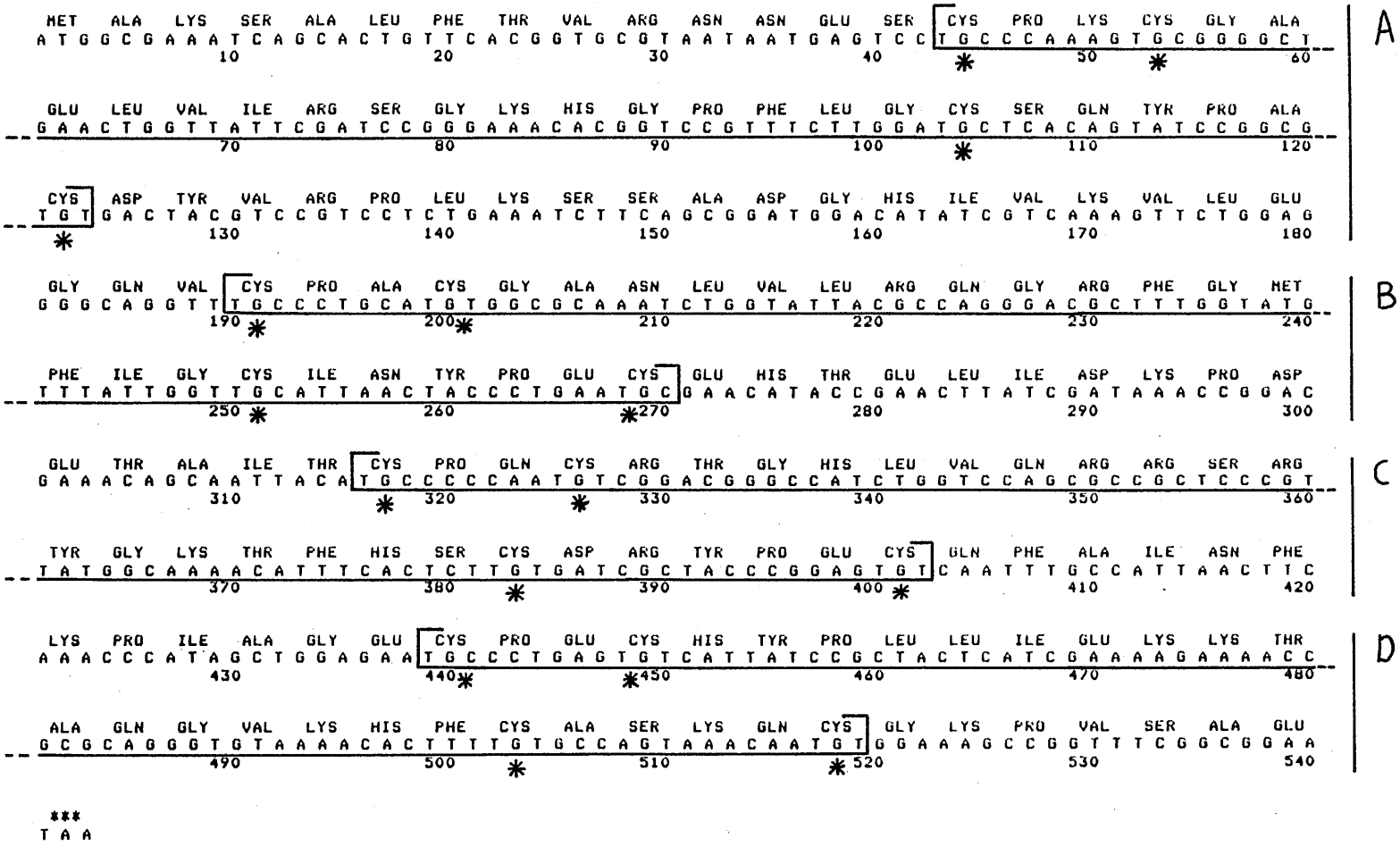
5.6.2A Four-fold repetition within the UPSORF 2 amino acid sequence

Inspection of the UPSORF 2 amino acid sequence revealed a very interesting structure: the 180 amino acid sequence (approximately 20 kDa) contains 4 imperfect repeats each of

Figure 5.6 Four-fold repetition within the UPSORF 2 amino acid sequence

The deduced amino acid sequence of UPSORF 2 is shown, starting from the first Met codon. The four homologous internal regions are underlined and labelled A, B, C, and D. All 16 cysteine residues are asterisked. The first nucleotide of the ClaI site is No.286 hence repeats A and B lie within the BamHI-ClaI region which was sequenced on only one strand. However, repeats C and D lie within the ClaI-HindIII region which was sequenced on both strands.

It should be noted that a four-fold structural repeat may actually represent, say, a three-fold functional repeat where each functional unit requires a "head" and a "tail" end of two structural repeats.


MET ALA LYS SER ALA LEU PHE THR VAL ARG ASN ASN GLU SER CYS PRO LYS CYS GLY ALA
ATGGCGAAATCAGCACTGTTCACGGTGCGTAATAATGAGTCCTGCCCAAAGTGCGGGGCT
10 20 30 40 50 60
GLU LEU VAL ILE ARG SER GLY LYS HIS GLY PRO PHE LEU GLY CYS SER GLN TYR PRO ALA
GAACTGGTTATTCGATCCGGGGAAACACGGTCCGTTTCTTGGATGCTCACAGTATCCGGCG
70 80 90 100 110 120
CYS ASP TYR VAL ARG PRO LEU LYS SER SER ALA ASP GLY HIS ILE VAL LYS VAL LEU GLU
TGT GACTACGTCCGTCCTCTGAAATCTTCAGCGGATGGACATATCGTCAAAGTTCTGGAG
130 140 150 160 170 180
GLY GLN VAL CYS PRO ALA CYS GLY ALA ASN LEU VAL LEU ARG GLN GLY ARG PHE GLY MET
GGGCAGGTTTGCCCTGCATGTGGCGCAAATCTGGTATTACGCCAGGGACGCTTTGGTATG
190 200 210 220 230 240
PHE ILE GLY CYS ILE ASN TYR PRO GLU CYS GLU HIS THR GLU LEU ILE ASP LYS PRO ASP
TTTATTGGTTGCATTAACTACCCTGAATGCGAACATACCGAACTTATCGATAAACCGGAC
250 260 270 280 290 300
GLU THR ALA ILE THR CYS PRO GLN CYS ARG THR GLY HIS LEU VAL GLN ARG ARG SER ARG
GAAACAGCAATTACATGCCCCCAATGTCGGACGGGCCATCTGGTCCAGCGCCGCTCCCGT
310 320 330 340 350 360
TYR GLY LYS THR PHE HIS SER CYS ASP ARG TYR PRO GLU CYS GLN PHE ALA ILE ASN PHE
TATGGCAAAACATTTCACTCTTGTGATCGCTACCCGGAGTGTCAATTTGCCATTAACTTC
370 380 390 400 410 420
LYS PRO ILE ALA GLY GLU CYS PRO GLU CYS HIS TYR PRO LEU LEU ILE GLU LYS LYS THR
AAACCCATAGCTGGAGAACTGCCCTGAGTGTCATTATCCGCTACTCATCGAAAAAGAAAAACC
430 440 450 460 470 480
ALA GLN GLY VAL LYS HIS PHE CYS ALA SER LYS GLN CYS GLY LYS PRO VAL SER ALA GLU
GCGCAGGGTGTAAAACACTTTTGTGCCAGTAAACAATGTGGAAAGCCGGTTTCGGCGGAA
490 500 510 520 530 540
***
TAA
A
B
C
D

27-29 amino acid residues. The positions of the 4 repeats within UPSORF 2 are shown in Figure 5.6 - they have been named A, B, C, and D. Repeats A and B lie within the BamHI-ClaI region which was sequenced mainly on one strand, while repeats C and D lie within the ClaI-HindIII region which was sequenced on both strands. Figure 5.7 shows the homologies between the 4 repeats.

The strongest homology is between repeats A and B. No "gapping" is required to achieve optimal alignment of these two repeats and there are 16 perfect matches out of 27 amino acids. There are also 4 mismatches between very similar amino acids: 2 Leu/Ile, 1 Lys/Arg, and 1 Gln/Asn. Although the overall homology between A, B, C, and D is poorer than that between A and B alone the basic structural motif is preserved:

...Cys Pro $\overset{1}{\text{—}}$ Cys ⟵ 16-18 a.a.'s ⟶ Cys ⟵ 5(4) a.a.'s ⟶ Cys

Note that there are four cysteine residues in each repeat, seemingly in two pairs. There are a few constant residues in addition to those in the basic motif.

The sequences which flank the repeats are of variable length being (in order from pre-A to post-D) 14, 22, 15, 12, and 7 residues long. There is no obvious homology between these flanking segments nor between them and the repeats, however, it may be significant that each of the four repeats is followed, at a variable distance, by ...Lys/Arg Pro...

The repeats presumably arose as the result of gene duplication events perhaps as a consequence of unequal crossing-over. Many similar cases of internal homologies

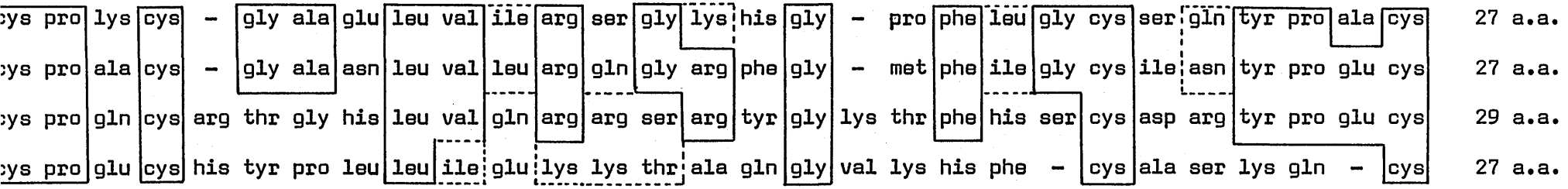


Figure 5.7 Homologies between the four UPSORF 2 repeats

The sequences of the four internal repeats A, B, C, and D were aligned by eye. Those columns of amino acids which are entirely boxed by solid lines are identical whereas those columns of amino acids whose boxing includes any broken lines are similar but not identical. Note that no gaps are required to align (optimally) A with B alone.

within proteins are known, for example, in anaerobic bacterial ferredoxins (Tsunoda and Yasunobu, 1968), Pseudomonas oleovorans rubredoxin (Yasunobu and Tanaka, 1973) and mitochondrial ADP/ATP translocase (Walker et al., 1982). The case of immunoglobulin heavy chains was mentioned in Chapter One.

The UPSORF 2 repeats and, in particular, the conserved pairs of cysteine residues initially suggested comparison with the small iron-sulphur proteins. The results of these comparisons will be considered in Section 5.6.2C. The intervening section contains a brief summary of the structure and role of iron-sulphur proteins.

5.6.2B Iron-sulphur proteins

Iron-sulphur (= nonhaem iron = FeS) proteins have been reviewed in the series edited by Lovenberg (1973): in particular, much of the amino acid sequence data discussed below is contained in the chapter by Yasunobu and Tanaka (1973). Bacterial FeS proteins have been reviewed by Yoch and Carithers (1979).

The simplest prosthetic group in FeS proteins is an iron ion (Fe^{2+} or Fe^{3+}) tetrahedrally coordinated to the sulphydryl groups of four cysteine residues in the polypeptide backbone (an "FeS centre"). Only in rubredoxins, however, are the FeS centres this simple. In the vast majority of cases inorganic sulphide is also liganded to the iron, either as an Fe_2S_2 centre (2 Fe, 2 sulphides,

4 backbone cysteines) or as an Fe_4S_4 centre (4 Fe, 4 sulphides, 4 backbone cysteines). Other more complex types also exist.

FeS proteins play crucial roles in a wide variety of biological redox processes, for example in:

(i) mitochondrial electron transport (NADH-Q reductase, succinate dehydrogenase)

(ii) photosynthesis (e.g. ferredoxin)

(iii) N_2 fixation

(iv) hydrogenase systems of many anaerobic bacteria

(v) oxygenase systems (e.g. adrenodoxin in steroid hydroxylation, P. oleovorans rubredoxin in omega-hydroxylation of alkanes and fatty acids)

(vi) ribonucleotide reductase

FeS proteins can be divided into two classes: firstly, the usually small electron transport proteins such as rubredoxins, ferredoxins, and adrenodoxin and, secondly, the more complex FeS redox enzymes. The latter often have additional prosthetic groups (e.g. flavins, copper, molybdenum, and perhaps selenium). We shall be mainly concerned with bacterial FeS proteins in the first category: rubredoxins and ferredoxins. These have been well studied at the sequence level.

5.6.2C Preliminary comparison of the UPSORF 2 amino acid sequence with known proteins

All the similarities outlined below between UPSORF 2 and various known proteins are very slight. Doolittle (1981)

has carefully considered the difficulties involved in establishing evolutionary relationships between distantly related proteins, the central problem being to distinguish chance similarity from common ancestry (or, rarely, convergent evolution). He has emphasised that the risk of chance "homology" between two amino acid sequences is often underestimated (especially where gaps must be inserted for optimal alignment) and the need for a rigorous statistical approach to such matters. The comparisons described here should therefore be regarded as preliminary. A more rigorous approach will be necessary once it is known if UPSORF 2 is expressed. Some discussion of quantitative methods of comparison will be found in the section on future work.

(i) Anaerobic bacterial ferredoxins.

There are minute but tantalising "homologies" between anaerobic bacterial ferredoxins (ABF's) and the UPSORF 2 amino acid sequence. All seven of the ABF sequences listed by Yasunobu and Tanaka (1973) show strong homology among themselves, and all have a two-fold internal repeat which accounts for most of the molecule. All have about 55 a.a.'s (6 kDa) hence the size of each repeat is roughly the same as that in UPSORF 2. Each of the ABF's contain two FeS centres and each repeat contains a cluster of four cysteine residues, a distribution different from that of the four cysteines in each UPSORF 2 repeat. This cluster and some

associated constant conserved residues are shown below:

...Cys$^{Val}_{Ile}$——CysGly——Cys———————Cys Pro...

... 1 2 3 4 5 6 7 8 9 10 11 12 ...

In the first ABF repeat position "6" is always Ala (and likewise in four out of seven cases for the second repeat) giving:

...Cys$^{Val}_{Ile}$—— Cys Gly(Ala)...

which is similar to the sequence around the first cysteine pair in repeats A and B of UPSORF 2:

...Cys Pro——Cys Gly Ala...

In the positions immediately preceding the cysteine cluster three of the seven ABF first repeats have:

...——Asp Ser...

-2 -1 0

of which two are:

...Asn Asp Ser...

-2 -1 0

Both the first and second repeats of all seven ABF's have Asp (or in one case Glu) in the "-1" position. All seven of the second repeats have Thr or Ser in the "0" position. In UPSORF 2 the very first cysteine pair is directly preceded by:

...Asn Glu Ser...

Thus the sequence around the first cysteine pair of UPSORF 2 resembles that around the first cysteine pair of ABF's (some more than others) over an interval of nine a.a. residues, respectively:

...Asn Glu Ser Cys Pro —— Cys Gly Ala...

compared with

...(Asn)$^{\text{Asp}}_{\text{(Glu)}}$ $^{\text{Ser}}_{\text{Thr}}$ Cys$^{\text{Val}}_{\text{Ile}}$—— Cys Gly Ala...

Five of the seven ABF's end with ...Glu-COOH (and one with Asp). Three end with ...Ala Glu-COOH. UPSORF 2 ends with ...Ala Glu-COOH.

In summary, there are traces of "homology" between UPSORF 2 and the ABF's as well as the common properties of internal repetition, roughly similar repeat sizes, and 4 cysteines per repeat. As stated at the beginning, some or all of the homology is quite possibly fortuitous. Also, the possibility of homology due to convergent rather than divergent evolution must always be remembered, especially in cases such as this where any sequence similarity is very slight.

UPSORF 2 is not the gene for the ferredoxin present in E.coli and of unknown function (Knoell and Knappe, 1974), as judged by the latter's different size, a.a. composition, and N-terminal a.a. sequence.

(ii) Rubredoxins.

The three anaerobic bacterial rubredoxins (ABR's) listed by Yasunobu and Tanaka (1973) - those from Micrococcus aerogenes, Peptostreptococcus elsdenii, and Clostridium pasteurianum - are all small, 6 kDa proteins (52-54 a.a.'s) with 1 FeS centre, no internal repeats, considerable homology

amongst themselves, and no known function. The a.a. sequence of all three has the pattern:

H_2N-Met...Cys$^{Thr}_{Ser}$—^{1}— Cys Gly⟵ 28-29 a.a.'s⟶Cys Pro —^{1}— Cys Gly$^{Ala}_{Val}$...$^{Asp}_{Glu}$-COOH

The sequence around the second cysteine pair:

...Cys Pro —^{1}— Cys Gly$^{Ala}_{Val}$...

is almost identical to that around the first cysteine pair in repeats A and B of UPSORF 2:

...Cys Pro —^{1}— Cys Gly Ala...

Like UPSORF 2 and the ABF's all three ABR's end with an acidic C-terminal residue.

Although the cysteine distribution in ABR's is qualitatively more similar to that in UPSORF 2 than that in ABF's, there are still very considerable differences. Firstly, in UPSORF 2 the Cys Pro - Cys pair comes first in the primary sequence. Secondly, the spacing is different (16-18 a.a's in UPSORF 2 v. 28-29 a.a.'s in the ABR's). Thirdly, the second cysteine pair in UPSORF 2 is not really homologous to the other cysteine pair in the ABR's, the former having 5 a.a.'s between the two cysteines.

The aerobic rubredoxin from P. oleovorans is a 20 kDa protein (174 a.a.'s) with two internal repeats and some homology to the ABR's. The first repeat has five cysteines arranged thus:

...Cys Pro AspCys⟵——— 29 a.a.'s ———⟶CysProCysCys...

The degree of "homology" between the ABR's and UPSORF 2 is clearly as weak as that between the ABF's and UPSORF 2, and the same strictures must be applied here also. There does seem to be a frequent use in FeS proteins of Cys—— ——Cys and, in rubredoxins, of CysPro —— Cys (ABF's also contain CysPro but in a different context).

(iii) Some other proteins.

Unfortunately the small FeS proteins are not the only proteins which contain important pairs of cysteine residues. For example, the small redox-active disulphide bridge proteins (= cysteine-hydrogen-donor proteins) glutaredoxin and thioredoxin both contain an essential pair of cysteine/half cystine residues arranged thus:

...Cys $\overset{1}{\text{——}}$ $\overset{2}{\text{——}}$Cys...

The sequence of the E.coli glutaredoxin active centre (Hoog et al., 1983) is:

...Cys Pro Tyr Cys...

This sequence is conserved in calf thymus glutaredoxin (Klintrot et al., 1984). There is no other immediately obvious homology to UPSORF 2 nor to FeS proteins. The existence of the conserved tetrapeptide emphasises the difficulties involved in interpreting the very faint resemblances between UPSORF 2 and the ABF's/ABR's.

Most thioredoxins contain ...Cys Gly Pro Cys... (Klintrot et al., 1984). There is no easily recognisable homology between E.coli thioredoxin and UPSORF 2.

Metallothionein has 20 cysteines out of 61 a.a's, no internal repeats, and no homology to UPSORF 2.

There is no homology detectable by eye between UPSORF 2 and the region around the 4 cysteine residues which coordinate the structural Zn^{2+} in liver alcohol dehydrogenase.

5.6.3 Conclusions

There is nothing visible in the UPSORF 1 sequence which adds weight to the hypothesis that it is a gene. In contrast, the 4-fold repetition within the UPSORF 2 sequence and the faint resemblances to some known proteins must make it very probable that UPSORF 2 is indeed a gene, especially when taken with the biased codon utilisation.

Speculation about the possible functions of UPSORF 2 will be left until after the next section which considers indirect evidence for an operon (with polycistronic mRNA) encompassing aroE and at least UPSORF's 1, 2, and 3. The possibility of such a link must clearly colour any discussion of the possible functions of UPSORF 2.

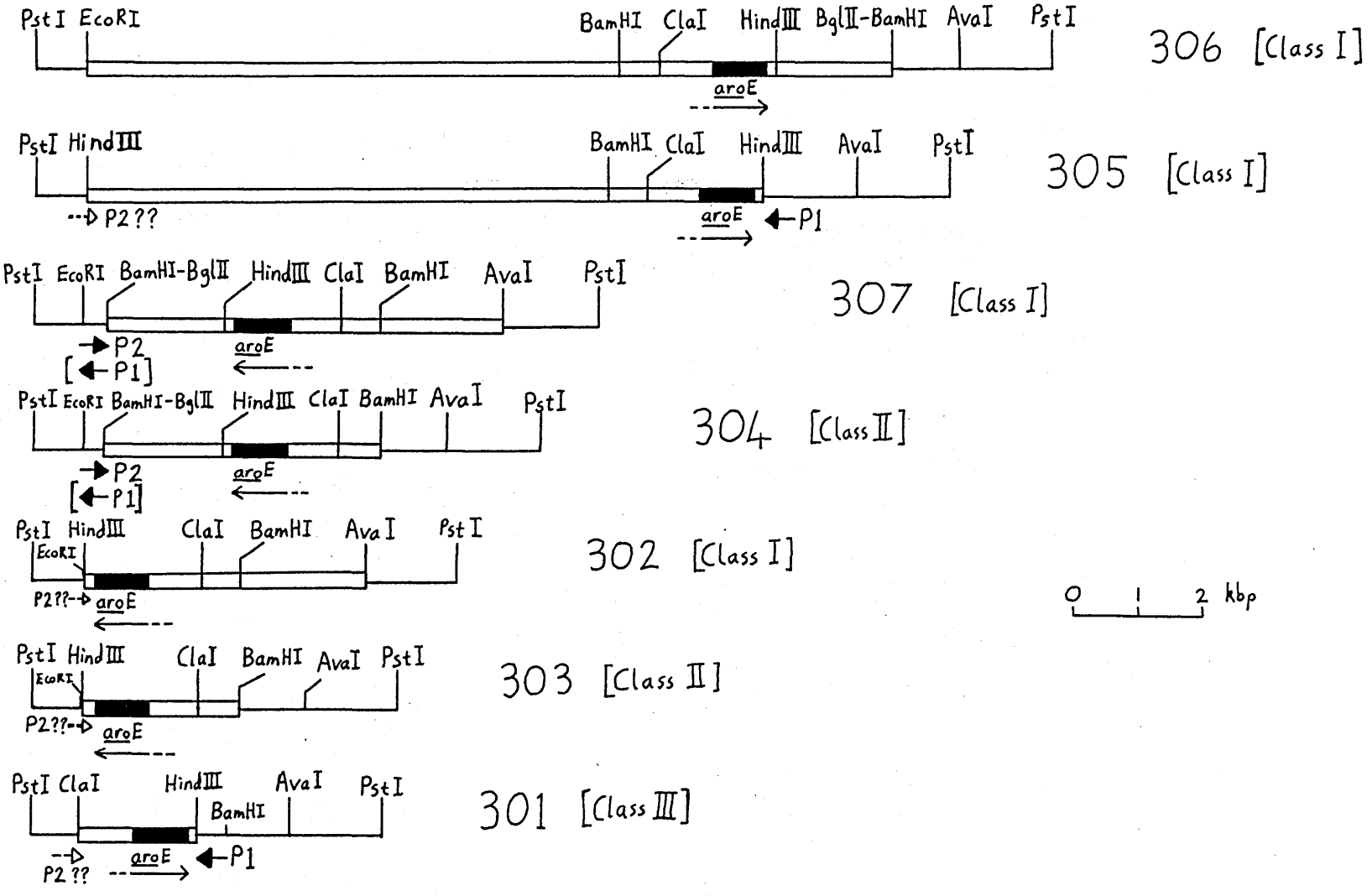
5.7 E3 specific activities of strains carrying particular aroE subclones

5.7.1 Background and summary of observations

This section considers again the wide variation in E3 specific activities found with different aroE subclones.

Figure 5.8 Position of the aroE gene within the various subclones and relationship to vector promoters.

The various aroE subclones described in Chapter Three are shown after linearisation at the PstI vector site. pAT153 vector sequences are represented by thin horizontal lines whereas insert sequences of E.coli DNA are represented by horizontal blocks. The aroE gene is indicated by the filled region of each block and its direction of transcription, from an unknown startpoint, is shown underneath by an open-headed arrow. Transcription from the vector promoters P1 and P2 (see text) is indicated underneath by the solid-headed arrows. Where there is the (remote) possibility of residual hybrid vector-insert P2 transcription, this is represented by a hollow-headed arrow with a dotted shaft. Key restriction sites are shown.


PstI EcoRI
BamHI ClaI HindIII BglII-BamHI AvaI PstI
306 [Class I]
aroE
PstI HindIII
BamHI ClaI HindIII AvaI PstI
P2??
aroE
P1
305 [Class I]
PstI EcoRI BamHI-BglII HindIII ClaI BamHI AvaI PstI
P2
[P1]
aroE
307 [Class I]
PstI EcoRI BamHI-BglII HindIII ClaI BamHI AvaI PstI
P2
[P1]
aroE
304 [Class II]
PstI HindIII ClaI BamHI AvaI PstI
EcoRI
P2??
aroE
302 [Class I]
PstI HindIII ClaI BamHI AvaI PstI
EcoRI
P2??
aroE
303 [Class II]
PstI ClaI HindIII AvaI PstI
BamHI
P2??
aroE
P1
301 [Class III]
0 1 2 kbp

The relevant data were presented in Section 3.3.5 and some preliminary discussion was given in Section 3.3.6. Now that the precise position, and direction of transcription, of the aroE gene are known it is possible to analyse the specific activity results more rigorously.

The relevant aroE subclones are shown in detail in Figure 5.8. Recall that the E3 specific activity (s.a.) values for E.coli strain AB2834 containing class I subclones are generally 4-5 times higher than those for class II subclones, and about 25 times higher than those for the class III subclone, in the same host strain.

5.7.2 Factors complicating the interpretation of the specific activity results

5.7.2A Interactions between vector and insert sequences

The first difficulty to be faced is that the pAT153 plasmid vector is not an inert container in which to study cloned E.coli DNA fragments.

Vector promoters are potentially able to distort the expression of cloned genes and there are two of relevance in the aroE subclones. Firstly, there is the tetracycline-resistance (tet^R) promoter "P2" (Stuber and Bujard, 1981). Its Pribnow box (see later sections) is just on the BamHI side of the HindIII site (see Figure 3.2). Hence, an insert extending clockwise from the HindIII or ClaI sites will usually, but not always, inactivate P2. If, by remote

chance, the insert fragment happens to contribute a passable Pribnow box, at the correct distance from the "-35" region remaining in the vector, then there may be some residual transcription from the resulting hybrid promoter. Secondly, there is the "anti-tet^R" promoter "P1" (Stuber and Bujard, 1981) which initiates transcription just clockwise from the HindIII site but in an anti-clockwise direction. This promoter is 1.5 times stronger than P2. Insertions that leave the vector's HindIII-BamHI region intact will leave this promoter intact.

These promoters P1 and P2 are significant because transcription of a cloned gene can be inhibited by a downstream opposing promoter unless there is a terminator to block the latter's effect. Fortunately, detailed consideration (see below) of the various subclones allows one to discount the vector promoters as a possible explanation for the wide range in s.a. values.

It is assumed that the s.a. value for pIA306, where there are no interfering vector promoters, represents the full expression level of aroE under its normal promoter(s), when carried on pAT153.

In pIA305 the anti-tet^R promoter P1 is in direct opposition to aroE transcription. However, pIA305 is a class I subclone so there is probably a protecting terminator between the HindIII site and aroE (see Figure 5.8). Whatever the reason, one may conclude that the close proximity of an opposing promoter to the HindIII site immediately downstream

from aroE does not have a significant deleterious effect on aroE expression, a point central to the argument as developed below. It also follows that sequences beyond (going in the direction of aroE transcription) this HindIII site are probably not required for maximal aroE expression. One can certainly rule out retroregulation (Gottesman et al., 1982) - reduced expression of a gene due to mRNA instability arising from loss of the normal 3' terminator structure - as a contributing factor in the reduced aroE expression from pIA303 and pIA301.

In pIA307 (Class I) aroE transcription is potentially opposed, at a distance of 2 kbp, by the tet^R promoter P2. Since P1 is 1.5 times stronger than P2 yet did not seem to affect aroE expression in pIA305 one would not expect any effect of P2 in pIA307 especially given the chance of terminators in the intervening 2 kbp of E.coli sequence. pIA305 and pIA307 being class I subclones implies that all sequences required for maximal aroE expression are located within the AvaI-HindIII region. This is confirmed by pIA302 being class I.

In pIA304 aroE is positioned relative to P2 exactly as in pIA307 yet 307 is class I while 304 is class II. The major difference between these two plasmids is that the AvaI-BamHI region is present in pIA307 but absent in pIA304.

pIA303 confirms the implication of the 304 result that sequences upstream (relative to aroE) of the BamHI site are required for full expression. There are 1.4 kbp between

the BamHI site and the start of the aroE coding sequence.

In pIA301 the aroE gene is situated exactly as in pIA305 with respect to P1 yet 305 is class I whereas 301 is class III. Any residual P2 transcription would, if anything, tend to support aroE transcription. The loss of E.coli sequences between the BamHI and ClaI sites appears to further reduce the expression of aroE relative to class II subclones.

In summary, one cannot explain any of the observed differences in E3 s.a.'s by invoking interactions with vector promoters. However, this conclusion should be treated with caution as there may be hidden complications. For example, there may be interactions between insert sequences and plasmid functions. One cannot exclude the possibility that the initial assumption made about pIA306 (and other class I subclones) is wrong and that, say, the copy number of the subclones having the ClaI-AvaI insert region is higher than for those lacking this region. This seems unlikely, however. None of the subclones have lost vector sequences during construction that are known to affect plasmid replication or partition.

5.7.2B Differences in plasmid size and host cell physiology

It was noted in Section 3.3.6 that the E3 s.a. was markedly influenced by the final A_{650} value to which cultures were grown, increasing as the final A_{650} increased (probably due to plasmid amplification). For this reason, s.a. values were only compared between cultures grown to similar final

A_{650}'s. All cultures were grown from a standard inoculum in the same volume of the same medium in similar flasks and aerated by (roughly) the same shaking speed on the same orbital shaker. However, there is some scatter in the results for a given subclone/host strain which may be partly due to the physiological state of the cells not being fully controlled. This scatter emphasises the need to treat the data cautiously. In particular there might be subtle threshold effects. Even so, the great difference between class I and class III subclones (about 25 fold) cannot easily be explained away.

One variable which can probably be eliminated as the source of the differences in E3 s.a. is plasmid size. It was at first thought that a larger plasmid might impose a greater "burden" on the host cell so that comparing A_{650} values for cultures of strains containing different sized subclones might be an invalid control for the cells' physiological state. However, pIA302 (class I) is smaller overall (6.7 kbp) than pIA304 (class II, 7.9 kbp), so that smaller size does not necessarily correlate with lower E3 s.a. pIA307 (class I, 8.8 kbp total) is not much larger than pIA304. pIA301 (class III, 5.5 kbp total) is only very slightly smaller than pIA303 (class II, 5.7 kbp total).

5.7.2C Conclusions

Leaving aside all the caveats and provisos in the above sections there is a simple hypothesis which explains the E3

s.a. results: there are promoter sequences far upstream from the aroE coding sequence which are required for the full expression of this gene. These sequences might either straddle the BamHI site as a single entity or be split between the ClaI-BamHI region and the BamHI-AvaI region. Both possibilities could explain the progressive reduction from class I → class II (BamHI truncation) → class III (ClaI truncation). This hypothesis is considered further in the next section.

5.7.3 Is aroE part of an operon?

There are two obvious interpretations of the hypothesis advanced in the previous section.

One could imagine that aroE is expressed as a monocistronic mRNA with an exceptionally long 5' untranslated leader sequence (or sequences, if there are multiple initiation sites). This cannot be completely discounted but is very implausible for two reasons. Firstly, most E.coli mRNA's have 5' untranslated leader sequences shorter than 200 bp (Kozak, 1983). For sequences around or upstream of the BamHI site to affect transcription directly would require a leader sequence of at least 1.4 kb, the distance from the BamHI site to the start of the aroE coding sequence. Secondly, such a leader sequence would inevitably include UPSORF 2 which is very probably a gene.

Alternatively and more plausibly, one could imagine that aroE is like the majority of E.coli genes and is expressed

as a polycistronic mRNA. This hypothetical operon would include at least aroE and UPSORF's 1, 2, and 3. If UPSORF 2 is a gene and UPSORF 1 is not then one would be left with a 584 bp intercistronic region and almost all such regions in E.coli are less than 400 bp (Kozak, 1983).

The operon hypothesis has been adopted as a model for further consideration. Experimental tests will be discussed later. A necessary feature of the model is that some promoter activity must remain after removing the region upstream from the BamHI site: in pIA304 and pIA303 (see Figure 5.8) there is no possibility of vector sequences contributing to the observed expression of aroE. If UPSORF 3 is a gene then the whole operon's promoter must be upstream from the BamHI site. One must therefore invoke a subsidiary promoter downstream from the BamHI site but before the ClaI site. Such subsidiary promoters within operons are known e.g. the trpP2 promoter within the E.coli trp operon (Lewin, 1983). The low expression of E3 from pIA301 could be solely due to residual activity from a hybrid vector-insert tet^R promoter. However, examination of the DNA sequence immediately downstream from the ClaI site shows this to be unlikely implying a second subsidiary promoter downstream from the ClaI site. As will be described shortly there are several sequences upstream from aroE, both before and after the ClaI site, which might act as weak promoters. However, this is undiluted speculation.

5.8 What might UPSORF 2 be?

Assuming that UPSORF 2 is a gene, and this seems very likely, one can speculate on the function of the gene product. There are only a few clues. Firstly, there are the very slight similarities between UPSORF 2 and various small electron carrier proteins, particularly FeS ones. Secondly, UPSORF 2 might be part of an *aroE* operon and it will be assumed initially that all the cistrons of such an operon would be functionally related in some way. Thus, one can start by posing a less difficult question: is there any possible undiscovered role in aromatic biosynthesis for a redox carrier protein?

Ubiquinone biosynthesis in *E.coli*, which proceeds from chorismate via 4-hydroxybenzoate, involves three hydroxylation steps of unknown mechanism; there is some evidence that a cytochrome P-450 system may be involved (Knoell *et al.*, 1978). The FeS protein adrenodoxin is required for steroid hydroxylation where it participates in a short non-phosphorylating electron transport chain to cytochrome P-450 (Lovenberg, 1973; Stryer, 1981). Camphor hydroxylation in *Pseudomonas putida* needs a ferredoxin ("putidaredoxin") which too is part of a cytochrome P-450 electron transport chain (Yoch and Carithers, 1979). As mentioned in Section 5.6.2C, the *P. oleovorans* rubredoxin is a component of an electron transport chain required for omega-hydroxylation of fatty acids and alkanes. Perhaps UPSORF 2 plays a role analogous to these in one or more of the hydroxylation steps

of ubiquinone biosynthesis. There is a precedent for an E.coli shikimate pathway gene being in an operon with a gene involved in a post-chorismate branch pathway, namely the aroF-tyrA "tyr" operon (Bachmann, 1983) which contains the genes for E1(Tyr) and chorismate mutase (T)-prephenate dehydrogenase.

A thought-provoking precedent is the first "mixed" operon in E.coli, the serC-aroA operon discovered by Duncan (1985). Here the two cistrons encode products involved in quite different pathways. However, there is thought to be a functional link between the serine and aromatic biosynthetic pathways since both serine and chorismate are required for the synthesis of the iron chelating agent enterochelin and E.coli has a pressing need to scavenge iron and has many systems for doing so. The connection between iron uptake and aromatic biosynthesis has been suspected for a long time (Gibson and Pittard, 1968) and recent work on enterochelin biosynthesis has confirmed the connection (Fleming et al., 1983). It is conceivable that the putative UPSORF 2 polypeptide resembles FeS proteins because it has an iron binding role in some iron uptake system and that the hypothetical operon is another example of the connections between aromatic biosynthesis and iron transport.

The E.coli genetic map (see Chapter One) was studied to see if UPSORF 2 could be correlated with any of the known genes adjacent to aroE.rrnD (an rRNA, tRNA operon) can be ruled out since it lies proximal to aroE and so must,

from the relative positions of genes on λ*spc*1, be downstream from *aro*E. *trk*A, which is involved in potassium transport, is known to be more than 5 kbp upstream from *aro*E (Meek and Hayward, 1984; D.W. Meek, personal communication). The position of *tol*M relative to nearby markers is considered uncertain (Bachmann, 1983). The original mapping data place it between *aro*E and *rps*L. Cotransduction frequencies imply that *tol*M is much closer to *rps*L than to *aro*E although this cannot yet be considered conclusive (Braun *et al*., 1980). However, despite the rather discouraging mapping data *tol*M merits further consideration in the light of the *ser*C-*aro*A operon.

The phenotype of *tol*M mutants is high level tolerance (not resistance) to colicin M (Braun *et al*., 1980). One of the many iron uptake systems of *E.coli* is the ferrichrome uptake system (Neilands, 1982). Ferrichrome is an iron chelator produced by fungi, often to be stolen by bacteria. The main component of the *E.coli* ferrichrome receptor (the *ton*A gene product) is also the unintended receptor for colicin M (Hantke and Braun, 1975; and references therein). There is tentative evidence that the wild-type *tol*M gene product (or that of a closely adjacent gene) is required for the inward movement of things bound to the ferrichrome receptor (Schaller *et al*., 1981) and thus it might be a a new component of the ferrichrome iron uptake system.

It is not inconceivable that UPSORF 2 might be an FeS protein involved in a complex "flavin reductase" (diaphorase) activity for chorismate synthase (see Section 1.8).

5.9 Possible transcriptional and translational control signals in the DNA sequence

5.9.1 Speculative promoters in the DNA sequence

E.coli promoters contain two conserved regions (Hawley and McClure, 1983): the "-10" region ("Pribnow box"; Pribnow, 1975) centred about 10 bp upstream from the transcriptional initiation site and, secondly, the "-35" region centred about 35bp upstream from the initiation site. The consensus sequences for these regions are shown below:

5' TTGACA ←—— 15-21 bp ——→ TATAAT ←— 4-8 bp ——→ A/G 3'

"-35" "-10"

The spacing between the "-35" and "-10" regions is usually 17 ± 1 bp and that between the "-10" region and the transcriptional initiation site (which is almost always a purine) is usually 6-7 bp. In Hawley and McClure's (1983) analysis of 112 well established *E.coli* promoters the individual bases in the consensus sequences were each present in the percentages shown:

$T_{82}\ T_{84}\ G_{79}\ A_{64}\ C_{54}\ A_{45}$ "-35"

and $T_{79}\ A_{95}\ T_{44}\ A_{59}\ A_{51}\ T_{96}$ "-10".

At present, attempts to locate promoters from DNA sequence alone are speculative since the context of each base-pair is important. However, analysis of promoter mutants shows that in most cases down-mutations decrease homology and up-mutations increase homology to the consensus sequence (Hawley and McClure, 1983).

The DNA sequence was examined for speculative promoters, as described in Chapter Two, but only for those acting in the BamHI to HindIII direction. Some sequences resembling consensus promoters are marked on Figure 5.9 (BamHI-ClaI) and on Figure 5.10 (ClaI-HindIII; no significant sequences straddle the ClaI site). These sequences are shown in more detail in Table 5.5. Sequences (2) and (3) in the BamHI-ClaI region look more convincing than the others but it must be emphasised that the matches to known "-10" and "-35" regions which are shown in Table 5.5 are of doubtful significance. A particular "-10"/"-35" region will probably be compatible with only certain "-35"/"-10" regions and only at certain spacings.

The aroE operon hypothesis probably requires a subsidiary promoter in the BamHI-ClaI region, as discussed in Section 5.7.3. Experiments designed to locate regions of promoter activity in the vicinity of aroE will be described later.

5.9.2 Palindromic structures and possible transcriptional terminators in the DNA sequence

5.9.2A Palindromes

Palindrome is used here only in the molecular biological sense. Palindromic structures in an E.coli DNA sequence ("inverted repeats"; "hairpin structures") can be fortuitous, or binding sites for regulatory proteins, or involved in mRNA secondary structure and transcriptional termination.

Table 5.5 Speculative promoters in the DNA sequence

(a) BamHI-ClaI (see Figure 5.9)

1. (5')T T G G T A ——15 bp——— T T A A C T - 6 bp - A

   "-10" identical to trp "-10".

2. (5')C T G A A A ——17 bp ——— T A T C G T - 4 bp - G

   "-10" identical to trp R "-10"; "-35" identical to proposed "-35" for cloacin.

3. (5')T T G A G C —— 20 bp ——— G A A A C T - 6 bp - A

   "-10" identical to mal EFG "-10"; "-35" identical to proposed "-35" for rpsA.

(b) ClaI-HindIII (see Figure 5.10)

1. (5')T T G T C G —— 20 bp ——

   T C G A A C —— 17 bp —— C A C A A T - 7 bp - G

   Closer "-35" identical to E.coli deoP1 and R100 RNA I "-35"'s.

2. (5')T T G C C A —— 21 bp —— G A G A A T - 6 bp - A

   T T A A C T —— 15 bp ——

   More distant "-35" identical to proposed "-35" for purF.

NOTES: (i) Matches with consensus sequence are underlined.

(ii) The speculative promoters are numbered as in the figures.

(iii) No speculative promoters straddle the ClaI site.

(iv) The significance of the identities given is doubtful - see text; for references see Hawley and McClure (1983).

Figure 5.9 Palindromic structures and speculative promoters in the BamHI-ClaI sequence.

The sequence of the BamHI-ClaI region is shown. It is written in the 5' to 3' direction starting at the BamHI end. Four palindromic structures are indicated by opposing sets of solid-headed arrows situated underneath the appropriate nucleotides, and each structure is named by a circled number. The number of bases in each symmetry related segment is given.

Three speculative promoters are marked by open-headed arrows located above the relevant nucleotides of each "-35" and "-10" region. Each "-10" region is labelled with a boxed number, the corresponding "-35" region having the same boxed number with a prime symbol. Possible purine start sites are indicated by a dot.

The first ATG codon of UPSORF 2 is indicated.

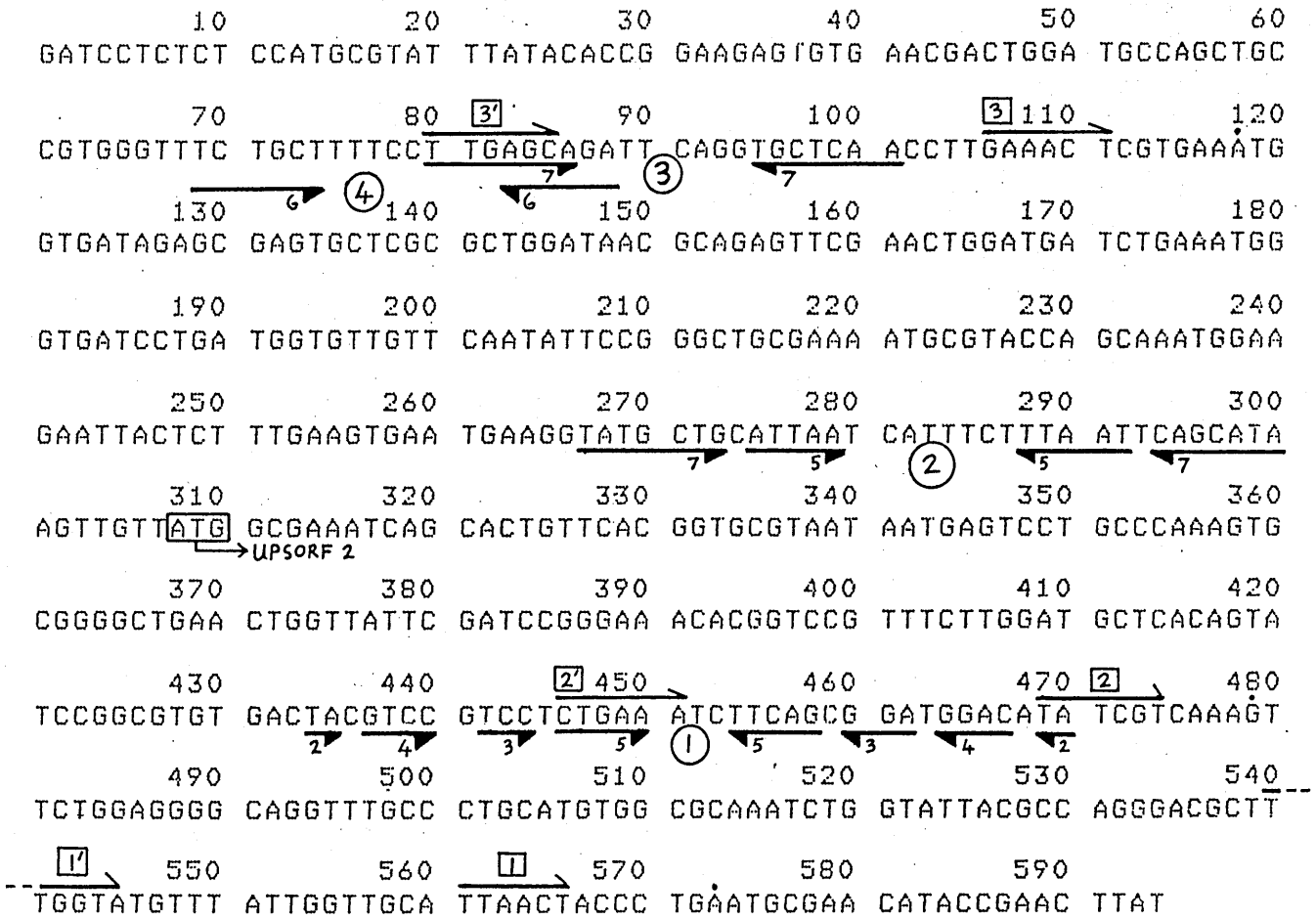
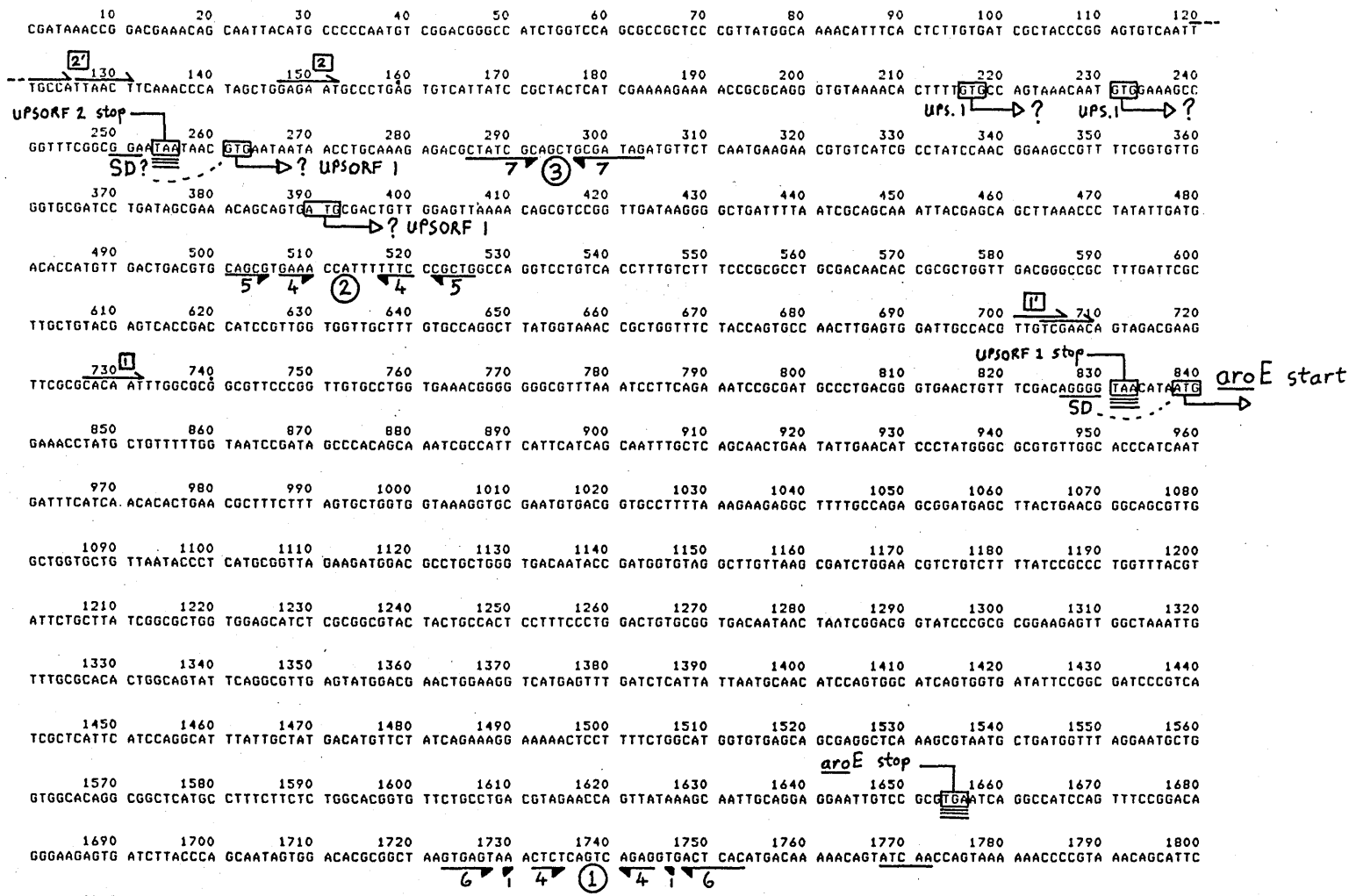

10 20 30 40 50 60
GATCCTCTCT CCATGCGTAT TTATACACCG GAAGAGTGTG AACGACTGGA TGCCAGCTGC
70 80 90 100 110 120
CGTGGGTTTC TGCTTTTCCT TGAGCAGATT CAGGTGCTCA ACCTTGAAAC TCGTGAAATG
130 140 150 160 170 180
GTGATAGAGC GAGTGCTCGC GCTGGATAAC GCAGAGTTCG AACTGGATGA TCTGAAATGG
190 200 210 220 230 240
GTGATCCTGA TGGTGTTGTT CAATATTCCG GGCTGCGAAA ATGCGTACCA GCAAATGGAA
250 260 270 280 290 300
GAATTACTCT TTGAAGTGAA TGAAGGTATG CTGCATTAAT CATTTCTTTA ATTCAGCATA
310 320 330 340 350 360
AGTTGTTATG GCGAAAATCAG CACTGTTCAC GGTGCGTAAT AATGAGTCCT GCCCAAAGTG
UPSORF 2
370 380 390 400 410 420
CGGGGCTGAA CTGGTTATTC GATCCGGGAA ACACGGTCCG TTTCTTGGAT GCTCACAGTA
430 440 450 460 470 480
TCCGGCGTGT GACTACGTCC GTCCTCTGAA ATCTTCAGCG GATGGACATA TCGTCAAAGT
490 500 510 520 530 540
TCTGGAGGGG CAGGTTTGCC CTGCATGTGG CGCAAATCTG GTATTACGCC AGGGACGCTT
550 560 570 580 590
TGGTATGTTT ATTGGTTGCA TTAACTACCC TGAATGCGAA CATACCGAAC TTAT

Figure 5.10 Palindromic structures, speculative promoters, and possible Shine-Dalgarno sequences in the ClaI-HindIII sequence.

The ClaI-HindIII sequence is written in the 5' to 3' direction starting from the ClaI site. Three palindromic structures are indicated by opposing sets of solid-headed arrows beneath the appropriate bases, and each structure is referenced by a circled number. The number of bases in each symmetry related segment is given.

Two speculative promoters are marked by open-headed arrows above the relevant nucleotides. Each "-10" region is labelled with a boxed number, the corresponding "-35" region(s) having the same boxed number with a prime symbol. Possible purine start sites are indicated by a dot.

The first ATG codon of aroE is marked and possible initiation codons for UPSORF 1 are indicated by "?" s. The termination codons of aroE, UPSORF 1 and UPSORF 2 are shown. The probable Shine-Dalgarno (S-D) sequence for aroE is underlined. Also shown is a possible S-D sequence upstream from the third GTG codon of UPSORF 1.


UPSORF 2 stop
SD?
? UPSORF 1
? UPSORF 1
UPS.1 ?
UPS.1 ?
UPSORF 1 stop
SD
aroE start
aroE stop

A limited search was carried out for palindromic structures, as described in Chapter Two. (For convenience these structures will be discussed using the stem/loop nomenclature for hairpins without necessarily assuming that they ever take this form). Only those with central loop sizes $\geqslant 3$ and $\leqslant 10$ bases would have been found by this search and, furthermore, only those where the symmetry related segments immediately flanking the central loop are $\geqslant 4$ bp long. The larger structures found are shown in Figures 5.9 and 5.10 (and none straddle the ClaI site).

In the BamHI-ClaI sequence the first palindrome structure ("1") is rather extensive having a 3 base loop and a 17 base long stem containing 14 bp in 4 segments. It lies wholly within UPSORF 2 and its functional significance, if any, is unknown. It overlaps one of the better speculative promoters. Structure "2" will be described shortly in the section on ribosome binding sites. It may be of considerable functional importance. There are two smaller structures ("3" and "4") which, like "1", happen to overlap one of the more plausible speculative promoters.

In the ClaI-HindIII sequence palindromic structure "1" may be the transcriptional terminator for the aroE gene/operon and will be discussed in detail below. Two other smaller structures of doubtful significance are also marked.

5.9.2B Transcriptional terminators in E.coli

There are two general classes of transcriptional termination signal in E.coli (Rosenberg and Court, 1979; Holmes et al., 1983). Factor-independent terminators all have a relatively G/C rich inverted repeat followed by a run of T's (in the sense strand). The second class, factor-dependent terminators, is rather diverse and few have been well characterised. Termination at these requires ancillary factors such as rho (or NusA). All contain palindromic structures, although the "hairpin" is often of doubtful stability, and some are associated with the sequence 5'... [CA] ATCAA...3' in the sense strand.

5.9.2C Possible transcriptional terminators in the DNA sequence

As mentioned above the palindromic structure labelled "1" in Figure 5.10 may be the transcriptional terminator for the aroE gene/operon. The loop size is 5 bases and the stem is 13 bases long, having 11 bp (5 G-C) and 2 G-A mismatches. (It is associated with a compression in the DNA sequencing ladder for template 24-2). 14 bp beyond the end of the palindrome there is the sequence 5'...ATCAA...3' (underlined in Figure 5.10) which is associated with some factor-dependent terminators. There are 66 bp between the aroE stop codon and the start of the stem. Although it is still impossible to recognise unambiguously a transcriptional terminator from sequence data alone, this structure is certainly a plausible candidate for a rho-dependent terminator.

It is worthy of note that 4 bases after the end of the stem is the sequence (in the sense strand) 5'...CAAAAA...3', followed 12 bases later by 5'...GTAAAAAA...3'. A putative terminator 45 bp after the stop codon of the aroD gene has the sequence 5'...CTAAAAAA...3' (in the sense strand) immediately following the 9 bp stem (Duncan, 1985).

None of the other palindromic structures described resemble known terminators.

5.9.3 Possible ribosome binding sites upstream from the large ORF's

5.9.3A Background

Shine and Dalgarno (1975) first noticed the imperfect complementarity between the 3' end of E.coli 16S rRNA and translational initiation regions. Their observation has proved to be generally valid (Stormo et al., 1982) and the term "Shine-Dalgarno" (S-D) sequence is now used for the region of a mRNA's ribosome binding site which shows this complementarity. However, adjacent regions are not totally random.

The 3' end of the 16S rRNA has the sequence:

3'AUUCCUCC...5'

which is complementary to:

5'...UAAGGAGG...3'.

In the compilation of Stormo et al. (1982) 83% of genes had at least either AGG or GGA or GAG 6-9 bases before the AUG,

although in one case there were 12 bases between the end-point of complementarity and the first base of the initiation codon. In addition to the S-D sequence there are often one or more stop codons in the ribosome binding site, at least in polycistronic mRNA's, and sometimes (particularly in highly expressed proteins) all or part of the heptanucleotide PuPuUUUPuPu. The "spacer" between the S-D sequence and the initiation codon is often A/U rich.

5.9.3B Results of sequence comparisons

There is a plausible S-D sequence upstream from the aroE initiation codon. It is marked on Figure 5.10 and is shown in detail below:

```
                S-D
5'...UGUUUCGACAGGGGUAACAUAAUG...3' (aroE) mRNA
        *** *
   3' AUUCCUCC...5'  16S rRNA
```

There are four base-pairs including three contiguous base-pairs. A stop codon (that for UPSORF 1) is underlined. The spacer region before the initiation codon (underlined twice) is A/T rich (6 out of 7). Just upstream from the S-D sequence is:

5'...UGUUUCG...3' (marked with a dashed line)

which resembles the heptanucleotide mentioned above.

With UPSORF 1 the first three GTG's and the first ATG will be considered as potential initiation codons. For each of the first and second GTG's and the first ATG there is no significant complementarity (at least 3 contiguous bp with at least one G-C bp) within a reasonable distance (less than

13 bases). However, for the third GTG (the first initiation codon after UPSORF 2) there is a possible S-D sequence:

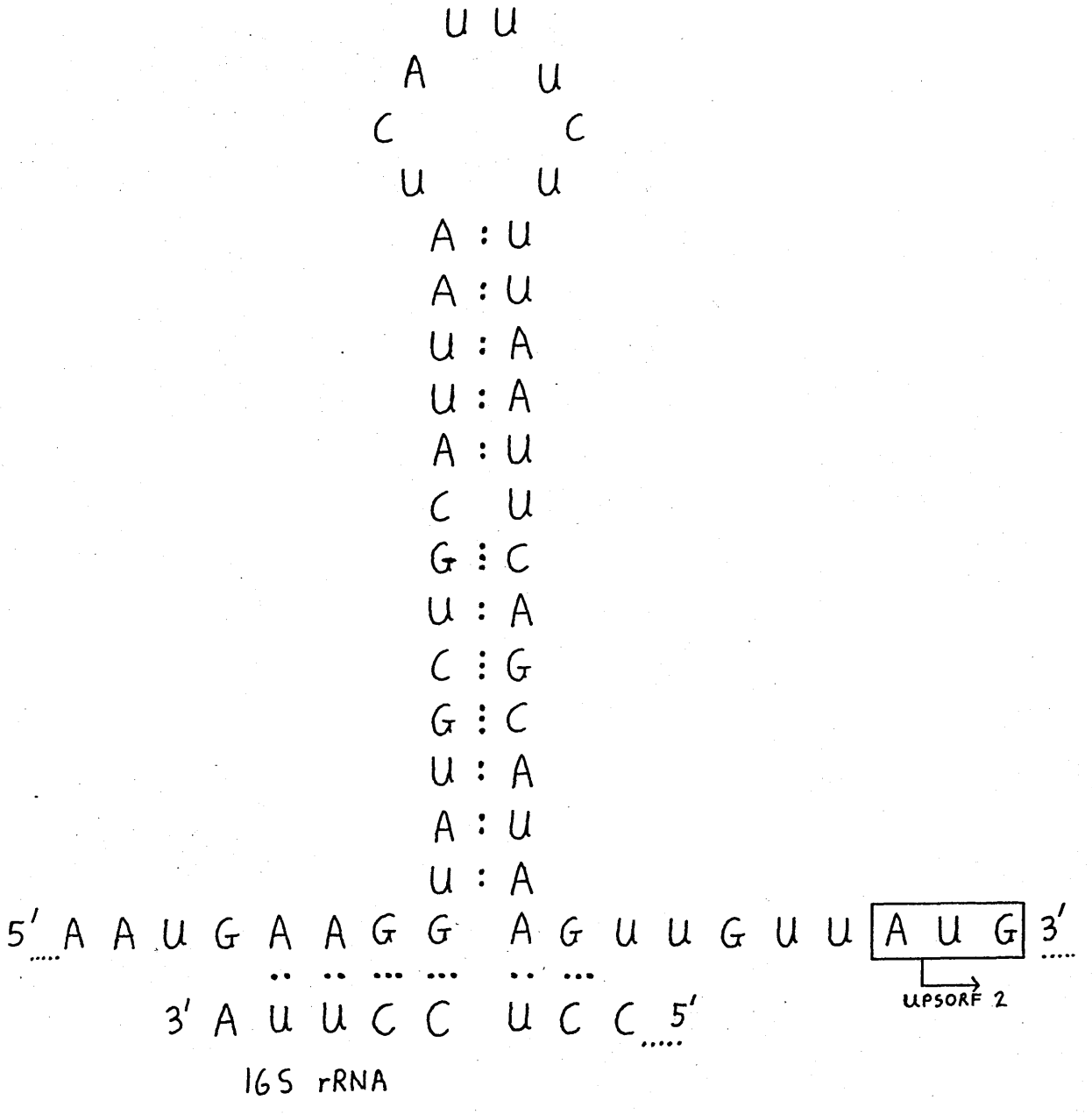
```
                  S-D
5'...GGUUUCGGCGGAAUAAUAACGUG...3' (UPSORF 1) mRNA
           ***
      AUUCCUCC...5'  16S rRNA
```

There are three contiguous base-pairs, two of which are G-C; two stop codons (those at the end of UPSORF 2); an A/T rich spacer (7 out of 8); and the heptanucleotide 5'...GGUUUCG...3' two bases upstream of the three contiguous base-pairs (as for aroE).

For UPSORF 2 there is a poor region of complementarity beginning five bases upstream from the ATG initiation codon. However, the palindromic structure numbered "2" in Figure 5.9 may provide an excellent S-D sequence for UPSORF 2 by generating a hairpin in the mRNA. This possibility is illustrated in Figure 5.11. Six base-pairs can be drawn between the mRNA and the 16S rRNA. A similar hypothesis has been put forward previously in the case of the phage T4 gene 38 mRNA (Gold et al., 1981). There are 23 bases between the gene 38 initiation codon and the first plausible S-D sequence but a hairpin in the mRNA is thought to reduce the separation to 5 bases. However, as here, no experimental tests have been done. It is not known, in general, whether the initiation complex could tolerate such a herniation, nor is it clear whether such unusual arrangements - if they exist - have some special import.

Figure 5.11 A hairpin in the mRNA may lead to a good S-D sequence for UPSORF 2.

Note that a stop codon (UGA) precedes the stem-loop structure.


U U
A U
C C
U U
A : U
A : U
U : A
U : A
A : U
C U
G : C
U : A
C : G
G : C
U : A
A : U
U : A
5′ A A U G A A G G A G U U G U U A U G 3′
3′ A U U C C U C C 5′
16S rRNA
UPSORF 2

5.9.4 Conclusions

The discussion in Section 5.9 has been speculative to varying degrees, entirely so in the case of possible promoters. However, two potentially functional palindromic structures have been noted: the putative terminator after aroE (which resembles a possible terminator after aroD) and the hairpin structure in Figure 5.11. The finding of almost equally clear-cut S-D sequences for aroE and UPSORF 1 is reassuring (as is the existence of at least some provision for UPSORF 2) but the best evidence that UPSORF 1 might be a gene is still its biased codon utilisation.

In the next section consideration is given to experiments which might reveal whether there are indeed genes immediately upstream from aroE and, if so, whether they are part of an aroE operon.

5.10 Possibilities for future work on the putative aroE operon

5.10.1 Are there upstream genes and, if so, are they part of an aroE operon?

The basic question here is whether UPSORF's 1 and 2 (and any others) are actually expressed as proteins. An answer could be obtained by analysis of the polypeptides produced by various aroE subclone plasmids when placed in an E.coli in vitro transcription/translation system (available commercially from Amersham). The most suitable subclone is pIA303 which carries only aroE, UPSORF 1, and UPSORF 2 on the HindIII-BamHI fragment. pAT153 could be used as a control for plasmid

encoded functions. It would be sensible to test pIA307 in such a system also, as well as pIA301. With the latter one would not expect to see any UPSORF 2 polypeptide, and reduced amounts of E3 and UPSORF 1 polypeptides.

The identification of any putative UPSORF polypeptides could be confirmed in various ways. Restriction site mutagenesis at the ClaI site, say, could be used to introduce a frameshift mutation in UPSORF 2 thus leading to a smaller product. Alternatively, suitable restriction fragments could be cloned in pKK223-3 and any overexpression of appropriately sized polypeptides detected by SDS PAGE of whole cell extracts. It would be important here not to discard the membrane fraction, most of which is spun out during the extraction procedure used for assays of E3 activity, in case the UPSORF gene products are associated with the cell membrane. Overexpressing constructs would be very useful for attempts at purifying the putative gene products (see below). A different approach would be to construct beta-galactosidase fusion proteins containing part or all of the UPSORF 1 or 2 sequences. This is relatively straightforward using the series of vectors developed by Ruther and Muller-Hill (1983). Such fusion proteins are often easily purified (by preparative SDS PAGE, for example) in a form suitable for the raising of antisera. The antisera can then be used as probes for Western blots of whole cell extracts of E.coli. As well as potentially providing further confirmation that UPSORF's 1 and 2 are expressed, such antisera could be exploited in other ways (see below).

If UPSORF's 1 and 2 are translated then the next question is whether they are transcribed as part of a polycistronic mRNA with aroE. A tentative answer could be quickly obtained by Northern blot analysis (Thomas, 1980) of RNA isolated both from wild-type E.coli cells and from strains carrying particular aroE subclones. M13 clones containing known insert fragments could easily be used to construct radioactive hybridisation probes. Particular templates specific for the non-overlapping regions of aroE, UPSORF 1, and UPSORF 2 are available (e.g. 27-4 for aroE, 21-3 for UPSORF 1, and 19-8 for UPSORF 2). If these three different classes of probe all detect a transcript of identical size then this is good evidence in favour of an operon rather than separate genes. It would then be necessary to confirm the operon model by mapping the location of transcriptional initiation and termination sites around aroE by S1 analysis (Maniatis et al., 1982).

At this stage one would probably wish to confirm the sequence of the ClaI-BamHI region on both strands. It would also be interesting to extend the sequence upstream (with respect to aroE) from the BamHI site.

5.10.2 If UPSORF's 1 and 2 are genes then how might their functions be determined?

It may not prove easy to determine the functions of the UPSORF 1 and 2 gene products. The option of raising antisera against the gene products by fusion protein techniques was mentioned above. The antibodies could be used as an assay for the purification to homogeneity of the gene products -

this would most sensibly be done after the construction of overproducing strains. N-terminal amino acid sequencing could be used to confirm that the purified proteins were encoded by UPSORF's 1 and 2.

Should fusion protein techniques prove unsuccessful then an alternative approach to raising antibodies would be to synthesise oligopeptides, corresponding to segments of the deduced amino acid sequences, and to use these as immunogens. If it worked this approach would allow the construction of an antibody affinity column from which specifically bound protein could be gently eluted by means of the appropriate oligopeptide. Antisera, produced by whatever route, could also be used to study the cellular localisation of the gene products.

It is quite possible that the hypothetical gene products could be purified without any assay, given a high level of overexpression. The chromatographic behaviour of the desired polypeptide could be determined by comparing the elution profiles of crude extracts in the presence and absence of high level overexpression. This assumes there will be no adventitious obscuring peaks.

The purification to homogeneity of the UPSORF 2 gene product would permit the testing of several possibilities discussed earlier. For example, if UPSORF 2 encodes an iron-sulphur protein then, firstly, atomic absorption spectroscopy should reveal the presence of iron and, secondly, the protein should display the electron spin resonance signature characteristic of FeS proteins. One could also test the protein's ability to bind iron.

Recently a powerful tool has been developed, analogous to some naturally occurring regulatory mechanisms, which is generally applicable to the problem of discovering a gene's function. This is the technique of antisense RNA and it offers the hope of being able to switch off the *in vivo* expression of particular cloned genes at will (Pestka *et al.*, 1984). A plasmid producing an antisense RNA specific for one of the UPSORF's could be produced quite easily. It would be interesting to see whether the introduction of such a plasmid into an *E.coli* K12 host conferred any new growth requirements. The possibility of polarity effects must, however, be remembered (Pestka *et al.*, 1984).

It would be useful to know whether copy number over-expression of the regions upstream from *aroE* in an *E.coli tolM* mutant host could confer near wild-type sensitivity to colicin M (see Section 5.8).

The deduced amino acid sequences of the UPSORF's should be compared with all known amino acid sequences by computerised searching of a protein sequence database. Also, as mentioned in Section 5.6.2C, the preliminary analysis of the very weak similarities between UPSORF 2 and various known proteins must be placed on a much more rigorous and quantitative footing. This also applies to any other faint similarities found by computer searches. A favoured method of rigorous analysis for faint homologies involves repeated scrambling (by computer) of the sequences to be compared and the construction of a

distribution curve for the number of random matches (Doolittle, 1981). One may then assess the statistical significance of the number of matches in the optimal alignment of the real sequences. However, a penalty must be introduced for every gap inserted in aligning the two sequences: this is a problematic area. An argument in favour of relatedness can be strengthened if a weak homology between two proteins is also found in a third distantly related protein.

5.11 Preliminary comparison of the E.coli shikimate dehydrogenase sequence with that of other proteins

5.11.1 N.crassa catabolic quinate/shikimate dehydrogenase

The N.crassa catabolic quinate/shikimate dehydrogenase - the qa-3 gene product - was described in Section 1.8.5B. This is the only protein with shikimate dehydrogenase activity, other than E.coli E3, for which the amino acid sequence is known. However, E.coli E3 only works with $NADP^+$ as cofactor whereas the qa-3 gene product prefers NAD^+. A very preliminary comparison of the two amino acid sequences was carried out by dot matrix analysis using the "DOTPLOT" program from the WISGEN software package (Devereux et al., 1984; see Chapter Two). The initial results are shown in Figure 5.12. There is clearly no strong overall homology between the two sequences. However, there is a weak, broken, and displaced diagonal indicating some short stretches of possible homology. These short regions are compared in detail in Figure 5.13.

Figure 5.12 Comparison of aroE and qa-3 at the amino acid level by dot matrix analysis

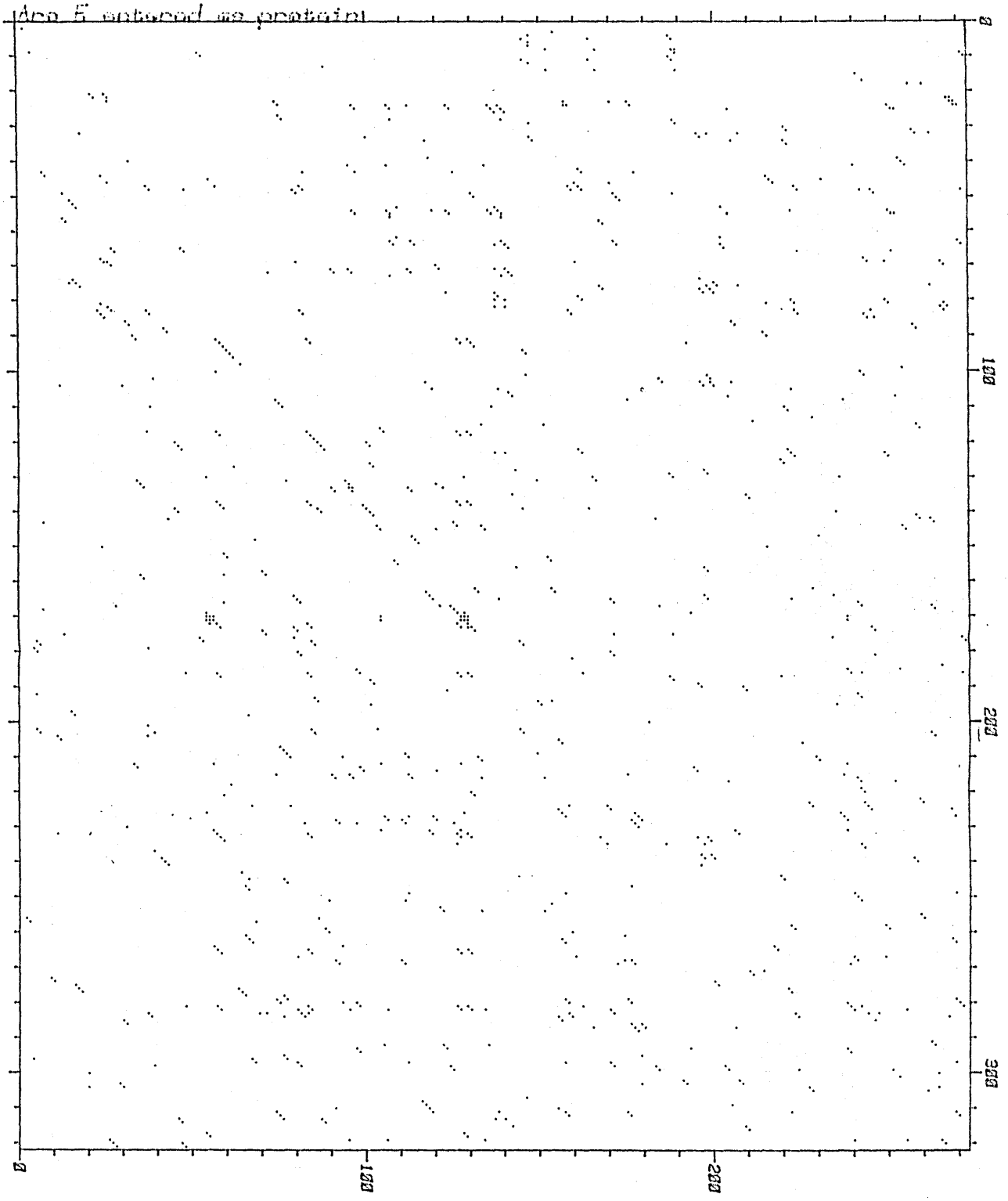
The aroE amino acid sequence lies along the vertical axis and the qa-3 amino acid sequence lies along the horizontal axis. The scales are both in units of amino acid residues. The analysis was carried out using the DOTPLOT program (see text). In this case the program compared the two sequences using segments of three residues and placed a dot where at least two out of three residues matched exactly. No allowance was made for conservative substitutions.

Window: 3 Stringency: 2 971 Points 11.9 bases/cm

12-JUL-94 16:41

aroa.pep 1647

Aroa E enterocol ms protein



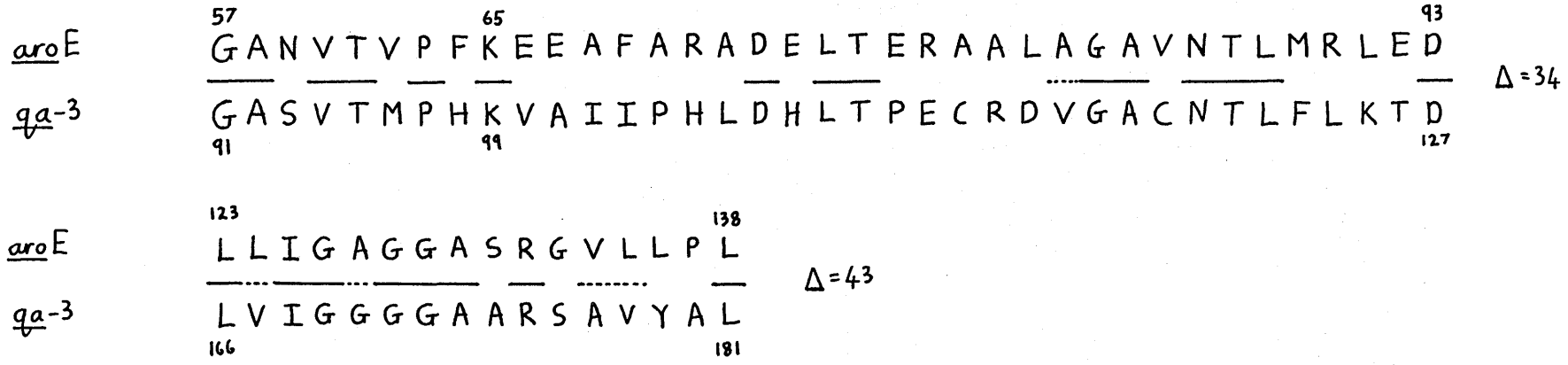


Figure 5.13 Regions of possible homology between aroE and qa-3 at the amino acid level

Amino acid residues are shown using the single letter code. The positions within the relevant a.a. sequence of the residues at the start and end of each region are indicated, as are the positions of a conserved lysine residue (see text). Perfect homology is shown by a continuous line between the two sequences whereas a broken line is used to indicate conservative substitutions. Note that for overall alignment a nine amino acid residue gap would have to be inserted in the aroE sequence between the two regions shown.

At amino acid residue 65 in aroE and 99 in qa-3 there is possibly a conserved lysine residue (see Figure 5.13). This is of interest because it has been shown previously that the E3 activity of N.crassa arom is very susceptible to inhibition by two different chemical modification procedures both of which are directed at lysine residues (J. Lumsden and J.R. Coggins, unpublished results). Treatment either with methyl acetimidate or with formaldehyde plus sodium borohydride leads to rapid, pseudo-1st order inactivation of arom E3 and in both cases the presence of shikimate protects against loss of activity. It is not known whether the putative lysine residue implied by these results has an essential role in the active site or whether modification merely obstructs the active site or changes its conformation. It will be interesting to see whether the lysine "conserved" between aroE and qa-3 is also conserved in the sequence of the E3 region of yeast arom. Clearly a quantitative analysis of the relatedness of aroE and qa-3 must also be carried out: at present it is quite possible that the apparent homology is fortuitous.

5.11.2 Other dehydrogenases

No attempts have yet been made to compare the amino acid sequence of E.coli E3 with that of other $NADP^+$-dependent dehydrogenases. However, a brief sketch of some relevant aspects of dehydrogenase structure is given below.

Early work concentrated on four NAD^+-linked enzymes: lactate dehydrogenase (LDH), malate dehydrogenase (MDH), liver alcohol dehydrogenase (LADH), and glyceraldehyde-3-phosphate dehydrogenase (GAPDH). The three-dimensional structures of these proteins were obtained by X-ray crystallography and protein sequencing (reviewed by Rossmann et al., 1975). The finding that LDH and MDH were similar in overall structure was not too surprising. However, of great interest was the discovery that although the overall structures of GAPDH and LADH were quite different from those of LDH and MDH (and from each other) there was a striking similarity in the conformation of the NAD^+ binding domain in all four enzymes. (The conformation of the bound NAD^+ is almost the same in each case.) At the sequence level there are very few identical residues between LDH, GAPDH, and LADH (although there are significant homologies between different dehydrogenases with the same function). Sequence comparisons made prior to the determination of the structures gave confusing and often erroneous results. Only after alignment of the 3-D structures was it possible to detect a few important and invariant residues required for coenzyme binding (Rossmann et al., 1975; Adams et al., 1981; Rossmann, 1983). Detailed analysis of the structural homology between NAD^+ binding domains suggests that they arose by divergence from a common ancestor.

Only more recently has attention turned to the structure of $NADP^+$-dependent enzymes such as 6-phosphogluconate dehydrogenase (6PGDH), dihydrofolate reductase (DHFR), and

glutathione reductase (reviewed by Adams et al., 1981; Adams et al., 1983). These are much more variable in structure than the known NAD^+-dependent enzymes. The structure of 6PGDH (from sheep liver) is almost entirely lacking in beta-sheet which is in immediate contrast with other $NADP^+$- or NAD^+-linked dehydrogenases. The essence of an NAD^+ binding domain is six strands of parallel beta-sheet with surrounding parallel alpha-helices. It was initially thought that the structures of $NADP^+$-dependent enzymes might fall into several different functional classes e.g. those catalysing direct hydride transfer and those utilising both $NADP^+$ and FAD as coenzymes. However, both DHFR and 6PGDH catalyse direct hydride transfer yet differ radically in structure, DHFR having six parallel strands of beta-sheet. Much more structural work is required on a variety of $NADP^+$-linked enzymes if any patterns are to be detected. The potential availability of large quantities of E.coli E3, and the known amino acid sequence, make this $NADP^+$-dependent enzyme a candidate worthy of consideration by X-ray crystallographers.

CHAPTER 6 GENERAL DISCUSSION AND FUTURE PROSPECTS

6.1 Sequence comparisons and the evolution of the arom multifunctional enzyme

The main justification for the work described in this thesis is that it will eventually allow the comparison of the sequence of a multifunctional enzyme with the sequences of its monofunctional counterparts, and thus permit conclusions to be drawn about the mode of evolution of the multifunctional enzyme.

The work described here has contributed the amino acid sequence of *E.coli* shikimate dehydrogenase, one of the monofunctional *E.coli* enzymes which correspond to the five activities of the *N.crassa* pentafunctional *arom* enzyme. The sequences of the other four *E.coli* enzymes have also been obtained in this laboratory (see Section 1.8). The *ARO1* gene of *S.cerevisiae*, which almost certainly encodes a pentafunctional *arom* enzyme similar to that of *N.crassa*, is presently being sequenced in this laboratory (K. Duncan, personal communication; see Section 1.4.3C). This work is well advanced and it should be possible to make sequence comparisons very soon. As described in Chapter 1, the finding of homologies between a multifunctional enzyme and its monofunctional counterparts is good evidence that the multifunctional version arose by gene fusion. It will be interesting to see if the order of activities within the *S.cerevisiae arom* polypeptide is the

same as that found in the N.crassa arom polypeptide. Ultimately it should be feasible to extend the comparison to the relevant plant enzymes, in particular the bifunctional plant E2/E3 enzyme (see below).

It is possible that one or more of the arom activities will have such slight homology to the appropriate E.coli enzyme that it is not possible to distinguish, from the sequence data alone, between chance similarity and common ancestry. If this should occur for E2 and/or E3 then assessment of the statistical significance of any weak homologies could be aided by comparison with the known qa-2 and qa-3 sequences (Doolittle, 1981). It would also be worth remembering that tertiary structure is often much more highly conserved than amino acid sequence in distantly related proteins (Doolittle, 1981; Phillips et al., 1983). The determination of protein structures by X-ray crystallography is still a major undertaking but one which has become steadily less laborious in recent years. The use of synchrotron sources and position-sensitive detectors has helped in this regard. Gene cloning has permitted the overexpression of many proteins not previously available in amounts sufficient for crystallisation, as was the case for the shikimate pathway enzymes. The determination of the arom structure would be a worthwhile exercise in its own right given the paucity of detailed structural information on multifunctional proteins.

The finding of homology between *arom* and its *E.coli* counterparts, each of the latter being an independent folding unit, would be consistent with the idea that each of the *arom* activities is located on a separate functional domain. However, it is possible that, as with some components of multienzyme complexes which do not function well in isolation, one or more *arom* domains might require some "quaternary" interactions for full activity.

Elucidation of the mode of evolution of the *arom* enzyme will leave open the question of whether there is any adaptive significance in this type of catalytic organisation. This area remains to be fully explored. The cloning of the relevant *E.coli* genes permits - if desired - the precise construction of artificial multifunctional enzymes (chimaeras).

6.2 Exploitation of the cloned aroE gene - present uses and future possibilities

The construction of pIA321, which greatly overproduces *E.coli* E3, was described in Chapter 4. Strains containing this plasmid have been used in the laboratory of J.R. Knowles (personal communication) to provide a rich source of E3 for use (together with overproduced E2) as a coupling enzyme in continuous spectrophotometric assays of E1.

Overproduced *E.coli* E3, purified to homogeneity by the author, has been used by M.S. Campbell and I.D. Hamilton (unpublished results) to raise polyclonal antisera in rabbits. Their preliminary experiments using Western blotting techniques

suggest that the antisera may recognise a band in E.coli K12 extracts, subjected to SDS PAGE, whose mobility is the same as a band of much greater intensity recognised in crude extracts of pIA321//AB2834. These bands have the same mobility as pure overproduced E.coli E3 which also appears to be recognised by the antisera. These antisera against E.coli E3 were made in the hope that they might cross-react (if only very slightly) with the plant E2/E3 enzyme and thus provide a route to cloning the gene for this bifunctional plant protein. It was also hoped that the availability of large amounts of overproduced E3 might permit the use of an "E.coli E3-Sepharose" column for the isolation of low affinity antibodies from a cross-reacting antiserum. These low affinity antibodies could then in turn be used to make an affinity column for the purification of the low abundance plant enzyme. Work on the antisera is still in progress but unfortunately initial results suggest that they do not cross-react significantly with the plant enzyme (M.S. Campbell and I.D. Hamilton, unpublished results).

It is conceivable that antisera against different E.coli shikimate pathway enzymes could be used to hunt for labile multienzyme complexes containing two or more of these enzymes. Antibodies against different enzymes could be bound to different size classes of colloidal gold particles for use in immuno-electron microscopy of sections of E.coli cells. This approach would rely on "coincidence counting" - detection of a higher than expected frequency of particles of different sizes occurring next to each other.

There is a second way in which the aroE gene (and other cloned E.coli shikimate pathway genes) could be used in an effort to clone the analogous plant gene. The aroE gene could be used as a heterologous hybridisation probe, at very low stringency, to screen a plant cDNA or genomic library. If regions of homology are found between aroE and yeast ARO1 then one could also try using as a probe a mixture of synthetic oligonucleotides made against the region most conserved between yeast and bacteria in the hope that this segment is also conserved in plants.

As described in Section 5.11.1, there is evidence from chemical modification studies for a lysine residue in or near the active site of N.crassa arom E3. If such a residue is also present in E.coli E3 then additional evidence of its importance (or not) could be obtained using the technique of site-directed mutagenesis (Winter et al., 1982).

6.3 Cloning of E.coli shikimate pathway genes

Almost all E.coli shikimate pathway genes have now been cloned. Only the gene for the shikimate kinase I isozyme and the mysterious aroI remain unaccounted for. The latter should be clonable by relief of auxotrophy. The discovery that aroA is part of a mixed operon, and the possibility that aroE might also be part of an operon, emphasises the need to examine the other E.coli shikimate pathway genes for similar associations. The possibilities for future work on the putative aroE operon have already been discussed.

In this study the next known region of E.coli DNA sequence lies about 9 kbp upstream from aroE (Meek and Hayward, 1984; D.W. Meek, personal communication). It is perhaps time for a database to be set up with the purpose of collecting and ordering all E.coli DNA sequence data. This would allow workers to send in pieces of sequence which might never otherwise be published. Such a centre could also notify independent groups that only a tiny segment remains unsequenced between their respective "territories". Information could also be held on restriction maps of clones thus permitting the gradual assembly of a physical map of regions of the E.coli genome. The discovery, "post-lac operon", of attenuation, antisense RNA, and mixed operons suggests that the E.coli genome may still contain many surprises. This is not a "simple" organism.

6.4 Renaturation of E3 activity after SDS PAGE

If generally applicable this technique could be used as a rapid screening method for the detection of large subunit m.w. forms of E3 after SDS PAGE of crude cell extracts. Such high m.w. forms would be rather suggestive of multi-functional proteins. Preliminary experiments with partially purified E2/E3 from peas were unsuccessful (M.S. Campbell, unpublished results). Obvious candidate species for further trials are S.cerevisiae, S.pombe, and Euglena gracilis. It would also be tempting to extend the phylogenetic range to Archaebacteria, since this third major class of organisms has not yet been looked at.

REFERENCES

Adams, M.J., Archibald, I.G., Helliwell, J.R., Jenkins, S.E. & White, S.W. (1981) in "Structural studies on molecules of biological interest" (Dodson, G., Glusker, J.P. & Sayre, D.;Eds.) 328-338, Oxford University Press, Oxford.

Adams, M.J., Gover, S., Pickersgill, R.W. & Helliwell, J.R. (1983) Biochem. Soc. Trans 11, 429-435.

Ahmed, S.I. & Giles, N.H. (1969) J. Bacteriol. 99, 231-237.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J.D. (1983) "Molecular Biology of the Cell", Chapter 17, Garland Publishing Inc., New York.

Alton, N.K., Buxton, F., Patel, V., Giles, N.H. & Vapnek, D. (1982) Proc. Nat. Acad. Sci. U.S.A. 79, 1955-1959.

Amersham's "M13 Cloning and Sequencing Handbook" (1983) Amersham International plc, Amersham, U.K.

Amrhein, N., Schab, J. & Steinrucken, H.C. (1980) Naturw. 67, 356-357.

Anton, I.A. & Coggins, J.R. (1983) Biochem. Soc. Trans. 12, 275-276.

Atkins, J.F., Gesteland, R.F., Reid, B.R. & Anderson, C.W. (1979) Cell 18, 1119-1131.

Bachmann, B.J. (1983) Microbiol. Rev. 47, 180-230.

Baillie, A.C., Corbett, J.R., Dowsett, J.R. & McCloskey, P. (1972) Biochem. J. 126, 21P.

Balinsky, D., Dennis, A.W. & Cleland, W.W. (1971) Biochemistry 10, 1947-1952.

Barea, J.L. & Giles, N.H. (1978) Biochim. Biophys. Acta 524, 1-14.

Berlyn, M.B. & Giles, N.H. (1969) J. Bacteriol. 99, 222-230.

Berlyn, M.B., Ahmed, S.I. & Giles, N.H. (1970) J. Bacteriol. 104, 768-774.

Biggin, M.D., Gibson, T.J. & Hong, G.F. (1983) Proc. Nat. Acad. Sci. U.S.A. 80, 3963-3965.

Birnboim, H.C. & Doly, J. (1979) Nucleic Acids Res. 7, 1513-1523.

Bisswanger, H. & Schmincke-Ott, E. (1980) "Multifunctional Proteins", John Wiley & Sons, Chichester.

Bolivar, F. & Backman, K. (1979) Methods Enzymol. 68, 245-267.

Bonner, D.M., DeMoss, J.A. & Mills, S.E. (1965) in "Evolving Genes and Proteins" (Bryson, V. & Vogel, H.J. Eds.) 305-318, Academic Press, New York.

Boocock, M.R. (1983) Ph.D. Thesis, Glasgow University.

Boocock, M.R. & Coggins, J.R. (1983) FEBS Lett. 154, 127-133.

Boudet, A.M. & Lécussan, R. (1974) Planta 119, 71-79.

Bowen, B., Steinberg, J., Laemmli, U.K. & Weintraub, H. (1980) Nuc. Acids Res. 8, 1-20.

Bradford, M.M. (1976) Anal. Biochem. 72, 248-254.

Braun, V., Frenz, J., Hantke, K. & Schaller, K. (1980) J. Bacteriol. 142, 162-168.

Burgoyne, L., Case, M.E. & Giles, N.H. (1969) Biochim. Biophys. Acta 191, 452-462.

Capaldi, R.A. & Vanderkooi, G. (1972) Proc. Nat. Acad. Sci. U.S.A. 69, 930-932.

Carter, P.E., Dunbar, B. & Fothergill, J.E. (1983) Biochem. J. 215, 565-571.

Catcheside, D.E.A. & Storer, P.J. (1984) Neurospora Newsletter Number 31, page 18.

Chaleff, R.S. (1974) J. Gen. Microbiol. 81, 337-355.

Chaudhuri, S. & Coggins, J.R. (1981) Biochem. Soc. Trans. 9, 193P.

Chaudhuri, S. & Coggins, J.R. (1985) Biochem. J. 226, 217-223.

Church, G.M. & Gilbert, W. (1984) Proc. Nat. Acad. Sci. U.S.A. 81, 1991-1995.

Clarke, L. & Carbon, J. (1979) Methods Enzymol. 68, 396-408.

Close, T.J., Christmann, J.L. & Rodriquez, R.L. (1983) Gene 23, 131-136.

Coggins, J.R., Boocock, M.R., Campbell, M.S., Chaudhuri, S., Lambert, J.M., Lewendon, A., Mousdale, D.M. & Smith, D.D.S. (1985) Biochem. Soc. Trans. 13, 299-303.

Comai, L., Sen, L.C. & Stalker, D.M. (1983) Science 221, 370-371.

Dagert, M. & Ehrlich, S.D. (1979) Gene 6, 23-28.

Dansette, P. & Azerad, R. (1974) Biochimie 56, 751-755.

Davies, D.D., Teixeira, A. & Kenworthy, P. (1972) Biochem. J. 127, 335-343.

Davies, W.D. & Davidson, B.E. (1982) Nuc. Acids Res. 10, 4045-4058.

Davis, B.D. (1951) J. Biol. Chem. 191, 315-325.

Davis, B.D. & Weiss, U. (1953) N.-S. Arch. Exper. Path. Pharmakol. 220, 1-15.

Davis, B.J. (1964) Ann. N.Y. Acad. Sci. 121, 404-427.

De Boer, H.A., Comstock, L.J. & Vasser, M. (1983) Proc. Nat. Acad. Sci. U.S.A. 80, 21-25.

De Leeuw, A. (1967) Genetics 56, 554-555.

Devereux, J., Haeberli, P. & Smithies, O. (1984) Nucleic Acids Res. 12, 387-395.

Donahue, T.F., Farabaugh, P.J. & Fink, G.R. (1982) Gene 18, 47-59.

Doolittle, R.F. (1981) Science 214, 149-159.

Dottin, R.P., Manrow, R.E., Fishel, B.R., Aukerman, S.L. & Culleton, J.L. (1979) Methods Enzymol. 68, 513-527.

Dowsett, J.R., Corbett, J.R., Middleton, B. & Tubbs, P.K. (1971) Biochem. J. 123, 23P.

Doy, C.H. & Brown, K.D. (1965) Biochim. Biophys. Acta 104, 377-389.

Dretzen, G., Bellard, M., Sassone-Corsi, P. and Chambon, P. (1981) Anal. Biochem. 112, 295-298.

Duncan, K. (1985) Ph.D. Thesis, University of Glasgow.

Duncan, K. & Coggins, J.R. (1983) Biochem. Soc. Trans. 12, 274-275.

Duncan, K., Lewendon, A. & Coggins, J.R. (1984a) FEBS Lett. 165, 121-127.

Duncan, K., Lewendon, A. & Coggins, J.R. (1984b) FEBS Lett. 170, 59-63.

Efstratiadis, A., Vournakis, J.N., Donis-Keller, H., Chaconas, G., Dougall, D.K. & Kafatos, F.C. (1977) Nucleic Acids Res. 4, 4165-4174.

Ely, B. & Pittard, J. (1979) J. Bacteriol. 138, 933-943.

Fickett, J.W. (1982) Nuc. Acids Res. 10, 5303-5318.

Fiddes, J.C. (1976) J. Mol. Biol. 107, 1-24.

Fleming, T.P., Nahlik, M.S. & McIntosh, M.A. (1983) J. Bacteriol. 156, 1171-1177.

Frost, J.W., Bender, J.L., Kadonaga, J.T. & Knowles, J.R. (1984) Biochemistry 23, 4470-4475.

Fulton, A.B. (1982) Cell 30, 345-347.

Gaertner, F.H. & Cole, K.W. (1977) Biochem. Biophys. Res. Commun. 75, 259-264.

Gentz, R., Langner, A., Chang, A.C.Y., Cohen, S.N. & Bujard, H. (1981) Proc. Nat. Acad. Sci. U.S.A. 78, 4936-4940.

Gibson, F. & Pittard, J. (1968) Bacteriol. Rev. 32, 465-492.

Giles, N.H., Case, M.E., Partridge, C.W.H. & Ahmed, S.I. (1967a) Proc. Nat. Acad. Sci. U.S.A. 58, 1453-1460.

Giles, N.H., Partridge, C.W.H., Ahmed, S.I. & Case, M.E. (1967b) Proc. Nat. Acad. Sci. U.S.A. 58, 1930-1937.

Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B.S. & Stormo, G. (1981) Ann. Rev. Microbiol. 35, 365-403.

Gollub, E., Zalkin, H. & Sprinson, D.B. (1967) J. Biol. Chem. 242, 5323-5328.

Gottesman, M., Oppenheim, A. & Court, D. (1982) Cell 29, 727-728.

Gould, S.J. (1984) in "Hen's Teeth and Horse's Toes", p.160, Penguin Books, Harmondsworth, England.

Gouy, M. & Gautier, C. (1982) Nuc. Acids Res. 10, 7055-7074.

Graziana, A., Boudet, A. & Boudet, A.M. (1980) Plant Cell Physiol. 21, 1163-1174.

Grosjean, H. & Fiers, W. (1982) Gene 18, 199-209.

Gross, S.R. & Fein, A. (1960) Genetics 45, 885-904.

Hagervall, T.G. & Bjork, G.R. (1984) Mol. Gen. Genetics 196, 201-207.

Hantke, K. & Braun, V. (1975) FEBS Lett. 49, 301-305.

Hasan, N. & Nester, E.W. (1978a) J. Biol. Chem. 253, 4987-4992.

Hasan, N. & Nester, E.W. (1978b) J. Biol. Chem. 253, 4999-5004.

Haslam, E. (1974) "The Shikimate Pathway", Butterworths, London.

Hautala, J.A., Jacobson, J.W., Case, M.E. & Giles, N.H. (1975) J. Biol. Chem. 250, 6008-6014.

Hawley, D.K. & McClure, W.R. (1983) Nucleic Acids Res. 11, 2237-2255.

Herrmann, K.M. (1983) in "Amino Acids: Biosynthesis and Genetic Regulation" (Herrmann, K.M. & Somerville, R.L.) 301-322, Addison-Wesley, London.

Hirs, C.H.W. (1967) Methods in Enzymol. 11, 197-199.

Holmes, D.S. & Quigley, M. (1981) Anal. Biochem. 114, 193-197.

Holmes, W.M., Platt, T. & Rosenberg, M. (1983) Cell 32, 1029-1032.

Hoog, J.-O., Jornvall, H., Holmgren, A., Carlquist, M. & Persson, M. (1983) Eur. J. Biochem. 136, 223-232.

Huiet, L. (1984) Proc. Nat. Acad. Sci. U.S.A. 81, 1174-1178.

Jaskunas, S.R., Lindahl, L. & Nomura, M. (1975a) Proc. Nat. Acad. Sci. U.S.A. 72, 6-10.

Jaskunas, S.R., Burgess, R.R. & Nomura, M. (1975b) Proc. Nat. Acad. Sci. U.S.A. 72, 5036-5040.

Katinka, M., Cossart, P., Sibilli, L., Saint-Girons, I., Chalvignac, M.A., Le Bras, G., Cohen, G.N. & Yaniv, M. (1980) Proc. Nat. Acad. Sci. U.S.A. 77, 5730-5733.

Kinghorn, J.R., Schweizer, M., Giles, N.H. & Kushner, S.R. (1981) Gene 14, 73-80.

Kinghorn, J.R. & Hawkins, A.R. (1982) Mol. Gen. Genet. 186, 145-152.

Kirschner, K. & Bisswanger, H. (1976) Ann. Rev. Biochem. 45, 143-166.

Klintrot, A.-M., Hoog, J.-O., Jornvall, H., Holmgren, A. & Luthman, M. (1984) Eur. J. Biochem. 144, 417-423.

Knoell, H.-E. & Knappe, J. (1974) Eur. J. Biochem. 50, 245-252.

Knoell, H.-E., Kraft, R. & Knappe, J. (1978) Eur. J. Biochem. 90, 107-112.

Koebner, R.M.D. & Shepherd, K.W. (1982) Genet. Res. 41, 209-213.

Koshiba, T. (1979) Plant Cell Physiol. 20, 667-670.

Kozak, M. (1983) Microbiol. Rev. 47, 1-45.

Laemmli, U.K. (1970) Nature 227, 680-685.

Lambert, J.M., Boocock, M.R. & Coggins, J.R. (1985) Biochem. J. 226, 817-829.

Larimer, F.W., Morse, C.C., Beck, A.K., Cole, K.W. & Gaertner, F.H. (1983) Molec. & Cellular Biol. 3, 1609-1614.

Levin, J.G. & Sprinson, D.B. (1964) J. Biol. Chem. 239, 1142-1150.

Lewendon, A. & Coggins, J.R. (1983) Biochem. J. 213, 187-191.

Lewin, B. (1983) "Genes", p.240, John Wiley & Sons, New York.

Llewellyn, D.J., Daday, A. & Smith, G.D. (1980) J. Biol. Chem. 255, 2077-2084.

Lovenberg, W. (1973) "Iron-Sulfur Proteins" Volumes I-III, Academic Press, London.

Lumsden, J. & Coggins, J.R. (1977) Biochem. J. 161, 599-607.

Lumsden, J. & Coggins, J.R. (1978) Biochem. J. 169, 441-444.

Maitra, U.S. & Sprinson, D.B. (1978) J. Biol. Chem. 253, 5426-5430.

Maniatis, T., Fritsch, E.F. & Sambrook, J. (1982) "Molecular Cloning: A Laboratory Manual", Cold Spring Harbor Laboratory.

Maxam, A.M. & Gilbert, W. (1977) Proc. Nat. Acad. Sci. U.S.A. 74, 560-564.

McCarthy, A.D. & Hardie, D.G. (1984) Trends Biochem. Sci. 9, 60-63.

Meek, D.W. & Hayward, R.S. (1984) Nucleic Acids Res. 12, 5813-5821.

Messing, J. & Vieira, J. (1982) Gene 19, 269-276.

Messing, J. (1983) Methods Enzymol. 101, 20-78.

Millar, G., Anton, I.A., Mousdale, D., White, P.J. & Coggins, J.R. (1985) Biochem. Soc. Trans., in press.

Miozzari, G.F. & Yanofsky, C. (1979) Nature 277, 486-489.

Morell, H., Clark, M.J., Knowles, P.F. & Sprinson, D.B. (1967) J. Biol. Chem. 242, 82-90.

Mousdale, D.M. & Coggins, J.R. (1984) Planta 160, 78-83.

Nakanishi, N. & Yamamoto, M. (1984) Mol. Gen. Genetics 195, 164-169.

Neidhardt, F.C., Vaughn, V., Phillips, T.A. & Bloch, P.L. (1983) Microbiol. Rev. 47, 231-284.

Neilands, J.B. (1982) Ann. Rev. Microbiol. 36, 285-309.

Nimmo, G.A. & Coggins, J.R. (1981) Biochem. J. 197, 427-436.

Ogino, T., Garner, C., Markley, J.L. & Herrmann, K.M. (1982) Proc. Nat. Acad. Sci. U.S.A. 79, 5828-5832.

Old, R.W. & Primrose, S.B. (1981) "Principles of Gene Manipulation" (2nd Edition), p.47 & p.182, Blackwell Scientific Publications, London.

Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G. & Steitz, T.A. (1985) Nature 313, 762-766.

Patel, V.B. & Giles, N.H. (1979) Biochim. Biophys. Acta 567, 24-34.

Pestka, S., Daugherty, B.L., Jung, V., Hotta, K. & Pestka, R.K. (1984) Proc. Nat. Acad. Sci. U.S.A. 81, 7525-7528.

Phillips, D.C., Sternberg, M.J.E. & Sutton, B.J. (1983) in "Evolution from Molecules to Men" (Bendall, D.S. - Ed.) 145-173, Cambridge University Press, Cambridge.

Pittard, J. & Wallace, B.J. (1966) J. Bacteriol. 91, 1494-1508.

Polley, L.D. (1978) Biochim. Biophys. Acta 526, 259-266.

Pribnow, D. (1975) Proc. Nat. Acad. Sci. U.S.A. 72, 784-788.

Rigby, P.W.J. & Lane, D.P. (1983) Adv. Viral Oncology 3, 31-57.

Rines, H.W., Case, M.E. & Giles, N.H. (1969) Genetics 61, 789-800.

Rosenberg, M. & Court, D. (1979) Ann. Rev. Genet. 13, 319-353.

Rosenthal, A.L. & Lacks, S.A. (1977) Anal. Biochem. 80, 76-90.

Rossmann, M.G., Liljas, A., Branden, C.-I. & Banaszak, L.J. (1975) in "The Enzymes" 11, 61-102 (P.D. Boyer, Ed.) Academic Press, London.

Rossmann, M.G. & Argos, P. (1981) Ann. Rev. Biochem. 50, 497-532.

Rossmann, M.G. (1983) Hoppe-Seyler's Z. Physiol. Chem. 364, 193.

Ruther, U. & Muller-Hill, B. (1983) EMBO J. 2, 1791-1794.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977) Proc. Nat. Acad. Sci. U.S.A. 74, 5463-5467.

Schaller, K., Krauel, A. & Braun, V. (1981) J. Bacteriol. 147, 135-139.

Schmincke-Ott, E. & Bisswanger, H. (1980) in "Multifunctional Proteins" (Bisswanger, H. & Schmincke-Ott, E., Eds.) 1-29, John Wiley & Sons, Chichester.

Schweizer, M., Case, M.E., Dykstra, C.C., Giles, N.H. & Kushner, S.R. (1981) Proc. Nat. Acad. Sci. U.S.A. 78, 5086-5090.

Shine, J. & Dalgarno, L. (1975) Nature 254, 34-38.

Shultz, J., Hermodson, M.A., Garner, C.C. & Herrmann, K.M. (1984) J. Biol. Chem. 259, 9655-9661.

Smith, M.A., Gerrie, L.M., Dunbar, B. & Fothergill, J.E. (1982) Biochem. J. 207, 253-260.

Smith, D.D.S. & Coggins, J.R. (1983) Biochem. J. 213, 405-415.

Staden, R. (1978) Nucleic Acids Res. 5, 1013-1015.

Staden, R. (1980) Nucleic Acids Res. 8, 3673-3694.

Stormo, G.D., Schneider, T.D. & Gold, L.M. (1982) Nucleic Acids Res. 10, 2971-2996.

Strauss, A. (1979) Mol. Gen. Genet. 172, 233-241.

Stryer, L. (1981) "Biochemistry", W.H. Freeman & Co., San Francisco.

Stuber, D. & Bujard, H. (1981) Proc. Nat. Acad. Sci. U.S.A. 78, 167-171.

Sutcliffe, J.G. (1979) Cold Spring Harbor Symp. Quant. Biol. 43, 77-90.

Thomas, P.S. (1980) Proc. Nat. Acad. Sci. U.S.A. 77, 5201-5205.

Tribe, D.E., Camakaris, H. & Pittard, J. (1976) J. Bacteriol. 127, 1085-1097.

Tsunoda, J.N. & Yasunobu, K.T. (1968) J. Biol. Chem. 243, 6262-6272.

Twigg, A.J. & Sherratt, D. (1980) Nature 283, 216-218.

Umbarger, H.E. (1978) Ann. Rev. Biochem. 47, 533-606.

Vapnek, D., Hautala, J.A., Jacobson, J.W., Giles, N.H. & Kushner, S.R. (1977) Proc. Nat. Acad. Sci. U.S.A. 74, 3508-3512.

Volz, K.W., Matthews, D.A., Alden, R.A., Freer, S.T., Hansch, C., Kaufman, B.T. & Kraut, J. (1982) J. Biol. Chem. 257, 2528-2536.

Walker, J.E., Saraste, M., Runswick, M.J. & Gay, N.J. (1982) EMBO J. 1, 945-951.

Watson, M.E.E. (1984) Nuc. Acids. Res. 12, 5145-5164.

Weber, K. & Kuter, D.J. (1971) J. Biol. Chem. 246, 4504-4509.

Weber, K. & Osborn, M. (1969) J. Biol. Chem. 244, 4406-4412.

Weiss, U. & Edwards, J.M. (1980) "The Biosynthesis of Aromatic Compounds", John Wiley and Sons, New York.

Winter, G., Fersht, A.R., Wilkinson, A.J., Zoller, M. & Smith, M. (1982) Nature 299, 756-758.

Yamamoto, E. (1980) Phytochemistry 19, 779-781.

Yaniv, H. & Gilvarg, C. (1955) J. Biol. Chem. 213, 787-795.

Yasunobu, K.T. & Tanaka, M. (1973) in "Iron-Sulfur Proteins, Volume II, Molecular Properties" (Lovenberg, W.) 27-130, Academic Press, London.

Yoch, D.C. & Carithers, R.P. (1979) Microbiol. Rev. 43, 384-421.

Zalkin, H. & Yanofsky, C. (1982) J. Biol. Chem. 257, 1491-1500.

Zalkin, H., Paluh, J.L., Cleemput, M., Moye, W.S. & Yanofsky, C. (1984) J. Biol. Chem. 259, 3985-3992.

Zurawski, G., Brown, K., Killingly, D. & Yanofsky, C. (1978) Proc. Nat. Acad. Sci. U.S.A. 75, 4271-4275.

Zurawski, G., Gunsalus, R.P., Brown, K.D. & Yanofsky, C. (1981) J. Mol. Biol. 145, 47-73.