THE D-ARABITOL OPERON OF *KLEBSIELLA AEROGENES*

by

Timothy John Knott

A dissertation submitted to the University of London in candidature for the degree of Doctor of Philosophy

Department of Biochemistry
Imperial College
University of London
London SW7 2AZ

October 1982

T.J. Knott

The D-Arabitol Operon of Klebsiella aerogenes

ABSTRACT

This dissertation describes the investigation, by DNA sequence analysis, of the d-arabitol operon of Klebsiella aerogenes. The starting material for this study was the specialised transducing phage λ p rbtdal, which carries both of the K.aerogenes pentitol operons. Experiments are described in which sub-clones of the operon are constructed using plasmid and phage vectors, and subsequently sequenced by the partial chemical degradation or dideoxy chain termination methods.

The sequence of a region of DNA which controls the expression of the two structural genes, d-arabitol dehydrogenase (ArDH) and d-xylulokinase (DXK), is presented together with preliminary data relating to the regulation of transcription from the dalDK promoter.

The complete amino acid sequence of ArDH has been derived from the DNA sequence and aspects of its primary and secondary structure are considered. The enzyme is compared to ribitol dehydrogenase from the closely linked ribitol operon and the bearing of the results on the speculative theories of invert gene duplication, believed to have been important in the natural evolution of the pentitol operons, discussed. Partial sequences for the DXK gene are also given.

The nucleotide sequence of the d-arabitol operon repressor gene has been obtained and parts of the predicted protein sequence are shown to be similar, at both the amino acid and secondary structure levels, to the operator-recognition regions of numerous other DNA-binding regulatory proteins. A common mechanism for protein-DNA interactions is therefore envisaged.

PREFACE

I should like to take this opportunity to thank my supervisor, Professor Brian Hartley, for his help in the form of support, encouragement and seemingly unshakeable optimism. Thanks are also due to :

Mrs. Thérèse Anderton for teaching me the necessary skills in nucleic acid sequencing, and for invaluable assistance throughout this project.

Dr. Michael Neuberger for introducing me to the basics of microbiology, nucleic acid biochemistry, cloning and constructive criticism of parts of the manuscript. I am grateful also to Dr. Mark Payton for 'volunteering' to read the remainder.

Drs. Andrew McLachlan and Alistair Bingham for help with the computer analysis of protein sequences, and Mr. Stuart Cox for rewriting Roger Staden's programme for compatibility with the Imperial College computer.

Mr. Glyn Millhouse for providing an excellent photographic service, and the staff of the Biochemistry Department workshop for the skilful manufacture of some of the specialist equipment used in this work.

Dr. Willhelm Ansorge and the European Molecular Biology Organisation for the opportunity to learn the techniques of electrophoresis on ultra-thin gels at laboratories in Heidelberg.

Finally, I would like to express my gratitude to Miss Deanne Eastwood for a rapid, accurate typing of this dissertation and for coping admirably with the vagaries of my handwriting.

Abbreviations and Symbols

Most are standard but attention is drawn to the following :-

Ap^R	Ampicillin resistance
APS	Ammonium persulphate
ArDH	Arabitol dehydrogenase
β·ME	β -mercaptoethanol
BAP	Bacterial alkaline phosphatase
bp	Base pair
BPB	Bromophenol Blue dye
BSA	Bovine serum albumin
CAA	Casamino acids (casein hydrolysate)
cAMP	Adenosine 3':5' cyclic monophosphate
cccDNA	Covalently closed circular DNA
cpm	Counts per minute
CTAB	Cetyltrimethlyammonium bromide
$d \cdot H_2O$	Deionised, distilled water
DEAE	Diethyl-aminoethyl ...
DNAase	Deoxyribonuclease
DRK	D-ribulokinase
dsDNA	Double-stranded DNA
DTT	Dithiothreitol
DXK	D-xylulokinase
EDTA	Ethylene diamine tetraacetic acid; disodium salt
EtBr	Ethidium Bromide
kb	1000 base-pairs, 1 kilobase
K_m	Michaelis-Menten constant
mRNA	Messenger RNA
NAD^+	β -nicotinamide adenine dinucleotide
NADH	β -nicotinamide adenine dinucleotide (reduced form)

OD_n	Optical density at wavelength n
PEG	Polyethylene glycol
Pu	Purine base (in DNA)
Py	Pyrimidine base (in DNA)
RDH	Ribitol dehydrogenase
rRNA	Ribosomal RNA
RNAase	Ribonuclease
SDS	Sodium dodecyl sulphate
ssDNA	Single-stranded DNA
Tc^R	Tetracycline resistance
TCA	Trichloroacetic acid
TdT	Terminal deoxynucleotidyl Transferase
TEMED	N,N,N',N' Tetramethyl-ethylene-diamine
Tris	Tris (hydroxymethyl) aminoethane
tRNA	Transfer RNA
uv	Ultra-violet
XC	Xylene cyanol FF dye

Genetic symbols

dalD, rbtD	ArDH and RDH structural genes
dalK, rbtK	d-xylulokinase and d-ribulokinase structural genes
dalR, rbtR	Repressor genes for d-arabitol and ribitol operons
dalDK p/o	d-arabitol operon promoter/operator region
rbtDK p/o	Ribitol operon promoter/operator region
rbt^c, dal^c	Constitutively expressed ribitol or d-arabitol operons.

Nomenclature

The nomenclature for restriction endonucleases is that of Roberts (1980). Genetic symbols are in accordance with Bachmann and Brooks-Low (1980) for E.coli K12 chromosome markers, and Hershey (1971) for λ genomic markers. A description of standard phenotypic designations may be found in Miller (1972).

Nucleoside triphosphates are assigned the prefix r, d or dd to indicate a ribose, deoxyribose or dideoxyribose sugar moiety.

The convention of Neuberger and Hartley (1979) has been retained for specifying restriction sites within cloned K. aerogenes DNA. Target sites for enzymes bear the suffix A, B, C, etc. (For example, BstA or HindB), and lettering is from left to right along the length of λ p rbtdal.

Plasmids bear the prefix RD to show that the insert is derived from λp rbtdal. Of the three numbers which follow, the first signifies the vector (1 = pAT153, 2 = pBR322, 3 = pBR313), and the next two represent the clone number. Clones from the dal operon are numbered from 50 upwards.

CONTENTS

List of Figures

List of Tables

List of Photographs

CHAPTER 1

INTRODUCTION

Molecular cloning of recombinant DNA has become an important tool in the study of prokaryotic and eukaryotic biology since reports first appeared in the literature (Cohen et al, 1973). There are five basic requirements for the propagation of foreign DNA in bacteria : 1) a vector capable of replicating within the recipient organism, 2) a means of cleaving DNA specifically and reproducibly, 3) a method for splicing DNA segments to the chosen cloning vector, 4) a procedure for introducing the composite molecule into recipient cells, and 5) a way of selecting those cells that have acquired the hybrid DNA species. The commonest vectors are derivatives of the plasmids pSC101 (Cohen et al, 1973) or Col E1 (Hershfield et al, 1974). Bacteria harbouring recombinant plasmids are generally identified through insertional inactivation of an antibiotic resistance marker on the vector. The use of plasmid and bacteriophage vectors permits the fractionation of individual DNA fragments from a more complex genome by their uptake into single cells, the amplification of those components and the opportunity to study their control mechanisms and expression.

Sufficient material may be obtained to permit analysis of the nucleotide sequence by one of several rapid DNA sequencing techniques currently available.

Examination of DNA sequences reveals much about the cellular mechanisms which are involved in the initiation and regulation of RNA and protein synthesis. Comparisons of gene

sequences derived from two or more independent sources can provide information on their probable evolutionary origins. The degree of sequence homology between two genes coding for the same protein not only reflects evolutionary relationships, but it can also give an important insight into the involvement of mutations in enzyme evolution and the way in which new specifities arise. A model system for such an investigation exists in this laboratory in the form of the pentitol operons of Klebsiella aerogenes.

Pentitol catabolism

The utilisation of pentitols by bacteria has been extensively studied over a number of years. Early work (Wood and Tai, 1958) established the presence of genes in Aerobacter aerogenes which allow the catabolism of two of the more common pentitols, ribitol and d-arabitol. Later experiments (Mortlock et al, 1965; Wood et al, 1961) were aimed at characterising the pathways and enzymes involved. Both ribitol and d-arabitol are initially oxidised to their 2-keto sugars by distinct dehydrogenases. The pentuloses are then phosphorylated, each by a different pentulokinase, to yield d-ribulose-5-phosphate and d-xylulose-5-phosphate (Fig. 1a) which can act as substrates for either the phosphogluconate pathway or nucleic acid biosynthesis.

Initial studies on the inducible Arabitol Dehydrogenase (ArDH) from A.aerogenes (Lin, 1961) established that among the possible physiological inducers only d-arabitol was effective to any degree. In addition, it was demonstrated that mannitol, although a good substrate for ArDH, is not an inducer of the enzyme. ArDH has no side specifity for any other pentitols. Synthesis of the A.aerogenes enzyme is 100% catabolite repressed during growth on glucose and to a lesser

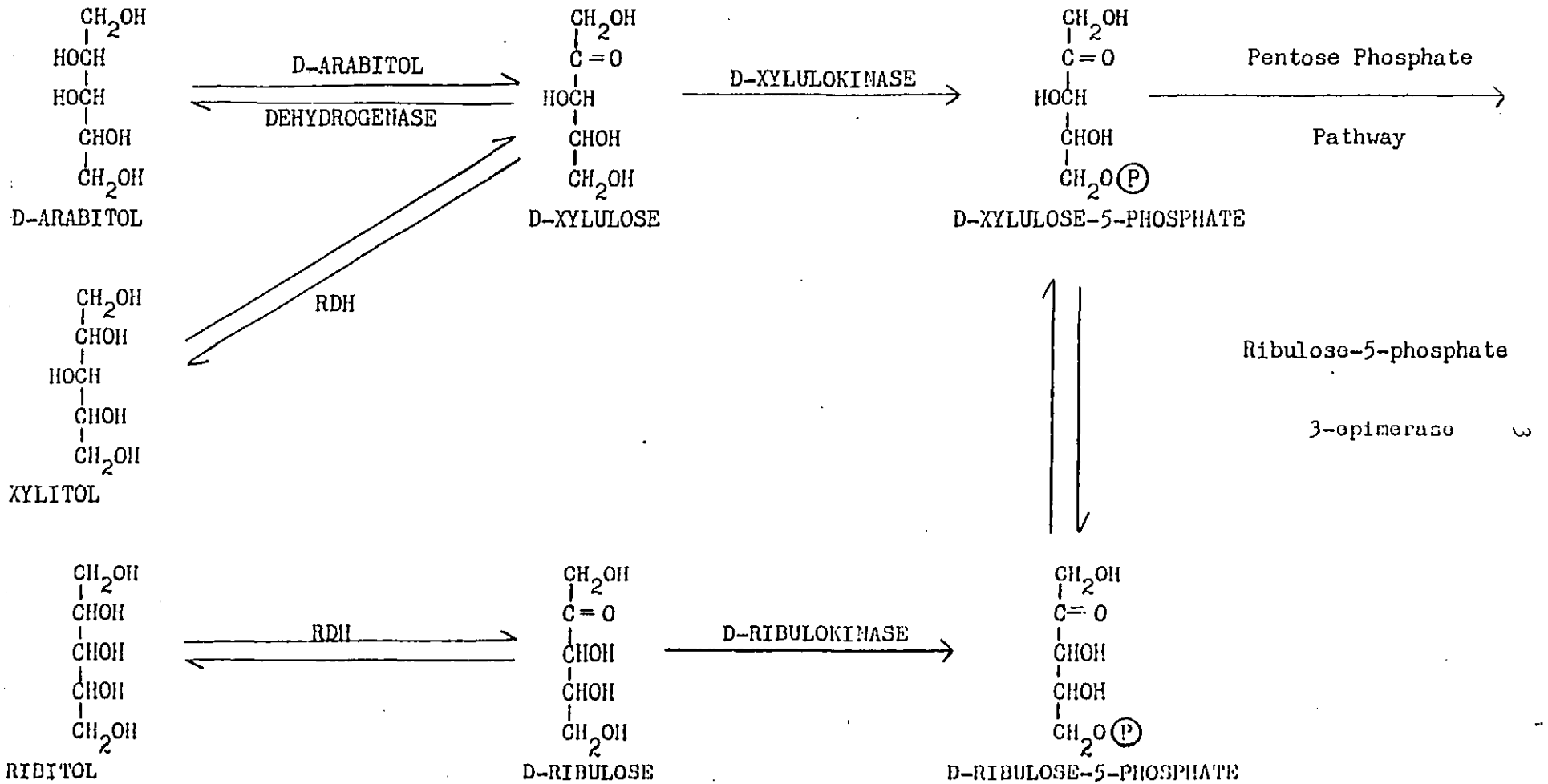


Fig. 1a Pentitol Catabolism in Klebsiella aerogenes

degree in the presence of mannitol (Lin, 1961).

An active transport system for d-arabitol was identified (Wu et al, 1968) which can also transport xylitol, enabling RDH constitutive mutants to take up xylitol and use it as a sole source of carbon and energy. It was noted that growth on xylitol leads to induction of the d-arabitol pathway since ArDH will catalyse the conversion of d-xylulose to d-arabitol if intracellular levels of this pentulose rise sufficiently.

Much of the early work on pentitol metabolism employed strains of A.aerogenes. Klebsiella aerogenes W70 was chosen in later experiments following a report of a transduction system using bacteriophage PW52, (McPhee et al, 1969). The two bacteria are very similar in their metabolism of the naturally occurring C_5 sugars, (Charnetzky and Mortlock, 1974a and 1974b), except that no L-xylulokinase activity is demonstrable in K.aerogenes. A change in nomenclature has meant that both are now classified as Klebsiella strains.

Genetic linkage of the ribitol and d-arabitol pathways of K.aerogenes

The close genetic linkage of the two pathways has been demonstrated by Charnetzky and Mortlock (1974c). Transduction experiments with phage PW52 indicated that rbt^+ and dal^+ phenotypes are more than 85% co-transducible. This caused speculation that the regulatory and catalytic functions for growth on ribitol and d-arabitol are encoded on a continuous stretch of DNA without any intervening genetic loci. In addition to dehydrogenase and kinase activities, two control sites were identified for each pathway and were tentatively assigned operator and repressor functions. The gene order for the region of the K.aerogenes genome essential to pentitol cata-

bolism was deduced by a series of genetic crosses (Charnetzky and Mortlock, 1974c). The sequence derived is rbtK.rbtD.rbtC. dalB.rbtB.dalC.dalD.dalK where K and D represent kinase and dehydrogenase genes respectively. B and C are control loci. A similar clustering of the genes for pentitol catabolism is found in E.coli C (Reiner, 1975). Here too the genetic elements of both pathways are closely linked, have mirror-image symmetry and are 85% co-transducible.

The Growth of K.aerogenes on Pentitols

The enterobacterium K.aerogenes is able to utilise the two most common pentitols as sole carbon sources. Ribitol and d-arabitol are both relatively abundant in nature whereas l-arabitol and xylitol occur less frequently. Several other organisms are known to be capable of growth on pentitols including some Pseudomonas and yeast species and E.coli C, but not E.coli K12 or B strains. The pathway for catabolism of these substrates is depicted in Fig. 1a.

In wild-type strains the synthesis of the necessary enzymes is coordinately controlled (Charnetzky and Mortlock, 1974a). Mutants selected for constitutive production of RDH are also constitutive for DRK, although DRK is not essential for growth on ribitol. Both of these enzymes are made in significant quantities only in the presence of their inducer, d-ribulose (Charnetzky and Mortlock, 1974a). Mutants defective in RDH synthesis cannot catalyse the oxidation of ribitol to d-ribulose and are thus unable to induce DRK activity. Similarly, ArDH and DXK are synthesised coordinately but the inducer is the initial substrate d-arabitol, (Wilson and Mortlock, 1973). In view of the close chemical and structural similarity of the intermediates in each pathway the choice of inducers may

be interpreted as being essential to achieve completely independent control. The K.aerogenes ribitol operon (rbt) and d-arabitol operon (dal) are both under negative control as are those of E. coli C (Neuberger and Hartley, 1979; Scangos and Reiner, 1979).

Mutant strains constitutive for ArDH can grow on d-mannitol using a side specificity of this enzyme which results in the formation of d-fructose. The V_{max} for this substrate is 45% of that for d-arabitol (Lin, 1961). Although K.aerogenes cannot normally grow on xylitol alone, mutants can be selected which are constitutive for rbt expression. Such mutants are able to grow on xylitol, but their mean doubling time is more than twice that on ribitol. RDH has a side specificity for xylitol (Km xylitol = 0.5M-1.0M, Km ribitol = 5 mM: Rigby et al, 1974), converting it to d-xylulose which is then phosphorylated by d-xylulokinase from the dal operon or the xylose operon; d-xylulose is capable of inducing the xylose operon (Neuberger and Hartley, 1979). If selective pressure is applied in the form of continuous culture in a chemostat, new phenotypes evolve which display faster growth rates on this novel substrate (Rigby, 1971). Such mutants fall into two classes : those making an RDH with a lower Km for xylitol, and strains carrying duplicate copies of all or part of the rbt operon (Rigby et al, 1974; Neuberger and Hartley, 1981). The wide range of mutants isolated from those chemostat experiments have formed the basis of an extensive project aimed ultimately at increasing our understanding of the mechanisms of gene duplication and enzyme evolution in prokaryotes.

Transfer of the Pentitol Genes into E.coli K12 and Phage λ

A detailed analysis of the K.aerogenes pentitol operons

in situ is complicated by the lack of available information regarding the genetic background of this bacterium. Rigby *et al* (1976) succeeded in moving the entire region into a more defined host, namely *E.coli* K12. Strains of *Klebsiella* sensitive to the generalised transducing phage P1 CM clr 100 were isolated. Lysates of infected cultures were used to transduce *E.coli* K12, resulting in *E.coli*/*K.aerogenes* hybrids capable of growth on ribitol and d-arabitol. Interestingly enough, the transferred genes mapped at position 40' on the K12 chromosome, precisely the position that they occupy on the *E.coli* C genome. Subsequently this hybrid (NC100) has been used in the construction of two specialised lambdoid transducing phages (Neuberger, 1978; Neuberger and Hartley, 1979). λp *rbt* and λp *rbt dal* carry genes for ribitol and ribitol + d-arabitol metabolism respectively. *E.coli* K12 lysogens of these phages acquire the ability to grow on pentitols as their sole carbon sources.

Construction of the recombinant phages required P1 transduction of the constitutive *rbt* operon from NC100 into an *E. coli* strain deleted for the λ attachment site. A transductant, NC 596, having the genotype *rbt*-101 *rbtD*$^+$ *rbtK*$^+$ *dal*$^+$ (*gal att*λ *bio*) was isolated. Infection of this strain with the thermoinducible phage λ627 produced lysogens which, upon induction of the prophage and re-infection into a new host, gave only rbt$^+$ phenotypes. It was, however, relatively straightforward using this phage (λ *prbt*) to derive a secondary transducing phage which carries both the *rbt* and the *dal* operons. λ *prbt dal* is thought to have arisen via *rec* mediated recombination between NC 596 chromosomal DNA and λ *prbt* followed by imperfect excision of the λ *prbt* prophage (Neuberger, 1978).

The transcribed regions of both operons have been localised by hybridisation of labelled mRNAs to restriction fragments of λ prbt and λ p rbt dal. In addition, hybridising dal or rbt-specific messengers to strand separated phage DNA revealed that transcription is bipolar (Fig. 1b) and occurs from opposite strands of the DNA (Neuberger and Hartley, 1979). Hence the possibility exists that a gene duplication and subsequent inversion event preceded the evolution of the pentitol operons in K.aerogenes.

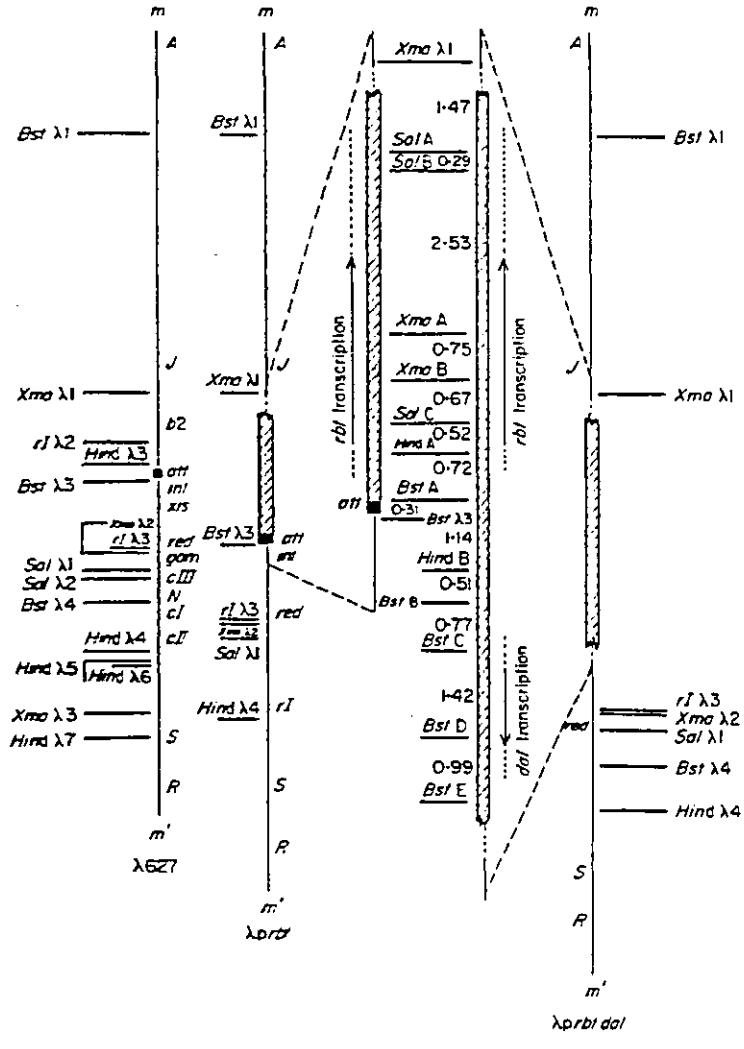
Operon Structure and the Regulation of Protein Synthesis in Prokaryotes

Most of the major concepts on operon structure, expression and regulation have arisen out of work with the E.coli lac operon. The choice of this particular model stems from the early pioneering work of Jacob and Monod (1961) on induced enzyme synthesis.

Construction of the specialised transducing phages λ lac and φ80 lac (Beckwith and Signer, 1966) was one of the first steps leading to a detailed genetic analysis of the lac operon. Similarly, the availability of F' hut and λ phut has been essential to investigations of the E.coli histidine utilisation system (Smith, 1971). λ p rbt and λ p rbt dal (Neuberger and Hartley, 1979) have proved to be equally useful in studying the pentitol operons.

Our current view of an operon is one of a contiguous set of genes coding for a group of enzymes which catalyse successive reactions in a multistep pathway. In the presence of an inducer, usually an intermediate of the pathway or the initial substrate itself, the rates of synthesis of all the enzymes required for metabolism of that substrate are increased. This stimulation

Fig. 1b Physical Structure of λprbt and λprbtdal



K.aerogenes DNA is represented as (▨). The directions of rbt and dal transcription are indicated together with the target sites for some of the major restriction endonucleases. All length measurements are given in kb.

(Diagram taken from Neuberger and Hartley,1979)

may be up to 1000-fold for the lac Z gene product.

The relative positions of structural genes within an operon need not reflect the order in which each product acts in the pathway. This is certainly true of the his operon. (Kasai, 1974).

The Control of Transcription Initiation

Three genetic elements are fundamental to the coordinate control of enzyme synthesis. These are the repressor protein and the promoter and operator sequences. The promoter is a region of DNA required for the initiation of transcription by RNA Polymerase . Early work on prokaryote transcription concentrated mainly on the lac operon (Miller, 1970) and the DNA sequence of the entire control region was later determined by Dickson et al (1975). The precise nucleotide sequence of a great number of individual promoter/operator sites have now been derived (see Siebenlist (1980) for a recent review and comparative study). Promoter structure is considered in detail in later chapters.

The operon structural genes form a single transcriptional unit initiated at the promoter. Mutations in this region impose a coordinate change in the levels of synthesis of all the operon products (Scaife and Beckwith, 1966; Reznikoff, 1972). Mutations in the repressor gene or operator generally result in constitutive synthesis (Jacob and Monod, 1961). Control of transcription initiation is mediated via repressor/DNA interactions at specific sequences in the operator site (Dickson et al, 1975; Ptashne et al, 1976; Gilbert and Müller-Hill, 1970; Otsuka and Abelson, 1978; Ogden et al, 1980). It is also dependent on interactions of RNA Polymerase with well defined promoter sequences (Pribnow,1975) and, in those operons subject to catabolite repression, the binding of a catabolite repressor

protein-cAMP complex is also a requirement for efficient mRNA synthesis (Reznikoff and Abelson, 1978). Catabolite repression is generally exerted over those pathways geared to energy production from carbohydrates other than glucose. Glucose constitutes a more efficient carbon source than most other sugars or sugar-alcohols, and growth on this substrate results in a lowering of intracellular cAMP levels (Makman and Sutherland, 1965). Hence the activation and binding of CRP to the promoter is prevented, an event which is essential for the formation of transcription initiation complexes in cAMP dependent operons (Willmund and Kneser, 1973). The exact mechanism by which CRP is able to bring about an increase in RNA Polymerase binding has long baffled researchers. Two models currently exist. According to the first, CRP destabilises the DNA helix to permit the establishment of an initiation complex between the polymerase and template. The alternative view is that protein-protein contacts between DNA-bound CRP and RNA Polymerase are needed before initiation can take place (Gilbert, 1976). Support for the first model is convincing: CRP can alter the confirmation of DNA as judged by enhancement of DNAase I cleavage. Helix-destabilising agents like glycerol or dimethyl sulphoxide facilitate the interaction of CRP with DNA and can partially replace it as an activator of transcription (Schmitz, 1981). CRP presumably promotes local melting of the DNA double helix, an essential step in the transition from a closed to open RNA Polymerase-promoter complex (Dickson et al, 1975; Chamberlin, 1976). Nakanishi et al (1975) have shown that cAMP.CRP shifts the transition temperature for open complex formation by about 10°C, mimicking the effects of glycerol or low Mg^{++} concentration. A detailed

molecular mechanism for the denaturant action of cAMP.CRP is provided by Ebright and Wong (1981). Following recognition of a specific base sequence (see Queen and Rosenberg, 1981) the adenine moiety exposed by the complex is inserted into the minor groove of the helix, thus breaking an A.T base pair and enhancing local denaturation which propagates to adjoining sites leading to an open complex. Initiation by RNA Polymerase at a site 15-100 base pairs downstream is enhanced.

Transcriptional Control, Attenuation and Polarity

The control of expression of most microbial genes is exerted by enhancement or inhibition of the rate of transcriptional initiation. The original view of the operon as a transcriptional unit where RNA synthesis, once started, continues through to the promoter-distal end has changed in recent years with the discovery that transcription may be terminated by specific signals distributed throughout the entire operon (de Crombrugghe et al, 1973). It is now evident that transcriptional control can be used to regulate the expression of genes distal to a termination signal.

Where polarity exists, the result is an inequality in the ratios of enzymes synthesised from a single polycistronic mRNA. The differences arise out of ribosomal preferences for certain initiation sites and the mechanism of degradation of labile bacterial messengers (Brown et al, 1967; Platt, 1978). In addition to the controls listed above, a further more complex regime may operate in certain cases; the process of attenuation. Studies on a number of operons have revealed that transcription termination can occur within genes as well as at the end of the operon (Jackson and Yanofsky, 1973; Kasai, 1974; Zurawski et al, 1978; Gardner, 1979; Gemmill et al, 1979). These internal termination sites are used for control of transcription

and appear sensitive to some features of metabolism that are unable to influence directly the initiation of transcription. Bacterial initiation and termination signals have been reviewed recently by Rosenberg and Court (1979). Attenuation regulates gene expression by selectively reducing transcription of the DNA in regions distal to the promoter. Operons concerned with the biosynthesis of amino acids are often controlled in this manner. The translation product of the initial portion of the mRNA is a peptide rich in the amino acid whose synthesis is determined by that operon. Scarcity of the amino acid causes the ribosome to stall long enough at the relevant codon in the transcript to permit distal parts of the messenger to adopt a conformation that allows RNA Polymerase to proceed through a site which would otherwise elicit termination (Platt, 1978; Keller and Calvo, 1979; Platt, 1981; Yanofsky, 1981). Attenuation in the trp operon reduces expression by a factor of about 10. Repression is far more effective and causes a 70-fold decrease in trp transcription.

Nucleotide sequences which cause stopping or pausing of RNA Polymerase

Termination of mRNA synthesis by E.coli RNA Polymerase occurs at specific sites. This event is modulated in vivo by the interaction of rho factor but many terminators can still function in vitro in the absence of rho (Roberts, 1969). Comparisons of rho-independent sites (Adhya and Gottesman, 1978; Rosenberg and Court, 1979) reveal common features which include a G+C rich dyad symmetry proximal to the termination point followed immediately in the transcript by a series of uridine residues. The symmetry element allows for the formation of stable RNA hairpins which have been demonstrated to cause pausing of

RNA Polymerase (Rosenberg et al, 1978; Farnham and Platt, 1981). Such a pause coupled with the instability of dA.rU base-pairing (Martin and Tinoco, 1980) is thought to enhance the release of the nascent transcript at the string of uridine residues. The efficiency of termination and release at rho-independent sites is directly related to the stability of the hairpin and the weakness of interactions between the DNA template and the 3' end of the newly synthesised mRNA. Farnham and Platt (1982) incorporated base analogues into the template and then transcribed from these strands. They were able to conclude that DNA-DNA interactions play no part in termination but that there is a dramatic effect when base analogues altering RNA-RNA intrastrand pairing are used. 5-iodocytidine-5'-triphosphate increases termination at rho-dependent terminators through enhancing RNA-RNA interactions in the hairpin. The incorporation of 5-bromouridine-5'-triphosphate into mRNA decreases termination because it stabilises rU.dA interactions of the nascent message with the template.

Those terminators which display an absolute requirement for rho-factor (Kupper et al, 1978; Rosenberg et al, 1978) exhibit weaker hairpins that are less G+C rich and lack the vital string of uridines. In these instances a common sequence 5'-CAAUCAA-3' is found at or near the 3' end of the transcript and a heptamer sequence resembling an RNA Polymerase binding site (Pribnow, 1975) is present about 30 bases upstream. The precise mechanism of termination in the presence or absence of rho protein is still unclear. Termination is often less than 100% efficient, causing polarity.

Bacterial operons have therefore proven to be more complex than originally expected. It is interesting to speculate on

the chain of molecular events leading to the evolution of such finely tuned systems. Luria (1965) looks at the possible origins of the lac, his and lys pathways. A process of retrograde evolution (Horowitz, 1945) based upon duplication and subsequent sequence divergence of ancestral genes is envisaged.

Gene Duplications and Enzyme Evolution

Among bacteria, if we exclude the possibility of the completely novel assembly of nucleotides to generate new genes and the acquisition of novel DNA via transduction or sexduction there remain only a few ways in which new enzymes or proteins can arise owing to the strict limitations imposed by the fidelity of DNA replication. Although extensive mutation of a redundant gene may yield a product with altered specificity or function or improved efficiency, it cannot lead to any increase in the overall gene numbers - a situation which must have occurred in order to explain the larger number of different proteins made, as a result of evolution, by more complex organisms. On the other hand, considerable evidence exists that the most important and perhaps the sole method of increasing the range of enzymes with novel functions is through gene duplication and the subsequent independent evolution of one or both sets of genes. The homologies between different haemoglobin chains are persuasive examples (Zuckerkandl and Pauling, 1965) and provide evidence that this mechanism has also been involved in the evolution of mammalian DNA. The importance of gene duplications in biological evolution has been extensively discussed (Ohno, 1970; Dayhoff, 1978; Koch, 1972; Zuckerkandl, 1975). In one model of enzyme evolution (Hartley, 1974) it is proposed that at some distant point in evolutionary history natural selection

favoured an organism which possessed several copies of a particular gene, thus "fixing" the duplication. Lifting the selective pressure would make one copy redundant and free to accumulate mutations which rendered it silent, while the other continued to make the original product. Reactivation of expression at a later date might result in a functional enzyme having a specificity for a novel substrate and which may enhance the survival of the organism.

Duplications arising in the absence of any selective pressure give rise to multiple gene copies and cause excess production of an enzyme. In this situation the wasteful synthesis of one protein may not be fully compensated for by the normal regulatory mechanisms and selection will act against the gene duplicated strain. Many duplications have a transient existence for precisely this reason, being maintained only as long as they confer some selective advantage. The duplication is eventually lost by excisional recombination between homologous regions of the genome. The segregation frequency decreases with time, following the initial duplication event, owing to DNA sequence divergence of the gene copies. It has been demonstrated that the synthesis of unnecessary amounts of one particular protein causes a reduction in growth rate (Andrews and Hegemann, 1976). It is logical to assume that the probability of a gene copy being retained is higher if it is unexpressed (Koch, 1972). Dothie (1974) showed that the loss of duplicated RDH genes from K.aerogenes strains is greater during growth on inositol than with glucose or xylitol as the substrate. Glucose catabolite represses the synthesis of RDH and so selective pressure to eliminate copies may be negligible. When the sole carbon source is inositol, catabolite repression is lower and there is evidence for very rapid selection of single-

gene strains.

Another way in which a potential gene might avoid segregational instability is to acquire a partial affinity for the new substrate (Orgel, 1977). Alternatively, should the duplication event be followed by an inversion of the repeated sequence, then such a structure will be rendered refractory to loss through generalised recombination and both copies will be free to evolve independently.

Gene duplication may occur as a result of unequal double crossing-over(e.g. Haptoglobulin),translocation (Strobel et al, 1979) or duplication of chromosomes (polyploidy). The latter options generally lower viability and so we presume that single locus duplication is the main route for increasing gene numbers.

Mutations in Protein Sequences

The divergence of two proteins will usually involve repeated single amino acid substitutions as a result of point mutations in the DNA. In silent gene copies where there is no selective pressure, all positions are equally mutable. If, however, a gene is being expressed, then it is found that selection is a non-random process. Some regions of a protein are unable to accept amino acid substitutions and still retain their full activity. This is illustrated by the invariance of 30-50% of all Cytochrome c sequences (Dickerson, 1972). There are few invariant sites in haemoglobin (Zuckerkandl and Pauling, 1965). The serine proteases too show wide varieties in primary structure, yet retain very similar tertiary structures (Hartley, 1970) and it is suggested by Dickerson (1971) that the invariant residues of a protein are those directly involved in interactions with co-factors, substrates, other proteins or intrachain linking.

Amino acid changes likely to alter drastically the functioning of protein occur in active site or prosthetic group residues, at bends in the polypeptide chain and in helices - resulting in destabilisation. Alterations to the interior hydrophobic regions may interfere with the close-packing of buried residues and disrupt the tertiary structure (Hartley, 1974). Exterior changes in polarity affect the solubility of the protein and can impair its activity.

Between 10 and 20% of DNA point mutations will be cryptic, i.e. resulting in no change of amino acid (Luria, 1965). Such a change, although neutral with respect to the protein, may not be neutral for the organism if it represents a switch to the use of a minor tRNA species (Richmond, 1970), or if it greatly affects the secondary structure of the DNA or RNA and by doing so interferes with normal transcription or translation processes.

Frameshift mutations are unlikely to be of great importance in enzyme evolution. The possibilities are remote that a gene could give rise to an active product following such an event; however, these mutations might prove very important if followed by a correcting frameshift, leading in particular to the creation of variable external loops.

Within a codon a single point mutation can give rise to nine possible alternative codons. Some will code for the identical amino acid and others will specify conservative changes, but a few will represent radical changes. These radical substitutions offer the greatest opportunity for the creation of new functional characteristics. Multichain enzymes may be of two types : those having identical subunits and those with different subunits. In the former class, the novel development was one which enabled aggregate formation between monomers

resulting from radical changes in certain surface residues. Single chain enzymes do not appear subject to allosteric effects, thus the development of a 'multichain' enzyme represents a major step in the evolution of regulatory processes. It is perhaps worth mentioning here that, whereas ArDH is active as monomer, RDH is a tetrameric enzyme composed of four identical subunits.

Evidence for the Involvement of Gene Duplications in Enzyme Evolution

Prokaryotes and Eukaryotes alike possess reiterated gene sequences. In higher organisms this is typified by the rRNA genes (Brown and Dawid, 1968), the haemoglobins (Clegg, 1970; Bishop and Robash, 1973) and the histone family (Weinberg et al, 1972). E.coli also has duplicated ribosonal RNA (rrn) genes, all of which show a very high degree of homology at the DNA level (Csórdas-Tóth et al, 1979; de Boer et al, 1979; Brosius et al, 1981).

Duplications usually fall into one of three classes. Quite large regions of the genome may be involved (Neuberger and Hartley, 1981) duplicating many genes. Alternatively, part of a gene may be repeated, leading to a protein having one or more regions of amino acid sequence homology such as Bovine Glutamate Dehydrogenase (Engel, 1973). The immunoglobulins are prime examples of internal duplication (Barker et al, 1978). Human serum albumin is comprised of three domains of 195 amino acids showing very strong homology to each other (McLachlan and Walker, 1977; see also Hood et al, 1978). Occasionally extensive internal duplication occurs resulting in a protein which is repetitive over much of its length. The size of the repeating unit varies widely. Rat collagen $\alpha 1$ chain is built up of 337

three-amino acid repeats (Barker et al, 1978). There may also be repetitious structures with repeated sequences. The prominent recurring unit of Tropomyosin is 42 amino acids long and is repeated seven times but contains a pattern of six 7 amino repeats (Barker et al, 1978). What is unclear is whether the smaller ancestral protein had a similar function to the present day polypeptide.

Gene elongation by duplication is known to occur, but the mechanisms of intragenic recombination involved are not well understood. Recent events of this nature have given rise to such proteins as Human Haptoglobin $\alpha 2$ chain (Smithies, 1962) and are believed to have been involved at some stage in the evolution of the structures of bacterial Cytochrome c and Ferridoxin (Dayhoff and Barker, 1972).

The findings of Edland and Normark (1981) suggest that tandem duplications up to 10 kb in length can result from recombination between randomly occurring short DNA homologies of only 12 or 13 base-pairs. Their data are derived from a study of DNA sequences at the novel joint of an E.coli K12 amp C duplication. The rec A protein is believed to be involved in the recombination event which occurs with the elimination of one of the repeat sequences. Bacterial insertion elements generate small specific DNA repeats following insertion (Calos et al, 1978; Grindley, 1978). A number of insertions of the chloramphenicol resistance transposon Tn9 into the lac I gene of E.coli have been sequenced and all reveal a 9bp repeat at the novel joint (Johnsrud et al, 1978). Thus it seems quite possible that recombinational events similar to these are involved in some duplications of DNA sequences. A study of the novel joint

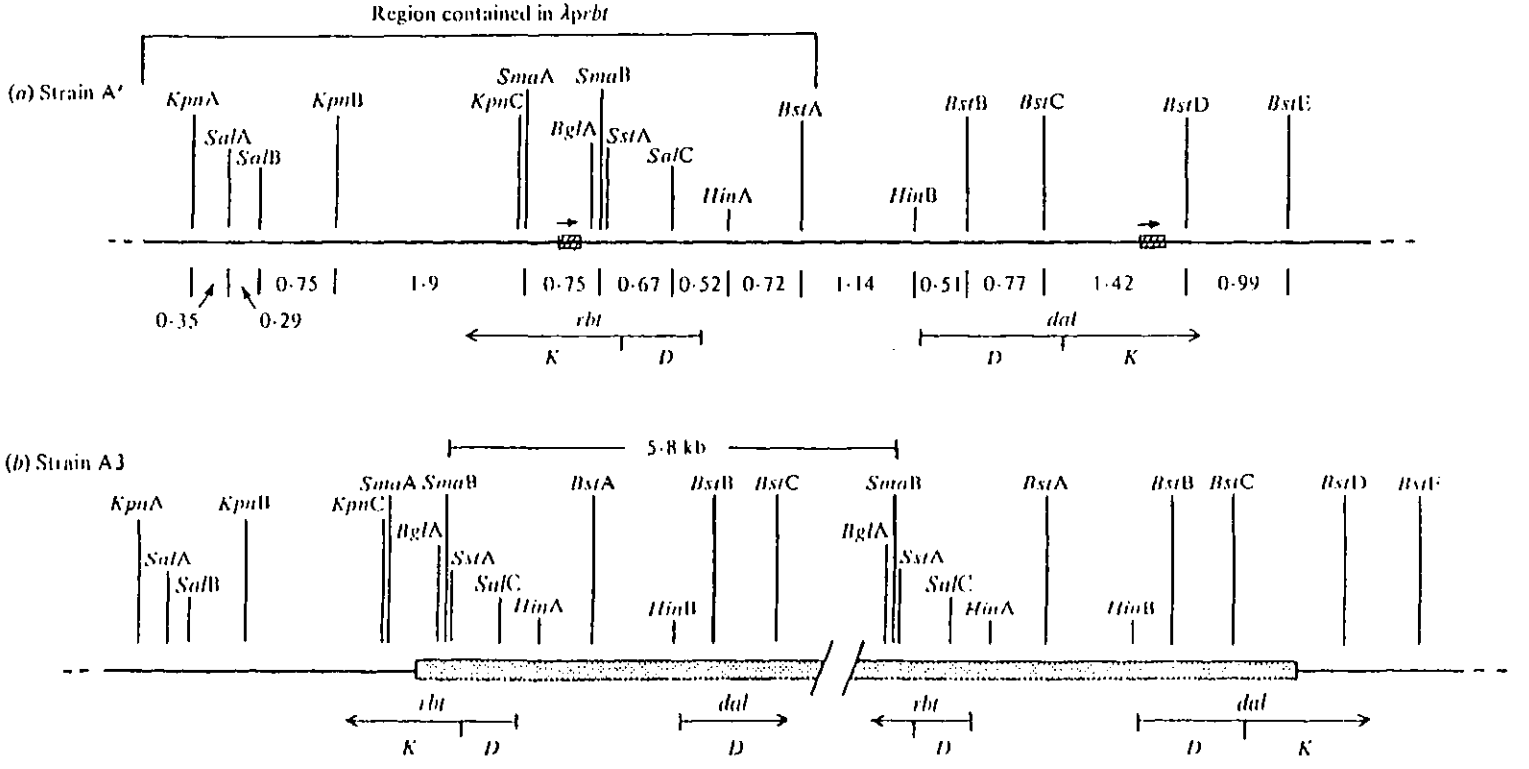
from A3, a gene duplicated RDH super-producing strain of K.aerogenes (Rigby et al, 1974; Neuberger and Hartley, 1981), reveals that recombination has taken place between two direct repeats of 9bp situated near the centre of the DRK gene and towards the end of DXK, (P. David, unpublished data; Fig. 1c).

The Evolution of Operons

With respect to operon evolution, two possibilities exist. Either the individual genes originated at different positions on the genome and were transposed, or they originated at their present sites. The former situation requires that each gene has a separate ancestry, whereas in the latter operons are presumed to arise by tandem duplication and subsequent functional differentiation. A number of considerations favour the second hypothesis (Horowitz, 1965). Enzymes in an operon form a reaction chain where one product acts as the substrate for the next stage. This implies overlapping specificity which in turn suggests some structural homology and probably common ancestry. The alternative hypothesis is both more complex and harder to substantiate. Inversion and translocation are rare events unless mediated by transposons, bacteriophages or F'-factors, yet studies with Drosophila reveal that inversions are the major cause of gene shuffling among different species of this organism (Sturtevant and Novitski, 1941).

Work on experimental evolution in K.aerogenes is well documented (Rigby, 1971; Hartley et al, 1972; Rigby et al, 1974; Hartley, 1974). These articles catalogue the spontaneous events leading to improved growth on a novel substrate. However, the lack of any detailed information on the genetics of K.aerogenes limited the depth in which these mutations could be examined. Transfer of the pentitol genes to E.coli K12

Fig.1c Physical Map of the K.aerogenes A3 Duplication



(a) Strain A' DNA. The positions of the two 7bp direct repeats responsible for the generation of the A3 duplication (b) are shown. Distances between restriction sites are in kb. The locations of genes encoding the pentitol dehydrogenase (D) and pentulokinase (K) of each operon are indicated.

(Diagram reproduced from Neuberger and Hartley,1981)

was seen as a step towards increasing knowledge in this area. Interestingly, Reiner (1975) had recently discovered natural strains of E.coli C capable of growth on ribitol and d-arabitol and showed that the genes for these functions mapped around 40' on the genome; almost equivalent to the insertion point of the K.aerogenes genes into E.coli K12. Reiner was unable to locate any natural E.coli B or K12 strains that could utilise pentitols and proved that the latter does not possess any DNA homologous to the pentitol operons. Further work by Scangos and Reiner (1978) showed the gene arrangement in E.coli C to be similar to that of K.aerogenes and raised speculation on the possibilities of interspecies transfer of the dal and rbt operons. It was considered that this may have been a relatively recent event. To test this hypothesis the amino acid sequences of RDH from both species were determined and shown to be 95% homologous, (Altosaar and Hartley, 1976), although it is not clear whether this degree of similarity is normal between proteins of these two very closely related enterobacteria.

The similarity of the two pathways for ribitol and d-arabitol metabolism and the clustered arrangement of the genes on the chromosome may signify that they share a common ancestor. In particular, a gene duplication and inversion are implied. The pentitol genes may therefore represent a model for the duplication and modification of an operon leading ultimately to the establishment of a new metabolic pathway.

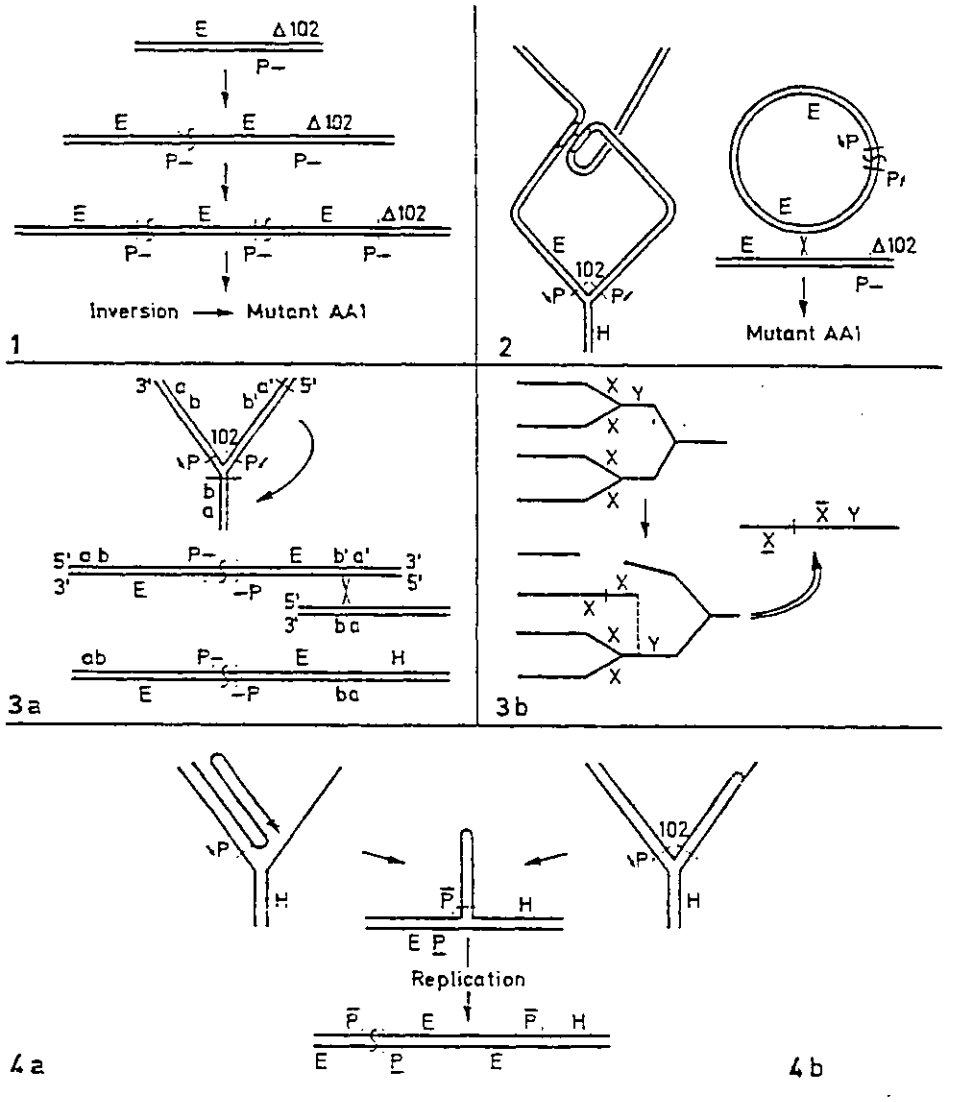
Charlier et al (1979) showed that duplications of arg genes in rec^+ and recA genetic backgrounds were strictly tandem and that short duplications (2kb) were much more stable than longer ones ($\approx$10kb). Analysis of one duplication,

however, revealed that it consisted of two inversely arranged *argE* genes which effectively constituted an artificial divergent operon.

What mechanisms are involved in the formation of invert repeats of the type postulated for the pentitol operons? Charlier *et al* propose a model requiring breakage and recombination between short identical or similar sequences, behind and upstream of the replication fork (Fig. 1d). This generates an inverted repeat from a single recombinational event. The combined duplication and inversion event results in an effective gene overlap at the centre where transcription could occur from both strands in opposite directions. Such a situation as this is implied in the gene order quoted by Charnetzky and Mortlock (1974c) for the pentitol operons. Their investigations suggest that putative repressor genes are interdigitated. However, it has since been demonstrated that this is not the case (C. Smith, unpublished data), so the Charlier model becomes less attractive as a means of explaining the pentitol gene order. It has not yet been proven conclusively that the two repressors do not overlap, but it seems certain that they are not interspersed.

If the *rbt* and *dal* operons are indeed related, albeit distantly, they would be expected to exhibit a greater degree of homology in nucleotide sequence than two totally unrelated systems. If the causal duplication occurred a considerable time ago, then similarities might be only slight. It was therefore decided to attempt to obtain DNA sequence data for parts of the *dal* operon which would allow comparison to equivalent regions of the *rbt* operon (T. Loviny, unpublished data).

Fig.1d Mechanisms for the Formation of Inverted Gene Duplications



(1) Formation of a tandem duplication followed by unequal crossing over and inversion of one copy. (2) A circular invert repeat arising from a break at the growing point of the replication fork and a recombination event upstream of the growing point. This is followed by reciprocal crossing over with the chromosome. (3a) Formation of a palindromic repeat by recombination between inverted repeats ahead of and behind the replication fork. (3b) Similar to (3a) but occuring between sequences in aligned replication forks. (4a) and (4b), Double palindromes formed by replication errors.

(Fig.11 of Charlier et al, 1979)

In addition, the amino acid sequences of structural proteins like ArDH and DXK could also be derived and compared to the available sequences of *rbt* enzymes. Aside from the evolutionary considerations, a study of *dal* operon expression was also envisaged, centred around sequence analysis of the promoter/operator region. This dissertation describes the work undertaken to try to resolve the evolutionary origins of the d-arabitol operon and investigate some aspects of its regulation.

CHAPTER 2

MATERIALS AND METHODS

I) Materials

Agarose by Marine Colloids was obtained through Miles Products Ltd., Stoke Poges, U.K.

d-arabitol and d-ribulose were from Cambrian Chemicals, Croydon, U.K.

Agar, tryptone and yeast extract were purchased from Difco Laboratories, Surrey, U.K.

BSA, Dowex 50W/X8, lysozyme, DNAase I, tRNA, Salmon sperm DNA, Heparin, chloramphenicol, tetracycline, vitamins, cAMP, IPTG, BCIG and d -xylulose were all purchased from Sigma, Poole, Dorset, U.K.

Ethidium bromide, caesium chloride, acrylamide and methylene-bisacrylamide (specially purified for electrophoresis) were obtained from BDH, Enfield, Middlesex, U.K.

Urea (ultra-pure) was from Schwartz-Mann, Orangeburg, N.Y., polyethylene glycol 6000 from Koch Light Labs., Colnbrook, Bucks., U.K., and Silane from Wacker Chemie, Munich, F.R.G.

Deoxynucleoside triphosphates and dideoxynucleoside triphosphates were supplied by P.L. Biochemicals, Windsor, Berkshire, U.K., and ribonucleoside triphosphates by Boehringer Corporation (London) Ltd., Lewes, E. Sussex, U.K.

Bacterial alkaline phosphatase and RNAase A were from Worthington Enzymes via Millipore (UK) Ltd., London. DNA Polymerase I and RNA Polymerase (*E.coli* MRE 600) were obtained from Boehringer.

Restriction endonucleases were purchased from New England Biolabs (CP Labs, Bishops Stortford, UK), Bethesda Research Labs (Cambridge, UK) or from Boehringer Ltd., with the exception

of Bst I, which was a generous gift from Dr. C.M. Clarke.

T4 polynucleotide kinase was a gift from Dr. D. Glover, and K.aerogenes RDH at 100u/mg was kindly donated by M. Cardosi.

Catabolite repressor protein was obtained through Calbiochem, Bishops Stortford, U.K.

All isotopes were bought from Amersham International, UK, except for cordycepin triphosphate, which was from New England Nuclear, Southampton, U.K. Tritiated xylitol and ribitol were made by the Amersham tritium-labelling service.

Piperidine AR (BDH) was re-distilled before use and stored at 4°C in the dark.

Phenol AR (BDH) was re-distilled and kept frozen at -20°C.

Formamide AR (BDH) was re-crystallised twice and deionised prior to use.

Hydrazine was bought from Pierce and Warriner, Chester, UK, and stored airtight at 4°C.

17mer primer was a gift from G. Winter (M.R.C. Cambridge) or was purchased from Collaborative Research Inc. through Uniscience Cambs., UK.

II) General Techniques

Spectrophotometry

Optical densities of solutions or bacterial cultures are measured in 1ml or 3ml silica cuvettes, using a Gilford 2000 recording spectrophotometer with a Pye Unicam monochromator.

Scintillation counting

Tritium is detected by immersing or dissolving the sample in a suitable cocktail and counting in an Intertechnique SL30 liquid scintillation counter. For ^{32}P, samples or filters are dried and counted for Cerenkov radiation without the addition of scintillant.

The scintillation cocktail consists of 20g of scintillation grade naphthalene and 7g of butyl PBD, [2-(4'-tert-butylphenyl)-5-(4"-biphenyl)-1,3,5 oxodiazole] dissolved in 1000mls of toluene and 1,4-dioxan (3:7 v/v).

Autoradiography

i) ^{32}P-labelled DNA or RNA is detected following electrophoresis in agarose or polyacrylamide gels by exposure to Kodak "Kodirex" or XH-1 film at room temperature for 1-2 hours. The gel is covered with Saran Wrap (Spondex Ltd., Croydon, UK). Lead/aluminium sheets are used to maintain close contact between the film and gel surfaces.

ii) DNA sequence gels are carefully transferred to the surface of a used X-ray film, covered with Saran Wrap and placed in a film cassette in contact with Kodax XH-1 or Fuji RX film. Where the radioactivity is weak, films are pre-fogged with an electronic flash-gun to give a background OD of 0.1 (Laskey and Mills, 1977). Exposure is at -70°C for between 3 days and 4 weeks.

Dideoxy sequencing yields a more radioactive product and much shorter exposures are possible. The gel is left attached to one glass plate and fixed for 20 minutes in 10% acetic acid, after which time it is blotted dry with tissues, covered with Saran wrap and exposed to un-fogged film at room temperature overnight.

iii) Ultra-thin 200µm or 100µm sequence gels are extremely fragile and are therefore covalently bound to their glass plate to facilitate handling. These gels are fixed for 10 minutes in 10% acetic acid, or until the marker dyes diffuse out completely. The gels are then rinsed gently with water from the tap and dried with tissues, before baking at 95-100°C for a minimum of one hour.

The dried gel remains firmly bonded to the glass plate and is exposed in direct contact to Kodak XH-1 at room temperature for 6-15 hours. It is important to allow sufficient time for all of the urea to diffuse from the gel before drying, or else it will crystallise out and disrupt the sequence ladder.

All X-ray films are developed automatically by a Kodak Industrial X-OMAT processor Model 3.

III) Microbiology and Genetics

A) Strains and Media

i) Bacteria

All K.aerogenes strains are derivatives of K.aerogenes 1033 (Neidhart and Magasanik, 1956). The parental strain for all evolvants is strain A. This is a prototroph derived from the ribitol operon constitutive mutant X1 (arg gua rbt-101) of Wu et al (1968). X1 itself originates from K.aerogenes FG5, which is inducible for the pentitol operons.

Those strains of E.coli designated 'CSH' are from Cold Spring Harbor. Full descriptions of their genotype may be found in Miller (1972). All other strains follow our laboratory classification unless otherwise stated. A list of those strains used in this work is presented in Table 2(i).

ii) Plasmids

The plasmid vector pBR322 (Bolivar et al, 1977) was used for all cloning experiments. Evidence exists that our stock of this plasmid has several mutations which alter its cleavage pattern by Alu I and Taq I restriction enzymes, (see text). The plasmid confers resistance to ampicillin (Ap) and tetracycline (Tc) at levels of 100µg/ml or 25µg/ml respectively. The complete DNA sequence of pBR322 is known (Sutcliffe, 1978a). All plasmids used in this work and their host strains are listed in Table 2(ii).

Table 2i Bacterial Strains

a) *K.aerogenes*

Strain	Genotype	Source
FG5	*arg gua*	E.C.C.Lin
A3	*rbtC*-101 $(rbtD)_n$	P.W.J.Rigby
A111	*rbtC*-101 $(rbtD)_n$	P.W.J.Rigby

b) *E.coli*

NC100	F^+ *gal tsx lpcA rpsL rbtC*-101 $rbtD^+$ $rbtK^+$ dal^+	Strain 1EA of Rigby *et al* (1976)
PS640	Hfr H *thi* (λ*prbt dal*)	Neuberger 1978
CSH 62	Hfr H *thi*	Cold Spring Harbor Labs.
DC10	*trpR gal* $hsdR^-$ $hsdM^+$	This Lab. P.David
SK1592	F^- *gal thi endA sbcB15* $hsdR_4$ $hsdM^+$	S.Kushner
JM101	*proAB* $lacI^q$ Z M15 *supE lac pro thi traD36* F^-	J.Messing
HB101	*proA2 leuB6 thi lacY rpsL* F^- *supE endA recA13* $hsdR^-$ $hsdM^-$ Sm^R $Su2^+$	H.Boyer
PS621	Hfr H *thi* (λ*prbt dal*)	This Lab. Neuberger(1978)

Table 2ii Bacterial Strains Harbouring Plasmids

Plasmid	Host Strain	Markers	Source
pBR322	HB101	Ap Tc	S.Kidd
pRD351	HB101	Ap *dal*^{c}	M.Neuberger
pRD251	HB101	Ap	This work
pRD252	HB101	Ap	This work
pRD253	HB101	Ap	This work
pRD256	SK1592	Ap	This work
pRD262	SK1592	Ap *dal*R^{+}	This work

iii) Bacteriophage

Phage M13 mp7 is a derivative of mp2 (Grönenburg and Messing, 1978). M13 mp2 has a DNA fragment encoding a partial lac repressor, complete lac o and lac p regions and the first 145 amino acids of the β-galactosidase gene inserted in the intergenic region. There is a unique EcoRI site positioned 12 base pairs into the lac Z gene. In mp7, a synthetic palindromic 42bp DNA fragment containing several unique restriction enzyme sites is cloned into this EcoRI recognition sequence. Phages mp8 and mp9 have smaller, asymmetric inserts which allow fragments to be cloned in one particular orientation relative to the M13 genome. The insertion of these synthetic pieces into lac Z does not inhibit ∝-complementation by the F factor in JM101. All three types of M13 have two amber mutations which remove the genomic Acc I/Hinc II and BamH I sites. All M13 and λ phage strains are listed in Table 2(iii).

B) Liquid Culture Media

i) Luria Broth (LB)

10g tryptone, 5g yeast extract and 10g NaCl are dissolved in d·H_2O and the solution is adjusted to pH7.4 with NaOH. The medium is made up to 1 litre and autoclaved. Antibiotics, if required, are filter-sterilised and added to the LB when cooled.

ii) M9 Salts

5.8g Na_2HPO_4, 3.0g KH_2PO_4, 0.5g NaCl and 1.0g NH_4Cl are dissolved in 1 litre of d·H_2O and autoclaved. Following sterilisation, 0.2ml of 1M $MgSO_4$, 2ml of 10mM $CaCl_2$ and 0.1ml 1% thiamine are added.

iii) Minimal Media

The basic M9 medium is supplemented to 0.2% w/v with the required carbon source. Vitamins and antibiotics are added

Table 2iii Bacteriophages

Phage	Host	Genotype	Source
627	CSH62	λcI857 Ssus7 ninΔ5 srIλ4° srIλ5° b2Δ(srIλ1-srIλ2)	N.E.Murray
λprbt	CSH62	λcI857 Ssus7 ninΔ5 srIλ4°,5° rbt-101 $rbtD^+$ $rbtK^+$	M.Neuberger
λprbtdal	CSH62	λcI857 Ssus7 ninΔ5 srIλ4°,5° rbt-101 $rbtD^+$ $rbtK^+$ $dalD^+$ $dalK^+$	M.Neuberger
M13 mp7	JM101	$lacI^-$ lac o/p $lacZ^-$	G.Winter
M13 mp8	JM101	$lacI^-$ lac o/p $lacZ^-$	R.Miller
M13 mp9	JM101	$lacI^-$ lac o/p $lacZ^-$	R.Miller

where necessary. L-amino acids are supplied at 40µg/ml, vitamins at 1µg/ml, Ap 100µg/ml and Tc at 25µg/ml. Thymine or any other essential nucleotides are present at 50µg/ml.

iv) TY medium

8g tryptone, 5g yeast extract and 2.5g NaCl are dissolved in d·H_2O, brought to pH7.4 with NaOH and adjusted to a final volume of 1 litre before sterilising.

C) Solid Media

i) TYE

10g tryptone, 5g yeast extract, 8g NaCl, 5mgs thymidine and 20g agar are dissolved in 1 litre of d·H_2O and autoclaved.

ii) Minimal Plates

These are basically M9 medium containing 1.5% w/v of agar. The agar, water and $MgSO_4$ are autoclaved together and combined with a 1/10 volume of sterile [10x]M9 salts just prior to pouring the plates. Sugars, sugar-alcohols, vitamins and antibiotics are added as for M9 media.

iii) Stab agar

10g tryptone, 8g yeast extract, 5g NaCl and 6g agar are autoclaved in 1 litre of d·H_2O and 3 ml portions dispensed into sterile glass bijou bottles.

iv) Soft agar

TY medium containing 6g agar per litre is autoclaved and stored for up to 1 week at 60°C.

v) H-Top agar

TY medium containing 15g agar per litre is sterilised. 2ml of 1M $CaCl_2$ and 5ml of 20% w/v glucose are added just before pouring the plates.

Fermenter medium

For growth of bacteria containing λ phage the following

medium is used.

0.5% peptone	0.26% KH_2PO_4
0.5% yeast extract	0.46% NaH_2PO_4
1% glucose	0.01% polyethylene
0.2% NaCl	glycol P-2000

The pH is adjusted to 7.0.

D) Storage of Bacteria

i) Stabs

Sterile, airtight glass bijous containing about 3 mls of a nutrient agar are prepared. A single colony from a fresh plate is stabbed into the medium, and the vial is incubated overnight at the appropriate temperature to initiate growth. Stabs are stored dark at 4°C and remain viable for several years.

ii) Glycerol Cultures

A single bacterial colony is grown overnight and 1 ml of the culture is mixed with an equal volume of sterile 30% glycerol. Strains are stored at -20°C and the cells are recovered by innoculating fresh medium with a few drops of the glycerol culture. Cells remain viable for several years.

Bacteria which contain plasmids are grown in the presence of antibiotics to prevent possible loss of the plasmid before storage.

E) Culture Conditions

Agar plates are incubated inverted at 37°C. Overnight growth is usually sufficient, but colonies on minimal media, particularly d-arabitol, may require several days. Thermoinducible λ lysogens are grown at 32°C.

Small liquid cultures (10-20 mls) are grown in 2.5 cm x 15.5 cm bubble-tubes immersed in a water bath. Aeration is provided by a 'Biotec' FE007 air pump via a manifold and sterile

cotton wool filters. Larger cultures (50-250 mls) are grown in conical flasks agitated in a Gallenkamp orbital incubator.

F) Microbiological Techniques

i) Preparation of Competent Cells

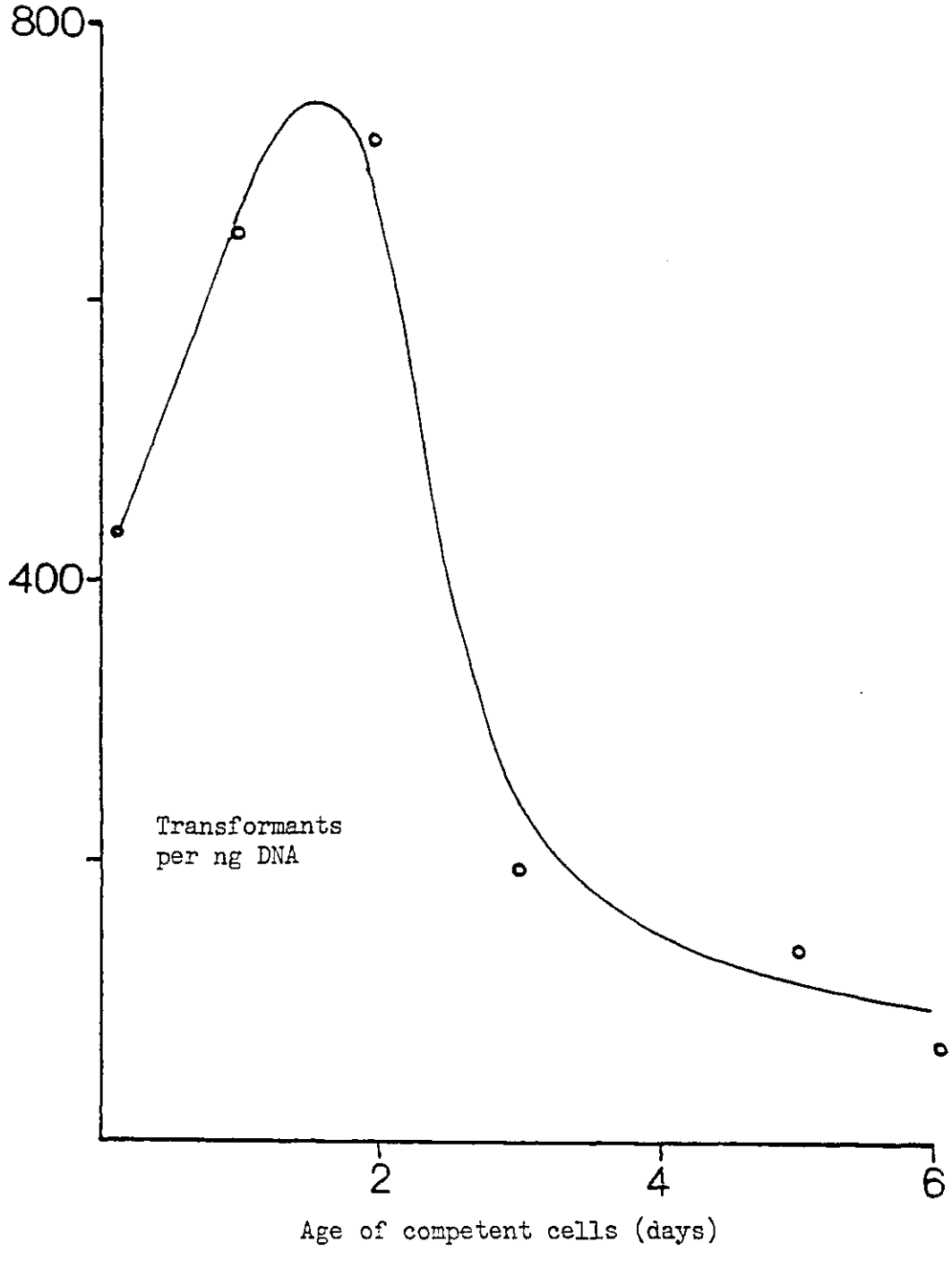
Cohen et al (1972) showed that wild type strains of E.coli can be transformed by cccDNA following a treatment with $CaCl_2$. Cells treated in this way are said to be 'competent'.

An overnight culture of the recipient strain is diluted 100 fold into fresh LB and grown with aeration at 37°C until the the OD_{600} is approximately 0.6. The cells are spun down gently at 3000g and 4°C to minimise breakage, then suspended in a one half culture volume of ice-cold 50mM $CaCl_2$. After standing on ice for 20 minutes, the cells are pelleted again and taken up in 1/10 of the original culture volume of $CaCl_2$. Competent cells may be used at once, although they retain competence for up to one week if stored at 0°C. The highest transformation efficiency is obtained with cells which have been stored for 24 hours (Fig. 2a) and declines steadily thereafter along with cell viability.

ii) Transformation of E.coli with plasmid DNA

DNA is dissolved in buffer containing 10mM $CaCl_2$, 10mM $MgCl_2$ and 10mM Tris-HCl pH7.5. 0.1ml of DNA is mixed with 0.2ml of competent cells and incubated with occasional shaking at 0°C for 40 minutes. After a 2 minute heat-shock at 42°C, 0.5ml of rich medium is added, and cells are allowed to grow for 1 hour at 37°C. This allows expression of the antibiotic resistance genes (Cohen et al, 1972). The culture is added to 2.5 ml of molten 0.8% TYE agar at 45°C, vortexed briefly and poured on to TYE plates containing the relevant selective antibiotics.

Fig.2a Variation of transformation efficiency of competent E.coli JM101 with time



Competent E.coli JM101 (0.2ml) were transformed at 24hr intervals with 1ng of M13 mp7 RF. Cells and DNA were incubated for 40 minutes at 0°C, heat-shocked for 2 minutes at 42°C then added to 3ml of molten agar containing 0.2ml JM101, 30µl BCIG (20mg/ml), 20µl IPTG (24mg/ml) for plating on minimal medium +glucose. Plates were incubated inverted for 15-20hrs at 37°C. Between transformations the cells were stored on ice in 50mM $CaCl_2$.

iii) Replica Plating

Up to 100 individual colonies are transferred to a "master" plate using sterile wooden cocktail sticks. A grid pattern (Miller, 1972) is used beneath the plate to ensure even spacing of the colonies. Clones are grown up at 37°C. The grid of colonies is then transferred in a single step to the surface of a piece of sterile velvet, stretched tightly over a cylindrical wooden block, by pressing the inverted master plate on to it. Up to six replicas of the master can usually be obtained by laying fresh plates on top of the velvet.

iv) D-cycloserine enrichment of transformants

This technique is used to select Ap^R Tc^S transformants from a mixed culture containing a significant proportion of Ap^R Tc^R cells. Treating the Bst I cut vector with BAP reduces the background of Tc^R cells arising from religation of the plasmid DNA, but even without this step d-cycloserine enrichment was found to give > 98% Ap^R Tc^S cells. The principle of the enrichment is quite straightforward. Actively growing cells are lysed by ampicillin at 100µg/ml. However, tetracycline at levels of 25µg/ml is bacteriostatic; sensitive strains are not killed, but cease to grow. d-cycloserine is toxic only to growing cells where cell wall mucopeptide synthesis is blocked by the inhibition of Alanine Racemase and d-alanyl-d-alanine synthetase (Curtiss et al, 1965).

Following transformation, cells are grown overnight in LB supplemented to 100µg/ml with ampicillin. A portion of the overnight culture is diluted 100 fold into 20ml LB/ampicillin in a bubble tube and grown at 37°C for one hour. Tetracycline is added and incubation continued for a further 45 minutes. d-cycloserine, dissolved in M9 salts, is added to 100µg/ml. The culture OD_{650}

falls as Tc^R cells are killed. After 2-3 hours incubation, the cells are pelleted at 3000g and 4°C for 5 minutes, washed once with LB to remove traces of d-cycloserine, then resuspended in 10ml LB + ampicillin (100µg/ml). The culture is grown to a suitable OD_{650} before serial dilutions are made and plated out on selective media.

IV) Enzymology

i) Preparation of Cell-Free Extracts

a) Sonication

Cultures (200ml) are grown in flasks in an orbital shaker to late log phase. The cells are then pelleted at 5000rpm and 4°C in the GSA rotor of a Sorvall RC-5 centrifuge. The supernatant is decanted off and the cells washed with 100mls of ice cold KPN buffer (100mM potassium phosphate pH7.0, 1mM NAD^+), pelleted again, and finally resuspended in 5ml of cold KPN. The suspension is subjected to sonication using a Soniprobe (Dawes Instruments, London W3) with the power setting at 5, and the current tuned to 6 Amps. Sonication, with the cells on ice, uses five bursts of 1 minute with a 1 minute pause between each. Cell debris is removed by a clearing spin at 27,000g for 30 minutes at 4°C in the Sorvall SS34 rotor. The supernatant may be assayed immediately or stored frozen at -20°C. If the extract is for assaying Kinase activity, then the KPN buffer is made 2mM in β-mercaptoethanol prior to rupturing the cells. In addition, for d-xylulokinase activity it is vital to maintain the temperature between 15°C and 25°C throughout the preparation, and to assay the extract at once, since this enzyme is cold-labile.

ii) Cell Rupture using Lysozyme

A 25ml culture is grown to late log phase in the appropriate medium. The bacteria are pelleted at 5000rpm for 5 minutes at 4°C, washed with 10mls of cold 100mM Tris-HCl pH8.0 containing 5mM

EDTA and finally resuspended in 5mls of the same buffer. In those preparations specifically intended for kinase assays 2mM β-mercaptoethanol is included. Cells are lysed by incubation with Lysozyme (50µg/ml) at 37°C for 15 minutes. $MgCl_2$ is added to a final concentration of 10mM and the viscosity is reduced by treatment with DNAase I (1µg/ml) at room temperature for 5 minutes. The cell debris is removed by centrifugation at 27,000g for 30 minutes at 4°C, or 15°C for d-xylulokinase extracts. The supernatant is used directly for enzyme assays.

Estimation of Protein Concentration by the Lowry Method

A modified Lowry method first cited by Miller (1959) was used. A solution containing 10% Na_2CO_3, 0.05% $CuSO_4$ and 0.1% potassium tartrate in 0.5M NaOH is prepared freshly for each set of assays, and 1ml of this mixture is added to 1ml of a neutral protein solution. The acceptable range of protein concentration lies between 40 and 200µg/ml. Samples are mixed well and allowed to stand at room temperature for 10 minutes before adding 3mls of a 1:11 dilution of Folin-Ciocalteau phenol reagent. Incubation is continued for a further 30 minutes during which time a blue colour develops. The OD_{560} of each solution is determined, and a standard curve is constructed for each set of assays using serial dilutions of BSA.

Estimation of Protein by the Dye-Binding Method

A protein determination method based on the binding of Coomassie Brilliant Blue G-250 to proteins is described by Bradford (1976). Binding results in a shift in the absorption maximum from 465nm to 595nm. The assay involves measuring the increase in OD_{595}.

The protein reagent is prepared as follows. Coomassie

Brilliant Blue G-250 (100mg) is dissolved in 50ml 90% ethanol, and 100ml of phosphoric acid is added. The resulting solution is made up to 1 litre with $d{\cdot}H_2O$.

For the assay, 0.1ml of a solution containing between 10µg and 100µg of protein is added to 3ml of protein reagent in a 4ml cuvette. The cuvette contents are mixed by inverting several times and the OD_{595} is measured after 2 minutes against a reference cell containing no protein. Samples may be left for up to one hour before reading the OD. A standard curve is constructed, using dilutions of BSA spanning a range of 0-100µg/cuvette.

V) Enzyme Assays

i) Ribitol Dehydrogenase and d-Arabitol Dehydrogenase Assays

Activity is measured by following the increase in OD_{340} resulting from the reduction of NAD^+. An increase in OD_{340} of 6.25 occurring in a sample volume of 1ml with a 1cm light-path is equivalent to the formation of 1µMole of NADH.

The assay mix is composed of 0.1M potassium phosphate pH 7.0, 0.83mM NAD^+ and 50mM d-arabitol or d-ribitol. Silica cuvettes containing 1ml of assay mix are equilibrated at 28°C in a water bath, and a sample of enzyme is added from a Hamilton syringe or Gilson 'Pipetman'. The contents are mixed throroughly by repeated inversion of the cuvette, and the appearance of NADH is monitored in a Gilford 2000 recording spectrophotometer with a Pye Unicam SP500 monochromator with the chart speed set at 5cm/min. and a full scale deflection of 0.1 OD units.

ii) Assay for d-ribulokinase using [1-^{3}H]-ribitol

This assay was developed by Neuberger (1978) and is a modification of the l-ribulokinase assay of Lee and Englesberg (1963), who incubated samples in the presence of [1-^{14}C]-l-ribulose

and rATP and measured the increase in barium/ethanol precipitable radioactivity resulting from formation of labelled ribulose-5-phosphate.

In the DRK assay labelled substrate is made *in situ* by the action of RDH on tritiated ribitol. The rate of formation of labelled ribulose-5-phosphate is an indication of DRK activity and remains linear up to 70 cpm/minute (Neuberger, 1978). Experimental details were identical to the published protocol. DRK activity may be expressed in units, where one unit of the enzyme phosphorylates 1µMole of d-ribulose per minute under the conditions of this assay. The RDH used in the assay mix should ideally be purified from a DRK^- strain but the sample used by these assays was demonstrated to be free of any contaminating DRK. When using cell-free extracts from *rbt* constitutive strains the addition of extra RDH is not essential for the DRK assay to proceed (Neuberger, 1978).

iii) *Assay for d-xylulokinase using [1-^{3}H]-xylitol*

This assay was developed as an alternative to the Lactate Dehydrogenase/Pyruvate Kinase coupled reaction, which is very sensitive to interference from the contaminating kinases and dehydrogenases in crude extracts. The procedure is similar to that described for assaying DRK, except that the assay mixture is 0.1M with respect to xylitol and contains 1.5×10^6cpm/ml [1-^{3}H]-xylitol. *K.aerogenes* A3 RDH is included at 0.5 units/ml. Tritiated d-xylulose is formed by the action of RDH on xylitol, and this serves as the substrate for DXK. The rate of production of [^{3}H]-d-xylulose-5-phosphate is monitored by removing samples from the reaction mixture at intervals and counting the barium/ethanol precipitable radioactivity on glass fibre filters as described (Neuberger, 1978).

VI) PREPARATION OF DNA

General

All glassware is chromic acid washed and sterilised. Buffers, wherever possible, are autoclaved. Double stranded plasmid or bacteriophage DNAs are stored at 4°C in Low Tris Buffer (10mM Tris-HCl pH7.5, 1mM EDTA) in siliconised glass tubes or Eppendorfs with a few drops of $CHCl_3$. For long term storage, DNA is frozen in LTB at -20°C. Repeated freeze-thawing of nucleic acids is best avoided as it causes shearing. Linearised plasmid DNA is more susceptible to damage by shear forces than supercoiled DNA and is always stored at 4°C. In some instances 10-50µg amounts of plasmid are precipitated with ethanol, washed, dried under vacuum and stored dry at -20°C. Samples can then be reconstituted in LTB as required.

For the determination of nucleic acid concentration it is assumed that an OD_{260} of 1.0 is equivalent to 50µg/ml of dsDNA, 50µg/ml RNA or 36µg/ml ssDNA. The ratio OD_{260}/OD_{280} is used as a criterion of purity; a value of 2.2 is considered to be pure DNA.

Sterile dialysis tubing is prepared by sequential boiling in i) $NaHCO_3$/5mM EDTA, ii) $d{\cdot}H_2O$, iii) 5mM EDTA, followed by several washes with $d{\cdot}H_2O$ and autoclaving for 15 minutes at 15 p.s.i. The tubing is then stored in $d{\cdot}H_2O$ at 4°C with a few drops of $CHCl_3$.

Preparation of Bacteriophage λ DNA

The method used is based upon those of Zubay (1973) and Haseltine and Müller-Hill (1972). Phages are prepared from frozen cell pastes of thermoinduced λ p *rbt* and λp *rbt dal* lysogens. Cells were grown in 60L fermenters in the Imperial College Pilot Plant to an OD_{650} of 1.5. Raising the temperature

from 30°C to 42°C causes induction of the prophage, after which growth is continued for a further 2½ hours at 37°C. Cells are harvested in a Sharples continuous flow centrifuge and the wet paste frozen at -20°C. Thermoinduction of λ cI857 Ssus7 lysogens causes accumulation of intact phage without resulting in cell lysis.

For the isolation of phage DNA 30g of cell paste is allowed to thaw at room temperature in 90ml of lysis buffer (10mM Tris-HCl pH7.9, 0.1mM EDTA) and 1ml of $CHCl_3$. The mixture is stirred gently and becomes viscous as lysis proceeds. After 30 minutes 1ml of 1M $MgCl_2$ and 100µl of 1mg/ml DNAase I are added. Incubation is at room temperature. When the viscosity of the lysate is greatly reduced it is centrifuged twice for 30 minutes at 12,000 rpm and 4°C in a Sorvall SS34 rotor. Phage particles are purified from the supernatant on a CsCl block gradient as follows. 2ml of 1.3g/ml CsCl made up in phage buffer (10mM Tris-HCl pH7.9, 10mM $MgCl_2$, 10mM NaCl) is placed in a Beckmann SW25 nitrocellulose tube. This is sequentially underlaid by 2ml of 1.5g/ml CsCl and 2ml of 1.7g/ml CsCl. Finally, the gradient is overlaid with 3ml of 30% Sucrose in phage buffer. The tube is then carefully filled to within a few mm of the top with the phage lysate. Two gradients are required for the amount of lysate produced from 30g of cells. Centrifugation, at 9°C, is for 100 minutes at 25,000 rpm in a Beckman SW25 rotor. A blue-white opalescent band in the 1.5g/ml block contains the phage and is best collected by puncturing the wall of the centrifuge tube with a 22-gauge syringe needle. To prevent leakage around the needle the surface of the tube is first roughened with acetone, and a small patch of cloth adhesive tape is applied. This acts

as an effective seal. The phages are then further purified by equilibrium centrifugation. Phage-containing fractions are adjusted to a CsCl density of 1.5g/ml in a volume of 4mls. Final accurate adjustments are performed using a refractometer. For λp rbt the correct index is 7°59', and for λp rbt dal 8°15'. These solutions are transferred to Beckman Ti50 nitrocellulose tubes and centrifuged at 25,000 rpm for 15 hours at 9°C in a Ti50 rotor. The phage band is harvested by side-puncture with a syringe as described above. Caesium chloride is removed from the purified phage sample by dialysis for 2-3 hours against 1 litre of Dialysis Buffer I (10mM Tris-HCl pH7.9, 50mM NaCl, 1mM EDTA) plus a few drops of $CHCl_3$. The OD_{260} of the phage suspension is measured and, if necessary, reduced to 10 by dilution with Dialysis Buffer I.

Deproteinisation of λ phage

SDS is added to the phage suspension to a final concentration of 0.5%w/v. This mixture is shaken gently at 65°C for 10 minutes and brought to 0.5M with KCl. After incubation on ice for 20 minutes the capsid proteins and potassium-SDS complexes are removed by centrifuging in the SS34 rotor for 45 minutes at 12,000 rpm (4°C). The supernatant is dialysed at 4°C against changes of 1 litre volumes of Dialysis Buffer II (10mM Tris-HCl pH7.5, 1mM EDTA). Phage DNA is removed from the dialysis tubing and stored at 4°C over $CHCl_3$ or is phenol extracted, precipitated with ethanol and stored dry at -20°C.

The yield of DNA varies according to the cell paste used but is generally between 0.5mg and 1mg per 30g wet weight. The OD_{260}/OD_{280} ratio of the pure DNA is around 2.1.

Plasmid DNA Preparation

The method used is based upon that of Clewell (1972). A

single colony is picked from a plate and grown to saturation in 10ml LB in a bubble tube. 2ml of this culture is used to innoculate another 200ml of LB. Those bacteria harbouring plasmids which carry antibiotic resistance markers are grown in the presence of the relevant antibiotic to prevent loss of the plasmid. When the culture OD_{650} reaches 1.3, chloramphenicol is added to 150µg/ml and shaking is continued at 37°C for a further 15 hours. The cells are pelleted in the cold at 6000 rpm for 5 minutes, using the SS34 rotor, then washed with cold TE Buffer (10mM Tris-HCl pH8.0, 1mM EDTA) and resuspended in 2.25mls of iced 25% sucrose/50mM Tris-HCl pH8.0. 0.75ml of freshly prepared Lysozyme (10mg/ml) is added, and the cells are swirled gently on ice for 5 minutes, or until the mixture becomes very viscous. 0.75ml of 0.5M EDTA is introduced, and the mixture is again shaken on ice for a further 5 minutes, after which time 3.75ml of cold Triton solution (5ml 1M Tris-HCl pH8.0, 12.5ml 0.5M EDTA, 1ml 10% Triton X-100, 80ml d·H_2O) is added. The lysate is swirled on ice for 10 minutes, then centrifuged at 17,000g, 4°C, for 30 minutes. Cell wall material and chromosomal DNA are pelleted together, and the remaining supernatant (cleared lysate) contains the plasmid DNA. The volume (x) of the lysate is measured, and xg of CsCl is added, along with 0.1x ml of 10mg/ml EtBr solution. When the CsCl is thoroughly dissolved, the lysate is sealed in nitrocellulose Ti50 or Ti60 tubes under a layer of mineral oil and spun at 38,000 rpm, 15°C, for 60 hours. Two fluorescent bands are visible under uv light. The upper layer of chromosomal DNA is removed first by side-puncture of the tube. The lower plasmid band is then taken out by the same method. It was found that any attempt to isolate the plasmid first usually causes some

mixing of the gradient as the chromosomal DNA tends to be pulled down towards the syringe needle. Chromosomal DNA, on the other hand, may easily be extracted from the gradient with no disturbance of the plasmid layer, which can then be removed completely and without contamination. The DNA containing EtBr is kept away from sources of bright light to minimise nicking of the duplex. EtBr is removed by passing the solution down a pasteur pipette half-filled with Dowex 50W/X8 (200-400 mesh) cation-exchange resin, or by repeated extraction into isopropanol equilibrated with saturated CsCl. The CsCl is removed through exhaustive dialysis against TE Buffer (10mM Tris-HCl pH 8.0 1mM EDTA) in the cold. DNA is kept over $CHCl_3$ at 4°C or precipitated and stored frozen at 0.5mg/ml.

Rapid Plasmid Preparation

For plasmid screening of large numbers of bacterial clones, a rapid means of preparing DNA is essential. Numerous techniques are available (Barnes, 1977; Telford et al, 1977; Klein et al, 1980). All of these procedures eliminate the time-consuming CsCl equilibrium centrifugation stage and yield enough DNA of sufficient purity to permit a small number of restriction digests. The plasmid is also fit for use as a hybridisation probe following removal of RNA and end-labelling with ^{32}P. The method most commonly applied in this work was a composite of those published by Hepburn and Hindley (1979) and Humphreys et al (1975). Basically, a scaled-down version of the 'cleared lysis' procedure, this preparation is performed on single colonies grown for 2-3 hours in 1ml cultures. All manipulations are carried out in plastic Eppendorf tubes. Two clearing spins at 9000g are employed to remove cell wall debris and chromosomal DNA after lysis. The supernatant is brought to 0.5M in NaCl and 30% PEG-6000 is added to a final concentration of 10%. The plasmid

DNA is precipitated in the Eppendorf microcentrifuge after standing for one hour on ice. The pellet is dissolved in LTB and used directly for restriction enzyme digests or gel electrophoresis.

VII) Purification and Concentration of DNA

Phenol Extraction

The removal of cytoplasmic proteins, restriction endonucleases and phosphatase from DNA samples involves extraction with phenol. The DNA should be in aqueous solution or neutral LTB. An equal volume of buffer-saturated (100mM Tris-HCl pH7.9, 5mM EDTA) re-distilled phenol is mixed with the DNA by vortexing for 1-5 minutes. The phases are separated by centrifuging at room temperature (cooling clouds the aqueous phase and obscures the interface). The upper aqueous phase, containing DNA, may require further extraction if the sample contains much protein. In removing the aqueous layer care should be taken to avoid disturbing any material deposited at the interface. The lower phenol phase is back-extracted by vortexing for 1 minute with 0.5 volumes of 100mM Tris-HCl pH7.9 and the DNA-containing phases are pooled. Phenol is removed by extraction with 0.5 volumes of d·H_2O-saturated $CHCl_3$. The upper phase is kept; phenol partitions into the lower $CHCl_3$ layer. Residual $CHCl_3$ may be evaporated in a stream of N_2. Dilute solutions of DNA are ready for concentration by ethanol precipitation or 2-butanol extraction.

Chromatography on DE52

DEAE-cellulose chromatography is used to obtain DNA of very high purity (Smith and Birnsteil, 1976).

The dry resin is pre-cycled by soaking for 30 minutes in 10 volumes of 0.5M HCl, filtering and washing with d·H_2O until

the pH is around 4. The resin is then subjected to two treatments with 10 volumes of 0.5M NaOH, each of 30 minutes, followed by water washes until the filtrate has a neutral pH.

A column with a 0.3ml bed volume is prepared in a 1ml plastic disposable syringe and equilibrated with DNA buffer (50mM Tris-HCl pH8.0, 150mM NaCl, 10mM EDTA). DNA in the same buffer is passed down the column at a flow rate of 0.1ml/minute. 5ml or more of a dilute DNA solution may be loaded. The binding capacity of DE52 is >200µg/ml bed volume. The column is washed extensively with 10 volumes of DNA buffer until the OD_{260} of the effluent stabilises. The bound nucleic acid is eluted with High-Salt buffer (50mM Tris-HCl pH8.0, 1M NaCl, 1mM EDTA). A small fraction (<1ml) containing all the DNA is collected and ethanol precipitated.

Ethanol Precipitation

A solution of DNA is brought to 0.3M by addition of a 3M stock solution of sodium acetate (pH5.5). 2.5 volumes of cold (-20°C) 99% ethanol are added, and, after mixing, the sample is placed in a dry-ice/isopropanol bath for 5-15 minutes. After freezing, the nucleic acid is pelleted by centrifugation at 12,000 rpm in the SS34 rotor at 4°C for 20 minutes or for 10 minutes at 9000g in an Eppendorf microfuge 5412. Excess salt is removed from the precipitate by washes of cold 95% ethanol. The DNA is dried briefly under vacuum and re-suspended in the required buffer. Recovery is >90%. With dilute solutions containing only a few µg of DNA per ml, precipitation may be enhanced by the inclusion of carrier tRNA at 40µg/ml or nuclease-free BSA at 100µg/ml.

Isopropanol Precipitation

As an alternative to ethanol precipitation, isopropanol

may be used where the volume of sample is very large. One volume of isopropanol is used in place of 2.5 volumes of ethanol. Precipitation may be carried out at room temperatures for 1 hour, but is best performed at -70°C or -20°C, as described above.

Concentration of DNA by 2-butanol

The technique used is described by Stafford and Bieber (1975). It is particularly useful for concentrating plasmid DNA solutions too dilute for efficient ethanol precipitation, and where it is undesirable to introduce carrier tRNA. The 2-butanol forms a biphasic system where water partitions into the upper butanol layer, while DNA and salts remain in the lower aqueous phase.

An aqueous DNA solution is first saturated by shaking with an equal volume of 2-butanol. The phases are separated by low speed centrifugation and the butanol layer discarded. Subsequent similar extractions give rapid reductions in the aqueous volume. A 100-fold concentration is attainable, but the relative ease of the process is determined by the buffer and salt concentration of the sample. High-salt decreases the solubility of water in the butanol phase. DNA is best concentrated from 10mM Tris-HCl pH8.0, 1mM EDTA. 2-butanol is easily removed by extraction with diethyl-ether or by dialysis. The latter simultaneously lowers the increased salt concentration.

Digestion of DNA with Restriction Endonucleases

Most enzymes were used in accordance with the manufacturers' recommendations. Digestion is at 37°C in capped plastic tubes or, for very small volumes, sealed glass capillaries. A full list of restriction enzyme buffers is given elsewhere (Fig. 2b).

Following incubation, the enzymes are inactivated by

Fig.2b Restriction Enzyme buffers

TMN	10mM Tris-HCl pH 7.5 10mM NaCl 10mM $MgCl_2$	Bst I Taq I Hinf I Hae III
Medium Salt TMN	10mM Tris pH 7.5 50mM NaCl 10mM $MgCl_2$ 1mM DTT	Hind III Alu I Pst I Rsa I Pvu II
High Salt TMN	10mM Tris-HCl pH 7.5 100mM NaCl 10mM $MgCl_2$	EcoR I
TMK	10mM Tris-HCl pH 7.8 20mM KCl 10mM $MgCl_2$ 1mM DTT	Sma I

heating at 65°C for 10 minutes or by extraction with phenol. Prior to electrophoresis, reactions are terminated by the addition of 0.1 volumes of 50% glycerol containing 100mM EDTA and 0.1% BPB. The non-covalently joined cohesive ends generated by restriction enzyme cleavage of DNA are then destroyed by heating to 65°C.

VIII) Recovery of DNA from Gels

High Salt Elution

The method of Gilbert and Maxam (1980) was used initially, but DNA isolated in this way often contains substantial quantities of low MW acrylamide, particularly when working with 4% gels, and so alternative methods were sought.

Electroelution of DNA fragments

DNA is eluted from agarose or acrylamide gels by a modification of the method employed by Galibert et al (1974). 5ml glass pipettes are cut to a length of 15cm and siliconised. A length of sterile, 1cm diameter dialysis tubing is fitted firmly over the pipette tip to give a watertight seal, and then secured with plastic adhesive tape. The tubing is tied to form a small bag capable of holding about 1ml of buffer. A siliconised glass wool plug is tamped down into the tip, separating the bag from the pipette, and the assembly is filled with elution buffer (40mM Tris-Acetate pH8.2, 0.1mM EDTA), taking care to remove any air bubbles. Gel slices containing DNA are inserted to rest on the plug, and the dialysis bag is immersed in a reservoir containing elution buffer and the anode of a D.C. power supply. The top of the pipette is connected via a capillary bridge to a second buffer trough and the cathode. Electrophoresis is carried out at 400V, 4 to 5 hours being sufficient for >95% recovery of fragments up to 1kb long.

The DNA solution is removed from the dialysis bag with a syringe, precipitated with ethanol and dried. Carrier tRNA is replaced by BSA if the DNA is to be subjected to Maxam and Gilbert sequencing.

Phenol Extraction of DNA fom Agarose Gels

This is a very rapid method, giving good yields of DNA of sufficient purity for cloning and sequencing purposes. The DNA is cut from the gel with the minimum of agarose around it. One gel-volume of phenol (saturated with 100mM Tris-HCl, pH7.9) is added, and the mixture is incubated at 65°C-70°C, with frequent vortexing, for 15-20 minutes. Longer incubations may be necessary if the gel contains >1% w/v of agarose. The resultant 'milky' liquid is centrifuged in an Eppendorf microfuge for 5 minutes, and the upper aqueous layer is removed. A second phenol extraction is usually required, after which the phenol layer is back extracted with 0.5 volumes of TE buffer. The pooled aqueous phases are precipitated with ethanol.

If gels are prepared with LGT Agarose, then the incubation step is performed at 50°C.

Extraction of DNA with cetyl-trimethylammonium bromide

The technique used by Langridge et al (1980) employs quaternary ammonium compounds to yield highly purified DNA in one straightforward operation. It is based on the partition of nucleic acids into 1-Butanol as their quaternary ammonium salts, leaving neutral agarose in the aqueous phase. DNA is then recovered into the aqueous phase using high salt concentrations. All solutions were prepared as described in the original paper, and the protocol was strictly adhered to.

IX) Labelling DNA with ^{32}P

5'-end labelling using T4 Polynucleotide Kinase

5'-single stranded termini are generated by restriction

enzyme cleavage, and the terminal 5' γ-phosphate groups are then removed by treatment with 0.1 units of alkaline phosphatase at 37°C for 30 minutes. The phosphatase is then removed by phenol extraction. After ethanol precipitation, dephosphorylated DNA is dissolved in 20µl of Kinase Buffer (50mM Tris-HCl pH7.6, 10mM $MgCl_2$, 5mM DTT, 0.1mM spermidine, 0.1mM EDTA, 5% glycerol) which contains 150-200µCi (50-65pMoles) of γ-[^{32}P]-rATP (Specific Activity >3000Ci/mMol). The final concentration of isotope is approximately 5µMolar. T4 Kinase (40 units) is added, and the reaction is allowed to proceed at 37°C for 30 minutes. Unlabelled rATP is added to 0.5mM, and the reaction is terminated by phenol extraction, then precipitated once more. DNA, ^{32}P labelled at both ends, is re-cut with a second restriction enzyme which will generate singly labelled molecules ready for Maxam and Gilbert sequence analysis. These ^{32}P-labelled fragments are resolved on 4% Acrylamide or 0.4-1.5% Agarose gels and eluted as described.

3'-end labelling with DNA Polymerase I

The 5' single stranded extensions produced by certain restriction enzymes form templates which may be "filled in" by DNA Polymerase. The polymerase used is the Klenow fragment (Klenow et al, 1971), which lacks 5' → 3' exonuclease activity.

Between 20µCi and 50µCi of ∝-[^{32}P]-dATP (2000Ci/mMol) is dried under vacuum and dissolved in 30µl of Polymerase Buffer (60mM Tris-HCl pH7.4, 125mM NaCl, 15mM $MgCl_2$, 2.5mM DTT and 125µM dNTP, if needed). This mixture is transferred to 10-50µg of restriction fragments and vortexed to dissolve the DNA. Incubation is performed at 20°C for 30 minutes in the presence of 0.5 units of DNA Polymerase I followed by a 10 minute chase with 100µM "cold" dATP. Labelled nucleic acids are precipitated at -70°C after the addition of 200µl of 2.5M ammonium acetate

pH5.5 and 3 volumes of 99% ethanol. The pellet is re-precipitated from 0.3M sodium acetate pH5.5 and finally washed once with cold 95% ethanol. Single-labelled DNA molecules are produced by further restriction enzyme digests.

3'-end labelling with cordycepin triphosphate and Terminal deoxynucleotidyl Transferase

This method was used to label Pst I fragments which have 3' ssDNA extensions and are not easily labelled by T4 Kinase and γ-[^{32}P]-ATP. The technique is based upon that of Tu and Cohen (1980).

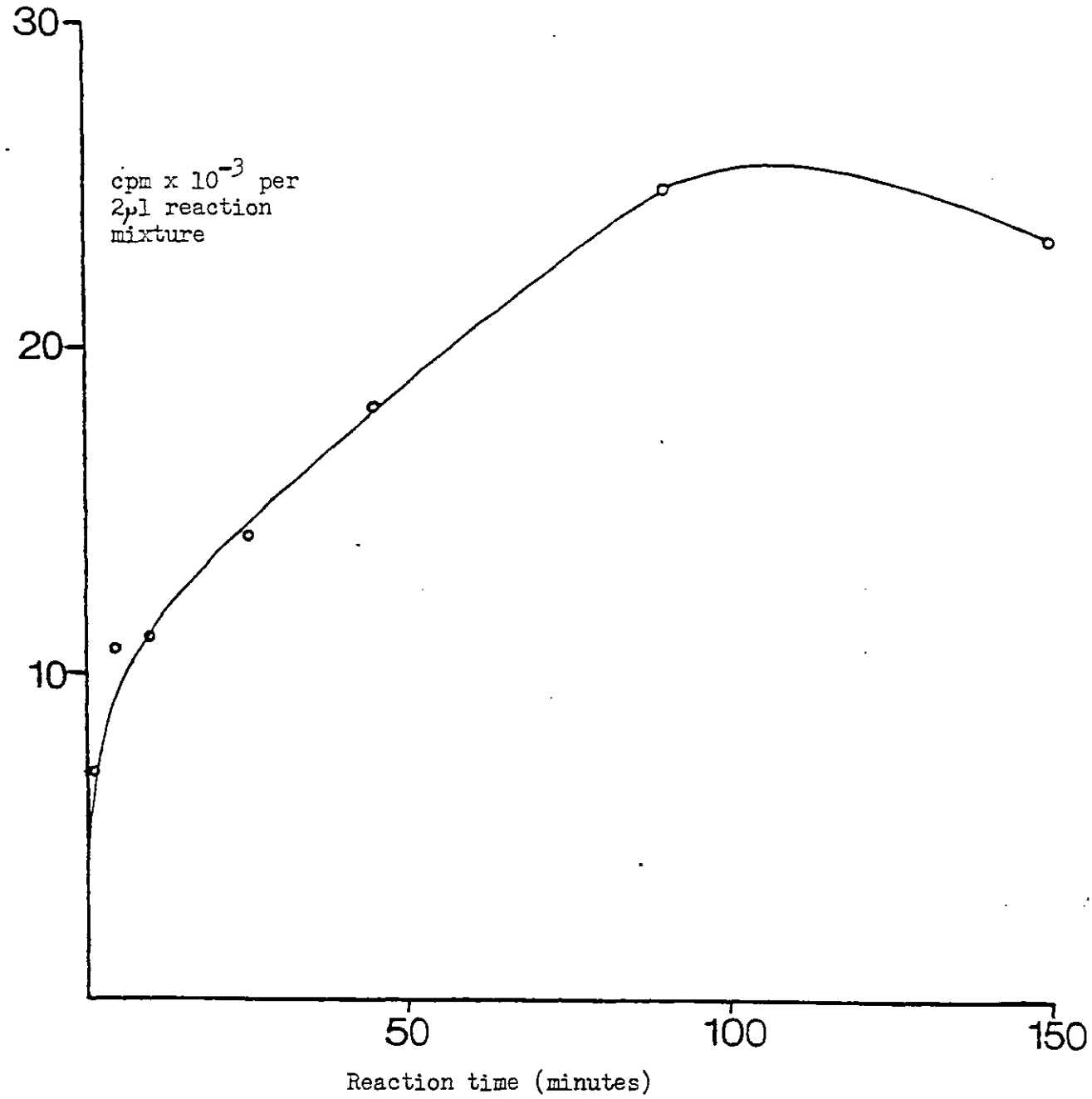
Pst I digested DNA is phenol extracted, precipitated and dried under vacuum. About 100µCi of α-[^{32}P]-3'dATP (390Ci/mMol) is dissolved in 200µl of Transferase buffer (140mM Potassium cacodylate, 30mM Tris, 1mM $CoCl_2$, 0.2mM DTT (pH6.9): prepared as described by Roychoudhury and Wu, 1980). The DNA pellet is taken up in the buffer/isotope mixture, and 50-80 units of TdT are introduced. Incubation is at 37°C until incorporation of ^{32}P ceases, as determined by essays on 2µl aliquots (Fig. 2c). The reaction is terminated by the addition of EDTA to 5mMolar, and the labelled DNA is ethanol precipitated. It is necessary to re-cut these fragments or separate the strands before Maxam and Gilbert sequencing can be started.

X) DNA Sequencing

Maxam and Gilbert Procedure : Partial Chemical Degradation

DNA fragments singly labelled at either the 3' or 5' terminus were sequenced by a modification of the published protocols (Maxam and Gilbert, 1980). Experimental details are summarised in Table 2(iv). Reaction times vary according to the size of the DNA being sequenced, smaller DNA's receiving shorter incubations.

Fig.2c Incorporation of radioactivity into PstI cut DNA by Terminal Deoxynucleotidyl Transferase and Cordycepin-5'-triphosphate.



PstI digested pRD351 was labelled with cordycepin triphosphate and TdT as described by Tu and Cohen (1980). 2µl aliquots of the reaction mixture were removed at intervals to 2.5cm Whatman 3MM discs, fixed in 5% TCA, dried and counted for Cerenkov radiation (Bollum,1959). A 0.2ml volume contained 100µCi cordycepin (400Ci/mMol), 85 units TdT and 125pMolar 3' ends in Transferase buffer (140mM potassium cacodylate, 30mM Tris, 1mM $CoCl_2$, 0.2mM DTT, pH 6.9.

Table 2iv Maxam and Gilbert Base-Specific Cleavage Reactions

G	G + A	T + C	C
200µl DMS buffer	10µl water	10µl water	15µl 5M NaCl
1µl carrier DNA	1µl carrier DNA	1µl carrier DNA	1µl carrier DNA
5µl ^{32}P-DNA	10µl ^{32}P-DNA	10µl ^{32}P-DNA	5µl ^{32}P-DNA
Chill to 0°C	*Chill to 0°C*	*Chill to 0°C*	*Chill to 0°C*
1µl DMS	2µl pyr-for, pH2	30µl hydrazine	30µl hydrazine
207 µl	23 µl	51 µl	51 µl
20°C, 10±5 min	*20°C, 60±20 min*	*20°C, 10±5 min*	*20°C, 10±5 min*
50µl DMS Stop	*Freeze*	200µl HZ Stop	200µl HZ Stop
750µl ethanol	*Lyophilize*	750µl ethanol	750µl ethanol
Chill 5 min	- - - - - -	*Chill 5 min*	*Chill 5 min*
Centrifuge 5 min	20µl water	*Centrifuge 5 min*	*Centrifuge 5 min*
Pellet	*Lyophilize*	*Pellet*	*Pellet*
250µl 0.3M NaAc	- - - - - -	250µl 0.3M NaAc	250µl 0.3M NaAc
750µl ethanol	- - - - - -	750µl ethanol	750µl ethanol
Chill 5 min	- - - - - -	*Chill 5 min*	*Chill 5 min*
Centrifuge 5 min	- - - - - -	*Centrifuge 5 min*	*Centrifuge 5 min*
Pellet	- - - - - -	*Pellet*	*Pellet*
Ethanol rinse	- - - - - -	*Ethanol rinse*	*Ethanol rinse*
Vacuum dry	- - - - - -	*Vacuum dry*	*Vacuum dry*
20 µl 1.0M piperidine	20 µl 1.0M piperidine	20 µl 1.0M piperidine	20 µl 1.0M piperidine
90°C, 30 min	*90°C, 30 min*	*90°C, 30 min*	*90°C, 30 min*
Lyophilize	*Lyophilize*	*Lyophilize*	*Lyophilize*
10µl water	10µl water	10µl water	10µl water
Lyophilize	*Lyophilize*	*Lyophilize*	*Lyophilize*
10µl water	10µl water	10µl water	10µl water
Lyophilize	*Lyophilize*	*Lyophilize*	*Lyophilize*
10µl formamide-buffer-dyes	10µl formamide-buffer-dyes	10µl formamide-buffer-dyes	10µl formamide-buffer-dyes
90°C, 1 min	*90°C, 1 min*	*90°C, 1 min*	*90°C, 1 min*
Quick-chill	*Quick-chill*	*Quick-chill*	*Quick-chill*
Load on gel	*Load on gel*	*Load on gel*	*Load on gel*

In some instances the 'A' and 'G' reaction mixtures are divided and one half given a longer incubation. The samples are then recombined prior to precipitation. This gives a more satisfactory size distribution of cleavage products.

M13 Cloning and Chain-Termination Sequencing

The ssDNA bacteriophage M13 has been developed into a cloning vector by the introduction of a region of the E.coli lactose operon (Messing et al, 1977). The original phage vector has been further improved by the insertion of a synthetic 42bp oligonucleotide containing useful restriction sites into the N-terminal portion of the β-galactosidase gene fragment (Messing et al, 1981). Recombinant phage carrying foreign DNA ligated into any one of these unique sites produce colourless plaques on the appropriate indicator medium, compared with the blue plaques arising from those host bacteria which are infected by the "wild-type" virus. The phage particles secreted by M13 infected cells in liquid culture conditions provide an ideal source of template DNA for analysis by the dideoxy chain-termination method of Sanger et al (1977). Since all DNA fragments are cloned into the same region of the M13 genome, all can be primed for the sequencing reaction by a single, universal primer.

The "shotgun" strategy of Sanger et al (1980) was applied to obtain sequences for some parts of the dal operon. DNA is cloned as randomly generated fragments into unique restriction sites in genetically engineered versions of the phage M13.

M13 mp7 replicative form (RF) DNA was prepared by the cleared lysate procedure of Hines and Ray (1980). Bacteriophage M13 mp7, E.coli K12 JM101 and synthetic 17mer single stranded primer were generous gifts from Dr. G. Winter.

The techniques for cloning in single-stranded phage vectors

have been extensively reported in the literature (Schreier and Cortese, 1979; Heidecker et al, 1980; Sanger et al, 1980; Messing et al, 1981). Any modifications to the published protocols are cited in the text.

Experimental details of the cloning and sequencing strategy are listed in Chapter 3. Transformation of E.coli JM101 with M13 and purification of SSDNA from recombinant phage are described by Winter and Fields (1980). Chain termination sequencing was carried out essentially as described by Sanger et al (1977 and 1980) using a universal primer, (Anderson et al, 1980). The techniques are summarised in Fig. 2d.

Computer Analysis of DNA and Protein Sequences

Computer analyses were performed with the aid of programs supplied by R. Staden (MRC, Cambridge). These programs were rewritten in FORTRAN IV for compatibility with the Imperial College CD6500 and IBM400 computers by S. Cox (Dept. of Computing, IC).

Comparisons of the amino acid sequences of ArDH and RDH were carried out at the MRC Laboratory of Molecular Biology, Cambridge, using the programs of Dr. A. McLachlan.

XI) Electrophoresis

Agarose Gel Electrophoresis

Horizontal slab gels are prepared in the Tris-Borate-EDTA (TBE) buffer system of Peacock and Dingman (1968). Agarose (M.E. grade) is used at concentrations of 0.4% to 2.5% w/v. EtBr is included at 0.5µg/ml. The gel is prepared by boiling the appropriate quantity of agarose powder in 90ml d·H_2O + 10ml [10x]TBE. [10x]TBE is 108g Tris base, 55g Boric acid and 9.3g EDTA in 1L d·H_2O (the final pH is 8.3). Once dissolved, the agarose is allowed to cool to around 45°C, and EtBr is

Fig. 2d Conditions for Chain-Termination Sequencing

a) Hybridisation of M13 ssDNA and Primer

The following are mixed in a glass capillary:

5μl ss M13 DNA (<0.5μg)
1μl (10x) Hin. Buffer
0.5μl 17mer Primer
3.5μl deionised water
10μl

The capillary is sealed and put at 100°C for 3 minutes then left to cool slowly to room temperature over 30 minutes.

b) Polymerisation Reactions

Each reaction mix (6.2μl) contains 2μl of annealed template/primer plus the following:

	T	C	G	A
α-[^{32}P]-dATP	1μCi	1μCi	1μCi	1μCi
dTTP	2.5μM	50μM	50μM	50μM
dCTP	50μM	2.5μM	50μM	50μM
dGTP	50μM	50μM	2.5μM	50μM
ddTTP	175μM	-	-	-
ddCTP	-	60μM	-	-
ddGTP	-	-	125μM	-
ddATP	-	-	-	60μM
DNA Pol I (Klenow)	0.1U	0.1U	0.1U	0.1U

Fig. 2d (contd)

The polymerisation reaction is allowed to proceed in open-ended capillaries or Eppendorf tubes at room temperature for 15 minutes, followed by a 15 minute chase with 2μl 0.5mM dATP.

4μl of Formamide containing 20mM EDTA, 0.03% XC, 0.03% BPB are added and the samples are boiled for 5 minutes. Between 1-2μl are used for each gel loading. DNA is stored at 4°C in Formamide + dyes or frozen without Formamide if they are not to be loaded immediately.

The hybridisation buffer (Hin) is 10mM Tris-HCl pH 7.4, 50mM NaCl, 10mM $MgCl_2$, 10mM DTT. Omission of salt from Hin. buffer sometimes reduces the background counts on sequence gels.

added before pouring into a shallow 25 x 15 x 0.6cm perspex tray. The gel sets around a perspex comb, which, when removed, creates the sample wells and is allowed to stand for up to one hour before use. Samples are loaded from finely drawn capillaries, a few µl of tracking dye (0.05% BPB in 50% glycerol, [1x]TBE) being added to each prior to loading. The electrophoresis buffer is [1x]TBE and separation is carried out in a horizontal slab gel apparatus, using Whatman 3MM paper wicks. Voltages of between 200V and 400V are applied, such that the BPB marker migrates to 3/4 of the gel length in 3-4 hours. The sample wells are filled with running buffer once the dye has entered the gel, and the whole gel, including wicks, is covered with 'Saran wrap' to prevent dehydration. Gels may also be run at 120V overnight to give better resolution, particularly where fragments are to be eluted for further analyses.

Low Gelling Temperature (LGT) Agarose Gels

Gels are prepared by heating the LGT Agarose to 67°C in [1x]TBE with EtBr and pouring as previously described for HGT gels. Polymerisation is allowed to proceed at 4°C for several hours before attempting to remove the comb and load any samples. Electrophesis may be carried out at room temperature, but the gel should not get too warm.

Submerged Agarose Gel Electrophoresis

Very small amounts (<0.1µg) of DNA are routinely screened on miniature agarose gels immersed in [1x]TBE. The separation time is extremely short, and this technique is ideally suited to checking for completeness of restriction digests or ligation reactions, since only a tiny amount of material is needed.

15ml portions of 1% w/v agarose in [1x]TBE are stored at 4°C and melted as required. When cooled to 45°C, the gel is poured into a 45 x 25mm template and allowed to set around a perspex

comb (slot size 4mm x 1mm). 25ml of [1x]TBE containing 0.5µg/ml EtBr is added to just cover the gel surface and DNA samples plus glycerol and BPB are loaded into the wells beneath the buffer. Electrophoresis is at 50mA for 20-40 minutes after which gels are examined under UV light.

Larger submerged gels (100ml volume) are run overnight at 20V in BRL submarine gel tanks (Model H2) for restriction mapping purposes.

Visualisation of DNA in Agarose Gels

Fluorescence of ssDNA or dsDNA as a result of EtBr binding is visualised by illumination with short-wave (300nm) uv light and photographed through a Kodak 'Wratten' yellow filter on Polaroid type 105 film. Exposure times of 10 seconds to 1 minute are sufficient. Long wave uv light and shorter exposure times are given if the DNA is subsequently to be excised for cloning or sequencing experiments.

Non-denaturing Polyacrylamide Gel Electrophoresis

^{32}P-labelled DNA fragments are separated using 4% or 8% w/v, non-denaturing polyacrylamide gels. The technique used is basically a modification of that for preparing and running DNA sequencing gels. The gel is formed between two glass plates (40 x 20 x 0.3cm), and is run vertically.

4% gels are prepared from 5.6g acrylamide, 0.4g methylene-bisacrylamide, 7.5ml [10x]TBE, and 1ml of freshly prepared 10% APS dissolved in d·H_2O to a final volume of 150 mls. The solution is degassed in a Buchner flask under vacuum. Polymerisation is initiated with 50µl of TEMED.

The template consists of two glass plates separated by 1.6mm thick perspex spacer strips held together and sealed along 3 sides by adhesive tape ("Tucktape", New Rochelle, N.Y.). At the unsealed end one plate is cut out to allow buffer contact

with the top reservoir. The template is held at about 45° to the horizontal, and the gel solution is poured in carefully from a beaker. A perspex comb of the same thickness as the spacers is inserted between the plates, and the gel is left to set for at least 2 hours before use at an inclination of about 5° to horizontal. Before running, the tape is slit along the bottom edge, and the assembly is then mounted in the electrophoresis tower. 300ml of [½x]TBE is poured into both buffer chambers. The comb is removed beneath the buffer surface, and the wells are flushed immediately to expel any unpolymerised acrylamide which may otherwise set. Pre-electrophoresis is not necessary, and samples are loaded straight away using finely drawn glass capillaries. A tracking dye (50% glycerol, 0.05% BPB, 0.05% XC) is included in all samples. Each individual well is washed out immediately before loading: This is particularly important for denaturing gels where the urea leaches out into the buffer, causing the DNA to smear as it enters the gel unevenly.

Electrophoresis is performed at 100V overnight or at 400V for a more rapid (3-4 hours) daytime separation. The run is ended when the XC marker has migrated to about half the gel length. The glass plates are separated by cutting the tape along the edges and prising them apart, when the gel should adhere to just one plate. Autoradiography is carried out, as described in "General Techniques", and portions of gel containing labelled DNA can be cut out as required.

Denaturing Polyacrylamide Gel Electrophoresis

The thin gel system of Sanger and Coulson (1978) was used. Preparation of the gels is essentially similar to that described for non-denaturing gels, but with the following exceptions.

Gels are made up in deionised water to a total volume of 50ml, and are comprised of 25g "Ultra Pure" Urea, 350µl 10% APS, 5ml [10x]TBE and sufficient of a 40% w/v acrylamide stock solution (20:1, Acrylamide : Bis) to give a final concentration of 8, 12 or 20%. The acrylamide stock is deionised by stirring for 30 minutes with 5g of Amberlite AMB-1 resin (BDH) and is filtered and stored at 4°C. Gel solutions are filtered through Nalgene 20µm filter units and de-gassed prior to pouring. The template is similar to those used for non-denaturing gels, except that the spacers are 0.35mm thick strips of "Plastikard" (Raven Scientific Ltd.). Gels are polymerised by the addition of 25-30µl of TEMED and allowed to set at a few degrees from the horizontal for at least one hour before use. Standing overnight does no harm as long as the gel around the comb does not dry out. The gel is set up and loaded as previously described. 0.5 volumes of formamide/dye mix (98% formamide twice re-crystallised, 0.02% XC, 0.02% BPB, 20mM EDTA) is added to each sample prior to loading, and the volume loaded does not exceed 2µl. The running buffer is [1x]TBE and is changed every 2 hours during electrophoresis. There is no period of pre-electrophoresis for 8% or 12% gels, but 20% gels are pre-run at 15mA for 1 hour. Separation is performed at 30-40mA with the power limited at 45W. DC power supplies for DNA sequence gels must be capable of delivering a stable high voltage output such as the LKB 2103 or Pharmacia ECPS 3000/150.

Separate applications of the same sample are usually made on one gel at intervals calculated to resolve the first 50-100 nucleotides from the labelled end, and intermediate sequences up to a maximum of 350 base pairs. Three loadings will normally be required. A greater proportion of the radioactivity tends to be present in small to intermediate sized oligonucleotides when

using the Maxam and Gilbert method, hence the amount of DNA per loading is varied to compensate for this. The running time to achieve any desired separation is variable and is determined from the migration of the marker dyes. This may take between 2 hours and 8 hours, depending on (a) the gel percentage, (b) how far into a sequence it is desirable to read, and (c) the voltage that can safely be applied without cracking the glass plates. For fixing and autoradiography of sequence gels, see "General Techniques" section.

Urea/Formamide Sequence Gels

For eliminating problems caused by stable secondary structures in DNA, 6% polyacrylamide gels containing 8M urea and 25%v/v formamide were used. Each gel is composed of 25g urea, 12.5ml formamide (twice re-crystallised from $d{\cdot}H_2O$), 10ml of 30% acrylamide/bis-acrylamide stock (19:1, acrylamide : bis). 5ml [10x]TBE, 350µl 10% APS made up to 50ml with deionised water. Polymerisation uses 35-40µl TEMED and procedures are identical to those for conventional sequence gels. Extra care is taken when prising apart the plates following electrophoresis, since these gels are softer and more fragile.

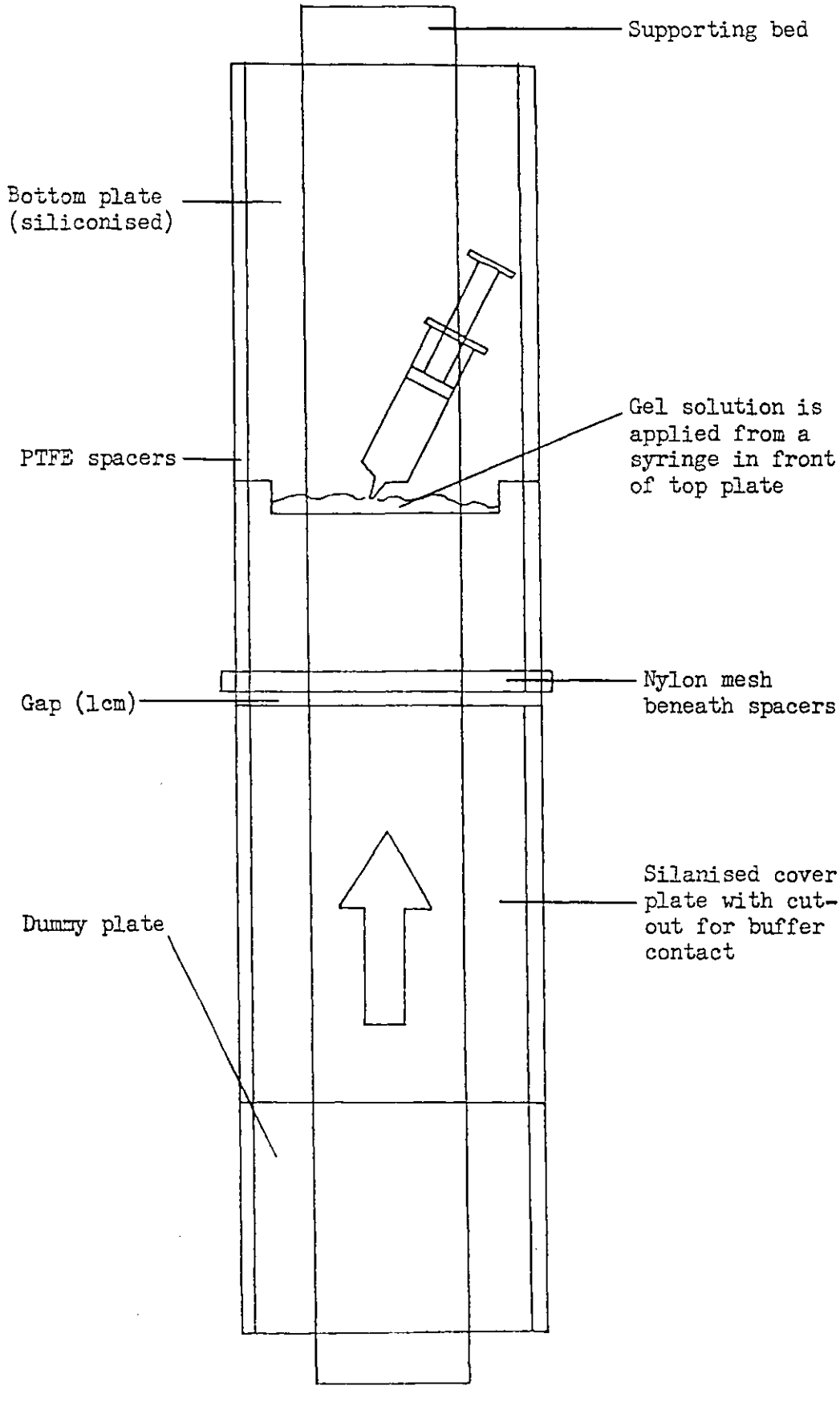
Ultra-Thin Temperature Controlled Polyacrylamide Sequence Gels

The technique of Ansorge and De Maeyer (1980) was applied to DNA sequenced by the dideoxynucleotide method of Sanger et al (1980).

The gel is cast directly between a glass plate and a thermostatting platten through which hot water circulates to maintain isothermal conditions during electrophoresis. This method is technically more demanding than the thin gels described by Sanger and Coulson (1978), but has certain advantages which will be discussed later. The procedure is as follows. The surfaces

of both plates are thoroughly cleaned with ethanol and the thermostatting plate is siliconised by wiping with "Repelcote" in order to prevent the gel from adhering to its surface and to facilitate pouring. The glass cover-plate (with a notch cut out to fit the upper buffer tank) is treated with a mixture of 25ml 99% ethanol, 750µl 10% acetic acid and 75µl silane (γ-methacryloxypropyl trimethoxysilane) which is allowed to dry and is then polished off with a lint-free tissue, and finally washed with ethanol. The prepared thermostatting plate is laid out on a flat horizontal supporting bed adjacent to a dummy glass plate of identical thickness, so that there exists between the two a gap of several cm. PTFE spacer strips of 100µm or 200µm thickness are laid along the plates at their outer edges, and the cover plate is laid on top (silanised side down), so that it covers most of the dummy plate and a few mm of the thermostatting plate (Fig. 2e). The gel solution, with TEMED added, is applied in front of cover plate, which is moved slowly forwards, thus spreading the gel evenly between the two plates. When the top and bottom plates are perfectly aligned, and the space between them is filled with gel solution, a 20-slot PTFE comb is inserted, and the assembly is clamped together with butterfly clips. Capillarity prevents the liquid from leaking out. When set, gels are placed in electrophoresis towers along with [1x]TBE and the slots are cleaned immediately the comb is removed. Water is circulated from bottom to top through the thermostatting plate by a Thermomix 1441 unit (B. Braun, Melsungen, FRG). The temperature is taken from 30°C to 65°C or higher in 5°C steps at 5 minute intervals, while pre-electrophoresis is carried out at 1000V. Once the required running temperature is attained, 1.5µl samples are loaded from 100µm diameter X-ray

Fig.2e Apparatus for casting ultra-thin sequence gels



analysis capillaries (A. Müller, Berlin, FRG), and the voltage is increased to 2000V. An even faster separation may be achieved at 3000V, but some resolution is sacrificed.

Following electrophoresis, the gel sticks firmly to the cover plate since it is covalently linked to the acrylic groups of the silane molecules. This enables the gel to be handled freely without separating from the glass, and it can be dried without shrinkage or cracking. Methods of fixing and autoradiography have already been described.

To remove the gel from the plate it is necessary to soak overnight in concentrated Decon 90 solution.

CHAPTER 3

CLONING AND SEQUENCING OF THE d-ARABITOL DEHYDROGENASE STRUCTURAL GENE

Although the entire d-arabitol operon is carried on the transducing phage λprbt dal, lambda is not an ideal vector for DNA that is to be sequenced by the Maxam and Gilbert method. End labelling of digested phage DNA generates many unwanted fragments and is wasteful of restriction enzymes. The isolation of λ DNA itself is a tiresome process and yields can be extremely variable. Thus, to facilitate DNA sequencing, restriction fragments from λp rbt dal were cloned into smaller, plasmid vectors.

Amplification of Col E1 type plasmids

Plasmid Col E1 and its derivatives replicate in the relaxed mode (Timmis et al, 1974). Exponentially growing cultures have about 20-30 copies of the plasmid per cell. Replication is not dependent on protein synthesis, but does require an active DNA Polymerase I. Chloramphenicol and other protein synthesis inhibitors which block chromosomal DNA replication permit each cell to accumulate several hundred plasmid molecules: up to 50% of the total cellular DNA content. Milligram quantities of plasmid DNA may be isolated from a 1 litre culture.

I) Sub-cloning fragments from λp rbt dal

It was decided to use the small Col E1-type plasmid pBR322 (Bolivar et al, 1977) as the cloning vehicle. This carries a single BamHI site in the Tc gene and unique sites for EcoRI, Pst I, Hind III and Sal I. The prime consideration in selecting this vector was that its complete nucleotide sequence had recently been deduced (Sutcliffe, 1978a and 1978b), making it

much simpler to restriction map cloned DNAs.

Phage DNA (40 µg) and pBR322 (4 µg) were digested with Bst I and the mixture re-ligated overnight at 10°C in 150 µl of C-Buffer (50mM Tris-HCl pH7.5, 10mM $MgCl_2$, 1mM rATP, 1mM DTT) as described in Materials and Methods. 0.1 ml of ligated DNA was used to transform 0.2 ml of $CaCl_2$-treated E.coli HB101. Cells were grown in LB + ampicillin to select transformants and then d-cycloserine enriched to eliminate Tc^R clones arising from re-circularisation of the vector. Appropriate dilutions of the enriched culture were plated on TYE + Ap (100 µg/ml). Single colonies (200) were toothpicked on to further TYE + Ap plates, incubated for 15 hours at 37°C, then replica-plated on to TYE + Tc (25 µg/ml). Out of 200 colonies screened in this way, 196 were $Ap^R Tc^S$ showing that a d-cycloserine enrichment step is extremely effective in reducing the background of $Ap^R Tc^R$ transformants.

Screening transformants

Mini plasmid preparations were performed on a number of $Ap^R Tc^S$ colonies to facilitate identification of the cloned fragments by restriction mapping. Clones containing four different sized inserts were chosen for full scale plasmid DNA preparations, as described in Materials and Methods, to provide sufficient material for DNA sequence analysis.

II) Characterisation of recombinant plasmids

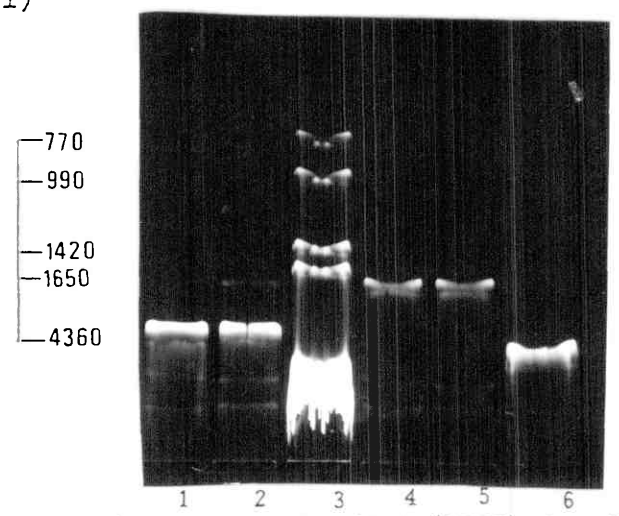
DNA sequencing by the Maxam and Gilbert (1980) method requires only a partial restriction map of the relevant DNA segment since it is hoped to discover new sites as the sequence is gradually compiled and to utilise these for further labelling and sequencing. Hence, the plasmids constructed here were not extensively mapped for all of the restriction enzymes which were to be used later in end-labelling experiments. Instead,

one of the main aims was to ascertain the orientation of each insert relative to the vector. Full restriction maps are most easily derived by computer-aided examination of the nucleotide sequence itself. Where the vector is already well characterised it is often sufficient to know whether or not an inserted DNA is cut by a particular enzyme and, if so, whether the fragments are a) of a size suitable for sequencing, and b) electrophoretically separable from other labelled species.

When digested with Bst I, plasmids were identified containing inserts of 0.77 kb, 0.99 kb or 1.42 kb and one species had both the 0.99 kb and 1.42 kb fragments. These sizes correspond to the λp rbt dal regions BstB/C, BstD/E and Bst C/D respectively. A fourth plasmid, pRD253, migrated as a single band of about 1.7 kb in both the uncut state and when digested with Bst I. Hind III or Pst I generated a slower moving fragent of about 3 kb (photo 3A), implying that this plasmid is deleted for a region around the Bst I site, inactivating the tetracycline resistance gene. Out of curiosity it was decided to investigate the extent of the deletion. Hinf I digests showed that pRD253 lacked 5 fragments compared with pBR322 (photo 3A). These pieces are all adjacent on the pBR322 map and are probably lost by a single deletion. The largest Hinf I fragment is bigger than the expected 1.63 kb, suggesting that the latter is fused to a smaller DNA by the deletion of an intervening Hinf I site. As this mini-plasmid derived from a Bst I cloning and knowing that Bst I* activity is very widespread (Clarke and Hartley, 1979), it is proposed that pRD253 was formed by joining of the Bst I site at position 375 in pBR322 with the Bst I* (Mbo I) site at 1666. This arrangement causes the loss of five Hinf I fragments (506, 298, 221, 220 and 154 bp), the formation of a large Hinf I piece (1.739 kb) and the

Photograph 3A

(i)



(1) pRD253 Hind III/Bst I

(2) pRD253 Hind III

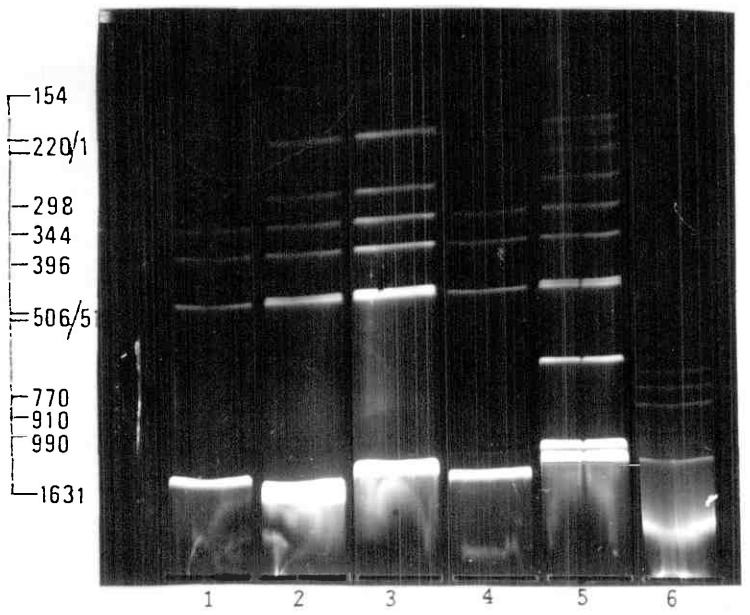
(3) λprbtdal Bst I

(4) pRD253 Bst I

(5) pRD253 Uncut

(6) pBR322 Bst I

1.4% agarose gel. 20mA (250V) for 5 hours



(1) pRD253 Hinf I

(2) pRD251 Hinf I

(3) pBR322 Hinf I

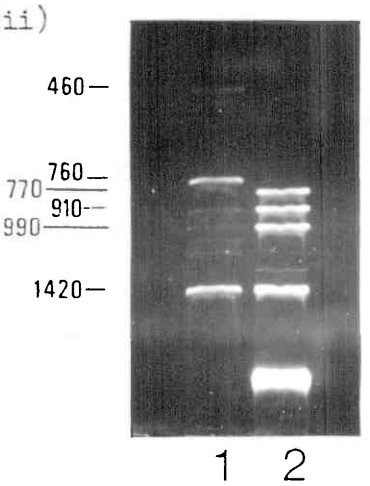
(4) pRD253 Hinf I/Bst I

(5) pRD251 Hinf I/Bst I

(6) pRD351 Bst I

2% agarose gel. 30mA (300V) for 3 hours

(ii)

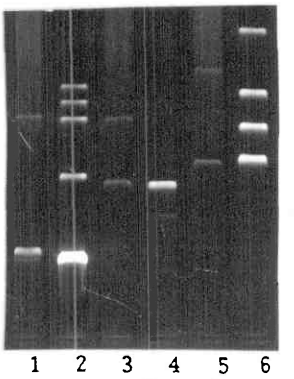


(1) Bst I digest of a 2.7kb HinfI fragment from pRD351

(2) pRD351 + Bst I

1.5% agarose gel. 200V for 3 hours

(iii)



1.5% agarose gel.4 hours at 100V

1) pRD257 + Bst I
2) pRD351 + Bst I
3) pRD257 + Hinf I
4) pBR322 + Hinf I
5) pRD257 + Bst I + Hinf I
6) ϕX174 + Hae III

destruction of the Bst I site. The resultant plasmid has a size of 2.96 kb and codes for resistance to ampicillin. The generation of pBR322 deletion mutants by BamHI has also been observed recently by other workers (Garaev et al, 1982).

Plasmid pRD251

This plasmid contains the BstD/E and BstC/D fragments. Bst I/Hinf I double digests reveal that the 0.99 kb BstD/E piece is cut by Hinf I to yield two smaller species running at 760 bp and 230 bp (Photo 3B). BstC/D has a single Pst I site 240 bp from BstD. The orientation of the two Bst I fragments is deduced from the Pst I and Hinf I digestion patterns (Photo 3B, Fig. 3a(i)) and from Pvu II/Rsa I mapping (Photo 3c). They do not retain the same arrangement as in the λp rbt dal genome or as in pRD351. A 2.7 kb Hinf I fragment can be isolated from pRD351 and re-cut with Bst I to generate 1500 bp, 760 bp and 460 bp fragments (Photo 3A). The 460 bp piece originates from BstB/C, and the 1500 bp fragment is BstC/D; hence, the Hinf I site in BstD/E must lie closer to BstE (Fig. 3a(ii)).

Plasmid pRD252

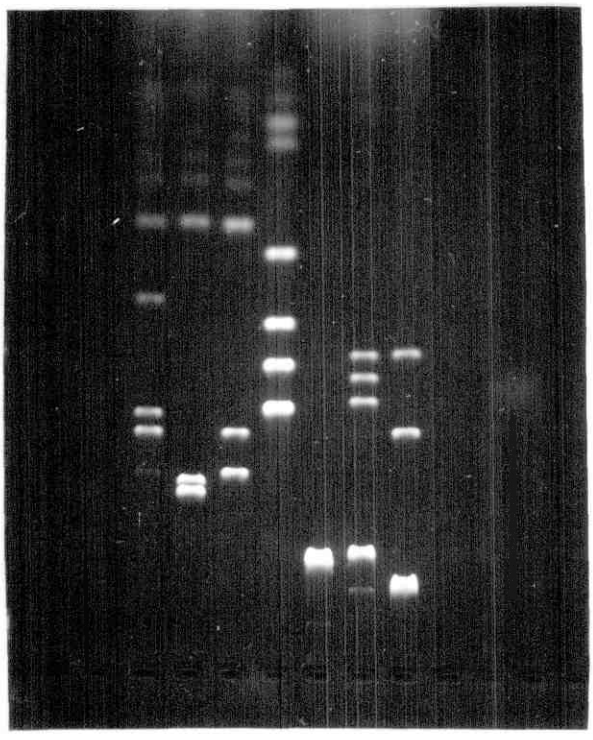
This plasmid possesses the 1.42 kb BstC/D fragment oriented as shown in Fig. 3a(iv). A Pst I digest of the plasmid (Photo 3D) generates two fragments only (2.3 kb and 3.4 kb), which is consistent with this structure. A map of Rsa I, Pvu II and Bgl I sites is shown along with Photo 3E.

Plasmid pRD256

pRD256 contains a 0.77 kb insert (Photo 3D) corresponding to BstB/C. 1.8 kb and 576 bp Hinf I fragments can be identified and serve to orient the insert as shown in Fig. 3a(iii).

Photograph 3B

Restriction digests of plasmid pRD251



1 2 3 4 5 6 7

1.4% agarose gel. 6mA (15V) for 15 hours

(1) Bst I + Hinf I — 1420, 1376 , 760, 256, 230

(2) Hinf I — 2130, 1980

(3) Pst I + Hinf I — 2130=1875 + 250, 1980 = 1420 + 496 (triplet)

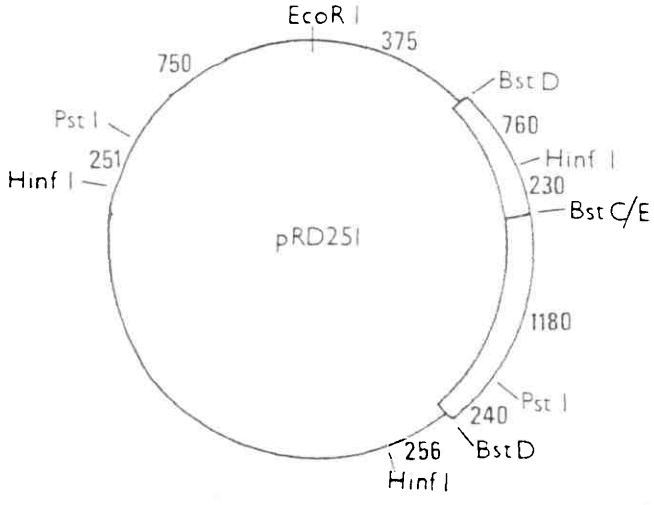
(4) ØX174 / Hae III markers

(5) Pst I — 3400 (doublet), 3300

(6) Bst I / Pst I — 3240, 1180, 1125, 990, 240

(7) Bst I — 4360, 1420, 990

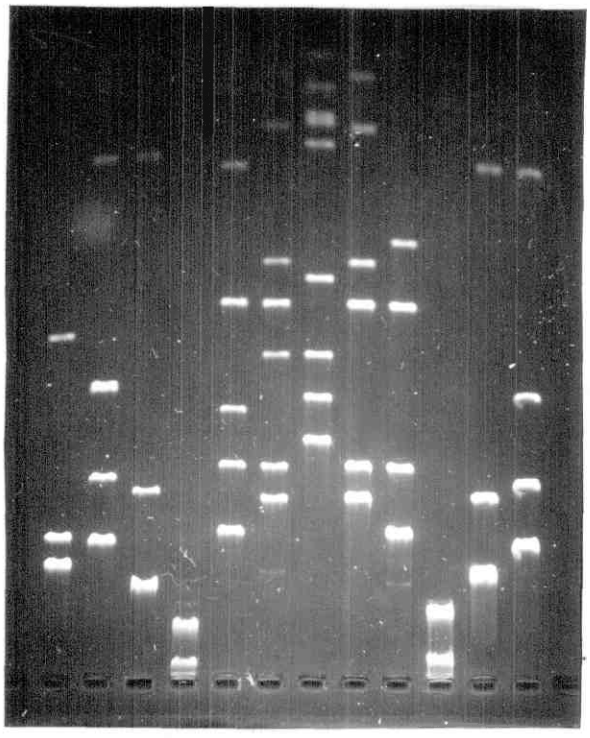
Circular map of pRD251



NB) All fragment sizes are in bp. Only relevant fragments are indicated.

Photograph 3C

Restriction mapping pRD251 and pRD252 with Rsa I and Pvu II



1 → 12

1.5% agarose gel. 100V for 4 hours

(1) Hind III + Pst I
(2) Pvu II + Bst I
(3) Pvu II
(4) Sma I
(5) Rsa I
(6) Rsa I + Bst I

pRD251

(7) ϕX174 + Hae III markers
(8) Rsa I + Bst I
(9) Rsa I
(10) Sma I
(11) Pvu II
(12) Pvu II + Bst I

pRD252

Linear map of the pRD251 insert

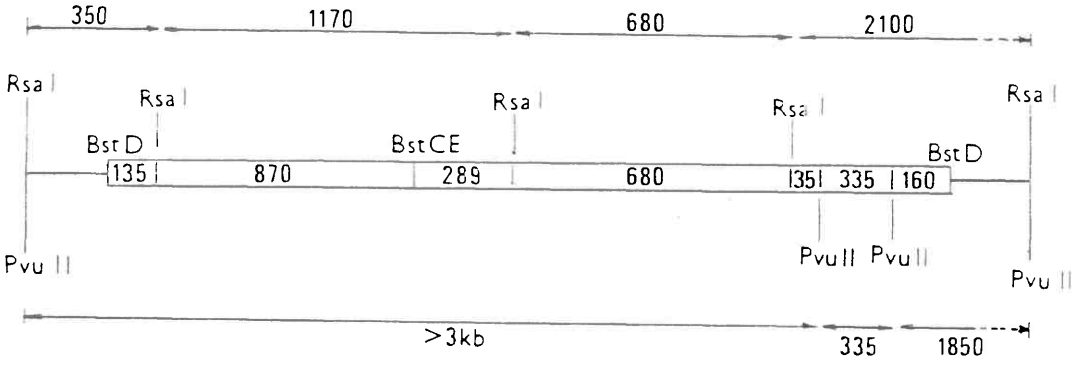
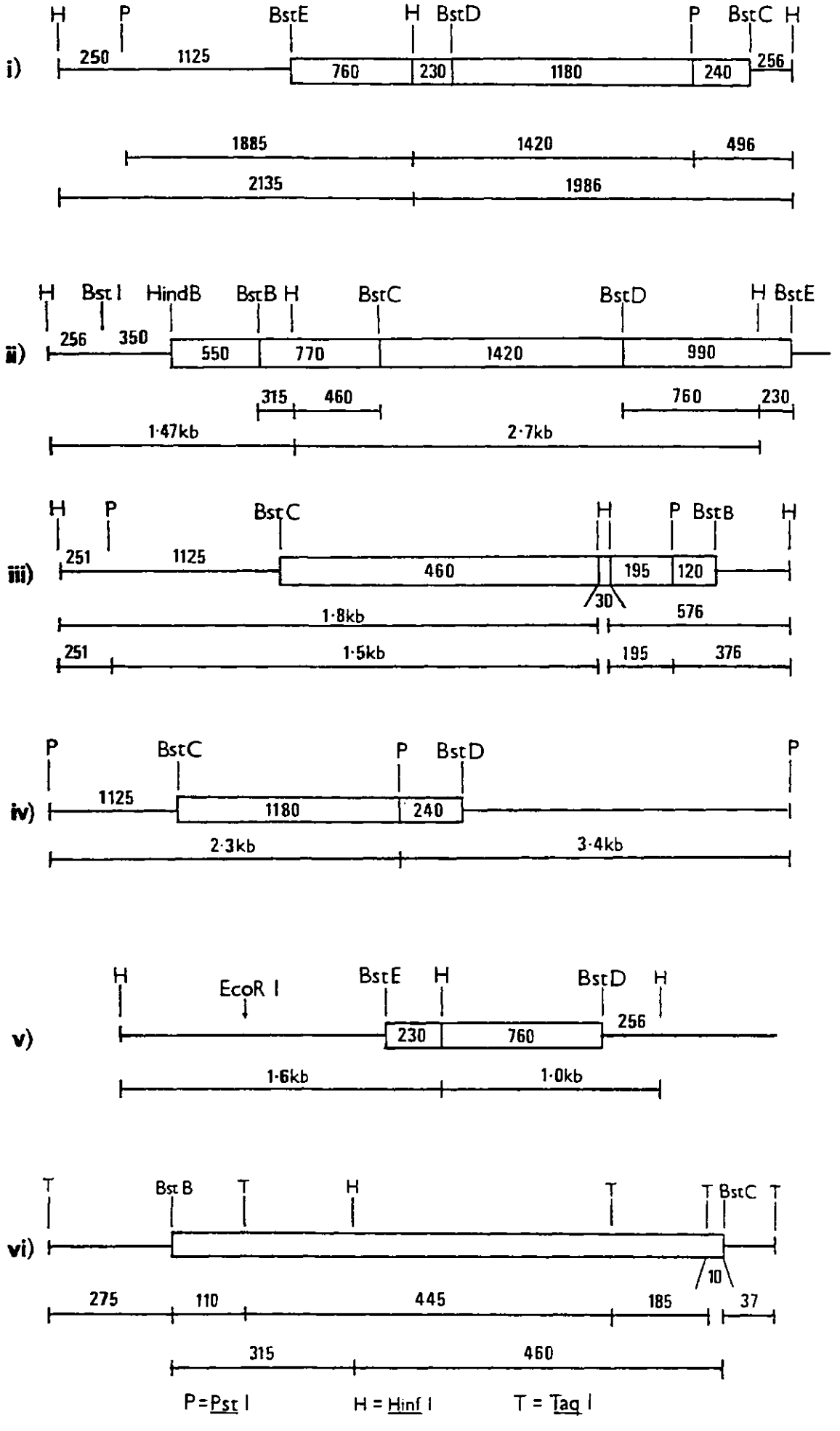
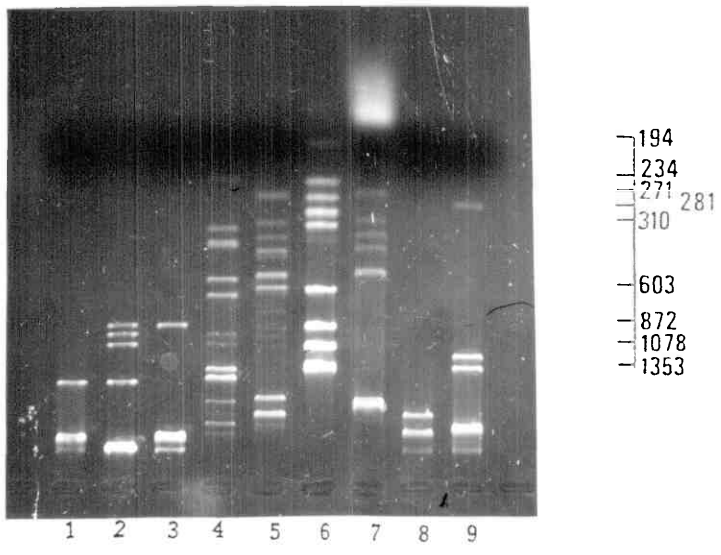


Fig. 3a Partial restriction maps of cloned regions of the dal operon



Photograph 3D

Plasmids pRD251, pRD252 and pRD256 cut with various restriction enzymes



(1) pRD252 + Bst I

(2) pRD351 + Bst I

(3) pRD256 + Bst I

(4) pRD256 + Taq I

(5) pRD256 + Hinf I

(6) ØX174 / Hae III markers

(7) pRD251 + Hinf I

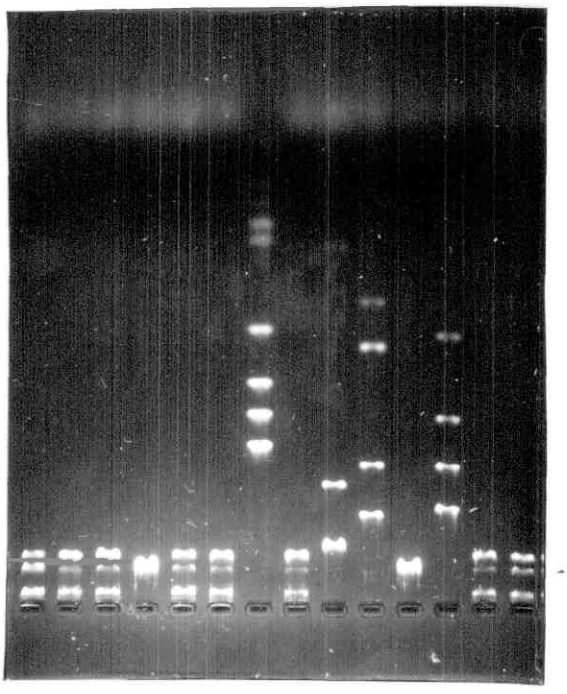
(8) pRD252 + Pst I

(9) pRD252 + Pst I and Bst I

Tracks 1,2 and 3 show that pRD252 and pRD256 contain inserts of 1.42kb and 0.99kb respectively. Lane 5 has the 1.8kb Hinf I fragment which confirms the orientation of BstB/C in pRD256. Lane 7 illustrates the 1.98kb and 2.13kb Hinf I fragments of pRD251 which verify the relative orientations of BstD/E and BstC/D. Lane 8 shows the 2.3kb and 3.4kb Pst I products from pRD252.

Photograph 3E

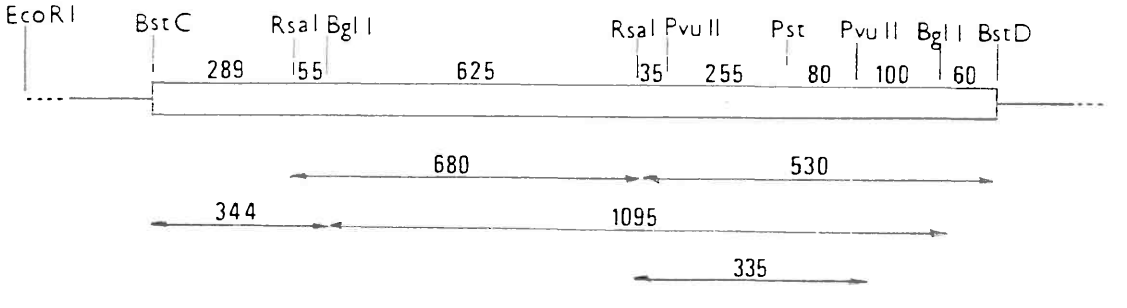
Restriction mapping the BstC/D insert in plasmid pRD252



1.2% agarose gel. 100V for 4 hours

(1) BstE II
(2) EcoR II
(3) Bgl II
(4) Ava I
(5) Bcl I
(6) Xba I
(7) ØX174 / Hae III
(8) Xho I
(9) Pvu II
(10) Rsa I
(11) Cla I
(12) Bgl I
(13) Hpa I
(14) Kpn I

Linear restriction map of the BstC/D fragment



Fragment sizes are in bp

BstB/C has two closely associated Hinf I sites, only 30 bp apart, approximately 315 bp from BstB. The three products of a Bst I/Hinf I double digest are clearly visible in Photo 3F. The 315 bp fragment is cut by Pst I to yield two fragments of 195 bp and 120 bp (Photo 3G). BstB/C has also been mapped for Taq I and contains three sites for this enzyme (Fig. 3a(vi) and Photo 3H).

Plasmid pRD257

The BstD/E 0.99 kb fragment is inserted such that BstE is nearest to the EcoRI site of the vector. Hinf I digests reveal two major brands of 1.6 kb and 1.0 kb (Photo 3A) which can only arise from the structure depicted (Fig. 3a(v)).

The structures of the five plasmids isolated from this Bst I cloning are represented in Fig. 3b.

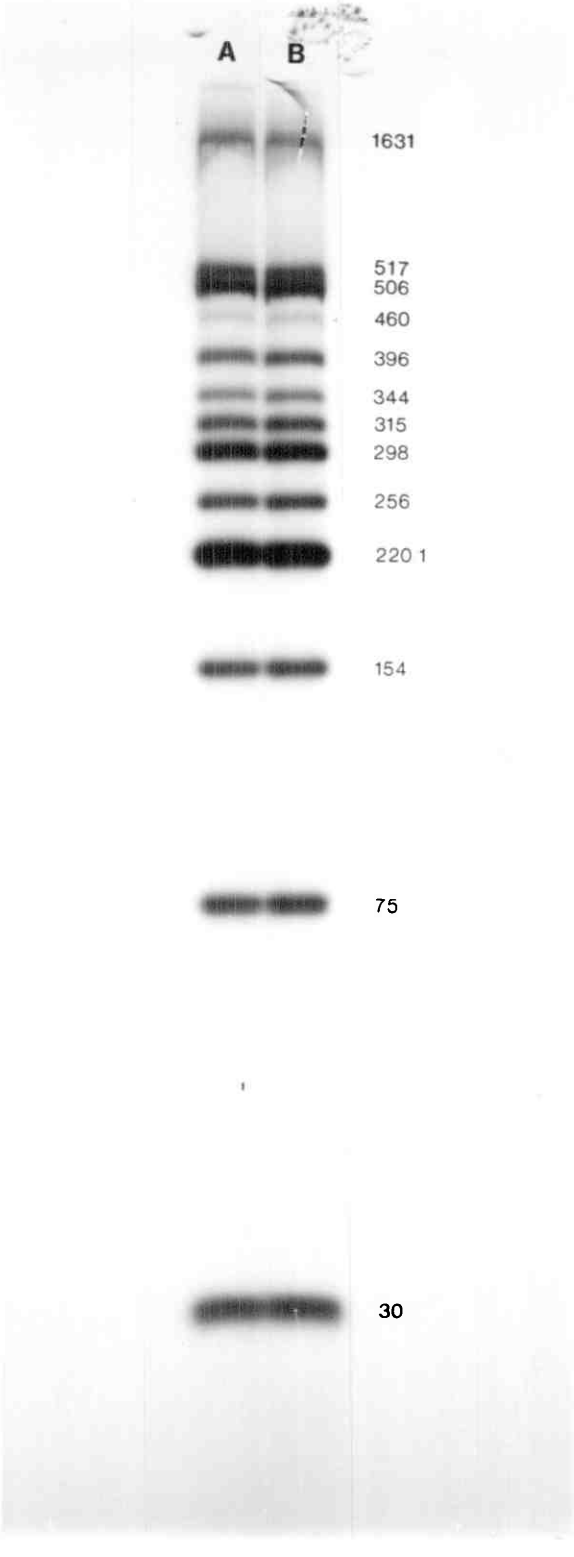
III) DNA Sequence Analysis of the ArDH Structural Gene

At the conception of this project the sequencing of large regions of DNA was just becoming feasible. The original paper by Maxam and Gilbert (1977) pointed the way towards rapid nucleotide sequence determination and it was hoped to utilise these relatively new techniques to answer some of the questions surrounding the evolutionary origins of the two pentitol operons in K.aerogenes. Since the complete amino acid sequence of RDH from the rbt operon was already available (Moore et al, unpublished data), it was decided that the derivation of the ArDH sequence should be the first step in order that a direct comparison might be made between these two potentially related proteins.

Structure of the dal operon DNA

A full genetic and physical analysis of the K.aerogenes DNA carried by λ prbt and λ prbt dal was undertaken by

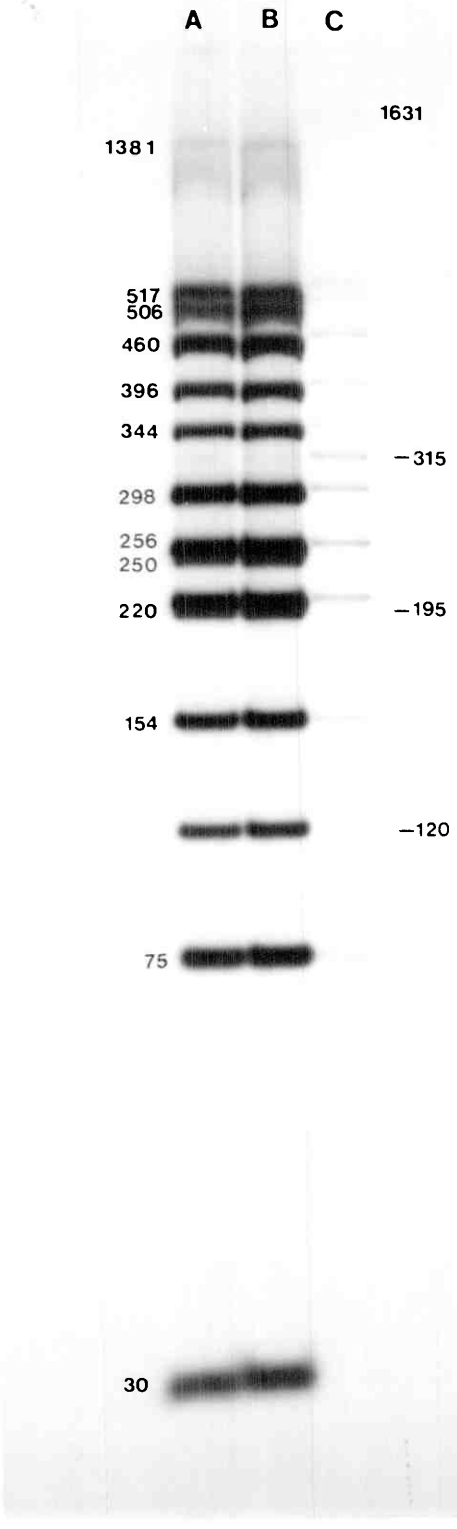
Photograph 3F Bst I + Hinf I digest of pRD256



Autoradiograph of an 8% non-denaturing acrylamide gel. Exposure was for 1 hour at room temperature on Kodak XH-1 film.

Lanes A and B are pRD256 3' labelled at Hinf I sites with α-^{32}P-dATP and Klenow polymerase, then re-cut by Bst I. The 315bp, 460bp and 30bp fragments from BstB/C are visible.

Photograph 3G Bst I / Hinf I / Pst I digests of pRD256



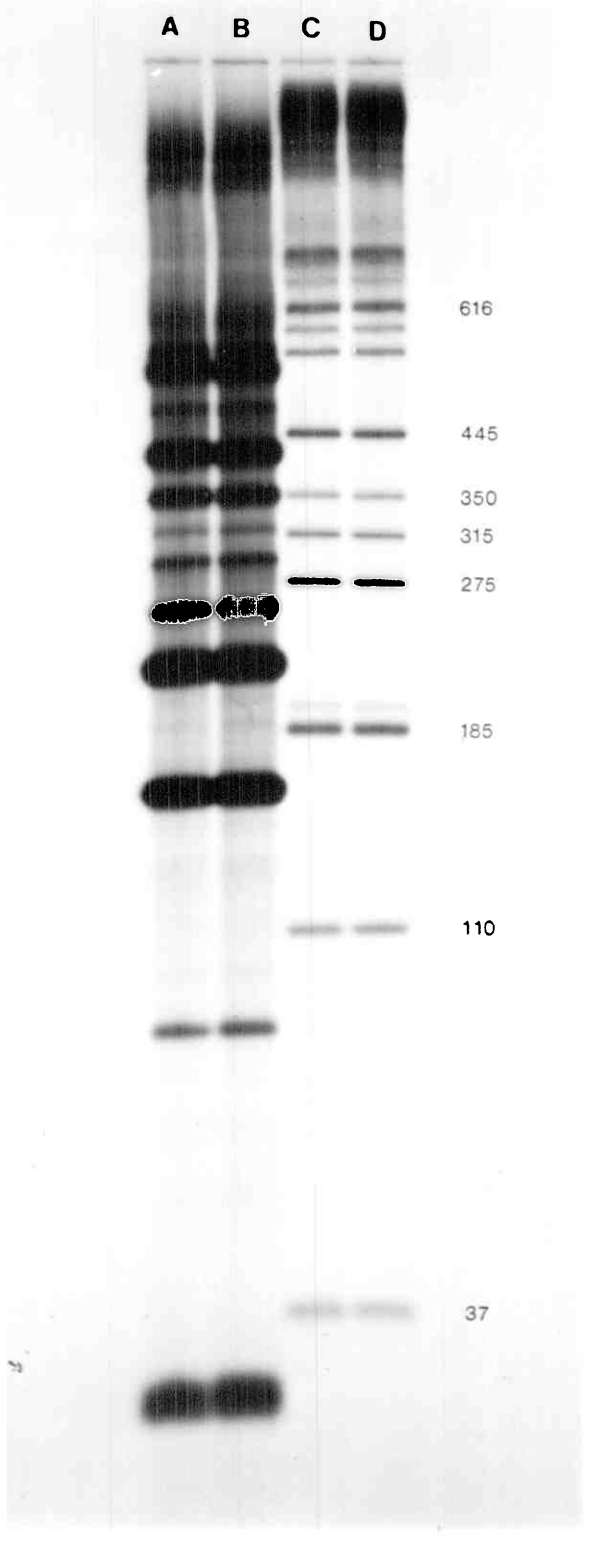
Autoradiograph of an 8% non-denaturing acrylamide gel. The exposure time was 1 hour at room temperature on Kodak XH-1 film.

A + B pRD256 3' labelled at Bst I and Hinf I. Recut by Pst I

C pRD256 3' labelled at Bst I and Hinf I.

The 315bp Bst I/Hinf I fragment is cut to generate two smaller species of 195bp and 120 bp.

Photograph 3H *Hinf* I and *Tac* I digests of pRD256



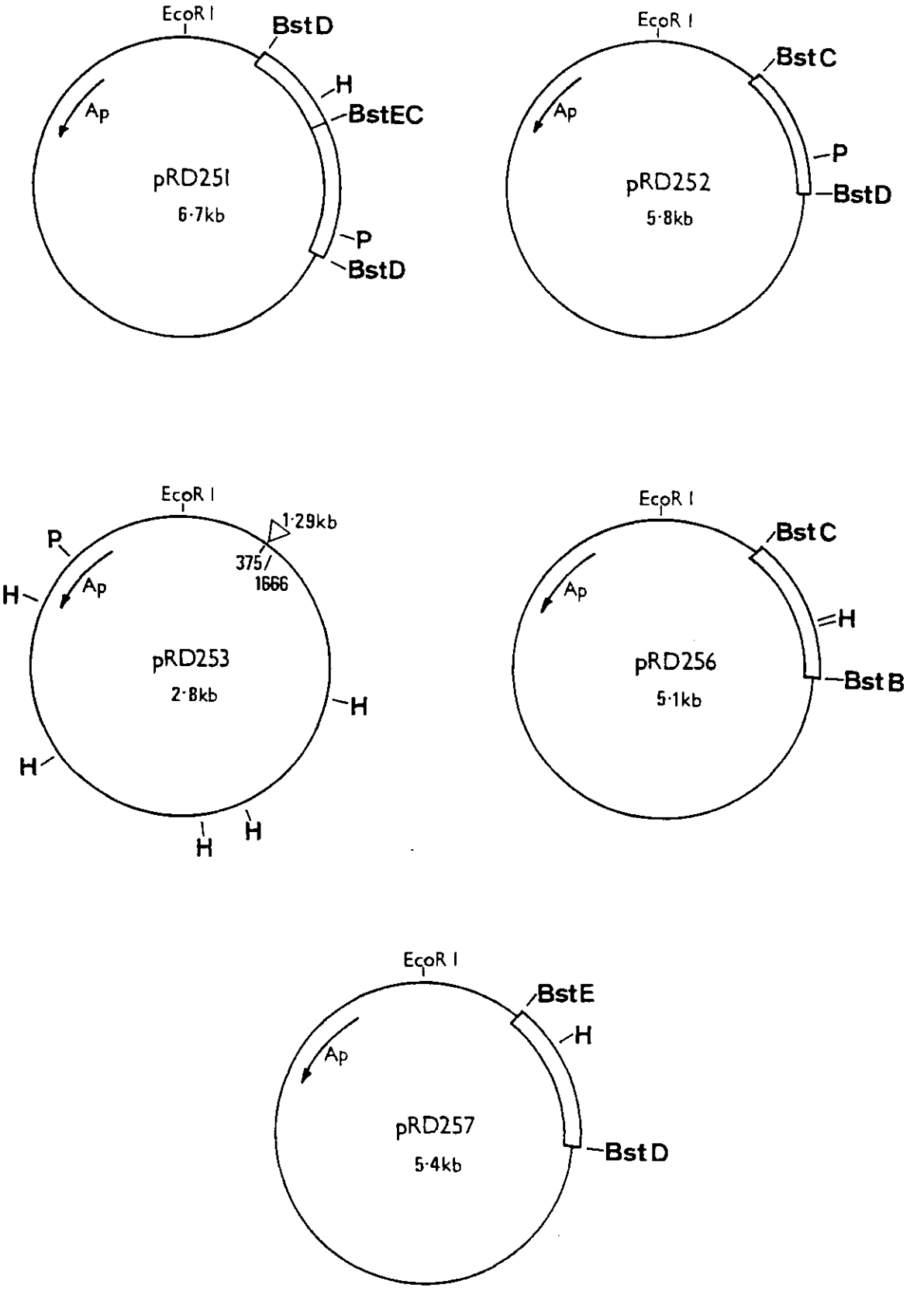
Autoradiograph of an 8% non-denaturing acrylamide gel. Eposure time was 2 hours on Kodak XH-1.

A + B pRD256 3' labelled at *Hinf* I with α-^{32}P-dTTP. Recut by *Bst* I

C + D pRD256 3' labelled at *Tac* I with α-^{32}P-dCTP and recut by *Bst* I, showing the 445,185 and 110bp fragments from *Bst*B/C.

The pBR322 475bp and 141bp *Tac* I fragments are present as a single 616bp species. *Tac* I shows some evidence of partial digestion.

Fig. 3b Circular restriction maps of five pRD plasmids, showing the orientation of insert DNA from λprbtdal.



H = Hinf I P = Pst I

Neuberger (1978) from which he was able to conclude that the former encodes genes for RDH and DRK, and possibly a ribitol permease. Hybridisation of ^{32}P-labelled *dal* and *rbt* mRNAs to separated strands of both phages revealed that transcription of the pentitol operon region is bipolar. Later work (Neuberger and Hartley, 1979) indicated the presence of functional ArDH and DXK genes on λ p *rbt dal* and localised the transcription start points of both operons. Since the approximate sizes of the four known structural proteins are available, it is possible to state tentatively which restriction fragments code for any particular gene. Firm assignments cannot be made in the case of ArDH or DXK as hybridisation experiments quoted above discovered an apparent shortage of coding potential. Strong hybridisation of *dal* mRNA was observed only with the *Bst*C/D fragment which is barely large enough to code for the dehydrogenase alone.

The absence of detectable cross-hybridisation between *rbt*-specific mRNA and DNA from the *dal* operon would seem to indicate a lack of extensive sequence homology within the limits of sensitivity of the technique. In addition, Neuberger and Hartley (1979) failed to observe heteroduplex formation by self-annealed λ p*rbt dal* ssDNA either by electron microscopy or S1 mapping under conditions calculated to allow for a 20% sequence divergence.

Maxam and Gilbert Sequencing Strategy

Initially it was assumed that the transcription start point for the ArDH gene (*dal*D) lay in the vicinity of *Bst*C and that sequencing in both directions from this site would therefore reveal the N-terminal part of the coding sequence. However,

it was found that all reading frames are closed within 185 bp of BstC in the BstC/D fragment. It is unlikely that the region rightwards from BstD has sufficient capacity to code for both ArDH and DXK. The problem was resolved quite easily. Pure ArDH was by this time available (Neuberger et al, 1979) and the sequence of the N-terminus and a number of small elastase-generated peptides (Fig. 3c) were determined (Neuberger, Walker, Dell and Hartley; unpublished data). It was envisaged that such peptide data would facilitate the detection of errors in the nucleotide sequence by giving an indication of the correct reading frame. A search of the translated DNA sequence from BstC towards BstD revealed three elastase peptides (Leu-Phe-Gly-Asp-Leu-Ala, Asp-Phe-Leu and Tyr-Thr-Leu-Ile) followed closely by the ochre stop codon, UAA, representing the C-terminal end of ArDH. It is indeed surprising that no hybridisation of dal mRNA to HindB/BstB or BstB/C was detected earlier since transcription must begin close to HindB. End-labelling at HindB and sequencing towards BstB revealed an amino acid sequence which matched the 20 N-terminal residues of ArDH, with f.Met positioned 122 bp from the Hind III site. The precise region of DNA spanned by dalD was now precisely defined, and the strategy employed to sequence the remainder of the gene is outlined in Fig. 3d.

By labelling at Hind III, Bst I and Pst I restriction sites it was possible to accumulate sequence data quite rapidly. Some fragments generated by these enzymes proved too large to be sequenced entirely since only about 200 bp of sequence could be reliably obtained from each end. It therefore became necessary to locate alternative sites for end-labelling in dalD. Problems were soon encountered as none of the restriction endo-

Fig. 3c Arabitol Dehydrogenase Elastase Peptides

Peptide Pool Number	
A	Thr-Leu-Pro-Tyr-Gln-Tyr Gly-Leu-Gln-Pro-Leu (Glx-Ala-Ala-Met-Gly-Ala)?
B	Leu-Leu-Gly Thr-Leu-Thr-Asp-Val-Leu Gln-Lys-Leu-Leu .....Asn-Pro-Tyr-Leu
C	Gly-Leu-Gln-Pro Asp-Phe-Leu
E	Met-Leu-Ala-Pro Met-Leu-Ala-Ome (Leu-Thr-Thr-Leu)?
G	Asn-Pro-Gln-Thr-Lys-Val-Leu Leu-Phe-Gly-Asp-Leu-Ala Ala-Asp-Leu-Pro-Ala
H	Tyr-Thr-Leu-Leu Met-Leu-Pro-Ala-Ome
L	Tyr-Ala-Leu-Ala-Ome
P	(Lys-Leu-Leu-Glu-Leu-Ala-Ome)?
R	Lys-Leu-Leu-Pro
T	Ala ... Asn-Pro-Tyr-Leu Tyr-Thr-Leu-Leu
W	Ala ... Trp-Tyr-Leu-Ome Ser-Leu-Gln-Lys

(contd)

Fig. 3c ArDH Elastase Peptides

Leu-Leu-Asn-Cys
Leu-Tyr-Gly
(Leu-Ala-Gly-Asn-Gln-Leu-Gly)?
Gln-Lys-Leu-Leu
Thr-Ser-Leu-Gln-Lys-Leu
Gly-Tyr-Tyr
Phe-Met-Glu-Gln
Gly-Lys-Gln

Elastase-digested ArDH was fractionated on a Dowex ion-exchange column. Fractions containing the fewest peptides, as judged by high-voltage paper electrophoresis using a pyridine-acetic acid buffer system (pH 6.5), were acetylated and permethylated for analysis on a Kratos MS 50 mass spectrometer, (Neuberger,Dell,Walker and Hartley, unpublished data).

NB) Leu = Leucine or Isoleucine

nucleases available at this time recognising hexanucleotide sequences had sites within the gene. (Many potentially useful enzymes have since been discovered and are now available commercially). As there are no sites for Eco RI, Sma I, Sst I, Bgl I, Bgl II, Hind III, Hinc II or Kpn I, it was necessary to use enzymes which cut quite frequently to provide the new material needed for sequencing. Hinf I and Taq I were selected because they were known to cleave the BstB/C fragment. Taq I labelling generated little new information since the sites lay within, or close to, regions already analysed, but the data served to confirm or correct the previous results. 3'-labelling of the two adjacent Hinf I sites near BstB provided good sequence information towards BstC but failed to overlap with sequences reading from that end. The small 31 bp Hinf I fragment could not readily be sequenced except by strand separation and this was not attempted. Oddly, the sequence was found to be unreadable from Hinf I towards BstB and at first it was thought that it could be due to incomplete "filling-in" of a Hinf I site whose sequence is 5'-GATTC-3', yielding two labelled species with either one or two d-ATP residues incorporated at the 3' end. However, chasing with "cold" dATP or using α-[^{32}P]-dTTP as the labelled nucleotide failed to improve the sequence ladder which persistently showed bands at every position in all four tracks. The same pattern was seen when labelling the 5' terminus with γ-[^{32}P]-rATP. New preparations of DNA and different batches of restriction enzymes also failed to alleviate the problem. The only satisfactory explanation is perhaps trace contamination of the DNA eluted from the gel by a similarly sized pBR322 fragment - possibly the 298 bp Hinf I fragment. The sequence was finally

completed later by dideoxy sequencing of Taq I clones from BstB/C, see later,(Fig. 3d).

Preliminary experiments utilising Alu I, Hae III or Hpa II as labelling sites were still less productive since two of these produce flush ends which are more difficult to label efficiently. Secondly, the labelled fragments of interest frequently co-migrated during electrophoresis with the numerous species originating from the vector DNA and if a fragment is not well separated from others a high background of spurious bands is often seen on the autoradiograph, making an unambiguous interpretation of the sequence ladder difficult. It was considered that one way to avoid this problem would be to isolate individual restriction fragments from plasmids and to label them separately. The more attractive alternative was to abandon Maxam and Gilbert sequencing in favour of the novel M13 cloning/dideoxy chain-termination methods. Approximately 75% of the ArDH sequence was compiled from Maxam and Gilbert analysis.

M13 Cloning and Dideoxy Sequencing of the dalD gene

The objective here was to clone and sequence pre-determined fragments corresponding to gaps in the ArDH sequence. The approach was to be less random than the original "shotgun" experiments described by Messing et al (1981). As a first step the individual fragments from a Bst I + Hind III digest of pRD351 were separated by agarose gel electrophoresis and purified on DEAE cellulose columns. Taq I and Sau 3A I sub-fragments from HindB/BstB and BstB/C were ligated with Acc I or Bst I digested M13 mp7 RF respectively. A typical 10 µl ligation comprised 1 ng of M13 mp7, a 5-10 fold excess of fragments and 0.5 u of T4 DNA Ligase in 1x C-Buffer, see Materials and Methods. Incubations were carried out overnight at 10°C or for five hours

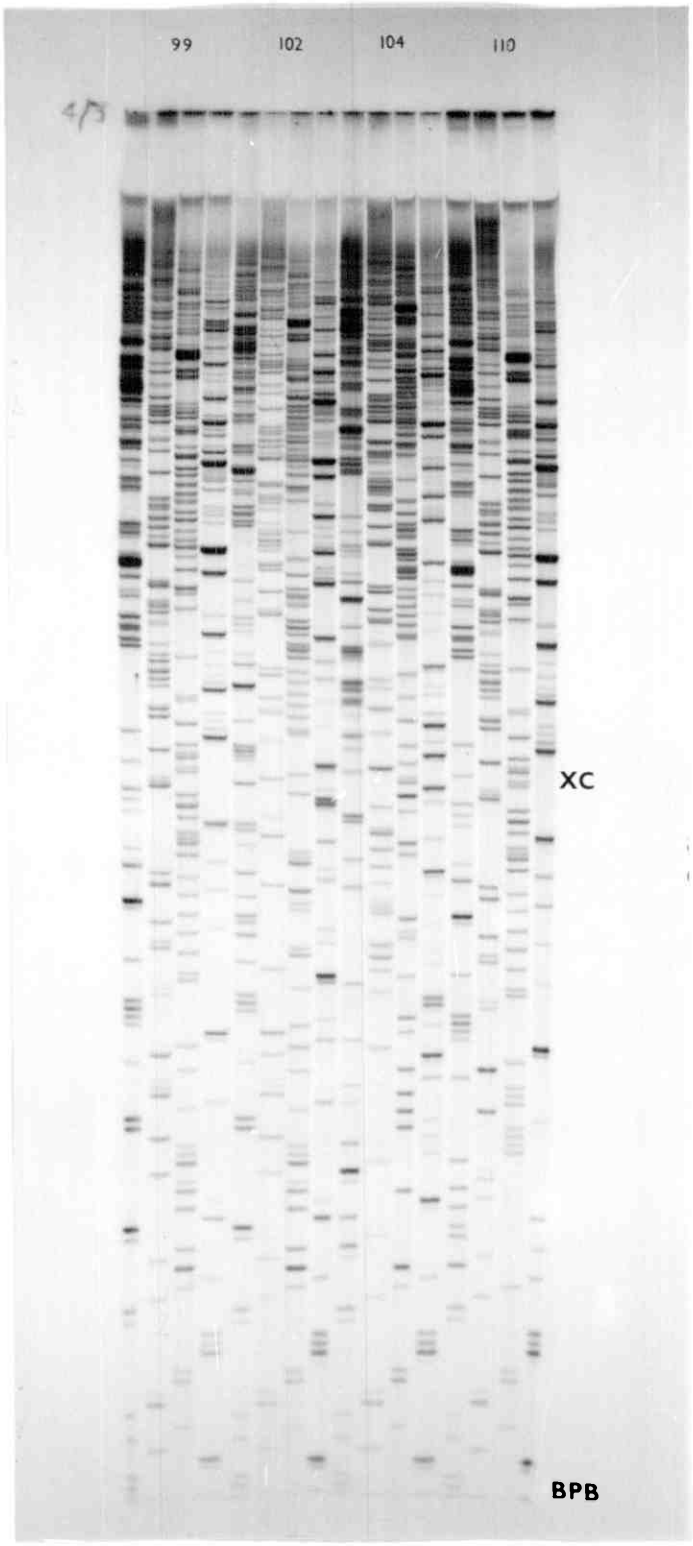
at 15°C. DNA was transformed into competent JM101 cells and recombinant phage identified by plaque morphology on M9 + glucose plates with IPTG and BCIG (Messing et al, 1977). Single stranded phage DNAs were prepared and those containing appropriately sized inserts were selected by electrophoresis on miniature, submerged agarose gels. Templates were then screened by ddT-tracking (Sanger et al, 1980) to eliminate identical clones and the dT ladders were compared to known sequences. Those M13 clones which overlapped regions adjacent to gaps in the Maxam and Gilbert derived data were subjected to full dideoxy sequencing (Photo 3I). Additional clones were used to confirm previous sequences and where information existed for only one strand (Fig. 3d).

The DNA and Amino Acid Sequence of Arabitol Dehydrogenase

A full translation of the ArDH gene is shown in Fig. 3e and a detailed restriction map showing the major fragment sizes has been constructed from the completed nucleotide sequence (Fig. 3f). The gene is 1362 bp long and specifies a protein of 454 amino acids, approximately 49,800 daltons. Earlier estimations of the molecular weight using gel filtration and equilibrium sedimentation set the size at 46,000 daltons or 411 amino acids (Neuberger et al, 1979). The sequence derived here shows that the protein is considerably larger than was originally thought, though the difference probably lies within the limits of experimental error. The amino acid composition (Table 3i) agrees well with the content predicted by amino acid analysis with the exceptions of lysine and alanine for which low and high values respectively are obtained.

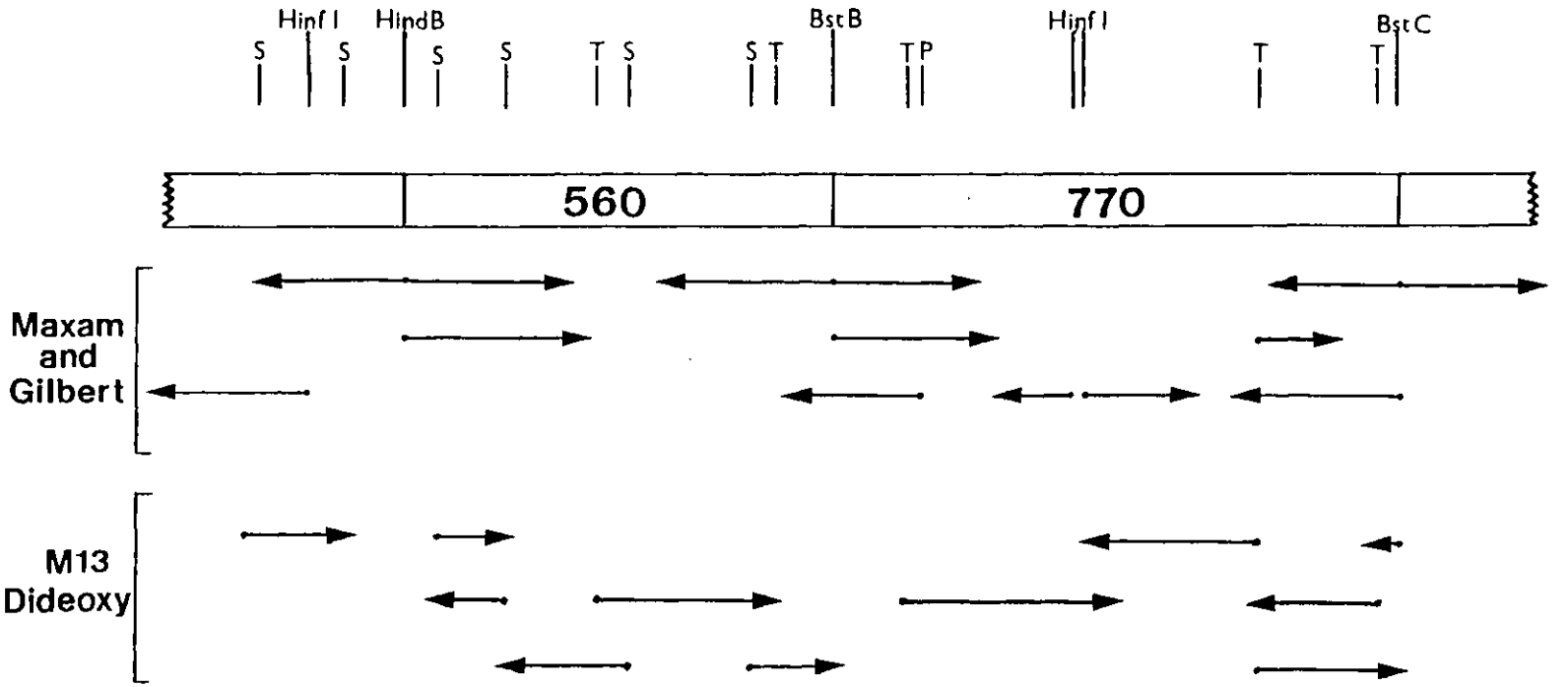
Of the 34 elastase peptides found by mass spectrometry

Photograph 3I



A 6% , 200μm thick acrylamide sequencing gel run for 1 hour at 1900V and regulated at 60°C. Four M13 clones (99,102,104 and 110) containing Sau3A I fragments from the HindB/BstB region of the dalD gene are displayed. The loading order is (left to right) GATC.

Fig.3d The strategy employed to sequence dalD, the Arabitol Dehydrogenase structural gene.



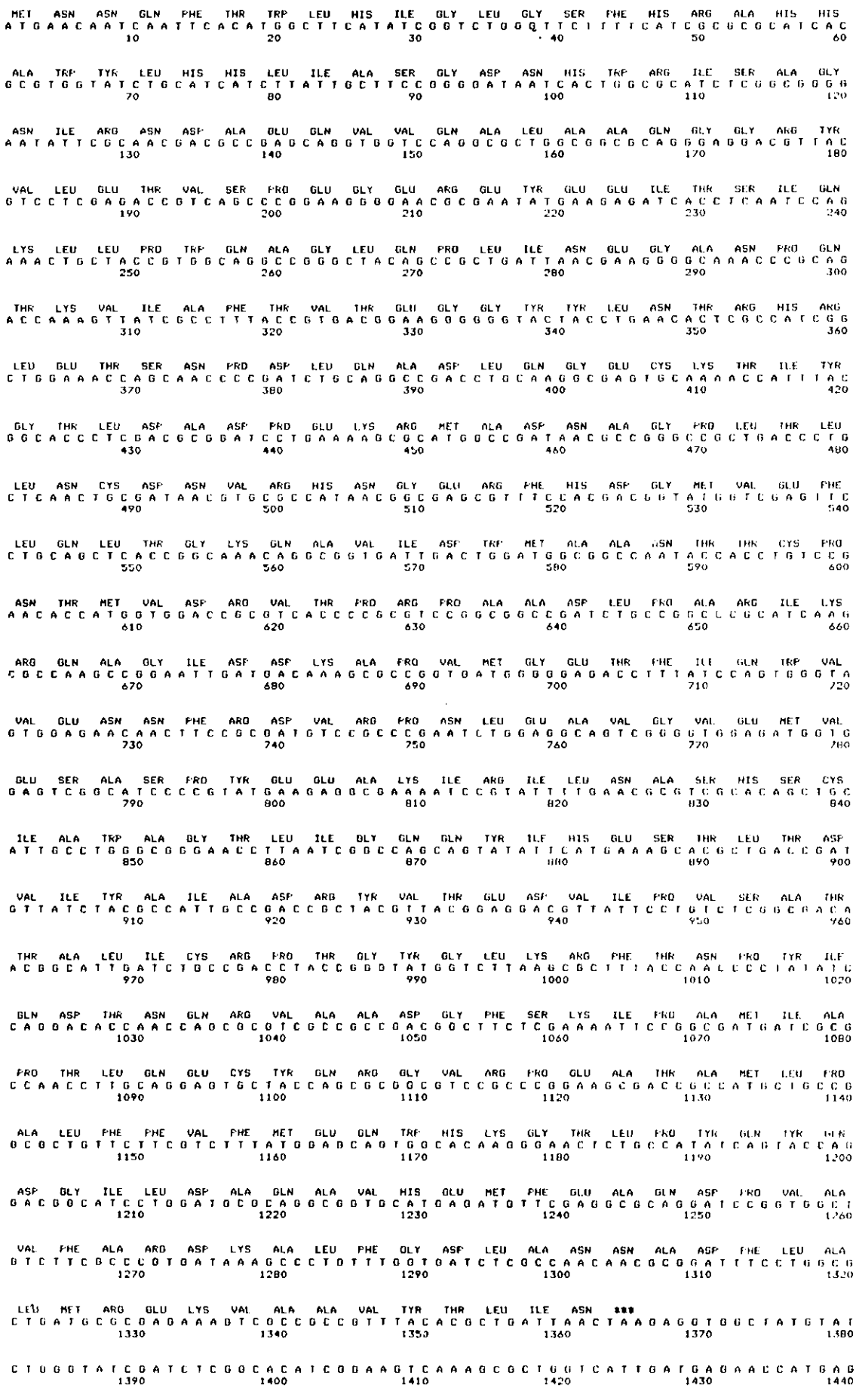


Fig. 3e DNA and amino acid sequence of ArDH

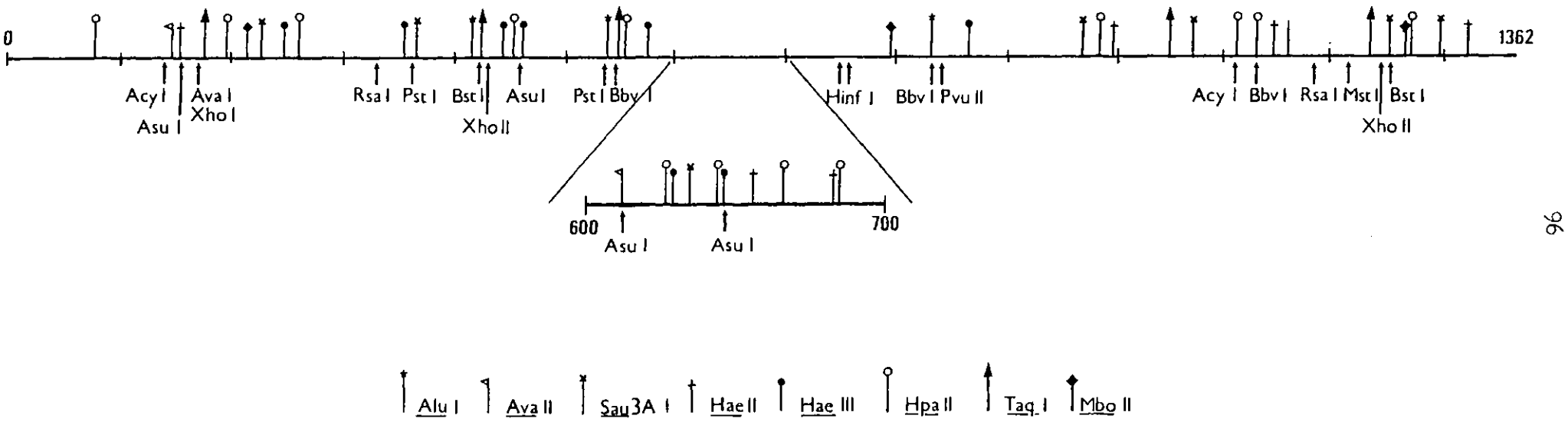


Fig. 3f A restriction enzyme map of the dalD gene compiled from the nucleotide sequence. The major fragment sizes are as follows : HinA/BstB = 557bp, BstB/BstC = 813bp, BstB/PstH = 110bp, BstB/Hinf I = 315bp, HinfI/HinfI = 31bp, HinfI/BstC = 467bp.

Table 3(i) The N-terminal sequence and amino acid composition of Arabitol Dehydrogenase

N-terminal sequence of the purified enzyme:

Met-Asn-Asn-Gln-Phe-Thr-Trp-Leu-His-Ile-Gly-Leu-Gly-(Ser)-Phe-His-Arg-(Ala)- ? -Glx- ? -Tyr-Leu- ....

Amino acid composition:

Amino acid	Residues/molecule (predicted)	Residues/molecule (actual)
Tryptophan	12.2	9
Lysine	21.0	13
Histidine	9.5	14
Arginine	21.0	25
Asp. + Asn.	43.7	52
Threonine	24.6	30
Serine	14.2	13
Glu. + Gln.	48.0	53
Proline	18.9	23
Glycine	27.6	29
Alanine	42.1	52
Valine	28.3	29
Methionine	11.3	12
Isoleucine	25.3	26
Leucine	33.4	36
Tyrosine	12.1	16
Phenylalanine	10.6	16
Cysteine	7.2	6
TOTAL	411	454

Molecular Weight (approx.) = 49,800 daltons

Length of structural gene = 1362bp

only four, about which there was some degree of uncertainty, were not located in the full protein sequence. The regular distribution of these peptides throughout the entire protein (Fig. 3g) lends support to the accuracy of the nucleotide sequence.

Codon Preferences in ArDH

An obvious feature of codon choice in this gene is the overall tendency towards G or C in the third "wobble" position (Table 3ii). This reflects the high G + C content of K.aerogenes chromosomal DNA in accordance with the "genome hypothesis" of Grantham et al (1980). K.aerogenes contains on average 56% G + C compared to 51% for E.coli and 52% for S.typhimurium (Sober, 1970). A higher than average figure of 59% is obtained for dalD (Table 3iii). This manifests itself predominantly in the third codon position where there is a greater degree of freedom and a total of 73% of the codons end in G or C. For the K.aerogenes trpA gene (Nichols et al, 1981a), this figure is still higher at 83%. The bias is most evident in those codon families that are 3 or 4 fold degenerate, particularly the valine, glycine, arginine, leucine and proline groups.

A number of factors serve to distinguish ArDH from very highly expressed genes. For serine, AGC and UGC are the most frequently used codons and AGU is never specified. Bacteria in general prefer UCU or UCC and E.coli messengers in particular select mainly UCU followed, at greatly reduced frequency, by AGC and UCG. The major glycine-tRNA in E.coli K12 is $tRNA_3^{Gly}$ which recognises GGPy : the codons GGU and GGC predominate in highly expressed mRNAs. In contrast, ArDH uses a high proportion of GGG and GGA codons which are recognised by two minor tRNAs present at only 1/6 of the levels of $tRNA_3^{Gly}$.

Fig. 3g Amino Acid Sequence of Klebsiella aerogenes Arabitol Dehydrogenase showing Elastase Peptides

MET ASN ASN GLN PHE THR TRP LEU HIS ILE GLY LEU GLY SER PHE HIS ARG [ALA] HIS HIS [ALA TRP TYR LEU] HIS HIS LEU ILE ALA SER GLY
ASP ASN HIS TRP ARG ILE SER ALA GLY ASN ILE ARG ASN ASP ALA GLU GLN VAL VAL GLN ALA LEU ALA ALA GLN GLY GLY ARG TYR VAL LEU
GLU THR VAL SER PRO GLU GLY GLU ARG GLU TYR GLU GLU ILE [THR SER ILE GLN LYS LEU LEU PRO] TRP GLN ALA [GLY LEU GLN PRO LEU] ILE
ASN GLU GLY ALA [ASN PRO GLN THR LYS VAL ILE] ALA PHE THR VAL THR GLU GLY [GLY TYR TYR] LEU ASN THR ARG HIS ARG LEU GLU THR SER
ASN PRO ASP LEU GLN ALA ASP LEU GLN GLY GLU CYS LYS THR [ILE TYR GLY] THR LEU ASP ALA ASP PRO GLU LYS ARG MET ALA ASP ASN ALA
GLY PRO LEU THR [LEU LEU ASN CYS] ASP ASN VAL ARG HIS ASN GLY GLU ARG PHE HIS ASP GLY MET VAL GLU PHE LEU GLN LEU THR [GLY LYS
GLN] ALA VAL ILE ASP TRP MET ALA ALA ASN THR THR CYS PRO ASN THR MET VAL ASP ARG VAL THR PRO ARG PRO ALA [ALA ASP LEU PRO ALA
ARG ILE LYS ARG GLN ALA GLY ILE ASP ASP LYS ALA PRO VAL MET GLY GLU THR PHE ILE GLN TRP VAL VAL GLU ASN ASN PHE ARG ASP VAL
ARG PRO ASN LEU GLU ALA VAL GLY VAL GLU MET VAL GLU SER ALA SER PRO TYR GLU GLU ALA LYS ILE ARG ILE LEU ASN ALA SER HIS SER
CYS ILE ALA TRP ALA GLY THR [LEU ILE GLY] GLN GLN TYR ILE HIS GLU SER [THR LEU THR ASP VAL ILE] [TYR ALA ILE ALA] ASP ARG TYR VAL
THR GLU ASP VAL ILE PRO VAL SER ALA ALA THR THR ALA TRP CYS ARG PRO THR GLY TYR GLY LEU LYS ARG PHE THR [ASN PRO TYR ILE] GLN
ASP THR ASN GLN ARG VAL ALA ALA ASP GLY PHE SER LYS ILE PRO ALA [MET ILE ALA PRO] THR LEU GLN GLU CYS TYR GLN ARG GLY VAL ARG
PRO GLU ALA THR ALA [MET LEU PRO ALA] LEU PHE PHE VAL [PHE MET GLU GLN] TRP HIS LYS GLY [THR LEU PRO TYR GLN TYR] GLN ASP GLY ILE
LEU ASP ALA GLN ALA VAL HIS GLU MET PHE GLU ALA GLN ASP PRO VAL ALA VAL PHE ALA ARG ASP LYS ALA [LEU PHE GLY ASP LEU ALA] ASN
ASN ALA [ASP PHE LEU] ALA LEU MET ARG GLU LYS VAL ALA ALA VAL [TYR THR LEU ILE] ASN

Elastase peptides are boxed.

MW = 49,800

[6] Amino Acids

Table 3(ii) Codon usage in the dalD gene

1 \ 2	T			C			A			G			3rd base
T	Phe	TTT	6	Ser	TCT	1	Tyr	TAT	7	Cys	TGT	1	T
		TTC	10		TCC	2		TAC	9		TGC	5	C
	Leu	TTA	1		TCA	1	-	TAA	1	-	TGA	0	A
		TTG	3		TCG	5	-	TAG	0	Trp	TGG	8	G
C	Leu	CTT	3	Pro	CCT	3	His	CAT	9	Arg	CGT	5	T
		CTC	5		CCC	2		CAC	5		CGC	18	C
		CTA	2		CCA	2	Gln	CAA	3		CGA	1	A
		CTG	23		CCG	16		CAG	22		CGG	1	G
A	Ile	ATT	13	Thr	ACT	2	Asn	AAT	5	Ser	AGT	0	T
		ATC	14		ACC	21		AAC	19		AGC	4	C
		ATA	0		ACA	2	Lys	AAA	9	Arg	AGA	0	A
	Met	ATG	12		ACG	5		AAG	4		AGG	0	G
G	Val	GTT	5	Ala	GCT	2	Asp	GAT	14	Gly	GGT	5	T
		GTC	13		GCC	21		GAC	12		GGC	8	C
		GTA	1		GCA	4	Glu	GAA	11		GGA	5	A
		GTG	11		GCG	24		GAG	18		GGG	11	G

Codons AGT (Serine) and AGA / AGG (Arginine) are not used. 68% of Leucine residues are specified by CTG and 68% of Prolines by CCG. 22 out of 25 Glutamines (88%) use CAG. 72% of all Arginines are coded by CGC and 79% of Asparagines by AAC. Overall there is a marked preference for codons termunating with G or C.

Table 3(iii) Base composition of the ArDH structural gene

800/1362 bp are G+C 58.7%

562/1362 bp are A+T 41.3%

Base	Occurence	% of total
T	263	19.2
C	400	29.4
G	400	29.4
A	299	22.0
TOTAL	1362bp	-

Whereas many bacteria prefer GUU or GUA, valine in ArDH is coded almost exclusively by GUC or GUG. The latter is recognised by the major $tRNA^{Val}$ but GUC requires two minor species, $tRNA^{Val}_{2A}$ and $tRNA^{Val}_{2B}$. Alanine codons also deviate from the ideal with GCC or GCG being favoured over the more usual bacterial choices GCA and GCU. In most other respects the codon preferences noted for E.coli and other bacteria are adhered to, for example, there is a very pronounced bias towards CUG for leucine, CCG (proline), AAC (asparagine) and ACC (threonine). The general rule that isoaccepting tRNAs interact more strongly with codons of the type A/T, A/T, Py is also supported by the distribution of such codons in ArDH. One exception is $tRNA^{Lys}$, where AAA is preferred to AAG. Note that 88% of glutamine residues are coded by CAG and 72% of arginines by CGC. The isoleucine codon ATA, recognised by a minor tRNA, is never used, and neither are the arginine codons AGA and AGG, whose accepting tRNA is also scarce.

Codon Distribution in Ribitol Dehydrogenase

In order to establish whether the selection of codons in ArDH might be typical of other K.aerogenes genes a survey of the RDH gene was also made (Table 3iv), but this was limited since the complete DNA sequence is still unknown. 169 codons were analysed, representing about 70% of the protein. The same preference for G and C in the third position is observed. Serine codon usage conforms to that generally found in E.coli with UCU and UCC being the most abundant among the UCX family. With respect to glycine, the rarer tRNAs are not specified as often as in ArDH. Otherwise, similar trends exist for the remaining codons. Some of the small differences might reflect the history of selection for fast growth on xylitol and

Table 3(iv) Codon usage in the rbtD gene

1 \ 2	T			C			A			G			3rd
T	Phe	TTT	1	Ser	TCT	2	Tyr	TAT	2	Cys	TGT		T
		TTC	3		TCC	4		TAC			TGC	2	C
	Leu	TTA			TCA		-	TAA		-	TGA	1	A
		TTG	1		TCG		-	TAG		Trp	TGG	2	G
C	Leu	CTT	4	Pro	CCT	1	His	CAT	1	Arg	CGT	1	T
		CTC	5		CCC			CAC	3		CGC	4	C
		CTA	1		CCA		Gln	CAA	1		CGA		A
		CTG	10		CCG	4		CAG	5		CGG		G
A	Ile	ATT	3	Thr	ACT	1	Asn	AAT	2	Ser	AGT		T
		ATC	9		ACC	3		AAC	6		AGC	3	C
		ATA			ACA	1	Lys	AAA	4	Arg	AGA		A
	Met	ATG	3		ACG	1		AAG	2		AGG	1	G
G	Val	GTT	2	Ala	GCT	2	Asp	GAT	5	Gly	GGT	4	T
		GTC	8		GCC	10		GAC	6		GGC	10	C
		GTA	1		GCA	1	Glu	GAA	4		GGA		A
		GTG	5		GCG	7		GAG	3		GGG	4	G

The complete nucleotide sequence of rbtD is not yet available. This table is based upon the 170 codons (including TGA) known to date.

improved RDH activities in this strain of K.aerogenes.

When compared with other K.aerogenes genes such as trpA, it emerges that the greater use of GUG/GUC for valine, GGG/GGA for glycine and UCG or AGC for serine is characteristic of this bacterium. Indeed, the codon usage of Klebsiella more closely resembles that of Salmonella species than that of E.coli.

The Intercistronic Region of the dal Operon

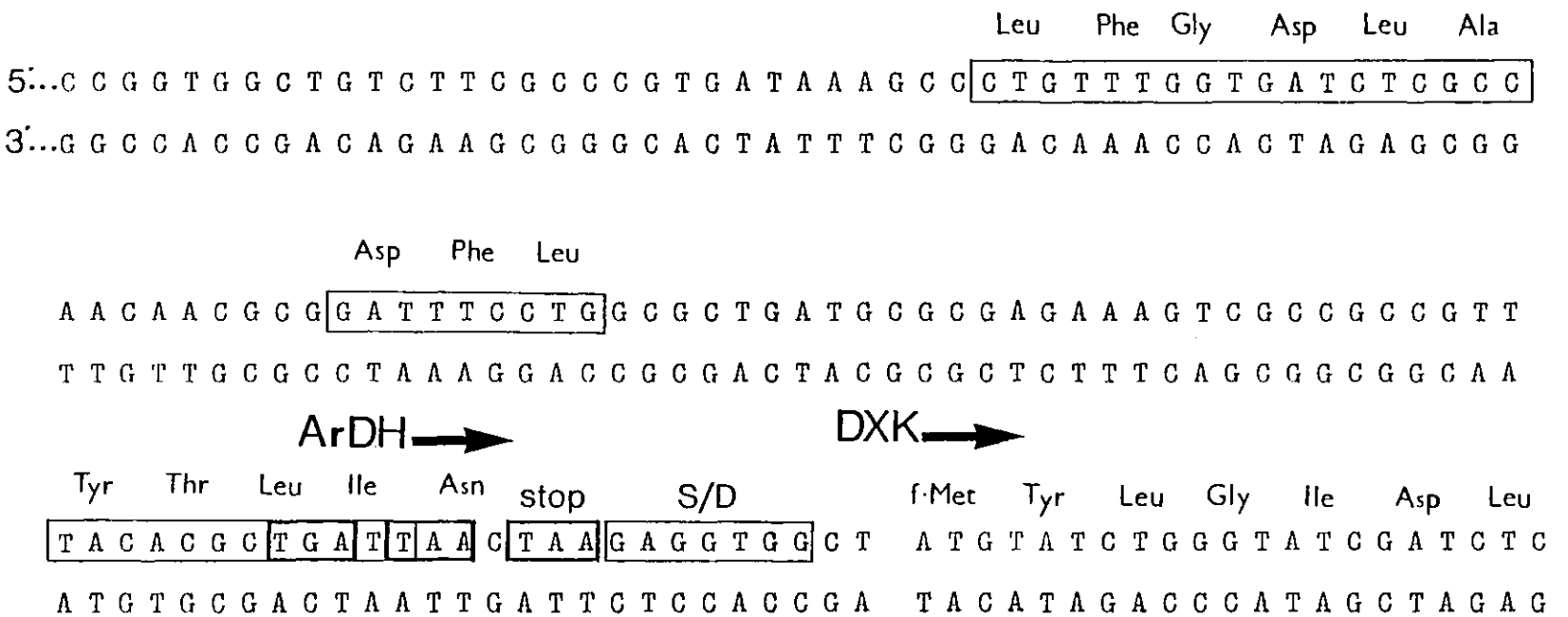
The nucleotide sequence spanning the C-terminus of ArDH and the N-terminus of DXK displays a number of interesting structural features. Firstly, translational termination signals occur in consecutive reading frames near the end of the dehydrogenase coding sequence, the final UAA codon marking the end of this gene (Fig. 3h). Such strong stop signals are common amongst prokaryotes. Secondly, only 9 nucleotides separate the ArDH terminator from the DXK initator AUG, and these bear substantial homology to a sequence at the 3' end of 16S rRNA. This region may serve as an attachment site for ribosomes translating the kinase gene - a possibility that will be discussed in more detail later, see Chapter 7.

The N-terminal amino acid sequence of DXK (Neuberger et al, 1979) correlates precisely with the DNA sequence in this region and allows us to pinpoint accurately the start of the gene.

DNA Sequencing of the d-Xylulokinase Gene

d-Xylulokinase is a protein of some 55,000 daltons, requiring approximately 1500 bp of DNA. The beginning of the gene is positioned 125 bp from BstC, which means that it must span the remainder of this fragment and terminate just distal to BstD. Some Maxam and Gilbert sequencing has already been carried out in this region and work involving the sub-cloning

Fig. 3h The dalD / dalK intercistronic region



The DNA sequence spanning the C-terminus of ArDH and the N-terminus of DXK is shown here. Sequences coding for elastase peptides are marked and their translation appears above each. Three nonsense codons at the end of dalD are indicated, together with a possible ribosome attachment site (S/D). The DXK N-terminal amino acid sequence agrees with the available protein sequence data.

of kinase coding sequences into the M13 vectors mp8 and mp9 is well advanced (Fig. 3i). The C-terminus has been located 57 bp beyond BstD and is consistent with a protein of 53,000 daltons. One major stumbling block in sequencing DXK has been an inability to clone the BstC/Pst i fragment into mp8, although the complementary strand is clonable at high frequency in mp9. This phenomenon has also been encountered by others (P. David, personal communication), and even clone turnaround in mp7 failed to generate clones with this 1.18 kb fragment in the reverse orientation. It may be possible to clone this region as smaller pieces by using Rsa I and Pvu II (Photo 3E) and by using the restriction minus host JM103 to pick up DNAs containing unmodified EcoK sites.

A 7 bp sequence between Pst i and BstD is repeated in the d-Ribulokinase gene of the rbt operon and is responsible for the formation of the gene duplication in K.aerogenes strain A3 (P. David, unpublished data; Neuberger and Hartley, 1981).

Accuracy of the Nucelotide Sequences

The usefulness of any sequence must depend ultimately upon its accuracy. Under circumstances where the protein sequence is unknown it is often necessary to sequence both DNA strands to minimise errors arising from strand-specific artefacts and to repeat each sequence run several times. Even though considerable peptide data was at hand, this procedure was nevertheless applied. Wherever possible, fragments were labelled and sequenced from their 3' and 5' ends and, in addition, sequences were obtained across all restriction sites used for labelling with the exception of BstC, where the sequence right up to the labelled nucleotide was determined for both 3' and 5' labelled fragments. The likelihood of two adjacent Bst I

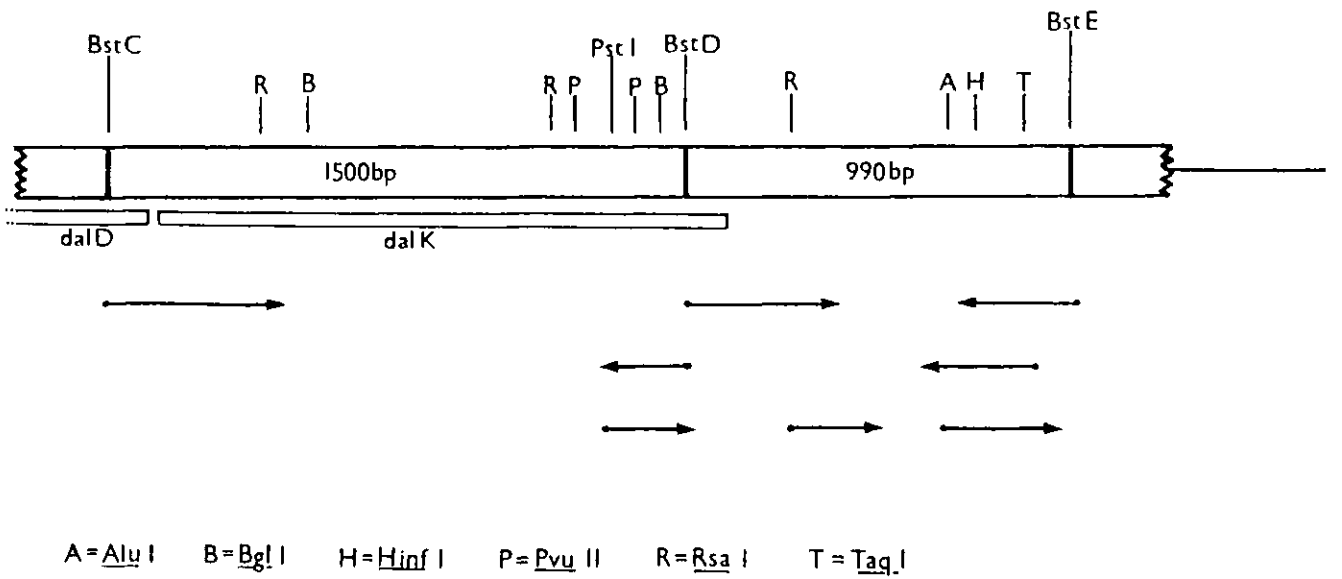


Fig. 31 Strategy for sequencing the dalK gene and the dal operon repeat sequence. Arrows indicate the sequence data accumulated from M13 clones of both regions. It is hoped to use the Rsa I and Bgl II sites in BstC/D to clone the remainder of the dalK gene. BstD/E carries an extensive stretch of sequence that is repeated on the complementary strand in the SalB/C fragment (see Chapter 7).

sites is remote, but not impossible. The reading frame is open on both sides of BstC and codes for known elastase peptides, showing that the phase is correct. Furthermore, no small ^{32}P-labelled species can be seen on 12% acrylamide gels.

When using the chain termination method it was usually possible to identify oppositely oriented clones by gel hybridisation (Hermann et al, 1980) and thus sequence both strands for any given region. Occasionally, the reverse clone could not be isolated. This is not uncommon : M13 inexplicably displays strong preferences for particular fragments and often one orientation is favoured exclusively (Gardner et al, 1981). In these instances the clone-turnaround technique (Winter, 1980) was applied to M13 RF isolated from 1 ml cultures of JM101 in order to provide the complementary sequence.

Sequence ladders were generally very clean, although resolution on Maxam and Gilbert gels was occasionally reduced by contaminating salt or low molecular weight acrylamide in the DNA. Polyethylene glycol carried over from M13 phage precipitations caused high backgrounds and artefact bands in a few dideoxy sequence gels. Where these problems occurred new template DNA was made and the experiments repeated.

The presence of 5-Methyl Cytosine, which reacts only slowly with hydrazine, was confirmed by the appearance of G on the complementary strand when using the Maxam and Gilbert technique. All EcoR II sites were found to be methylated as shown :

$$
\begin{array}{ccccccc}
5' & C & \overset{*}{C} & N & G & G & 3' \\
 & G & G & N & \underset{*}{C} & C &
\end{array}
$$

This sequence occurs six times in dalD. Only one other position was found to be modified, namely :

5' C C G C* C 3' Position 1221 in dalD

G G C G G

An artefact that cannot be circumvented directly is caused by intrastrand pairing during electrophoresis. The resulting "compression" often leads to superimposed bands and may render that area of the sequence ladder unreadable. Compressions arise when a sequence and its complement appear on the same strand, allowing hairpin loops to develop. This is generally only a problem where runs of four or more G•C pairs can form. Weak compressions were normally relieved by simply running gels at higher voltages (i.e. hotter) to melt out secondary structures, or by sequencing the opposite strand, where the compression is often displaced slightly. Five very bad compressions were encountered and, in general, tended to pose more of a problem during dideoxy sequencing, owing to the sensitivity of DNA Polymerase I to variations in the template sequence (Sanger and Coulson, 1975; Sanger et al, 1977). Compressions in dalD around positions 360 and 790 were overcome by running gels as hot as possible (up to 45W) without cracking the plates. Artefacts stemming from a dyad symmetry centred around position 1211 bp and a G + C rich string at 250/280 bp were resolved on ultra-thin gels regulated at 65°C. The most stubborn compression was very extensive and situated around position 70 bp and across the region 90/120 bp. Electrophoresis on standard sequence gels or 200µm gels at 65°C did not significantly improve matters. Only by employing a 200 µm gel containing 8M urea and 25% w/v formadide thermostatted at 70°C could the sequence be read clearly.

Two instances of sequence heterogeneity were discovered, at positions 235 bp and 938 bp. In both cases the nucleotide

was identified as A by the chemical method, and G by dideoxy sequencing. Neither alteration is neutral. At 235 bp, A→G changes Ile to Val, and at 938 bp Asp becomes Gly. Earlier work used DNA isolated from strains selected for growth on d-arabitol. It is likely that small fragments of DNA cloned in non-essential regions of M13 will accumulate mutations more rapidly as no selective pressure exists. It was therefore decided to maintain A at both positions. Winter and Fields (1980) detected nucleotide changes at the low level of 1 in 3700 while sequencing Influenza virus cloned in M13. This figure included population variations and copy errors. The *in vitro* error rate of DNA Polymerase I is estimated to be 10^{-4} (Weymouth and Loeb, 1978). Some heterogeneity is normally observed in phage populations (Domingo *et al*, 1978) In all, a great deal of care has been taken to ensure that the sequences presented here are free of errors, but small changes may well have occurred since the *dal* operon was originally cloned from *K.aerogenes* into *E.coli*.

CHAPTER 4

ARABITOL DEHYDROGENASE : ANALYSIS OF THE PROTEIN AND NUCLEIC ACID SEQUENCES. COMPARISONS WITH RDH AND OTHER PROTEINS

One of the original objectives of this project was to cast light on the evolutionary origins of the K.aerogenes pentitol operons. It was thought that a comparison of the protein and nucleotide sequences might provide conclusive evidence of common ancestry. From physical data alone, the pentitol dehydrogenases appear quite distinct : RDH is a tetramer of subunit MW 27,000 (Neuberger et al, 1979). Both proteins are also dissimilar with respect to their reaction kinetics. Close structural similarities are also ruled out by the lack of common antigenic determinants, but the absence of any detectable level of immunological cross-reaction does not totally exclude the possibility of amino acid sequence homologies. In view of these differences, one would expect to find few extensive similarities between ArDH and RDH, despite some tempting gene duplication theories put forward to account for the compact structural arrangement of the pentitol operons.

The Primary Structures of ArDH and RDH

Both proteins display fairly typical amino acid compositions (Table 4i), particularly ArDH, whose make-up matches almost exactly with that of the "average" protein (Dayhoff et al, 1978). RDH tends towards a higher than normal content of hydrophobic residues (Table 4ii), amounting to 53% of the total. The distribution of chemically similar amino acids throughout the polypeptide chain is not noticeably non-random, except that hydrophobic groups in ArDH are more numerous in the C-terminal region (Fig. 4a). RDH shows no obvious trends aside from a slightly higher proportion of polar and charged residues among the

Table 4(i) The amino acid compositions of Arabitol Dehydrogenase and Ribitol Dehydrogenase

Amino acid	Average occurence (%)	RDH	ArDH
Met	1.7	6 (2.4)	12 (2.6)
Val	6.6	31 (12.6)	30 (6.6)
Leu	7.4	28 (11.3)	37 (8.1)
Ala	8.6	30 (12.1)	51 (11.2)
Ile	4.5	14 (5.7)	27 (4.8)
Trp	1.3	3 (1.2)	8 (1.7)
Phe	3.6	7 (2.8)	16 (3.5)
Pro	5.2	9 (3.6)	23 (5.1)
Tyr	3.4	3 (1.2)	16 (3.5)
Ser	7.0	13 (5.3)	13 (2.8)
Asn	4.3	9 (3.6)	24 (5.3)
Thr	6.1	11 (4.5)	30 (6.6)
Cys	2.9	2 (0.8)	6 (1.3)
Gln	3.9	11 (4.5)	25 (5.5)
Gly	8.4	20 (8.1)	29 (6.4)
Lys	6.6	10 (4.0)	13 (2.9)
His	2.0	5 (2.0)	14 (3.1)
Arg	4.9	11 (4.5)	25 (5.5)
Asp	5.5	15 (6.1)	26 (5.7)
Glu	6.0	9 (3.6)	29 (6.4)
TOTAL		247	454

Values for the average occurence of each amino acid are taken from Dayhoff et al,(1978) .The bracketted figures represent the % composition of that particular residue.

Table 4(ii) The abundance of chemically similar amino acid groups in ArDH and RDH

Type	Amino acid	Average (%)	RDH	ArDH
Small aliphatic	A+G	16.9	20.2	17.8
Hydroxyl	S+T	13.1	9.7	9.5
Acidic	D+E	11.6	9.7	11.9
Acidic and Acid amide	DBNEZQ	19.8	17.8	23.1
Basic	K+R+H	13.5	10.5	11.5
Hydrophobic	L+V+I+M	20.2	32.0	22.7
Aromatic	F+Y+W	8.3	5.3	9.0

Arabitol Dehydrogenase

A.A. Group	No. Residues	% of total
Hydrophobic	222	49.0
Polar	127	27.9
Basic	50	11.0
Acidic	55	12.1

Ribitol Dehydrogenase

A.A. Group	No. Residues	% of total
Hydrophobic	131	53.0
Polar	66	26.7
Basic	26	10.5
Acidic	24	9.8

Fig. 4a Amino acid sequence of Arabitol Dehydrogenase showing the distribution of (a) charged and (b) hydrophobic residues

(a)

```
        10         20         30         40         50         60
HNNQFTWL*I GLGSF**A** AWYL**LIAS G*N*W*ISAG NI*N*A*QVV QALAAQGG*Y

        70         80         90        100        110        120
VL*TVSF*G* **Y**I*SIQ *LLPWQAGLQ PLIN*GANPQ T*VIAFTVT* GGYYLNT***

       130        140        150        160        170        180
L*TSNF*LQA *LQG*C*TIY GTL*A*F*** MA*NAGPLTL LNC*NV**NG **F**GMV*F

       190        200        210        220        230        240
LQLTG*QAVI *WMAANTTCF NTHV**VTP* PAA*LPA*I* *QAGI***AP VMG*TFIQWV

       250        260        270        280        290        300
V*NNF**V*P NL*AVGV*MV *SASPY**A* I*ILNAS*SC IAWAGTLIGQ QYI**STLT*

       310        320        330        340        350        360
VIYAIA**YV T**VIPVSAT TALIC*PTGY GL**FTNPYI Q*TNQ*VAA* GFS*IPAMIA

       370        380        390        400        410        420
PTLQ*CYQAG VAP*ATAMLF ALFFVFM*QW **GTLPYQYQ *GIL*AQAV* *MF*AQ*PVA

       430        440        450
VFA***ALFG *LANNA*FLA LM***VAAVY TLIN
```

(b)

```
        10         20         30         40         50         60
*NNQ*T**H* G*GS*HR*HH ****HH***S GDNH*R*S*G N*RNI*EQ** Q****QGGR*

        70         80         90        100        110        120
**ET*S*EGE RE*EE*TS*Q N****Q*G*Q ***NEG*N*Q TN****T*TE GG***NTRHR

       130        140        150        160        170        180
*ETSN*I*Q* L*QGECN**** GT*D*I*EKR **IN*G**T* *NCIN*RHNG ER*HIG**F*

       190        200        210        220        230        240
*G*TGKQ*** D*****VTTC* N**IR*T*R ***I***R*L RQ*G*DIK** **GET**Q**

       250        260        270        280        290        300
*ENN*RD*R* N*E**G*E** ES*S**EE*N *R**N*SHSC ****GT**GQ Q**HEST*TI

       310        320        330        340        350        360
******IR** TEI*****S*T T***CR*TG* G*KR*TN*** QDTNQR***I G*SK******

       370        380        390        400        410        420
*T*QEC*Q*S ***E*T*** *******EQ* HKGT***Q*Q DG**I*Q**H E**E*QD***

       430        440        450
***RDK***G I**NN*I*** **REK***** T**N
```

N-terminal 40 amino acids (Fig. 4b).

The Secondary Structure of Dehydrogenases

In all known NAD^+-dependent enzymes, those parts of the polypeptide chain involved in coenzyme-binding show a remarkable similarity in their secondary structural organization, although the remainder of the protein may be quite variable. Regions outside the nucleotide-binding site are important for substrate-binding, catalysis, specificity and subunit interactions. The major structural elements of the coenzyme domain consist of an evolutionarily conserved/converged pattern of parallel, pleated β-sheet flanked by α-helices (Fig. 4c). At the amino acid level this takes the form of an alternating sequence of helix and sheet units (Rossman et al, 1975), and such arrangements may be diagnostic of a nucleotide-binding fold comprised of repetitive $\beta\alpha\beta$ units. Brändén et al (1973) have shown that the coenzyme-binding region of Liver Alcohol Dehydrogenase (LADH) has a main chain conformation similar to the corresponding region of Lactate Dehydrogenase (LDH) and Malate Dehydrogenase (MDH), (Rossman et al, 1971; Hill et al, 1972). Jörnvall (1973) showed that a significant degree of homology exists between residues 30-85 of both LADH and Yeast Alcohol Dehydrogenase (YADH). The conformation of the coenzyme-binding site of these four proteins is conserved and the consensus domain of six parallel β-sheets joined by helices or loops emerges as a general structure for the binding of nucleotides. In LADH, the amount of secondary structure in this domain is considerable. Helix structures account for 44% of the total, β-sheet 32% and reverse turns 13%, leaving only 10% of residues in irregular forms of which no string is greater than four amino acids long. The sequences of ArDH and RDH have been analysed in the hope of

(a)

```
        10         20         30         40         50         60
*KHS*SS*NT S*SGK****T G**SG*G*EC *RT**G*G*K ****DREGEK *NK***E*GQ

        70         80         90        100        110        120
N****Q*D** Q*DQ*DN**Q G**Q*TGR*D **H*N*G*** GG***EGD*D **DR**H*N*

       130        140        150        160        170        180
N***RC*RS* **H***QKSG D***T****G ****E***T* SK***Q***H TTRRQ**Q*G

       190        200        210        220        230        240
*R*G****G* **T***DD** K*K*DE***D GS**Q**E** ES*****TRS KN*T*RD***

**NS*D*
```

(b)

```
        10         20         30         40         50         60
M**SVSSMNT SLSG*VAAIT GAASGIGL*C A*TLLGAGA* VVLI***G** LN*LVA*LGQ

        70         80         90        100        110        120
NAFALQV*LM QA*QV*NLLQ GILQLTG*L* IF*ANAGAYI GGFVA*G*F* VW**VL*LNI

       130        140        150        160        170        180
NAAF*CV*SV LF*LLAQ*SG *IIFTAVIAG VVIW*FVYTA S*FAVQAFV* TT**QVAQYG

       190        200        210        220        230        240
V*VGAVLFGF VVTALL**WF *A*M**ALA* GSLMQFI*VA *SVLFMVT*S *NVTV**IVI

LFNSV*L
```

Fig. 4b Amino acid sequence of RDH showing the distribution of (a) hydrophobic and (b) charged residues.

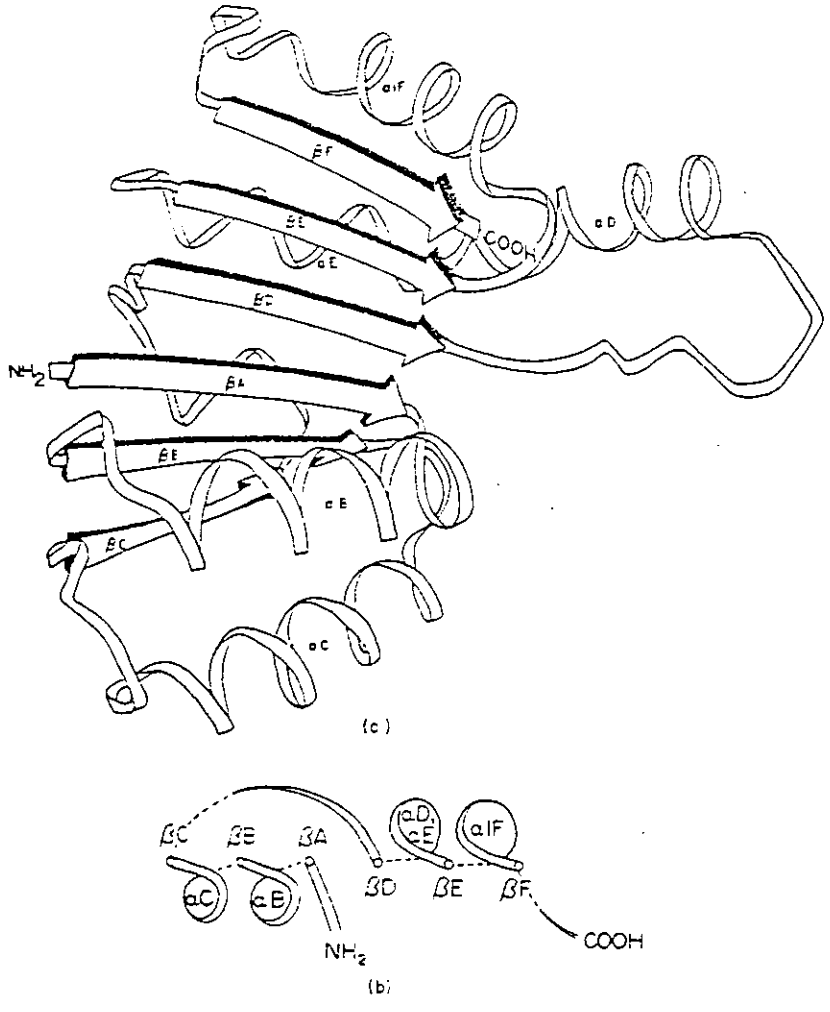


Fig. 4c Schematic drawings of the NAD^+ binding domain of dehydrogenases (a) viewed from above and (b) seen end-on, looking along the sheet elements with their amino terminal ends closest to the observer.

(Diagram reproduced from Rossman et al, 1975)

identifying regions which may be involved in coenzyme-binding. X-ray crystallographic data are at present unavailable.

Secondary Structure Analyses of ArDH and RDH

A possible secondary structure for ArDH was determined by the methods of Chou and Fasman (1974), and by using the PREDICT computer program of Dr. A. McLachlan. Both gave quite similar results and suggest a high degree of helix and sheet structure with very little of the polypeptide chain remaining in an irregular form (Fig. 4d, Table 4iii). For comparison, a computer prediction for RDH was also made (Fig. 4e). In earlier work Dothie (1974) attempted to relate the occurrence of single amino acid changes in mutant RDHs to a disruption of secondary structure elements predicted by the Chou and Fasman algorithm. The computer-generated structures differ slightly from his forecasts, but not in the vicinity of the mutations. Overall, Dothie's analysis underestimates regions of β-sheet and overestimates α-helices.

The Secondary Structure of ArDH

Computer analysis predicts a prominent stretch of alternating β-sheet and α-helix units in the N-terminal part of the sequence. The first 155 residues constitute a region composed of a tandem array of five $\beta\alpha$ structures. As mentioned above, a feature of this type exists near the N-terminus of many NAD^+-linked enzymes, although the precise location is variable. In order to verify this region as the cofactor-binding site, sequences from proteins of known tertiary structure were compared to the ArDH sequence. In addition, Jörnvall (1981) presented evidence which assigns residues 14-50 of RDH a role in NAD^+-binding, and so this region and the flanking sequences were also aligned against ArDH. Residues Lys·15 - Leu·34 of

Table 4(iii) Predicted secondary structure content of Arabitol Dehydrogenase and Ribitol Dehydrogenase

Ribitol Dehydrogenase (residues 1 - 247)

Structure	No. Residues	Percentage
β-sheet	97	39.0
β-turns	17	6.8
α-helix	130	52.6
Random	3	0.01

Arabitol Dehydrogenase (residues 1 - 156)

Structure	No. Residues	Percentage
β-sheet	60	38.4
β-turns	20	12.8
α-helix	74	47.4
Random	2	0.01

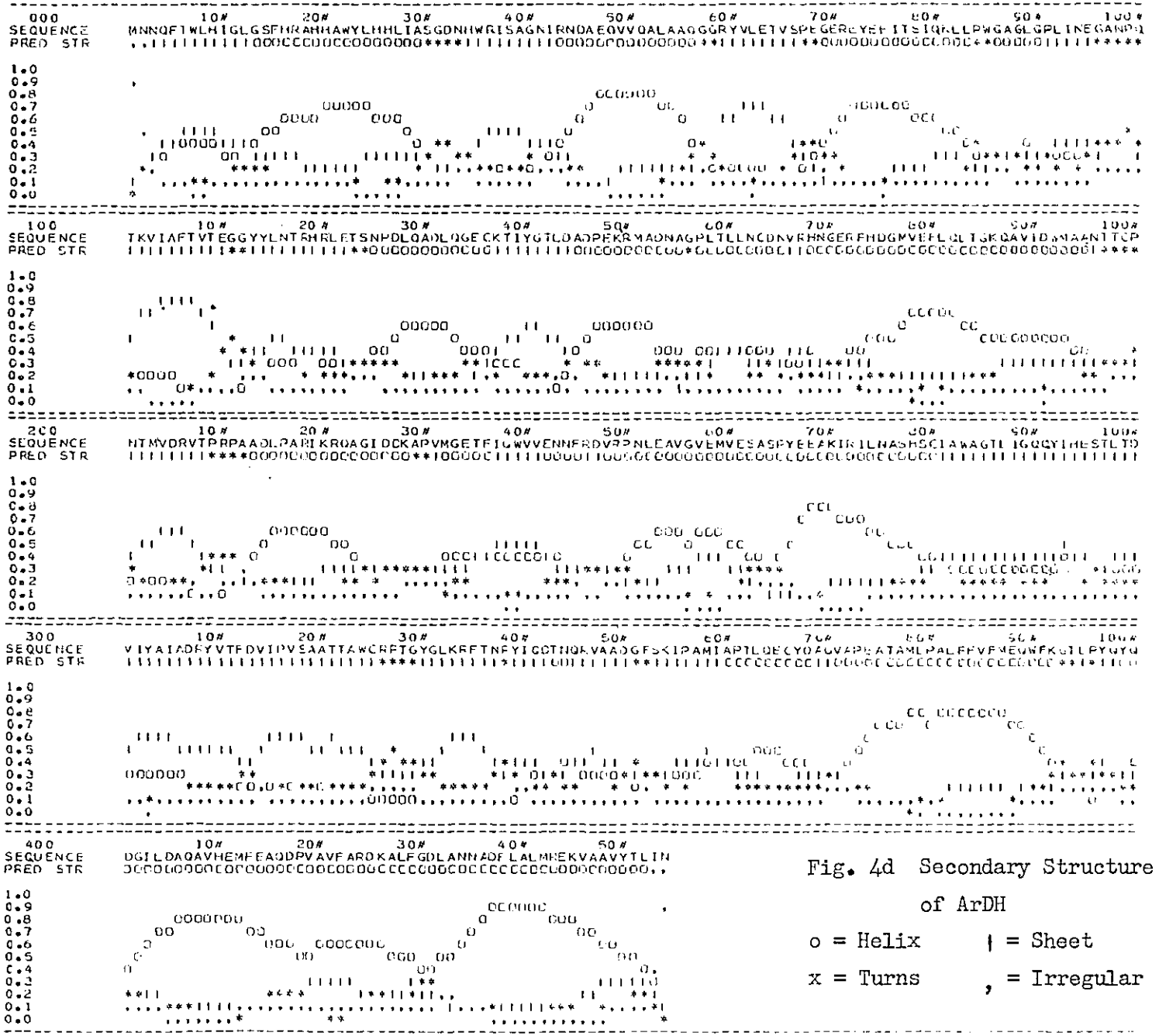


Fig. 4d Secondary Structure of ArDH

o = Helix | = Sheet

x = Turns , = Irregular

Fig. 4e Secondary Structure of RDH

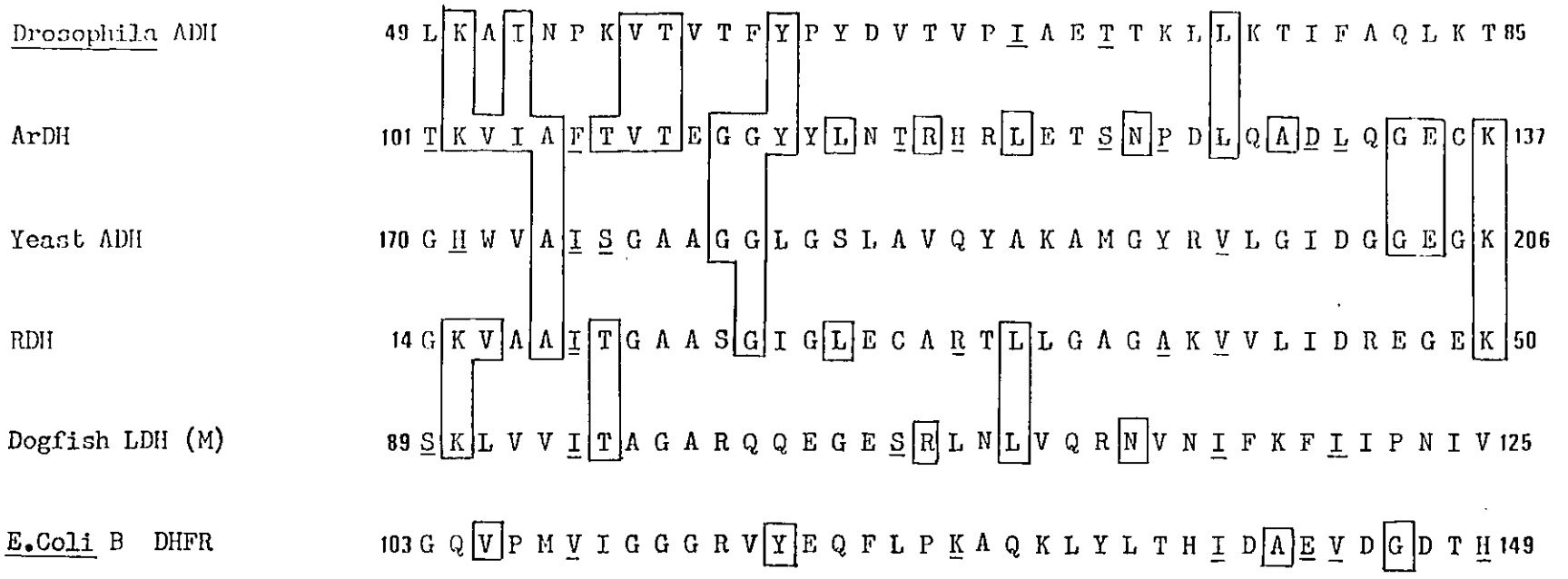
o = Helix | = Sheet x = Turns , = Irregular

RDH can be matched with ArDH Lys·102 - Leu·121 to give a total of seven identities. Over the greater distance screened by Jörnvall (RDH Lys·15 - Lys·50), only one extra match is obtained with ArDH (Lys·102 - Lys·137), but these additional 16 residues show homology to YADH His·171 - Lys·206 (Fig. 4f). The probability of seven identities in twenty residues occurring by chance is 3.1×10^{-5} (see Engel, 1973), but this calculation does not take into account the different frequencies at which each amino acid occurs in a polypeptide. When compared to the NAD^+-binding regions of Dihydrofolate Reductase (DHFR), RDH, LDH, YADH and Drosophila Alcohol Dehydrogenase (DADH) collectively, the 36 amino acid string form Lys·102 - Lys·137 of ArDH shows 16 identities and 3 conservative changes. Hence, there is strong evidence that this region represents part of the domain for NAD^+ binding.

In Bovine Glutamate Dehydrogenase (GDH) and Dogfish LDH the residues involved in direct interactions with the cofactor are Lys·126 and Arg·101 - Arg·109 respectively. Bennett (1974), in comparing sequences bordering these essential amino acids with other putative NAD^+ domains, observed two features which were considered to be important. Firstly, the active site lysine or arginine is preceded by three consecutive Gly or Ala residues. Secondly, a further charged amino acid positioned seven or eight places distal. If the coenzyme-binding domain of ArDH has been correctly identified, then Lys·120 and Asp·127 are the only possible candidates for this role, but neither ArDH, RDH nor YADH shows any evidence of a Gly/Ala tripeptide followed <u>immediately</u> by lysine or arginine.

As there is some amino acid homology between the NAD^+ domains of ArDH and RDH, this might also be manifested in a similar distribution of helix and sheet structures. The

Fig. 41 Homology of ArDH with the NAD^+ binding regions of some other dehydrogenases



The boxes indicate conserved amino acids. Chemically similar residues or conservative replacements found at the same position in ArDH and any of the other five sequences are underlined. These comparisons are strictly linear. No attempt has been made here to optimise matching by inserting spaces.

relevant regions of both enzymes were compared and found to share some common features. The NAD^+ region of each is preceded by a short stretch of β-sheet (amino acids 1-10 of RDH and 91-95 of ArDH) and then a sequence with strong reverse turn potentials. There next follows, in both instances, a 10 residue β-sheet terminated again by a reverse turn. ArDH has an additional ßαß sub-structure at this point (RDH lacks the first sheet element), but the final β-sheet in both cases (AA 40-45 RDH, AA 137-144 ArDH) is followed by an extraordinarily long segment of α-helix. In ArDH this helix is 50 residues long and contains a single reverse turn, but in RDH is predicted as an unbroken helix of 44 residues with no internal prolines and most of the amino acids having high helix-stabilising potentials. This is quite exceptional since such a helix must represent a length approximating to the total molecular diameter ($\approx$5 nm) of RDH.

The remainder of the ArDH polypeptide chain from around position 200 to the C-terminus appears to be structured quite differently from the N-terminal part. Residues 212-280 are predominantly helix-forming and are followed by 80 residues of β-sheet units broken only by reverse turns. Ala·360 through to the C-terminus adopts a conformation that is almost entirely helical. The region Tyr·399 - Leu·452 is predicted to be an α-helix of 53 residues, but position 430/431 has quite a high ß-turn potential and the helix-breaking amino acid proline occurs at 418. This part of the protein may, therefore, exist as distorted helix.

In contrast, RDH exhibits a very pronounced ßαß structure throughout its entire length. Each helix or sheet is, for the most part, greater than 15 residues long, except when punctuated by ß-turns. The C-terminal end of the protein contains another

long helix of some 30 amino acids. There is no evidence of similarity in the secondary structures of ArDH and RDH outside their N-terminal regions which probably serve the common function of cofactor-binding. Differences in the amino acid sequences of the proposed NAD^+ sites can be viewed in terms of base changes at the DNA level (Fig. 4g). It is apparent that many of the differences could have arisen out of a single point mutation but the time since divergence must be very long indeed.

Amino Acid Sequence Homology Between ArDH, RDH and Other Dehydrogenases

If gene duplications were at some stage involved in the evolution of the pentitol dehydrogenases, then it might still be possible to detect faint homologies. The evidence accumulated to date argues strongly against a recent duplication event. The entire sequences of both proteins were compared by computer. A span length of 29 residues and a significance level of 1% were the specified parameters. Numerous weak homologies were detected, the best of which are illustrated in Fig. 4h. The results are not striking, and it appears ArDH and RDH are no more similar than one might reasonably expect of any two dehydrogenases. The N-terminal 115 residues of RDH display many faint homologies to the first 200 of ArDH, but in the C-terminal half of ArDH there are very few sequences having even weak homology with any RDH sequence. This supports the idea that the two share a common, conserved, functional region near their N-termini, but differ significantly in the remaining (catalytic) regions. Jörnvall (1981) finds some similarity between sequences in the catalytic region of DADH and RDH residues Ile·143-Glu·206. One of the better matches

Fig. 4g Homologies at the DNA and amino acid level between the putative coenzyme binding domains of Arabitol Dehydrogenase and Ribitol Dehydrogenase

(a)

```
RDH   GGTAAAGTCGCCGCGATCACCGGCGCGGCGTCCGGTATCGGCCTCGAGTGCGCGAGGACCCTGCTTGGGGCTGGCGCAAAAGTGGTACTGATCGACCGCGAAGGCGAAAAG
         |||||   |||  | ||||   |||     ||    |  |||  |              |||           | | |  ||        |   |  |  |    |||
ArDH  ACCAAAGTTATCGCCTTTACCGTGACGGAAGGGGGTACTACCTGAACACTCGCCATCGGCTGGAAACCAGCAACCCCGATCTGCAGGCCGACCTGCAAGGCGAGTGCAAA
```

(b)

Ribitol Dehydrogenase	14	G	K	V	A	A	I	T	G	A	A	S	G	I	G	L	E	C	A	R	T	L	L	G	A	G	A	K	V	V	L	I	D	R	E	G	E	K	50
							·		·	·	·	·					·			·	·				·		·	·	·					·	·	·			
Arabitol Dehydrogenase	101	T	K	V	I	A	F	T	V	T	E	G	G	Y	Y	L	N	T	R	H	R	L	E	T	S	N	P	D	L	Q	A	D	L	Q	G	E	C	K	137

(a) DNA homology between regions 40-150bp of RDH and 301-411bp of ArDH. There are a total of 37 matches (33%).

(b) Amino acid homology between residues 14-50 of RDH and 101-137 of ArDH. Identical amino acids are boxed.

Differences which could arise from single base changes in either gene are marked (·).

Fig. 4h Amino acid homologies between ArDH and RDH , detectable at a significance level of 1%

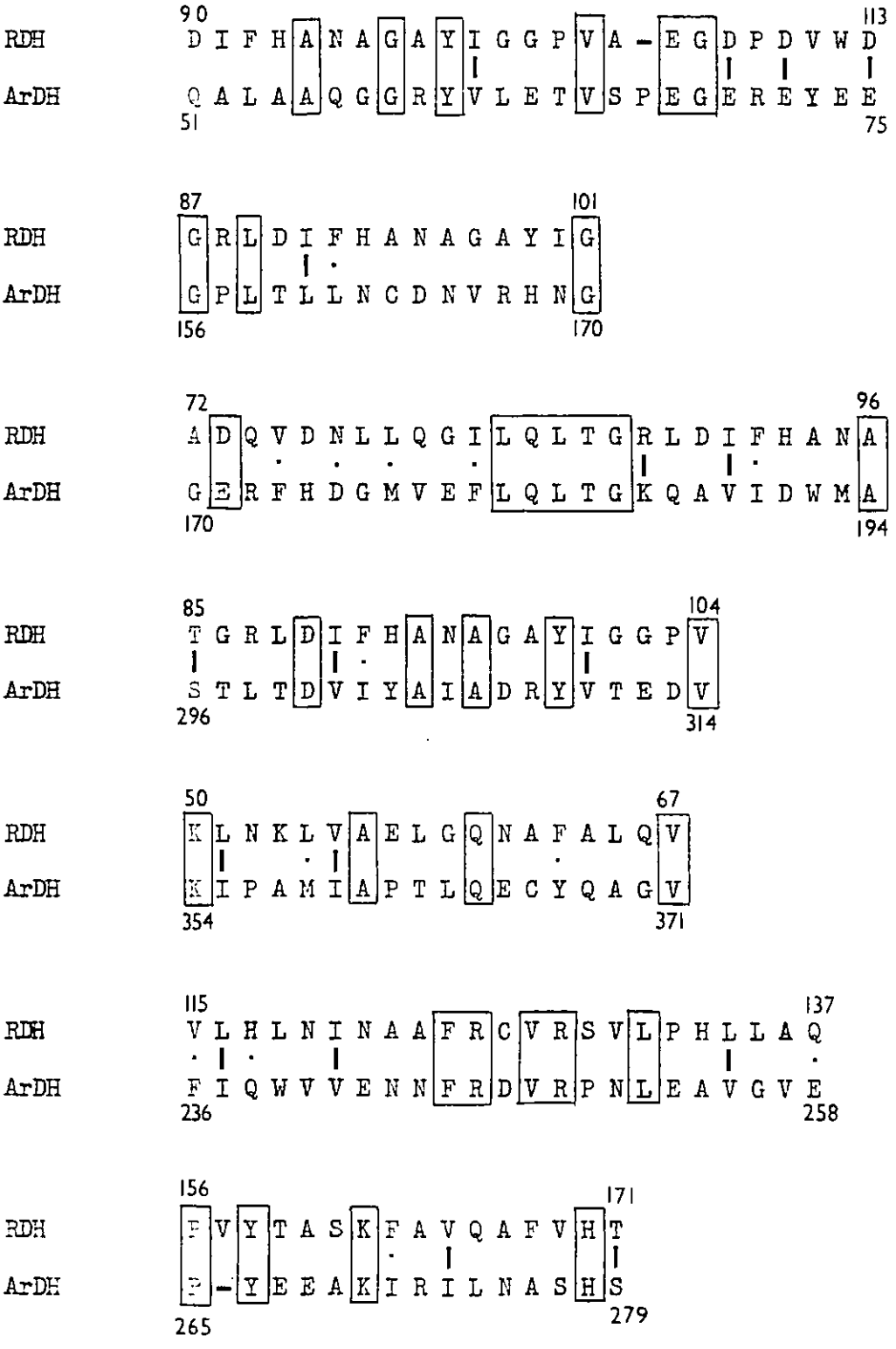
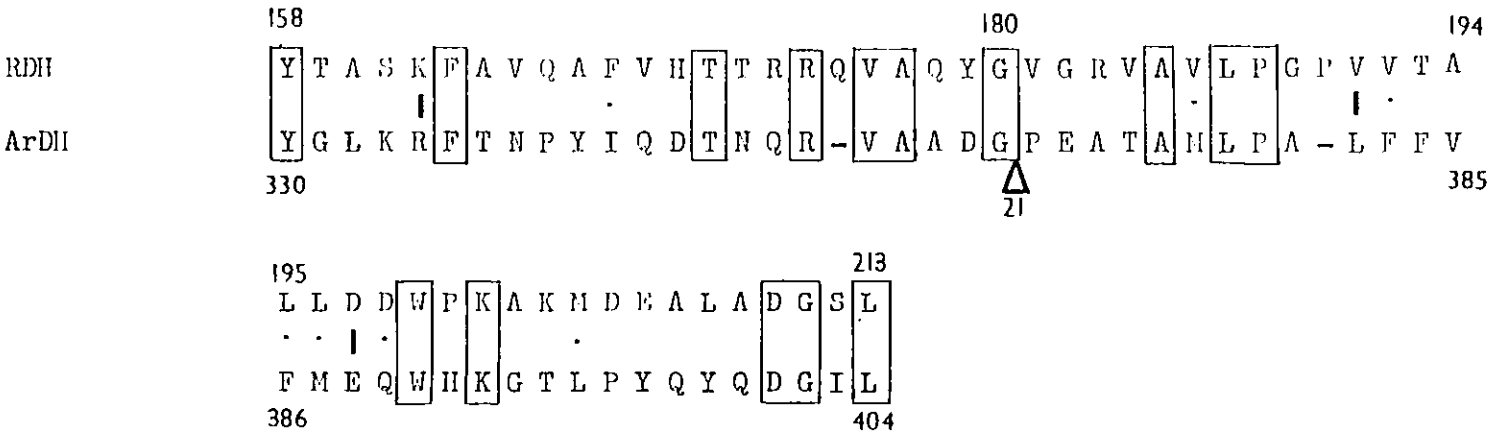


Fig. 4h (contd.)



Identical residues are boxed. Conservative replacements (—) are shown and differences requiring only one base-change are also indicated (•).

with ArDH is given by RDH Tyr·158 - Leu·213, and may indicate a slight similarity in the catalytic site of these proteins. McCarthy (1967) suggests that certain codon sequences will occur quite frequently as a result of similarities in the active sites of many enzymes. The significance of these data remains unclear until X-ray crystallography provides us with a view of the tertiary structures involved.

Repetition of the Nucleotide Binding Site of ArDH?

Engel, (1973) provided evidence that partial gene duplication may play a part in the evolution of regulatory sites. The sequence of Bovine GDH reveals a 50 residue duplicate copy of the NAD^+-binding region which has a regulatory function, but no catalytic activity. The two sequences share 24% identity, and a further 38% of the residues could have resulted from single DNA base-changes.

The DNA sequence centred around nucleotides 360-410 of ArDH appears partially repeated three times (Fig. 4i). In each case the homology in the core region is > 50%. The strongest match gives 29% amino acid identity with a further 35% of residues being accounted for by single nucleotide changes in the codons. There is a close similarity in the predicted secondary nature of these "repeats" and any homologies are probably attributable to this factor alone (Fig.4j).The regions involved are much less extensive than the GDH repeated sequence, and no duplication or partial duplication of the ArDH coenzyme-binding domain is suggested.

Are the dalD and rbtD genes related?

To date, approximately 70% of the rbtD nucleotide sequence is known (Fig. 4k), and so far it tallies with the amino acid sequence derived by classical methods. The entire sequence of

Fig. 4i Partially repeated sequences in dalD

(a)

```
610         620         630         640         650         660
  GTGGACCGCG TCACCCCGCG TCCGGCGGCC GATCTGCCGG CCCGCATCAA GCGCCAAGCC
   ****    *    ** ***    ** *  **** ** ****  *  *     ***    *     *
  CTGGAAACCA GCAACCCCGA TCTGCAGGCC GACCTGCAAG GCGAGTGCAA AACCATTTAC
361         371         381         391         401         411
670         680         690         700         710         720
  GGAATTGATG ACAAAGCGCC GGTGATGGGG GAGACCTTTA TCCAGTGGGT AGTGGAGAAC
  ** *     * **   *  **  *  * * *    * **  **  *    **     **
  GGCACCCTCG ACGCGGATCC TGAAAAGCGC ATGGCCGATA ACGCCGGGCC GCTGACCCTG
421         431         441         451         461         471
```

(b)

```
448         458         468         478         488         498
  CGCATGGCCG ATAACGCCGG GCCGCTGACC CTGCTCAACT GCGATAACGT GCGCCATAAC
  *       **    *** ***   * ** * **    **  *   ****   *       *  * **
  CTGGAAACCA GCAACCCCGA TCTGCAGGCC GACCTGCAAG GCGAGTGCAA AACCATTTAC
361         371         381         391         401         411
508         518         528         538         548         558
  GGCGAGCGTT TCCACGACGG TATGGTCGAG TTCCTGCAGC TCACCGGCAA ACAGGCGGTG
  ***   *     *   **    *           *     *    * ****     * * *  **
  GGCACCCTCG ACGCGGATCC TGAAAAGCGC ATGGCCGATA ACGCCGGGCC GCTGACCCTG
421         431         441         451         461         471
```

(c)

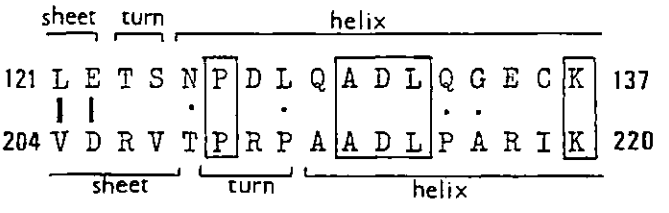
```
930         940         950         960         970         980
  TACGGAGGAC GTTATTCCTG TCTCGGCGAC AACGGCATTG ATCTGCCGAC CTACCGGGTA
    ** **  *     **   *  **   *     **** *  *    * *   *   * *
  CCCGCAGACC AAAGTTATCG CCTTTACCGT GACGGAAGGG GGGTACTACC TGAACACTCG
294         304         314         324         334         344
990        1000        1010        1020        1030        1040
  TGGTCTTAAG CGCTTTACCA ACCCCTATAT CCAGGACACC AACCAGCGCG TCGCCGCCGA
     **    *       * ** ***** ** *  **** *  *    **  ***     *
  CCATCGGCTG GAAACCAGCA ACCCCGATCT GCAGGCCGAC CTGCAAGGCG AGTGCAAAAC
354         364         374         384         394         404
```

The numbers represent the distance in bp from ATG (f-Met).The bottom string in each case is from the NAD^+ binding domain of ArDH.The number of matches are (a) 53, (b) 47, (c) 48.(See Fig. 4j)

Fig. 4j

Amino acid and structural similarities between regions of ArDH which show weak homologies at the DNA level

(a)

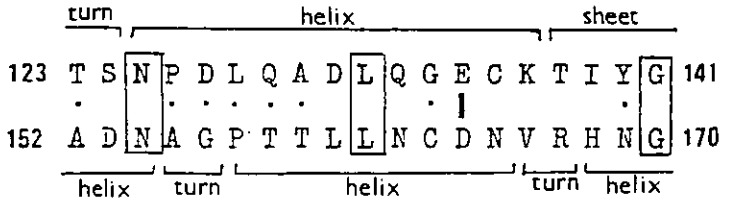


5 identities (29%)

2 conservative changes (|)

4 one-base changes (•)

(b)



3 identities (16%)

1 conservative change

8 one-base changes

Fig. 4k Partial DNA sequence of the rbtD gene from Met. - 1 to Thr. - 161

```
        10          20          30          40          50          60
ATGAAGCACT CTGTCTCCTC TATGAATACT TCCCTCAGCG GTAAAGTCGC CGCGATCACC
TACTTCGTGA GACAGAGGAG ATACTTATGA AGGGAGTCGC CATTTCAGCG GCGCTAGTGG

        70          80          90          100         110         120
GCCGCGGCGT CCGGTATCGG CCTCGAGTGC GCGAGGACCC TGCTTGGGGC TGGCGCAAAA
CCGCGCCGCA GGCCATAGCC GGAGCTCACG CGCTCCTGGG ACGAACCCCG ACCGCGTTTT

        130         140         150         160         170         180
GTGGTACTGA TCGACCGCGA AGGCGAAAAG CTCAACAAAC TGGTCGCCGA ACTTGGCGAA
CACCATGACT AGCTGGCGCT TCCGCTTTTC GAGTTGTTTG ACCAGCGGCT TGAACCGCTT

        190         200         210         220         230         240
AACGCCTTCG CCCTGCAGGT CGACCTGATG CAGGCGGACC AGGTCGATAA CCTACTGCAG
TTGCGGAAGC GGGACGTCCA GCTGGACTAC GTCCGCCTGG TCCAGCTATT GGATGACGTC

        250         260         270         280         290         300
GGCATTTTGC AGCTTACCGG GCGTCTCGAT ATCTTCCACG CCAACGCCGG CGCCTATATC
CCGTAAAACG TCGAATGGCC CGCAGAGCTA TAGAAGGTGC GGTTGCGGCC GCGGATATAG

        310         320         330         340         350         360
GGCGGGCCGG TGGCCGAGGG CGACCCGGAC GTCTGGGACC GCGTGCTGCA CCTTAATATC
CCGCCCGGCC ACCGGCTCCC GCTGGGCCTG CAGACCCTGG CGCACGACGT GGAATTATAG

        370         380         390         400         410         420
AACGCCGCCT TCCGCTGCGT GCGCAGCGTC CTGCCGCATC TGATCGCGCA AAAATCCGGG
TTGCGGCGGA AGGCGACGCA CGCGTCGCAG GACGGCGTAG ACTAGCGCGT TTTTAGGCCC

        430         440         450         460         470         480
GATATTATCT TCACAGCTGT -ATCGCGGGC GTTGGTGCCG GTGATTGGGA GCCTGTCTAT
CTATAATAGA AGTGTCGACA -TAGCGCCCG CAACCACGGC CACTAACCCT CGGACAGATA

        6           16          26          36          46          56
ACGG
TGCC
```

dalD has determined (Chapter 3) and the computer program SEQFIT was used to search for homologies between the two genes. No statistically significant alignments are possible. A number of matches averaging 40% homology are found, but the values are artificially raised by short (10-15 bp) strings of matching bases which correspond to isolated pockets of similar amino acids. Overall the degree of similarity between dalD and rbtD is around 30%, close to the value one would expect for two random sequences. The conclusion must be that the pentitol dehydrogenases are not products of an invert gene duplication. The implications of this finding with respect to the evolutionary origins of the rbt and dal operons are considered in the Discussion section.

CHAPTER 5

THE *dal*DK PROMOTER/OPERATOR REGION AND THE CONTROL OF *dal* EXPRESSION

A large number of prokaryotic promoter sequences have now been studied, and it is possible through comparison to pick out certain common characteristics. It is noticeable that within those regions of DNA controlling transcription some nucleotide sequences appear to be conserved. The identification of numerous mutations which influence transcription have enabled us to assign a functional role to many of these. In the light of these findings, this chapter presents a detailed structure of the *dal* operon control region and discusses those features believed to be important in the regulation of *dal*DK expression. The system is shown to be sensitive to catabolite repression, and preliminary experiments have been carried out which illustrate the binding of RNA Polymerase and Catabolite Repressor Protein (CRP) to the promoter.

I) The structure of the *dal*DK promoter/operator region.

Derivation of the nucleotide sequence

The presence of plasmid pRD351 in *E.coli* HB101 brings about constitutive synthesis of ArDH (Neuberger and Hartley, 1979), but this constitutive expression is repressed by lysogenisation of HB101 (pRD351) by phage λ p *rbt* *dal*$^+$. This illustrates that a functional promoter and operator are present in pRD351. The control region must therefore be confined to the area between *Hind*B and the start of the ArDH gene, a span of 122bp. Maxam and Gilbert sequencing of both strands provided the nucleotide sequences across this region, and these data were confirmed by sequencing M13 clones carrying *Sau*3A I fragments from the promoter area. The involvement of sequences to the left of *Hind*B

in promoter function cannot be totally excluded, and so an additional 180bp were sequenced leftwards from this site towards BstA. Later work on the sequencing of the d-arabitol repressor gene (Chapter 6) extended the sequence still further in this direction. The complete DNA sequence of the promoter and flanking regions is presented (Fig. 5a).

Structural Features of dalDK p/o

Siebenlist et al (1980) analysed a large number of known bacterial and viral promoters and have compiled a consensus sequence based upon the frequency with which each of the four bases occurs in any one position. In particular, it was found that regions around -45, -35, -10 and -1bp (relative to the start point of transcription at +1) tend to be conserved. Some of these trends are visible in the dalDK promoter (Fig. 5b) with several areas having at least partial homology to known RNA Polymerase recognition or binding sites (Pribnow, 1975; Schaller et al, 1975; Rosenberg and Court,1979). An 8-nucleotide string 5'-CCATCAAC-3', commonly associated with transcription initiation is present only 3bp downstream of the putative Pribnow box signal 5'-TACAGTG-3'. If this represents the true start of mRNA synthesis, then the messenger has a short leader sequence of 22 bases preceding the N-terminal AUG codon of dalD. Positioned 8 nucleotides 5' to the initiation codon is the hexamer 5'-AAGGAG-3', which may be involved with interactions between the 30S ribosomal subunit and the mRNA (Shine and Dalgarno, 1974).

The entire region between HindB and the proposed RNA Polymerase binding site at -10 is A+T rich (Fig. 5c), a feature characteristic of many promoters (Rosenberg and Court, 1979), although detailed sequence homologies are not generally observed, and its significance remains unclear. A+T rich

HinfI

```
        10         20         30         40         50         60
GATTCTGGCC GGCGATGTAG TACATCCAGC GGCACGCACC TTCTGATCCA ACCGGATATC
CTAAGACCGG CCGCTACATC ATGTAGGTCG CCGTGCGTGG AAGACTAGGT TGGCCTATAG

        70         80         90        100        110        120
GTCTTCTTTA CTCATGGTTT CCCGCCCTAA GTTACTGGCC AGGGCACGCA ACCGGCCCGG
CAGAAGAAAT GAGTACCAAA GGGCGGGATT CAATGACCGG TCCCGTGCGT TGGCCGGGCC

                                          HindB
       130        140        150        160        170        180
CGAATACCCG CATTCTCGCC GTGGTTGGCA TAACTGGCAA GCTTTTGCTC TTTTCTGGTC
GCTTATGGGC GTAAGAGCGG CACCAACCGT ATTGACCGTT CGAAAACGAG AAAAGACCAG

       190        200        210        220        230        240
ATTTGTAATT TAATTGGGTA ATTGCTCTTT TGTGATCTAT GGCTCTTATT TAGGTCAAAT
TAAACATTAA ATTAACCCAT TAACGAGAAA ACACTAGATA CCGAGAATAA ATCCAGTTTA

                                   dalD
       250        260        270        280        290        300
GATCAATTAC AGTGGCGCCA TCAACTCAAG GAGAGCAGAA CATGAACAAT CAATTCACAT
CTAGTTAATG TCACCGCGGT AGTTGAGTTC CTCTCGTCTT GTACTTGTTA GTTAAGTGTA
```

Fig. 5a DNA Sequence of the dalDK promoter region and N-terminus of the ArDH gene

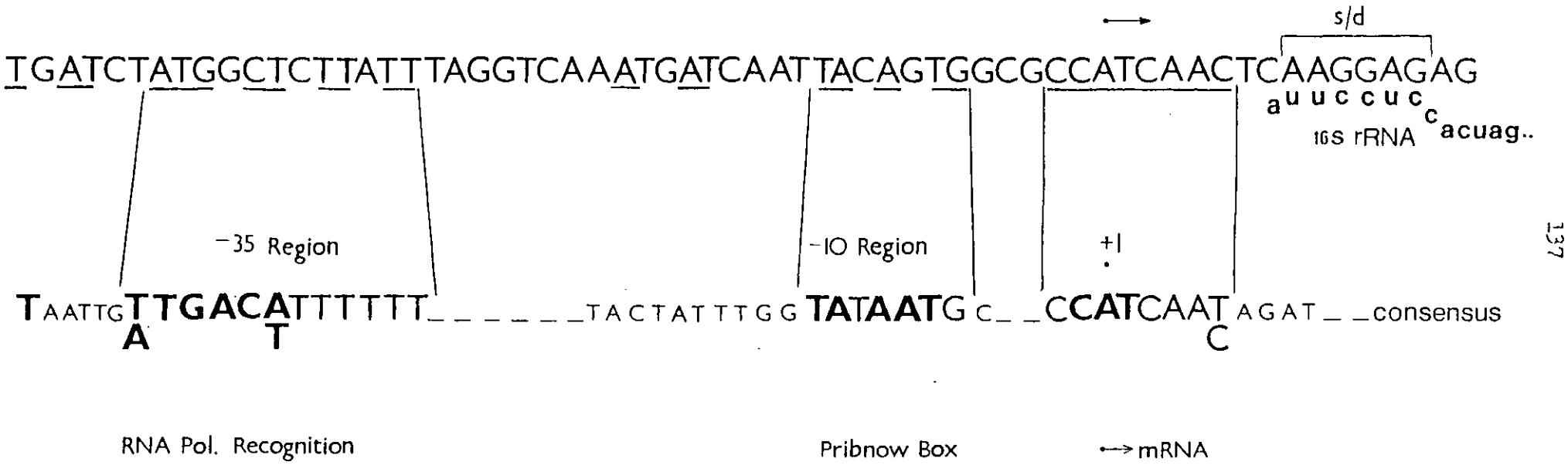


Fig. 5b A comparison of the dalDK promoter region with a promoter consensus sequence. Bold letters represent those bases which are most strongly conserved. The dal Shine/Dalgarno sequence is marked s/d. The consensus sequence is that of Siebenlist et al, (1980).

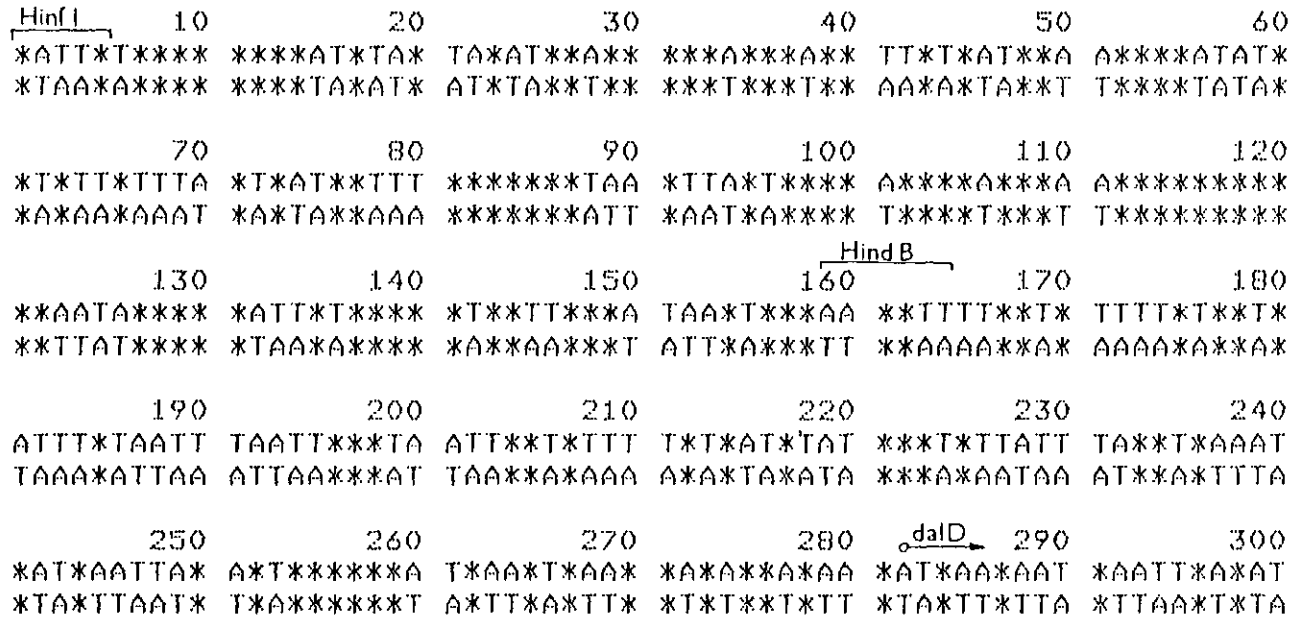


Fig. 5c Base composition of the dalDK promoter rgegion. G and C residues are replaced by asterisks. Note the A+T rich stretch between positions 120 - 250 and the blocks of G.C pairs between 80 and 120bp relative to the Hinf I site.

regions may facilitate local melting of the duplex and allow entry of RNA Polymerase. Nakamura and Inouye (1979) observe that for the E.coli outer membrane lipoprotein promoter the bases A and T tend to occur in blocks of 8-14bp within which one strand contains overwhelmingly either A or T. This is true to a lesser degree of dalDK p/o, where A·T pairs aggregate in groups of 4 to 10. The same authors also looked at the A+T content of regions preceding the start of transcription in various other operons. Between -1 and -45 the figures for araBAD are 58% A+T, gal 64%, trp 64% and lac 58%. The dalDK promoter compares favourably with two of the catabolite repressed promoters at 58% A+T across the same region. Base distribution in the promoter is uneven. The antisense strand is 65% pyrimidine rich (Fig. 5d), a phenomenon also noted by Rosenberg and Court (1979).

An extensive dyad symmetry stretches from positions -10 to -62. Such structures are not uncommon in prokaryotic promoters, but their importance is largely unknown. The whole region may be represented as a stem and loop structure (Fig. 5e) in which the stem, containing only two mismatches and a small bulge loop, is 20bp in length. The terminal loop consists of 11 unpaired bases which include the important -35 region. In the argTr and dhuA promoters (the control regions of argT and hisJQMP, two transport operons under "nitrogen control"), stem and loop structures carry the -35 region and fall in the approximate centre of a sequence having mirror-image symmetry (Higgins and Ames, 1982). Sequences either side of the "foot" of the stem are identical in both argTr and dhuA and homologous to a region of the regulatory locus of the histidine biosynthetic operon, in which bases -30 to -44 are also enclosed in a hairpin loop structure. The discovery of such structures in all three

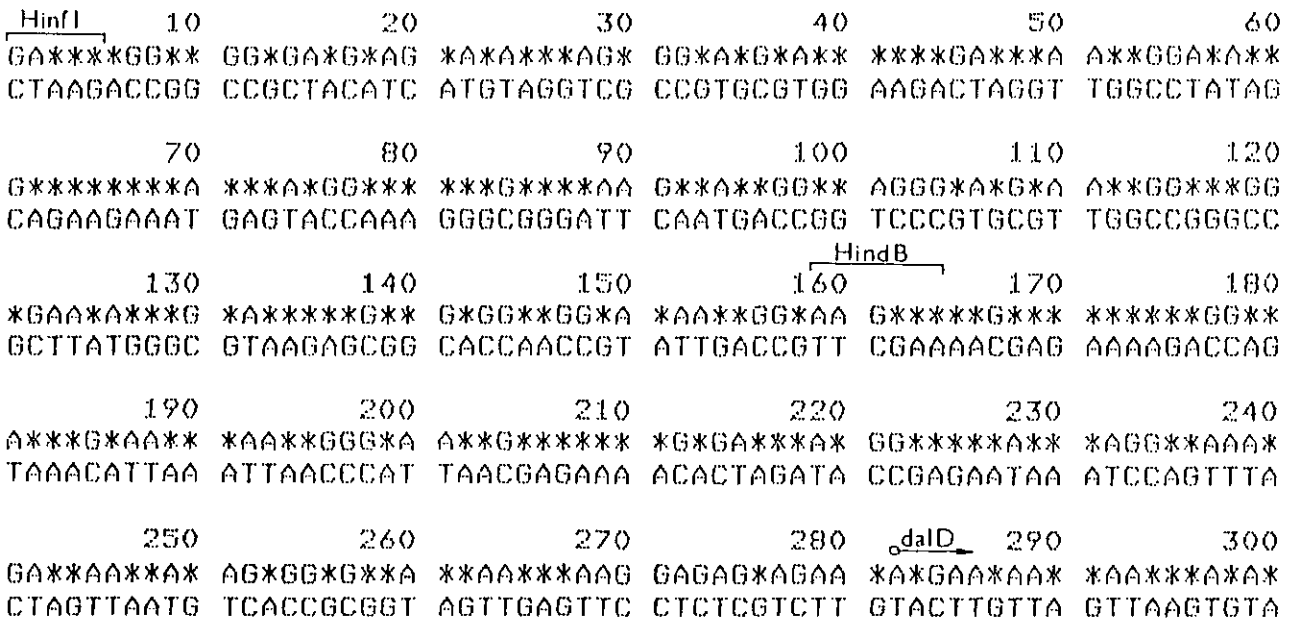


Fig. 5d The dalDK promoter, demonstrating the pyrimidine-rich antisense strand. C and T residues are replaced by asterisks.

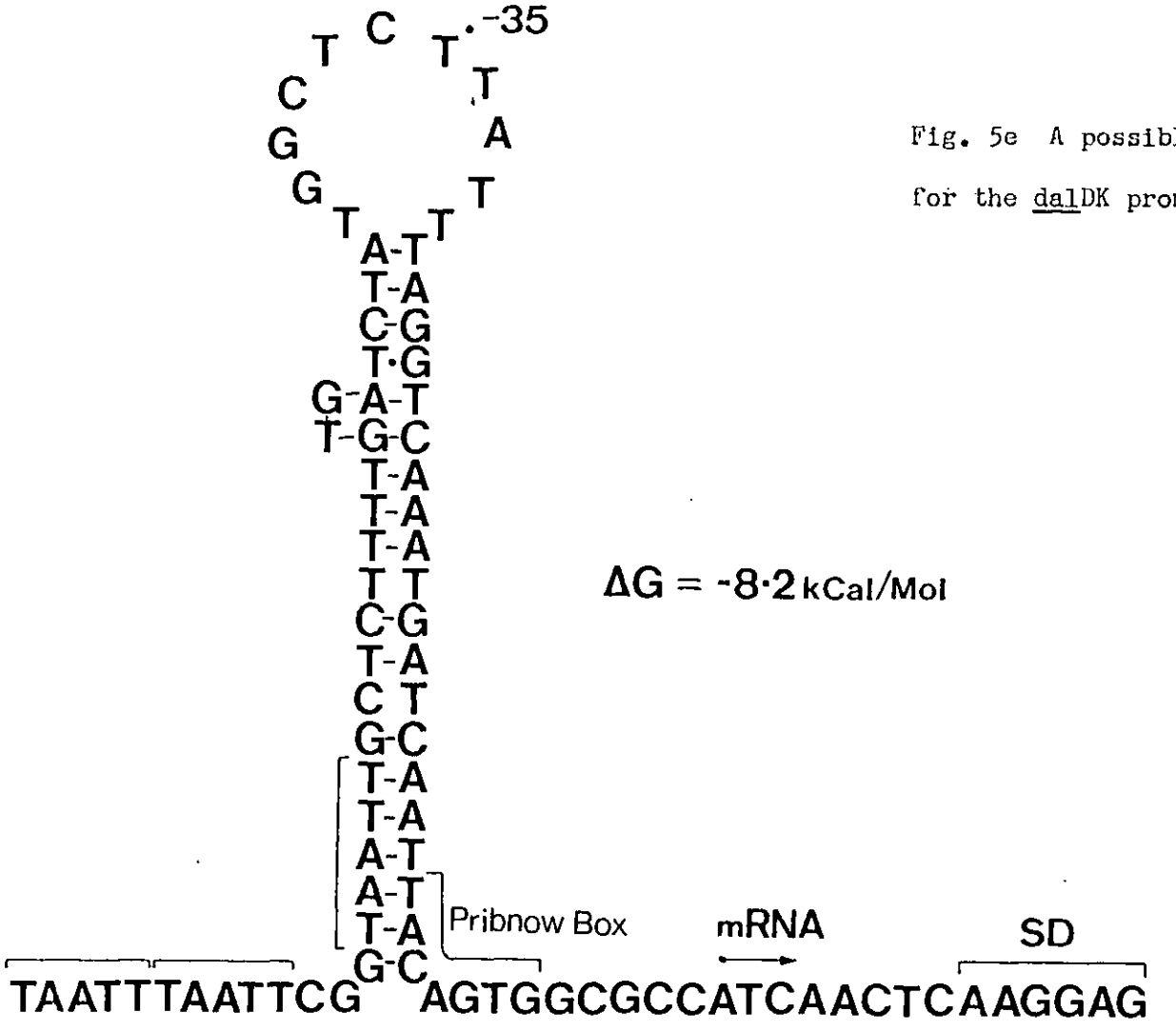


Fig. 5e A possible secondary structure for the dalDK promoter region.

promoters strongly suggests a specific function. The mirror symmetries of argTr and dhuA are sites for the recognition and binding of the ntrB and ntrC gene products which activate or repress transcription according to nitrogen availability. The symmetry implies a dimeric or multimeric protein and any interaction must be independent of the polarity of the sugar-phosphate backbone since this is reversed in the two copies. This represents a new class of protein-DNA interaction, since no secondary structures are possible, and therefore the primary nucleotide sequence must be the important factor. These results are interesting since a large mirror symmetry is also found in dalDK p/o (Fig. 5f). It is similar to the one in argTr, being T-rich overall, and having a run of contiguous T residues in the largest symmetry element. Both are also A-deficient. The dal mirror symmetry has 15/29 residues identical either side of the central A·T pair at -69bp. This compares with 13/24 and 11/25 in argTr and dhuA respectively. Higgins and Ames have also found mirror image symmetries in the neighbourhood of the promoters for his, trp, lacI and glnA.

In dalDK p/o, sequences similar to those in the -35 region are partially re-iterated five times (Fig. 5g), something that is true also of the gal operon in E.coli (Willmund and Kneser, 1973). One of the most striking features of the promoter is a conserved sequence of 17bp, which is repeated three times in the same orientation (Fig. 5l). Similar direct repeats in other operators are implicated in the binding of regulatory proteins (Maniatis et al, 1975a; Pirotta, 1975; Tsurimoto and Matsubara, 1981).

The remainder of this chapter will deal in more detail with some of the points raised here, and present evidence to support the functions assigned to specific sequences.

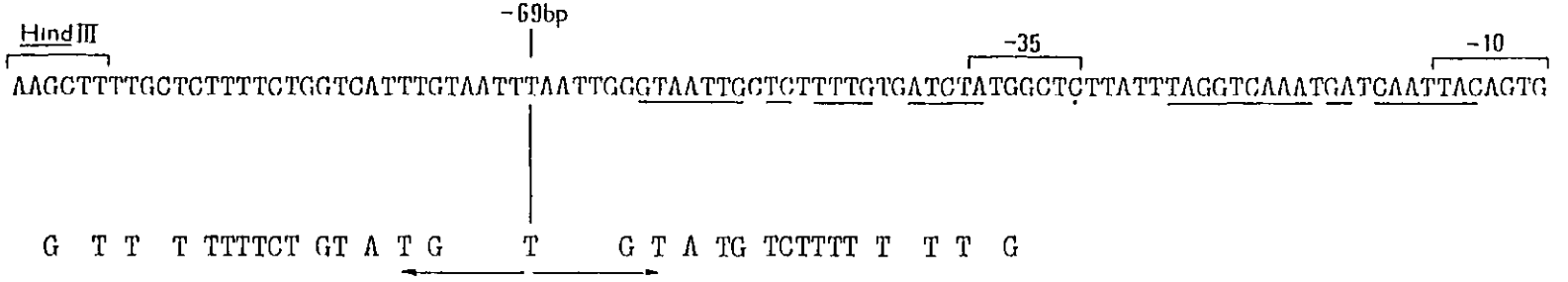


Fig. 5f A large mirror-symmetry element in the dalDK promoter region.The positions of RNA Polymerase binding and recognition sequences and of an extensive dyad symmetry are shown.

Fig. 5g Repeated RNA polymerase binding sites in the dalDK promoter. Sequences homologous to the -35 region of a number of E.coli promoters are reiterated 5 times(A to E above).

II) The control of transcription in the dal operon.

Catabolite Repression

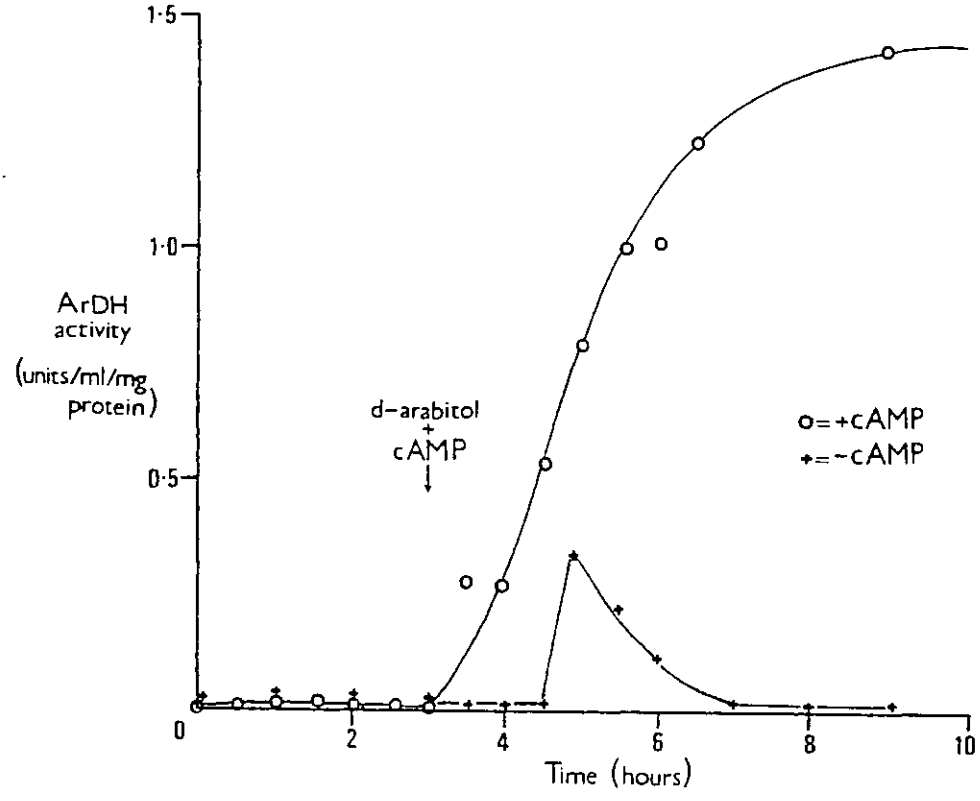
A promoter is a region of DNA required for the initiation of transcription. As such it must contain all the information necessary for the interaction of RNA Polymerase and regulatory proteins. The latter may take the form of repressors or positive stimulatory molecules such as CRP. The catabolite repression of inducible enzyme synthesis by glucose, glucose-1-phosphate, fructose or glycerol has been recognised for some time (Magasanik, 1961; Monod, 1947) and affects pathways involved in the degradation of a number of carbohydrates. The literature surrounding the effects of cAMP and its receptor protein CRP is extensive, but good reviews may be found in de Crombrugghe and Pastan (1978) and Zubay (1973). Glucose lowers the intracellular concentration of cAMP in E.coli (Makman and Sutherland 1965), and addition of cAMP to cells growing on glucose can increase the rate of synthesis of certain enzymes (Perlman and Pastan, 1968; Varmus et al, 1970). To date, the control regions of a number of catabolite sensitive operons have been sequenced, and those sections of the DNA important for binding the cAMP·CRP complex identified, (lac, Simpson, 1980; ara,Ogden et al, 1980; gal, Taniguchi et al, 1979). Catabolite sensitive promoters thus comprise a special class, whose activity is regulated through the coordinate action of RNA Polymerase and a cAMP·CRP complex. Various models have been put forward for the mechanisms behind CRP-stimulated enhancement of transcription (Gilbert, 1976), and very recently the detailed molecular mechanism by which cAMP·CRP interacts with the template DNA has been published (Ebright and Wong, 1981).

Preliminary work by Neuberger (1978) suggests that the

synthesis of RDH is subject to repression during growth on glucose and earlier experiments (Lin, 1961) imply that ArDH synthesis in A.aerogenes is similarly affected. Lin found that mannitol too was capable of suppressing ArDH activity to 1/3 of the normal induced levels when included in the growth medium along with d-arabitol. Strong evidence existed, therefore, that dal expression might be catabolite sensitive, but conclusive tests had not been performed.

The addition of cAMP to an exponentially growing culture of PS640 (a λ p rbt dal lysogen inducible for the dalDK operon) in the presence of both glucose and inducer resulted in an increase in ArDH synthesis relative to a control culture (Fig. 5h), illustrating that dal expression is linked to intracellular cAMP levels. The lac, gal and araC/araBAD operons are among many known to bind CRP and much is now known about their sequence organisation (Reznikoff and Abelson, 1978; Ogden et al, 1980; Taniguchi et al, 1979). Recently, new data arising from a study of a hitherto unknown catabolite sensitive promoter on pBR322 have resulted in a consensus sequence for the cAMP·CRP interaction site being formulated (Queen and Rosenberg, 1981). Within the dalDK promoter a stretch of bases showing considerable homology to this prototype sequence and to the sequences of other catabolite sensitive promoters, particularly gal, can be found (Fig. 5i). The basic structural element common to each is a small dyad symmetry of the general form 5'-TGTG N_8 CACA-3', but its position relative to transcription initiation varies and indicates perhaps a slightly different role in each promoter. The CRP site of pBR322 P4, gal and araC overlaps the -35 region, whereas in lac and araBAD it precedes this area by 30 and 60 nucleotides respectively. In the dalDK promoter, the putative CRP site lies between positions -33 and -50, overlapping the

Fig. 5h The effects of cAMP on ArDH levels in the λ*prbtdal* lysogen PS640



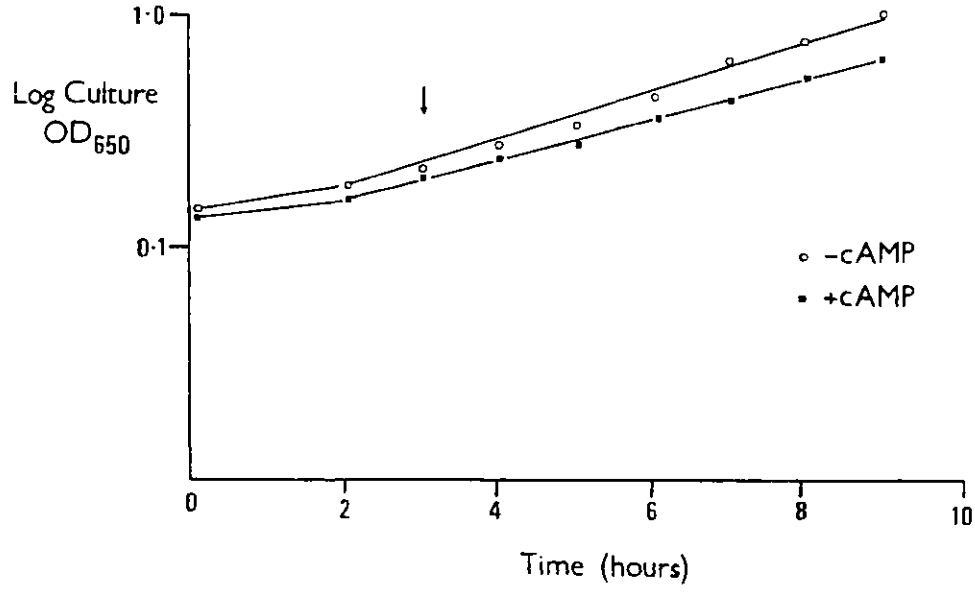


Figure 5h

A single colony of the λprbtdal lysogen PS640 was grown in M9 + arabitol and thiamine, subcultured into 200ml of M9 + CAA +thiamine and grown to late log phase at 32°C. The cells were pelleted and resuspended in 20ml of 10mM $MgSO_4$. 5ml of these cells were then innoculated into two parallel culture flasks containing 200ml of M9 glucose + 0.2% sodium gluconate and 1μg/ml thiamine. Cultures were shaken at 32°C to prevent induction of the prophage. During exponential growth, 10ml samples were removed at 30 minute intervals and cell-free extracts prepared by lysozyme + DNAase I treatment (see Materials and Methods). Extracts were then assayed for ArDH activity. D-arabitol (0.2% w/v final conc.) was added to both cultures at t=180 minutes and cAMP (15mM final) to one culture only. Shaking continued at 32°C with 1ml and 10ml samples being taken at 30 minute intervals for OD_{650} measurements and enzyme assays respectively.

Protein estimations on the cell-free extracts were made using the Coomassie method described in Materials and Methods, and ArDH activity was determined spectrophotometrically at 340nm.

The results are plotted as ArDH units / ml culture / mg protein. The addition of solid cAMP (free acid form) to 15mM resulted in a pH drop of 0.3 units in that culture and a marginally slower growth rate thereafter, although the latter may have occured due to induction of non-essential, energy-consuming pathways normally catabolite repressed.

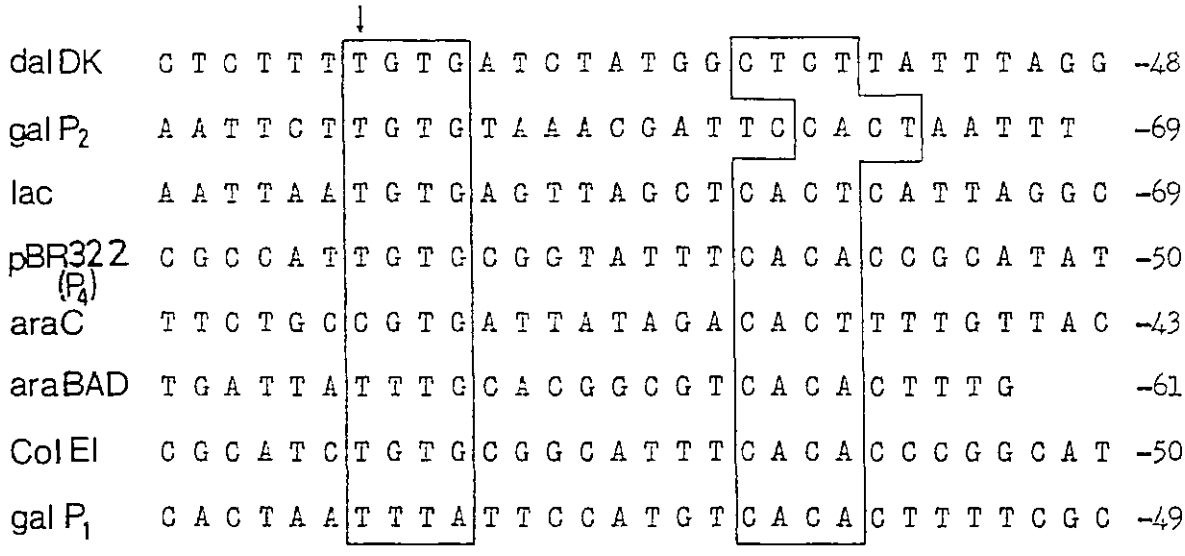


Fig. 5i The DNA sequences of the CRP-binding regions of a number of catabolite sensitive E.coli promoters. A sequence resembling 5'-TGTG(N_8)CACA-3' is conserved throughout. The figures in the right hand margin give the position of the first base (arrowed) relative to the origin of transcription.

polymerase recognition site.

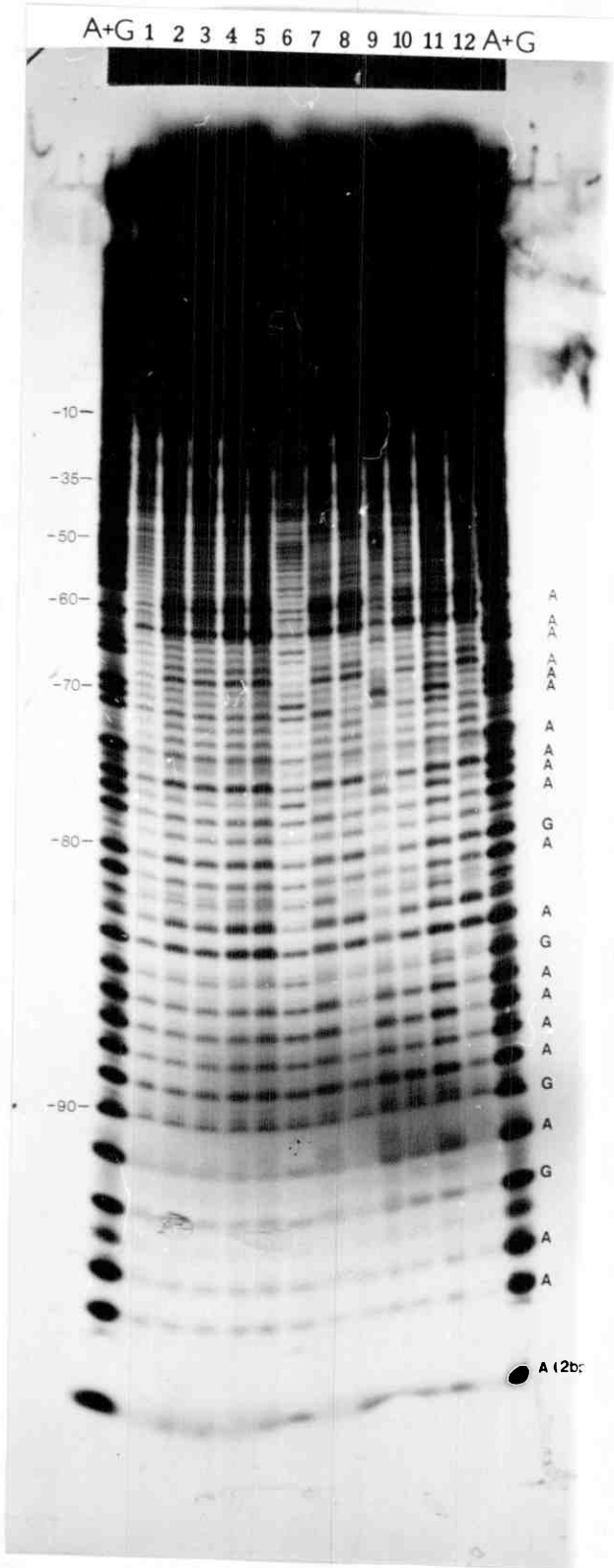
An Attempt to Localise the Binding Sites for RNA Polymerase and CAMP CRP

Although the DNA sequence of the dalDK p/o region displays many familiar structural features, so far these have been assigned functional roles on a purely comparative basis. It was felt that experiments were needed to support and verify the accuracy of these statements, in particular the suggestion of multiple RNA Polymerase binding sites. To study protein-DNA interactions, the DNAase I footprinting methods of Galas and Schmitz (1978) were used. Experimental procedures were essentially as described in the published material, but it was necessary to reduce the EDTA in stop-solutions to 100mM in order to minimise the formation of an insoluble pellet after ethanol precipitation.The promoter was isolated in the form of the 0.57kb HindB/BstB fragment, 3' labelled at the Hind III site by the Klenow fragment of DNA Polymerase I and α-[^{32}P]-dATP. DNA was isolated from gels by electroelution or butanol/CTAB extraction, as outlined in the "Materials and Methods" section. The protocols of Willmund and Kneser (1973) were used as a guide-line for the binding conditions for RNA polymerase and CRP. Precise experimental details are included in the figure legends. DNA size markers were provided by the products of a pyridinium formate (A+G) sequence ladder (Maxam and Gilbert, 1980), and electrophoresis was performed on urea/acrylamide sequence gels (Materials and Methods).

Protein-DNA interactions in the dalDK promoter

Photo. 5A shows the effects of bound RNA Polymerase and CRP on DNAase I partial digests of promoter-carrying DNA fragments. It is apparent that although there are no extensive regions of the template completely shielded from DNAase I

Photograph 5A



Photograph 5A

Protection of the dalDK promoter region from DNAase I by RNA polymerase and CRP.

The promoter-carrying HinB/BstB fragment was end-labelled at the Hind III site by DNA polymerase I and $\alpha[^{32}P]$dATP. Labelled DNA (20,000cpm) was then taken up in 40µl of Binding buffer (20mM Tris-HCl pH7.9, 10mM $MgCl_2$, 5mM $CaCl_2$, 0.1mM DTT, 0.1mM EDTA, 100mM KCl) 0.2mM cAMP. Further additions were made as indicated below, (volumes are in µl):

										-cAMP		
—	1	2	3	4	5	6	7	8	9	10	11	12
Pol	5	0.5	1.5	2.5	5	-	2.5	2.5	-	2.5	2.5	2.5
CRP	-	4	4	4	4	-	-	4	4	-	4	4
rATP (0.5mM)							33 (200µM)			33		33
rGTP (2mM)							10 (200µM)			10		10
heparin								2				
dH_2O	55	55.5	54.5	57.5	51	60	14.5	51.5	56	14.5	53.5	10.5
Final Conc. Heparin, 100µg/ml; CRP, 75µg/ml; RNA Pol, 10-100µg/ml												

In reactions 2,3,4,5,11 and 12, CRP was added first, followed by a 5 minute pre-incubation step before the addition of RNA polymerase. Incubation was then continued for a further 10 minutes at 25°C. Incubations with CRP or polymerase alone were for 10 minutes at the same temperature. In 8, the order of addition was polymerase, heparin, CRP.

DNAase I digestion was performed at 25°C for 30 seconds using 0.15µg/ml DNAase I. The reaction was terminated by adding 25µl of 0.5M EDTA containing 100µg/ml tRNA. DNA was then ethanol precipitated, taken up in 4µl of formamide + BPB, boiled for 5 minutes and electrophoresed on a 20% acrylamide sequence gel at 2000V. Autoradiography was at -70°C on pre-fogged Kodak XH-1 film with a phosphotungstate screen for 5 days.

cleavage, there are significant differences in the overall distribution and intensity of bands. The absence of total protection is attributable, I believe, to several factors. Firstly, non-saturating quantities of RNA Polymerase or a deficiency of the more labile sigma factor, essential for binding. Secondly, the impurity of the commercial preparation of CRP. Electrophoresis on SDS-Polyacrylamide gels showed that the CRP was not homogeneous, but contained large quantities of extraneous protein matter. This set of experiments should ideally be repeated using freshly prepared CRP (Anderson et al, 1971), but despite some problems, much information can still be derived from the footprinting data.

Regions of the promoter partially protected from DNAase I digestio

[N.B. In this section the lane numbers refer to Photo 5A]

RNA Polymerase alone cannot protect dalDK promoter DNA from DNAase I digestion (Lane 1 and 6). Compared to ^{32}P-labelled fragments digested under identical conditions, pre-incubation with polymerase has the following effects : a) diminished band intensities at -77, -70 and -65bp; b) partial protection of the region between -50 and -57 (this area is fainter in Lane 1); c) some evidence of two regions of enhanced sensitivity to DNAase I between -10 and -30bp. A more accurate interpretation is hampered by the closeness of bands in the upper portion of the gel.

Inclusion of rATP and rGTP to 150μM in the binding buffer results in several differences in the cleavage pattern. Polymerase binding now enhances cutting around -25 and -10bp, but cleavage is increased at -60 and -63bp and, to a lesser extent, at -67/68, -76 and -78/79bp. Attack by DNAase is reduced at -50/55, -65/66, -70, -77 and -85bp.

When cAMP and CRP are present together, then the pattern of

digestion products is identical to that of DNA alone, except for small differences around -50bp. This does not preclude the possibility that CRP binding is exclusively protecting the complementary strand. The cAMP dependence of polymerase and CRP binding is seen in a comparison of lanes 4 and 11, the latter lacks cAMP but is identical otherwise. In the absence of this factor, the ladder resembles the control (lane 6) in many respects, yet enhancement at -60 and -63bp is retained.

The order of addition of the components is important also. The products of reaction 8, where polymerase was allowed to bind before CRP, show that bases -71, -82 and -85 to -88 inclusive are partially protected. This is not true if CRP is added first (lanes 2-5). This protection does seem to be CRP dependent, since polymerase alone does not prevent attack at these positions in the presence (or absence) of cAMP (lanes 7 and 10). A cAMP·CRP complex may therefore permit RNA Polymerase to partially protect two regions of DNA which the polymerase alone is unable to recognise or bind.

Protection of some regions of the dal promoter by CRP and polymerase has been shown to be cAMP dependent. The results also suggest that the effects of cAMP are mimicked to some extent in vitro by rATP and rGTP (lanes 11 and 12). Addition of these triphosphates at the binding stage leads to partial polymerase protection of regions between -65/-71, -77/-82 and -85/-90. This cannot be achieved by polymerase, rATP and rGTP alone (lane 9). In the E.coli gal operon, Willmund and Kneser (1973) showed that rATP and rGTP, in the absence of cAMP and CRP, permit one polymerase molecule to bind at a position from which gal transcription may be initiated.

A follow-up of these very exploratory experiments is

essential. Conclusive comments cannot really be made in the absence of total protection from DNAase digestion, but the data do seem to support the theory of multiple RNA Polymerase interaction sites as far back as -90bp from the transcription start. The "protected" areas agree quite well with repeated -35 regions depicted in Fig. 5g. The conditions for obtaining the best DNAase I partials were determined in advance of these experiments but attempts were not made to optimise cAMP, CRP or polymerase concentrations; factors which are also known to influence the extent of protection (Taniguchi et al, 1979; Queen and Rosenberg, 1981). CRP binding in vitro is inhibited by KCl (which stabilises the DNA helix) and enhanced by the mild denaturant glycerol (Dove and Davidson, 1962; Nakanishi et al, 1974; Schmitz, 1981). Future experiments should take these variables into consideration. Protection studies must also be performed on the complementary strand. Regulatory proteins bound at operators and promoters are known to make specific contacts with both strands (Majors, 1975). It is hoped to carry out these experiments in conjunction with repressor/operator-binding studies. Work is already under way to superproduce the dalR protein with a view to purification and eventual crystallisation.

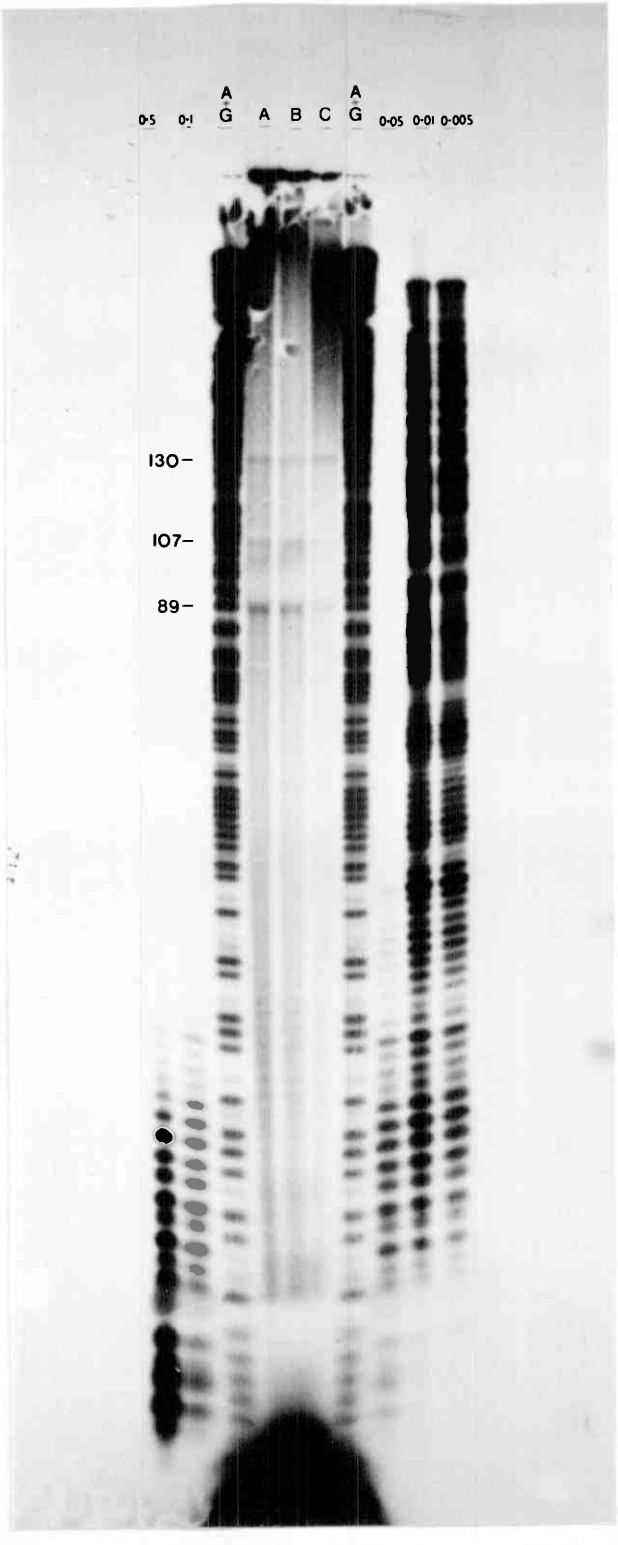
Identifying the start point of dal mRNA transcription

A number of well established methods are available for determining RNA sequences based upon limited enzymic and chemical fragmentation of the RNA, followed by electrophoretic or chromatographic separation of the ^{32}P-labelled products (Sanger et al, 1965; Jay et al, 1974). A simpler scheme was devised for localising the position of transcriptional initiation in the dalDK promoter. This method would rely on the accurate sizing of a ^{32}P-labelled run-off transcript made in vitro from a promoter-carrying restriction fragment. The HindB/BstB fragment isolated

from pRD351 was cut with Hpa II. This enzyme does not cleave within the promoter itself; the nearest site is 89bp into the ArDH gene. If the Pribnow box has been correctly identified, then the major RNA synthesised from a HindB/Hpa II fragment will be 110 nucleotides long. In vitro transcriptions were performed, and the products sized on 12% and 20% 8M urea sequence gels against size markers generated by Pyridinium formate degradation of the same end-labelled fragment. Experimental details can be found in the relevant figure legends.

A number of small transcripts are produced (Photo. 5B) corresponding to approximately 89, 107 and 130 nucleotides in length. On a higher percentage gel additional oligonucleotides of 4 and 10 residues are also resolved (Photo. 5C). The transcript of about 107 bases probably represents the run-off product initiated at +1 in the sequence CCATCAAC (see Fig. 5b). The smaller species are most likely to be paused transcripts or products of faulty initiation (Gilbert, 1976). Nicks in the DNA duplex, sometimes caused by restriction enzymes at site analogous to their recognition sequences, can serve as good initiators for core-polymerase in vitro, particularly if the specificity of binding is reduced by a deficiency of sigma factor (Vogt, 1969). The large transcript could have arisen from an initiation event nearer to HindB. It has already been suggested that the promoter region has a number of possible recognition sequences, and some of these precede potential Pribnow boxes. For example, TATTTAG centred around -30bp could function with a "-35 region" between -58 and -49bp to promote initiation from a string of adenine residues near -20bp. The product (130 bases) is consistent with the larger transcript made in vitro from the Hind III/Hpa II fragment. RNA Polymerase binding experiments indicated some degree of protection across

Photograph 5B



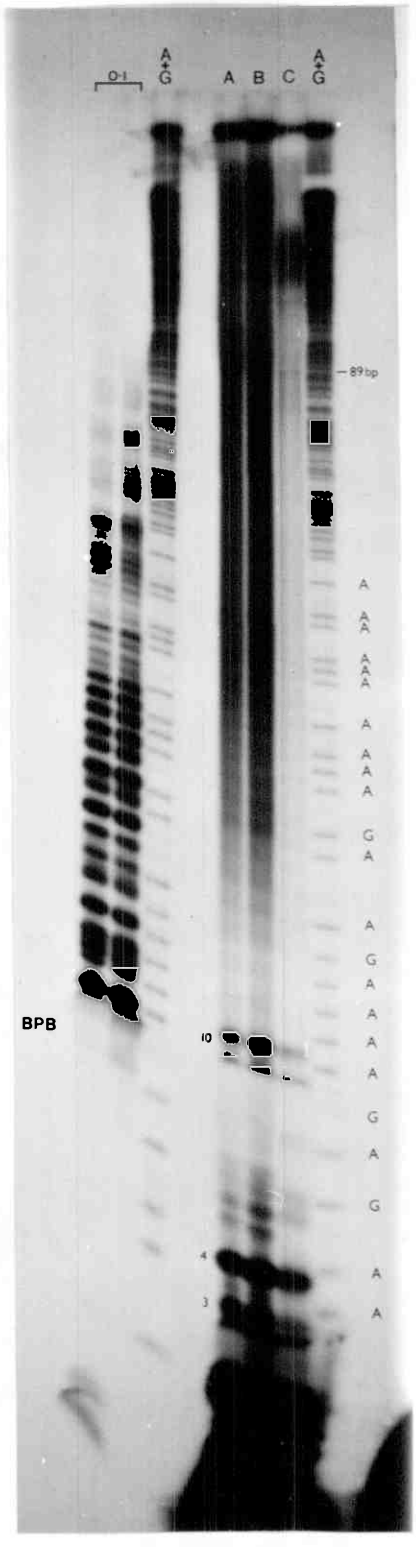
Photograph 5B

In vitro transcription of the *Hin*B/*Bst*B fragment.

Lanes A,B and C show the products of *in vitro* transcription of 1μg, 0.5μg and 0.2μg of the isolated fragment. Transcriptions were performed in 50μl of 20mM Tris-HCl pH7.9,80mM KCl,10mM $MgCl_2$,1mM DTT,0.1mM EDTA, 0.2mM UTP,CTP and TTP containing 10μCi ATP (400i/mMol).Reactions were started by the addition of 1μg of RNA polymerase and,after 30 minutes at 37°C, terminated by ethanol precipitation. Samples were resuspended and boiled for 5 minutes in 10μl of formamide + BPB then separated on a 12% urea/acrylamide sequence gel with partial DNAase I digests of the 0.35kb *Hin*B/*Bst* I fragment from pRD351 as size markers.Autoradiography was overnight on Fuji RX film.

Partial digests were carried out using 30,000cpm of the *Hin*B/*Bst* I fragment (3' labelled at *Hind* III) in 100μl of DNAase I buffer (20mM Tris-HCl pH8,10mM $MgCl_2$,0.1mM DTT,50mM NaCl). The amounts of DNAase I used are shown in the photograph(μg/100μl).Reactions were performed at room temperature for 30 seconds and stopped by adding 25μl of 0.5M EDTA containing 100μg/ml tRNA.

Photograph 5C



Photograph 5C

In vitro transcription of a *Hind* III / *Hpa* II fragment carrying the *dal*DK promoter.

Lanes 1 and 2 are DNAase I digests of a 0.35kb *Hin*B/*Bst* I fragment from pRD351 prepared as described in the legend to Photograph 5B. Tracks 3 and 7 are A + G sequence ladders of the *Hin*B/*Bst*B fragment, labelled at *Hind* III. Lanes 4,5 and 6 show the transcripts of:

A 0.5μg *Hin*B / *Hpa* II DNA

B 1μg *Hin*B / *Hpa* II DNA

C 1μg *Hin*B / *Hpa* II DNA transcribed at 3μM GTP

Transcriptions were as detailed in the legend to Photograph 5B. Electrophoresis was on a 20% urea / acrylamide sequence gel at 2000V and the gel was exposed to pre-fogged Kodak XH-1 film at -70°C with an intensifying screen for 3 days.

the region -55 to -50 and, although the alternative Pribnow box is not as good as the one at -10bp, initiation from a second site *in vitro* remains a distinct possibility. Indeed, expression of the *E.coli gal* operon is controlled by two overlapping promoters; one stimulated, and the other inhibited by cAMP·CRP (Musso *et al*, 1977).

The very small RNAs of about 4 and 10 nucleotides have yet to be fully accounted for, but are almost certainly paused transcripts, like those found in the *E.coli lac* operon of between 4 and 17 bases long by Maizels (1973). Conditions of limiting rGTP permit limited chain elongation (Krakow *et al*, 1976), but RNA Polymerase may pause before incorporating GTP into the growing chain. One transcription was carried out at limiting GTP concentration, and the same small transcripts are seen. No additional RNAs can be produced by inducing the polymerase to stall at G-rich regions in the template.

To summarise, one *in vitro* transcript is consistent with an initiation event 100bp from HindB, but other sites may also have promoter activity. Specificity might be improved by including CRP in the transcription reactions. Ideally, the mRNA should be isolated, end-labelled and fingerprinted to place this particular issue beyond doubt.

A Ribosome Attachment Site on *dal*DK mRNA

The 5' end of the *dal* messenger predicted by the DNA sequence includes a string of six purine nucleotides which are complementary to a region near the 3' end of *E.coli* 16S rRNA (Shine and Dalgarno, 1974; Steitz and Jakes, 1975). Current theories support the view that during the initial stages of protein biosynthesis the 3' end of 16S rRNA pairs with a polypurine string common to the 5' ends of bacterial and phage mRNAs. This involves Watson-Crick base-pairing

between four or more nucleotides. The Shine-Dalgarno (S/D) sequence is usually situated within 3-15 bases of the initiator AUG codon. The S/D sequence alone is often insufficient to account for the variation in translational efficiency of messengers with equally good ribosome-binding sites (Hall et al, 1982). An RNA hairpin loop is also an important factor and the S/D sequence facilitates recognition of this structure and stabilisation of the interaction. Messengers having strong S/D sequences may lack a RNA hairpin altogether, and many of the bacteriophage ØX174 and G4 mRNAs fall into this category. The dalDK mRNA has a very good S/D sequence (AAGGAG) and the untranslated leader can be arranged into a small, although not very stable, hairpin in which this sequence is exposed in the unpaired loop (Fig. 5j).

Repeated Sequences within the dalDK Promoter

Three direct repeats occur within the dalDK p/o region. The first of these (R1) is situated between positions -75 and -95bp, 20 nucleotides upstream of the other repeats (R2 and R3), which span the region from -23 to -58bp. R1 and R2 are both 17bp in length, and are A+T rich. R3 is longer by 1bp, the difference resulting from the insertion of an A.T pair at the centre of the sequence (Fig. 5k). R2 and R3 both differ from R1 at three positions and, with only one exception, these minor changes occur in the second half of the repeats. The first half of all three sequences is therefore conserved and may be generalised as 5'-TTGCTC-3' followed by a string of four or more T residues. R1 and R2 are separated by 20bp, of which 16 are A.T pairs. R3 includes the -35 region of the promoter, and between them R2 and R3 encompass the CRP binding site. Collectively, the repeat sequences span three out of

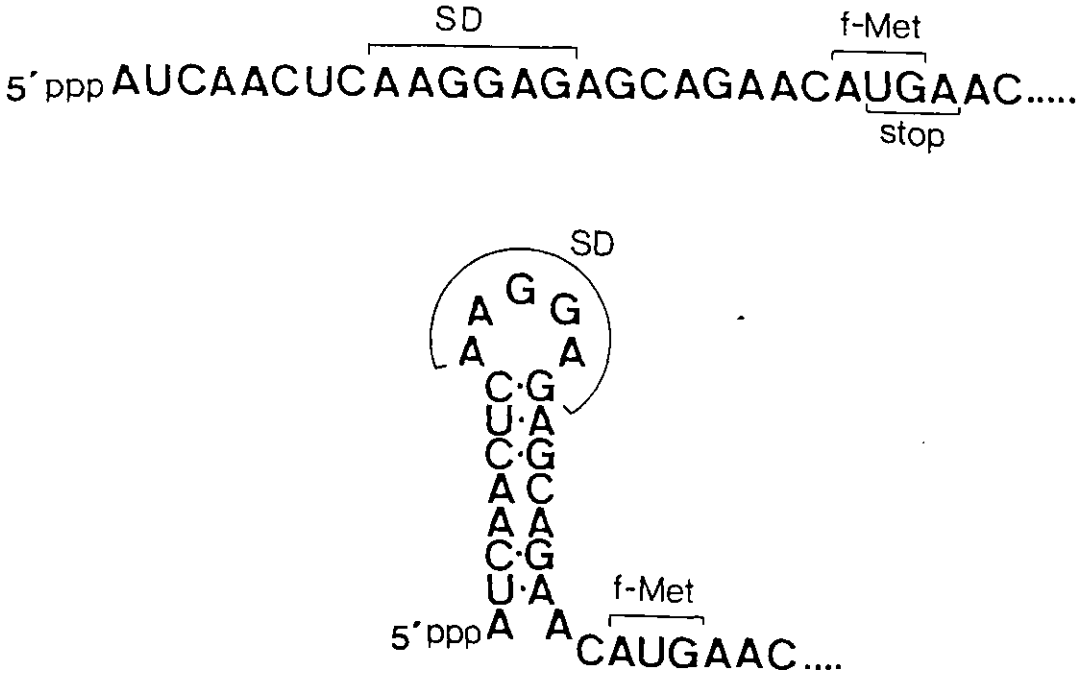


Fig. 5j A possible secondary structure at the 5' end of dalDK mRNA

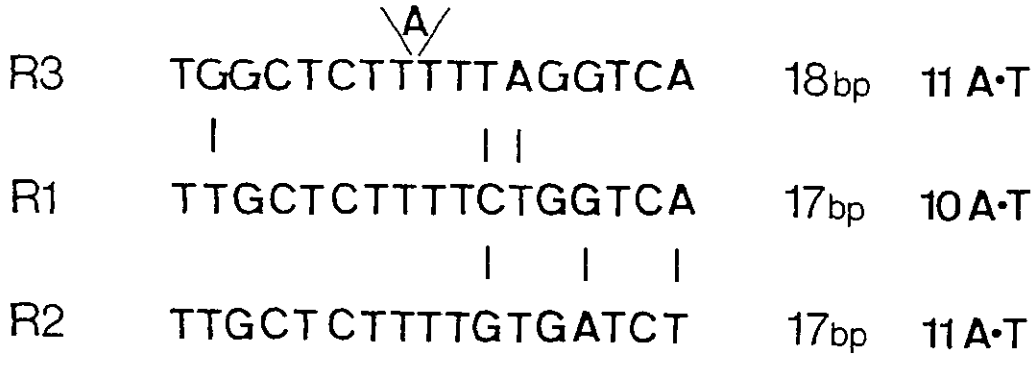


Fig. 5k A comparison of the three repeated sequences in the dalDK promoter. R1 and R2 are separated by 21bp, of which 17 are A+T. Both R2 and R3 differ from R1 at only 3 positions, excluding the extra A residue at the centre of R3. Note that TTGCTC followed by a string of T's is a common feature of each repeat.

the five regions previously put forward as possible RNA Polymerase recognition sites, and in each instance these sites fall within the first (conserved) half of the repeat (Fig. 51).

It is tempting to speculate on the function of these repeats, particularly with respect to the binding of a regulatory protein. Is there any evidence consistent with these sequences being the recognition sites for the dal repressor protein? The trp operon repressor binds to a site overlapping the promoter and blocks the binding of RNA polymerase (Squires et al, 1975). The operator, lying between -2 and -21bp, exhibits a pronounced two-fold symmetry, and the central core of 8bp (one helix turn) is essential for promoter function (Bennett et al, 1978). The λ repressor protein recognises six sites on the phage genome, three within the left operator region (O_L1, O_L2, O_L3) and three in the right operator (O_R1, O_R2, O_R3). Each binding site consists of 17bp with approximate two-fold symmetry. The sequences of all six sites are similar, but not identical (Maniatis et al, 1975a and 1975b; Ptashne et al, 1976). Binding is asymmetric within each repeat; the sequence 5'-CGGTG-3' in the right hand half is essential in this respect (Bailone and Galibert, 1980). Hence, the dalDK promoter repeats are similar in size to the trp and λ_{PL} + λ_{PR} operators, and, although they display no dyad symmetry, attention must be drawn to the A+T rich regions separating R1 and R2, and to the evidence of a conserved, possibly essential, sequence in the left half of each repeat. Four 19bp repeating units, again with an internal symmetry, are responsible for the binding of λ O protein at the phage replication origin (Tsurimoto and Matsubara, 1981), but symmetry need not be a strict requirement for all DNA-binding proteins. Murotsu et al (1981) have identified a 19bp

Fig. 51 Tandem repeats in the dalDK promoter. Three repeats (R1, R2 and R3) occur between HindB and the Pribnow box TACAGTG. These repeats overlap a number of the putative RNA polymerase recognition sequences including the -35 region (shaded areas). Asterisks mark three pentamer repeats (TAATT) and the large dyad symmetry is underlined.

sequence repeated nine times in the plasmid "mini-F" which has no rotational symmetry and is believed to represent the interaction site for the replication initiator protein π. As yet, the dal repressor protein has not been isolated, but when this is achieved the first priority will be to ascertain the true operator sequence. The DNA and protein sequences of the repressor are presented in Chapter 6, and a further consideration of its recognition sequence is included in the Discussion section in the light of fresh data.

III) The Intercistronic Region of the dal Operon.

The Primary and Secondary Structures of the dalD/dalK Junction

The nucleotide sequence spanning the dalD/dalK junction (Fig. 3h) displays a number of interesting features. Between BstC and the beginning of the DXK gene, there are numerous blocks of G.C pairs separated, in many instances, by shorter A+T rich spacers. There are no less than seven translation termination signals; 4 UGA and 3 UAA. The ArDH gene terminates with three stop codons in different reading frames. Only 12bp separates the C-terminal ArDH codon (Asn·454) from the DXK initiator AUG. At the mRNA level, seven out of these twelve bases show homology to the 16S rRNA 3' end. A second good ribosome binding site is positioned 70bp upstream and within the dalD gene itself. Translation of DXK messenger must occur either by readthrough from ArDH or via re-initiation at the S/D sequence between the two genes. Initiation of translation at the upstream S/D sequence would give rise to a highly charged, 12 amino acid peptide (Met-Ala-Arg^{+}-Glu^{-}-Ser-Arg^{+}-Arg^{+}-Arg^{+}-Leu-His^{+}-Ala-Asp^{-}), terminating one 1bp before the end of ArDH (see McConnell, 1979).

Secondary structures for the dalDK mRNA around the intercistronic region have been determined (Fig. 5m). Prominent are the two stable hairpin loops with 7 and 15bp stems (ΔG values of -10.5 and -21 kCal/Mol respectively). The smallest hairpin contains the first S/D sequence, and the unpaired loop of the larger structure bears a UGA stop codon overlapping an AUG triplet. Considering their position, the similarity of these structures to known transcription terminators and RNA processing sites may be significant. These ideas are pursued further in the final Discussion chapter.

The Intercistronic Region of the rbt Operon

The DNA sequence around the rbtD/rbtK junction has recently been determined (T. Loviny, unpublished data). Severe compressions in the sequence ladder have prevented accurate reading of bases beyond the limits shown (Fig. 5n). There are no homologies at the DNA level with the corresponding region of dalDK mRNA and the N-terminal kinase sequences are also different.

The intergenic region consists of 28bp and has stop codons in all three frames like its counterpart in dal. An 8bp sequence thought to constitute a ribosome-binding site is found immediately 5' of the DXK initiator AUG. A second possible S/D sequence is located 41bp upstream in the RDH coding sequence. The arrangement of translational initiation and termination signals is analogous to the dalD/dalK junction, but no RNA hairpins can be formed from any of the rbt sequences currently available. Note that if initiation of translation could begin at the upstream S/D sequence, the dipeptide Val-Asp$^-$ might be synthesised (McConnell, 1979).

Enzyme Levels : Evidence for Polarity in the Pentitol Operons.

Early work on the d-arabitol pathway of K.aerogenes W70

Fig. 5m Secondary structures for the mRNA around the dalD/dalK intercistronic region. All translation termination signals are bracketted. Arrows indicate the beginning and end of a 12 amino acid peptide which could be synthesised following an initiation event at the first SD sequence

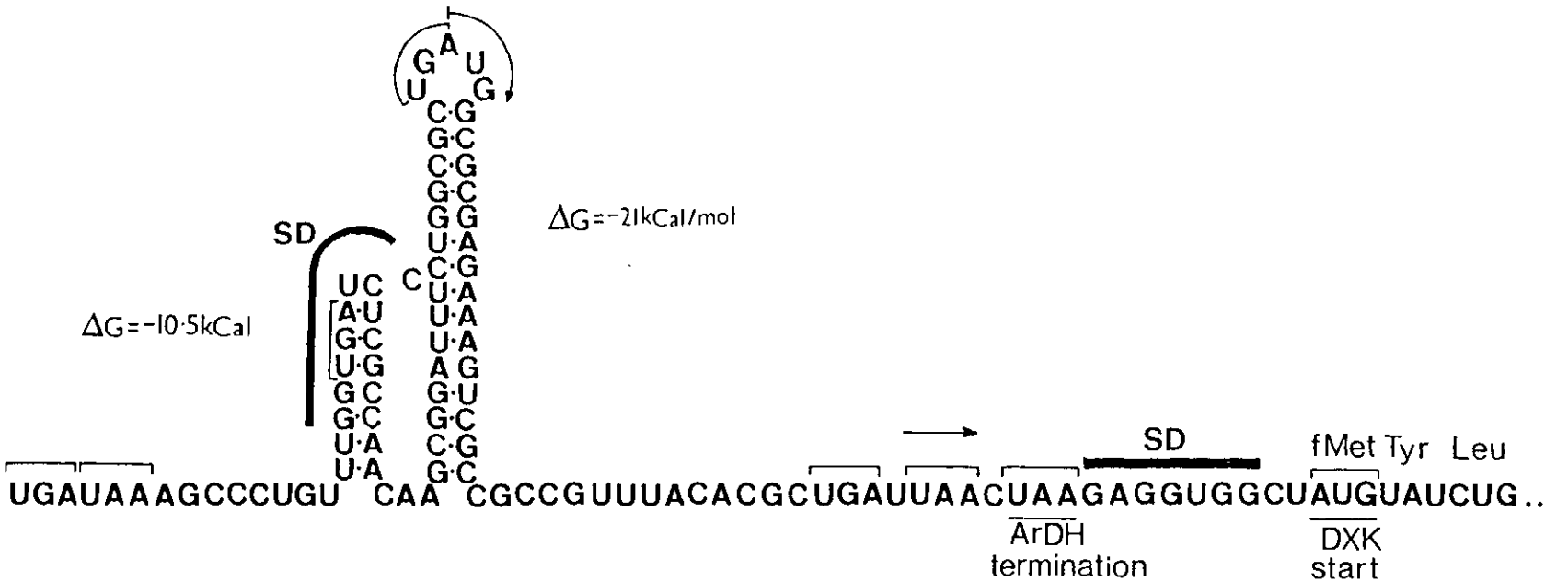
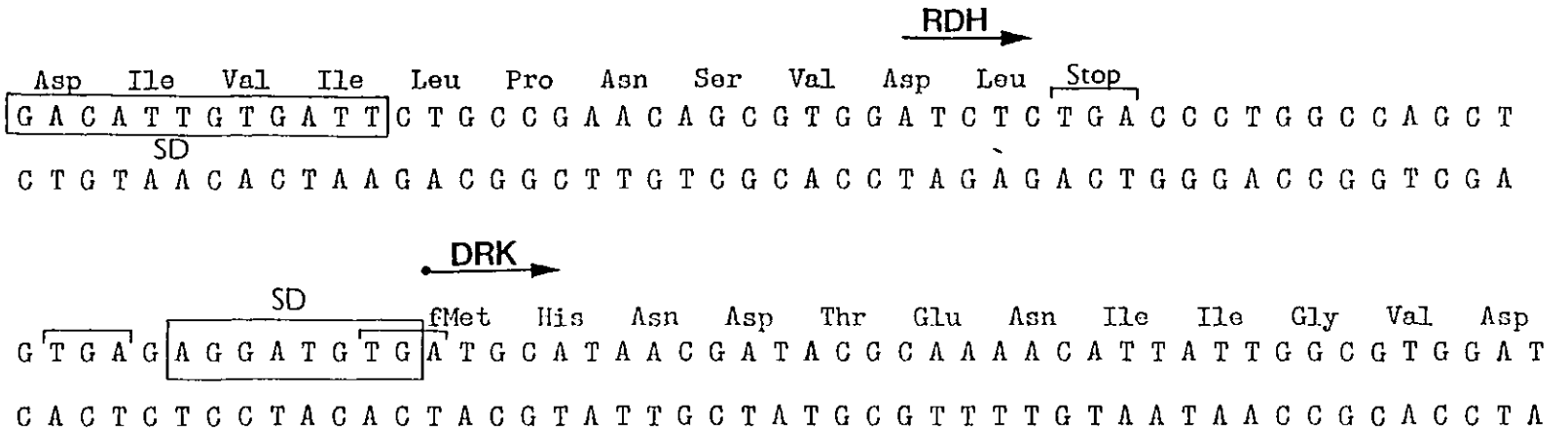


Fig. 5n The rbtD / rbtK intercistronic region



28bp separate Leu.247 of RDH from the N-terminal Met. of DRK . This region contains termination codons in all three reading frames and a possible ribosome attachment site immediately 5' of the rbtK initiator. A second ribosome binding site may be present near the end of the rbtD gene.

suggested that the dehydrogenase and kinase are not made in equimolar amounts (Charnetzky and Mortlock, 1974b). The activity of ArDH in cell-free extracts is, on average, twice that of DXK for d-arabitol induced cultures. Allowing for differences in the specific activities of the two enzymes (Neuberger et al, 1981), considerably more ArDH monomers must be made. However, DXK levels in such extracts may be slightly elevated by induction of the xylose operon during growth on d-arabitol through synthesis of the inducer, d-xylose, from d-xylulose by d-xylose isomerase (Neuberger et al, 1981). In the E.coli lac operon approximately 3-5 times more β-galactosidase monomers are made than transacetylase (Brown et al, 1967) as a result, it is thought, of the preferential attachment of ribosomes to 'z' message and the lability of 'a' message.

Enzyme assays on cell-free extracts of a number of dal constitutive and inducible strains suggest that the ratio of ArDH:DXK may vary substantially (Table 5i), a fact that is supported by SDS polyacrylamide gel electrophoresis of crude extracts (M. Neuberger, personal communication). If the same genes are present at the same dosage and under the control of the same promoter in two different organisms, and if the quantitities and ratios of the operon products are not equal, then this is evidence of species differences at the transcriptional or translational level (assuming that the steady state intracellular levels of inducer are similar also). Within the same organism enzyme ratios should be similar, unless selective pressure or mutation have brought about changes. The K.aerogenes strain A111, selected for fast growth on xylitol, and duplicated for the RDH gene, makes much less ArDH than any of the other strains tested. Prolonged selection for improved xylitol dehydrogenase activity, the use of UV mutagenesis and extended

Table 5i Levels of pentitol dehydrogenases and pentulokinases in various strains of E.coli and K.aerogenes.

Enzyme	Strain	Specific Activity (units/mg protein)
ArDH	FG5	0.624 (2)
	NC100	0.576 (3)
	A3	0.577 (3)
	A111	0.29 (3)
	PS640	0.611 (3)
DXK	FG5	7.5 (3)
	NC100	15.7 (2)
	A3	ND (3)
	A111	20.15 (2)
	PS640	22.04 (2)
RDH	FG5	0.33 (2)
	NC100	0.358 (3)
	A3	1.07 (2)
	A111	0.696 (2)
	PS640	0.902 (3)
DRK	FG5	0.172 (2)
	NC100	0.081 (2)
	A3	0.046 (2)
	A111	ND (2)
	PS640	0.024 (2)

50ml cultures were grown at 37°C(32°C for PS640).The culture medium was M9 + CAA and supplemented to 0.2% w/v with d-arabitol or ribitol for inducible strains.Cells were harvested in late log phase and cell-free extracts prepared as described in Materials and Methods.Protein concentrations were determined by the Coomassie method.ArDH and RDH were measured spectrophotometrically but the [^{3}H] assay was used for DXK and DRK (see Materials and Methods).The DXK figures are in arbitrary units since no pure enzyme was available to standardise the assay.All values are averages of several independent assays(bracketted figures).

Table 5ii The ratios of the specific activities of pentitol catabolic enzymes in cell-free extracts of E.coli and K.aerogenes strains.

Strain	ArDH/RDH	ArDH/DXK	RDH/DRK
FG5	5.7	0.25	0.27
NC100	4.8	0.1	0.1
A3	1.6	-	3.3
A111	1.3	0.05	-
PS640	2.0	0.1	5.36

The enzyme levels presented in Table 5i have been corrected to allow for differences in the specific activities of the pure enzymes, namely:

ArDH $V_f^{arabitol}$ = 55μMol/min/mg

RDH $V_f^{ribitol}$ = 166μMol/min/mg

DXK $V_f^{xylulose}$ = 150μMol/min/mg

DRK $V_f^{ribulose}$ = 71μMol/min/mg

Genotypes:

K.aerogenes	FG5	rbt^+ dal^+
	A3	rbtC-101 $(rbtD)_n$ dal^+
	A111	rbtC-101 $(rbtD)_n$ dal^+
E.coli	PS640	rbtC-101 dal^+
	NC100	rbtC-101 dal^+

redundancy of the dal operon may have led to changes which have reduced the efficiency of dal expression or allowed deleterious lesions to accumulate in the ArDH structural gene itself.

With K.aerogenes FG5, the d-arabitol inducible levels of ArDH are comparable to those of E.coli NC100, but less DXK is made by the former. Assuming the kinases to be identical in each case, this phenomenon may be indicative of variations in the relative frequency of transcription or translation of the ArDH and DXK genes in these species, but could equally be due to differences in the induction of the xyl operon DXK. As there is no significant difference in the levels of ArDH (or RDH) synthesised by FG5 and NC100, the E.coli machinery must adequately transcribe and translate the Klebsiella DNA. Hence, the somewhat poor -35 region of the dalDK promoter does not apparently reflect any altered specificity by K.aerogenes RNA Polymerase.

As expected, strains duplicated for the rbt operon make more RDH, but the specific activity remains lower than that of ArDH (Table 5ii). PS640, an E.coli λp rbt dal lysogen carrying a constitutive rbt operon, synthesises 3 times more RDH than the ribitol-induced FG5, and also higher than the rbt^{c} NC100, and could possibly be a polylysogen.

The data gathered by these experiments is far from conclusive, and much more work is obviously required in this area. Bahramian and Hartley (1980) have observed an apparent switch from transcriptional to translational control of rbt expression during late log phase, and similar time-course experiments using dal^{+} K.aerogenes and E.coli strains could yield valuable information about the control mechanisms operating here.

To eliminate interference by non-arabitol induced DXK, a

xyl$^-$ host ought to be used in determining fresh values for ArDH/DXK ratios. Plasmid pRD351 has now been transformed into R8 (a *xyl*$^-$ *E.coli*) for this purpose. In addition, a comparison of ArDH and DXK levels with those of a *rho*$^-$ strain harbouring the same plasmid should reveal whether rho-dependent termination is involved in *dal* expression.

CHAPTER 6

CLONING AND SEQUENCING OF THE d-ARABITOL OPERON REPRESSOR

When Charnetzky and Mortlock (1974c) performed classical genetic mapping of the pentitol operons they concluded that genes in the central control region were interdigitated, (i.e. rbtD-rbtO-dalR-rbtR-dalO-dalD) with dalR lying closer to rbtDK o/p than to dalDK o/p. This arrangement is not consistent with the mirror symmetry of the remainder of the operons and is somewhat harder to explain in terms of gene duplications. Subsequent work in this laboratory (C. Smith, unpublished data) has shown their genetic map to be incorrect, the revised gene order being rbtK-rbtD-rbto/p-rbtR-dalR-dalo/p-dalD-dalK. In this chapter I will discuss briefly the work involved in localising the dal repressor gene and demonstrating its expression in vivo, and then go on to describe the cloning and sequencing of a 1.2kb restriction fragment which carries the repressor gene. The full nucleotide sequence and resultant predicted amino acid sequence of the dalR gene product is presented, and the protein is compared with a number of other repressors from E.coli and the phages P_{22} and λ.

Localisation of the dalR coding sequence

The bulk of this work has been performed by Chris Smith and is presented here to provide the necessary background to my own studies. Subfragments from the SalC/BstB region of λ prbt dal were cloned into pBR322 and transformed into the E.coli host SK1592. Recombinant plasmids were characterised by restriction digests using Bst I, Hind III or Pst I (Fig. 6a). This small "library" of plasmids was used to transform the d-arabitol constitutive E.coli SC24 which was constructed by transducing the mtl^A mtl^D host WI485 to rbt^c dal^+ and then

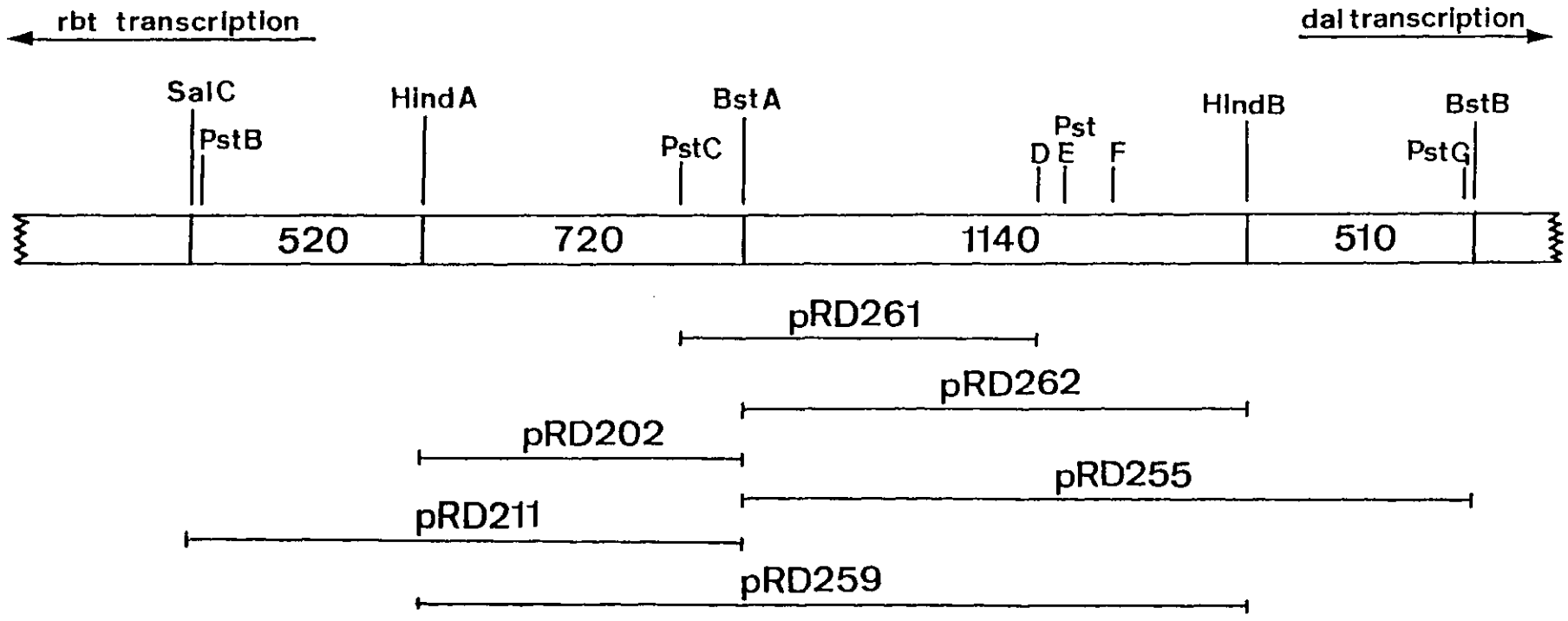


Fig. 6a Sub-clones of the rbtdal control region

Table 6i *In vivo* assays for *dal* repressor activity

Host strain	Plasmid	Growth on mannitol	ArDH (U/mg)	RDH (U/mg)
SC20	-	-	0	1.7
SC24	-	+	0.89	1.8
SC24	pBR322	+	0.90	1.5
SC24	pRD202	+	1.10	1.7
SC24	pRD211	+	0.87	2.1
SC24	pRD255	-	0	1.8
SC24	pRD259	-	0	1.3
SC24	pRD262	-	0	1.6
SC24	pRD261	+	0.89	1.9

'0' indicates an activity of less than 0.05U/mg protein

SC20 = F^- SupE *mtl*A *mtl*D *rbt*^{c} *dal*$^+$; a derivative of W1485

SC24 = F^- SupE *mtl*A *mtl*D *rbt*^{c} *dal*^{c} ; spontaneous from SC20

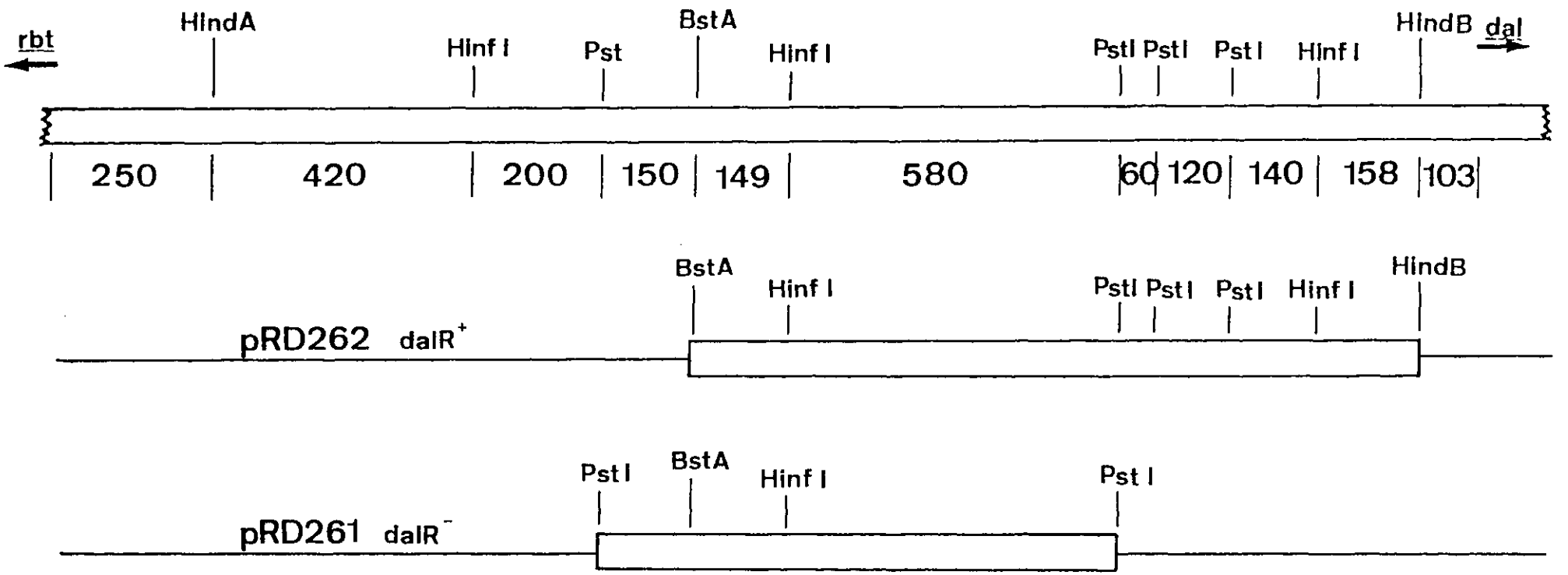
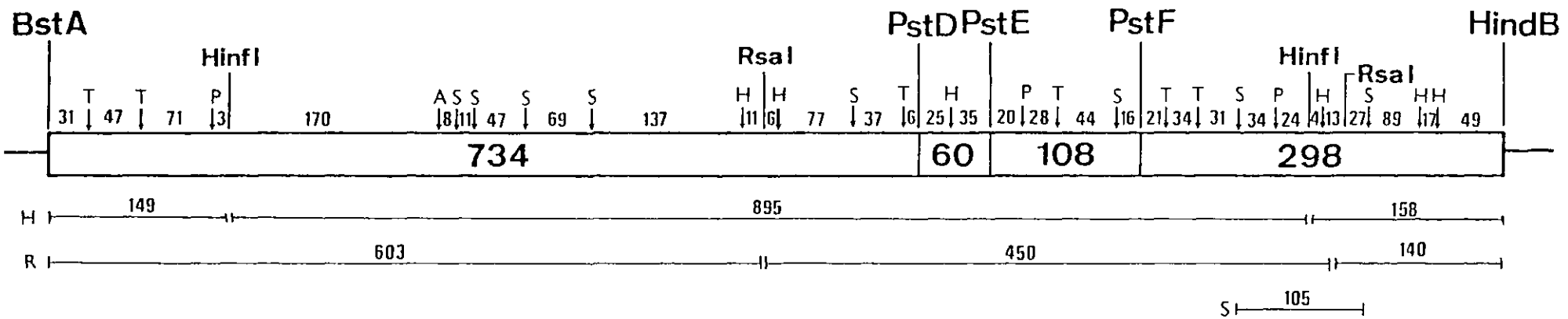


Fig. 6b Sub-clones of the dalR gene

Fig. 6c Restriction map of the BstA / HindB fragment carried by pRD262



P = PvuII+AluI H = HaeIII T = TaqI S = Sau3AI A = AluI R = RsaI

selecting for spontaneous mutants capable of growth on mannitol, based on the fact that ArDH will oxidise mannitol, but that this substrate cannot induce the dal operon. Strains with plasmids encoding an intact dal repressor were identified by their mtl$^-$ phenotype and by the abolition of ArDH activity in SC24 crude extracts (Table 6i). Three plasmids, pRD255, pRD259 and pRD262 were found to repress ArDH to background levels. None of these had any effect on constitutive RDH synthesis by this strain. The BstA/HindB segment contains all the information necessary for the synthesis of an active dal repressor protein and the gene straddles the PstD site (Fig. 6b) since pRD261 is dalR$^-$. This being the case, then rbtR must be situated wholly or in part to the left of Bst A, since there is insufficient coding sequence between PstF and HindB. The mapping experiments of Charnetzky and Mortlock which place rbtR between dalR and dalDK o/p are therefore at fault. It is not immediately clear whether, in these plasmids, transcription of the dal repressor is occurring from its own promoter or whether expression arises as a consequence of readthrough from the pBR322 Tc promoter upstream. pRD255, pRD259 and pRD262 all carry the HindB/BstA fragment in the same orientation, HindB being closest to the vector EcoRI site. However, pRD259 has an additional 0.56kb of the dalD gene between the Tc promoter and the dal repressor-carrying fragment and is still dalR$^+$. The amount of DNA available (1200bp, maximum) sets an upper size limit of about 40,000 daltons for the dalR gene product.

Restriction mapping of the plasmid pRD262

To facilitate cloning, a partial restriction map was compiled for the BstA/HindB fragment (Fig. 6c). The original mapping for Pst I, Taq I and Alu I performed by Chris Smith was repeated, and several extra Taq I sites were discovered. The

sites for the enzymes Pst I and Alu I were confirmed as being correct. In addition, sites were located for Pvu II, Rsa I and Hinf I. Details of some of the mapping procedures are to be found in the legends to Photos. 6A and 6B.

Cloning fragments from the dalR gene into M13 vectors

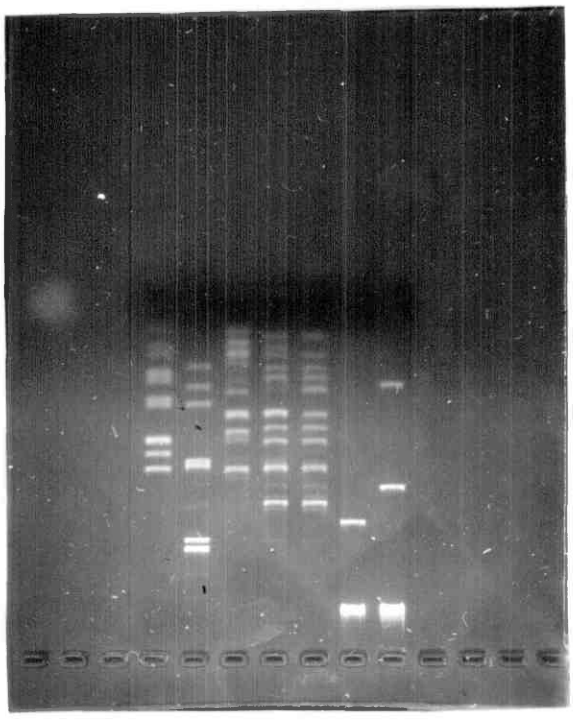
A) The plasmid pRD262 was also used as a source of DNA for cloning into the directional vectors M13 mp8 and mp9 (Messing and Veira, unpublished data). 5µg of plasmid was cut with Pst I, Hind III and Bam HI simultaneously in 20µl of TMN + 50mM NaCl. 0.2µg of each vector RF was digested with combinations of Pst I, Bst I and Hind III. Ligations were performed at 15°C for five hours using 1ng of vector, a 5-10 fold excess of DNA fragments and 0.1 Weiss units of T4 DNA Ligase, as described in Materials and Methods.

Similarly, Rsa I and Rsa I + Bst I fragments from the same plasmid were cloned into mp8/Sma I and Sma I + Bst I-cut mp8 and mp9. Ligations were left overnight at 15°C with 1.0 Weiss unit of ligase. Transformation and plating procedures were standard. From each of the eight separate ligation experiments twelve recombinants were selected and template DNA prepared.

B) Alu I, Hpa II and Sau3A I subclones of the repressor region were constructed using the M13 phage vector mp7.3. About 10µg of pRD262 was digested with Hind III and Bst I and the 1.2kb fragment carrying the gene was isolated from a 1.2% LGT agarose gel. Following electroelution, the DNA was further purified by DE52 chromatography and ethanol precipitation. The fragment was divided and digested with either Sau3A I, Alu I or Hpa II, and the resultant pieces ligated into Bst I, Hinc II or Acc I - cleaved Mp7 RF respectively. Some batches of commercially

Photograph 6A

Restriction digests of plasmid pRD262



1 2 3 4 5 6 7

1) Taq I + Hinf I
2) Taq I
3) Alu I + Hinf I
4) Alu I
5) Alu I + Pst I
6) Pst I
7) Pst I + Hind III

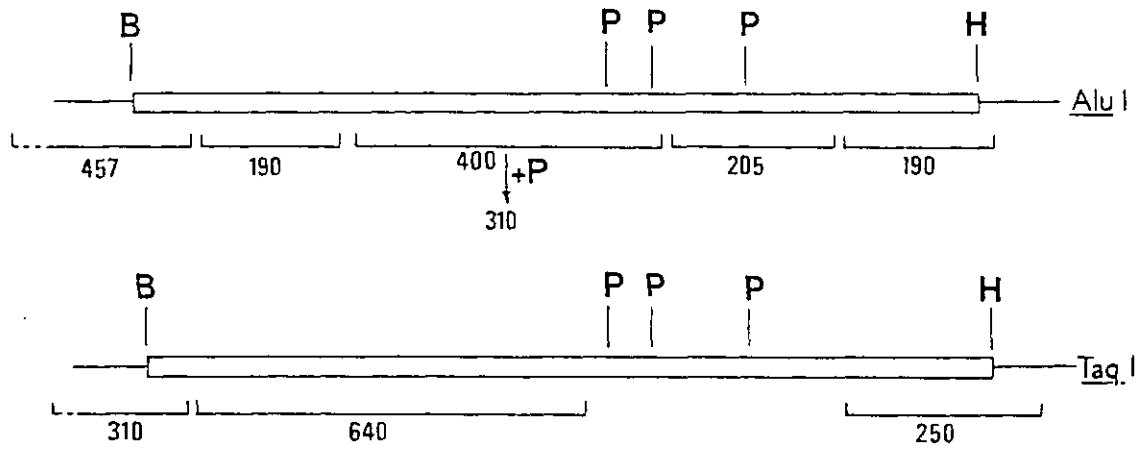
Digests were carried out in 20µl of medium-salt TMN at 37°C for 30 minutes using 1µg pRD262 and 2U of each enzyme. Reactions were stopped by adding 0.1 volumes of BPB/EDTA/glycerol mixture and a 10 minute incubation at 65°C. Electrophoresis was performed on a 1.5% agarose gel, run submerged at 100V until the tracking dye had migrated about 10cm.

Photograph 6A

Restriction digests of pRD262 : Fragment sizes.

Tao I	Alu I	Pst I
1450	910	1008
1300	659	
640	520	
615	457	Pst I + Hind III
370	400 (d)	760
310 (d)	281	290
250	257	
	226	
	205	
	190 (d)	

Fragments smaller than 150bp are not resolved. Those arising wholly or partly from BstA/HindB are underlined. Doublets are marked (d).

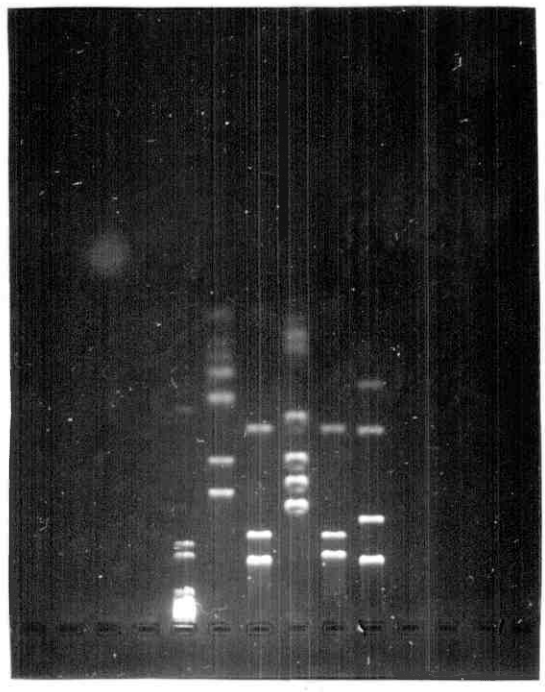


B = Bst I , P = Pst I , H = Hind III

For a complete restriction map see Fig. 6c

Photograph 6B

Mapping pRD262 for Pvu II and Rsa I



1) λDNA + Hind III
2) Hinf I
3) Pvu II
4) ØX174 + Hae III
5) Pvu II + Hind III
6) Rsa I

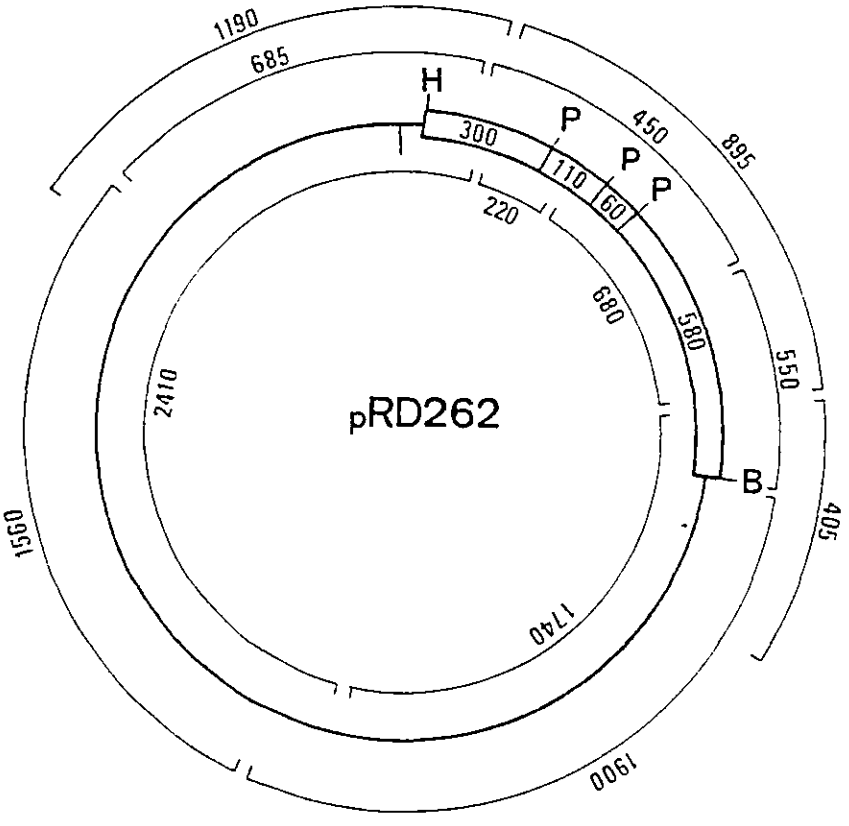
Digests were carried out using 0.5μg pRD262 and 2U of restriction enzyme in 20μl of medium-salt TMN at 37°C for 30 minutes. Separation was achieved on a 1.5% agarose gel (submerged) at 6mA for 18 hours. The major fragment sizes are listed below.

Hinf I	Pvu II	Rsa I	Pvu II/Hind III
1190	2400	2460	2300
900	1700	1560	1850
516/506	680	685	680
405	220	450	210
396			180
344			
298			
221/220			
154			

(continued..)

Photograph 6B contd.

Circular restriction map of pRD262



Fragment sizes are given in bp.

Outer fragment sizes are Hinf I, followed by Rsa I / Bst I. The innermost circle gives the Pvu II fragments.

available Hinc II contain contaminating nucleases which cause a highish background of white plaques upon self-ligation of the RF, but this preparation generated only 4-5 whites per 20 ng DNA. Single strand phage templates were prepared from 24 Sau3A I clones, 57 Alu I clones and 35 Hpa II clones. Electrophoresis was used to identify those phage having inserts whose size was inconsistent with them originating from BstA/HindB - a useful preliminary, since this particular vector preparation contained traces of chromosomal DNA and sometimes tended to give irrelevant sequences. ddT-tracking allowed elimination of identical DNAs prior to sequencing (Photos 6C and 6D).

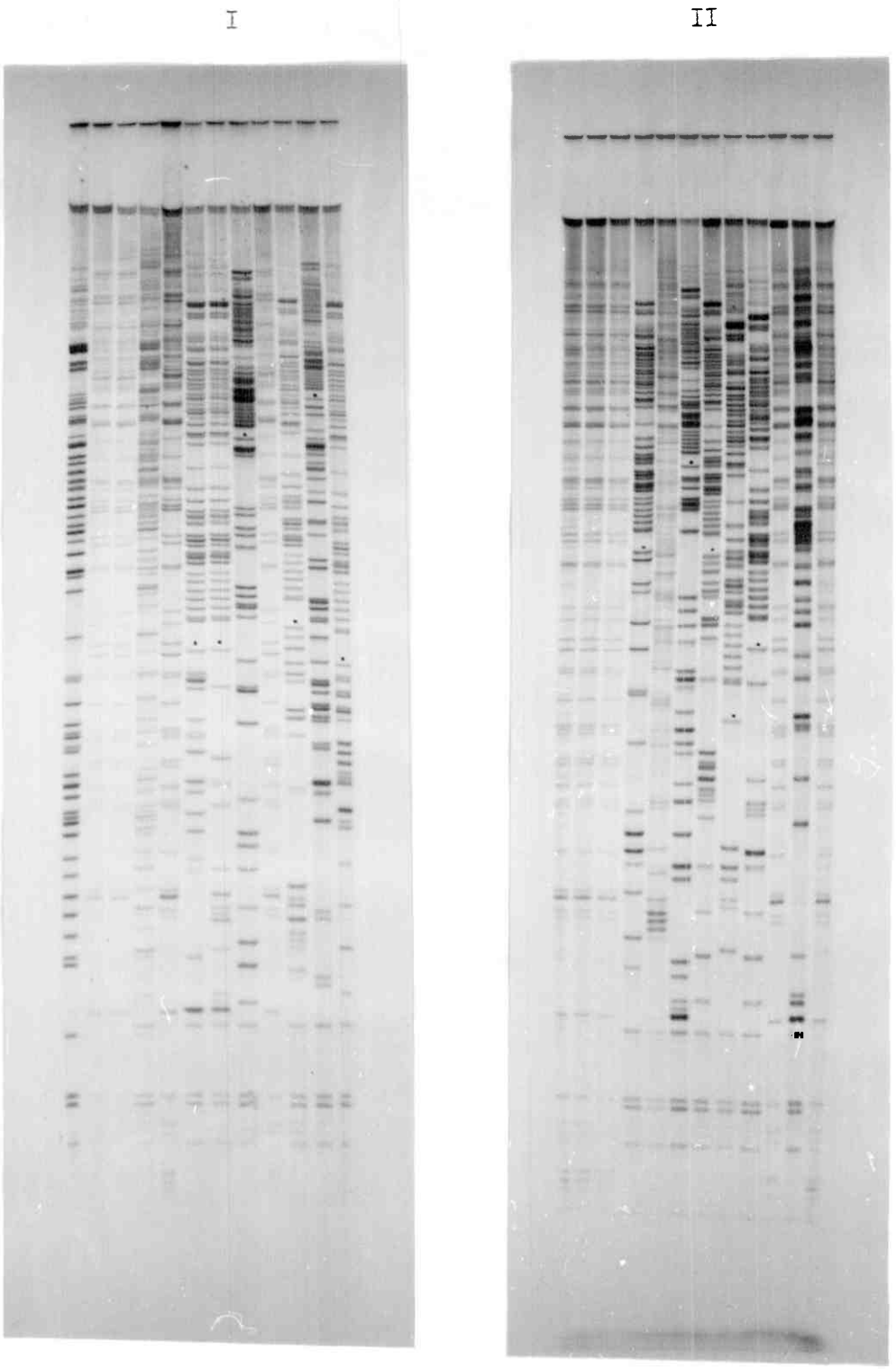
The vast majority of Hpa II clones proved virtually useless since they contained only very small inserts (5-20bp) due to the high frequency with which this enzyme cuts G+C rich K.aerogenes DNA and the preference of M13 for small inserts (Gardner et al, 1981).

Sequencing Strategy : Characterisation of M13 Clones

The major clones used to derive the sequence of the BstA/HindB region are shown in Fig. 6d. Individual clones were identified through overlaps with known sequences near to BstA (Loviny et al, 1981) or HindB (see Figure 5a) and by using the DNA/DNA hybridisation techniques of Herrman et al (1980).

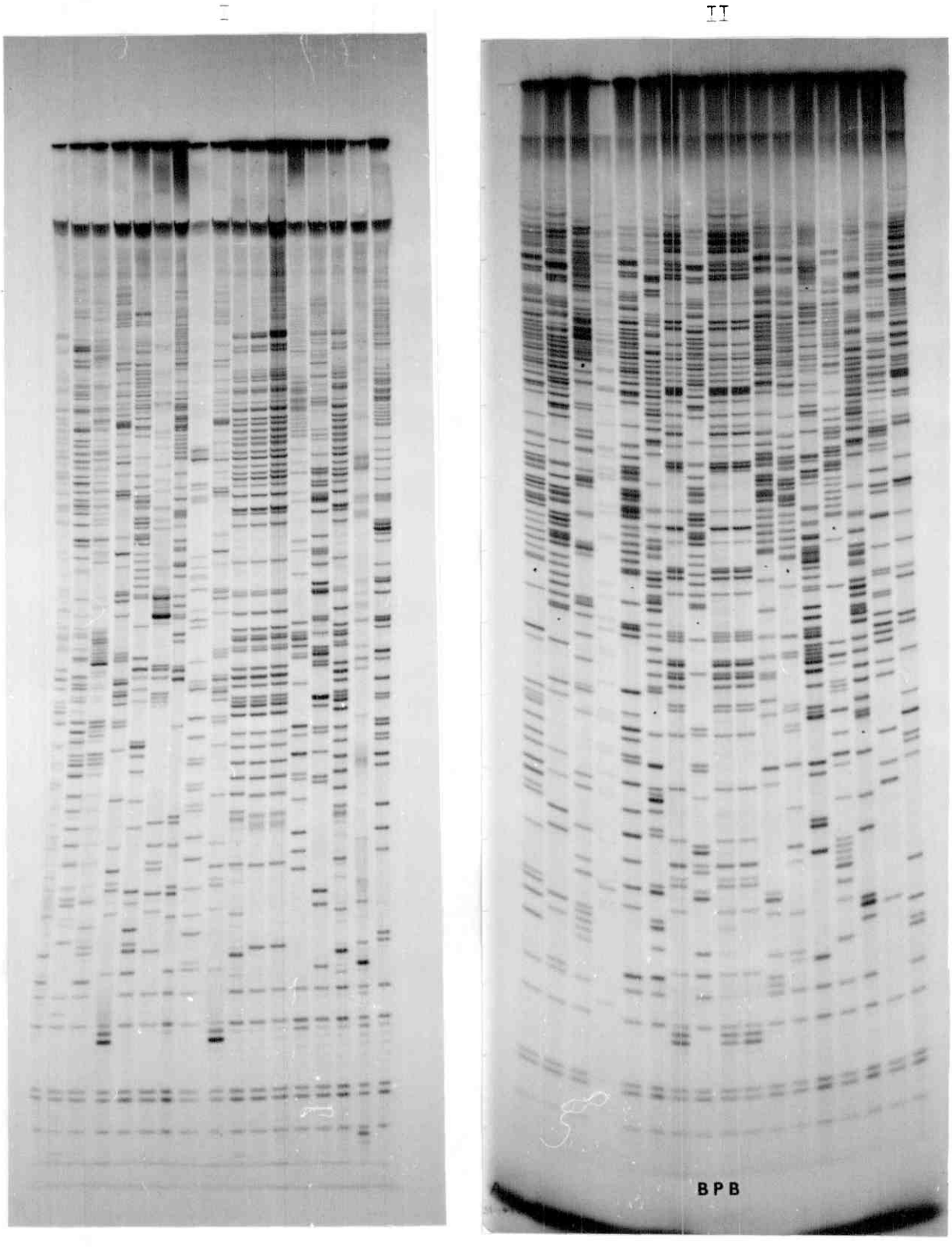
Among the 12 Pst/Bst mp8 clones screened, 6 were found to contain PstD/BstA and a further 3 had, in addition, either a 60bp or 108bp Pst I fragment situated before PstD. These hybrids presumably resulted from concatemerisation of the donor DNA during ligation. There appears to be a preferential arrangement of the two small Pst I fragments : PstD/E is always found in the same orientation as in clone 2A (Fig. 6d) and PstE/F is always in the reverse orientation to clone 1D.

Photograph 60



ddT screening of M13 clones containing Sau3A I fragments from the HindB/BstA region. I) Clones a → l, II) clones m → x. Separation was for 50 minutes at 2000V and 60°C on 200μm, 6% acrylamide sequencing gels.

Photograph 6D



ddT screening of M13 Alu I (I) and Hpa II (II) clones from the Hind3/BstA fragment. Separation was at 2000V for 90 minutes on 6% acrylamide sequence gels.

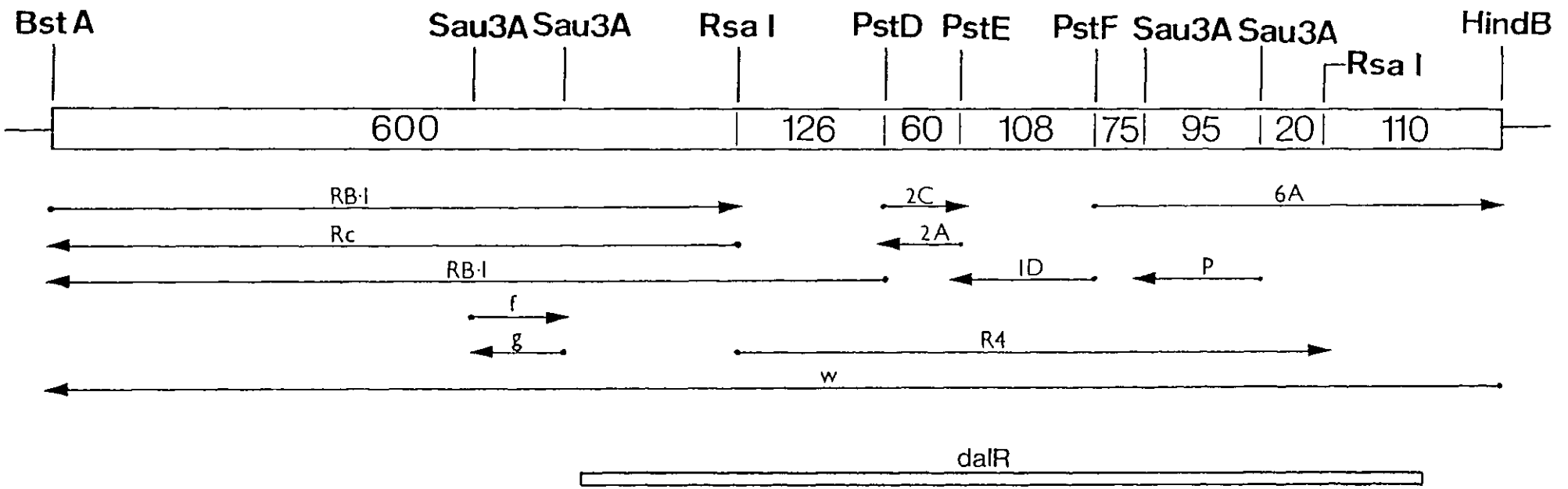


Fig. 6d The major clones used to derive the DNA sequence of the dalR gene. Clone names are referred to in the text. The extent of dalR coding sequence is shown.

Recombinant phage carrying the PstF/HindB region were located by matching DNA sequences with those obtained earlier in this work using the Maxam and Gilbert method. Out of 6 clones screened, 4 contained this fragment, and the remainder gave a sequence recognisable as belonging to the vector pBR322. The Sau3A I clone 'p' was found to hybridise to templates bearing PstF/HindB (Photo 6E(i)).

The 450bp long Rsa I fragment was detected in clone R4, which hybridised to clones 2C and 6A, but not to PB·1, 1D or p (see Fig. 6d and Photo 6E(ii)). The BstA/Rsa I region was picked up by ddT-tracking in 2 out of 12 clones selected for screening, namely RB·1 and RB·6. Once identified, these phage were used to probe Rsa/Bst mp9 clones for the complementary sequence - positive results were obtained for the clones Rc and Re. PstD/BstA was found in 6 out of 12 mp8 Pst/Bst clones, also by hybridisation to RB·1 (Photo 6E(ii)). The small internal fragments PstD/E and PstE/F were present in the vast majority of phage arising from a simple mp8/Pst I cloning, where only large plaques were selected from the plates.

All sequences were determined by the Sanger dideoxy method using a universal 17mer primer as described in Materials and Methods. Each sequence was derived several times from different clones and wherever possible both strands were sequenced. No severe problems attributable to compressions were encountered when reading the autoradiographs (Photo 6F). The use of thermostatted gels and formamide/urea gels removed any minor ambiguities.

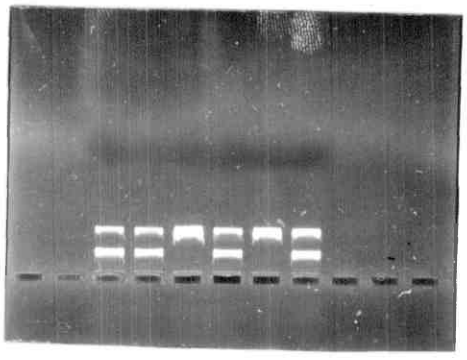
The nucleotide sequence of a 1.2kb fragment encoding the d-arabitol repressor

The complete nucleotide sequence of the BstA/HindB region

Photograph 6E

Hybridisation of M13 ssDNA to detect cloned complementary strands

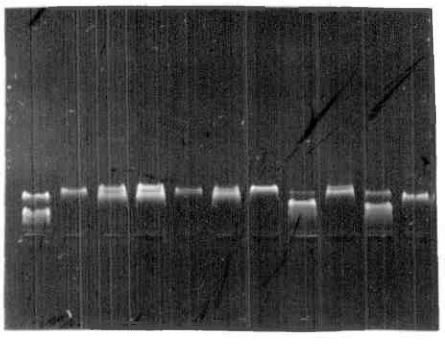
(i)



1 2 3 4 5 6

1% agarose minigel showing Sau3A I clone 'p' hybridised with six PstF/HindB clones (6A→6F, left to right). Clones 6A,B,D and F all hybridise to 'p' .

(ii)



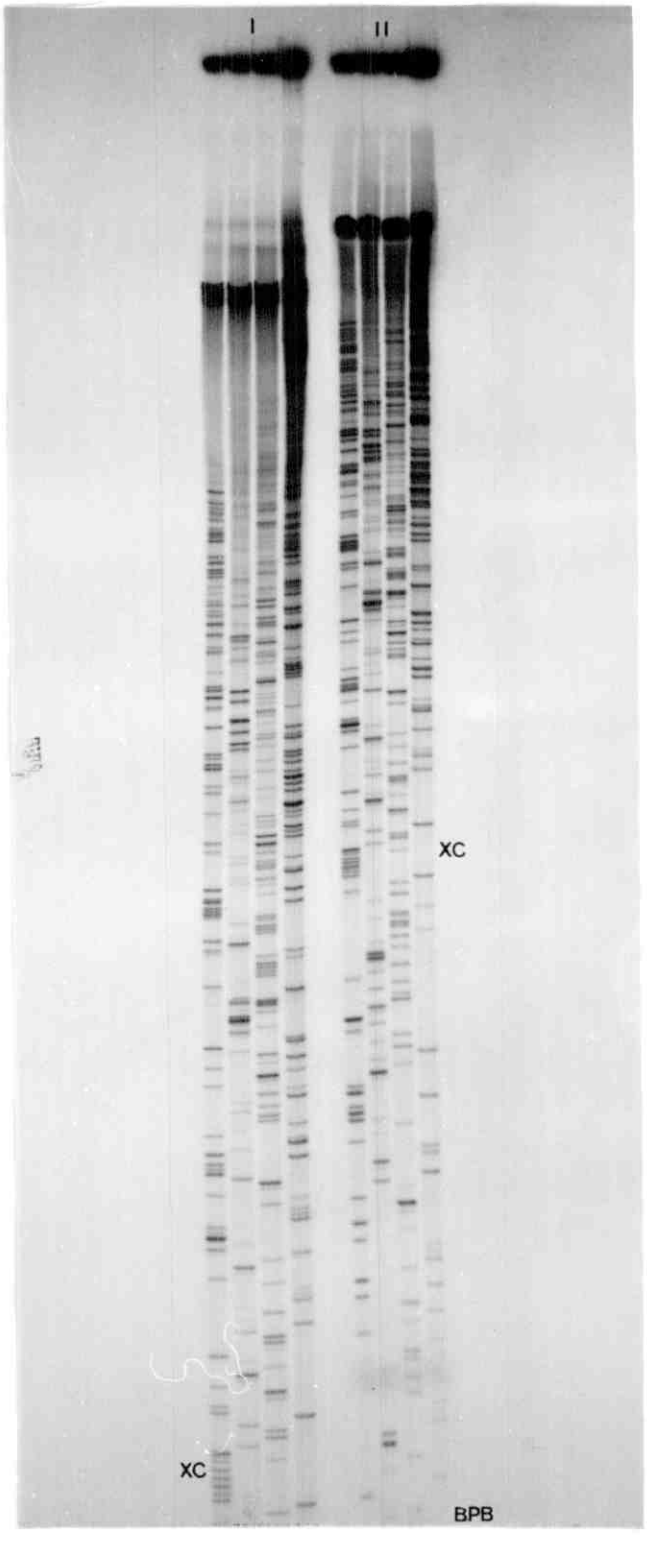
1 2 3 4 5 6 7 8 9 10 11

1% agarose minigel

1) 6B + R4
2) 6B
3) R4 + 2A
4) R4 + 1D
5) 'p'
6) R4 + 'p'
7) RB1
8) RB1 + PB1
9) PB1
10) PB1 + RB6
11) RB6

Hybridisations were comprised of 1µl of each DNA, 4µl 10x Hin buffer, 3µl dH_2O and 1µl of BPB/SDS/glycerol. The mixture was heated to 100°C for 3 minutes, transfered to 80°C for 10 minutes then allowed to cool to room temperature over 30 minutes before loading directly on to a gel.

Photograph 6F



A 6% acrylamide sequence gel of M13 clone Rc which contains the Rsa I/BstA region. Two separate loadings have been made and the order (left to right) is TCGA. The positions of the two marker dyes are shown. Electrophoresis was carried out at 1500V for 3 hours.

has been determined (Fig. 6e) and spans 1202bp of which 58% are G+C pairs - slightly above the average for the K.aerogenes genome. An analysis of the sequence using the TRNTRP programs of Staden (1977) reveals only one open reading frame of significant length, and since this is present on the opposite strand to the ArDH and DXK genes, bipolar transcription within the dal operon is implied. The 615bp open frame extends from an ATG triplet positioned 88bp from HindB. The translation termination codon is followed immediately by a string of five T residues and is preceded by a G+C rich region capable of folding into a stable RNA hairpin, a typical transcription termination signal (Fig. 6f). Beyond this point all three frames are closed at regular intervals, and codon usage is atypical, suggesting that this region is untranscribed. There is sufficient coding capacity for a protein of 205 amino acids (Fig. 6g), but for reasons which will be discussed later it is believed that translation of the messenger may be initiated at a second in-phase methionine 140bp from HindB, resulting in a slightly smaller protein of 188 residues or approximately 21,000 daltons. The gene is preceded by a promoter-like sequence (Fig. 6h), whose role in the regulatic of dalR expression is considered in detail in the following chapter. At first sight the dalR promoter appears to have a very good -35 region and Pribnow box, although overall it contains a rather high proportion of G+C pairs that may lower its efficiency.

The size of the dal operon repressor is of the same order as many of the regulatory proteins sequenced to date. The phage P22 c2 protein, 216 residues (Sauer et al, 1981), TnpR, 185 residues (Chou et al, 1979) and LexA from E.coli,

Fig. 6e DNA sequence of the 1.2kb HindB/BstA fragment encoding the d-arabitol repressor protein.

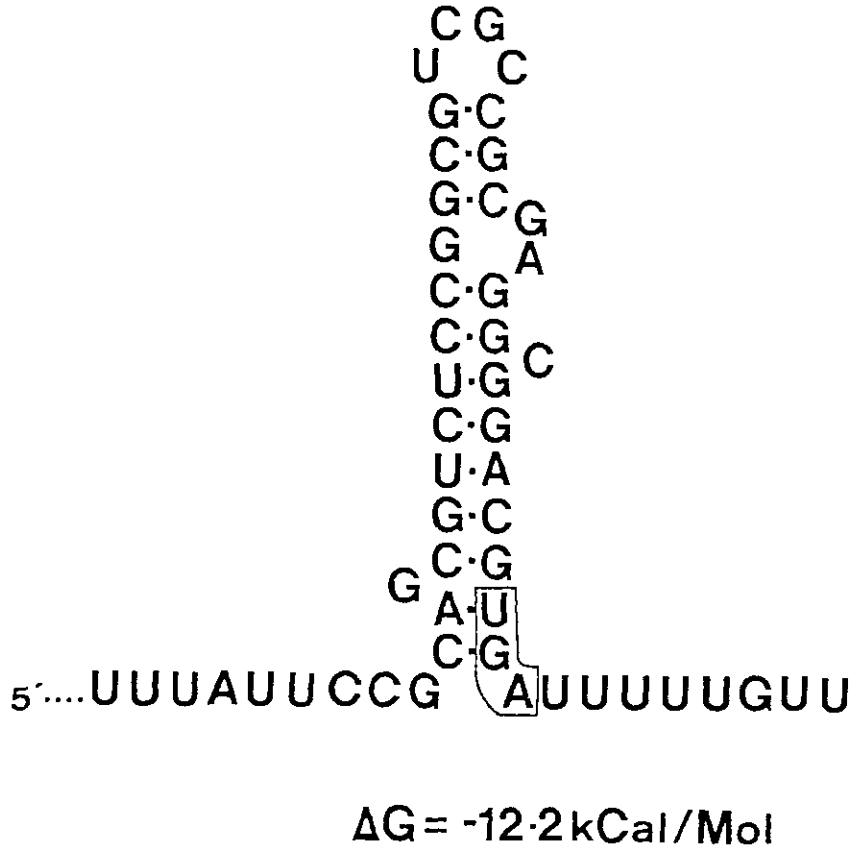


Fig. 6f A possible secondary structure for the 3' end of the dalR mRNA. The G+C rich hairpin and string of Uridines constitute a transcription termination signal.

```
  1
MET SER LYS GLU ASP ASP ILE ARG LEU ASP GLN LYS VAL ARG ALA ALA TRP MET TYR TYR
 M   S   K   E   D   D   I   R   L   D   Q   K   V   R   A   A   W   M   Y   Y

 21
ILE ALA GLY GLN ASN GLN SER GLU ILE ALA SER GLN LEU GLY THR SER ARG PRO VAL VAL
 I   A   G   Q   N   Q   S   E   I   A   S   Q   L   G   T   S   R   P   V   V

 41
GLN ARG LEU ILE ALA ALA ALA LYS GLU GLU GLY ILE VAL SER ILE ASN LEU HIS HIS PRO
 Q   R   L   I   A   A   A   K   E   E   G   I   V   S   I   N   L   H   H   P

 61
VAL ALA ASN CYS LEU ASP TYR ALA GLN LEU LEU GLN GLU LYS TYR GLY LEU ILE GLU CYS
 V   A   N   C   L   D   Y   A   Q   L   L   Q   E   K   Y   G   L   I   E   C

 81
ASN VAL VAL PRO ALA PHE SER GLU GLU SER THR LEU ASP SER VAL SER PHE GLY CYS TYR
 N   V   V   P   A   F   S   E   E   S   T   L   D   S   V   S   F   G   C   Y

101
GLN LEU MET ALA ARG TYR LEU GLN ASP ASP LYS GLU LYS ILE ILE CYS LEU GLY SER GLY
 Q   L   M   A   R   Y   L   Q   D   D   K   E   K   I   I   C   L   G   S   G

121
LEU THR LEU LYS LYS ALA LEU GLN ARG ILE ASP PHE ASP SER LEU ASN THR ARG CYS VAL
 L   T   L   K   K   A   L   Q   R   I   D   F   D   S   L   N   T   R   C   V

141
ALA LEU ILE SER ALA MET ASN ALA ASP GLY GLN CYS ASN TYR TYR ASP ASP VAL PRO LEU
 A   L   I   S   A   M   N   A   D   G   Q   C   N   Y   Y   D   D   V   P   L

161
LEU LEU THR ARG LYS ILE LYS ALA LYS TYR TYR GLN TRP PRO ALA PRO ARG TYR ALA GLN
 L   L   T   R   K   I   K   A   K   Y   Y   Q   W   P   A   P   R   Y   A   Q

181
SER ALA ASP GLU TYR GLU MET TRP CYS THR ASN ARG LEU PHE ARG SER VAL SER GLY VAL
 S   A   D   E   Y   E   M   W   C   T   N   R   L   F   R   S   V   S   G   V

201
ALA ALA ARG ARG THR
 A   A   R   R   T
```

Fig. 6g The amino acid sequence of the d-arabitol repressor. The positions of two alternative translation start points are indicated.

Fig. 6h The dalR structural gene and promoter region

```
  3
GCT TGC CAG TTA TGC CAA CCA CGG CGA GAA TGC GGG TAT TCG CCG GGC CGG TTG CGT GCC
    -35                              -10          +1

 63
                                        MET SER LYS GLU ASP ASP ILE ARG LEU ASP GLN
CTG GCC AGT AAC TTA GGG CGG GAA ACC ATG AGT AAA GAA GAC GAT ATC CGG TTG GAT CAG

123
LYS VAL ARG ALA ALA TRP MET TYR TYR ILE ALA GLY GLN ASN GLN SER GLU ILE ALA SER
AAG GTG CGT GCC GCA TGG ATG TAC TAC ATC GCC GGC CAG AAT CAG AGC GAG ATT GCC AGC

183
GLN LEU GLY THR SER ARG PRO VAL VAL GLN ARG LEU ILE ALA ALA ALA LYS GLU GLU GLY
CAG CTG GGC ACC TCC AGA CCG GTG GTG CAA CGG CTG ATC GCC GCC GCG AAA GAA GAA GGG

243
ILE VAL SER ILE ASN LEU HIS HIS PRO VAL ALA ASN CYS LEU ASP TYR ALA GLN LEU LEU
ATT GTG TCG ATT AAT CTG CAC CAT CCG GTA GCG AAC TGC CTC GAT TAT GCG CAG TTG CTG

303
GLN GLU LYS TYR GLY LEU ILE GLU CYS ASN VAL VAL PRO ALA PHE SER GLU GLU SER THR
CAG GAA AAA TAC GGC TTG ATC GAG TGC AAT GTG GTC CCC GCC TTT AGC GAA GAA AGC ACC

363
LEU ASP SER VAL SER PHE GLY CYS TYR GLN LEU MET ALA ARG TYR LEU GLN ASP ASP LYS
CTC GAC AGT GTG TCA TTT GGC TGC TAT CAG CTG ATG GCT CGC TAT CTG CAG GAT GAT AAA

423
GLU LYS ILE ILE CYS LEU GLY SER GLY LEU THR LEU LYS LYS ALA LEU GLN ARG ILE ASP
GAG AAA ATC ATC TGT CTG GGC TCG GGC CTG ACC CTG AAA AAA GCG CTG CAG CGC ATC GAT

483
PHE ASP SER LEU ASN THR ARG CYS VAL ALA LEU ILE SER ALA MET ASN ALA ASP GLY GLN
TTT GAC AGC CTG AAT ACC CGC TGC GTG GCG TTG ATC AGC GCC ATG AAC GCC GAC GGG CAG

543
CYS ASN TYR TYR ASP ASP VAL PRO LEU LEU LEU THR ARG LYS ILE LYS ALA LYS TYR TYR
TGC AAT TAT TAC GAT GAC GTG CCC CTG CTG CTG ACC CGC AAA ATT AAG GCC AAG TAC TAT

603
GLN TRP PRO ALA PRO ARG TYR ALA GLN SER ALA ASP GLU TYR GLU MET TRP CYS THR ASN
CAG TGG CCG GCG CCG CGC TAC GCC CAA AGC GCG GAT GAG TAT GAA ATG TGG TGC ACC AAT

663
ARG LEU PHE ARG SER VAL SER GLY VAL ALA ALA ARG ARG THR ***
CGT TTA TTC CGC AGC GTC TCC GGC GTC GCC GCG AGG CGG ACG TGA TTT TTG TCG GCA TTG
```

202 amino acids (Horii et al, 1981) are of a comparable size, whereas E.coli LacI and Gal R are significantly larger; 360 and 343 amino acids respectively (Farabaugh, 1978; Wilcken-Bergmann and Müller-Hill, 1982). In common with other repressors, the amino acid composition (Table 6ii) of DalR is not noticeably different from proteins involved in general cellular metabolism. Two notable exceptions are the large aromatic residue tyrosine, for which a rather high frequency is observed, and cysteine, where again the content is about twice the normal value for a protein of this size. In common with GalR and LacI the repressor shows a clustering of basic amino acids near the C-terminus, but neither DalR nor any of the other repressors that have been sequenced to date display any imbalance in the ratio of basic:acidic residues (Table 6iii) which might be expected of a protein that binds to nucleic acids.

Codon usage in dal repressor reflects the high (58%) G+C content of the DNA. Those codon families having a 3 or 4-fold degeneracy reveal a distinct preference for codons ending in G or C; this is particularly prevalent among the Ala, Val, Gly and Arg groups (Table 6iv). Surprisingly, AAU is preferred over AAC for Asn, and GAU is the most common Glu triplet. Serine is largely coded by AGC, and CAG is the overwhelming choice for Gln. The usual bacterial trends towards ACC (Thr) and CUG (Leu) are also present in the dalR gene.

A comparison of DalR with other DNA-binding regulatory proteins

In bacteria, the control of transcription initiation often involves the interaction of repressor or activator proteins with specific sites on the DNA, and although a number of these operators have now been sequenced little is known of the DNA-binding sites on the proteins themselves. Gicquel-Sanzey and Cossart (1982), in a comparison of 13 prokaryotic regulatory

Table 6ii Amino acid composition of the d-arabitol repressor

Amino acid	No. of Residues	No. of Residues	
Met	4	5	
Ala	19	21	
Ile	11	12	
Leu	20	21	
Pro	6	6	Hydrophobic 47%
Tyr	12	12	
Trp	2	3	
Phe	4	4	
Val	11	12	
Ser	14	15	
Thr	7	7	Polar 31.4%
Asn	8	8	
Gln	12	13	
Cys	7	7	
Gly	9	9	
Asp	10	13	Acidic 11.6%
Glu	10	11	
Lys	9	11	Basic 14.1%
Arg	11	13	
His	2	2	
Total	188	205	

Figures in the third column are calculated from the longer open reading frame which begins only 88bp from HindB, whereas those in the second column are based on translational initiation at the second in-phase AUG codon.

Table 6iii A comparison of the amino acid composition of DalR with other fully sequenced repressor proteins

Amino acid	Number of residues per molecule				
	TrpR	λcI	LacI	GalR	DalR
Met	4 (3.7)	8 (3.3)	10 (2.7)	11 (3.2)	5 (2.4)
Ala	10 (9.2)	19 (8.1)	44 (12)	44 (12.8)	21 (10.2)
Ile	3 (2.7)	10 (4.2)	18 (5.0)	18 (5.2)	12 (5.8)
Leu	19 (17.6)	19 (8.1)	41 (11.3)	32 (9.3)	21 (10.1)
Pro	4 (3.7)	15 (6.4)	14 (3.9)	14 (4.1)	6 (3.0)
Tyr	2 (1.8)	7 (2.9)	8 (2.2)	10 (2.9)	12 (5.8)
Trp	2 (1.8)	3 (1.3)	2 (0.5)	1 (0.3)	3 (1.4)
Phe	1 (0.9)	12 (5.1)	4 (1.1)	8 (2.3)	4 (1.9)
Val	5 (4.6)	15 (6.4)	34 (9.4)	28 (8.2)	12 (5.8)
Ser	6 (5.5)	22 (9.3)	32 (8.8)	24 (7.0)	15 (7.3)
Thr	4 (3.7)	10 (4.2)	19 (5.2)	19 (5.5)	7 (3.4)
Asn	5 (4.6)	8 (3.3)	12 (3.3)	17 (4.9)	8 (3.9)
Cys	0 -	3 (1.3)	3 (0.8)	4 (1.2)	7 (3.4)
Gln	6 (5.5)	11 (4.6)	27 (7.5)	12 (3.5)	13 (6.3)
Gly	5 (4.6)	17 (7.2)	22 (6.1)	24 (7.0)	9 (7.5)
Asp	4 (3.7)	9 (3.8)	17 (4.7)	18 (5.2)	13 (6.3)
Glu	12 (11.1)	22 (9.3)	16 (4.4)	18 (5.2)	11 (5.3)
Lys	4 (3.7)	17 (7.2)	11 (3.1)	6 (1.7)	11 (5.3)
Arg	9 (8.3)	8 (3.3)	19 (5.2)	23 (6.7)	13 (6.3)
His	2 (1.8)	1 (0.4)	7 (1.9)	12 (3.5)	2 (0.9)
Total	108	236	360	343	205
BASIC	13.8%	11.0%	10.2%	11.9%	14.1%
ACIDIC	14.8%	13.1%	9.2%	10.4%	11.6%

Figures in brackets are percentages of the total number of residues per molecule.

Table 6iv Codon usage in the dalR gene

1st \ 2nd	T		C		A		G		3rd
T	Phe	3	Ser	0	Tyr	6	Cys	1	T
		1		2		6		6	C
	Leu	1		1	stop	0	stop	1	A
		3		2	stop	0	Trp	2	G
C	Leu	0	Pro	0	His	1	Arg	1	T
		2		2		1		6	C
		0		0	Gln	2		0	A
		14		4		10		2	G
A	Ile	4	Thr	0	Asn	6	Ser	1	T
		7		6		2		8	C
		0		0	Lys	7	Arg	1	A
	Met	4		1		2		1	G
G	Val	0	Ala	1	Asp	7	Gly	0	T
		3		10		4		7	C
		1		0	Glu	6		0	A
		7		8		3		2	G

TOTAL = 188 Amino acids

proteins, find two distinct regions of homology within all of them, and, in the light of similarities disovered also in their target DNA sequences, a general model for all protein-DNA interactions is put forward. Both areas of homology are located in the known DNA-binding domains of LacI (reviewed by Beyreuther, 1978) and CRP (Aiba and Krakow, 1981) from E.coli and the cI protein of phage λ (Pabo et al, 1979). I^{-D} mutations which impair the binding of Lac repressor to its operator are concentrated in these two conserved sequences (Müller-Hill, 1975), thus confirming their importance in DNA interactions. The DalR protein sequence, towards its N-terminus, bears strong resemblances to numerous other repressors and activators (Fig. 6i). Over a range of 20 residues, (Ala-13 to Glu-32), 10 occur at the same position 3 or more times in the other 13 sequences and correspond to the strongest zone of homology (Region I) described by Gizquel-Sanzey and Cossart. Some similarity to the second sequence (Region II) is also apparent (Fig. 6i) and points to the involvement of the N-terminal 60 or so amino acids of DalR in operator recognition and binding. The same authors also found common features in the DNA at the operator sites of each of those proteins studied. In general, these all have two-fold rotational symmetry and fit with one of two consensus sequences. Up to this point it has been postulated that the three 17bp repeats in the dalDK promoter might bind the repressor, but as these lack any symmetry element and do not contain sequences in agreement with the consensus for other regulatory proteins, their importance is questionable. A suitable operator sequence can, however, be found spanning the transcription start for dalDK mRNA (Fig. 6j), and if this is the DalR binding site then the situation is similar to that in the lac operon, where the repressor is positioned to

Fig. 6i

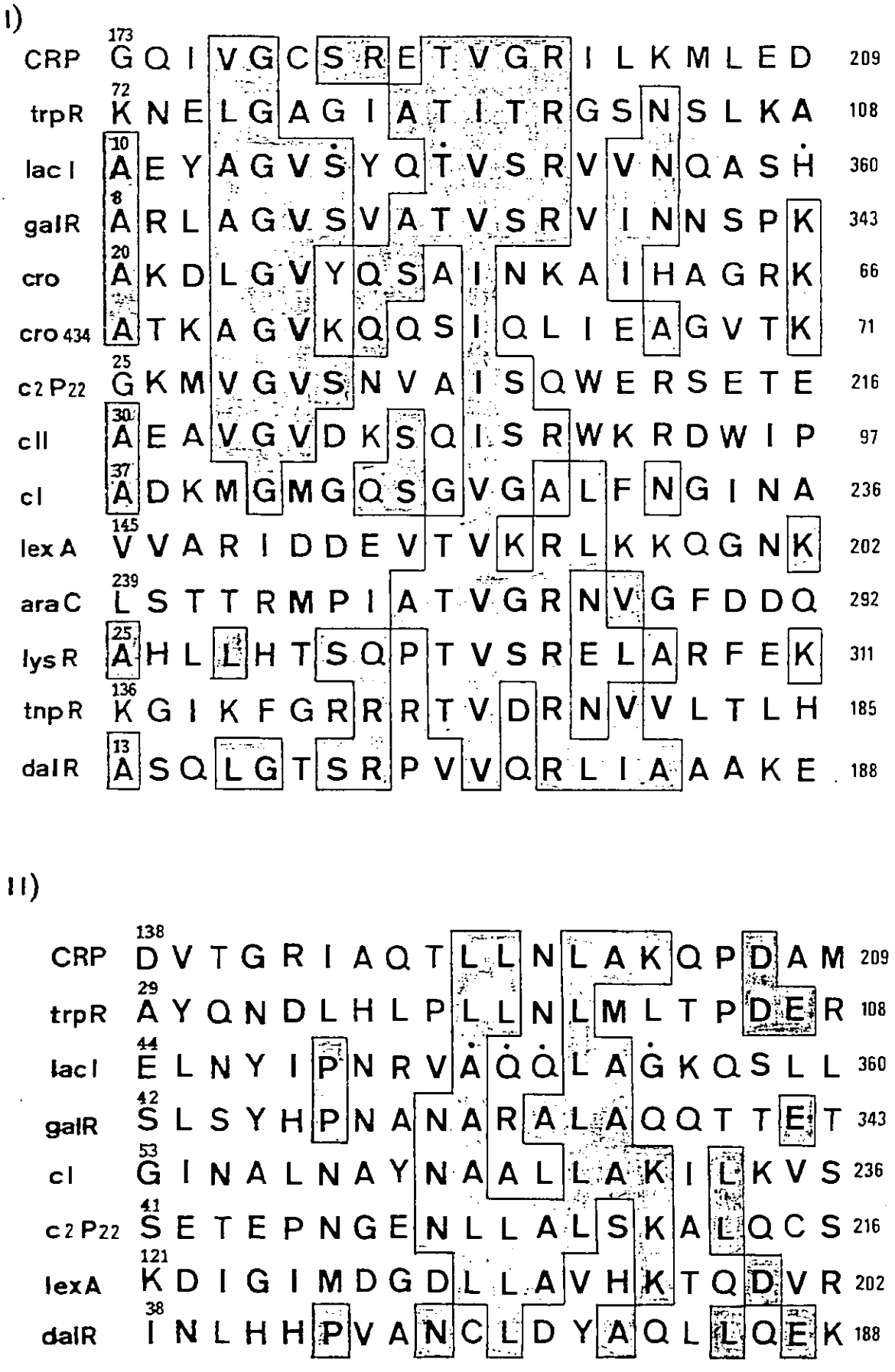
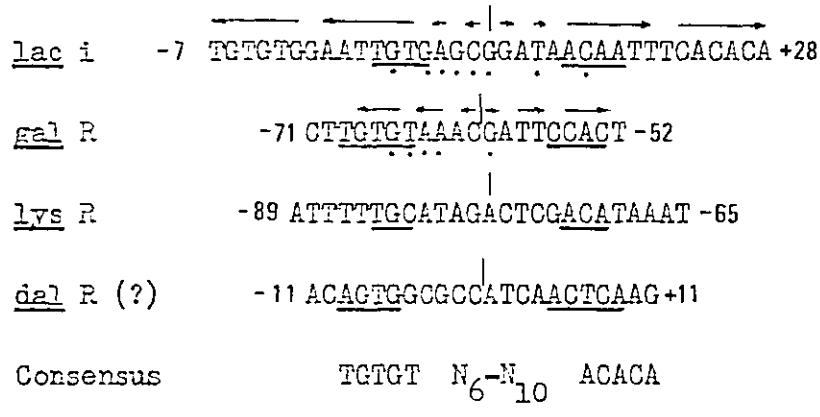

I)
CRP 173 G Q I V G C S R E T V G R I L K M L E D 209
trpR 72 K N E L G A G I A T I T R G S N S L K A 108
lac I 10 A E Y A G V S Y Q T V S R V V N Q A S H 360
galR 8 A R L A G V S V A T V S R V I N N S P K 343
cro 20 A K D L G V Y Q S A I N K A I H A G R K 66
cro 434 A T K A G V K Q Q S I Q L I E A G V T K 71
c2 P22 25 G K M V G V S N V A I S Q W E R S E T E 216
cII 30 A E A V G V D K S Q I S R W K R D W I P 97
cI 37 A D K M G M G Q S G V G A L F N G I N A 236
lexA 145 V V A R I D D E V T V K R L K K Q G N K 202
araC 239 L S T T R M P I A T V G R N V G F D D Q 292
lysR 25 A H L L H T S Q P T V S R E L A R F E K 311
tnpR 136 K G I K F G R R R T V D R N V V L T L H 185
dalR 13 A S Q L G T S R P V V Q R L I A A A K E 188
II)
CRP 138 D V T G R I A Q T L L N L A K Q P D A M 209
trpR 29 A Y Q N D L H L P L L N L M L T P D E R 108
lacI 44 E L N Y I P N R V A Q Q L A G K Q S L L 360
galR 42 S L S Y H P N A N A R A L A Q Q T T E T 343
cI 53 G I N A L N A Y N A A L L A K I L K V S 236
c2 P22 41 S E T E P N G E N L L A L S K A L Q C S 216
lexA 121 K D I G I M D G D L L A V H K T Q D V R 202
dalR 38 I N L H H P V A N C L D Y A Q L L Q E K 188

Fig. 6i Homologies between the amino acid sequence of DalR and those of 13 other DNA-binding regulatory proteins. Lac i^{-D} mutations are marked by an asterisk.Residues which appear three times or more in any one position are shaded.The first amino acid of each string is numbered according to its position in the protein.The number in the right-hand column is the total number of residues per subunit.

Note that for DalR,positions 12,16 and 26 are occupied by hydrophobic amino acids and 9,10,11,14,15,19,20,24 and 25 by polar or charged residues,(see text).

Fig. 6j A comparison between a number of operator sites and a consensus sequence for repressor/DNA interactions



• = mutations which affect the binding or action of repressor. Arrows mark the position and extent of dyad symmetry elements.

The putative dalR operator site shown above is centred on the transcription start point of the dalDK operon. It bears a remarkable similarity to the CRP site at position -41bp (TGTG N_8 CTCT). The repressor consensus sequence is taken from Gicquel-Sanzey and Cossart, (1982).

effectively block the initiation of transcription. In the dalDK promoter, though, DalR would also cover the RNA Polymerase binding site at -10, since this constitutes the first element of the recognition sequence, centred between positions +1 and -1. Despite a good match with the consensus sequence, the dalDK "operator" lacks any obvious dyad symmetry, but this is not an absolute requirement, as demonstrated by the sites for cII protein in the λP_i and λP_E promoters and by the lysR operator, which are all deficient in this respect (Gicquel-Sanzey and Cossart, 1982). Note that the proposed dalDK operator does not differ enormously from the CRP site centred at -42/-41 (see Fig. 6j). Given the homologies that exist between their recognition sequences, then it should be possible for any regulatory protein to interact with the binding site of another, and indeed this phenomenon has been observed in the lac and ara operons where CRP is able to bind, at reduced efficiency, to both of these operators (Schmitz, 1981; Ogden et al, 1980). Subtle differences in the DNA domains of the proteins and in the nucleotide sequence itself govern the strength and specificity of binding and are sufficient to ensure that any cross-reactions of this type do not pose problems for the organism.

A recent model for the binding of Lac repressor to DNA (Wilcken-Bergmann and Müller-Hill, 1982) proposes that α-helical arms from two subunits (residues 50-60) are positioned in the major groove with Lys·59 near the symmetry centre of the operator. Steitz et al (1982) have found striking structural similarities between CRP and λ cro proteins, and each binds such that two helices (helix α3 of cro and helix F of CRP) from two monomers also make contacts in major groove. These helices

span Region I (Fig. 6i) and are believed to be the main determinant for recognition and binding of CRP and cro. In the same paper, Steitz et al showed that CRP and cro both have three consecutive α-helices in the N-terminal region, and the remainder of the polypeptide chain folds into a series of anti-parallel β-sheets. This structure compares favourably with that predicted for the Dal repressor (Table 6v, Fig. 6k) by the Chou and Fasman method, using the IntelliGenetics computer program.

Among the lysogeny establishment proteins of λ, P22 and 434 phages Sauer et al (1982a) observe a number of highly conserved residues which are important in maintaining the essential spatial arrangement of α-helices 2 and 3 in the operator-binding domain. λ repressor Gly·41 forms part of the turn between the helices while Ala·37 and Val·47 make significant van der Waal's contacts and help determine the angle between α-helix 2 and 3 (Pabo and Lewis, 1982). These three amino acids are retained in LacI, GalR and, significantly, by DalR (residues Ala·13, Gly·17 and Val·23), suggesting a helix-turn-helix super-structure for each. Residues 36, 40, 42 and 50 of the λ repressor are hydrophobic, while 33-35, 38/39, 43-46 and 48/49 are hydrophilic, and the same is also true of the phage P22 and 434 proteins. Equivalent positions in DalR are identical in this respect (Fig. 6i), which strongly supports the view that it adopts a similar structure in this region of its DNA-binding domain. By analogy with λ repressor and LacI (Beyreuther, 1978), Ile·12, Leu·16, Thr·18 and Leu·26 might constitute part of the hydrophobic core of the domain, while the hydrophilic or charged amino acids in positions 9-11, 14/15, 19-21 and 24/25 remain solvent exposed. Residues Ser·19 to Arg·25 are predicted to include the DNA-contact residues. Serine, arginine and glutamine are found in the DNA-binding sites

```
SECONDARY STRUCTURES PREDICTED BY CHOU AND FASMAN ALGORITHM:

KEY:     AAAAAA  ALPHA HELIX
         BBBBB   BETA SHEET
         TTTT    TURN (TURN FREQUENCY >= ABSTURNMIN)
         T?T?    OTHER POSSIBLE TURN SITES (TURN FREQUENCY > RELTURNMIN)


         10        20        30        40        50        60        70
MSKEDDIRLDQKVRAAMMYYIAQQNQSEIASQLGTSRPVVQRLIAAAKEEGIVSINLHHPVANCLDYAQL
AAAAAAAAAAAAAAAAAA     AAAAAAAAAA        AAAAAAAAAA              AAAAAAA
      BBBBB   BBBBBBBB               BBBBBBBBBB     BBBBBBBB BBBBB BBBB
 T?T?           T?T?  TTTTTT     TTTT                          TTTT?T?

         80        90        100       110       120       130       140
LQEKYGLIECNVVPAFSEESTLDSVSFGCYQLMARYLQDDKEKIICLGSGLTLKKALQRIDFDSLNTRCV
AAAA       AAAAAAAAAAAAAA    AAAAAAAAAAAAAAAAAA   AAAAAAAAAAAAAAAA    A
BB    BBBBBBB              BBBBBBBBBBBBB   BBBBB          BBBBBB  BBBBBB
  TTTT                   TTTT        TTTTTT    TTTT          TTTT TTTT

        150       160       170       180       190       200
ALISAMNADGQCNYYDDVPLLLTPKIKAKYYQWPAPRYAQSADEYEMWCTNRLFRSVSGVAARRT
AAAAAA             AAAAAAAAAAA          AAAAAA                AAAAAA
BBB      BBBBBB  BBBBBBB   BBBBBB           BBBBBBBBBBB
      TTTTTTTTT?T?         T?T?   TTTT  TTTT    TTTT   TTTT
```

Fig. 6k The secondary structure of the arabitol repressor protein

Table 6v Secondary structure predictions for the arabitol repressor .

SECONDARY STRUCTURES PREDICTED BY CHOU AND FASMAN ALGORITHM:

BEGIN	END	TYPE	<PA>	<PB>	<PT>
1	18	ALPHA	1.159	0.939	
2	5	RELTURN	1.113	0.600	1.160
7	11	BETA	1.078	1.094	
15	22	BETA	1.156	1.181	
17	20	RELTURN	0.978	1.340	0.960
23	26	TURN	0.865	0.960	1.270
24	33	ALPHA	1.076	0.979	
25	28	TURN	1.015	0.778	1.178
34	37	TURN	0.788	0.905	1.225
37	46	BETA	1.089	1.147	
41	50	ALPHA	1.282	0.890	
52	59	BETA	0.984	1.198	
61	65	BETA	1.012	1.182	
61	64	TURN	0.963	1.153	0.978
64	74	ALPHA	1.122	1.022	
64	67	RELTURN	0.903	1.125	1.095
67	72	BETA	1.125	1.183	
73	76	TURN	0.983	0.833	1.113
77	83	BETA	1.041	1.250	
82	95	ALPHA	1.049	0.991	
96	99	TURN	0.793	1.018	1.195
97	104	BETA	1.035	1.134	
100	105	ALPHA	1.143	1.113	
103	117	ALPHA	1.119	1.020	
105	109	BETA	1.000	1.068	
108	111	TURN	1.073	0.730	1.228
110	113	TURN	1.210	0.598	1.055
113	117	BETA	1.046	1.286	
118	121	TURN	0.780	0.888	1.285
121	135	ALPHA	1.100	1.036	
127	132	BETA	1.087	1.142	
131	134	TURN	0.980	0.803	1.238
135	143	BETA	1.018	1.214	
136	139	TURN	0.795	1.050	1.165
140	146	ALPHA	1.201	1.151	
147	150	TURN	0.918	0.753	1.310
150	155	BETA	0.738	1.145	
150	153	TURN	0.763	0.983	1.323
152	155	TURN	0.688	1.255	1.258
155	158	RELTURN	0.943	1.063	1.140
158	164	BETA	1.010	1.181	
159	169	ALPHA	1.090	1.020	
168	173	BETA	1.025	1.163	
168	171	RELTURN	0.990	1.128	0.988
175	178	TURN	0.915	0.945	1.068
179	184	ALPHA	1.207	0.737	
181	184	TURN	1.178	0.623	1.073
185	195	BETA	1.021	1.097	
189	192	TURN	0.795	1.050	1.165
196	199	TURN	0.793	0.988	1.230
199	204	ALPHA	1.080	0.995	
207	214	BETA	1.108	1.195	
208	217	ALPHA	1.129	1.068	
207	234	ALPHA	1.063	1.035	

of most regulatory proteins, and, being hydrophilic, they are capable of forming hydrogen bonds with the DNA.

The phage repressors share extensive sequence homology in their C-terminal regions, which is probably related to RecA-binding (Sauer et al, 1982b). This is peculiar to the λ , 434 and P22 proteins and to LexA and is not exhibited by GalR, LacI or DalR.

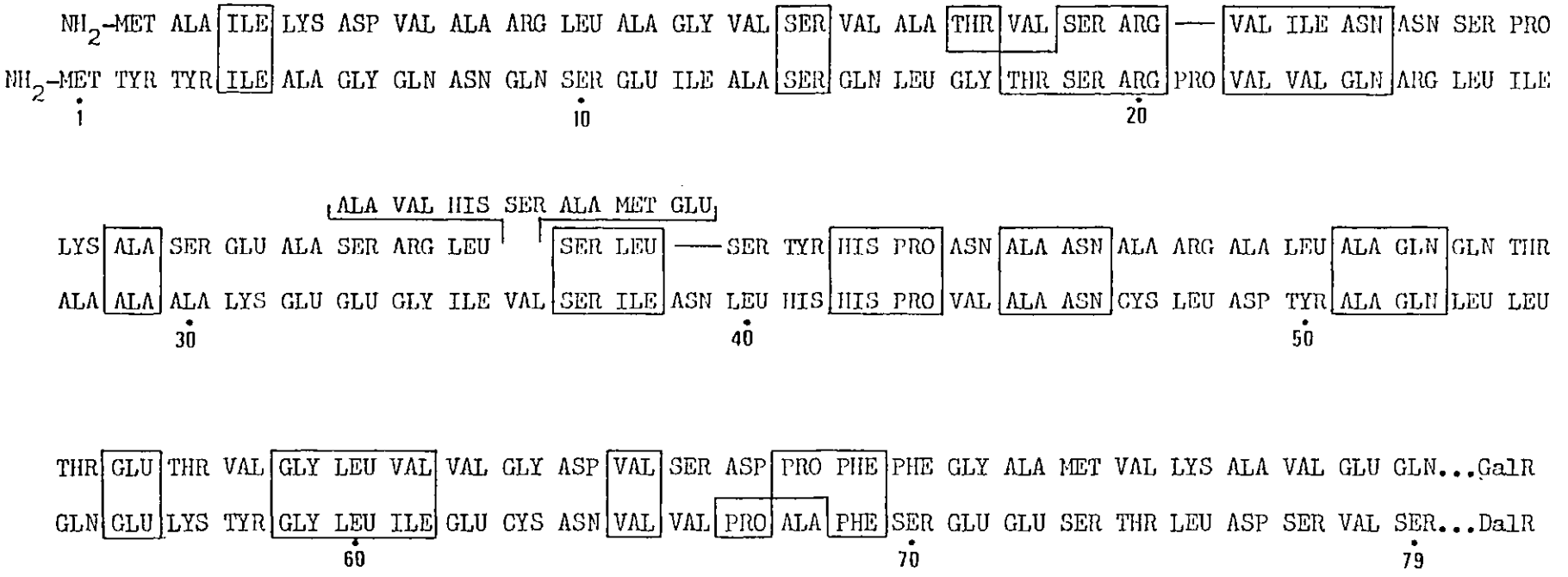
Considering the presence of strong helix structures in the N-terminal region of the Dal repressor, its similarity to other regulatory proteins and the discovery of a potential operator in the dalDK promoter having homology to a consensus sequence, a similar mode of action to that of cro or CRP is anticipated; helices from two subunits interacting with successive major grooves of a right-handed B DNA. Since all the repressors or activators studied to date are either dimers (e.g. CRP; Eilen and Krakow, 1977) or tetramers (e.g. LacI ; Riggs and Bourgeois, 1968), it is fairly safe to assume that DalR is no exception. The C-terminal domain of LacI is involved in oligomerisation (Beyreuther, 1978), but shows no obvious homology to similar regions of DalR.

Evolution of the Dal repressor

Sequence homologies clearly show that the repressor and cro proteins of phages λ, 434 and P22 have evolved from a common ancestor (Sauer et al, 1982a).

In a recent report (Wilcken-Bergmann and Müller-Hill, 1982) a common evolutionary origin is suggested for E.coli Lac and Gal repressors, based upon a comparison of their amino acid sequences. By aligning the DalR sequence with that of the galactose repressor, it is possible to pick out a region in both which appears to be conserved (Fig. 61). These amino acid

Fig. 61 A comparison of the N-terminal sequences of DalR and GalR

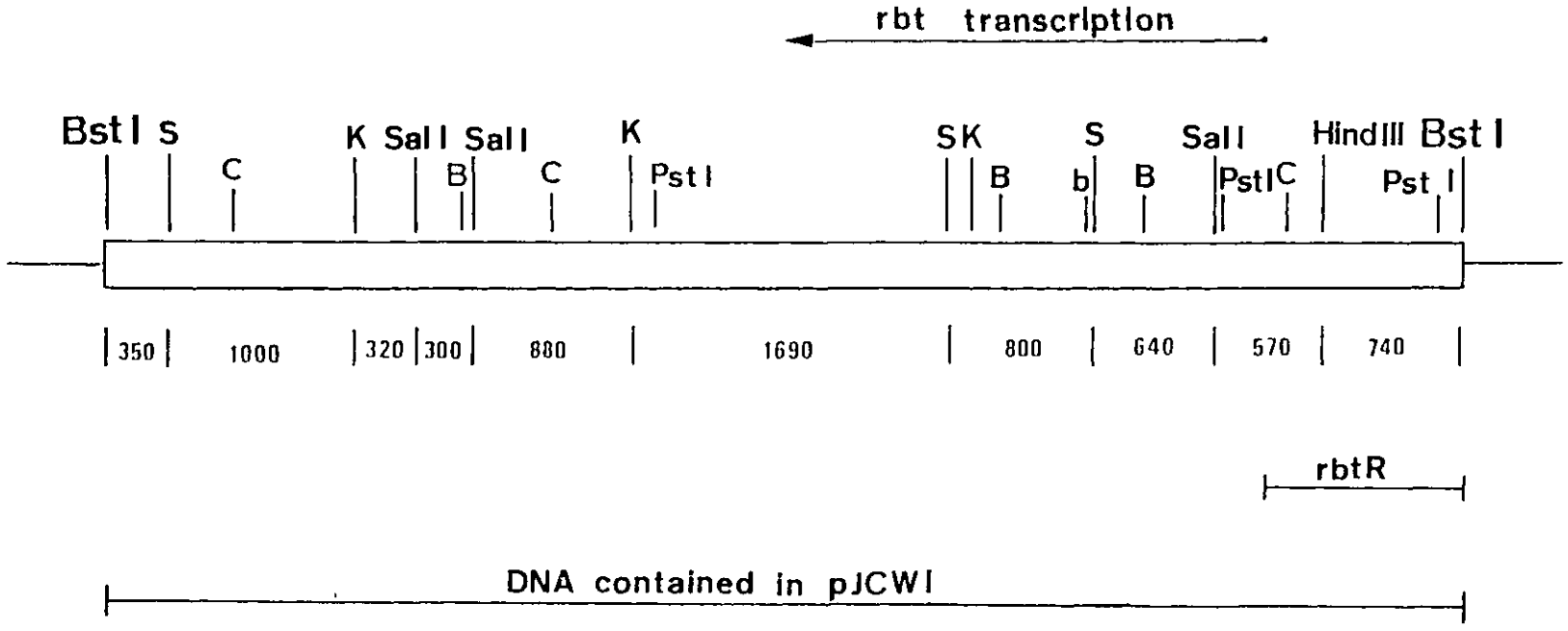


Identical or similar amino acids are boxed. It has been necessary to insert two spaces and delete seven residues from the GalR sequence in order to maximise the homologies.

homologies occur, in a substantial number of cases, at positions where matches are also strongest between GalR and LacI. The N-terminal sequences do not align perfectly, and it has been necessary to loop out seven amino acids from GalR and insert two spaces to maximise the matches. Homology between the Dal and Gal repressors is confined to the N-terminal 70 residues, and no similarities are visible between the remaining 118 amino acids of DalR and the region spanning AA160-260 of either LacI or GalR, believed to have a role in inducer-binding (Miller, 1978). Note that in this alignment the first part of the DalR sequence is shifted by five amino acids relative to GalR, compared to that in Fig. 6i. The alternative arrangement still gives quite a number of matches but requires additional gaps to be inserted in the DalR sequence.

Despite amino acid homologies between their N-termini, computer analyses show the DNA sequences of dalR and galR and lacI to be unrelated. If common ancestry is proposed for these proteins, then the absence of detectable DNA homology must reflect that a long time has elapsed since divergence. It is more likely that any similarities arise out of the need to preserve an established arrangement of α-helices in the DNA-binding domain; a situation analogous to the sequence homologies within the NAD^+ sites of various dehydrogenases. It is obviously going to be of immense interest from an evolutionary point of view to compare the Dal repressor sequence with that of the closely related rbt operon when it becomes available. λ p rbt dal has a mutation in rbtR, believed to be a deletion, which confers constitutivity, but the wild-type sequence from K.aerogenes FG5 has now been cloned and is currently being sequenced (J-C. Wu, unpublished data). With the demonstration that plasmid pJCW1 (Fig. 6m) carries an inducible

Fig. 6m A restriction map of the rbt region of the K.aerogenes PG5 chromosome



rbt operon, and therefore makes a functional repressor, we now know that *rbt*R does *not* span the *Bst*A site. This leaves almost 0.5kb of putatively non-coding DNA separating the pentitol operons. The secondary λ *att* site responsible for the generation of λp *rbt* is contained in this region (Loviny *et al*, 1981).

CHAPTER 7

DISCUSSION

Pentitol Dehydrogenase Amino Acid Sequences

Now that the complete amino acid sequences have been analysed, it is clear that ArDH and RDH are no more similar than one might expect of two NAD^+-linked dehydrogenases. At the DNA level, the structural genes too appear totally unrelated, and only a few small, isolated stretches of the sequence show more than 30% identity. All dehydrogenases show a degree of structural homology in their NAD^+-binding domains (Rossman et al, 1975), but this is not generally notable, even at the amino acid sequence level. Hence, homology at the DNA level will naturally be even harder to detect.

The region between Thr·101 and Lys·137 of ArDH is believed to represent part of the cofactor-binding site by analogy to sequences in the known NAD^+ domains of various other dehydrogenases These comparisons (see Fig. 4f) show a number of potentially significant features. A basic amino acid, usually lysine, followed by a triplet, comprised largely of the three hydrophobic residues, valine, isoleucine and alanine, commonly occurs at the N-terminal end of each 37 amino acid string. Six places further on, one finds a region rich in glycine and/or tyrosine residues, then another basic amino acid four or five places downstream. The C-terminal regions of the compared sequences frequently have a hydrophobic residue (leucine, isoleucine or valine) six positions to the left of a tetrapeptide containing glycine, glutamine and lysine (or histidine). This strongly suggests that all six polypeptides shown in Fig. 4f adopt similar tertiary structures and that the conserved regions outlined above are important either in this respect or are involved directly in binding the coenzyme

NAD^+ but until X-ray diffraction data for the pentitol dehydrogenases are available, these functional assignments remain speculative.

The ArDH N-terminal 150 residues display the familiar $\beta\alpha\beta$ repeating structures common to NAD^+-linked dehydrogenases, a feature which is less prominent in RDH and may account for the lower affinity of the latter towards NAD^+. The higher degree of folding of the ArDH N-terminus could be indicative of a superior nucleotide-binding pocket.

There are few similarities between the two proteins outside their N-terminal regions, and weak amino acid homologies detected by computer alignments of the sequences probably arise from constraints imposed by the requirement to maintain similar secondary structures, and are not sufficiently extensive to be significant in evolutionary terms. No evidence can be seen of repeated DNA or amino acid sequences within ArDH. Internal duplications, as a result of unequal crossing over (Smithies et al, 1962) were probably not important in the evolution of this enzyme.

Most dehydrogenases coupled to NAD^+ possess essential sulphydryl groups reactive with alkylating agents or mercury compounds and against which substrate-binding offers little protection. Burleigh et al (1974) found that RDH has cysteine residues which are important for catalysis, but similar requirements for ArDH activity have yet to be demonstrated. No homology can be seen from comparisons of the sequences flanking ArDH cysteines with those bordering known active cysteines in lactic dehydrogenase or liver alcohol dehydrogenase (Jörnvall, 1981).

These results are somewhat disappointing in terms of the original objective of uncovering potential evolutionary relationships between the rbt and dal operons. Since the most obvious theory of a recent duplication event cannot adequately account

for the compact, symmetric arrangement of the pentitol catabolic unit, an alternative mechanism must be sought. The explanation may prove, ultimately, to be of even greater interest than a simple invert gene duplication, and a number of possibilities will be discussed later.

Codon usage in dal operon proteins

Fiers et al (1976) suggested that the use of codons recognised by tRNAs of limited intracellular availability might influence translation rates, and although the choice of any particular synonymous codon will not alter the nature of the gene product, it may influence the level of expression. Hence, the expressivity of a mRNA may be under the control of degenerate base usage. Genes which are expressed at high levels (e.g. the ribosomal protein genes, Nomura et al (1980)) selectively employ codons specifying the major tRNA species (Ikenura, 1981) and, according to Grantham et al (1980), translational efficiency is the constraint governing codon choice in these instances, whereas genes expressed at only moderate levels or intermittently show a pattern of codon usage reflecting the organisms' genomic G+C content. The structural genes for ArDH and RDH fall into this final category, and indeed codon usage in dalD and rbtD displays many of the non-random features observed by other workers (Elton et al, 1976; Berger, 1978; Crawford et al, 1980; Nichols et al, 1981a and 1981b). The avoidance of minor tRNAs is not absolute and may be explained in several ways. It is possible that K.aerogenes may differ with respect to intracellular levels of certain amino-acyl tRNAs, for example valine and alanine are two instances where the codon profile is at variance with the majority of the E.coli data. Alternatively, it is likely that those genes which need to be expressed only at moderate levels or during times of metabolic deficiency are able to tolerate

the limited use of less abundant tRNAs without greatly influencing cell viability or fitness. Additionally, some small deviations from the ideal choices might signify a functional or structural role for particular codons which overrides the consideration of translational efficiency. The general preference for Phe, Ile and Asn codons terminating with C is common and serves to compensate for the intrinsically weak interaction of the first two bases with the anticodon, but AAA is preferred over AAG for lysine as the thiolation of uridine residues tends to restrict G·U wobble-pairing. Efficient translation selects degenerate codons which give codon-anticodon interactions of intermediate strength in preference to those that bind very strongly or very weakly, but codon usage in weakly expressed genes follows opposite rules (Grosjean and Fiers, 1982). Repressors, for example, often show a much more random codon profile overall, suggesting that this factor is less restrictive than in more frequently transcribed genes, and there also tends to be a higher proportion of "rate-limiting" codons (i.e. those recognised by minor tRNAs or by a weakly interacting isoaccepting tRNA). Some examples of the latter, amongst E.coli species, include AUA (Ile), CGG/CGA/AGA and AGG (Arg), CUA (Leu) and GGA (Gly), and although dalR shows considerable agreement with the genome hypothesis of Grantham et al (1980) in overwhelmingly choosing codons terminated by C or G, it cannot be said to comply with the rules of Grosjean and Fiers (1982) with respect to the use of these rarer, "modulating" codons. This suggests that the dalR messenger may be translated quite efficiently and that controls, if present, may depend upon mRNA secondary structure. Rapid translation of mRNAs specifying even rare proteins may still be desirable to ensure

efficient use of the ribosome, and the bias towards major tRNA species might represent a means of minimising translational errors. Several experimenters have shown that when decreased levels of an aminoacyl tRNA limit the rate of protein synthesis, the occurrence of translational errors increases due to more effective competition from incorrect tRNAs, (Edelmann and Gallant, 1977; O'Farrell, 1978; Parker *et al*, 1978). Choosing abundant tRNA species reduces this risk.

The *dal*DK operon promoter/operator region

The full nucleotide sequence of the *dal*DK control region has been determined. No similarity is seen with sequences which precede the origin of *rbt*DK transcription (Fig. 7a). The DNA reveals a total of five possible RNA Polymerase recognition sequences, some of which are overlapped by three A+T rich repeated sequences that may be important in repressor binding. An element of dyad symmetry is found centred at -35 (Fig. 5e) relative to the putative transcription start, and a good Pribnow box sequence is situated around -10bp which differs in only one position from the ØX174B promoter, 5'-TACAGTA-3'. Between -50bp and -33bp a sequence believed to be involved in CRP·cAMP binding has been located.

Although dyad symmetries are not general features of prokaryote promoters, many striking examples are known, including the $tRNA^{Tyr}$, *str* and *spc* promoters (Sekiya and Khorana, 1974; Post *et al*, 1978), but such structures have not yet been proven to be important in the selection of sites by RNA Polymerase. For the *rplKA*(L11) and *rplJ*(β) promoters it is thought that dyad symmetries may be involved in the termination of upstream transcription to prevent readthrough into the operon. A relationship between transcription initiation and termination signals has been proposed (Rosenberg *et al*, 1978). To be significant *in vivo*

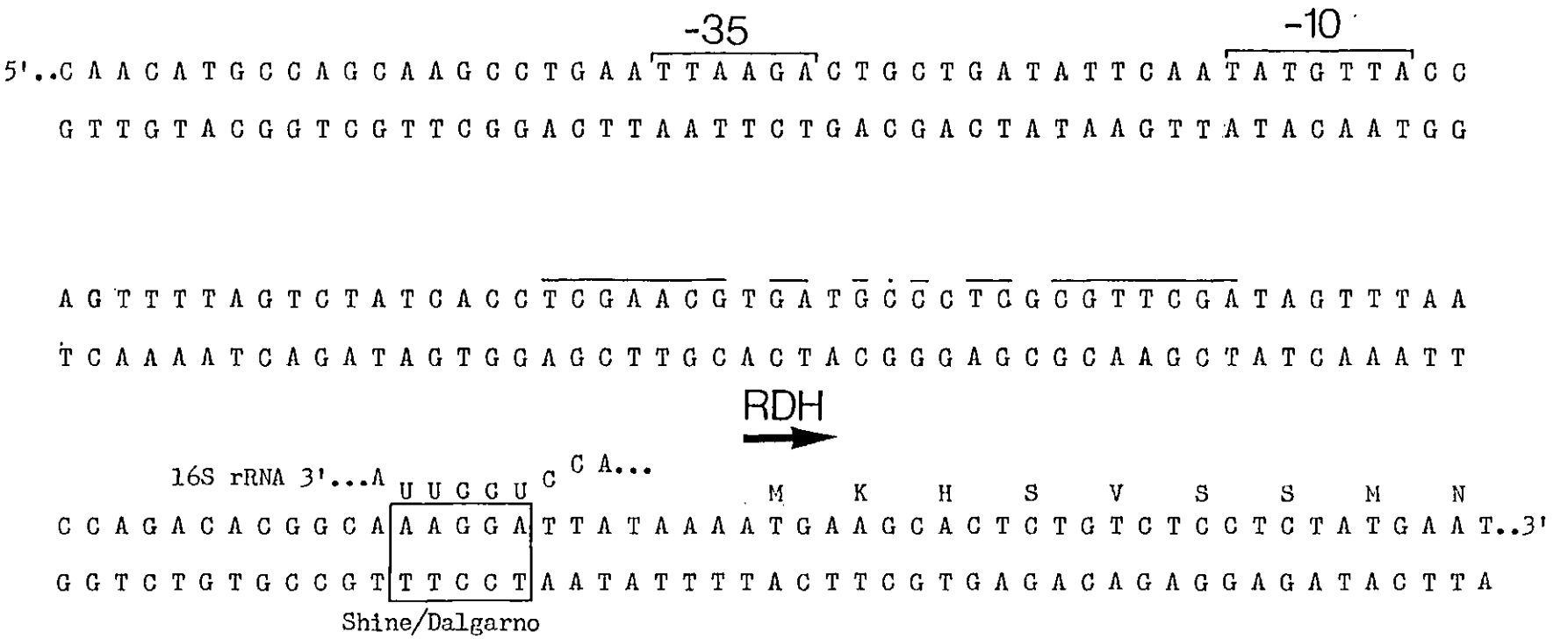


Fig. 7a The rbtDK promoter/operator region. The RDH coding sequence is preceded by a good ribosome-binding site 7bp upstream. The best candidates for -35 and -10 sequences are indicated together with a dyad symmetry which could form a stable RNA hairpin.

a stem and loop structure must be thermodynamically at least 8kCal/mol more stable than the linear DNA (Gralla and Crothers, 1973). The hairpin which can be formed from the dalDK promoter sequence might fulfil this requirement (ΔG = -8.2 kCal/mol), but is unlikely to be involved in transcription termination since the adjacent gene (dalR) is transcribed in the opposing direction.

Insufficient data exist at present for the rbtDK promoter to permit any meaningful comparisons to be made, but the Pribnow box and transcription start point have been arbitrarily assigned to the positions indicated in Fig. 7a. If correct, the rbt message has a leader sequence 73 bases long and good Shine-Dalgarno (SD) sequence 7 nucleotides 5' to the AUG initiation codon. A small dyad symmetry in the template DNA corresponding to the mRNA 5' end might represent an operator site positioned so as to block transcription initiation, by analogy to the lac operon (see Fig. 6j). The best candidates for dalDK operator sequences are perhaps the three 17bp direct repeats spread throughout the promoter region. Although not symmetrical, they are similar in size to other known operators, and it is perhaps worth noting that a sequence with considerable homology to these repeats occurs 5' to 3' on the sense strand of the DNA spanning the translation start point (Fig. 7b). An alternative operator is given by the sequence AGTGG N_8 ACTCA, which is centred at the origin of dalDK transcription and closely resembles a proposed repressor consensus sequence TGTGT N_{6-10} ACACA (Gicquel-Sanzey and Cossart, 1982).

Catabolite Repression in the dalDK operon

Evidence has been presented which shows that in vivo expression of the dalDK operon is modulated by cAMP, and this is supported by by exploratory protein-DNA binding studies. The interaction site

Fig. 7b A sequence homologous to the R2 repeat is present on the sense strand spanning the translation start point of the dalD gene. A total of 11 out of 17 bases are identical.

for CRP is believed to be situated between positions -33 and -50 relative to the start of dalDK transcription. Rosenberg and Court (1979) find that in addition to the usual -35 polymerase recognition sequence a small symmetry element nearby is conserved in catabolite-sensitive promoters. This takes the general form 5'-TGTG N_8 CACA-3' on the antisense strand, and a similar string occurs in the dalDK promoter (Fig. 5i). The position of CRP sites can vary, suggesting that direct CRP-polymerase contacts may not be involved in the enhancement of transcription. Recent re-examinations of the lac o/p region have uncovered a second CRP site in addition to the one previously known to exist around -60 (Schmitz, 1981). The new site, which lies within the operator, shows considerable homology to the consensus sequence. The CRP-binding sites of galP1, araC, araBAD and cat (chloramphenicol acetyl transferase) all overlap the promoter -35 region (Le Grice and Mutzura, 1981) and, in the case of cat, it has been demonstrated that CRP increases DNAase I protection by RNA Polymerase at the Pribnow box, although the -35 region is inaccessible to polymerase. These data imply that binding of the cAMP·CRP complex at or near -35 may obviate the requirement for RNA Polymerase to recognise this region and subsequently stimulates transcription by encouraging polymerase to bind directly to the Pribnow box.

The nucleotide sequence near -35 in the dalDK promoter is not strikingly similar to what is now considered the ideal sequence for polymerase recognition (Siebenlist et al, 1980). It is significant to note that this is true for most catabolite-sensitive promoters sequenced to date. Only the lac promoter gives a reasonable fit (TTTACA), but in this system neither of the two CRP sites overlap the -35 region. The lac promoter is

known to bind polymerase and exhibit low-level transcription, even in the absence of any positive effector (Majors, 1975). This is certainly consistent with the view that the -35 area is essential for polymerase recognition and that, among those operons which display an absolute requirement for cAMP·CRP, this positive regulator is able to compensate for its absence, i.e. CRP itself is able to perform the recognition function normally associated with the RNA Polymerase holoenzyme. This tallies with observations of very low levels of *in vitro* *dal* transcription when CRP is excluded, presumably resulting from limited recognition of the weak -35 sequence. The promoter is capable of directing high levels of expression in *dal* constitutive strains (ArDH >5% of total soluble protein) grown in M9+CAA where catabolite repression is relieved. In non catabolite repressed operons,levels of transcription are dependent upon direct interaction of polymerase with the template DNA around -35bp. A very strong Pribnow box signal can successfully bypass this requirement, as illustrated by the *lac* UV5 promoter (Schmitz and Galas, 1979).

Transcription from the *dal*DK promoter

The *dal*DK promoter -10 region is arguably the best fit to the consensus sequence TATPuPuTPu of Pribnow (1975), and an RNA transcript is synthesised *in vitro*, whose size is consistent with a transcription event initiated 5bp downstream from this sequence. A number of other products also result during transcription of a restriction fragment carrying the intact promoter. These may occur through pausing of the polymerase or initiation at one or more alternative sites on the template. Csordás-Tóth *et al* (1979) found several sites resembling the RNA Polymerase recognition signal upstream from the P1 promoter of the ribosomal

RNA gene rrnB in E.coli and postulate that polymerase molecules queue up in order to effect rapid, efficient initiation of transcription. Multiple polymerase-binding sites are known for the E.coli galactose operon (Willmund and Kneser, 1973), where up to six molecules may interact simultaneously. An examination of the DNA sequence up to -120bp in the dalDK control region reveals at least five stretches with some homology to the consensus -35 region of Siebenlist et al (1980). DNAase I footprinting experiments suggest that RNA Polymerase can bind to some of these sites. In addition to the Pribnow box at -10, five other sequences can be identified, which are suitably positioned to act as polymerase-binding sites and perhaps function as weak promoters (Fig. 7c). Transcription initiated from one or more of these "promoters" may explain some of the larger transcripts made in vitro in the absence of any positive effector. In view of the closeness of these sites it is unlikely that all can be utilised simultaneously. The importance of sequences between -50 and -100 might best be assessed by measuring the levels of enzymes made by mutants deleted in all or part of this area. Suitable restriction sites exist for the construction of the relevant plasmids. Alternatively, progressive digestion from HindB with the double-strand specific nuclease Bal 31 might be a better approach. Note that for each of the proposed polymerase recognition sites the initial TTG triplets are separated by about 20bp. This distance is equivalent to two helix turns, and hence all of these sequences lie on the same side of DNA duplex. Siebenlist et al (1980) suggest that RNA Polymerase aligns and makes specific contacts along only one side of the double helix; the same side to which the CRP protein binds in catabolite repressed promoters. Polymerase-polymerase interactions

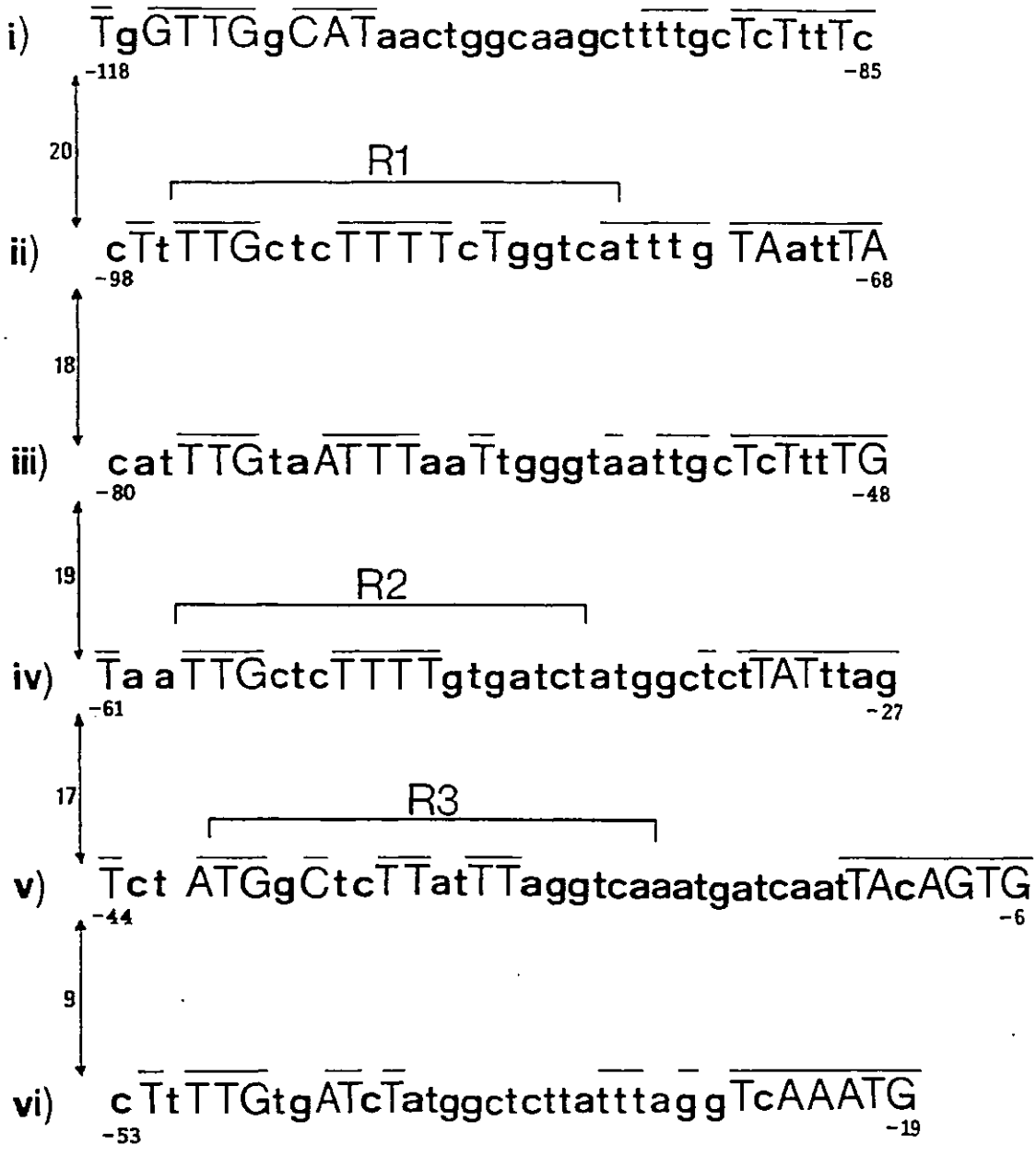

i) TgGTTGgCATaactggcaagcttttgcTcTttTc
-118
-85
20
R1
ii) cTtTTGctcTTTTcTggtcatttg TAattTA
-98
-68
18
iii) catTTGtaATTTaaTtgggtaattgcTcTttTG
-80
-48
19
R2
iv) Ta aTTGctcTTTTgtgatctatggctctTATttag
-61
-27
17
R3
v) Tct ATGgCtcTTatTTaggtcaaatgatcaatTAcAGTG
-44
-6
9
vi) cTtTTGtgATcTatggctcttatttaggTcAAATG
-53
-19

Fig. 7c

Fig. 7c

Alternative promoters for the dalDK operon.Bases are numbered relative to the preferred origin of transcription (see Chapter 5).Capitals are used where there is agreement with the -35 or -10 regions of the canonical promoter sequence (Fig. 5b).Other positions also matching the consensus sequence are marked (——).R1,R2 and R3 are the three repeated sequences of the dalDK promoter.The figures to the left of the diagram show the spacing (in bp) between each of the six elements and with the exception of (v) and (vi) this distance approximates to two helical turns.Hence, the repeats and the putative -35 regions are on the same side of the DNA duplex.

might also be important. The putative polymerase recognition sites I-IV (Fig. 7c) show substantial divergences from what is considered the statistically ideal sequence. This is not uncommon, and similar deviations exist in weak promoters or those requiring activation by a cofactor. According to Pirrotta et al (1980), a non-ideal sequence makes strand separation and transcription initiation difficult in the absence of these additional factors. It should be noted, however, that a good Pribnow box does not necessarily specify the position of a strong start site. The rplJ β promoter initiates from a sequence differing in two positions from the consensus, situated only a few nucleotides upstream of a "perfect" Pribnow box. This promoter also lacks a good -35 region (Post et al, 1979).

Pausing of RNA Polymerase in vitro is well documented. Two small transcripts of 6 and 17 nucleotides are synthesised from the lac promoter (Gilbert et al, 1974), and point mutations causing a decrease in the G+C content between +8 and +10 (some being lac O^C mutations) weaken this pausing, which occurs at an A+T rich sequence approximately one helix turn downstream. This implies that any A+T rich string following a G+C rich region may constitute an effective stop signal in vitro. In the dalDK promoter, bases -5 to -1 are G·C pairs and 7 out of the next 10 base pairs are A·T. This too may represent a pause or stop signal for RNA Polymerase to account for the two minor transcripts of 4 and 10 nucleotides which are synthesised in vitro (Photo 5C). Normally pausing would occur 8-10 bases beyond a transcribed sequence of G·C pairs (Gilbert, 1976), but in this particular case the G+C rich stretch immediately precedes the mRNA start. Whether this arrangement can also elicit pausing is unknown. A G+C rich sequence spanning -5 to +2bp is involved

in transcriptional regulation of operons under stringent control (Travers, 1980), but the pentitol operons do not fall into this category. Maizels (1973), working with lacUV5 mutants, was able to demonstrate that paused RNA synthesis generated a set of transcripts ranging in size from 7 to 100 nucleotides. Synthesis of a 10 nucleotide transcript from dalDK would theoretically yield the oligomer 5'-AUCAACUCAA-3', which shows some similarity to the sequence CAAUCAA found at the 3' end of many rho-terminated messengers. An exact match for this heptamer does actually occur further downstream and is formed from the 2nd, 3rd and 4th codons of the dalD gene itself, viz.

Met Asn Asn Gln .....

AUG AAC AAU CAA .....

but it appears not to function in vitro.

If, as already suggested, several sites exist for the initiation of mRNA syntheis from the dalDK promoter, then CRP-binding might influence these too. For example, in the E.coli galactose operon two interspersed promoters, S_1 and S_2, control expression (de Crombrugghe and Pastan, 1978). Both CRP and cAMP are needed for transcription from S_1, but together they abolish all transcription from S_2. It is possible that binding of CRP to dalDK around -40bp promotes transcription primarily from the position designated +1, but that very low levels of initiation occur at this and other sites during growth on glucose + d-arabitol. These possibilities await further investigation.

Linked to observations that CRP acts as a DNA-denaturing agent are the findings of Willmund and Kneser (1973) that, for the E.coli gal operon, binding of this factor defines a "storage

stretch" of up to 6 polymerase molecules and that in the absence of the protein or cAMP only one molecule can be bound in a stable, heparin-resistant complex capable of initiating transcription. One might generalise and propose that CRP allows more RNA Polymerase molecules to bind at the promoter of any catabolite-sensitive operon by destabilising and opening the DNA helix. This idea is especially attractive in the case of the dal operon due to the existence of a number of possible recognition sequences in extremely A+T rich regions.

Regulatory signals at the intercistronic junction of dalDK mRNA

The intercistronic region of the dalDK operon DNA shows a number of interesting features which may be of importance in regulating transcription or translation of the mRNA. Transcriptional control might involve pausing and/or partial termination of RNA synthesis resulting from specific signals contained in this region. In vitro transcriptions are not a viable proposition for answering these questions, since the behaviour of RNA Polymerase in in vitro systems is known to vary in comparison to its action in vivo (Bertrand et al, 1976; Lee et al, 1976; Wu and Platt, 1978; Wu et al, 1981).

Those structures commonly associated with the termination of RNA synthesis have been described in Chapter 5. The nucleotide sequence at the 3' end of the ArDH gene is extremely G+C rich, and its transcript is capable of forming a very stable hairpin (ΔG = -21kCal/mol. See Fig. 5m) characteristic of rho-independent terminators; however, this would normally be followed by a string of 4-6 uridine residues, whereas only 3 consecutive uridines are found, and these are 6bp downstream. This would probably not constitute an efficient termination signal. More-

over, termination at this site would result in the loss of the five C-terminal amino acids of ArDH, four of which constitute the tetrapeptide sequence Tyr-Thr-Leu-Ile identified by mass spectrometry in elastase digests of the purified enzyme.

Whereas rho-independent termination occurs about 20bp downstream of a dyad symmetry element, rho-mediated termination does not require a strikingly good hairpin or a string of U residues, and may take place at any number of positions within an A+T rich region which is preceded by a G+C rich sequence (Adhya and Gottesman, 1978). A sequence having homology to the consensus CAAUCAA is normally found at the 3' end of rho-terminated messages, but is not present in the vicinity of the dalD/dalK junction. Cytosine rich ribopolymers make excellent substrates for the RNA-dependent ATPase action of rho protein, suggesting a preferential interaciton with the G+C rich region which usually precedes a termination site (Lowery and Richardson, 1977). The involvement of rho as a modulator of dal transcription cannot be completely ruled out. The behaviour of RNA Polymerase on templates lacking a string of T residues (such as strict rho-dependent terminators and sites of termination responsible for polarity) is thought to require additional factors (Christie et al, 1981). The precise nature of such factors and their role in the cessation of RNA synthesis remains undetermined. Some information regarding the possible involvement of rho factor might be gained by transferring the d-arabitol operon on pRD351 into a rho^- host and comparing the relative levels of dehydrogenase and kinase produced. At the moment we can only speculate that the stem and loop structures at the 3' end of the dalD transcript can cause pausing of RNA Polymerase and partial termination.

Aside from transcriptional controls, several other

mechanisms can be involved in polarity effects. Messenger degradation models exist for a number of operon systems. Random fragmentation at limited sites is proposed for lac and gal mRNAs (Achord and Kennell, 1974; Blundell and Kennell, 1974), where each intercistronic region, or site nearby, is subject to cleavage from the time of synthesis. The best hypothesis for the disappearance of trp mRNA is sequential exonucleolytic degradation with "hold-up" points at or near the intercistronic boundaries (Schlessinger et al, 1977). Sequences around the dalD/dalK junction might function in such processes as these. Structural differences in mRNAs cause different stabilities (half-lives) and may also influence their affinity for ribosomes. This may take the form of variations in the stability of the initiation complex - a phenomenon that can be investigated using agents such as kasugamycin which inhibit the initiation of protein synthesis (Hirasima et al, 1973). A "stable" mRNA might be expected to have a specific structure or base sequence rendering it resistant to degradative nucleases. Do the hairpins in the dalDK mRNA represent this type of structure, allowing the DXK gene to be degraded quite rapidly, but protecting the ArDH gene from 3' attack, and thus permitting more frequent translation of the latter?

Translational control is another alternative. The precise mechanism by which a ribosome translates a polycistronic mRNA is still unclear, but there are three possibilities. A ribosome may traverse the entire messenger or new ribosomes may initiate at each cistron. Alternatively, there may be readthrough into distal genes at <100% efficiency. Experiments by Zalkin et al (1974) showed that in the presence of kasugamycin ribosomes cannot travel the entire length of trp mRNA, but cross into

distal genes at less than 60-70% efficiency. Stable secondary structures in the mRNA at intercistronic boundaries could conceivably interfere with translation and cause a proportion of the ribosomes to dissociate. On the dalDK mRNA the unusual clustering of termination and initiation codons together with stable hairpins and two ribosome-binding sites is interesting in this respect. The hairpins might serve to regulate the rate of translation across this region of biological importance. The first SD sequence (Fig. 5m) is part of a small hairpin, and is preceded by UAA and UGA triplets (see Atkins, 1979). The synthesis of a highly charged 12 amino acid peptide from the downstream AUG codon could easily occur and ought to be verifiable by in vitro translations.

Hairpin loops at intercistronic regions are also characteristic of RNAase III processing. This was first discovered by Young and Steitz (1978) in the ribosomal RNA operons rrnD* and rrnX. This enzyme is specific for dsRNA (Robertson et al, 1968), and, although the exact mechanism of site selection is unclear, it appears to cleave mismatched pairs within an RNA hairpin. The hairpin need not be very stable and can have few G·C pairs and a high proportion of U·G matches. Work by Barry et al (1980) has shown that RNAase III processing of the E.coli rplJL-rpoBC transcript occurs at a stem and loop structure closely associated with a SD sequence and a number of translational start/stop signals similar to those found between the ArDH and DXK genes. Further similarities are found with the RNAase III site between phage T7 genes 0.7 and 1.0 (McConnell, 1979). This cleavage point follows a ribosome-binding site, and just to the 3' side there are start and stop codons. McConnell speculates that f.Met tRNA may be

involved in an interaction between ribosomes and the RNA, and that a short peptide is synthesised: a situation mirrored by dalDK mRNA? RNAase III processing also occurs in the E.coli his operon. A 180bp sequence in the hisJ/hisQ junction forms a stable (ΔG = -54 kCal/mol) hairpin with several mismatches. Despite this structure, no termination of transcription is seen in vivo or in vitro and, as a result of processing, J protein outnumbers Q protein by a factor of ten (Higgins et al, 1982). The juxtaposition of sequences in the dalDK intercistronic region suggests that functional interactions between transcriptional and translational machinery may occur during the expression of this operon. In the case of the trp operon there is convincing in vivo evidence that translation can modulate the frequency of transcription termination at the attenuator (Zurawski et al, 1978). Interaction of ribosomes at a site preceding the large stable hairpin between the ArDH and DXK coding sequences could serve to modulate transcription termination somewhere in this region, but much more work is obviously required to discover whether transcription, translation, mRNA degradation or nucleolytic processing is responsible for the variations in enzyme levels observed in vivo.

Several recent discoveries have also pointed to nucleic acid secondary structures as an additional dimension in gene expression (Wells et al, 1980; Cantor, 1981). As potential binding sites for proteins, such structures have been proposed to participate in the regulation of replication, transcription and RNA-processing. The intercistronic regions of ØX174 and SV40 comprise only 4% and 14% of the genome respectively, but contain 20-40% of all major hairpins (Müller and Fitch, 1982).

The predominance of hairpins at these positions supports speculations that they have evolved as regulatory loci which exert their influence via RNA (or DNA) secondary, rather than primary, structure.

Expression of the dal repressor gene

Kelley and Yanofsky (1982) have shown recently that the trpR promoter performs at about 10% of the efficiency of the lac promoter in trpR-lacZ fusions. This is a very high value for a gene whose product is required in only small amounts, and it is apparently essential since the mRNA lacks any SD sequence at its 5' end and is translated very badly in vivo. The dalR promoter, based on nucleotide sequence data alone, would also appear to represent quite a good initiator of transcription. Unlike the low-level constitutive lacI promoter, dalR possesses an excellent -35 region and a good Pribnow box (Fig. 7d), a far better fit with the consensus sequence than even trpR. Experiments to test the efficiency of this promoter in vivo are planned for the future.

The trpR operon is regulated autogenously (Gunsalus and Yanofsky, 1980). The repressor binds to a trpR operator which has considerable DNA homology to both the trp and aroH promoters. The operator for the dalDK operon has not yet been characterised beyond doubt, but no sequences in common with the dalR promoter region can be identified. No strings homologous to the three 17bp repeats of dalDK exist in dalR, and there is no other noticeable similarity between the two promoters at the DNA level (throughout the region -1 to -45 the dalDK promoter is 58% A+T rich compared to only 38% for dalR). It therefore seems unlikely that DalR protein regulates transcription from its own promoter.

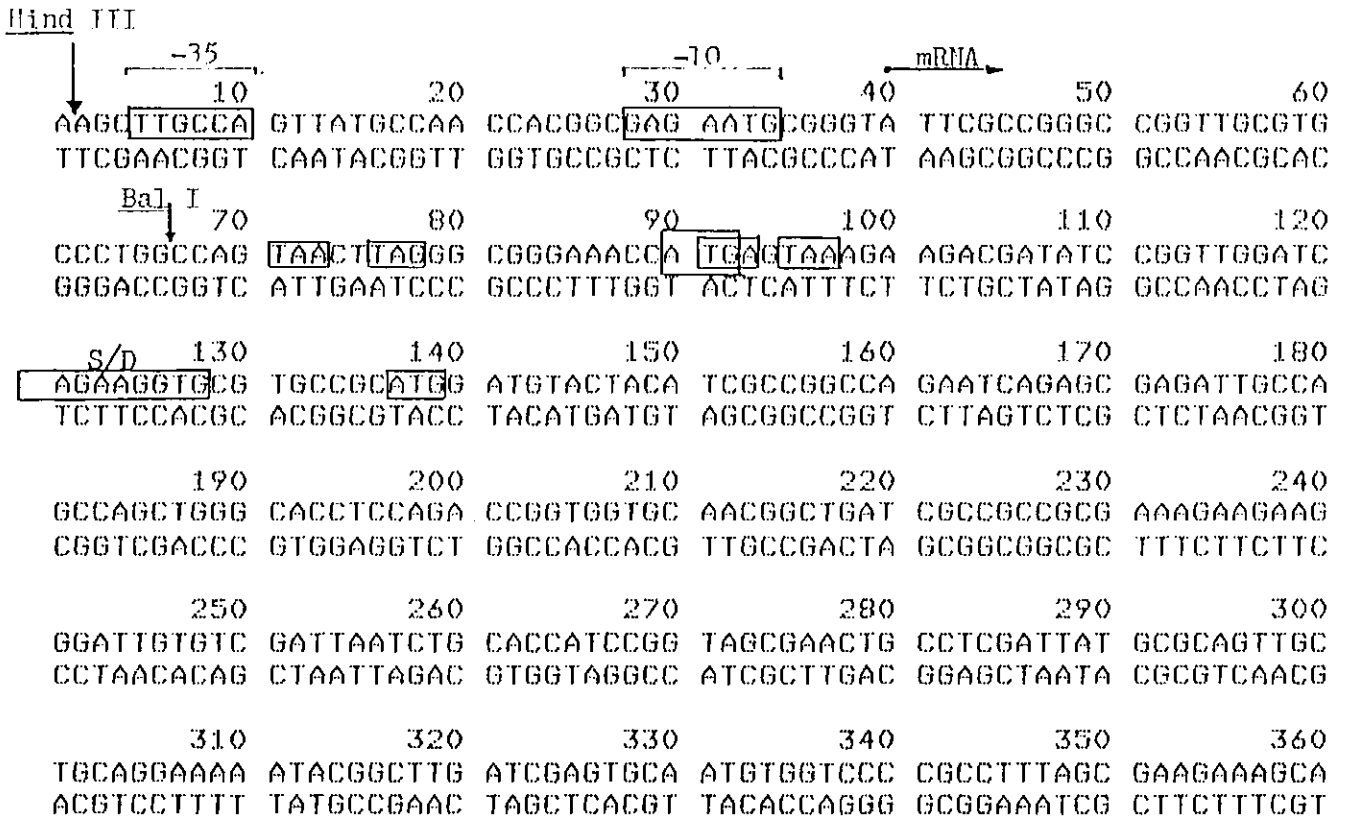


Fig. 7d The DNA sequence of the dalR promoter region.

Potential -35 and -10 sequences are indicated. Translational initiation and termination codons are boxed, together with the most obvious Shine/Dalgarno sequence.

An imperfect palindrome extends across the -10 region and the transcription start point of the repressor gene (Fig. 7e), but its significance, if any, is unknown. No sequence with homology to this dyad symmetry can be found in the dalDK promoter or the first 100 bases of the mRNA. Is there some advantage in having a good promoter for the dal repressor? Several reasons can be envisaged. The repressor itself may be unstable and susceptible to proteolysis, necessitating a higher level of expression to maintain adequate repression of the dalDK operon. Secondly, the dalR messenger might have a very short half-life or be translated inefficiently for some other reason. Finally, the promoter might have to compete for RNA Polymerase. The very close proximity of the dalDK promoter and its putative multiple polymerase-binding sites makes this an interesting possibility and, in fact, the -35 region of dalR is overlapped by the most upstream of these sites. Position -35 of the dalR promoter is equivalent to -100 of the dalDK promoter. Even when the d-arabitol operon is uninduced, stacking of polymerase molecules at these sites could interfere with transcription of the repressor gene, which must then have a good -35 region of its own in order to be transcribed frequently enough to maintain intracellular levels of repressor protein. The control of dalR expression could also be regulated like trpR at the translational level by means of sequences or secondary structures within the 5' leader of the mRNA, an idea that is discussed later.

The Recognition of Base Sequences by Regulatory Proteins

E.coli RNA Polymerase can protect 70bp (23nm) of DNA from attack by DNAase I (Schmitz and Galas, 1979) and makes actual contact with the template over a range of 43bp or 14.6nm (Siebenlist et al, 1980). A protein of the size of the dal repressor ($\approx$20,000 daltons) will have a molecular diameter of

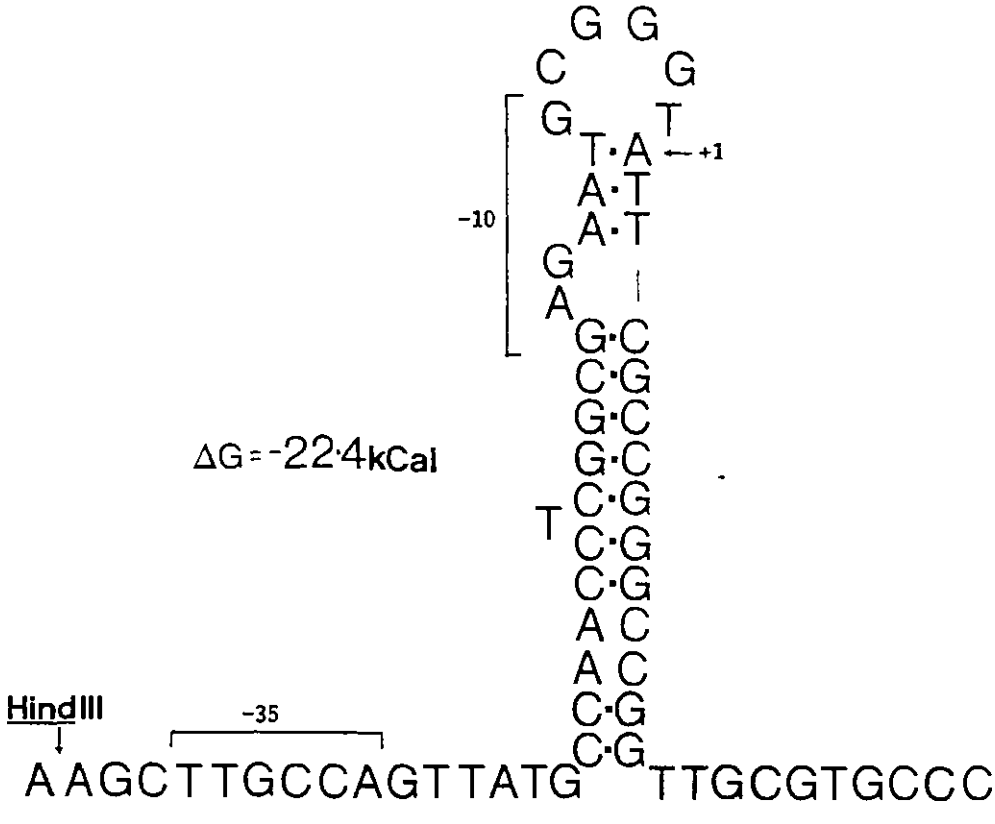


Fig. 7e An imperfect palindrome in the dalR promoter. The structure is represented as a DNA hairpin and encloses both the Pribnow box and origin of transcription.

5-6nm if globular, and should therefore be able to protect a minimum of 20bp of DNA. Binding to sequences within the left half of the three 17bp dalDK promoter repeats could effectively prevent interaction of RNA Polymerase at all five "recognition sequences" by steric hindrance. O'Neill (1976) proposes that a 6bp sequence in the left half of lac and λ operators constitute the recognition sequences for their repressors. All or part of the sequence $TTGCTC(T)_n$ which is conserved within the three repeats (Fig. 5k) may be responsible for dal repressor binding. This can only be investigated once the protein has been isolated and partially purified. Repressor interactions at the other potential dalDK operator spanning position +1 would prevent the initiation of transcription.

A number of repressors are known to bind as oligomers (Pirrotta et al, 1970; Johnson et al, 1981) at operators which display an element of dyad symmetry. The spatial arrangement is such that the sequence and its inverse both appear on the same side of the helix, yet in lac and phage λ the major contact points suggest that neither repressor fully exploits the symmetrical features of the operator (Miller, 1978). If the three repeated sequences in the dalDK promoter are operator sites, then the absence of any inverse symmetry may reflect the ability of the dal repressor to recognise a linear sequence or that it binds as a monomer. Both of these ideas contradict the bulk of data derived from studies on other prokaryote repressor proteins and point in favour of the sequence AGTGG N_8 ACTCA at position +1 as the true operator.

Experiments aimed at probing contact points between sequence-specific proteins and DNA reveal that regulatory proteins like CRP (Simpson, 1980) and lac repressor (Barkley and Bourgeois, 1978) recognise only one side of the helix. Here it is interesting

to note that the spacing of the dal promoter repeats would allow for a similar mode of interaction (Fig. 7c). In E.coli the DNA appears to be wrapped around a core comprised of DNA-binding proteins (Hélène, 1981). This may explain why regulatory proteins must recognise only one face of the helix as the other remains hidden. Phasing of operator sequences then becomes essential in order that protein contact sites are exposed.

How do the dalDK and dalR messenger RNAs relate to current models of translational control?

Synthesis of a protein to high levels in E.coli depends upon efficient transcription from a "strong" promoter and requires also that the mRNA bears a ribosome binding site in close proximity to the initiator AUG (or GUG). Current models suggest that such a site consists of the translational start codon itself and another sequence that is complementary to the 3' end of the 16S rRNA (Shine and Dalgarno, 1974). SD sequences have been found in almost all E.coli messages sequenced to date. They appear to vary greatly in size (3-9 bases) and precede the initiator codon by between 3 and 12 bases (for a review, see Steitz, 1979). More recently it has been suggested that there may be general requirement for the mRNA to adopt a suitable secondary structure which exposes the SD sequence to an incoming ribosome and facilitates the formation of a stable initiation complex (Bahramian, 1980; Hall et al, 1982). Although every AUG triplet is potentially capable of initiating protein synthesis, most are unable to do so because they lack essential structural features upstream, or because they are sequestered in double-stranded regions (Platt et al, 1976; Dunn et al, 1978).

What constitutes an efficient initiator region? Experimental evidence (Hall et al, 1982) suggests that a small hairpin

located immediately 5' to the AUG triplet is important. So, too, is sequence complementarity between the 5' shoulder of the hairpin and the 3' end of 16S rRNA. If the hairpin is excessively large, then a second SD sequence is commonly located in a non hydrogen-bonded loop within it. Atkins (1979) also implicates the involvement of the termination signals UAA and UGA upstream of the initiator AUG. Ganoza et al (1978) propose that a protein is involved in the recognition of these signals, and may form part of the initiation complex. There is also evidence that secondary and tertiary structures in mRNAs negatively influence ribosomal recognition of initiation sites (Lodish and Robertson, 1969; Steitz, 1973). This may be due to competition with genuine initiation hairpins.

Let us now look at the two dal operon transcripts in the light of this information. Translation of ArDH is initiated at the AUG triplet 8 bases downstream of the hexamer sequence 5'-AAGGAG-3' near the 5' terminus of the messenger. Transcription initiation 6 bases away from the Pribnow box (TACAGTG) allows for the formation of a small RNA hairpin carrying this SD sequence in a non hydrogen-bonded loop (Fig. 5j). The termination codon UGA overlaps the initiator triplet, and 8 nucleotides separate the SD sequence from the AUG codon. Experiments aimed at optimising expression of cloned genes (Backmann and Ptashne, 1978; Guarente et al, 1980) have examined the effect of "moving" the ribosome-binding site relative to AUG codons and show that a spacer sequence of around 8 bases gives maximal translation. These are all factors which suggest that dalDK mRNA ought to be translated efficiently. Once induced, it is important that the products of an operon are made rapidly; these signals may help to ensure that this objective is achieved.

The translation of dalR message may be a far more complex event. The putative transcription start for this mRNA is 50 bases prior to the first AUG, and 101 bases upstream of a second in-phase AUG. The whole of this leader sequence is able to fold into two stable hairpin structures (Fig. 7f) which, it is believed, may play an important role in expression. The first 32 nucleotides of the messenger constitute a hairpin having a large negative free energy of formation (ΔG = -23.4 kCal, see Tinoco et al, 1973). The G+C content of this structure is greater than 75%. The second hairpin is larger and contains two internal loops, one of which carries a string of 10 bases homologous to the 3' end of 16S rRNA. This SD sequence lies 11 residues in front of the second of the two AUG triplets. The first AUG is hydrogen bonded in the stem of this hairpin, and possesses no recognisable SD sequence immediately upstream. Applying the "rules" of translational initiation described earlier, it is proposed that the repressor is translated at the second AUG triplet, adjacent to the hairpin carrying the unbonded SD sequence. The other initiator codon, being sequestered in a double stranded region, is deemed to be unavailable, although final confirmation must wait until the N-terminal protein sequence is obtained.

I have already mentioned that there is evidence that hairpins negatively influence ribosome-binding, and it is interesting to speculate upon the possible involvement of the G+C rich structure at the 5' end of dal mRNA in this context.

The trpR messenger has a leader sequence of 56 bases and, despite a fairly high rate of transcription, it is translated very poorly since it lacks any SD sequence. Near the 5' end of the molecule, there is a G+C rich dyad symmetry element

Fig 7f

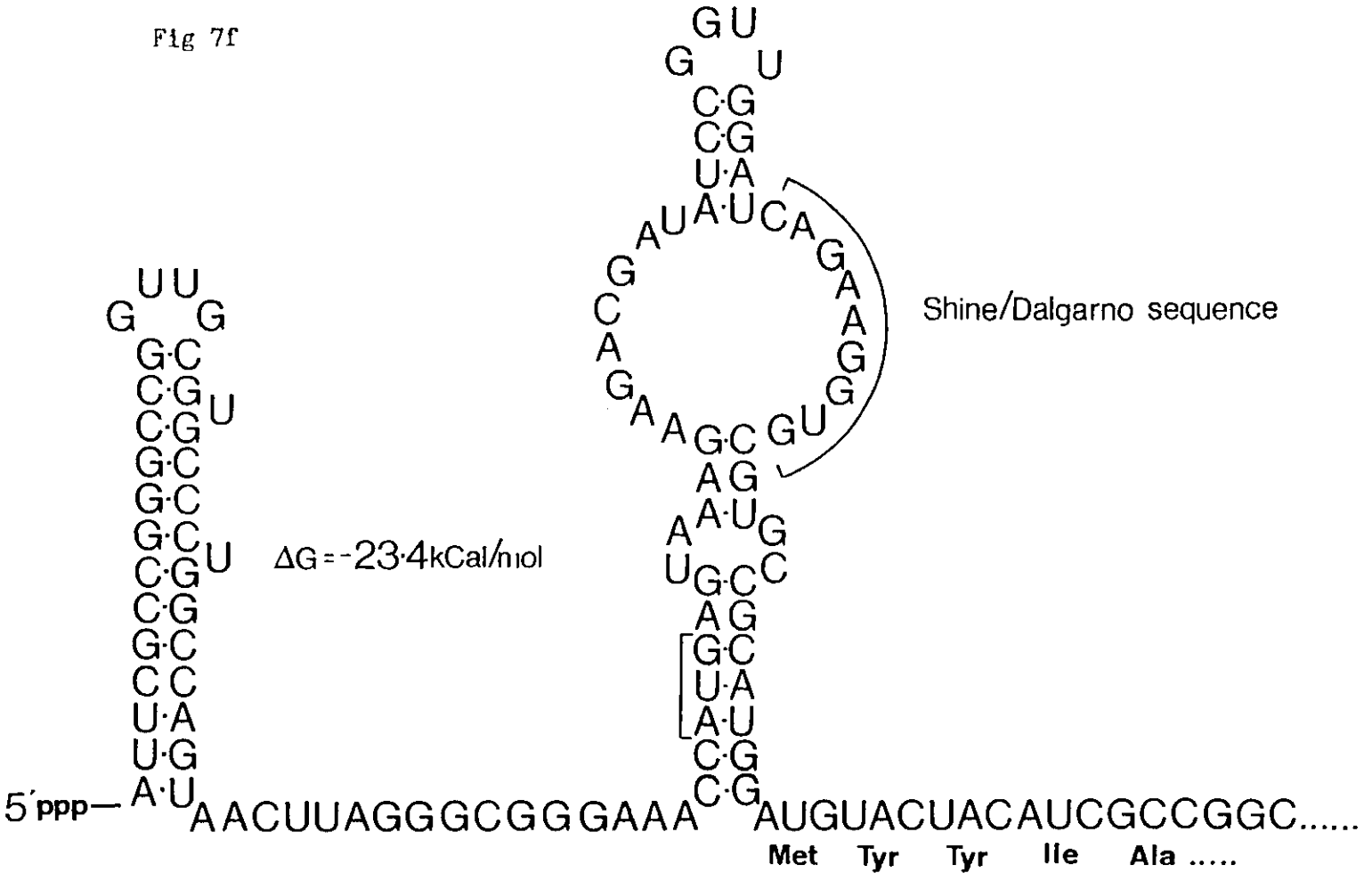

Shine/Dalgarno sequence
ΔG = -23·4kCal/mol
5'ppp— AACUUAGGGCGGGAAA AUGUACUACAUCGCCGGC.....
Met Tyr Tyr Ile Ala .....

Fig. 7f

Secondary structures at the 5' end of dalR mRNA.The proposed ribosome-binding site and N-terminal amino acid sequence of DalR are shown.An alternative AUG initiator codon occurs upstream but is sequestered in the stem of the second hairpin and lacks a suitably positioned Shine/Dalgarno sequence.

(ΔG = -26kCal/mol) which has been observed to give terminated transcripts *in vitro*. The similarity to *dalR* is obvious. Both operons seem to have good promoters and in each case the long leader sequences are capable of folding into G+C rich hairpins which, in *trpR* at least, may serve to terminate transcription. The *dalR* message would appear to possess an excellent SD sequence and initiation hairpin, and more investigation will be required to assess the role, if any, of the terminal stem and loop structure.

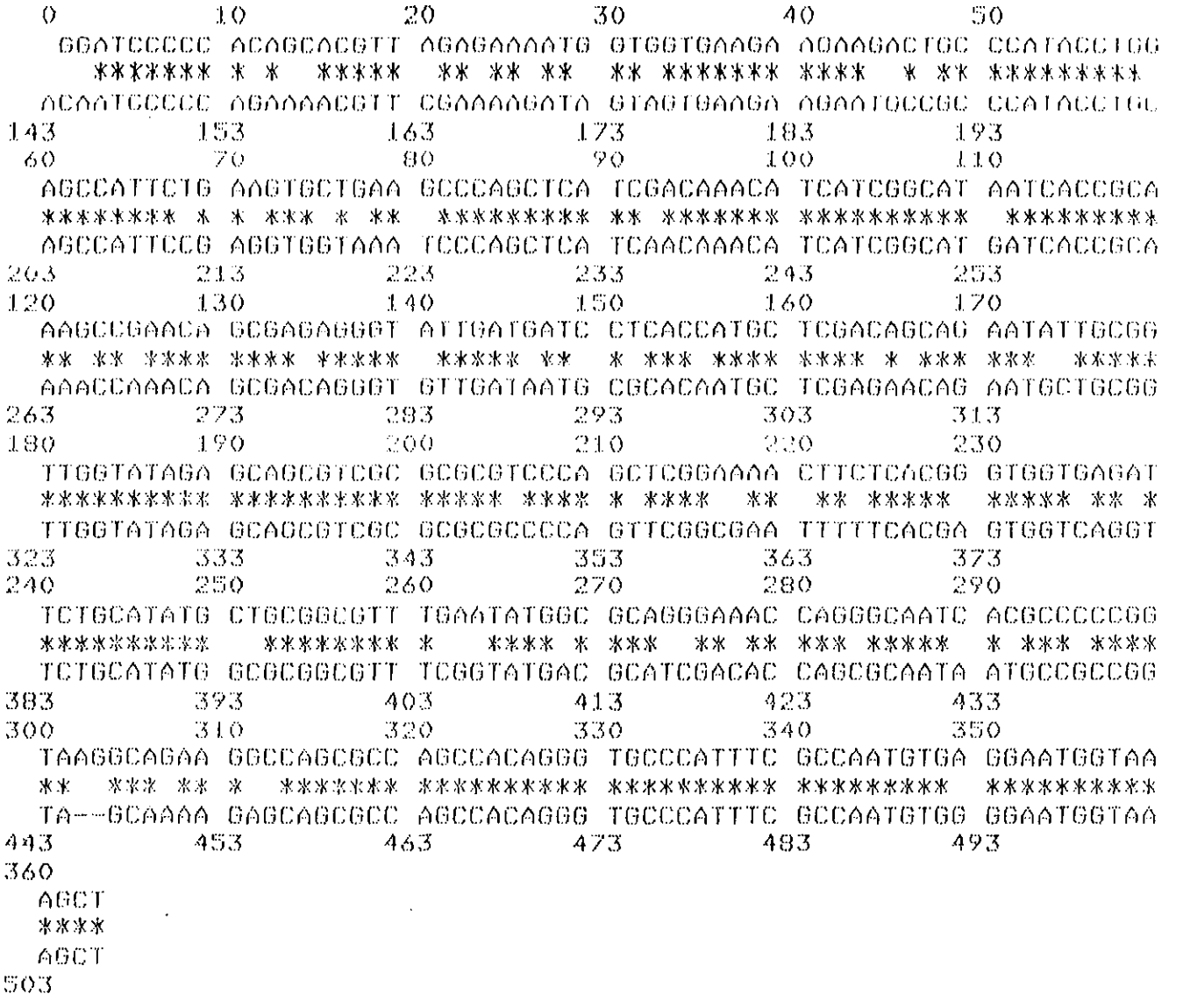
The Evolution of the Pentitol Operons

The structural arrangement of the pentitol operons promotes speculation regarding their evolutionary origins. The two loci, *rbtR* and *dalR*, encode repressor proteins, and it has been demonstrated here that the genes are not interdigitated, contrary to earlier beliefs (Charnetzky and Mortlock, 1974c). Mirror imagery of the *rbt* and *dal* operons is complete, and an evolutionary origin based on gene duplication and inversion is therefore attractive. Events of this nature render the duplicate copy immune to excision through recombination and allow time for significant sequence divergence. Despite its obvious attractions, this explanation would appear to be unsatisfactory. Previous studies on the pentitol dehydrogenases have revealed wide differences in quaternary structure, subunit size and kinetic parameters. In addition, there is no immunological cross-reaction, thus ruling out extensive structural similarities. Hybridisations aimed at detecting DNA homologies within the operons also proved negative (Neuberger and Hartley, 1979). The DNA and amino acid sequences presented here prove conclusively that there is no evolutionarily significant homology between the two dehydrogenases at either level. The control regions for *rbt* and *dal* transcription

differ considerably, and the overall situation has been further confused with the discovery that the dal repressor and structural genes are transcribed in a bipolar fashion. The probability of common ancestry for these operons is thus diminished, and the idea of a recent, simple duplication/inversion event cannot adequately account for the arrangement of genes on the K.aerogenes genome. The dissimilarity of the dehydrogenases does not preclude the possibility that the pentulokinases are related. Neuberger and Hartley (1979) demonstrated the existence of a sequence on λp rbt, which is repeated between BstD/BstE on λp rbt dal, and might represent similarities in the distal parts of the kinase coding sequences, or might fall just outside the transcribed regions of both operons. The exact nature and extent of these repeats was until very recently unknown, but it is now clear that inverted repeats of at least 0.5kb (and possibly as much as 1.5kb) are present beyond the kinase genes of both rbt and dal operons (Brannigan, Knott, Loviny and Hartley, unpublished data). Within those regions sequenced to date, there is 85% homology (Fig. 7g), and both copies show evidence of a long open reading frame. Irrespective of the imperfect nature of these repeats, the polypeptides which each potentially encodes are very similar, owing to the fact that most of the mismatches are concentrated in the third codon position. The degree of mismatching explains why early attempts at S1 mapping were fruitless (Neuberger and Hartley, 1979). The discovery that these sequences may code for proteins rekindles speculation about the existence of pentitol-specific permeases, since the coding strands are the same as those of the rbt and dal operons. Are these perhaps homologies between two permease genes? Certainly beyond the C-terminus of DXK there is another open reading frame, but as yet gaps still remain in the nucleotide

Fig. 7g

Homology between 360bp of sequence from the 0.99kb BstD/E region and a sequence just distal to the rbtK gene. The uppermost sequence begins at the BstE site and runs towards BstD, the other starts near KpnB and continues towards SalC. The two share 84% base sequence homology.



sequence, and it cannot be linked in-phase to the polypeptide coded by the dal repeat. Ribitol and arabitol permeases are not essential to growth on these substrates (Neuberger, 1978), but no evidence has ever been put forward to deny their existence. One might reasonably expect a high degree of amino acid homology between two such proteins.

Flanked by large inverted repeats, the pentitol operons could represent a type of "metabolic transposon", in which the functional units are neatly sandwiched between two repeated sequences by analogy with certain bacterial transposons (Review:- Cohen and Shapiro, 1980). It may be significant that this unit is absent from E.coli K and B strains, yet occurs in K.aerogenes and E.coli C at equivalent map positions. Lateral transfer between Enterobacteria is a distinct possibility, and the absence of the genes from most species implies that this may have been a relatively recent event. Cornelis et al (1978) have isolated, from Yersinia enterocolitica, a transposon-like sequence (Tn951) responsible for lactose fermentation. Tn951 carries lac i, z and y genes homologous to those of E.coli K12 and to lac genes from a wide variety of other bacteria. The DNA segment is 16.6kb long and flanked by identical 41bp inverted repeats from Tn3, but is defective in transposition. Complementation is possible by the tnpA product (a transposase) of Tn3 and transposition is then recA independent and temperature sensitive. In addition, Tn951 carries a single copy of IS1 which can cause deletions and inversions within the transposon itself (Cornelis, 1980; Cornelis and Saedler, 1980; Cornelis et al, 1981). Supported by more recent work on the sequences flanking the E.coli lac operon, the interspecies transfer of these genes now seems certain to have occurred. Lampel and Riley (1982) have

demonstrated that, although Salmonella typhimurium possesses DNA homologous to regions on either side of the E.coli lac operon, it has no sequences resembling the lac genes themselves, and so these must have been added to E.coli or lost from the S.typhimurium genome since their divergence from a common ancestor.

Link and Reiner (1982) recently published data showing that the E.coli C rbt and dal genes are surrounded by 1.4kb inverted repeats of imperfect homology and suggest that these too are vestigial transposable elements, but as no transpositions of these genes could be detected, the authors propose that the Tn elements have been rendered non-functional by sequence divergence. This hypothesis will soon be tested since the DNA sequences of the K.aerogenes repeats are nearing completion. At present the available data do not correlate with any of the published Tn or IS sequences; nevertheless the possibility still remains of a distant interspecies transfer of both the lac and rbt/dal operons.

Roughly 18% of the E.coli chromosome is organised into units of 22kb or 27.5kb ± 1.5kb flanked by inverted repetitious sequences (Chow, 1977). Some of these correspond to known IS elements, but others, like the pentitol repeats, are of unknown nature. Such DNA segments may have the potential to translocate like Tn5 or Tn10 or to invert like MuG or P1G. Alternatively, this arrangement may reflect the evolution of the bacterial chromosome from DNA pooled from donor ancestors by translocation. The wide variety of translocatable elements, their common occurrence and their ability to mediate recombination and rearrangement of non-homologous DNA, suggest that they may play a major role in bacterial evolution.

Assuming that the pentitol gene cluster is not a result

of simple duplication and inversion, other models must be devised to explain their close proximity in the genome. One can envisage a translocation or inversion event which brings together two regions of DNA, each of which code for a dehydrogenase and repressor. This might then be succeeded by recruitment of a kinase from elsewhere on the genome, and subsequent divergence of the two operons. This hypothesis might be favoured if any significant homology is found between the two pentitol operon kinases. An alternative DXK exists in the d-xylose operon, and the E.coli enzyme has now been cloned in this laboratory (K. Briggs, unpublished data). It will be of great interest to compare this protein with both DRK and DXK.

The compactness of the dalDK operon intercistronic region, with its intricate lay-out of hairpins, stop/start codons, SD sequences and only 11bp separating the genes is inconsistent with a chance translocation event of the type proposed above and perhaps argues against this operon being a product of gene shuffling. An ancestral repressor/RDH unit positioned adjacent to an already intact dal operon through a fortuitous genomic rearrangement might subsequently have evolved into the rbt operon via recruitment and adaptation of the dal DXK gene.

Other possible origins for the pentitol catabolic unit include a recombination mechanism involving two independent crossover events (Fig. 7h) occurring between short DNA homologies flanking two dissimilar ancestral operons. This theory does not invoke gene duplication and seems the most likely explanation at present. Such short homologies would not be detectable by hybridisation studies or electron microscopy and would probably have been lost or eroded due to sequence divergence since the original event took place. It is therefore almost impossible to

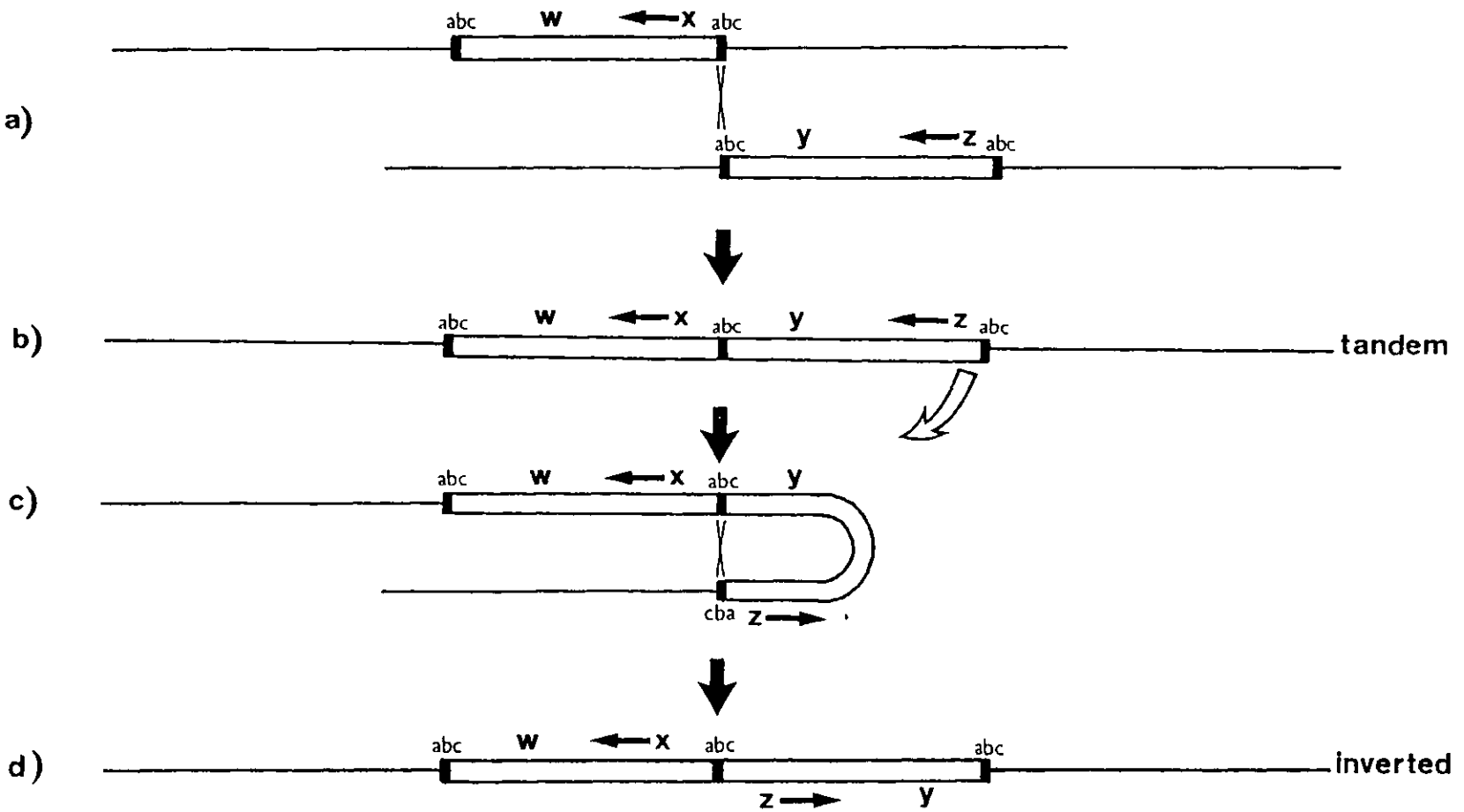


Fig. 7h Two independent crossover events between short homologous sequences leading to the establishment of a divergent operon "unit".

substantiate, but may be the only practical explanation if the pentitol kinases turn out to be as different as the dehydrogenases are. Edland and Normark (1981) have successfully demonstrated that tandem duplications can arise by crossovers between very short (10-12bp) regions of DNA sequence homology. This process is considered to be recA independent. Charlier et al (1979) point out that such a crossover is most likely to occur between two arms of a replication fork rather than two chromosomes.

From the situation three years ago when it seemed that the pentitol genes represented a unique opportunity to study an example of a duplication event involving an entire operon system, our understanding has progressed to a point where it now becomes apparent that the ribitol and d-arabitol operons of K. aerogenes might have totally separate genetic backgrounds. Despite this disappointment, the pentitol operons have always proved a very interesting area of study and much more work still needs to be carried out before the file can be closed on what has become affectionately known as "The RDH Project". Work on the sequencing of the ribitol repressor gene and the two pentulokinase genes is already nearing completion, leaving only the large, invert repeat sequences to be studied. Approximately 90% of the sequence of the d-arabitol element of this repeat is already known.

Experiments are now in progress to place the dalR gene under the control of the λp_L promoter in pPLc236 (Remaut et al, 1981), with the aim of producing sufficient quantities of the repressor for operator-binding studies. The DalR recognition sequence needs to be identified. Protection of regions of the dalDK promoter from attack by DNAase I or dimethyl sulphate will identify the operator site. It is expected that the operator will span the putative origin of transcription, but

the latter has not yet been conclusively identified, a situation which must be rectified by RNA sequencing of the 5' end of dal mRNA.

When purified DalR protein becomes available, the N-terminal sequence will reveal whether translation initiates at the second of the two AUG codons, as proposed in this work.

What is the role of the G+C rich hairpin at the 5' end of dalR mRNA? We have the trp operon promoter available on a 42bp Pvu II fragment (Russell and Bennett, 1982) which can be inserted at sites in the dalR promoter or at the unique Bal I site which forms the 3' end of the hairpin (see Fig. 7d). Any differences in the levels of dalR expression may tell us something of the importance of this structure.

Finally, one of the prime aims of this part of the project will be to attempt crystallisation of the repressor for X-ray diffraction analysis.

In summary, although the work presented here has gone a long way towards answering some of the questions surrounding the origins of the pentitol catabolic unit, we cannot yet state, with confidence, exactly how these genes have evolved to reach their current state, but it is hoped that this will soon become possible, once the precise nature of the repeated sequences which flank the operons and the non-coding sequences which separate them are known.

REFERENCES

Achord D. and D.Kennell (1974), J.Mol.Biol. 90 581-599

Adhya S. and M.Gottesman (1978), Ann.Rev.Biochem. 47 967-996

Aiba H. and J.S.Krakow (1981), Biochemistry(Washington) 20 4774-4780

Altosaar I. and B.S.Hartley (1976), Proc.Int.Congress Biochem. 10 200

Anderson S., M.J.Gait , L.Mayol and I.Young (1980), Nuc.Acids Res. 8 1731-1743

Anderson W.B., A.B. Schneider, M.Emmer, R.L. Perlman and I.Pastan (1971), J.Biol.Chem. 246 5929-5937

Anrews K.J. and G.D. Hegeman (1972), J.Mol.Evol. 8 317-328

Ansorge W. and L de Maeyer (1980), J.Chromatography 202 45-53

Atkins J.F. (1979), Nuc. Acids Res. 7 1035-1041

Bachmann B.J. and K.Brooks-Low (1980), Microbiology Rev. 44 1-56

Backmann K. and M.Ptashne (1978), Cell 13 65-71

Bahramian M.B. (1980), J.Theoret. Biol. 84 103-118

Bahramian M.B. and B.S.Hartley (1980), Eur. J.Biochem. 110 507-519

Bailone A. and F.Galibert (1980), Nuc. Acids Res. 8 2147-2164

Barnes W.M. (1977), Science 195 393-395

Barker W.C., L.K.Ketcham and M.O.Dayhoff (1978), in "The Atlas of Protein Sequence and Structure" 5 (ed. M.O.Dayhoff), National Biomedical Research Foundation, Washington, N.Y.

Barkley M.D. and S.Bourgeois (1978), p177-220 in "The Operon", (eds. J.H.Miller and W.S.Reznikoff), Cold Spring Harbor Labs., N.Y.

Barry G., C.Squires and C.L.Squires (1980), Proc.Nat.Acad.Sci. 77 3331-3335

Beckwith J.R. and E.R.Signer (1966), J.Mol.Biol. 19 254

Bennett C.D. (1974), Nature 248 67-68

Bennett G.N., K.D.Brown and C.Yanofsky (1978), J.Mol.Biol. 121 139-152

Berger E.M. (1978), J.Mol.Evol. 10 319-323

Bertrand K., C.Squires and C.Yanofsky (1976), J.Mol.Biol. 103 319-337

Beyreuther K. (1978), p128-154 in "The Operon", (eds. J.H.Miller and W.S.Reznikoff), Cold Spring Harbor Labs., N.Y.

Bishop J.O. and M.Robash (1973), Nature New Biology 241 204-207

Blundell M. and D.Kennell (1974), J.Mol.Biol. 83 143-161

de Boer H.A., S.F.Gilbert and M.Nomura (1979), Cell 17 201-209

Bolivar F., R.L.Rodriguez, P.J.Greene, M.C.Betlach, H.L.Heyneker, H.W.Boyer (1977), Gene 2 95-113

Bollum F.J. (1959), J.Biol.Chem. 234 2733-2734

Bradford M.M. (1976), Anal.Biochem. 72 248-254

Brändén C-I., H.Eklund, B.Nordström, T.Boiwe, G.Söderlund, E.Zeppezauer, I.Ohlsson and Å.Åkeson (1973), Proc.Nat.Acad.Sci. 70 2439-2442

Brosius J., T.J.Dull, D.D.Sleeter and H.F.Noller (1981), J.Mol.Biol. 148 107-127

Brown D.D. and I.B.Dawid (1968), Science 160 272-280

Brown J., D.Brown and I.Zabin (1967), J.Biol.Chem. 242 4254-4258

Burleigh B.D., P.W.J.Rigby and B.S.Hartley (1974), Biochem. J. 143 341-352

Calos M.P. (1978), Nature 274 762-765

Calos M.P., D.Galas and J.H.Miller (1978b), J.Mol.Biol. 126 865-869

Calos M.P., L.Johnsrud and J.H.Miller (1978), Cell 13 411-418

Cantor C.R. (1981), Cell 25 293-295

Chamberlin M.J. (1976), p159-191 in "RNA Polymerase", (ed. R.Losick), Cold Spring Harbor Labs., N.Y.

Charlier D., M.Crabeel, R.Gunin and N.Glansdorf (1979), Mol. Gen. Genet. 174 75-88

Charnetzky W.T. and R.P.Mortlock (1974a), J.Bacteriol. 119 162-169

Charnetzky W.T. and R.P.Mortlock (1974b), J.Bacteriol. 119 170-175

Charnetzky W.T. and R.P.Mortlock (1974c), J.Bacteriol. 119 176-182

Chou J., M.Casadaban, P.Lemaux and S.N.Cohen (1979), Proc.Nat.Acad.Sci. 76 4020-4024

Chou P.Y. and G.D.Fasman (1974), Biochemistry 13 222-245

Chow L.T. (1977), p73-79 in "DNA Insertion Elements,Plasmids and Episomes", (eds. A.I.Bukhari, J.A.Shapiro and S.C.Adhya), Cold Spring Harbor Labs.

Christie G.E., P.J.Farnham and T.Platt (1981), Proc.Nat.Acad.Sci. 78 4180-4184

Clarke C.M. and B.S.Hartley (1979), Biochem.J. 177 49-62

Clegg J.B. (1970), Proc. Royal Society London,(series B), 176 235-246

Clewell D.B. (1972), J.Bacteriol. 110 667-676

Cohen S.N., A.C.Chang and L.Hsu (1972), Proc.Nat.Acad.Sci. 69 2110-2114

Cohen S.N., A.C.Chang, H.W.Boyer and R.B.Helling (1973), Proc.Nat.Acad. Sci. 70 3240-3244

Cohen S.N. and J.A.Shapiro (1980), Scientific American 242 36-45

Cornelis G., D.Ghosal and H.Saedler (1978), Mol.Gen.Genet. 160 215-224

Cornelis G. (1980), J.Gen.Microbiol. 117 243-247

Cornelis G. and H.Saedler (1980), Mol.Gen.Genet. 178 367-374

Cornelis G., H.Sommer and H.Saedler (1981), Mol.Gen.Genet. 184 241

Crawford I.P., B.P.Nichols and C.Yanofsky (1980), J.Mol.Biol. 142 489-502

de Crombrugghe B., S.Adhya, M.Gottesman and I.Pastan (1973), Nature New Biology 241 260-264

de Crombrugghe B. and I.Pastan (1978), p303-334 in "The Operon",(eds. J.H.Miller and W.S.Reznikoff), Cold Spring Harbor Labs., N.Y.

Csordás-Tóth E., I,Boros and P.Venetianer (1979), Nuc.Acids Res. 7 2189-2197

Curtiss R., L.J.Charamella, C.M.Berg and P.E.Harris (1965), J.Bacteriol. 90 1238-1250

Dayhoff M.O. (1978), ed. Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington DC.

Dayhoff M.O. and W.C.Barker (1972), in Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington DC.

Dayhoff M.O., L.T.Hunt and S.H.Calderone (1978), in Atlas of Protein Sequence and Structure.

Dickerson R.E. (1972), Scientific American 226 58-72

Dickerson R.E. (1971), J.Mol.Evol. 1 26-45

Dickson R.C., J.N.Abelson, W.M.Barnes and W.S.Reznikoff (1975), Science 187 27-35

Dove W.F. and F.Davidson (1962), J.Mol.Biol. 5 467-478

Domingo E., D.Sabo, T.Taniguchi and C.Weissman (1978), Cell 13 735-744

Dothie J. (1974), Ph.D. Thesis, University of Cambridge.

Dunn J.J., E.Buzash-Pollert and F.W.Studier (1978), Proc.Nat.Acad.Sci. 75 2741-2745

Ebright R.H. and J.R.Wong (1981), Proc.Nat.Acad.Sci. 78 4011-4015

Edelman P. and J.Gallant (1977), Cell 10 131-137

Edland T. and S.Normark (1981), Nature 292 269-271

Eilen E. and J.S.Krakow (1977), J.Mol.Biol. 114 47-60

Elton R.A., G.J.Russell and J.H.Subak-Sharpe (1976), J.Mol.Evol. 8 117- 135

Engel P.C. (1973), Nature 241 118-120

Farabaugh P.J. (1978), Nature 274 765-769

Farabaugh P.J., U.Schmeissner, M.Hofer and J.H.Miller (1978), J.Mol.Biol. 126 847-863

Farnham P.J. and T.Platt (1981), Nuc.Acids Res. 9 563-567

Farnham P.J. and T.Platt (1982), Proc.Nat.Acad.Sci. 79 998-1002

Fiers W., R.Contreras, F.Duerinck, G.Haegeman, D.Iserentant, J.Merregaert, W.Jou, F.Molemans, A.Raeymaekers, A.Van de Berghe, G.Volckaert and M.Ysebaert (1976), Nature 260 500-507

Galas D.J. (1978), J.Mol.Biol. 126 858-863

Galas D.J. and A.Schmitz (1978), Nuc.Acids Res. 5 3157-3170

Galibert F., J.Sedat and E.Ziff (1974), J.Mol.Biol. 87 377-407

Ganoza M.C., A.R.Fraser and T.Neilson (1978), Biochemistry 17 2769-2775

Garaev M.M., A.F.Bobkov, A.F.Bobkova, V.N.Kalinin, V.D.Smirnov, Y.E.Khudakov and T.I.Tikchonenko (1982), Gene 18 21-28

Gardner J.F. (1979), Proc.Nat.Acad.Sci. 76 1706-1710

Gardner R.C., A.J.Howarth, P.Hahn, M.Brown-Lendi, R.J.Shepherd and J.Messing (1981), Nuc.Acids Res. 9 2871-2887

Gemmill R.M., S.R.Wessler, E.B.Keller and J.M.Calvo (1979), Proc.Nat. Acad.Sci. 76 4941-4945

Gicquel-Sanzey B. and P.Cossart (1982), EMBO Journal 1 591-595

Gilbert W. and B.Müller-Hill (1970),p93-110 in "The Lactose Operon", (eds. J.R.Beckwith and D.Zipser), Cold Spring Harbor Labs., N.Y.

Gilbert W., N.Maizels and A.Maxam (1974), CSH Symp. Quant. Biol. 38 845-855

Gilbert W (1976), p193-205 in "RNA Polymerase" ,(ed. R.Losick), Cold Spring Harbor Labs., N.Y.

Grantham R., C.Gautier, M.Gouy, R.Mercier and A.Pavé (1980), Nuc.Acids. Res. 8 r49-r62

Gralla J. and D.M.Crothers (1973), J.Mol.Biol. 78 301-319

Grindley N.D. (1978), Cell 13 419-426

Grönenborn B. and J.Messing (1978), Nature 272 375-377

Grosjean H. and W.Fiers (1982), Gene 18 199-209 .

Guarente L., G.Lauer, T.M.Roberts and M.Ptashne (1980), Cell 20 543-553

Gunsalus R.P. and C.Yanofsky (1980), Proc.Nat.Acad.Sci. 77 7117-7121

Hall M.N., J.Gabay, M.Débarbouillé and M.Schwartz (1982), Nature 295 616-618

Hartley B.S. (1970),"Homologies in Serine Proteases" in Phil. Trans. of the Royal Society of London 237 77-87

Hartley B.S., B.D.Burleigh, G.G.Midwinter, C.H.Moore, H.R.Morris, P.W.J Rigby, M.J.Smith and S.S.Taylor (1972), in "Enzymes: Structure and Function", 8th FEBS Meeting 29 151-176 , eds. J.Drenth, R.A. Oosterbaan and C.Vieger, Amsterdam ; North Holland.

Hartley B.S. (1974), Symp.Soc.Gen.Microbiol. 24 151-182 , Cambridge Univ. Press, (ed. M.J.Carlile).

Hartley B.S., I.Altosaar, J.M.Dothie and M.S.Neuberger (1976), p191-200 in "Structure-Function Relationships in Proteins", North Holland, Amsterdam. (eds. R.Markham and R.W.Horne).

Haseltine B. and B.Müller-Hill (1972), in "Experiments in Molecular Genetics", (ed. J.H.Miller), Cold Spring Harbor Labs., N.Y.

Heidecker G., J.Messing and B.Grönenborn (1980), Gene 10 69-73

Hélène C (1981), Bioscience Reports 1 477-483

Hepburn A.G. and J.Hindley (1979), J.Biochem.Biophys.Methods 1 299-308

Hershey A.D. (1971), ed."The Bacteriophage Lambda", Cold Spring Harbor Labs., N.Y.

Hershfield V., H.W.Boyer, C.Yanofsky, M.A.Lovett and D.R.Helinski (1974), Proc.Nat.Acad.Sci. 71 3455-3459

Herrman R., K.Neugebauer, E.Pirkl, H.Zentgraf and H.Schaller (1980), Mol.Gen.Genet. 177 231-242

Higgins C.F. and G.F-L Ames (1982), Proc.Nat.Acad.Sci. 79 1083-1087

Higgins C.F., P.D.Haag, K.Nikaido, F.Ardeshir, G.Garcia and G.F-L Ames (1982), Nature 298 723-727

Hill E., D.Tsernoglou, L.Webb, L.J.Banaszak(1972), J.Mol.Biol. 72 577-591

Hines J.C. and D.S.Ray (1980), Gene 11 207-218

Hirasima A., G.Childs and M.Inouye (1973), J.Mol.Biol. 79 373-389

Hood J.M., A.V.Fowler and I.Zabin (1978), Proc.Nat.Acad.Sci. 75 113-116

Horii T., T.Ogawa and H.Ogawa (1981), Cell 23 689-697

Horowitz N.H. (1945), Proc.Nat.Acad.Sci. 31 153-157

Horowitz N.H. (1965), in "Evolving Genes and Proteins", (ed. V.Bryson and H.Vogel), Academic Press.

Humphreys G.O., G.A.Willshaw and E.S.Anderson (1975), Biochem. Biophys. Acta 383 457-463

Ikenura T. (1981), J.Mol.Biol. 146 1-21

Jackson E.N. and C.Yanofsky (1973), J.Mol. Biol. 76 89-101

Jacob F. and J.Monod (1961), J.Mol.Biol. 3 318-356

Jay E., R.Bambara, R.Padmanabhan and R.Wu (1974), Nuc.Acids Res. 1 331- 353

Johnson A.D., A.R.Poteete, G.Lauer, R.T.Sauer, G.K.Ackers and M.Ptashne (1981), Nature 294 217-223

Johnsrud L., M.P.Calos and J.H.Miller (1978), Cell 15 1209-1219

Jörnvall H. (1973), Proc.Nat.Acad.Sci. 70 2295-2298

Jörnvall H., M.Persson and J.Jeffrey (1981), Proc.Nat.Acad.Sci. 78 4226-4230

Kasai T. (1974), Nature 249 523-527

Katz L., D.T.Kingsbury and D.R.Helinski (1973), J.Bacteriol. 114 577-591

Keller R.L. and J.M.Calvo (1979), Proc.Nat.Acad.Sci. 76 6186-6190

Kelley R.L. and C.Yanofsky (1982), Proc.Nat.Acad.Sci. 79 3120-3124

Klein R.D., E.Selsing and R.D.Wells (1980), Plasmid 3 88-91

Klenow H., K.Overgaard-Hansen and A.Patkar (1971), Eur.J.Biochem. 22 371-381

Koch A.L. (1972), Genetics 72 297-316

Krakow J.S., G.Rhodes and T.M.Jovin (1976), p127-157 in "RNA Polymerase", (ed. R.Losick), Cold Spring Harbor Labs., N.Y.

Kupper H., T.Sekiya, M.Rosenberg, J.Egan and A.Landy (1978), Nature 272 423-428

Lampel K.A. and M.Riley (1982), Mol.Gen.Genet. 186 82-86

Langridge J., P.Langridge and P.Bergquist (1980), Anal.Biochem. 103 264-271

Laskey R.A. and A.D.Mills (1977), FEBS Lett. 82 314-316

Lee F., C.Squires, C.L.Squires and C.Yanofsky (1976), J.Mol.Biol. 103 383-393

Lee N. and E.Englesberg (1963), Proc.Nat.Acad.Sci. 50 696-702

Le Grice S.E. and H.Mutzura (1981), J.Mol.Biol. 150 185-196

Lin E.C.C. (1961), J.Biol.Chem. 236 31-36

Link C.D. and A.M.Reiner (1982), Nature 298 94-96

Lodish H.F. and H.D.Robertson (1969), CSH Symp.Quant.Biol. 34 655-673

Lowery C. and J.P.Richardson (1977), J.Biol.Chem. 252 1381-1385

Loviny T., M.S.Neuberger and B.S.Hartley (1981), Biochem.J. 193 631-637

Luria S.E. (1965), p357-358 in"Evolving Genes and Proteins",(ed. H.Bryson and H.Vogel), Academic Press.

MacPhee D.J., I.W.Sutherland and J.F.Wilkinson (1969), Nature 221 475-476

Magasanik B. (1961), CSH Symp.Quant.Biol. 26 249-256

Maizels N. (1973), Proc.Nat.Acad.Sci. 70 3585-3589

Majors J. (1975), Proc.Nat.Acad.Sci. 72 4394-4398

Makman R.S. and E.W.Sutherland (1965), J.Biol.Chem. 240 1309-1314

Maniatis T., M.Ptashne, K.Backmann, D.Kleid, S.Flashman, A.Jeffrey and R.Maurer (1975a) , Cell 5 109-113

Maniatis T., A.Jeffrey and D.Kleid (1975b), Proc.Nat.Acad.Sci. 72 1184-1188

Martin F. and I.Tinoco (1980), Nuc. Acids Res. 8 2295-2300

Maxam A. and W.Gilbert (1977), Proc.Nat.Acad.Sci. 74 560-564

Maxam A. and W.Gilbert (1980), p499 in Methods in Enzymology 65

McCarthy B.J. (1967), Bacteriol.Rev 31 215-229

McConnell D.J. (1979), Nuc. Acids Res. 6 3491-3503

McDonell M.W., M.N.Simon and F.W.Studier (1977), J.Mol.Biol. 110 119-146

McLachlan A.D. and J.E.Walker (1977), J.Mol.BIol. 112 543-558

Messing J., B.Grönenborn, B.Müller-Hill and P.H.Hofschneider (1977) Proc.Nat.Acad.Sci. 74 3642-3646

Messing J., R.Crea and P.H.Seeburg (1981), Nuc. Acids Res. 9 309-321

Miller G.L. (1959), Anal.Chem. 31 964

Miller J.H. (1970), p173-188 in "The Lactose Operon", (ed. J.R.Beckwith and D.Zipser), Cold Spring Harbor Labs., N.Y.

Miller J.H., T.Platt and K.Weber (1970), p343-352 in "The Lactose Operon",

Miller J.H. (1972), ed. "Expts. in Molecular Genetics", CSH Labs., N.Y.

Miller J.H. (1978), p31-88 in "The Operon",(ed. J.H.Miller and W.S. Reznikoff), Cold Spring Harbor Labs., N.Y.

Monod J. (1947), Growth 11 223-289

Mortlock R.P., D.D.Fossitt and W.A.Wood (1965), Proc.Nat.Acad.Sci. 54 572-579

Müller U.R. and W.M.Fitch (1982), Nature 298 582-585

Müller-Hill B. (1975), Prog.Biophys.Mol.Biol. 30 227-252

Musso R.E., R Di Lauro, S.Adhya and B de Crombrugghe (1977), Cell 12 847-854

Murotsu T., K.Matsubara, H.Sugiaski and M.Takanami (1981), Gene 15 257-271

Nakamura K. and M.Inouye (1979), Cell 18 1109-1117

Nakanishi S., S.Adhya, M.Gottesman and I.Pastan (1974), J.Biol.Chem. 249 4050-4056

Nakanishi S., S.Adhya, M.Gottesman and I.Pastan (1975), J.Biol.Chem. 250 8202-8208

Neidhart F.C. and B.Magasanik (1956), Biochem.Biophys.Acta 21 324-334

Neuberger M.S. (1978), PhD. Thesis, University of London.

Neuberger M.S. and B.S.Hartley (1979), J.Mol.Biol. 132 435-470

Neuberger M.S., R.A.Patterson and B.S.Hartley (1979), Biochem.J. 183 31-42

Neuberger M.S. and B.S.Hartley (1981), J.Gen.Microbiol. 122 181-191

Neuberger M.S., B.S.Hartley and J.E.Walker (1981), Biochem.J. 193 513-524

Nichols B.P., M.Blumenberg and C.Yanofsky (1981a), Nuc. Acids Res. 9 1743-1755

Nichols B.P., M Van Cleemput and C.Yanofsky (1981b), J.Mol.Biol. 146 45-54

Nomura M., L.E.Post and S.Jinks (1980), p315-328 in "RNA Polymerase,tRNA and Ribosomes",(ed. S.Osawa), Univ. of Tokyo Press, Japan.

O'Farrell P.H. (1978), Cell 14 545-557

Ogden S., D.Haggerty, C.Stoner, D.Kolodrubetz and R.Schlief (1980), Proc.Nat.Acad.Sci. 77 3346-3350

Ohno S. (1970), in "Evolution By Gene Duplication", Springer-Verlag, Berlin, Heidelberg, N.Y.

O'Neill M.C. (1976), Nature 260 550-554

Orgel L.E. (1977), J.Theoret.Biol. 67 773

Otsuka A. and J.Abelson (1978), Nature 276 689-694

Pabo C.O., R.T.Sauer, J.N.Sturtevant and M.Ptashne (1979), Proc.Nat.Acad. Sci. 76 1608-1612

Pabo C.O. and M.Lewis (1982), Nature 298 443-447

Parker J., J.Pollard, J.Friesen and C.Stanners (1978), Proc.Nat.Acad. Sci. 75 1091-1095

Peacock A.C. and C.W.Dingman (1968), Biochemistry 7 668-674

Perlman R.L. and I.Pastan (1968), J.Biol.Chem. 243 5420-5427

Pirrotta V. (1975), Nature 254 114-117

Pirrotta V., P.Chadwick and M.Ptashne (1970), Nature 227 41-44

Pirrotta V., K.Ineichen and A.Walz (1980), Mol.Gen.Genet. 180 369-376

Platt T. and C.Yanofsky (1975), Proc.Nat.Acad.Sci. 72 2399-2403

Platt T., C.Squires and C.Yanofsky (1976), J.Mol.Biol. 103 411-420

Platt T. (1978), p263-302 in "The Operon",(ed. J.H.Miller and W.S.Reznikoff), CSH Labs., N.Y.

Platt T. (1981), Cell 24 10-23

Post L.E., A.E.Arfsten, F.Reusser and M.Nomura (1978), Cell 15 215-229

Post L.E., G.D.Strycharz, M.Nomura, H.Lewis and P.Dennis (1979), Proc. Nat.Acad.Sci. 76 1697-1701

Pribnow D. (1975), Proc.Nat.Acad.Sci. 72 784-788

Ptashne M., K.Backmann, M.Humayan, A.Jeffrey, R.Maurer, B.Meyer and R.T.Sauer (1976), Science 196 156-161

Queen C. and M.Rosenberg (1981), Nuc. Acids Res. 9 3365-3377

Reiner A.M. (1975), J.Bacteriol. 132 166-173

Remaut E., P.Stanssens and W.Fiers (1981), Gene 15 81-93

Reznikoff W.S. (1972), Ann.Rev.Genet. 6 133-156

Reznikoff W.S. and J.N.Abelson (1978), p221-243 in "The Operon", (ed. J.H.Miller and W.S.Reznikoff), CSH Labs., N.Y.

Richmond R.C. (1970), Nature 225 1025-1028

Rigby P.W.J. (1971), PhD. Thesis, University of Cambridge.

Rigby P.W.J., B.D.Burleigh and B.S.Hartley (1974), Nature 251 200-204

Rigby P.W.J., M-J Gething and B.S.Hartley (1976), J.Bacteriol. 125 728-738

Riggs A.D. and S.Bourgeois (1968), J.Mol.Biol. 34 361-364

Roberts J.W. (1969), Nature 224 1168-1174

Roberts R.J. (1980), Nuc. Acids Res. 8 r63-r80

Robertson H.D., R.E.Webster and N.D.Zinder (1968), J.Biol.Chem. 243 82-91

Rosenberg M. and D.Court (1979), Ann.Rev.Genet. 13 319-353

Rosenberg M., D.Court, H.Shimatake, C.Brady and D.Wulfe (1978), Nature 272 414-423

Rossman M.G., M.J.Adams, M.Buehner, G.C.Ford, M.L.Hackert, P.J.Lentz Jr, A.McPherson Jr, R.W.Shevitz and I.E.Smiley (1971),CSH Symp. Quant.Biol. 36 179-191

Rossman M.G., A.Liljas, C-I.Brändén and L.J.Banaszak (1975), p61-102 in "The Enzymes", (ed. H.Boyer), 3rd Edition, Vol 11

Roychoudry M. and R.Wu (1980), in Methods in Enzymology 65

Russell D.R. and G.N.Bennett (1982), Gene 17 9-18

Sanger F., G.G.Brownlee and B.G.Barrell (1965), J.Mol.Biol. 13 373-398

Sanger F., A.R.Coulson (1975), J.Mol.Biol. 94 441-448

Sanger F., S.Nicklen and A.R.Coulson (1977), Proc.Nat.Acad.Sci. 74 5463-5467

Sanger F. and A.R.Coulson (1978), FEBS Lett. 87 107-110

Sanger F., A.R.Coulson, B.G.Barrell, A.J.Smith and B.A.Roe (1980), J.Mol.Biol. 143 161-178

Sauer R.T., J.Pan, P.Hopper, K.Hehir, J.Brown and A.R.Poteete (1981), Biochemistry (Washington), 20 3591-3598

Sauer R.T., R.R.Yocum, R.F.Doolittle, M.Lewis and C.O.Pabo (1982a), Nature 298 447-451

Sauer R.T., M.J.Ross and M.Ptashne (1982b), J.Biol.Chem. 257 4458-4462

Scaife J. and J.R.Beckwith (1966), CSH Symp.Quant.Biol 31 403-408

Scangos G.A. and A.M.Reiner (1978), J.Bacteriol. 134 492-500

Scangos G.A. and A.M.Reiner (1979), J.Mol.Evol. 12 189-195

Schaller H., C.Gray and K.Herrmann (1975), Proc.Nat.Acad.Sci. 72 737-741

Schlessinger D., K.A.Jacobs, R.S.Gupta, Y.Kano and F.Imamato (1977), J.Mol.Biol. 110 421-439

Schmitz A. and D.J.Galas (1979), Nuc. Acids Res. 6 111-137

Schmitz A. (1981), Nuc. Acids Res. 9 277-291

Schreier P.H. and R.Cortese (1979), J.Mol.Biol. 129 169-172

Sekiya T. and H.G.Khorana (1974), Proc.Nat.Acad.Sci. 71 2978-2982

Selker E. and C.Yanofsky (1979), J.Mol.Biol. 130 135-143

Shapiro J.A. (1979), Proc.Nat.Acad.Sci. 76 1933-1937

Shine J. and L.Dalgarno (1974), Proc.Nat.Acad.Sci. 71 1342-1346

Siebenlist U., R.B.Simpson and W.Gilbert (1980), Cell 20 269-281

Simpson R.B. (1980), Nuc. Acids Res. 8 759-766

Smith G.R. (1971), Virology 45 208-223

Smith H.O. and M.L.Birnsteil (1976), Nuc. Acids Res. 3 2387-2399

Smithies O., G.E.Connell and G.H.Dixon (1962), Nature 196 232-236

Sober H.A. (1970), ed. "The Handbook of Biochemistry:Selected Data for Molecular Biology", CRC Press, West Palm Beach, Florida, U.S.A.

Squires C.L., F.D.Lee and C.Yanofsky (1975), J.Mol.Biol. 92 93-111

Staden R. (1977), Nuc. Acids Res. 4 4037-4051

Stafford D. and D.Bieber (1975), Biochem. Biophys. Acta 378 18-21

Steitz J.A. (1973), Proc.Nat.Acad.Sci. 70 2605-2609

Steitz J.A. (1979), p349-399 in "Biological Regulation and Development", Vol. 1 , (ed. R.F.Goldberger), Plenum Press, N.Y.

Steitz J.A. and K.Jakes (1975), Proc.Nat.Acad.Sci. 72 4734-4738

Steitz J.A., D.H.Ohlendorf, D.B.McKay, W.F.Anderson and B.W.Matthews (1982), Proc.Nat.Acad.Sci. 79 3097-3100

Strobel E., P.Dunsmuir and G.Rubin (1979), Cell 17 429-439

Sturtevant A.H. and E.Novitski (1941), Genetics 26 517

Sutcliffe J.G. (1978a), CSH Symp.Quant.Biol. 43 77-90

Sutcliffe J.G. (1978b), Nuc. Acids Res. 5 2721-2728

Taniguchi T., T.M.O'Neill and B de Crombrugghe (1979), Proc.Nat.Acad. Sci. 76 5090-5094

Taylor S.S., P.W.J.Rigby and B.S.Hartley (1974), Biochem.J. 141 693-700

Telford J., P.Boseley, W.Schaffner and M.L.Birnsteil (1977), Science 195 391-392

Timmis K., F.Cabello and S.N.Cohen (1974), Proc.Nat.Acad.Sci. 71 4556-4560

Tinoco I., P.N.Borer, B.Dengler, M.D.Levine, O.C.Uhlenbeck, D.M.Crothers and J.Gralla (1973), Nature New Biology 246 40-41

Travers A.A. (1980), J.Bacteriol. 141 973-976

Tsurimoto T. and K.Matsubara (1981), Nuc. Acids Res. 9 1789-1801

Tu C-P. and S.N.Cohen (1980), Gene 10 177-183

Varmus H.E., R.L.Perlman and I.Pastan (1970), J.Biol.Chem. 245 2259-2267

Vogt V. (1969), Nature 223 854-855

Weinberg E.S., M.C.Birnsteil, I.F.Purdom and R.Williamson (1972), Nature 240 225-228

Wells R.D., T.C.Goodman, W.Hillen, G.T.Horn, R.D.Klein, J.E.Larson, U.R.Müller, S.K.Nuendorf, N.Panoyotatos and S.M.Stirdivant (1980), Prog.Nuc.Acids Res. and Mol.Biol. 24 167-267

Weymouth L.A. and L.A. Loeb (1978), Proc.Nat.Acad.Sci. 75 1924-1928

Willmund R. and H.Kneser (1973), Mol.Gen.Genet. 126 165-175

von Wilcken-Bergmann B. and B.Müller-Hill (1982), Proc.Nat.Acad.Sci. 79 2427-2431

Wilson B.L. and R.P.Mortlock (1973), J.Bacteriol. 113 1404-1411

Winter G. (1980), EMBO Cloning Course, Cambridge MRC.

Winter G. and S.Fields (1980), Nuc. Acids Res. 8 1965-1974

Wood W.A. and J.Tai (1958) , in Bact. Proc, Soc. American Bacteriologists, Baltimore. p99.

Wood W.A., M.J.McDonough and L.B.Jacobs (1961), J.Biol.Chem. 236 2190-2195

Wu A.M. and T.Platt (1978), Proc.Nat.Acad.Sci. 75 5442-5445

Wu A.M., G.E.Christie and T.Platt (1981), Proc.Nat.Acad.Sci. 78 2913-2017

Wu T.T., E.C.C.Lin and S.Tanaka (1968), J.Bacteriol. 96 447-456

Yanofsky C. (1981), Nature 289 751-758

Young R.A. and J.A.Steitz (1978), Proc.Nat.Acad.Sci. 75 3593-3597

Zalkin H., C.Yanofsky and C.L.Squires (1974), J.Biol.Chem. 249 465-475

Zubay G. (1973), Ann.Rev.Genet. 7 267-287

Zuckerkandl E. and L.Pauling (1965), p97-166 in "Evolving Genes and Proteins", (ed. H.Bryson and H.Vogel), Academic Press.

Zuckerkandl E. (1975), J.Mol.Evol. 7 1-57

Zurawski G., D.Elseviers, G.V.Stauffer and C.Yanofsky (1978), Proc.Nat. Acad.Sci. 75 5988-5992