GENETICS OF ABEQUOSE BIOSYNTHESIS IN THE *rfb* REGION OF *Salmonella typhimurium* LT2

Paul Wyk, B.Sc. (Hons.) (Adelaide)
Department of Microbiology and Immunology
The University of Adelaide

A thesis submitted for the degree of
Doctor of Philosophy
February 1988

TABLE OF CONTENTS

		page
	Table of contents	i
	Abstract	vii
	Statement	ix
	Acknowledgements	x
	Index of figures and tables	xi
CHAPTER 1:	INTRODUCTION	1
1.1	Cell walls	1
1.1.1	Peptidoglycan function	1
1.1.2	Periplasm	2
1.1.3	Fusion sites between outer and inner membranes	2
1.2	The outer membrane	3
1.3	Lipopolysaccharide	4
1.3.1	Lipid A	5
1.3.2	R-core structure	6
1.3.3	O-antigen	7
1.3.3.1	Modifications	8
1.4	Polysaccharide components other than LPS	9
1.4.1	T antigens	9
1.4.2	Vi antigen	10
1.4.3	Colanic acid	10
1.4.4	Enterobacterial Common Antigen	11
1.5.1	LPS function - Permeability	11
1.5.2	LPS function - Resistance to host defences	12

1.6	LPS biosynthesis	14
1.6.1	LPS biosynthesis - Lipid A	15
1.6.2	LPS biosynthesis - The Inner Core	16
1.6.3	LPS biosynthesis - The Outer Core	18
1.7	O-Chain Biosynthesis	20
1.7.1	O-Chain Biosynthesis - Acyl lipid carrier	21
1.7.2	O-Chain Biosynthesis - Transferases	22
1.7.3	O-Chain Biosynthesis - Polymerization	22
1.7.4	O-Chain Biosynthesis - Transloction	24
1.7.5	O-Chain Biosynthesis - Transport across to the Outer Membrane	24
1.8	The *rfb* locus	25
1.8.1	Gene order, and direction of transcription in the *rfb* locus	26
1.8.2	The *rfb* locus - Biosynthetic pathways	29
1.8.3	The *rfb* locus - UDP-galactose biosynthesis	30
1.8.4	The *rfb* locus - GDP-mannose biosynthesis	30
1.8.5	The *rfb* locus - TDP-rhamnose biosynthesis	31
1.8.6	The *rfb* locus - CDP-abequose biosynthesis	33
1.9	Aims of this thesis	35
CHAPTER 2:	<u>MATERIALS AND METHODS</u>	36
2.1	Reagents	36
2.2	Growth media	37
2.3	Bacterial strains, Bacteriophages and Plasmids used	37
2.4	DNA techniques	38
2.4.1.1	DNA isolation - Chromosomal	38
2.4.1.1	DNA isolation - Small scale plasmid	38

2.4.1.3	DNA isolation - Large scale plasmid	38
2.4.1.4	DNA isolation - Lambda	38
2.4.1.5	DNA isolation - M13 ss DNA	39
2.4.2	Restriction endonuclease digestion	39
2.4.3	Agarose/Acrylamide gel electrophoresis	39
2.4.4	Purification of DNA fragments :- Agarose gels	40
2.4.5	Purification of DNA fagments :- Acrylamide gels	40
2.4.6	Endfilling and Endlabelling	40
2.4.7	Nick Translation	41
2.4.8	Ligation of DNA	41
2.4.9	Transformation	41
2.4.10	Plaque and colony hybidization	42
2.4.11	Bal31 deletion mutagenesis	42
2.4.12	DNAseI deletion mutagenesis	42
2.4.13	Tn1725 mutagenesis	42
2.4.14	Cosmid band construction	42
2.4.15	Sequencing methods	43
2.4.16	Oligonucleotide synthesis	43
2.5	Enzyme assays	43
2.5.1	Cell extracts	43
2.5.2	ABE-1	43
2.5.3	ABE-2	44
2.6	Protein techniques	44
2.6.1	PAGE electophoresis	44
2.6.2	Fractionation of Cell envelope and cytoplasm	45
2.6.3	Minicell methods	45
2.7	Immunological techniques	45

2.7.1 Whole cell lysate 45

2.7.2 Western Blot 45

2.7.3 Slide agglutination 46

CHAPTER 3: CLONING AND IDENTIFYING THE ABEQUOSE REGION IN *S. typhimurium* 47

3.1 Introduction 47

3.2 Cloning - using a lambda bank 48

3.3 Cloning - construction and the use of a cosmid bank 50

3.4 Summary and conclusions 53

CHAPTER 4: LOCALIZING DNA RESPONSIBLE FOR THE SPECIFICITY IN ABEQUOSE BIOSYNTHESIS 56

4.1 Introduction 56

4.2 Expression of 0-4 specific LPS in ~~*Salmonellae*~~ salmonellae of groups A and D 57

4.3 Localizing DNA that specifies the O-4 epitope 58

4.4 Summary and conclusions 61

CHAPTER 5: SEQUENCE AND ANALYSIS OF PART OF THE REGION INVOLVED IN ABEQUOSE BIOSYNTHESIS 62

5.1 Introduction 62

5.2 Sequencing strategy 63

5.3.1 Sequence analysis - Predicting proteins 64

5.3.2 Sequence analysis - Properties of the predicted proteins 66

5.3.3 Sequence analysis:- Homologies wih other proteins 70

5.3.4	Sequence analysis:- Homologies betweenpredicted proteins	71
5.3.5	Sequence analysis:- comparison of *rfbJ* and *galE*	73
5.3.6	Sequence analysis - Dehydrogenase stucture	75
5.3.7	Sequence analysis:- accuracy of the sequence data	77
5.3.8	Sequence analysis - Special properties; low G+C content	77
5.3.9	Sequence analysis - Codon usage and amino acid composition	80
5.3.10	Sequence analysis - G+C content of the 1st, 2nd and 3rd bases	82
5.3.11	Sequence analysis - Special structures	83
5.4	Summary and conclusions	84
CHAPTER 6:	PROTEINS IN THE REGION ENCODING ABEQUOSE BIOSYNTHESIS	90
6.1	Introduction	90
6.2.1	Detection of Enzymes encoded by the original clones	92
6.2.2	Enzyme activity of tn1725 mutants of pPR586	93
6.2.3	Localisation of the ABE-1 and ABE-2 enzyme activity	94
6.3.1	^{35}S labelled proteins seen in minicells	95
6.3.2	Proteins visualized by Coomassie Brilliant Blue Staining in the region 9.57 to 15.28	97
6.4	Summary and conclusion	99

CHAPTER 7: DISCUSSION 103

7.1 Introduction 103

7.2.1 Cloning *rfb* DNA 103

7.2.2 Localizing the region encoding Abequose biosynthesis genes 103

7.2.3 Asymmetrical distribution of cosmid clones 104

7.3.1 Localizing *rfbJ* - the gene responsible for 0-4 antigen specificity 106

7.3.2 ORF224 may be an abequosyl transferase 108

7.4.1 Possible *rfbJ* regulation 109

7.4.2 Tyvelose expression in hybrid strains 111

7.5 Proteins 111

7.5.1 Protein function - *rfbJ*, *rfbG* and *rfbF*. 112

7.5.2 Protein function - *rfbH* and *rfbI* 114

7.5.3 Protein function - proteins of unknown function seen by PAGE 116

7.5.4 Protein function - other proteims identified by sequence data 117

7.6 An NAD+ dehydrogenase binding domain in *rfbJ* and *galE*. 118

7.7 Much of the sequenced DNA has an extremely low G+C content 119

7.8 Promoters of the *rfb* gene cluster in the region examined 121

7.9 The genetic structure of the abequose biosynthesis region 125

BIBLIOGRAPHY 127

ABSTRACT

The aim of this thesis was to characterize the biosynthesis of abequose as encoded in the *rfb* region of *Salmolnella typhimurium*. Some of the *rfb* region has already been cloned (Brahmbhatt *et al.*, 1986) and some of this DNA was used to probe a lambda 1059 gene bank of *S. typhimurim* LT2 for an *rfb* carrying clone. The extent of the cloned DNA was measured by its ability to probe to a series of chromosomal deletion mutants and insufficient DNA was cloned to cover the abequose synthesis genes of this region, so a cosmid bank was constructed using pcos2EMBL as a vector. Nine *rfb* containing cosmids were isolated and subclones made from these were used to probe the same series of chromosomal deletion mutants to show that sufficient DNA had been cloned to cover the abequose biosynthetic portion of the *S. typhimurium rfb* region. A restriction endonuclease map of a substantial part of the abequose biosynthesis region was compiled.

A series of plasmid constructs made with DNA from the abequose biosynthesis region were transformed into Salmonellae of groups A and D, and *S. typhimurium*-like (0-4 specific) LPS was produced; a function that was localised to a 1.23kb DNA fragment. A part of the abequose biosynthetic region was sequenced to produce a protein map of seven open reading frames. Homology between these proteins was analysed and several of these proteins have a segment that shows homology with the other proteins. Two other open reading frames ORF311 and ORF344 are shown to have other more extensive homologies. The gene that confers 04 specificity on LPS, *rfbJ*, are shown to have similarities to the recently sequenced *galE* protein and both proteins were found to have

significant similarities with the NAD binding domain identified by Rossman *et al.*, (1975).

The sequenced region is found to have an extremely low G+C content, and one boundary of the low G+C content DNA is probably identified, a second boundary is postulated on less firm data. Amino acid usage is found to be significantly different to amino acid usage seen in genes with a G+C content closer to 50% in keeping with the low G+C content of the region.

The genes for the first two enzymes of the abequose biosynthetic pathway *rfbF* and *rfbG* have been localized using sublconed fragments, and enzyme assays, and *rfbJ*, the gene encoding the last enzyme in the pathway has been localized by sequence data. Their products have been seen as radioactive proteins in minicell systems and Coomassie Brilliant Blue stained proteins on SDS-PAGE. The two other proteins involved in CDP-abequose biosynthesis have not been positively identified, although it seems very likely that one of the open reading frames is *rfbH*, however, *rfbI* has not been localized.

Two promoter are localized in the abequose biosynthetic region, a strong promoter at the likely beginning of the DNA encoding abequose biosynthesis and another weaker one after the second gene of the pathway, *rfbG*, in front of *rfbJ*.

STATEMENT

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university, and to the best of my knowledge and belief, it contains no material previously published or written by another person, except where due reference is given in the text.

Paul Wyk

February 1988

ACKNOWLEDGEMENTS

I am deeply indebted to Professor Peter Reeves for his supervision, support and inspiration.

Thanks are due to the students and post-doctoral fellows associated with Professor Reeves, to Dr. Paul Manning and to members of his group, for assistance and helpful discussion. Many thanks to Dr. Michael Heuzenroeder for his invaluable help. Thanks are also due to Dr. A. Sivaprasad for assistance in sequencing techniques and to Mr. Chris Cursaro for occasional technical assistance.

Lastly I am deeply indebted to Kerri and Michael for their patience and understanding.

FIGURES AND TABLES

FIGURES page

1.1 location of *rfb* genes defined by *S. typhimurium* strains with deletion endpoints in the *rfb* region. 25

3.1 Restriction map of *rfb* DNA in the region encoding abequose biosynthesis 49

3.2 Extent of cloned DNA in relation to deletion mutants of *S. typhimurium* as shown by southern hybridization 49

3.3 Initial restriction map of *rfb* DNA cloned in several cosmids and the extent of *rfb* DNA in those cosmids 51

4.1 Clones that express the antigenic factor 0-4 in *Salmonella dublin* M6 58

4.2 Polymerization of O-units with antigen factor O-4 in *Salmonella dublin* M6 61

5.1 Clones used in sequencing 63

5.2 Sequence of *rfb* from 3.28 to 5.97 64

5.3 Coding regions in the DNA sequence as predicted by Shepherd's RNY preference 67

5.4 Charge and Hydrophobicity plots of predicted proteins 67

5.5 Possible secondary structures in the predicted proteins 67

5.6 Alignment of amino acids in a region that shows homologies in all proteins 71

5.7 Relationship of amino acid sequence of two predicted proteins; ORF311 and ORF334 72

5.8 Relationship of the amino acid sequence of two proteins; *rfbJ* and *galE* 74

5.9 Amino acid alignment of *rfbJ* and *galE* in relation

to amino acids of NAD^+ binding domains 76

5.10 Restriction enzyme sites predicted for the sequenced region 77

5.11 Variation in G+C content of the sequenced DNA and associated parameters 77

6.1 Transposon tn1725 insertion mutants of pPR586 93

6.2 Proteins produced by minicells in various clones containing *rfb* DNA 96

6.3 Proteins produced in strains carrying clones with *rfb* DNA 98

6.4 Proteins encoded in that part of the *rfb* region encoding abequose biosynthesis 99

7.1 Proteins and potential promotes in that part of the *rfb* region encoding genes for abequose biosynthesis 111

TABLES

2.1 Bacterial strains used in this thesis 38

2.2 Plasmids used in this thesis 38

5.1 Synthetic oligonucleotides used in sequencing 64

5.2 Suggested ribosome binding sites in the region sequenced 65

5.3 Transmembrane segment predictions for proteins of the regions sequenced 68

5.4 Homologies between the predicted proteins 71

5.5 Base composition of coding and non-coding regions inthe sequenced DNA 78

5.6 Codon usuage tables for each of the predicted proteins 80

5.7 Amino acid content of the proteins predicted

from the sequence data 80

5.8 G+C content of the 1st, 2nd and 3rd bases in coding triplets of the sequenced region 80

6.1 ABE-1 and ABE-2 enzyme activity in strains carrying a variety of *rfb* clones 92

CHAPTER ONE

INTRODUCTION

1.1. Cell Envelopes.

The surface structures of bacteria are at the interface between the organism and their environment and as such play a primary role in the interaction of bacteria with their environment. Bacteria produce walls at this interface outside of a lipid bilayer membrane which separates cytoplasm from the outside. Gram positive bacteria have one lipid bilayer with other cell wall structures external to that whereas gram negative bacteria have two lipid bilayer membranes, an inner membrane and an outer membrane. Proteins with a variety of functions constitute a large proportion of both membranes (Braun and Hantke, 1974). Many of the proteins in the inner membrane are involved in the transport of metabolites across the barrier set up by this membrane. Some of the transport systems require a proton gradient across the inner membrane that is generated with the help of the ATPase (Senior, 1985) and the electron transport chain; other transport systems derive energy for transport chemically (Postma and Langeler, 1985). A variety of proteins function to transport and assemble the structures associated with the membrane systems. The way in which the proteins themselves are exported is attracting a considerable amount of research (Oliver, 1985).

1.1.1. Peptidoglycan function.

A layer of peptidoglycan surrounds the outside of this plasma

membrane in most gram-negative bacteria. In *Escherichia coli* it takes the form of a thin layer composed of chains comprising two sugar derivatives, *N*-acetylglucosamine and *N*-acetylmuramic acid, that are cross linked by a tetrapeptide bridge of L-alanine, D-alanine, D-glutamic acid, and diaminopimelic acid but variations of this configuration do exist in other organisms. This functions to give the bacterial cell mechanical rigidity and to overcome the osmotic stresses that the cell experiences.

1.1.2. Periplasm.

The presence of the outer membrane in gram-negative bacteria creates two extra functional regions in the cell, one is the outer membrane itself, and the other is the periplasmic space which lies between the membranes. The periplasmic space has a significant proportion of the total cellular protein and these periplasmic proteins are usually involved in functions which are cell-surface related, such as transport, or degradation of metabolizable compounds which are too large or too highly charged to enter the cell intact (Ferro-Luzzi Ames and Higgins 1983, Ferro-Luzzi Ames, 1986). Peptidoglycan also lies within the periplasm.

1.1.3. Fusion sites between outer and inner membranes.

Under certain conditions, sites where the plasma membrane and the outer membrane are fused can be seen. It appears that a number of phages use these sites to attach to cells, and these sites are probably the route by which their DNA enters the cell on infection (Bayer, 1975; Bayer, 1979; Bayer *et al.*, 1982). The adhesion sites are thought to be the regions through which outer membrane components are transported. Some speculation has it that the

adhesion sites may also be involved in the uptake of various substances, especially macromolecules such as colicins and DNA, but little data is available to support this contention.

1.2. The outer membrane.

The outer membrane is, like the inner membrane, essentially a lipid bilayer, with a large number of associated proteins. One of the major components of the outer membrane protein profile is the murein lipoprotein which is substituted at its N terminus with fatty acids thought to anchor the lipoproteins to the outer membrane and in a proportion of cases is covalently linked via a cysteine residue on its carboxy terminus to the peptidoglycan layer. It thus seems to function in anchoring the outer membrane onto the peptidoglycan (Braun, 1975). Another major outer membrane protein, the *ompA* protein, appears to have a role in conjugation (Skurray *et al.*, 1974). The major outer membrane proteins coded for by the *ompF*, *ompC* and *phoE* genes in *E. coli* K-12 produce relatively non-specific pores or channels that allow the passage of small hydrophilic molecules, of less than 600 daltons, across the outer membrane. Other enteric bacteria sometimes produce additional porins, for example the *ompD* porin of *S. typhimurium* LT2 (Nikaido and Vaara, 1985).

In addition to these specific diffusion channels several proteins function to transport specific compounds across the outer membrane. The PhoE porin synthesis is deregulated under low phosphate conditions and is more efficient at transporting phosphorylated compounds, so that in addition to its general porin function it is also has the additional function of transporting

phosphate. In *E. coli* the *lamB* and *malE* proteins cooperate in the transport of maltodextrins (Bavoil and Nikaido, 1981); the *tsx* protein is involved in the diffusion of nucleosides across the outer membrane (Koch, 1971); the collaboration between the *tonB* protein and the outer membrane protein *btuB* protein results in the accumulation of very large numbers of vitamin B_{12} molecules in the periplasmic space (Reynolds *et al.*, 1980). There are a number of outer membrane proteins involved in the transport of various chelates of ferric ion; the *tonA* protein, necessary for the uptake of ferrichrome, (Luckey *et al.*, 1975; Wayne and Neilands, 1975), the *fepA* protein, required for the transport of ferric enterochelin and citrate (Hollifield and Neilands, 1978).

The outer membrane has few enzymes associated with it and the only one that has been identified is phospholipase A_1 (Nishijima *et al.*, 1977), although it appears that the outer membrane also possesses protease activity (Melchior and Stein, 1976).

1.3. Lipopolysaccharide.

The outer membrane of gram negative bacteria is a very atypical membrane being asymmetrical in the distribution of its lipid components. It appears that little or no phospholipid is present in the outer leaflet of this membrane. In place of the phopholipids are lipopolysaccharide (LPS) molecules which are peculiar to the outer leaflet of the outer membrane of gram-negative bacteria (Mühlradt and Golecki, 1975; Funahara and Nikaido, 1980).

LPS structure is most thoroughly described for *Salmonella typhimurium*, and the structure of the LPS of this organism serves

as a basis for comparison of results obtained in other species. LPS from many members of the family *Enterobacteriaceae* has been analysed, and the structures found in all of them have similarities. Bacteria that carry mutations in genes encoding LPS have colonies that are rough and irregular instead of entire and even and because of this difference bacteria (and also LPSs) which have complete LPS are denoted as smooth (S) and those which are mutant are denoted rough (R) (Hitchcock *et al.*, 1986). A large range of mutants in LPS related genes has been isolated and most of these produce truncated LPS molecules, since omission of any sugar in the core will interfere with the addition of distal sugars (Kauffmann, 1966). These mutations have helped workers recognize the tripartite structure of the LPS molecule, the three regions being; i) the lipid A moiety, ii) the core region, synthesized by enzymes encoded mostly by the *rfa* gene cluster, iii) and the O-polysaccharide, synthesized by enzymes encoded mainly in the *rfb* gene cluster (Lüderitz *et al.*, 1968).

1.3.1. Lipid A.

Lipid A has a very highly conserved structure in the enteric organisms examined so far. The lipid A moiety anchors the LPS molecule into the outer leaflet of the outer membrane (Rietschel *et al.*, 1984) and also possesses the enterotoxigenic properties of LPS. However its structure has only recently been elucidated (Sidorcyk *et al.*, 1983; Strain *et al.*, 1983a, 1983b Takayama *et al.*, 1983). Some contention still exists over substitutions of the lipid A, due largely to technical difficulties in working with isolated material, and to the lethality of mutations in genes responsible for biosynthesis of this region.

Lipid A is a diglucosamine in beta 1-6 linkage with ester-, amide- and diester-linked fatty acids (the type and position of fatty acids varies among species) and with phosphate and ethanolamine substituents in *E. coli* or 4-amino-arabinose and phosphate substituents in *S. typhimurium*, on the terminii of the dissacharide backbone. Unlike phospholipids, which have only two fatty acid chains connected to the backbone structure, an LPS molecule has up to seven fatty acid chains, with carbon length of 12 to 16, linked to the glucosamine dissacharide backbone. Unlike phospholipids, all of the fatty acid chains in an LPS molecule are saturated under normal cultural conditions, and some are 3-hydroxy fatty acids. Furthermore, some fatty acid residues are linked to the 3-hydroxy group of other fatty acids, producing the characteristic 3-acyl-oxy-acyl structure.

1.3.2. R-core structure.

The R-core is taken to mean the chains of sugars between lipid A and the O-antigen and is often described in two parts; the inner core and the outer core. The outer core has been characterized reasonably thoroughly and is quite uniform, with one common core structure present in all Salmonellae examined and minor variations of this core present in other *Enterobacteriaceae*. For example five different outer core types have been identified in various strains of *E. coli* (Jann and Jann, 1984); all are made of similar sugars to those of the Salmonella core. The Salmonella core has a distal hexose oligosaccharide made of a glucose-galactose-glucose backbone with galactose attached to glucose I and *N*-acetylglucosamine attached to glucose II, this region is sometimes referred to as the outer core.

The lipid A proximal inner core region consists of an oligosaccharide of the core-specific sugars, L-glycero-D-*manno*-heptose (Heptose) and 2-keto-3-deoxy-D-*manno*-octonate (KDO), each probably forming a branched trisaccharide. One of the KDO molecules forms part of the main chain linking lipid A to heptose, a second KDO is linked to the first as a branch, and there is some uncertainty as to whether a third is linked to the second as a continuation of the branch (Taken *et al.*, 1986). KDO I links the polysaccharide to lipid A in a relatively acid-labile (ketosidic) linkage. The O chains are attached to the subterminal glucose II unit of the core. A number of substitutions to the core region have been postulated, that are not necessarily present in molar amounts. Ethanolamine and phosphate substitutions to the heptose region do exist but their exact configuration is unclear.

1.3.3. O-antigen.

In Salmonellae and other genera of *Enterobacteriaceae*, O-chains are extremely polymorphic in their constituent sugars and linkages between them. Although salmonellae have one common R core structure, great variation is seen in the composition of the O-antigen of the salmonellae. Salmonella strains have been sorted into groups (initially from A to Z, but extending beyond that to groups 51-56) (Kauffmann, 1966) on the basis of shared antigen factors. Not all of these antigen factors are attributable to LPS but many of them are, additionally not all antigenic factors are unique to any one particular group (Lüderitz *et al.*, 1968), thus these groups give us an idea of the variation that exists in LPS structure but does not define it.

The O-unit of an LPS can vary in its constituents with up to seven sugars. This is often branched with one or more side-chain sugars (Jann and Jann, 1984). An example of an O-unit is that of *S. typhimurium* which has a backbone of galactose-rhamnose-mannose with an abequose residue branching from the mannose. Group A, B and D salmonellae have a backbone of this sort with a different sugar branching from the backbone, being paratose, abequose and tyvelose respectively.

1.3.3.1. Modifications.

The variation in the structure of the O-antigen is further extended by modifications and substitutions that can be made on the basic O-unit. The O-antigen of Salmonella strains can be glucosylated, O-acetylated, or the linkage of the backbone sugars may be altered. The O-antigen in *S. typhimurium,* can be taken as an example of this. O-acetylation of the abequose is usual among *S. typhimurium* strains and this substitution requires an active *oafA* gene (Mäkelä and Stocker, 1984). Another O-acetylation is due to lysogenic conversion by temperate bacteriophage A3 and A4 which cause O-acetylation of either the O-2 or O-3, or both, of the L-rhamnose (Wollin *et al.*, 1987).

Glucose can be added, in this case onto the galactose molecule. The *oafR-oafE* gene cluster is responsible for the substitution onto the C4 of the galactose residue (Mäkelä and Stocker, 1984). The resulting alpha-glucosyl-1,4-galactose epitope is the O-antigen factor 12_2 (Jann and Westphal, 1975; Lüderitz *et al.,* 1971). The *oafR* gene interconverts between two states at quite a high frequency; in one state substitution occurs and in the other it does not (Kauffmann,

1961, 1966), a phenomenon called form variation. The substitution of an O-specific polysaccharide by glucose via its ACL derivative can also be induced, or repressed, by temperate (converting) bacteriophages. Glucosylation of the galactose residue O-6 through an alpha linkage, results in antigen factor 1 (Lüderitz *et al.*, 1966). Similarly alteration of the galactose to mannose linkage between adjacent repeat units of the O-chain from alpha 1,2 to alpha 1,6 (with the appearance of antigenic factor O-27) as a result of lysogenic conversion by phage P27 reduces the rates of adsorption of phage P27 (Wollin *et al.*, 1981).

1.4. Polysaccharide components other than LPS.

A variety of other polysaccharides may form a capsule or a slime layer on the outside of the ~~*Salmonellae*~~ salmonellae. With the exception of LPS the way in which the capsular material is attached to the cell wall is not understood. The distinction between slime and capsular polysaccharides is not sharp, and the same bacteria may switch from producing one type to producing another (Sutherland, 1977). Even the distinction between LPS and capsular polysaccharide is not always sharp. Many of the polysaccharides found in capsules may use part of the existing biosynthetic machinery involved in biosynthesis of LPS.

1.4.1. T antigens.

Polysaccharides other than the O chain are in some circumstances attached to the LPS core. T1 chains, polymers of ribose and galactose in the unusual furanose form (Berst *et al.*, 1971) encoded by the *rft* (at map position 12) genes have been found in some isolates of *S. paratyphi* B unable to make O chains. T1 chains can

be linked to the core LPS and this requires both the complete core and a functional *rfaL* gene for attachment (Sarvas, 1967). Strains with the T1 antigen become T1-negative at a high frequency by mutation at or near *rft* but the reverse change does not occur with detectable frequency. Another T antigen, T2, also attaches to LPS core (Bruneteau *et al.*, 1974). This incompletely characterized polymer of *N*-acetylglucosamine depends for expression on the *rfu* gene(s) (Valtonen *et al.*, 1976a, 1976b).

1.4.2. Vi antigen.

Salmonella typhi is one of the few Salmonellas to have a capsular polysaccharide, the Vi antigen; this is a polymer of O- and N-acetylated galactosaminuronic acid (Clark *et al.*, 1958). Two gene loci *viaB* and *viaA*, at map position about 92 and 46 are associated with this polysaccharide (Johnson *et al.*, 1965, 1966). This capsule type appears to undergo variation of colony form (Kauffmann, 1954).

1.4.3. Colanic acid.

All *E. coli*, Salmonella and *Enterobacter cloacae* strains can synthesize an additional, mucoid, surface polysaccharide, which is antigenically very similar in all of them. This polysaccharide, colanic acid (M antigen), consists of hexasaccharide repeating units. Colanic acid isolated from different strains can vary in its O-acetylation and in the pyruvyl substituents peculiar to this molecule (Sutherland, 1977). In wild-type strains, the synthesis of large amounts of the M antigen can be turned on by environmental conditions such as low temperature, high concentration of salts and excess of fermentable sugars (Markovitz, 1977). Many genes are

involved in its synthesis some of which may be *rfb* genes involved in the pathway for biosynthesis of the mannose residue. Synthesis of colanic acid is regulated by the *lon* gene product, which seems to function by limiting, by proteolysis, the *rcsA* gene product which acts as a positive regulator (Torres-Cabassa and Gottesman, 1987).

1.4.4. Enterobacterial Common Antigen.

Enterobacterial common antigen (ECA) is an acidic polysaccharide containing *N*-acetyl-D-glucosamine, *N*-acetyl-D-mannosaminuronic acid and 4-acetamido-4,6-dideoxy-D-galactose, and is present in a significant amount in all members of the *Enterobacteriaceae*. ECA appears to be linked to a phospholipid "anchor", which presumably attaches ECA to the outer membrane. In the so-called ECA-immunogenic strains, however, ECA is linked to the LPS core; this form, in contrast to the haptenic form, is only found in certain rough mutants. The genes involved are close to *ilv* (*rfe* and *rff* at position 83) or in the *rfb* cluster at position 44. The *rff* genes determine enzymes required for the synthesis of UDP-*N*-acetylmannosaminuronic acid (Lew *et al.*, 1978); the function of the two other gene groups is not known. Recently, *in vitro* biosynthesis of ECA has been established (Barr and Rick, 1987), and this together with the use of the antibiotic tunicamycin, a specific inhibitor of ECA biosynthesis, will contribute greatly to our understanding of ECA biosynthesis.

1.5.1. LPS function - Permeability.

It would seem logical that the LPS plays a role in the impermeability of the outer membrane to "hydrophobic" molecules (a "hydrophobic" molecule is defined by Nikaido as one that

partitions 1% or more into octanol from water) (Nikaido and Vaara, 1985) considering that the external leaflet of the outer membrane contains LPS and is almost entirely free of phospholipid. LPS because of polar substitutions by phosphate groups and ethanolamine groups, and because of the negatively charged carboxyl groups of the KDO residues can be considered a polyanion. Very strong ionic interactions occur between the LPS molecules and between LPS molecules and proteins. These interactions are mediated by inorganic and organic cations (Coughlin *et al.*, 1983, Galanos and Lüderitz, 1975) which are, at least in part, the cause (Nikaido and Vaara, 1985) of the highly ordered state in which LPS exists in the outer leaflet of the outer membrane (Labishinski *et al.*, 1985). The highly ordered state is presumably what makes this membrane so impermeable to "hydrophobic" molecules. This property of the outer membrane can be viewed as a defence against the hostile environment found inside the intestinal tract, the first barrier of their hosts that these enteric organisms have to cross, by limiting the uptake of toxic compounds such as bile salts and other "hydrophobic" molecules which are able to diffuse directly across the inner membrane which is less ordered.

1.5.2. LPS Function - Resistance to host defences.

Another major role of LPS in enteric bacteria, in conjunction with capsular components, is to afford some protection against their host's immune response. Salmonellae are intracellular parasites; only a very short initial phase is spent in the body fluids, and the humoral response has very little effect on the overall survival rate of *S. typhimurium* during the greater part of experimental infection (Collins, 1969, Blanden *et al.*, 1966, Saxen, 1984). Complement does

seem to play a role in this initial phase; it can be activated directly by components of the LPS, without an immune response being mounted (Morrison and Kline, 1977, Plushke and Achtman, 1984, Tenner *et al.*, 1984). It appears that the quality of the polysaccharide affects the activation of complement, some polysaccharides being better at avoiding the role of activator and cover the lipid A portion well enough to avoid complement activation by that part of the molecule. For example a comparison of isogenic derivatives of *S. typhimurium* differing in the structure of their O polysaccharide showed that they differed in their ability to activate complement via the alternative pathway (Liang-Takasaki *et al.*, 1982, 1983; Valtonen, 1970, 1977). Derivatives of the same strains without any O polysaccharide were nonvirulent and easily phagocytosed, and activated complement rapidly (Valtonen, 1970, Grossman and Leive, 1984). In some cases it appears that *Salmonellae* avoid the affects of activation; Joiner *et al.*, (1982a, 1982b) found that activation of complement takes place in a hydrophilic environment on the O-polysaccharide of smooth *Salmonellae* away from the surface of the outer membrane so that the membrane-acting complement is ineffective.

Complement can also activate phagocytes via the C3b component of complement for which phagocytes have a receptor. The O-polysaccharide is important for the resistance of smooth bacteria to killing by macrophages, and mutants with a defective LPS core are very sensitive (Stocker and Mäkelä, 1986). Once the macrophages have been activated however, *Salmonellae* are no longer able to resist their battery of killing agents. The main multiplication of *Salmonellae* in mouse tissues takes place in the macrophages of the

liver and spleen (Collins, 1969). It is therefore an advantage for *Salmonellae* to be injested by macrophages. *Salmonellae* seem to have evolved elaborate O polysaccharide structures that allow complement activation, without subsequent cellular lysis, so they can be injested by phagocytes and multiply within them. Provided that the macrophages that have been infected are not activated by antibody or some other means the bacteria have a short while to multiply before the host's reaction to the infection causes activation of the macrophages.

In contrast to the humoral response, T-cell-mediated immune mechanisms can activate macrophages to kill even intracellular *Salmonellae* (MacKaness *et al.*, 1966). The development of effective immunity takes quite some time (Collins, 1969, MacKaness *et al.*, 1966) so bacteria that can grow to a sufficient population size and that can relatively efficiently avoid the non-specific host defences in this time period will kill the host. On the other hand, if a host has previously been infected T-cell immunity is activated beforehand then only *Salmonellae* with a different serotype can survive. This may be the evolutionary pressure that is responsible for the great diversity of Salmonella serotypes, and perhaps this diversity is responsible for the success of the *Salmonellae* as a genus.

1.6. LPS biosynthesis.

The lipid A with LPS core is made separately from the O-antigen and the two components are joined together after they are completed, or nearly completed. Lipid A is made from two monosaccharide precursors, which are substituted with fatty acids, KDO, phosphates, ethanolamine and 4-amino-D-arabinose and this

moiety serves as an acceptor for the assembly of the R core oligosaccharide. Sugars for the O-units are synthesized and assembled on an acyl carrier lipid (ACL), and then polymerized on the same carrier. The ACL-linked polysaccharide is subsequently transferred to core-lipid A and the whole LPS molecule is then transported across the membrane system.

1.6.1. LPS Biosynthesis - Lipid A.

Until recently little was known about the biosynthesis of lipid A as only very few mutants have been found that are affected in its biosynthesis. Mutations that affect biosynthesis or assembly of the KDO units and lipid A are lethal, and any information relating to that region comes from work with conditional mutants.

The chemistry of the lipid A moiety has also lagged because such complicated lipids are difficult to work with, and its structure was not correctly characterized until very recently (Sidorcyk *et al.*,1983; Strain *et al.*,1983a, 1983b Takayama *et al.*, 1983), although two classes of mutants had given some indication of the biosynthetic pathway. Mutants deficient in KDO biosynthesis accumulate a disaccharide precursor of lipid A, (Rick *et al.*, 1977). These mutants have been more rigorously characterized and analyses of accumulated intermediates in these mutants (Raetz *et al.*, 1985) suggest that addition of KDO occurs prior to the addition of the all the fatty acid chains and therefore prior to the completion of the lipid A structure. The addition of phosphatidylethanolamine via the ester bound pyrophosphate at position 4', and presumably 4-amino-4-deoxy-L-arabinose on the ester bound phosphate at position 1 (Raetz *et al.*, 1985) also occurs before all the fatty acids are added to

lipid A. The dissacharide is made with UDP-2,3,-diacylglucosamine and 2,3-diacylglucosamine 1-phosphate as intermediates. Anderson *et al.*, (1985) examined reactions in crude extracts of *pgsB* mutants and suggested a pathway for the biosynthesis of UDP-2,3,-diacylglucosamine. An important feature of this pathway is the precursor, UDP-*N* -acetylglucosamine, a compound known to be a precursor of peptidoglycan in both Gram-positive and Gram-negative bacteria. Enzymes that function in adding fatty acids to the lipid A precursors have very recently been characterized (Anderson and Raetz, 1987, Brozek *et al.*, 1987).

1.6.2. LPS biosynthesis - The Inner Core.

Some details of the structure of the LPS core oligosaccharide of the salmonellae are uncertain and some substitutions of the basic structure are known to be incomplete, so that the sugar unit (or other group) concerned is not present in stoichiometric proportion. Furthermore, the extent of a particular substitution may be altered, either by mutation or by conditions of growth. With these limitations the structure of the core is said to be the same in Salmonella of all serotypes, or, at least, for all those falling in the major subgenus I.

KDO is added before the lipid A is complete (Raetz *et al.*,1985) and its linkage to lipid A is not dependent on full substitution of the *N*-acetylglucosamine dimer with fatty acid chains. The third KDO, if it exists, requires more than just two KDO molecules as the receptor for its transferase (Strain *et al.*, 1983a, 1983b) and may need phosphate or ethanolamine substitutions (Prehm *et al.*, 1975). The KDO part of the molecule is genetically very poorly known; only

mutants of two genes participating in the synthesis of the KDO donor have been isolated. Phosphate and ethanolamine phosphate occur as branches on this section of the core, but in respect to them no mutants are available. The linkage of the KDO to lipid A has been identified recently (Strain *et al.*, 1983a, 1983b, Takayama *et al.*, 1983, Sidorczyk *et al.*, 1983) as an ester linkage to the C-6' residue. The two known mutants with defects in the formation of the KDO region have blocks in a pathway that seems likely to use D-ribulose-5-phosphate as a precursor for the biosynthesis of CMP-KDO, which is known to be the immediate precursor for assembly on the core (Heath *et al.*, 1966). One of the mutants is thought to be defective in the structural gene of KDO-8-P synthetase, *kdsA*, and has been mapped at position 39 (Lehmann *et al.*, 1977) and the second is a mutant in the KDO-CMP synthetase gene and is known as *kdsB*, which maps at position 14-17 (Sanderson and Roth, 1983). These mutants are conditional lethal mutants and presumably the synthesis of the LPS molecule to the lipidA plus KDO stage is essential for the survival of the cell, this would account for the lack of mutants.

Biosynthesis of heptose (L-glycero-D-*manno*-heptose) is characterized by more mutants (*rfaC, D, E* and *F)* (Mäkelä and Stocker, 1984), and also the more recently isolated *rfa2* mutant (Coleman and Deshpande, 1985). All but one of them, *rfaE*, are in the cluster of *rfa* genes that map at position 79. The phenotypes resulting from these mutations, however, have allowed no more than the fairly reliable identification of two of these (*rfaF* for the transferase of heptose II and *rfaD* for ADP-heptose-epimerase) (Mäkelä and Stocker, 1984). The precursor of heptose addition

appears to be an ADP derivative of the sugar (Kontrohr and Kosics, 1981; Coleman and Leive, 1979; Colemman, 1983). The biosynthesis of this sugar is postulated to derive from Sedoheptulose-7-phosphate, isomerised to D-glycero-D-*manno*-heptose-7-phosphate, converted to D-glycero-D-*manno*-heptose-7-phosphate, made into ADP-D-glycero-D-*manno*-heptose, and then epimerised to ADP-L-glycero-D-*manno*-heptose (Shibaev *et al.*, 1986). Two heptoses form part of the main sugar chain and a third heptose, a phosphate, and ethanolamine diphosphate are all present as branches to either heptose II or I, in most cases in non-stoichiometric proportions. Mutants unable to add glucose I of the outer core to heptose also lack the branch heptose; an *rfaP* mutation in *Salmonella minnesota* abolishes all these branches as well as the completion of the core beyond the addition of glucose I; presumably these substitutions are made after the addition of the distal glucose (Mäkelä and Stocker, 1984).

1.6.3. LPS Biosynthesis - The Outer Core.

The structure of the outer part of the core polysaccharide varies somewhat in strains of ~~*Salmonella*~~ salmonellae, and *E. coli* (Jann and Jann, 1984). Essentially the same sugars are used but they are arranged differently. The biosynthesis of the outer part of the core region of at least *S. typhimurium* is better understood than that of the inner region. The intermediates of the substituent parts are mostly common cellular components, UDP-glucose, UDP-galactose and UDP-*N*-acetylglucosamine. These compounds have functions in the cell economy unrelated to LPS core biosynthesis, and structural genes for the enzymes for biosynthesis of these compounds are not closely linked to genes concerned only with the LPS core synthesis. Each of

the five transferases required for the assembly of this part of the core has been identified through mutations in the corresponding gene; however a temperature sensitive mutation in *rfaI* (Mäkelä and Stocker, 1984) provides the only unambiguous identification. All of these genes (*rfaK, J, I, G* and *B*) are located in the main *rfa* cluster in the region between *cysE* and *pyrE* at position 79. These genes, apart from *rfaK*, have been cloned from *E. coli* K12 (Creeger and Rothfield, 1982) and from *S. typhimurium* (Kadam *et al.*, 1985) and their gene order has been determined, but it is not known whether they are organized into a single operon.

There are no major uncertainties about the structure of this part. A minor point concerns an acid-labile substituent postulated on the branch galactose (II) (Funahara and Nikaido, 1980). Although O-acetyl substitution has not been described for the core of *S. typhimurium*, such substituents have been found in some other Salmonella strains (Hellerqvist *et al.*, 1969; Rudén and Mäkelä, 1974).

It appears that biosynthesis of this region is under positive regulatory control. The product of the *rfaH* gene in *S. typhimurium* acts as a positive regulatory protein, probably as an antiterminator, of the expression of the transferase genes *rfaJ* and *rfaK*, for the formation of the glucose II and *N*-acetylglucosamine units, and probably also for gene *rfaL*, necessary for attachment of O polysaccharide. Termination in the operon affected is not complete, so that even in the absence of *rfaH* some read-through occurs, and genes distal to the first termination loop still get expressed, although in lower amounts. A heterogenous population of LPS core

molecules is found in *rfaH* mutants, reflecting this read-through (Creeger *et al.*, 1984).

1.7. O-Chain Biosynthesis.

The structure of the O-unit of *S. typhimurium* has a backbone of galactose-rhamnose-mannose with abequose attached to the mannose in an alpha 1-2 linkage. The sugar components needed in the O-unit are made as dinucleotide sugar precursors and are transferred to a lipid carrier in the cytoplasmic membrane by specific transferases (Robbins and Wright, 1971). The O-units are assembled onto this lipid carrier, and then one O-unit is transferred onto another, releasing one of the lipid carriers while extending the chain of another. After successive rounds of extension the polymerized O-polysaccharide is transferred to the glucose II residue of the outer core.

1.7.1. O-Chain Biosynthesis - Acyl lipid carrier.

ACL is needed in the biosynthesis of peptidoglycan, yeast cell wall mannan and bacterial capsules as well as in O-antigen synthesis (Goldmann and Strominger, 1972; Scher *et al.*, 1968; Troy *et al.*, 1971, 1975; Weisgerber and Jann, 1982). ACL has been isolated from *S. anatum*, and its structure has been determined by mass spectroscopy to be the monophosphoester of undecaprenol, a C_{55}-polyisoprenoid alcohol (Wright *et al.*, 1967). ACL is synthesized in the cell from mevalonic acid as a pyrophosphate monoester (Hemming, 1970).

A key enzyme in the use of the ACL is a membrane-bound phosphatase, before ACL-P-P can enter into the biosynthetic

pathway of polysaccharides, it is dephosphorylated, and on recycling of ACL the ACL pyrophosphate (ACL-P-P) is dephosphorylated to ACL-P by the specific membrane-bound phosphatase (Weiner *et al.*, 1965; Robbins and Wright, 1971). This enzyme, which is also involved in the biosynthesis of peptidoglycan (Ghuysen and Shockman, 1973) is inhibited by the peptide antibiotic bacitracin (Siewert and Strominger, 1967; Stone and Strominger, 1971). In the presence of bacitracin, ACL-P-P accumulates and O-antigen polymerization, as well as transfer to the core, is inhibited.

1.7.2. O-Chain Biosynthesis - Transferases.

In *S. typhimurium* the galactose is bound to the ACL with a diphosphate linkage, one of the phosphates of the UDP-galactose contributing to this bond, whereas all subsequent sugars are added to the O-unit without incorporation of one of their phosphates into the product (Osborn and Tze-Yuen, 1968; Robbins and Wright, 1971; Wright *et al.,* 1967). The transfer of galactose phosphate into the assembly is reversible (Nikaido and Nikaido, 1971; Nikaido *et al.,* 1971) also in contrast to subsequent transfers. The specificity of the rhamnosyl transferase has been studied in some detail (Shibaev, 1978, Shibaev *et al.*, 1982, Danilov *et al.*, 1980, 1981) and a degree of non-specificity is found; the rhamnose residue can be added to fucose and talose, as well as galactose attached to lipid carrier. Presumably the other transferases will have a similar degree of non-specificity as far as the acceptor molecule is concerned. As judged by intergenic crosses, genes for these transferases map in the *rfb* gene cluster.

1.7.3. O-Chain Biosynthesis - polymerization.

Two different systems are known for the polymerization of the O-units; in *Salmonellae* of groups A, B, D, and E the *rfc* gene is required, whereas *Salmonellae* of groups C1 and L and *E. coli* 08 and 09 strains need a functional *rfe* gene (Mäkelä *et al.*, 1970; Stocker and Mäkelä, 1978). The *rfe* gene also plays a role in the biosynthesis of the enterobacterial common antigen (ECA) (Makelä and Mayer, 1974). Strains which depend on a functional *rfe* gene for O-antigen expression do not need the *rfc*-directed oligosaccharide polymerase.

In the group of bacteria characterized by *S. anatum*, the polysaccharide chain is added on top of the newly synthesized O-unit which is linked to ACL, that is to say that it is elongated at the reducing end (Bray and Robbins, 1967). In *S. typhimurium* the O-unit is polymerized as a tetramer, abequose being transferred to the oligosaccharide before polymerization (Yuasa *et al.*, 1969, Kanegasaki and Wright, 1970; Osborn and Weiner, 1968; Weiner *et al.*, 1965). Glucosylation of the galactose residue and O-acetylation of the abequose occur at a later stage and are thus considered to be modifications. The polymerase in this system is not absolutely specific, lipid-linked oligosaccharides generated in the membranes of *S. anatum*, *S. typhimurium* and *S. senftenberg*, which contain analogs of galactose or mannose, can act as substrates of the polymerase (Shibaev, 1978; Shibaev *et al.*, 1982).

Most of our understanding of the *rfe* dependent systems relies on the analysis of *E. coli* 08 and 09 antigen biosynthesis. Alpha-glucose-1-phosphate is reversibly transferred to ACL-P, followed by

irreversible transfer of alpha-mannose residues directly from GDP-mannose, in a single chain mechanism; it doesn't seem to need transfer of O-units from one ACL to another to elongate the O side chain. Since only one molecule of ACL-P is needed per mannan chain, the polymerization in not as sensitive to inhibition by bacitracin as the *rfe* independent system (Siewert and Strominger, 1967; Stone and Strominger, 1971).

The regulation of polymerase activity is not well understood. The chain lengths of the products were thought to be in the region of 10-30 repeating units (Nikaido, 1973) but SDS-PAGE patterns of complete O-antigens show great heterogeneity with chain length ranging from no O-chain attached to cores with much more than 40 repeating units (Goldmann and Leive, 1980; Jann *et al.*, 1975; Palva and Mäkelä, 1980). A high proportion (up to 40 or 50%) of LPS molecules have only core or one O-antigen repeat unit (Hitchcock and Brown, 1983).

The substitution of the O-antigen with alpha-glucose also takes place at the level of the membrane-associated, ACL-linked polysaccharide (Mäkelä, 1973, Takeshita and Mäkelä, 1971; Nikaido *et al.*, (1971). A glucosylated lipid is an intermediate which has been characterized as beta-glucosyl-P-ACL (Nikaido and Nikaido, 1971). Here, glucose is transferred to ACL-P, from UDP-glucose in an manner different to the transfer of galactose-1-phosphate as the first sugar of the O-unit. The transfer of galactose-1-phosphate from UDPGal occurs with the retention of the anomeric galactosyl configuration, that of glucose from UDPGlc occurs with inversion of this configuration to give an alpha glucosyl susbstitution.

1.7.4. O-chain Biosynthesis - Translocation.

Two genes are involved in the translocation of the completed core to the polymerized O-unit, *rfbT* from the *rfb* gene cluster, and *rfaL* from the *rfa* gene cluster. The O-antigen is attached to the C-4 position of the subterminal glucose of the core (Nikaido, 1970; Yuasa, 1970) by an enzyme(s) that has (have) not been characterized. The galactose unit of the O-specific polysaccharide linked to the core-lipid A moiety is not glucosylated in postpolymerization (O-antigen) modification (Nikaido, 1973). Translocation has been achieved *in vitro*, and has been studied in cell free systems with a *S. typhimurium* mutant defective in sugar activation (Nikaido, 1965), in a *galE* mutant of *S. typhimurium* (Osborn and Weiner, 1986) and in wild-type bacteria of *S anatum* and its bacteriophage-converted derivatives (Losick and Robbins, 1967; Bray and Robbins, 1967; Robbins and Wright, 1971). In *S. anatum*, during the translocation, the last molecule of ACL participating in the biosynthesis of one O-antigen molecule is liberated in the form of its pyrophosphate derivative. The energy expenditure of the translocase reaction is covered by the cleavage of the glycosyl-pyrophosphate bond with the formation of a glycosyl bond of lower energy content.

1.7.5. O-chain biosynthesis - Transport across to the Outer membrane.

The transport of the LPS molecule from its place of synthesis on the inner side of the cytoplasmic membrane to the outer surface of the outer membrane is poorly understood. Mühlradt *et al.*, (1973) demonstrated that newly formed O-antigen appeared after about 2

minutes at about 200 to 300 distinct patches over the entire cell surface corresponding to regions of the bacterial cell surface where the cytoplasmic and the outer membranes come in contact, known as Bayer's junctions. The assembly of the LPS of *S. typhimurium* starts on the cytoplasmic face of the inner membrane (Osborn *et al.*, 1972b) and the completed molecule can be localized on the periplasmic face of the inner membrane, at a later time in its biosynthesis (Mulford and Osborn, 1983). It is thought that KDO is needed for transport across the membrane perhaps as a recognition signal for translocation. This hypothesis is based mainly on evidence that this is the minimal sized LPS known to translocate (Jones and Osborn, 1977, Knox *et al.*, 1967, Walenga and Osborn, 1980, Osborn *et al.*, 1980). Very recent work suggests that LPS transport is an energy dependent process requiring maintenance of both proton motive force and high energy phosphate pools (Marino *et al.*, 1985).

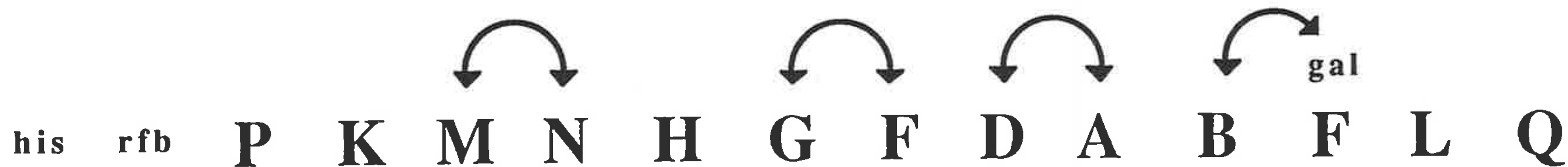
1.8. The *rfb* locus.

Most of the enzymes used in the biosynthesis of an O-unit are encoded by one gene cluster, the *rfb* gene cluster at position 42 on the *S. typhimurium* chromosome. Transconjugants of crosses exchanging the *rfb* regions between *Salmonellae* of different O groups usually results in a change of the O-antigen expressed (Kochibe, 1970, Kishi and Iseki, 1973a, 1973b, Mäkelä, 1965, 1966, Nikaido *et al.*, 1966, Jousimies and Mäkelä, 1974), thus the *rfb* region must contain all the specific information for making the O-antigen. Occasionally, recombinants were found to have rough type LPS; in these a crossing-over probably occurred within the *rfb* cluster, resulting in an incompatible combination of monosaccharide

FIGURE 1.1

LOCATION OF *rfb* GENES DEFINED BY *S. typhimurium* STRAINS WITH DELETION ENDPOINTS IN THE *rfb* REGION

Eight strains are represented, being a selection of those used in generating a genetic map of the *rfb* genes in *Salmonella typhimurium*. The solid horizontal lines represent DNA missing from the chromosomes of the strains with deletion mutations. The dotted lines to the right side of some of these solid lines represent the extent to which chromosomal DNA may be missing. A strain number is given for all of the strains represented as well a *his* number, defining the deletion isolate. The row of letters on the top of the diagram are letters representing gene functions as defined in the early genetic mapping work (Nikaido *et al.*, 1967, Levinthal *et al.*, 1969). The arrows between the symbols for gene functions indicate that the order of the two gene functions has not been determined.



his1462, P9077

his660, P9033

his515, P9031

his801, P9034

his695, P9074

his809, P9035

his101, P9051

his388, P9029

synthesis and transfer genes (Nikaido *et al.*, 1966).

1.8.1. Gene order, and direction of transcription.in the *rfb* locus.

A partial gene map of the *rfb* gene cluster has been determined with the help of a series of chromosomal deletion mutants extending from the *his* operon into and past the *rfb* cluster (Nikaido *et al.*, 1967) (Fig 1.1). Intermediates in pathways for the nucleotide sugar synthesis were assayed in these mutants, and the progressive loss of function indicated the relative position of genes encoding these functions. This work suggests that genes of individual pathways encoded exclusively by the *rfb* gene cluster are grouped together, all abequose functions being together as are all the rhamnose functions, suggesting that the *rfb* region may be a group of operons.

The expression of the *rfb* genes in deletion mutants that have the *his* operon fused to the *rfb* operon, suggests that the direction of transcription in the *rfb* gene cluster is towards the *his* operon (Levinthal and Nikaido, 1969). This early genetic work covered only part of the *rfb* gene cluster, the ends were not characterized, and no clues are given as to the number of operons involved in this cluster (Fig 1.1). Although it does seem clear that the region between and including genes *rfbA* and *rfbM*, is read towards the *his* operon, other parts may be read in the opposite direction.

The map produced does not order the genes completely, as not quite enough deletion mutants were available. It should be remembered that mutations in several of the genes directing the synthesis of the

O unit can be growth-inhibiting in indirect ways and therefore it may not be possible to have deletion endpoints in all of the genes. A difficulty in characterizing an unknown gene cluster is determining the number and function of the constituent genes; it is quite likely that other enzymes are present in this region that have unknown function, and thus could not be located by enzyme assay. Some functions such as the transferases for CDP-abequose and GDP-mannose have definitely not been mapped. The genes encoding these transferases are encoded in the *rfb* region as shown by the assembly of O-units in transconjugants in crosses between O groups B, C_1, D, E, G, L, R and U of the ~~*Salmonellae*~~ salmonellae (Kochibe, 1970, Kishi and Iseki, 1973a, 1973b, Mäkelä, 1965, Mäkelä, 1966, Nikaido *et al.*, 1966, Jousimies and Mäkelä, 1974). A gene required for the transfer of galactose phosphate, *rfbP*, is located at the right hand end of the *rfb* cluster closest to *his* (Yuasa *et al.*, 1969). A point mutation defective in the transfer of rhamnose from dTDP-rhamnose to the precursor served to define and locate the *rfbN* gene (Levinthal and Nikaido, 1969). The genes for the other transferases have however not been located. In addition *rfbT*, the *rfb* gene involved in the translocation of the O-antigen onto the core, has not been mapped in the deletion analysis but has been mapped in this region by the intergenic crosses.

Some other features of the region need clarification; the so-called 'X', 'Y' and 'Z' regions are hazy in their definition, they may be control regions or structural proteins. The 'Z' region functions in some unknown way on dTDP-rhamnose. Deletions covering locus 'Y' caused increased production of phosphomannomutase, whereas those covering locus 'X' led to increased levels of dTDP-rhamnose

oxidoreductase activity (Nikaido *et al.*, 1967). Because in the last-mentioned deletions gene *rfbB*, the presumed structural gene for this enzyme, was also deleted, the activity in the X mutants was thought to be based on a different structural gene determining a second enzyme with at least partial dTDP-rhamnose oxidoreductase activity. This gene would most likely be located outside the *rfb* cluster since the enzyme was present in strains with the longest *rfb* deletion. The present knowledge of the structure of ECA (Lugowski *et al.*, 1983) suggests that it might be part of the *rfe-rff* cluster concerned with the synthesis of ECA (Mäkelä and Mayer, 1976).

One of the major difficulties in working on this cluster has been the difficulty in getting stable point mutations; very few exist and point mutants tend to develop secondary mutations in *rfbP* or *galE*, the instability seems to be related to the fact that of the ACL is shared with peptidoglycan biosynthesis (Anderson *et al.*, 1972). Mutants in O-antigen biosynthesis apparently deplete the pool of ACL available for use in the biosynthesis of peptidoglycan, because incomplete O-units are bound irreversibly to this carrier, the cell wall thus becomes quite fragile. Mutants with secondary defects grow better than original mutants of the O-antigen biosynthesis and thus cultures with secondary mutations tend to be maintained. The result of this drawback is that conventional genetic studies have, to date, been impossible in the *rfb* gene cluster, and consequently the genetics of this region has lagged behind that of other regions more amenable to study. The deletion mutants in which the deletion started in the *his* operon and extended to varying lengths of the *rfb* did not appear to have the problem of poor growth due to accumulation of incompleted O units linked to undecaprenol,

because fortuitously the *rfbP* gene is located at the end of the *rfb* cluster closest to *his* (Yuasa *et al.*, 1969), and thus the deletion mutants would not even start the assembly. Unexpectedly, longer deletions extending midway into the *rfb* cluster were sensitive to bile acids and sodium dodecylsulphate (SDS) and accumulated secondary mutations, which this time were located in the *rfe-rff* cluster (map position 84) (Mäkelä *et al.*, 1976). In this case the deletion eliminated the genes for the first two enzymes of the rhamnose pathway needed to convert glucose-1-phosphate to dTDP-4-keto-6-deoxyglucose; thereby preventing the synthesis of dTDP-amino-dideoxygalactose (thomasamine), a component of ECA. If this was not the first residue of ECA assembly, its absence could deplete the pool of undecaprenol phosphate carrier with an incomplete oligosaccharide unit in a manner analogous to that described above for incomplete O-unit assembly; this occurrence could be alleviated by an *rff* mutation blocking the synthesis of *N*-acetylmannosamine-uronic acid (Lew *et al.*, 1978).

1.8.2. The *rfb* locus - Biosynthetic pathways.

The *S. typhimurium rfb* region encodes only the genes needed to complete the biosynthesis of the four nucleoside precursors of the O-antigen. Enzymes that are necessary for biosynthesis of these four nucleosides but also take part in biosyntheses related to other aspects of the cell economy are usually encoded by genes present elsewhere on the chromosome and are not duplicated in the *rfb* region. The substrate glucose-1-phosphate can be taken as a common intermediate for all four sugars. In the case of galactose the transferase, and one function proported to modify an existing enzyme have genes in the *rfb* cluster, but intermediates of the

pathway to UDP-galactose are used in other cellular functions and the genes map elsewhere. In the case of mannose incorporation, more steps are encoded by the *rfb* cluster but they start with mannose-6-phosphate as a substrate. In contrast genes for all the known steps in the biosynthesis of both rhamnose and abequose are encoded by *rfb* genes.

1.8.3. The *rfb* locus - UDP-Galactose biosynthesis.

The genetic material required to synthesize UDP-galactose is outside of the *rfb* region in the *gal* operon at position 18. Two *rfb* genes are involved in the biosynthesis of the galactose portion of the O-unit in *S. typhimurium*, *rfbP* encodes the transferase that transfers UDP-galactose to the ACL, and *galF* (also part of the *rfb* gene cluster) participates in the synthesis of UDP-galactose in an indirect way. The enzyme UDP-glucose pyrophosphorylase (first enzyme of the UDP-galactose pathway) is a general housekeeping enzyme of the cell, and accordingly its structural gene *galU* (at map position 35) is not in the *rfb* cluster. The *galF* gene in this cluster, however, modifies UDP-glucose pyrophosphorylase as indicated by an altered electrophoretic mobility when compared to the form of the enzyme in a strain lacking the *galF* gene because of a deletion mutation in the *rfb* (Nakae and Nikaido, 1971a, 1971b; Nakae, 1971). It is not clear why such a modification is needed, since the end-product is the same with or without the *galF*, it is, however, hypothesized that this modification assists the use of the enzyme for purposes of O unit synthesis, in some unspecified way (Nakae, 1971).

1.8.4. The *rfb* locus - GDP-Mannose biosynthesis.

All the reactions for the biosynthesis of GDP-mannose are found in

E. coli K-12 as well as *S. typhimurium*. This biosynthetic pathway is used for the synthesis of GDP-L-fucose in *E. coli*, an intermediate in the biosynthesis of colanic acid (Markovitz, 1977). The first step in the GDP-mannose pathway is the isomerization of fructose-6-phosphate to mannose-6-phosphate catalyzed by the isomerase specified by the gene *pmi* at map position 32 (in *E. coli* the corresponding gene is called *manA*). This enzyme is also used in the reverse orientation for the metabolism of exogenous mannose. The next two enzymes in the pathway to GDP-mannose, phosphomannomutase (encoded by *rfbK* and *rfbL*) and guanosine diphosphomannose pyrophosphorylase (encoded by *rfbM*), are not required for the catabolism of exogenous mannose and are needed, as far as is known, for the synthesis of GDP-mannose only. As expected, the genes determining their synthesis are located in the *rfb* cluster. The enzyme guanosine diphospho-mannose pyrophosphorylase is subject to a feedback inhibition control exercized by the end-product of its pathway (Kornfeld and Ginsburg, 1966).

Phosphomannomutase occurs in multiple eletrophoretic forms as does UDP-glucose pyrophosphorylase (Nikaido *et al.*, 1967). The structural gene for this enzyme, *rfbL*, is located in the *rfb* cluster, and its deletion eliminates both forms, as expected. The gene *rfbK* probably modifies phosphomannomutase however the possibility that this represents a separate enzyme or that phosphomannomutase is composed of nonidentical subunits has not formally been excluded as has been done for UDP-glucose pyrophosphorylase.

1.8.5. The *rfb* locus - TDP-rhamnose biosynthesis.

Genes specifying the conversion of glucose-1-phosphate to dTDP-rhamnose all lie in the *rfb* cluster. The genes involved in this pathway are found in *E. coli* K12 as well as in *S. typhimurium*. Two of the identified intermediates, dTDP-glucose and dTDP-4-keto-6-deoxyglucose, are also intermediates on the pathway to dTDP-aminodideoxygalactose (Matsuhashi and Strominger, 1964), recently found (Lugowski *et al.,* 1983) as part of the glycolipid of the enterobacterial common antigen (ECA), (Mäkelä and Mayer, 1976, Mayer and Schmidt, 1979). Other genes involved in ECA biosynthesis are close to *ilv* (*rfe* and *rff* at position 84) (Lew *et al.*,1978).

The enzymes involved in the biosynthesis of dTDP-rhamnose have been reasonably well characterized, dTDP-glucose pyrophosphorylase (encoded by *rfbA*) requires magnesium as a cofactor and appears to be expressed constitutively. It is competitively inhibited by dTDP-rhamnose in *Pseudomonas aeruginosa* (Melo and Glaser, 1965) and it is inhibited by both UDP-glucose and TDP-glucose as well as dTDP-rhamnose in *E. coli* and *S. anatum* (Bernstein and Robbins, 1965). dTDP-oxidoreductase catalyzes the next stage and requires NAD as a cofactor, it has been crystalized (Wang and Gabriel, 1969) and functions in a dimeric form with an apparent M_r 78,000d, with subunits of M_r40,000d in *E. coli* B (Zarkowsky *et al.,* 1970). This enzyme occurs in multiple eletrophoretic forms in *S. typhimurium* (Nikaido *et al.,* 1967). The structural gene for this enzyme in *S. typhimurium*, *rfbB* is located in the *rfb* cluster, and its deletion eliminates both forms, the reason for the different electrophoretic forms is not known and the gene(s) responsible for the modification of the dTDP-glucose oxidoreductase

has not been identified. The dTDP-rhamnose synthetase is proposed to require two steps each requiring a separate enzyme. Two enzyme fractions are required, E-II, is postulated to be an epimerase which forms enzyme bound TDP-4-keto-6-deoxy-L-rhamnose, an intermediate that is stereospecifically reduced by E-I and NADPH to TDP-L-rhamnose, which is then released from the enzyme (Glaser *et al.*, 1972); these two functions were assayed as one in defining *rfbD* (Nikaido *et al.*, 1967; Levinthal *et al.*, 1969).

1.8.6. The *rfb* locus - CDP-abequose biosynthesis.

Abequose is a specific component of the O unit, in the O group B to which *S. typhimurium* belongs, in the O groups C_2 (O antigen 6,8 as e.g. in *Salmonella bareilly*), and C_3 of Salmonella (Lüderitz *et al.*, 1968), and in some serogroups of *Citrobacter* (Jann and Jann, 1978) and *Y. pseudotuberculosis* (Samuelsson *et al.*, 1974). CDP-abequose is synthesized from glucose-1-phosphate via a series of specific reactions. No other functions for these intermediates are known and, as expected, all the genes involved are in the *rfb* cluster.

Cytidine diphospho-D-glucose pyrophosphorylase catalyses the first step of the biosynthesis of abequose. It adds cytidine diphosphate onto glucose-1-phosphate to form cytidine diphospho-D-glucose (CDP-glucose), with the release of pyrophosphate; this enzyme requires magnesium and is encoded by the *rfbF* gene in *S. typhimurium* (Nikaido *et al.*, 1967). It has been purified and characterized most fully in *Y. pseudotuberculosis* type V as a protein, of M_r110,000d on a 5% acrylamide gel or 120,000d on Sephadex G150, and functions as a monomer (Rubenstein and Strominger, 1974b). This compares with a molecular weight of

approximately M_r100,000d found for the enzyme in *S. typhimurium* using Sephadex G200 (Chojnacki *et al.*, 1968). This enzyme is subject to feedback control by the end product of the pathway it initiates. In the case of the 3,6-dideoxyhexoses, the inhibitor is the CDP-3,6-dideoxyhexose, the last soluble form of the sugar before it is transferred to membrane bound intermediates in lipopolysaccharide biosynthesis (Nikaido and Nikaido, 1966; Mayer and Ginsburg, 1965; Kimata and Suzuki, 1966).

The second step, the conversion of CDP-glucose to CDP-4-keto-6-deoxy-D-glucose, is catalyzed by the gene product of *rfbG* in *S. typhimurium* (Nikaido *et al.*, 1967). This protein has been purified and characterized in *Y. pseudotuberculosis* type V. Cytidine diphosphate-D-glucose oxidoreductase requires NAD for activity (Matsuhashi *et al.*, 1966b) is M_r43,000d and functions as a dimer (Gonzalez-Porque and Strominger, 1972a).

What was thought to be the final step of the biosynthesis of CDP-abequose, catalysed by the *rfbH* gene product in *S. typhimurium* (Nikaido *et al.*, 1967), was found to use two steps with the involvement of three proteins (Matsuhashi and Strominger, 1967) in *Y. pseudotuberculosis*, and later in *S. typhimurium* (Pape and Strominger, 1969). CDP-4-keto-6-deoxy-D-glucose is converted to CDP-4-keto-3,6-dideoxy-D-glucose in *Y. pseudotuberculosis* by two proteins, E_1 and E_3. They have molecular weights of M_r61,000d and M_r41,000d respectively (Gonzalez-Porque and Strominger, 1972b). The mechanism of this step has been studied in some detail (Gonzalez-Porque and Strominger, 1972b, 1972c, Rubenstein and Strominger, 1974a). The E_1 protein adds pyridoxamine-5'-

phosphate to the substrate and the E_3 possess NAD(P)H oxidase activity working as the reductase releasing the pyridoxamine-5'-phosphate, a sulphydryl group on E_3 seems vital to the reaction. An *rfbH* mutant of *S. typhimurium* defective in the E_1 enzyme has been described (Yuasa *et al.*, 1969; Pape and Strominger, 1969), thus *rfbH* was redefined as the gene encoding enzyme E_1. The last step, for the biosynthesis of CDP-3,6-dideoxyhexoses from CDP-4-keto-6-deoxy-D-glucose in various serotypes of *Y. pseudotuberculosis*, is catalysed by the E_2 protein, which performs the required reduction step. This enzyme requires NADPH for activity but has not been purified and therefore has not been as well characterized (Matsuhashi and Strominger, 1967).

1.9. Aims of this thesis.

The only enzyme of the *rfb* cluster that has been purified from *S. typhimurium* is the *rfbF* gene product. Since the time of the last major genetic analysis of the *rfb* region new technical developments have occurred, particularly with the advent of molecular biology. These technical advances can overcome the problems faced by earlier workers in their analysis of the *rfb* region. The aim of the work in this thesis was to characterize, using some of these new techniques, the region in the *rfb* locus that directs the biosynthesis of CDP-abequose, using as a starting point some of the material already cloned from the *S. typhimurium* *rfb* region (Brahmbhatt *et al.*, 1986).

CHAPTER TWO

MATERIALS AND METHODS

2.1. Reagents.

Reagents were obtained from the following sources:- Acrylamide and Bis acrylamide grade from Biorad or BDH; Urea from BDH was used for sequencing gels; Sarkosyl, from Ciba-Geigy; mixed bed resin (MB-1), sodium dodecyl sulphate, from BDH; Caesium chloride was from BRL; LGT and HGT agarose from Seakem; nitrocellulose from Schleicher and Schuell; deoxy nucleotide triphosphates and dideoxy-nucleotide triphosphates from Boehringer Mannheim; Glucose-1-phosphate, Nicotinamide Adenine Dinucleotide and CDP-glucose from (Calbiochem); CTP, rATP, Dithiothreitol, ampicillin, kanamycin sulphate and tetracycine, from Sigma; chloramphenicol, from Calbiochem. DNaseI, lysozyme, Sigma; T4 DNA ligase, from New England Biolabs; large fragment of DNA polymeraseI (Klenow), and Bal31(slow), from IBI; T4 DNA polymerase, from Amersham, Restriction enzymes (Biolabs, Pharmacia, and Boehringer Mannheim); Restriction enzyme linkers, from Biolabs; proteinase K from Boehringer Mannheim; sequencing primer from BRESA. [^{32}S] methionine (1000 Ci/mM) from Amersham; [alpha-^{32}P] dCTP (2000Ci/mM) from BRESA. Xray film, from Kodak. All other chemicals were of analytical grade. Antisera to LPS determinants (04, 05, 02, 09), coupled to *Staphylococcus aureus* and uncoupled rabbit antisera, were the generous gift of Mr. Chris Murray of Institute of Medical and Veterinary Science Salmonella typing

laboratory, Adelaide. Goat anti-rabbit Horse radish peroxidase (Heavy and light chain IgG) was obtained from Nordic Immunology.

2.2. Growth media.

Double strength Difco nutrient Broth (Difco 003, 16g/L) with 5g/L NaCl was used routinely for growth of bacteria in broth cultures. Bacteria were generally grown on solid media made up of nutrient agar (Blood base agar, Difco, 0045) prepared without the addition of blood. Soft agar contained equal volumes of nutrient broth and nutrient agar.

Where a defined medium was required the minimal medium of Davis and Mingoli, (1950) was used. Carbon sources for minimal medium were glucose at 5mg.ml or maltose 10mg/ml. Growth factors were added to a final concentration of 20 micro g/ml, except vitamin B1 (thiamine) which was added to a final concentration of 1micro g/ml.

The "T" broth of (Karn *et al.*, 1980) was used for liquid propagation of lambda or recombinant lambda clones. The 2x"YT" of (Sanger *et al.*, 1980) was used in propagating phage M13 and its derivatives.

Antibiotics were used in nutrient media at the following final concentrations:- Tetracycline 16micro g/ml, Kanamycin 25 micro g/ml, chloramphenicol 25 micro g/ml and Ampicillin, 25 micro g/ml.

2.3. Bacterial strains, Bacteriophages and Plasmids used.

The bacterial strains and plasmids used are listed in tables 2.1 and

2.2 respectively.

Bacterial colonies were purified by streaking on solid media and phage plaques were purified by the overstreaking procedure. Bacterial cultures were routinely grown in nutrient broth shaking at 37 ^{o}C. Overnight cultures of JM101 were prepared in minimal medium supplemented with vitamin B1.

2.4. DNA techniques.

2.4.1.1. DNA isolation - Chromosomal.

Chromosomal DNA was prepared according to the method used by Manning *et al.*, (1986), where lysis of bacterial cells is achieved with the use of lysozyme treatment followed by sarkosyl/EDTA/pronase treatment, the lysed mixture is subsequently dialysed.

2.4.1.2. DNA isolation- small scale plasmid.

Plasmid DNA preparations were made according to the method of Kahn *et al.*, (1979) modified to improve its efficiency, by reducing the concentration of triton-X-100 used to lyse lysozyme treated cells by 5 fold, and not ether extracting the phenol extract.

2.4.1.3. DNA isolation- Large scale plasmid preparations.

Large scale purification of plasmid DNA was carried out by the two-step caesium chloride gradient method of Garger *et al.*, (1983).

2.4.1.3 DNA isolation - lambda.

Lambda and derivative recombinant lambda phage were propagated essentially according to the method of Karn *et al.*,

TABLE 2.1

BACTERIAL STRAINS USED IN THIS THESIS.

Strain	Characteristics	Reference/Source
***Salmonella typhimurium* LT2 derivatives**		
P9003	*hsdL trpB2 nmlH(b) flaA66 H2(enx) rpsL xylT404 ilvE452 metA22 hsdA*	Dr. I. Beacham
P9077	del *(hisD-rfbK)*-1462	(Nikaido *et al.*, 1967)
P9033	del *(hisD-rfbM)*-660	
P9031	del *(flaA-rfbH)*-515	
P9034	del *(hisO-rfbF)*-801	
P9074	del *(hisB-rfbA)*-695	
P9035	del *(hisO-galF)*-801	
P9051	del *(hisA-rfbL)*-101	
P9029	del *(hisO-metG)*-388	
LB5010	*metA22 metE551 trpD2 leu* r_{sa}^{-} m_{sa}^{+} r_{sb}^{-} m_{sb}^{+} r_{lt}^{-} m_{lt}^{+} *galE.*	Bullas *et al.*, (1983)
Various *Salmonella* strains		
S. dublin M6	J84 0551	S. Dixon
S. dublin M22	J84 5174	S. Dixon
S. typhi Ty21a	*galE*	Germanier and Furer, (1975)
S. paratyphi M8	IMVS 1316	S. Dixon

***Escherichia coli* K12 derivatives**

DH1	F$^-$ *gyrA96 recA1 relA1 endA1 thi-1 hsdR17 supE44* lambda$^-$	Hannahan, (1983)
JM101	*supE thi* del(*lac-pro*) [F' *traD36 proA+ proB+ laqIq* Z M15]	B. Egan Vieira, (1982)
P2039 (DS410)	*azi tonA lacY minA minB rpsL xyl mtl thi ara.*	Frazer and Curtis, (1975).

TABLE 2.2

PLASMIDS USED IN THIS THESIS

Plasmid	Characteristics	Source/reference
pUC18/19	cloning vector	Yannish-Perron *et al.*, (1985)
M13mp18/19	cloning vector	Norrander *et al.*, (1983)
pJRD158	cloning vector	Davison *et al.*, (1984)
pEMBL8/9	cloning vector	Dente *et al.*, (1984)
pPR328	cloning vector	Quigley and Reeves, (1987)
pcos2EMBL	cloning vector	Poutska *et al.*, (1984)
lambdaAD6	lambda 1059 based *rfb* clone	
pPR259	Ap^r; vector pUC8; with an *rfb* fragment.	Brahmbhatt *et al.*, (1986)
pPR300	Ap^r; vector, pJRD158; with *rfb* DNA from EcoRI (9.57) to XbaI (12.38).	
pPR301	Ap^r; vector, pEMBL8; with *rfb* DNA from EcoRI (0) to EcoRI (9.57).	
pPR302	Ap^r; vector, pEMBL8; with *rfb* DNA from EcoRI (0) to EcoRI (9.57).	
pPR303	Ap^r; vector, pEMBL8; BglII/BamHI *cutdown* of pPR301.	
pPR340	Ap^r, vector, pEMBL8; with	

	rfb DNA from EcoRI (9.57) to EcoRI (15.28).
pPR341	Cm^r; vector, pPR328; HindIII (15.28) to EcoRI (9.57) subclone of pPR340 (Fig. 3.3).
pPR398	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR399	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR400	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR401	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR402	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR403	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR404	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR405	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR406	Km^r; vector, pcos2EMBL; with cloned *rfb* DNA (Fig. 3.3).
pPR586	Ap^r; vector, pUC18; SphI (4.19) to SphI (14.85) *rfb* fragment cloned from pPR405.
pPR587	Ap^r; vector, pUC18; SphI (4.19) to SphI (14.85) *rfb* fragment cloned from pPR405.

pPR589	Ap^r; vector pUC18; BglII (7.57) to NruI (5.50) *rfb* fragment cloned from pPR586.
pPR590	Ap^r; vector pUC18; HpaI (4.87) to HpaI (6.78) *rfb* fragment cloned from pPR586.
pPR591	Ap^r; vector pUC18; HpaI (4.87) to HpaI (6.78) *rfb* fragment cloned from pPR586.
pPR592	Ap^r; vector pUC18; NruI (5.50) to EcoRI (9.57) *rfb* fragment cloned from pPR586.
pPR593	Ap^r; vector pUC18; HpaI (6.78) to HpaI (8.11) *rfb* fragment cloned from pPR586.
pPR594	Ap^r; vector pUC18; HpaI (6.78) to HpaI (8.11) *rfb* fragment cloned from pPR586.
pPR595	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR596	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR597	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR598	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR599	Ap^r; vector pUC18; Tn1725 derivative of pPR586.
pPR600	Ap^r; vector pUC18; Tn1725

	derivative of pPR586 (Fig. 6.1).
pPR601	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR602	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR603	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR604	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR605	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR606	Ap^r; vector pUC18; Tn1725 derivative of pPR586 (Fig. 6.1).
pPR873	Ap^r; vector pUC18; BglII (7.57) to NruI (5.50) *rfb* fragment cloned from pPR586
pPR874	Ap^r; vector pUC18; HpaI (8.31) to HpaI (9.54) *rfb* fragment cloned from pPR586.
pPR875	Ap^r; vector pEMBL8; Bal31 deletion derivative of pPR303 (Fig 4.1).
pPR876	Ap^r; vector pEMBL8; Bal31 deletion derivative of pPR303. (Fig 4.1)
pPR877	Ap^r; vector pEMBL8; Bal31 deletion derivative of pPR303 (Fig 4.1).
pPR878	Ap^r; vector pEMBL8; Bal31 deletion derivative of pPR303 (Fig 4.1).
pPR879	Ap^r; vector pEMBL8; Bal31 deletion derivative of pPR303 (Fig 4.1).

pPR880	Ap^r; vector pEMBL8; *Bal*31 deletion derivative of pPR303 (Fig 4.1).
pPR881	Ap^r; vector pEMBL8; *Bal*31 deletion derivative of pPR303 (Fig 4.1).
pPR882	Ap^r; vector pUC18; *Bgl*II (7.57) to *Eco*RI (9.57) *rfb* fragment cloned from pPR303.
pPR883	Ap^r; vector pUC18; *Bgl*II(7.57) to *Eco*RI (9.57) *rfb* fragment cloned from pPR303.
pPR884	Ap^r; vector pUC19; *Hin*dIII (15.28) to *Eco*RI (9.57) fragment cloned from pPR341.
pPR885	Ap^r; vector pUC19; *Mlu*I (15.54) to *Hin*dIII (15.28) *cutdown* of pPR884.
pPR886	Ap^r; vector pUC19; *Bgl*II (12.55) to *Hin*dIII (15.28) *cutdown* of pPR884.
pPR887	Ap^r; vector pUC19; *Xba*I (12.38) to *Hin*dIII (15.28) *cutdown* of pPR884.
pPR888	Ap^r; vector pUC19; *Hin*dII (10.67) to *Hin*dIII (15.28) *cutdown* of pPR884.
pPR889	Ap^r; vector pUC19; *Dra*I (11.40) to *Hin*dIII (15.28) *cutdown* of pPR884.
pPR890	Ap^r; vector pUC19; *Hpa*I(13.0)

	to HindIII (15.28) *cutdown* of pPR884.
pPR891	Vector, M13mp18; EcoRI (9.57) to PstI (3.27) *rfb* cloned from pPR301.

(1980) on the *E. coli* K12 strain DH1. Phage were concentrated by centrifugation overnight at 12,000xg for 16 hrs and DNA was prepared by the method of Young *et al.*, (1979).

2.4.1.4. DNA isolation-M13 ssDNA.

Single stranded DNA was prepared from a polyethylene glycol pellet of M13 or M13 derivative phage as propagated on JM101 essentially by the method of Sanger *et al.*, (1980).

2.4.2. Restriction endonuclease digestion.

DNA digestions of DNA were usually performed in one of three buffers; 'High Salt', 'Medium Salt' or 'Low Salt' buffers prepared according to Maniatis *et al.*, (1982). The choice of buffer used was based on the recommendation in the "New England Biolabs" catalogue available at the time. Digestion with restriction endonuclease Sma I was made according to manufacturer's specifications.

Usually 0.1 to 1micro g of plasmid DNA was digested with 2 units of enzyme at 37 ^{o}C for 1 hour. For chromosomal DNA digestions, 5-20 units of enzymes were used. Digestions were terminated where necessary either by heating the digestion mixture at 65 ^{o}C for 15 minutes or phenol extracting the mixture with tris-equilibrated phenol, followed by ethanol precipitation if heat stable enzymes were used.

2.4.3. Agarose/Acrylamide gel electrophoresis.

Restriction enzyme digested or undigested DNA samples were analysed on 0.7-1.0% horizontal agarose gels in TBE buffer. Smaller

fragments of DNA were separated on acrylamide gels in TBE buffer (Maniatis *et al.*, 1982). Gels were stained with ethidium bromide (0.5-1.0 micro g/ml) and the bands were visualized on a UV transilluminator. EcoRI digests of DNA prepared from bacteriophage SPPI were routinely used as markers for estimating the size of DNA fragments. pBR322 digested with HinfI was occasionally used as a marker for DNA fragments of smaller sizes.

2.4.4. Purification of DNA fragments-Agarose gels.

DNA fragments of sizes down to 0.5kb were purified from agarose gels in one of two ways. Electroelution in dialysis tubing from a fragment of gel, followed by ethanol precipitation, to concentrate the electroeluent (Maniatis *et al.*, 1982) was one method used. The preferred method was to cut fragments from gels of LGT (low gelling temperature) agarose after staining in ethidium bromide, followed by melting of the LGT agarose fragment and phenol extraction, as described in Maniatis *et al.*, (1982)

2.4.5. Purification of DNA fragments-Acrylamide gels.

Extraction of DNA fragments of up to a size of 1.5kb was possible from acrylamide gels. The restriction enzyme digested fragments were routinely endlabelled for this procedure, the gel was autoradiographed and bands located by the superimposition of the autoradiograph over the gel, usually after 4 hour exposure, bands were eluted as described by Maniatis *et al.*, (1982).

2.4.6. Endfilling and Endlabelling.

Endfilling of ends of DNA molecules was done essentially as described by Maniatis *et al.*, (1982). Where a 5' overhang was

created Klenow fragment of DNA polymerase was used in TM buffer with dA- ,dC-, dG- and dTTP, where a 3' overhang was made, T4 DNA polymerase was used to endfill it in T4 DNA polymerase buffer along with the four nucleotides. For endlabelling ^{32}P labeled dCTP was added to the DNA/buffer/enzyme mixture and incubated at 37 ^{o}C for 15 minutes prior to the addition of the four nucleotides.

2.4.7. Nick translation.

Nick translation of plasmid DNA or DNA fragments extracted from agarose gels was performed essentially by the method of Rigby *et al.*, (1977), but the DNA was first treated with DNaseI (10 micro g/ml) for 10 minutes at 37 ^{o}C.

2.4.8. Ligation of DNA.

Digested DNA samples were ligated in a 20 micro L volume, if a DNA fragment was to be ligated into another, or in 100-200 micro L volume if a plasmid was to be cut-down. For insertion of DNA fragments into a vector a molar ratio of approximately 1:3 of vector to insert was used. A typical ligation reaction contained 20 to 100ng of DNA, 1mM rATP, 10mM $MgCl_2$, 10mM dithiothreitol, 50mM Tris-HCl pH 7.4 and 1 unit (for sticky end ligations) to 40 units (for blunt end ligations) of T4 DNA ligase. Sticky-end ligations were carried out a 4 ^{o}C for 1-2 hours or overnight (14-18hrs.) blunt-end ligations were carried out at 4 ^{o}C overnight. This mixture was used directly to transform competent cells.

2.4.9. Transformation.

Bacterial strains were made competent and transformed with plasmid DNA routinely by the method of Lederberg and Cohen,

(1974).

2.4.10. Plaque hybridization Colony hybridization.

Colony and plaque hybridizations were carried out as described by Grunstein and Hagness, (1975) and Benton and Davis, (1977) respectively.

2.4.11. Bal31 deletion mutagenesis.

The method used for Bal31 deletion mutagenesis was the procedure described by Maniatis *et al.*, (1982), for isolating a range of deletion sizes using time sampling and ligating a restriction enzyme site linker between the deletion ends.

2.4.12. DNAseI deletions.

Deletions were made in M13 clones according to an adaptation of the method of Hong, (1982). DNaseI treated DNA was not purified from an agarose gel, and plaques were screened by running single stranded DNA on 0.7% agarose gels to ascertain the relative sizes of the deletions produced.

2.4.13 Tn1725 mutagenesis.

A temperature sensitive conjugative derivative of RtsI carrying Tn1725 was used to mobilize Tn1725 into target strains. An adaptation of the method of Ubben and Schmidt, (1986) was used to isolate transposon mutants of the target strains.

2.4.14. Cosmid Bank construction.

High molecular weight chromosomal DNA was partially digested with Sau3A1 to give fragment sizes in the range 35-45kb. These

fragments were cloned into the BamHI site of cosmid vector pcos2EMBL essentially as described by Poutska *et al.*, (1984) for pcos2EMBL cosmid libraries, but without use of phosphatase on the chromosomal DNA.

2.4.15. Sequencing method.

Sequencing was carrier out by the dideoxy method of Sanger *et al.*, (1977 and 1980).

2.4.16. Oligonucleotide synthesis.

Oligonucleotides were synthesized on an "Applied Biosystems" 381A DNA synthesizer; all reagents being supplied by Applied Biosystems. The concentrations of the synthesized oligonucleotides were measured by reading absorptions at 260nm and diluted for use as primers in sequencing reactions.

2.5. Enzyme assays.

2.5.1. Cell extracts.

100ml of cells were washed in 50mM Tris pH7.5, lysed by sonication, whole cell debris was spun out by a low speed spin and the membrane fraction was spun out at 30,000xg for 60 min. Cell extracts were used either immediately, or stored in small aliquots at -20 ^{o}C, or at 4 ^{o}C if storage was required for less than a week.

2.5.2. ABE-1.

CDP-D-glucose pyrophosphorylase (ABE-1) activity, was measured as a combined ABE-1 and ABE-2 assay, using a method adapted from that of Matsuhashi *et al.*, (1966b). The final product measured

was CDP-4-keto-6-deoxy-D-glucose, which absorbs light at a wavelength of 320nm. Reaction tubes had 50 mM Tris pH7.5, 10 mM $MgCl_2$ 1 mM cytidine triphosphate, 4.5 mM glucose-1-phosphate, 0.1mM nicotinamide Adenine dinucleotide, and 30 micro L of cell extract in a final volume 200 micro L. After the reaction mixture was incubated for 30 minutes at 37 ^{o}C 0.7ml of 0.1N NaOH was added, and left at room temperature for 15 mins. The mixtures was spun in an eppendorf 5414S centrifuge for 2 minutes to pellet the floculating organic debris, the absorbance of the supernatant at a wavelength of 320nm was then measured. A control without glucose-1-phosphate was always included to determine the background.

2.5.3. ABE-2.

The CDP-D-glucose oxidoreductase activity was measured in an assay adapted from that of Matsuhashi *et al.*, (1968b). Reaction tubes had 50 mM Tris pH7.5, 0.1mM cytidine diphospho-D-glucose, 0.1mM nicotinamide adenine dinucleotide, and 30 micro L of cell extract in a final volume 200 micro L. Incubation, developing and measuring the levels of CDP-4-keto-6-deoxy-D-glucose was done in the same way as the ABE-1 assay.

2.6. Protein techniques.

2.6.1. PAGE electrophoresis.

Proteins were analysed essentially as described by (Lugtenberg *et al.*, 1975) on 11-20% gradient SDS-Polyacrylamide gels. Samples were heated in a boiling water bath for 5 minutes before loading. Gels were stained with Coomassie Brilliant Blue G250 (0.06%, w/v)

in 3.5% perchloric acid and destained in 5% (v/v) acetic acid according to the method of Achtman *et al.*, (1978).

2.6.2 Fractionation of Cell envelope and Cytoplasm

Whole cell envelopes from 10ml bacterial cultures were isolated essentially by the lysozyme-sonication method described by Morona and Reeves, (1982) with the modification that after sonication of sphaeroplasts, unbroken cells were removed by low-speed centrifugation before envelopes were pelleted by high-speed centrifugation (30,000 xg, 1hr). The supernatant of this spin was retained as the cytoplasmic fraction.

2.6.3. Minicell methods.

Minicells were purified from DS410 and derivatives of that DS410 carrying a variety of plasmids on a sucrose step gradient and ^{35}S methionine was incorporated into proteins produced as described by Dougan and Kehoe, (1984). Labelled minicells were electrophoresed on 11-20% gradient SDS-Polyacrylamide gels. Gels were dried down at $80^{o}C$ onto whatman 3MM paper, and autoradiographed at room temperature, until sufficiently exposed.

2.7 Immunological techniques

2.7.1 Whole cell lysate

The procedure of Hitchcock and Brown, (1983) was used as a rapid means of preparing samples to electrophorese on acrylamide gels to detect LPS.

2.7.2 Western blot

The "Western Blot" procedure of Burnette, (1982) was followed. To inhibit non-specific binding of antibodies to the nitrocellulose sheets, they were soaked for 1 hr. in 5%(w/v) aqueous skim milk. The procedures of De Joungh-Leuvenik *et al.*, (1985) for adsorption of secondary antibodies and for developing the horse radish peroxidase assay were used.

2.7.3. Slide agglutination.

Agglutination of cells by antisera was usually tested on single colony isolates of the strains of interest, emulsified in a loopful of distilled water, and a loopful of antisera coupled to *S. aureus* cells was added to test agglutination. Positive and negative controls were included to check the validity of the test.

CHAPTER THREE

CLONING AND IDENTIFYING THE ABEQUOSE BIOSYNTHESIS REGION IN *S. TYPHIMURIUM*

3.1. Introduction.

This part of the work was done in parallel with another PhD student, H. Brahmbhatt who has extended some of the work presented in this chapter. Some overlap with the work of H. Brahmbhatt has occurred although the work diverged after the initial cloning described in this chapter and I will acknowledge where the results of his efforts are used.

The initial cloning of a part of the *S. typhimurium rfb* region has been described (Brahmbhatt *et al.*, 1986). The aim of the work in this chapter was to clone DNA that covered the region known to encode enzymes necessary for biosynthesis of abequose. We chose to initially define the extent of these genes by genetic means based on mapping the *rfb* endpoints of deletion mutants, extending from *his* into the *rfb* gene cluster (Nikaido *et al.*, 1967) (Fig. 1.1), in relation to a restriction map. The genes involved in abequose biosynthesis, as far as examined, appear to be grouped together with no genes for other functions among them (Nikaido *et al.*, 1967). If we, therefore, define deletion endpoints of the mutants affecting genes to either side of the abequose region we would define the abequose region. Strain P9033 has a deletion mutation (*his*660) that inactivates the *rfbM* gene involved in mannose biosynthesis, but not "*rfbH*", the *his* proximal gene(s) of the abequose gene group (Nikaido *et al.*, 1967).

By defining the deletion endpoint in the *rfb* region of strain P9033 the maximum possible extent of abequose biosynthesis genes on the left side of the rfb gene cluster has been defined. That endpoint has been shown to lie between 4.36kb and 4.52kb from the reference EcoRI site that is defined as position 0 Brahmbhatt *et al.*, (1986). Only DNA to the right of 4.36 is necessary for abequose biosynthesis because abequose biosynthesis has been demonstrated in P9033 (Nikaido *et al.*, 1967).

The deletion mutation in strain P9074 (*his695*) inactivates some of the rhamnose biosynthetic pathway as well as the abequose biosynthetic pathway, and by defining this endpoint we can define the *his* distal boundary of abequose genes in the region. The DNA inbetween the two endpoints should encode all the abequose genes in the region.

3.2. Cloning - using a lambda bank

Plaques from a *S. typhimurium* LT2 gene bank constructed using the vector lambda 1059 (Karn *et al.*, 1980) obtained as a kind gift from Dr. Russell Maurer (Maurer *et al.*, 1984) were probed with the MluI-HindIII (*rfb* map position 4.36-5.72) insert fragment of plasmid pPR259 (Brahmbhatt *et al.*, 1986). Phages from seven positive plaques were purified and an EcoRI digest of their DNA was probed with the same insert DNA from pPR259, this confirmed the positive result of the plaque hybridization. DNA preparations from the plaques were also nick translated and hybridized to DNA of the deletion strains (Fig. 1.1), using colony blotting. DNA from lambda AD6 hybridized to DNA from *S. typhimurium* wild-type strain P9003 and deletion strains up to and including P9034 (*his201*). This

recombinant phage was chosen for further analysis.

A preliminary restriction enzyme map of lambda AD6 was made using enzymes EcoRI, XbaI, NruI, KpnI, HindIII and SmaI. No SmaI sites were found and lambda AD6 was found to have all the restriction sites of pPR281 (Brahmbhatt *et al.*, 1986) in respect to these enzymes and had some additional sites indicating additional *rfb* DNA. For ease of analysis *rfb* DNA from lambda AD6 was subcloned. Preliminary restriction analysis indicated the presence of an insert containing two EcoRI sites and a single XbaI site where the EcoRI fragment was 9.57kb in length and the EcoRI-XbaI fragment was 2.81kb in length. Both of these fragments were subcloned into plasmid vector pEMBL8, pEMBL9 (Dente *et al.*, 1984) using the EcoRI site and into pJRD158b (Davison *et al.*, 1984; Heusterspreute and Davison, 1984) using the EcoRI and XbaI sites to yield plasmids pPR301, pPR302 and pPR300 respectively. Thus the inserts in pPR301 and pPR300 contained DNA from positions 0 to 9.57 and 9.57 to 12.38 respectively on the *rfb* map. A more thorough restriction map of the cloned pieces was collated from restriction enzyme analyses of pPR300, and pPR303; the latter being a BamHI-BglII cut-down of pPR301 (Fig. 3.1).

HindIII digests of DNA from chromosomal deletion strains were probed with a nick translated preparation of the plasmid pPR300. A HindIII band with an estimated length of 6.3kb from the chromosomal DNA of *S. typhimurium* deletion strains P9077 (*his1462*) and P9033 (*his660*) hybridized to the probe; P9031 (*his515*) has a band of much higher molecular weight (Fig 3.2), which is presumably a hybrid band bridging the deletion endpoints in *his*

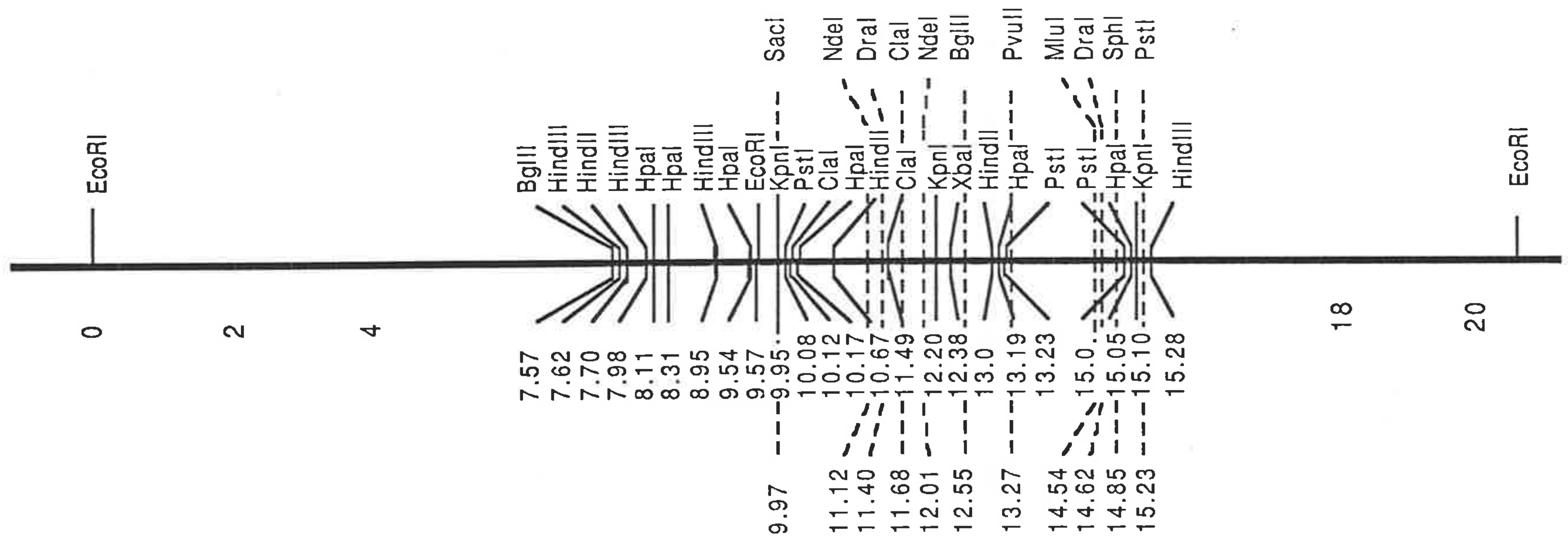
FIGURE 3.1

RESTRICTION MAP OF *rfb* DNA IN THE REGION ENCODING ABEQUOSE BIOSYNTHESIS

A is a map of restriction data assembled from restriction analyses of clones lambdaAD6, pPR301, pPR303, pPR300, pPR340, and pPR341. Numbers refer to the distance from the EcoRI reference site at position 0. Those sites named closest to the central line with solid pointer lines are derived from this analysis. Those sites further out from the central line with dashed pointer lines are sites derived from the analysis of H. Brahmbhatt.

B shows the extent of the *rfb* DNA contained within the clones cited in this chapter.

A



B

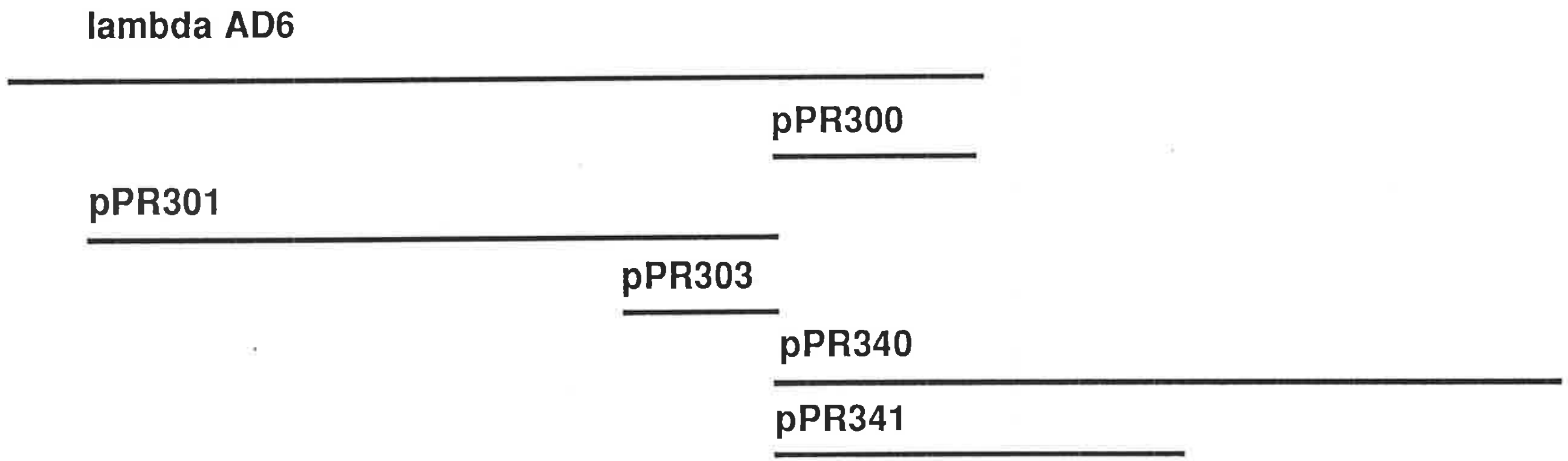


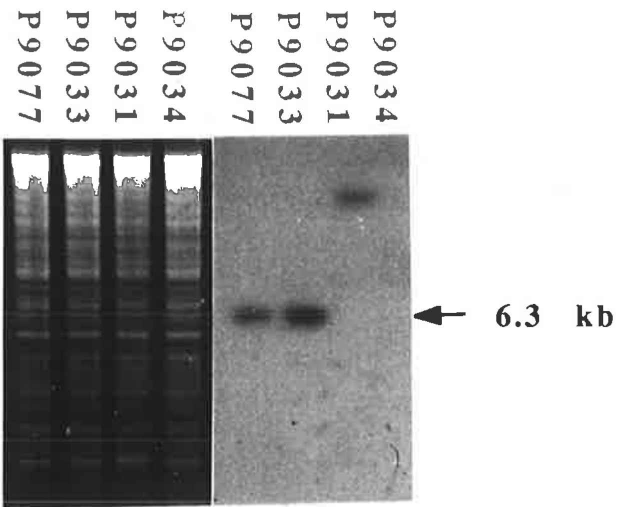
FIGURE 3.2

EXTENT OF CLONED *rfb* DNA IN RELATION TO DELETION MUTANTS OF *S. typhimurium* AS SHOWN BY SOUTHERN HYBRIDIZATION

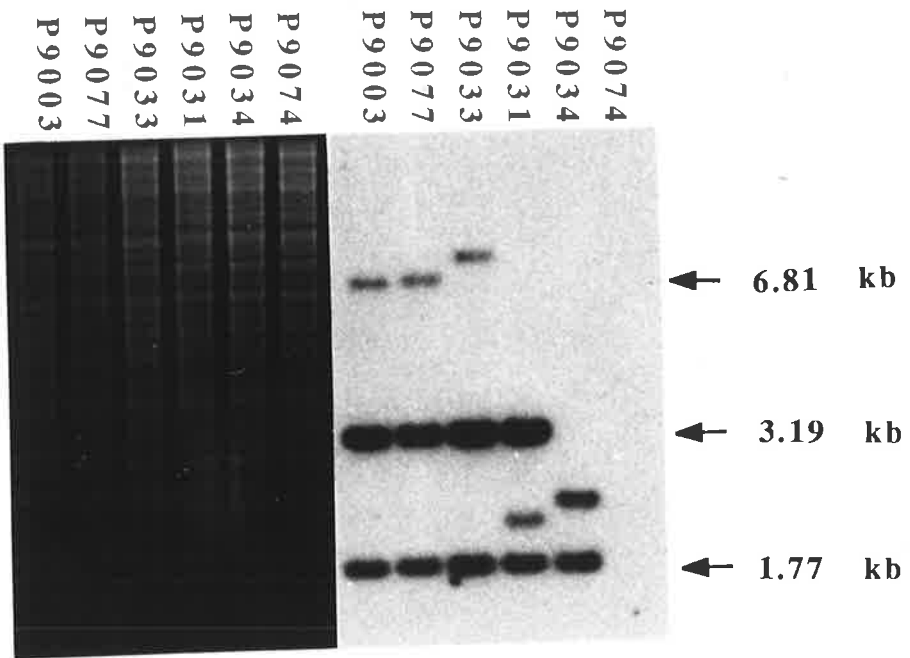
Shown are autoradiograms of southern blots of chromosomal DNA preparations of several deletion mutants extending into the *rfb* region from *his* (Fig 1.1) digested and electrophoresed on an agarose gel. Next to the autoradiograms are photographs of the Ethidium-bromide-stained gels before transfer of DNA to nitrocellulose paper. The DNA on the nitrocellulose paper was probed with one of three probes, and autoradiographed, photographs of the autoradiograms are shown. **A** shows HindIII digests of several deletion strains probed with ^{32}P labelled pPR300, **B** shows PstI digests of several deletion strains probed with ^{32}P labelled pPR341, and **C** shows HindIII digests of several deletion strains probed with ^{32}P labelled pPR340.

DIRECTION OF ELECTROPHORESIS ▼

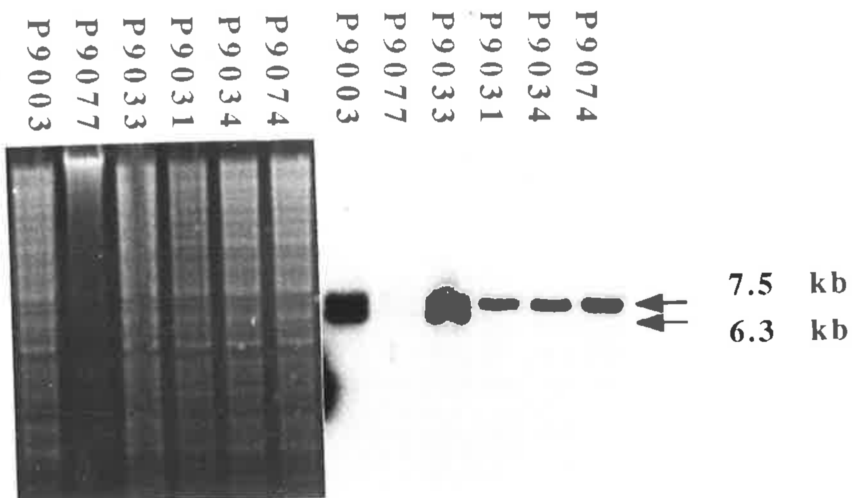
A



B



C



and *rfb* hybridized to the probe, thus the *rfb* end of the deletion in P9031 lies within the 6.3 kb HindIII fragment. No DNA fragment from P9034 DNA hybridized with the probe; presumably DNA from pPR300 does not extend past the deletion endpoint of this strain. Note that this is different to what is seen with the colony blotting experiment using lambda AD6 as a probe, a result that suggests the deletion endpoint of P9034 lies in the *rfb* DNA in lambda AD6 to the right of DNA cloned in pPR300. Restriction enzyme analysis suggests that the *rfb* DNA in this region is only about 0.1kb in length; possibly the P9034 deletion endpoint lies within that region. However we have no way of subcloning *rfb* DNA to include this fragment in a more manageable vector. None of the cloned DNA hybridized to P9074 (*his*695), therefore the amount of cloned *rfb* DNA must be extended in order to ensure inclusion of all of the region that encodes abequose biosynthesis genes.

3.3. Cloning - construction and use of a cosmid bank

A cosmid bank was constructed from *S. typhimurium* P9003 DNA, using pcos2EMBL as a vector (Poutska *et al.*, 1984). Colonies from the bank were probed with a purified 2.34kb KpnI fragment isolated from pPR300 (DNA from 9.95 to 12.29). A series of positive colonies were identified and their DNA purified. An initial restriction enzyme analysis of the cloned DNAs indicated that they could be arranged with respect to each other on the the basis of their common EcoRI fragments (Fig. 3.3). EcoRI digests of these clones were probed with the KpnI fragment that was originally used as a probe, and with pcos2EMBL to show which fragments correspond to *rfb* and vector DNA. An 11kb fragment hybridized to the KpnI fragment in 6 of the 9 cosmid clones examined and in 3 of the cosmids a smaller fragment

was probed (Fig. 3.3). An indication of the overall amount of DNA cloned is seen, more than 40kb of DNA to the right of map position 0 has been cloned. An intial, incomplete, restriction map has been constructed from this data (Fig. 3.3), starting from position 0. The four EcoRI fragments, between positions 20 to 24 are not ordered, but a 15.5kb fragment can be mapped distal to these four fragments; an 11kb fragment is just to the left of the cluster of four EcoRI fragments and the previously identified 9.57kb fragment is seen. Two other EcoRI fragments, of 8.9kb, and 9.5kb are seen to the left of position 0, we are not able to order these two, and no further attempts have been made to do so.

All nine of the cosmid isolates were transformed into LB5010 (Bullas *et al.*, 1983), a *S. typhimurium* strain, that has all three of the restriction systems inactivated, but the modification enzymes functional. This strain is used to increase the efficiency of transformation; ordinarily much of the DNA that is transformed into *S. typhimurium* from a non-*S. typhimurium* source is degraded by DNA restriction enzymes and therefore transformation efficiency is usually very low. However if a restrictionless strain is used the transformation efficiency is dramatically increased. After passage of the cosmids through LB5010 the cosmids were then transformed into P9029, a strain whose deletion (*his388*) extends beyond the rfb region (Fig 1.1). Ex-transformants from this experiment were tested for function using an anti-04 and an anti-05 antiserum; none of them were positive. It would appear therefore that none of the cosmids had carried all of the *rfb* cluster.

For ease of further analysis, an 11kb EcoRI fragment (positions 9.57

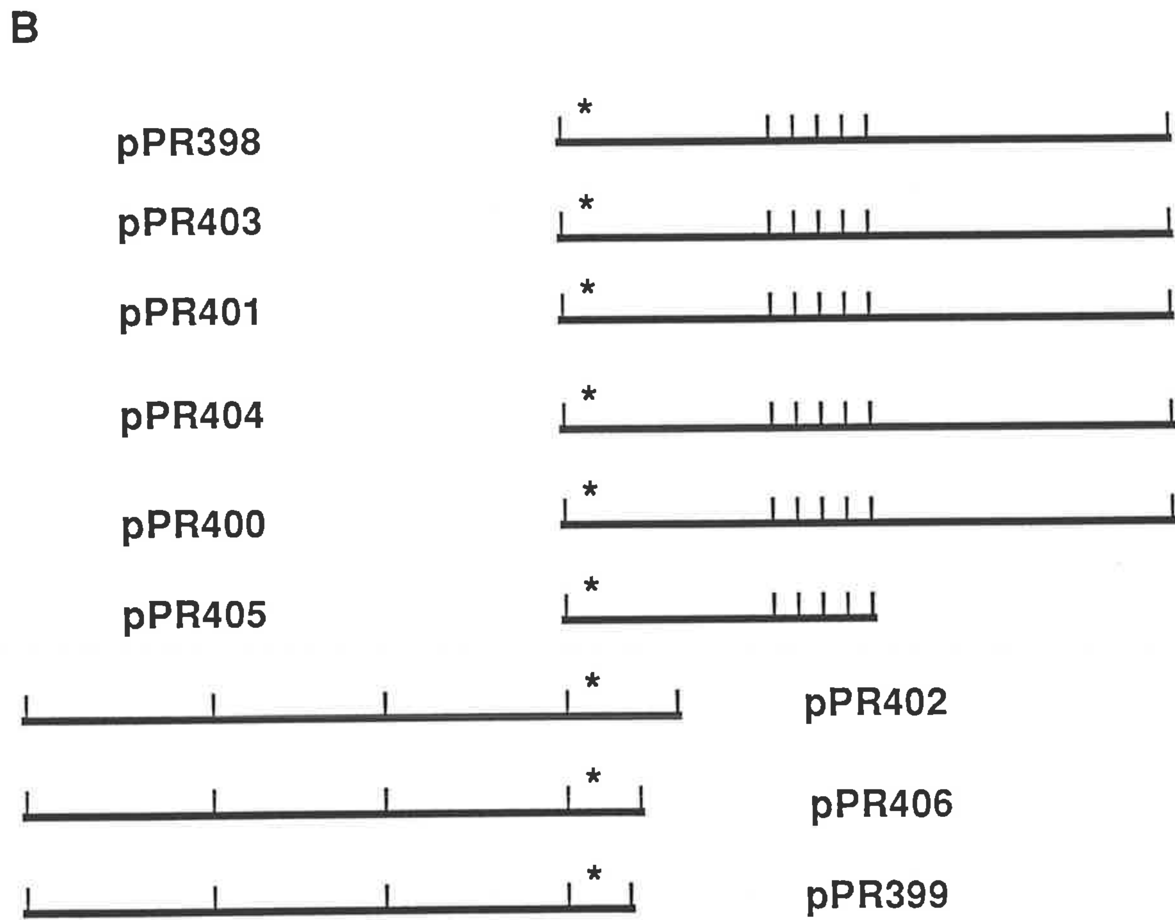
FIGURE 3.3

INITIAL RESTRICTION MAP OF *rfb* DNA CLONED IN SEVERAL COSMIDS AND THE EXTENT OF *rfb* DNA IN THOSE COSMIDS

A shows a map of the *rfb* and surrounding regions of *S. typhimurium,* derived from restriction analysis of 9 cosmid clones with DNA from that region. The fine vertical lines indicate EcoRI restriction sites. The numbers refer to distance in kilobases from the reference EcoRI site at 0. The two fragments to the left of 0 are 8.9 and 9.5 kb in size but their order has not been determined.

B shows the extent of the cloned material in each of the cosmid clones. The fine vertical lines again indicate the position of EcoRI sites. Other EcoRI DNA fragments are seen in all of the cosmid clones, but are not shown because they correspond to hybrid fragments that comprise part chromosome and part vector. Asterisks (*) in **B** indicate those restriction fragements that hybridized to the KpnI fragement (9.95 to 12.29) that was used as a probe.

A
0
9.57
20.6
24.4
40.1


B
pPR398
pPR403
pPR401
pPR404
pPR400
pPR405
pPR402
pPR406
pPR399

to 20.59) was subcloned from pPR403 into the EcoRI site of plasmid vector pEMBL8 to give pPR340, this being the most convenient fragment to clone. HindIII digested whole cell DNAs of the deletion strains (Fig 1.1) were probed with pPR340 after gel electrophoresis and transferal to nitrocellulose by the method of Southern, (1975). Nick translated pPR340 probed to two DNA fragments in tracks with DNA from *S. typhimurium* P9003, P9077 (*his1462*) (this is not seen clearly in figure 3.3 because of a poor restriction digest for this DNA) and P9033 (*his660*); the 6.3kb fragment HindIII seen in DNA cloned from lambda AD6 and another 7.5kb HindIII fragment (Fig. 3.2). Only the 6.3kb fragment was seen in tracks with DNA from P9031 (*his515*), P9034 (*his801*) and P9074 (*his865*), and no other deletion strains had homology with pPR340. P9074 (*his895*) had an intact 7.5 kb HindIII fragment indicating that the endpoint of this deletion lies within the 6.3 kb HindIII fragment. Therefore all the genes related to abequose biosynthesis, in this region, lie to the left of the distal HindIII site and therefore DNA cloned in pPR340 extends well beyond the genes encoding abequose biosynthesis.

An initial restriction enzyme analysis was made of pPR340, using enzymes NruI, HincII, HindIII, PstI, BglII, HpaI, BamHI and ClaI; not all restriction enzyme sites were placed. The unique HindIII site shown in chromosomal DNA by probing, in the previous section, was mapped to position 15.28 and was used to subclone the 5.71kb EcoRI-HindIII fragment (positions 9.57 to 15.28) from pPR340 into plasmid vector pPR328 (Quigley and Reeves, 1987), the chloramphenicol resistance marker of which was used for ease of subcloning, the resultant plasmid being pPR341. This plasmid was used to produce a more detailed restriction endonuclease map (Fig 3.1). Note that this

restriction map has been extended by H. Brahmbhatt and a map has been published jointly (Brahmbhatt *et al.*, 1988).

PstI digested DNA from deletion strains (Fig. 1.1) was probed with nick translated pPR341 after gel electrophoresis and transfer to nitrocellulose by the method of Southern, (1975). Three PstI fragments hybridize with the probe in the wild type *S. typhimurium* P9003, being 1.77kb, 3.19kb and 6.81kb in size. These bands disappear with increasing length of deletion in the strains selected. No bands are probed in P9074 (*his895*), but the 1.77kb PstI fragment is intact in 9034 (*his801*) (Fig. 3.2). The endpoint for deletion in strain P9034 (*his801*) is therefore to the left of 13.23. This concurs with the endpoint of this deletion as mapped to between 12.38 and 12.48 using the DNA from lambda AD6. Note that the position of the endpoints of these deletion mutations have since been more precisely mapped by H. Brahmbhatt.

3.4. Summary and conclusions.

DNA in the area around the *S. typhimurium* *rfb* region has been cloned, firstly using a lambda 1059 bank of *S. typhimurium* LT2 DNA (Maurer *et. al.*, 1984) and probing with a fragment of cloned *rfb* DNA. A subclone, pPR300, of this *rfb* DNA was found not to probe chromosomal DNA from P9034, a strain which has a deletion extending past the abequose biosynthesis region. The initial lambda isolate did probe P9034 on a colony blot, but a subclone of the DNA distal to the reference EcoRI site of the lambda subclone did not. The initial lambda clone had 0.1 kb more *rfb* DNA to the right than did pPR300, thus it seems likely that the endpoint of the deletion in P9034 (*his801*) lies within that 0.1kb fragment, that is, the endpoint

of the deletion lies between the XbaI site at 12.38 and 12.48, the end of cloned DNA in lambda AD6.

A cosmid bank constructed using chromosomal DNA from *S. typhimurium* LT2 strain P9003 and vector pcos2EMBL (Poutska *et al.*, 1984), was probed with a KpnI fragment from pPR300. Nine different isolates were analysed to give an initial EcoRI restriction map spanning 40kb of DNA in the *S. typhimurium rfb* region. Using pPR340 and pPR341, derivatives of these cosmids, as probes the deletion endpoint in P9034 (*his801*) was mapped to the left of 13.23, and the deletion endpoint of P9074 (*his895*) was mapped to the left of the HindIII site at position 15.28. This suggests that abequose encoding genes in this region lie to the left of position 15.28.

The assumption made in defining an abequose biosynthesis region is that all these genes are grouped in one cluster with no interruptions by genes involved in rhamnose biosynthesis or any other gene. This appears to be the case (Nikaido *et al.*, 1967; Levinthal *et al.*, 1969) at the level of the genetic analysis that has so far been done. Since that time the involvement of two extra proteins in abequose biosynthesis has been shown in *Yersinia pseudotuberculosis* (Matsuhashi and Strominger, 1967) and in *S. typhimurium* (Pape and Strominger, 1969). Two of the three proteins, E_1, E_2, and E_3 that are now known to catalyze the last three steps in CDP abequose biosynthesis have not been mapped in relation to the deletions extending into *rfb* from *his*. The E_1 protein, the product of the redefined *rfbH* gene maps in the abequose region between the *rfb* deletion endpoints for *his660* and *his515* (Yuasa *et al.*, 1969), as expected and all three enzymes may map here but there is no evidence for that. With that limitation we

can say that genes encoding abequose biosynthesis lie within the region 4.36 to 15.28.

CHAPTER FOUR

LOCALISING DNA RESPONSIBLE FOR SPECIFICITY IN THE BIOSYNTHESIS OF ABEQUOSE

4.1. Introduction.

Salmonellae of groups A, B and D produce O-antigens of very similar structure; they differ in the sugar that protrudes from the O-antigen backbone. This branch sugar is a dideoxy sugar; in group A strains it is paratose, in group B strains (e.g. *S. typhimurium*) it is abequose and in group D it is tyvelose. The polymorphism found in groups A, B and D of the salmonellae is also found in the different types of *Y. pseudotuberculosis*; abequose, tyvelose and paratose are made in *Y. pseudotuberculosis* as well as the additional 3,6-dideoxyhexose, ascarylose, each characteristic of a specific serotype. It seems that the biosynthetic pathway is the same in the biosynthesis of all of these 3,6-dideoxy hexoses and only at the very last step(s) is the specificity defined. The last reaction from CDP-4-keto-3,6-dideoxy-D-glucose to a CDP-3,6-dideoxyhexose is catalysed by the E_2 protein, and this protein, therefore, determines specificity. In an *in vitro* assay system, supplying the appropriate intermediate and cofactors, purified E_2 from *Y. pseudotuberculosis* type II makes CDP-abequose, that from type III yields CDP-paratose, and that from type V leads to the formation of CDP-ascarylose (Matsuhashi and Strominger, 1967). A further NAD dependent epimerization step is required to convert paratose to tyvelose, in type IV *Y. pseudotuberculosis* and the salmonellae of group D (Matsuhashi, 1966).

The *rfb* region has been localized on the *S. typhimurium* chromosomal map by genetic studies on hybrid strains derived by inter-crossing a variety of Salmonella species with others differing in their O-antigen (Kochibe, 1970, Kishi and Iseki, 1973a, 1973b, Mäkelä, 1965, 1966, Nikaido *et al.*, 1966, Jousimies and Mäkelä, 1974). These workers had to rely on conjugation systems that transfer large numbers of genes in any one cross, and thus the effect of the transfer of any individual gene could not be determined. We were in a position to look at this at a much finer level using cloned DNA from *S. typhimurium*. We hoped that *S. typhimurium rfb* DNA transformed into strains of groups A and D would express, and thereby substitute the mannose residue with abequose in addition to the 3,6-dideoxy hexose normally produced in those strains.

4.2. Expression of 0-4 specific LPS in *Salmonellae* of groups A and D.

Preliminary sequence data showed a SphI site at position 4.19, this together with the location of the SphI site at 14.85 seemed the most promising restriction fragment to clone in an attempt to get expression of the entire abequose pathway. The SphI site at 14.85 is 0.4kb to the left of the outer limit of the deletion endpoint in P9074 (*his* 695) and therefore may still lie within the right boundary of DNA encoding abequose biosynthesis but the SphI site at 4.19 is to the left of the endpoint of the deletion in strain P9033 (*his* 660) and therefore is well beyond the left boundary of DNA encoding abequose biosynthesis. The 10.66kb SphI fragment was cloned into pUC18 (Yanisch-Perron *et al.*, 1985) from pPR405. Two constructs were isolated with the *lac* promoter reading in opposite directions into the

cloned DNA; with pPR587 being read in the direction that the data of Levinthal *et al.*, (1969) suggest transcription takes place and pPR586 being read opposite to that.

These two plasmids as well as pPR301 were transformed into representatives of group D (*S. dublin* M6, *S. dublin* M22, *S. typhi* Ty21a), and a representative of group A (*S. paratyphi* M8). Transformants were purified and single colonies were tested for their ability to agglutinate anti-04 antisera. All transformants agglutinated anti-04 antisera, indicating that we could use this complementation test at the level of cloned DNA (Fig. 4.1). Hybrid strains generated from the group D strains could also agglutinate anti-09 antisera, and hybrid strains generated from *S. paratyphi* could also agglutinate anti-02 antisera, indicating that the two different O-units are expressed in the one strain. The membrane protein profile on SDS-polyacrylamide gels of hybrid strains were identical to the profile of the host strain, confirming that these strains were not *S. typhimurium* based, but based on the strain into which these plasmids were transformed. Several new proteins were produced in all the hybrid strains and some of these have been analyzed in more detail in *E. coli* K12 strains as described in chapter 6.

4.3. Localizing DNA that specifies the 0-4 epitope.

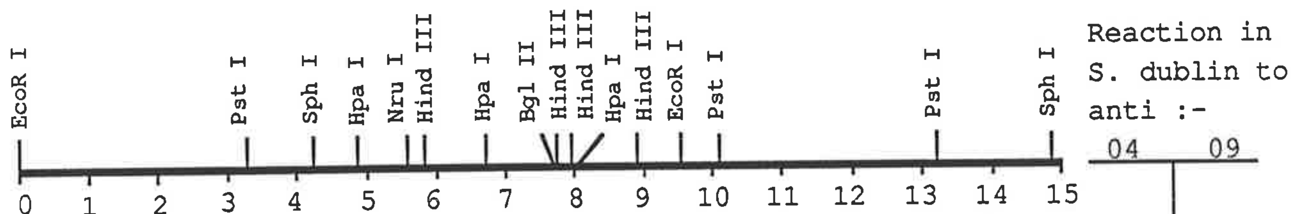
These results indicate that the gene(s) responsible for this complementation lie(s) within the region between map positions 4.36 and 9.57, being the lowest possible map position for the deletion endpoint in *rfbM* and the highest map position in pPR301. Fragments from within these two points were purified from agarose gels and subcloned into the SmaI site of pUC18 (Yanisch-Perron 1985)

FIGURE 4.1

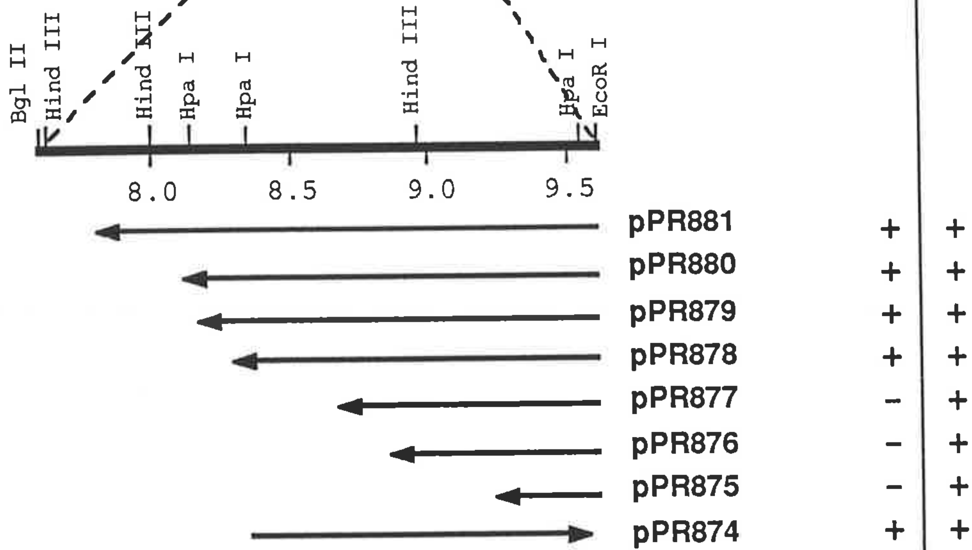
CLONES THAT EXPRESS THE ANTIGENIC FACTOR 04 IN *Salmonella dublin* M6.

A represents subclones used in the localization of *S. typhimurium* DNA required for expression of antigen factor 04 in *S. dublin* M6. Arrowheads on lines representing these clones indicate the direction of transcription from the external *lac* promoter. Arrowheads on both ends of these lines indicate that the insert is cloned in two orientations in the vector: names of the plasmids that carry these inserts are indicated to one side of the lines, the first name, if there are two, indicates the clone that has the external promoter reading into the insert material from right to left. **B** shows the clones derived from pPR303 using Bal31 deletion mutagenesis. The bottom clone in this section represents a cloned HpaI fragment that most closely defines the minimal amount of DNA needed to express 04 specificity.

A



B



(Fig. 4.1). The HpaI piece, bounded by map positions 4.87 and 6.78, was subcloned in both orientations and the two orientations were designated pPR590 and pPR591 with pPR590 being oriented such that the *lac* promoter would read from 6.78 to 4.87. The HpaI fragment bounded by map positions 6.78 and 8.11, was subcloned in both orientations resultant plasmids were designated, pPR593 and pPR594, with pPR593 being in the same orientation as pPR590. The fragment bounded by NruI (5.50) and BglII (7.57) was cut from pPR586 end filled, and then purified from a LGT agarose gel and subcloned into the SmaI site of pUC18, isolates with the insert in both orientation were purified to give pPR589 and pPR873 with pPR589 being in the same orientation as pPR590 and pPR593. A clone containing insert DNA bounded by the NruI site at 5.50 and the EcoRI site at 9.57 was derived by isolating the fragment from an agarose gel after cutting it from pPR586, endfilling it and cloning into the SmaI sites of pUC18 in the same orientation as pPR590; this was dubbed pPR592.

These newly derived clones, and pPR303 were transformed into *S. dublin* M6. This strain was used for screening large numbers of clones, because it transformed more easily than did any of the other Salmonellae used for the complementation test. This strain may have one of its restriction endonucleases inactive or may have give a higher proportion of competent cells (Bullas *et al.*, 1980). These hybrid strains were checked for expression by the slide agglutination test described above and only pPR303 expressed antigen 04 in *S. dublin* M6.

A family of Bal31 deletion mutants were made in pPR303 using PstI

linkers to join the deletion ends. A selection of these clones were transformed into *S. dublin* M6 and checked for their ability to produce LPS reactive to 04 antisera (Fig. 4.1). pPR878 which had an *rfb* DNA insert of 1.40kb (from 8.17 to 9.57) was able to produce anti-04 reactive LPS in *S. dublin* M6 whereas pPR877 which had an *rfb* fragment of 1.13 (from 8.44 to 9.57), was unable to produce anti-04 reactive LPS.

The DNA necessary for this complementation was further refined by subcloning the HpaI fragment bounded by 8.31 and 9.54. This was subcloned from purified DNA derived from pPR303 into the SmaI site of pUC18. Only clones where this DNA was aligned in an orientation opposite to that expected to give expression from the *lac* promoter were obtained. Several attempts to subclone this in the other orientation were unsuccessful, a result that will be discused in a later part of this thesis. When this plasmid, pPR874 was transformed into *S. dublin* M6, the resultant strain produced LPS that was reactive with anti 04 antisera, as well as LPS that was reactive to anti-09 antisera. Thus the DNA necessary for complementation has been localized to a 1.23kb DNA fragment.

Whole cell lysates (Hitchcock and Brown, 1983) were made of selected hybrid strains electrophoresed through an 11-20% SDS Polyacrylamide gel, transferred to nitrocelluose paper and probed with anti-04 antisera, in a western blot system. Plasmids used in this experiment were pPR874, pPR303, and the PstI-EcoRI fragment of pPR303 cloned into pUC18 and pUC19, to give plasmids pPR882 and pPR883 respectively. The last two plasmids were constructed in an attempt to demonstrate expression in plasmids with an external

promoter in both directions. All strains with *rfb* DNA showed the typical ladder of LPS gels seen in silver staining (Goldmann and Leive 1980, Jann *et al.*, 1975; Palva and Mäkelä, 1980) (Fig. 4.2). Clearly the 04 antigen is polymerized and expressed on the surface of hybrid strains.

4.4. Summary and Conclusions.

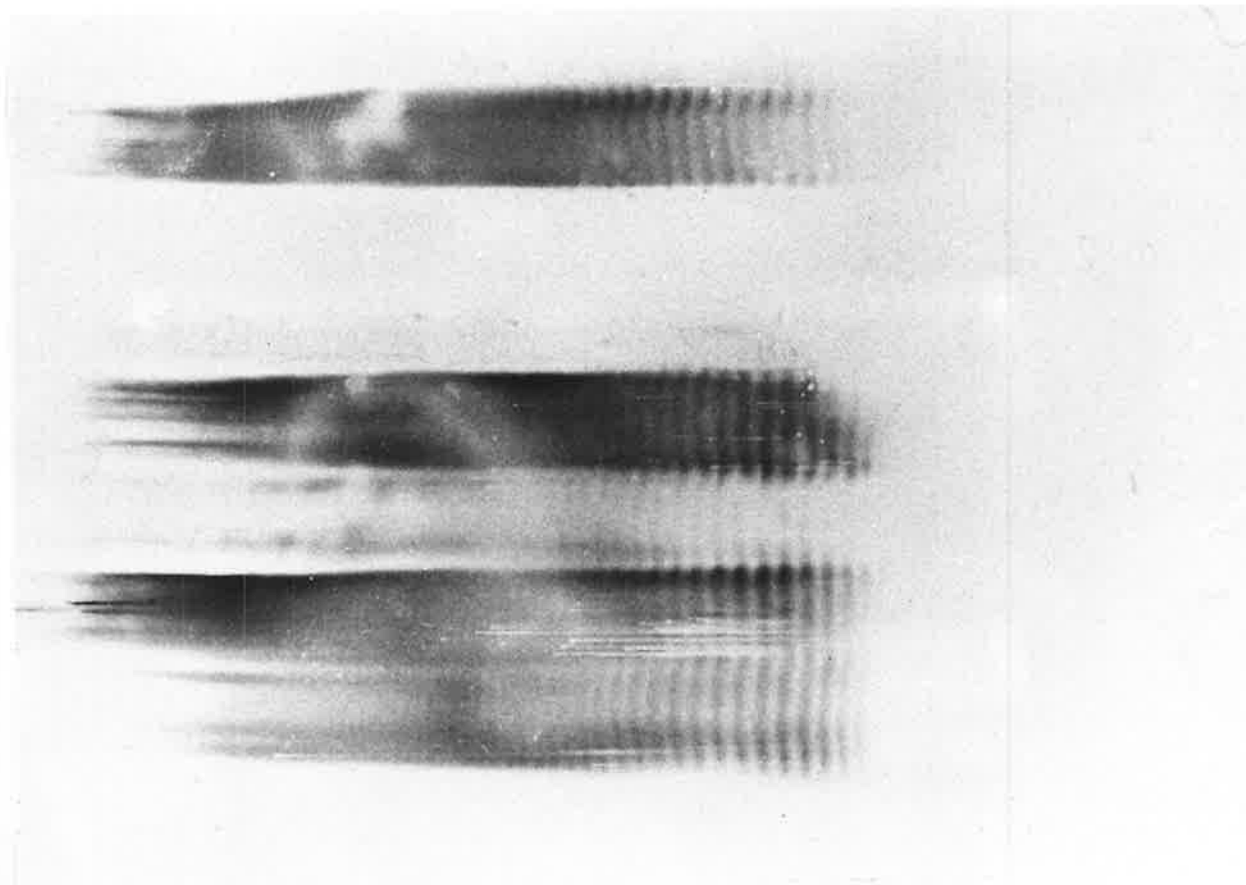
A 1.23kb *S. typhimurium* DNA fragment encodes enough information to make a *S. dublin* strain produce abequose which is then incorporated into the O-antigenic repeat unit and polymerized. We have presumably localized the gene that encodes the E_2 enzyme of *S. typhimurium* (Pape and Strominger, 1969); given the small size of the fragment it is unlikely to contain more than one gene and the transfer and polymerization steps are presumably carried out by *S. dublin* genes which do not discriminate between abequose and tyvelose.

Expression of LPS that reacts with anti 04 antisera in *S. dublin* M6 carrying pPR882 and pPR883 is *prima facie* evidence that a promoter exists between the beginning of the gene encoding the E_2 enzyme and EcoRI at 9.57 as the inserted DNA is in both orientations. Expression in pPR874 narrows the location of this promoter down further still; the promoter is localized to between 9.54 and the beginning of the gene for E_2.

FIGURE 4.2

POLYMERIZATION OF O-UNITS WITH ANTIGEN FACTOR O4 IN *Salmonella dublin* M6.

This western blot is from whole cell lysates of *S. dublin* M6 strains carrying plasmids that have some *rfb* DNA from *S. typhimurium*. Samples of these cell lysates were run on an 11-20% SDS-Polyacrylamide gradient gel and transferred to nitrocelluose paper. This was probed with anti-04 antisera as the primary antisera, and developed with a commercially available goat anti-rabbit horse radish peroxidase conjugate. Samples loaded from left to right were derivatives of *S. dublin* M6 with:- pPR883, pPR882, pPR874, pPR303 and pUC18; the next track was *S. dublin* M6 with no extra plasmids and in the track on the right side was *S. typhimurium* LT2 strain P9003.


P9003
M6
pUC18
pPR303
pPR874
pPR882
pPR883
DIRECTION OF ELECTROPHORESIS

CHAPTER FIVE

SEQUENCE AND ANALYSIS OF PART OF THE REGION INVOLVED IN ABEQUOSE BIOSYNTHESIS

5.1. Introduction

As shown in chapter four, the gene(s) responsible for the enzyme conferring specificity on the 3,6, dideoxyhexose is(are) encoded in a region of DNA that is at most 1.23 kilobases, bounded by HpaI sites at positions 8.31 and 9.54. The importance of the DNA in conferring this specificity as well as the potential for future comparative work with similar gene(s) from *Salmonellae* of groups D or A makes the DNA sequence of this region of interest.

N. Verma has, using clones obtained from myself and H. Brahmbhatt of LT2 DNA, identified the *rfb* regions of *S. typhi* Ty21a and *S. paratyphi* A (Verma *et al.*, 1988). He has shown that there is a stretch of non-homology in the DNA cloned from *S. typhi* Ty21A and *S. paratyphi* A, when compared with DNA from *S. typhimurium* LT2. The former two do not have DNA homologous to a region that lies between 5.8 and 8.17 (*S. typhimurium* restriction map units) based on heteroduplex studies, and only partial homology in the region from 8.17 to 9.57, as this region shows little conservation of restriction endonuclease sites, but still shows homology by heteroduplex. This data confirms the likely location of the genes encoding abequose biosynthesis given that it is a *S. typhimurium* specific function. Sequencing the non-homologous region may also give clues to the

nature of rearrangements that have taken place.

5.2. Sequencing strategy

We chose to sequence to the right from the PstI site at 3.27, because that was a convenient restriction site that extends beyond the end point of the chromosomal deletion ending in *rfbM* (*his660* in P9033) (Brahmbhatt *et al.*, 1986). Genes potentially involved in abequose biosynthesis may be encoded by DNA that lies anywhere to the right of the deletion endpoint in *rfbM* (*his660* in P9033). Thus by sequencing from the PstI site, any genes involved in abequose biosynthesis in this region would be sequenced.

We used the dideoxy sequencing method (Sanger *et al.*, 1977, 1980) to determine the DNA sequence of this region. The PstI (3.27) to EcoRI (9.57) *rfb* fragment was cloned from pPR301 into the PstI and EcoRI sites of the M13 vector mp18 (Norrander *et al.*, 1983) to make plasmid pPR891. A family of deletions was made in pPR891 which extended from the unique EcoRI site using a method that used the random cutting property of DNAseI (Hong, 1983). Selected deletion derivatives were sequenced and the extent of that sequence is shown in figure 5.1.

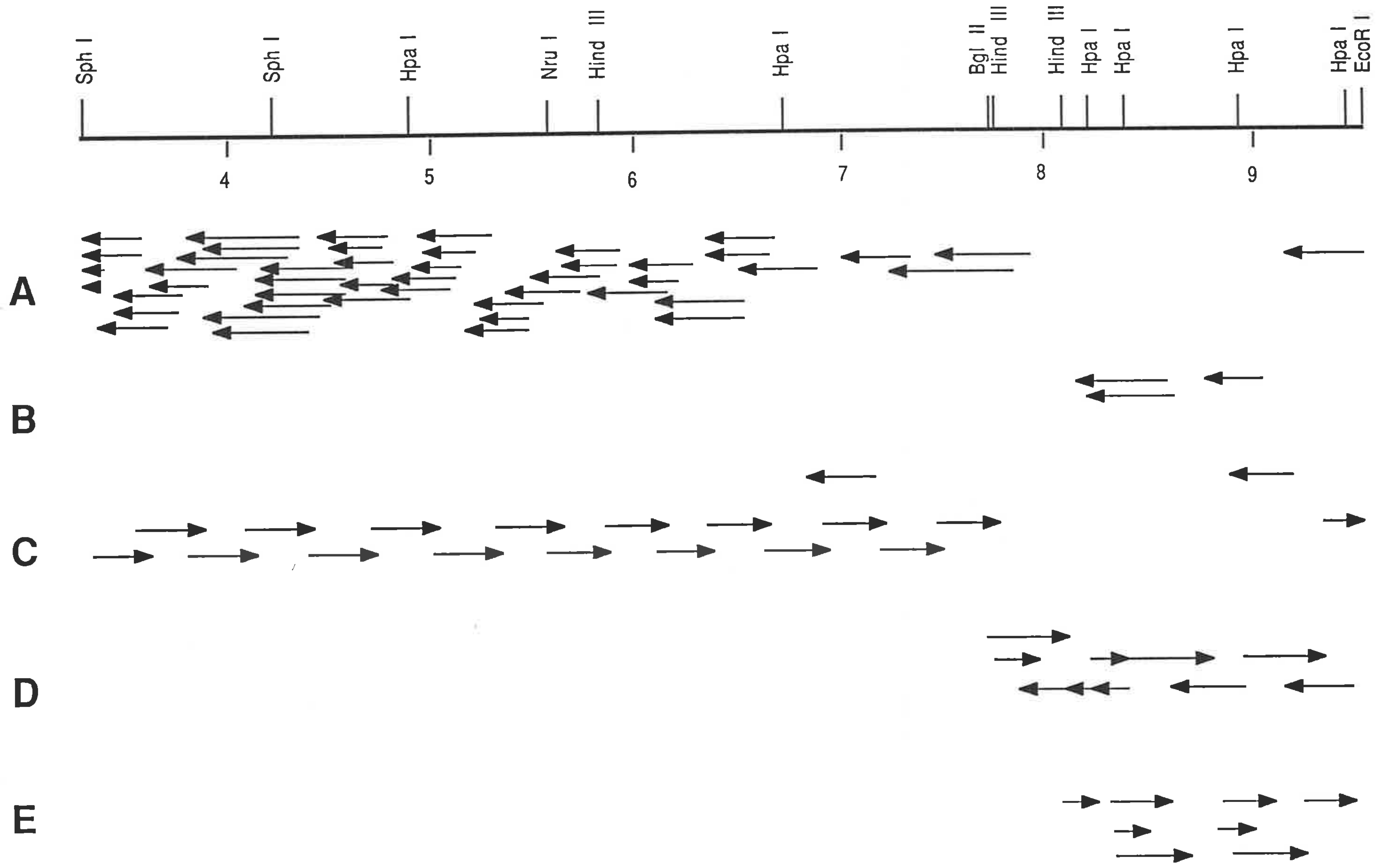
Very few deletions ended to the right of the BglII site (7.57). The BglII (7.57) to EcoRI (9.57) fragment was cloned from pPR303 into the PstI-EcoRI sites of M13mp18 and M13mp19, deletions were again generated according to the method of Hong (1983) and were sequenced (Fig. 5.1). A preparative double digest of pPR303 using HpaI and HindIII was end-filled and fragments were extracted from an acrylamide gel after electrophoresis (Maniatis *et al.*, 1982), these were cloned into the SmaI site of M13mp18, and isolates of clones

FIGURE 5.1

CLONES USED IN SEQUENCING

This figure shows the sequence data obtained from clones derived in one of five ways. An abbreviated restriction map is shown, with the distance from 0 shown in kilo bases. The arrows denote the direction in which the sequence was obtained and the length of the lines indicate the length of the sequence that was read from each M13 clone.

Those lines to the right of **A** represent sequence derived from deletion clones made from pPR891 using the method of Hong (1983) including sequence derived from pPR891. The lines to the right of B represent sequence derived from deletion clones made from an mp18 and an mp19 clone of BglII (7.57) to EcoRI (9.57) *rfb* DNA using the method of Hong (1983). Those lines to the right of **C** represent sequence derived from synthetic oliognucleotides used as primers for the sequencing reaction. Those lines to the right of **D** represents sequence derived from M13 mp18 clones of HpaI/HindIII blunt-ended fragments cloned from pPR303. Those lines to the right of **E** represent sequence derived from M13mp18 clones of the *rfb* region of Bal31 deletion mutants made in pPR303.


Sph I
Sph I
Hpa I
Nru I
Hind III
Hpa I
Bgl II
Hind III
Hind III
Hpa I
Hpa I
Hpa I
Hpa I
EcoR I
4
5
6
7
8
9
A
B
C
D
E

that lay in both orientation were sequenced (Fig 5.1). More M13 clones lying in the region from BglII (7.57) to EcoRI (9.57) were obtained by transferring the *rfb* containing PstI-EcoRI fragments (these restriction enzyme sites defined the extent of *rfb* DNA in these clones) from Bal31 deletion derivatives of pPR303 described in chapter four into the PstI-EcoRI sites of M13mp18. To complete the sequence in both directions a series of oligonucleotides were synthesized, and used as primers, a list of these and their positions in relation to the nucleotide sequence is shown in table 5.1. EcoRI at restriction endonuclease map position 9.57 is taken as sequence position 0, PstI is taken as 6,314.

The DNA sequence of the various clones was handled using the DB system of Staden, (1982, 1986). The sequence data was thoroughly checked and no errors were detected. The sequence obtained is shown in figure 5.2.

5.3.1. Sequence Analysis - Predicting proteins.

Seven open reading frames were found in the sequence, all of them read in the same direction and several of them were butt-jointed. The direction in which these open reading frames are read is that predicted by Levinthal *et al.*, (1969).

The nomenclature that we chose to adopt for the open reading frames is that which Sanger *et al.*, (1982) used for bacteriophage lambda, referring to each open reading frame by the number of amino acids it comprises, in the form ORF#, where the last character represents the number of amino acids predicted for the protein.

TABLE 5.1

SYNTHETIC OLIGONUCLEOTIDES USED IN SEQUENCING

This table shows the synthetic oligonucleotides used as primers in sequencing DNA lying between the PstI site at 3.27 and the EcoRI site at 9.57 map units. The oligonucleotides are listed, and an indication is given as to whether they are the reverse complement of the sequence (rcm) or not. The position on the sequence that the oligonucleotides represent, or would bind to is shown.

OLIGONUCLEOTIDE	ORIENTATION	POSITON
TAT TTG TCA ACT CAC CCA	rcm	198-215
TTG GTA AGC ATT TAC TCG		405-422
ATC AGT GGA CTG CTG CTT		2537-2554
ATA ATC AAG CCA AGC AGC	rcm	2548-2565
GGA AAT TAG TCT TGT TCC	rcm	2797-2815
TGA GAT AGA TTT CCC TAT	rcm	3091-3108
ATT AAA ATT CGT CTG GCC	rcm	3338-3356
AAA ATT CTT AAG CTC GGT	rcm	3606-3620
GGA TAA ATA ATA TTA CGA	rcm	3896-3913
TAA TGG TCT AAG TCG CGA	rcm	4158-4175
TAC ATT ATA AGT TTT AGC	rcm	4425-4442
CGC GTA GGC TTT AGA ATA	rcm	4727-4745
CCG CCA ACC AAT TGC CAA	rcm	4996-5013
TAA CAT CTG CGT CCA GAG	rcm	5254-5271
CTC TTG GAA TCC AGA ACG	rcm	5800-5817
CCA AGC AAA ATA TTG ACA CCT	rcm	2254-2274
GAA CGG CGA TAG GCA GCA AAA	rcm	5501-5521
TTG ATA GTT TAC CTT CAA CGC	rcm	6055-6075

FIGURE 5.2

SEQUENCE OF *rfb* FROM 3.28 TO 9.57

The sequence obtained is presented here, as are the proposed proteins with their predicted amino acids located within the open reading frames shown in figure 5.3. The amino acids are denoted by their international three letter code, stops are indicated by three asterisks. Proposed ribosome binding sites are underlined.

```
EcoRI
GAATTCCTCGAATTACCAGAAGCAACAGAGAAATCAGATCCATCCTGGTTTGGCTTCCCTATCACCCTGAAAGAAACTAGCGGTGTTAAC
        10        20        30        40        50        60        70        80        90

TGTCGAACTGGTGAAATTCCTTGATGAAGCAAAAATCGGTACACGTTTACTGTTTGCTGGAAATCTGATTCGCCAACCGTATTTTGCTAA
       100       110       120       130       140       150       160       170       180

TGTGAAATATCGTGTAGTGGGTGAGTTGACAAATACCGACCGTATAATGAATCAAACGTTCTGGATTGGTATTTATCCAGGCTTGACTAC
       190       200       210       220       230       240       250       260       270

                                                                                     MetThr
AGAGCATTTAGATTATGTAGTTAGCAAGTTTGAAGAGTTCTTTGGTTTGAATTTCTAATTCAATTTATTCTATCTGGTGATTGCGATGAC
       280       290       300       310       320       330       340       350       360

PheLeuLysGluTyrValIleValSerGlyAlaSerGlyPheIleGlyLysHisLeuLeuGluAlaLeuLysLysSerGlyIleSerVal
CTTTTTGAAAGAATATGTAATTGTCAGTGGGGCTTCCGGCTTTATTGGTAAGCATTTACTCGAAGCGCTAAAAAAATCGGGGATTTCAGT
       370       380       390       400       410       420       430       440       450

ValAlaIleThrArgAspValIleLysAsnAsnSerAsnAlaLeuAlaAsnValArgTrpCysSerTrpAspAsnIleGluLeuLeuVal
TGTCGCAATCACTCGAGATGTAATAAAAAATAATAGTAATGCATTAGCTAATGTTAGATGGTGCAGTTGGGATAATATCGAATTATTAGT
       460       470       480       490       500       510       520       530       540

GluGluLeuSerIleAspSerAlaLeuIleGlyIleIleHisLeuAlaThrGluTyrGlyHisLysThrSerSerLeuIleAsnIleGlu
CGAGGAGTTATCAATTGATTCTGCATTAATTGGTATCATTCATTTGGCAACAGAATATGGGCATAAAACATCATCTCTCATAAATATTGA
       550       560       570       580       590       600       610       620       630

AspAlaAsnValIleLysProLeuLysLeuLeuAspLeuAlaIleLysTyrArgAlaAspIlePheLeuAsnThrAspSerPhePheAla
AGATGCAAATGTTATAAAACCATTAAAGCTTCTTGATTTGGCAATAAAATATCGGGCGGATATCTTTTTAAATACAGATAGTTTTTTTGC
       640       650       660       670       680       690       700       710       720

LysLysAspPheAsnTyrGlnHisMetArgProTyrIleIleThrLysArgHisPheAspGluIleGlyHisTyrTyrAlaAsnMetHis
CAAGAAAGATTTTAATTATCAACATATGCGGCCTTATATAATTACTAAAAGACACTTTGATGAAATTGGGCATTATTATGCTAATATGCA
       730       740       750       760       770       780       790       800       810

AspIleSerPheValAsnMetArgLeuGluHisValTyrGlyProGlyAspGlyGluAsnLysPheIleProTyrIleIleAspCysLeu
TGACATTTCATTTGTAAACATGCGATTAGAGCATGTATATGGGCCTGGGGATGGTGAAAATAAATTTATTCCATACATTATCGACTGCTT
       820       830       840       850       860       870       880       890       900

AsnLysLysGlnSerCysValLysCysThrThrGlyGluGlnIleArgAspPheIlePheValAspAspValValAsnAlaTyrLeuThr
AAATAAAAAACAGAGTTGCGTGAAATGTACAACAGGCGAACAGATAAGAGACTTTATTTTTGTAGATGATGTGGTAAATGCTTATTTAAC
       910       920       930       940       950       960       970       980       990

IleLeuGluAsnArgLysGluValProSerTyrThrGluTyrGlnValGlyThrGlyAlaGlyValSerLeuLysAspPheLeuValTyr
TATATTAGAAAATAGAAAAGAAGTACCTTCATATACTGAGTATCAAGTTGGAACTGGTGCTGGGGTAAGTTTGAAAGATTTTCTGGTTTA
      1000      1010      1020      1030      1040      1050      1060      1070      1080

LeuGlnAsnThrMetMetProGlySerSerSerIlePheGluPheGlyAlaIleGluGlnArgAspAsnGluIleMetPheSerValAla
TTTGCAAAATACTATGATGCCAGGTTCATCGAGTATATTTGAATTTGGTGCGATAGAGCAAAGAGATAATGAAATAATGTTCTCTGTAGC
      1090      1100      1110      1120      1130      1140      1150      1160      1170

AsnAsnLysAsnLeuLysAlaMetGlyTrpLysProAsnPheAspTyrLysLysGlyIleGluGluLeuLeuLysArgLeu***
AAATAATAAAAATTTAAAAGCAATGGGCTGGAAACCAAATTTCGATTATAAAAAAGGAATTGAAGAACTACTGAAACGGTTATGAGATTT
      1180      1190      1200      1210      1220      1230      1240      1250      1260

                                                                            ValPheValGlnLeu
TCATGATCTTTTAATAAATAAATCGTTAACAAATTAGTCGCGTTATGTTGTAAAAACTAAGTCGTTTAATTGCATAGTGAAAGTTCAATT
      1270      1280      1290      1300      1310      1320      1330      1340      1350

LeuLysIleProSerHisLeuIleValAlaGlySerSerTrpLeuSerLysIleIleIleAlaGlyValGlnLeuAlaSerIleSerTyr
GTTAAAAATTCCGAGTCATTTAATTGTTGCAGGTTCATCATGGTTATCCAAAATAATAATTGCCGGGGTGCAGTTAGCAAGTATTTCATA
      1360      1370      1380      1390      1400      1410      1420      1430      1440

LeuIleSerMetLeuGlyGluGluLysTyrAlaIlePheSerLeuLeuThrGlyLeuLeuValTrpCysSerAlaValAspPheGlyIle
TCTTATTTCTATGCTAGGTGAAGAGAAATATGCAATCTTTAGTTTGTTAACTGGTTTATTAGTATGGTGTAGCGCTGTTGATTTTGGCAT
      1450      1460      1470      1480      1490      1500      1510      1520      1530

GlyThrGlyLeuGlnAsnTyrIleSerGluCysArgAlaLysAsnLysSerTyrAspAlaTyrIleLysSerAlaLeuHisLeuSerPhe
AGGTACAGGACTGCAAAATTATATATCAGAATGCAGAGCCAAAAACAAAAGTTATGATGCATATATTAAATCAGCATTACATCTAAGCTT
      1540      1550      1560      1570      1580      1590      1600      1610      1620
```

```
IleAlaIleIlePhePheIleAlaLeuPheTyrIlePheSerGlyValIleSerAlaLysTyrLeuSerSerPheHisGluValLeuGln
TATAGCTATTATTTTTTTTATTGCTTTATTTATATTTTTCTGGGGTAATTTCCGCTAAATATCTTTCTTCTTTTCATGAGGTATTACA
  1630      1640      1650      1660      1670      1680      1690      1700      1710

AspLysThrArgMetLeuPhePheThrSerCysLeuValPheSerSerIleGlyIleGlyAlaIleAlaTyrLysIleLeuPheAlaGlu
GGACAAAACCAGAATGCTCTTTTTTACCTCATGTCTGGTTTTCAGTTCTATTGGAATCGGAGCTATTGCTTATAAAATACTTTTTGCCGA
  1720      1730      1740      1750      1760      1770      1780      1790      1800

LeuValGlyTrpLysAlaAsnLeuLeuAsnAlaLeuSerTyrMetIleGlyMetLeuGlyLeuLeuTyrIleTyrTyrArgGlyIleSer
ATTGGTCGGGTGGAAAGCTAATCTATTAAACGCATTATCTTATATGATAGGTATGCTCGGCTTGCTATATATATACTATAGGGGGATCTC
  1810      1820      1830      1840      1850      1860      1870      1880      1890

ValAspIleLysLeuSerLeuIleValLeuTyrLeuProValGlyMetIleSerLeuCysTyrIleValTyrArgTyrIleLysLeuTyr
AGTTGACATAAAATTATCACTAATAGTCCTGTATCTTCCAGTGGGTATGATTTCATTGTGCTATATTGTATATAGATACATAAAGCTTTA
  1900      1910      1920      1930      1940      1950      1960      1970      1980

HisValLysThrThrLysSerHisIle***
TCATGTTAAAACAACAAAATCTCATATATAGCAATTTTACGTAGATCTTCAGGGTTTTTTCTTTTTACTTTATTATCGATAGTGGTGCTT
  1990      2000      2010      2020      2030      2040      2050      2060      2070

CAAAGATTATATTGTTGTCATTTCTCTAAAGACTAACTCCTGCTGATATTGTTCAATATACAGTAACGATGAAAATTTTTGGTTTAGTCT
  2080      2090      2100      2110      2120      2130      2140      2150      2160
                                                                                        Met
TTTTATTTATACTGCTATTTTGCAAGGCATTATGGCCTATATGTGCTGAATTGAGAGTCAAACAGCAAATGGAAAAAACTTAACAAAAT
  2170      2180      2190      2200      2210      2220      2230      2240      2250

IleGlyValAsnIleLeuLeuGlySerLeuTyrValValGlyCysThrIlePheIleTyrLeuPheLysGluGlnIlePheSerValIle
GATAGGTGTCAATATTTTGCTTGGCTCACTATATGTTGTTGGATGTACAATATTTATTTATTTATTTAAAGAACAGATATTTTCAGTAAT
  2260      2270      2280      2290      2300      2310      2320      2330      2340

AlaLysAspIleAsnTyrGlnValSerIleLeuSerPheMetLeuIleGlyIleTyrPheCysIleArgValTrpCysAspThrTyrAla
AGCCAAAGATATTAATTATCAAGTTTCTATTTTATCTTTTATGTTAATTGGCATATATTTCTGTATTCGCGTTTGGTGTGACACTTATGC
  2350      2360      2370      2380      2390      2400      2410      2420      2430

MetLeuLeuGlnSerMetAsnTyrLeuLysIleLeuTrpIleLeuValProLeuGlnAlaIleIleGlyGlyIleAlaGlnTrpTyrPhe
AATGTTATTGCAAAGTATGAATTATTTAAAAATACTTTGGATATTAGTACCACTACAAGCAATAATTGGTGGAATAGCACAATGGTATTT
  2440      2450      2460      2470      2480      2490      2500      2510      2520

SerSerThrLeuGlyIleSerGlyValLeuLeuGlyLeuIleIleSerPheAlaLeuThrValPheTrpGlyLeuProLeuThrTyrLeu
TTCTAGTACGCTTGGAATCAGTGGAGTGCTGCTTGGCTTGATTATATCTTTTGCTTTAACTGTTTTTTGGGGGCTTCCACTAACTTACTT
  2530      2540      2550      2560      2570      2580      2590      2600      2610

IleLysAlaAsnLysGly***        MetLeuIleSerPheCysIleProSerTyrAsnArgLysGlnTyrLeuGluGluLeuLeuAsnSer
AATTAAGGCAAATAAGGGATAATCATATGCTTATATCATTTTGTATTCCAAGTTATAATAGAAAACAATATCTTGAAGAGTTGTTGAATA
  2620      2630      2640      2650      2660      2670      2680      2690      2700

IleAsnAsnGlnGluLysPheAsnLeuAspIleGluIleCysIleSerAspAsnAlaSerThrAspGlyThrGluGluMetIleAspVal
GTATAAATAATCAGGAAAAATTTAATTTAGATATTGAGATATGTATATCAGATAATGCCTCTACTGATGGTACAGAGGAAATGATTGATG
  2710      2720      2730      2740      2750      2760      2770      2780      2790

TrpArgAsnAsnTyrAsnPheProIleIleTyrArgArgAsnSerValAsnLeuGlyProAspArgAsnPheLeuAlaSerValSerLeu
TTTGGAGGAACAATTATAATTTCCCAATAATATATCGGCGTAATAGCGTTAACCTTGGGCCAGATAGGAATTTTCTTGCTTCAGTATCCC
  2800      2810      2820      2830      2840      2850      2860      2870      2880

AlaAsnGlyAspTyrCysTrpIlePheGlySerAspAspAlaLeuAlaLysAspSerLeuAlaIleLeuGlnThrTyrLeuAspSerGln
TTGCGAATGGGGATTATTGTTGGATATTTGGCAGTGATGATGCTCTTGCGAAAGACTCGTTAGCGATATTACAAACTTATCTCGATTCTC
  2890      2900      2910      2920      2930      2940      2950      2960      2970

AlaAspIleTyrLeuCysAspArgLysGluThrGlyCysAspLeuValGluIleArgAsnProHisArgSerTrpLeuArgThrAspAsp
AAGCAGATATATATTTATGTGACAGAAAAGAGACCGGGTGTGATTTAGTTGAGATTAGAAACCCTCATCGTTCTTGGCTCAGAACAGATG
  2980      2990      3000      3010      3020      3030      3040      3050      3060

GluLeuTyrValPheAsnAsnAsnLeuAspArgGluIleTyrLeuSerArgCysLeuSerIleGlyGlyValPheSerTyrLeuSerSer
ATGAACTTTATGTGTTTAATAATAATTTAGATAGGGAAATCTATCTCAGTAGATGCTTATCTATTGGTGGTGTATTTAGCTATCTAAGTT
  3070      3080      3090      3100      3110      3120      3130      3140      3150

LeuIleValLysLysGluArgTrpAspAlaIleAspPheAspAlaSerTyrIleGlyThrSerTyrProHisValPheIleMetMetThr
CTTTAATAGTAAAAAAAGAACGATGGGATGCCATTGATTTTGATGCGTCCTATATTGGCACTTCCTATCCTCATGTATTTATCATGATGA
  3160      3170      3180      3190      3200      3210      3220      3230      3240
```

```
  TyrLeuIleArgHisGlyCysLeuLeuProLeuTyrIleLysThrThrArgAsnMetProThrGluIleMetIleValSerArgThrLys
CGTATTTAATACGCCATGGGTGCCTTTTGCCATTATATATCAAAACCACTCGTAATATGCCGACGGAGATAATGATAGTTTCGAGAACGA
    3250      3260      3270      3280      3290      3300      3310      3320      3330

  GlyLysAlaArgArgIleLeuIleAspPheIleAlaTyrLeuLysLeuAlaAsnAspPheTyrSerLysAsnIleSerLeuLysArgAla
AAGGAAAGGCCAGACGAATTTTAATTGATTTTATTGCATATTTAAAATTAGCTAATGATTTTTACAGTAAAAATATATCTTTAAAACGAG
    3340      3350      3360      3370      3380      3390      3400      3410      3420

  PheGluAsnValLeuLeuLysGluArgProTrpLeuTyrThrThrLeuAlaMetAlaCysTyrGlyAsnSerAspGluLysArgAspLeu
CATTTGAAAATGTTTTGCTAAAAGAGAGACCATGGTTATATACAACTTTGGCTATGGCATGTTATGGCAATAGTGATGAAAAAAGAGATT
    3430      3440      3450      3460      3470      3480      3490      3500      3510

  SerGluPheTyrAlaAsnValGlyCysAsnLysAsnMetIleAsnThrValLeuArgPheGlyLysLeuAlaTyrAlaValLysAsnIle
TATCTGAATTTTATGCAAACGTAGGTTGTAATAAAAATATGATCAACACTGTACTTCGATTTGGGAAACTAGCATATGCAGTGAAAAATA
    3520      3530      3540      3550      3560      3570      3580      3590      3600

  ThrValLeuLysAsnPheThrLysArgIleIleLys*********
TTACCGTGCTTAAGAATTTTACTAAACGGATAATTAAGTAGTAGTAAGTTATTATATTGAGATTAAATGTAGATTTAACCTTTCTGGATT
    3610      3620      3630      3640      3650      3660      3670      3680      3690

CAGCTAGATTTACGTTACTGACTTTTCTTTTTAATGAAAATCATATTTGATATATATAAATAAATTTGGATAGCTTAACTACTTAGATGT
    3700      3710      3720      3730      3740      3750      3760      3770      3780

TTTTTTCTGGGAATGTTAGTGATAATAATATATTTCTTTATGATTGTTTTTGTAGGTGTTTTACTGCCGGTATTACATTACCTCTATTAA
    3790      3800      3810      3820      3830      3840      3850      3860      3870

TTAAGAATTACACCTAGTGTAAGCTTCGTAATATTATTTATCCTTATGATTATTGCTTTAAAGATGCGTATGGAAAAACGGAGAGCTATT
    3880      3890      3900      3910      3920      3930      3940      3950      3960

  MetIleValAsnLeuSerArgLeuGlyLysSerGlyThrGlyMetTrpGlnTyrSerIleLysPheLeuThrAlaLeuArgGluIleAla
CAATGATCGTAAACCTATCACGTTTAGGTAAAAGTGGTACGGGAATGTGGCAATACTCGATTAAATTTTTAACGGCACTGCGAGAAATAG
    3970      3980      3990      4000      4010      4020      4030      4040      4050

  AspValAspAlaIleIleCysSerLysValHisAlaAspTyrPheGluLysLeuGlyTyrAlaValValThrValProAsnIleValSer
CTGATGTTGACGCAATAATCTGTAGCAAGGTACACGCTGATTATTTTGAAAAGCTCGGTTATGCAGTAGTTACTGTTCCGAATATTGTTA
    4060      4070      4080      4090      4100      4110      4120      4130      4140

  AsnThrSerLysThrSerArgLeuArgProLeuValTrpTyrValTyrSerTyrTrpLeuAlaLeuArgValLeuIleLysPheGlyAsn
GCAACACATCAAAAACATCGCGACTTAGACCATTAGTATGGTATGTATATAGTTACTGGCTTGCGCTGAGGGTTTTAATTAAGTTTGGTA
    4150      4160      4170      4180      4190      4200      4210      4220      4230

  LysLysLeuValCysThrThrHisHisThrIleProLeuLeuArgAsnGlnThrIleThrValHisAspIleArgProPheTyrTyrPro
ATAAAAAAATTGGTGTGTACTACACATCACACTATCCCCTTACTGAGAAACCAAACGATAACCGTACATGATATAAGACCTTTTTATTATC
    4240      4250      4260      4270      4280      4290      4300      4310      4320

  AspSerPheIleGlnLysValTyrPheArgPheLeuLeuLysMetSerValLysArgCysLysHisValLeuThrValSerTyrThrVal
CAGATAGTTTTATTCAGAAAGTGTATTTTCGCTTTTTATTAAAAATGTCCGTTAAGCGATGTAAGCATGTTTTAACGGTATCTTATACCG
    4330      4340      4350      4360      4370      4380      4390      4400      4410

  LysAspSerIleAlaLysThrTyrAsnValAspSerGluLysIleSerValIleTyrAsnSerValAsnLysSerAspPheIleGlnLys
TTAAAGATAGCATTGCTAAAACTTATAATGTAGATAGTGAGAAAATATCAGTAATTTATAATAGTGTTAATAAATCTGATTTTATACAAA
    4420      4430      4440      4450      4460      4470      4480      4490      4500

  LysGluLysGluAsnTyrPheLeuAlaValGlyAlaSerTrpProHisLysAsnIleHisSerPheIleLysAsnLysLysValTrpSer
AAAAAGAAAAAAGAGAATTACTTTTTAGCTGTTGGTGCAAGTTGGCCACATAAAAATATTCATTCATTCATAAAAAATAAAAAAGTTTGGT
    4510      4520      4530      4540      4550      4560      4570      4580      4590

  AspSerTyrAsnLeuIleIleValCysGlyArgThrAspTyrAlaMetSerLeuGlnGlnMetValValAspLeuGluLeuLysAspLys
CTGACTCTTATAATTTAATTATTGTATGTGGTCGTACTGACTATGCAATGTCTCTCCAACAAATGGTCGTTGATCTGGAACTAAAAGATA
    4600      4610      4620      4630      4640      4650      4660      4670      4680

  ValThrPheLeuHisGluValSerPheAsnGluLeuLysIleLeuTyrSerLysAlaTyrAlaLeuValTyrProSerIleAspGluGly
AAGTGACTTTTTTACATGAAGTCTCATTTAATGAATTAAAGATTTTATATTCTAAAGCCTACGCGCTTGTTTATCCATCTATTGATGAGG
    4690      4700      4710      4720      4730      4740      4750      4760      4770

  PheGlyIleProProIleGluAlaMetAlaSerAsnThrProValIleValSerAspIleProValPheHisGluValLeuThrAsnGly
GTTTTGGTATACCTCCTATTGAAGCGATGGCATCAAATACTCCAGTTATAGTGTCCGATATACCAGTATTTCATGAAGTGTTAACCAATG
    4780      4790      4800      4810      4820      4830      4840      4850      4860
```

AlaLeuTyrValAsnProAspAspGluLysSerTrpGlnSerAlaIleLysAsnIleGluGlnLeuProAspAlaIleSerArgPheAsn
GTGCATTATATGTGAATCCGGATGATGAAAAAAGCTGGCAGAGTGCAATTAAAAATATAGAGCAGTTGCCTGATGCAATTTCCCGATTTA
4870 4880 4890 4900 4910 4920 4930 4940 4950

AsnTyrValAlaArgTyrAspPheAspAsnMetLysGlnMetValGlyAsnTrpLeuAlaGluSerLys***MetLysIleThrLeuIle
ACAACTATGTCGCACGGTATGACTTTGATAATATGAAGCAGATGGTTGGCAATTGGTTGGCGGAATCAAAATAAATGAAAATAACATTAA
4960 4970 4980 4990 5000 5010 5020 5030 5040

IleProThrTyrAsnAlaGlySerLeuTrpProAsnValLeuAspAlaIleLysGlnGlnThrIleTyrProAspLysLeuIleValIle
TTATTCCCACATATAATGCAGGGTCGCTTTGGCCTAATGTTCTGGATGCGATTAAGCAGCAAACTATATATCCGGATAAATTGATTGTTA
5050 5060 5070 5080 5090 5100 5110 5120 5130

AspSerGlySerLysAspGluThrValProLeuAlaSerAspLeuLysAsnIleSerIlePheAsnIleAspSerLysAspPheAsnHis
TAGACTCAGGTTCTAAAGATGAAACGGTTCCGTTAGCCTCAGACCTGAAAAATATATCAATATTTAATATTGACTCTAAAGATTTTAATC
5140 5150 5160 5170 5180 5190 5200 5210 5220

GlyGlyThrArgAsnLeuAlaValAlaLysThrLeuAspAlaAspValIleIlePheLeuThrGlnAspAlaIleLeuAlaAspSerAsp
ATGGAGGAACCAGAAATTTAGCAGTTGCAAAAACTCTGGACGCTGATGTTATAATTTTTCTAACGCAAGATGCAATTCTCGCGGATTCGG
5230 5240 5250 5260 5270 5280 5290 5300 5310

AlaIleLysAsnLeuValTyrTyrPheSerAspProLeuIleAlaAlaValCysGlyArgGlnLeuProHisLysAspAlaAsnProLeu
ATGCAATTAAAAATTTGGTTTATTATTTTTCAGATCCATTGATAGCAGCGGTTTGTGGTAGACAACTTCCTCATAAAGATGCTAATCCCT
5320 5330 5340 5350 5360 5370 5380 5390 5400

GlnCysMetProGluIleLeuIleIleValGlnAsnLeuLeuLeuSerLysAlaAspIleGluLysLeuGlyIleLysThrValPheMet
TGCAGTGCATGCCAGAAATTTTAATTATAGTTCAAAATCTATTGTTAAGTAAGGCAGATATAGAAAAATTGGGTATTAAAACTGTATTTA
5410 5420 5430 5440 5450 5460 5470 5480 5490

SerAsnSerPheAlaAlaTyrArgArgSerValPheGluGluLeuSerGlyPheProGluHisThrIleLeuAlaGluAspMetPheMet
TGTCCAATTCTTTTGCTGCCTATCGCCGTTCCGTTTTTGAAGAGTTAAGTGGGTTTCCTGAACATACAATTCTTGCCGAGGATATGTTTA
5500 5510 5520 5530 5540 5550 5560 5570 5580

AlaAlaLysMetIleGlnAlaGlyTyrLysValAlaTyrCysAlaGluAlaValValArgHisSerHisAsnTyrThrProArgGluGlu
TGGCGGCTAAGATGATTCAGGCGGGTTATAAGGTCGCCTACTGCGCTGAAGCGGTGGTAAGACACTCCCATAATTATACCCCGCGAGAAG
5590 5600 5610 5620 5630 5640 5650 5660 5670

PheGlnArgTyrPheAspThrGlyValPheHisAlaCysSerProTrpIleGlnArgAspPheGlyGlyGlyGlyGluGlyPheArgPhe
AGTTTCAACGATATTTTGATACTGGTGTATTTCATGCTTGTTCTCCGTGGATTCAGCGTGACTTTGGCGGCGGTGGTGAGGGTTTCCGCT
5680 5690 5700 5710 5720 5730 5740 5750 5760

ValLysSerGluIleGlnPheLeuLeuLysAsnAlaProPheTrpIleProArgAlaLeuLeuThrThrPheAlaLysPheLeuGlyTyr
TCGTAAAATCAGAGATTCAATTCCTGCTTAAAAATGCACCGTTCTGGATTCCAAGAGCTTTATTAACAACCTTTGCTAAATTCTTGGGTT
5770 5780 5790 5800 5810 5820 5830 5840 5850

LysLeuGlyLysHisTrpGlnSerLeuProLeuSerThrCysArgTyrPheSerMetTyrLysSerTyrTrpAsnAsnIleGlnTyrSer
ACAAATTAGGCAAGCATTGGCAATCTTTACCGTTGTCTACATGTCGCTATTTTAGCATGTACAAGAGTTATTGGAATAATATCCAATATT
5860 5870 5880 5890 5900 5910 5920 5930 5940

SerSerLysGluArg***
CTTCGTCAAAAGAGAGATAAAATAAATGTCTTTTCTTCCCGTAATTATGGCTGGCGGCACAGGTAGCGGCTTTATGCGCGCTTTCACGCG
5950 5960 5970 5980 5990 6000 6010 6020 6030

MetLeuGlnAsnThrIleLysArgLeuAlaPheThrPheTyrArgArg
AATATCATCCGAAGCAGTTTCTAAGCGTTGAAGGTAAACTATCAATGCTGCAAAATACTATAAAGCGATTAGCGTTCACTTTCTACAGAA
6040 6050 6060 6070 6080 6090 6100 6110 6120

ThrArgPheIleCysAsnAspArgHisArgPheLeuValAlaGluGlnLeuArgGluIleAspLysLeuAlaAsnAsnIleIleLeuGlu
GAACCCGGTTTATTTGCAATGACAGACACCGTTTCTTAGTCGCTGAACAACTCCGTGAAATTGACAAGTTAGCAAATAATATTATTCTCG
6130 6140 6150 6160 6170 6180 6190 6200 6210

ProValGlyArgAsnThrAlaProSerAspProLeuAlaAlaPheCysAlaLeuGlnAsnAlaAspAsnGlySerIleLeuPheCysTrp
AACCGGTAGGCCGTAATACTGCACCTAGCGATCCTCTTGCCGCGTTTTGTGCGCTCCAGAATGCTGATAATGGCTCGATCCTCTTTTGTT
6220 6230 6240 6250 6260 6270 6280 6290 6300

PheLeuLeuGln
GGTTCTTGCTGCAG
6310 PstI

The first full open reading frame starts 355 bases from the map position 9.57 but it is possible that the carboxy-terminal portion of a protein is encoded by this DNA, but the open reading frames crossing the EcoRI position are far too short to be identifiable as the carboxy terminus portion of a protein.

The ribosome binding sites thus far reported in the literature have had a median spacing between the two known elements (i.e., the start codon and the Shine-Dalgarno sequence) of 7 nucleotides. It is suggested that the distances between these two elements outside of the range of 5 to 9 decreases the efficiency of ribosome binding, however, functional ribosome binding sites with a distance as low as 4 and as high as 13 have been shown to function (Kozak *et al.*, 1983). Many of the ribosome binding sites that are the best approximations of ribosome binding sites found in our sequence are shown in table 5.2. The suggested ribosome binding sites have quite long spacings between their two elements, *rfbJ* 8 nucleotides, ORF224 has 11 nucleotides, ORF127 has 14 nucleotides, ORF334 has 8 nucleotides, ORF353 has 8 nucleoties, ORF 311 has 11 nucleotides, and ORF80 has 9 nucleotides. The ribosome binding site suggested for ORF127 seems quite unlikely, but the long open reading frame in this region suggests that a protein is likely to exist in this region, and the ribosome binding site suggested is the best fit that could be found in the sequence to correspond to a ribosome binding site for this open reading frame. All open reading frames have a predicted methionine start with the exception of ORF224 which uses the codon GUG, encoding a valine residue.

The ANALYSEQ DNA sequence analysis program was used to implement Shepherd's RNY preference method to predict proteins in

TABLE 5.2

SUGGESTED RIBOSOME BINDING SITES IN THE REGION SEQUENCED.

Shown are the suggested genes together with the sequence between the two elements of their respective ribosome binding sites; the Shine-Dalgarno sequence and the start codon. In a third column is shown the distance between the two elements.

GENE	INTERVENING SEQUENCE	DISTANCE
rfbJ	TGGTGATTGCGATG	8
ORF224	TAATCGTTTAATTGCATAGTG	11
ORF127	AATGGAAAAAACTTAACAAAATG	14
ORF334	TAAGGGATAATCATATG	8
ORF353	GGAGAGCTATTCAATG	9
ORF311	GGCGGAATCAAAATAAATG	11
ORF80	AGGTAAACTATCAATG	9

this region (Shepherd, 1981), in an attempt to see if proteins are predicted by more indirect means. Shepherd has found that many genes have a preference for the use of codons in the form RNY where R= purine Y= pyrimidine, and N= any base. He has attributed this to being due to remnants of a primitive genetic code. Strong potential protein coding regions are predicted, by this method, in all but one of the proteins predicted, ORF127 (Fig. 5.3). Note that as a convention figures are drawn to fit in with the restriction map numbers such that DNA closer to the 0 point of the restriction map is shown on the left hand side of the figure. The poor fit of the predicted ribosome binding site due to the long interval between the two constituent elements, and the indefinite prediction of coding sequence in this region by the method of Shepherd, (1981) cast doubt on the existence of ORF127. It is possibile that the open reading frame could start earlier; there are only very few stops in frame two for the interval between ORF224 and ORF127 (Fig. 5.3) and it is possible that we have a mutation in the sequenced DNA. The analysis also suggests that the third base of the EcoRI site should be used as the first base of a triplet codon, and that the terminal portion of the open reading frame extends to base 77.

5.3.2. Sequence Analysis - Properties of the predicted proteins.

We predict that several proteins are encoded by the DNA in this region. The following describes the physical properties of these proteins as predicted by a number of programs. ALOM predicts transmembrane segments on the basis of differences in the distribution of hydrophobic amino acids (Klein *et al.*, 1985). Molecular weights and a hydrophobicity index was calculated by the program

ANALYSEQ (Staden, 1982). The hydrophobicity and charge distribution was plotted also using ANALYSEQ which bases its calculations on the values of Kyte and Doolittle, (1982). The prediction of the secondary structure of proteins, is usually very unreliable, with only a proportion of structures correcty predicted in new proteins (Kabsch and Sander, 1983; Taylor and Thornton, 1983; Kabsh and Sander, 1984). We used the program PREDICT, which combines eight different methods to produce a joint prediction (Eliopoulos *et al.*, 1982).

The initial segment that may be the carboxy terminal portion of a protein has as yet not been examined because we are uncertain whether this portion does code for part of a protein and which of these short open reading frames extend further, if in fact any of them do.

The protein encoded by the first open reading frame, is very closely bounded by the HpaI sites on either sides that were used to make clone pPR874, which expresses antigen 04 in *S. dublin* M6. Were therefore conclude that the protein produced by this open reading frame is the enzyme E_2. We chose to allocate the name *rfbI* to enzyme E_3 and *rfbJ* to E_2 because these were two names not yet allocated to *rfb* genes and the names *rfbH*, *rfbI* and *rfbJ* would reflect the order in which genes acted in the biosynthesis of CDP-abequose. *rfbJ* has a predicted molecular weight of 34,106 daltons comprising 299 amino acids, Shepherd's RNY prediction shows quite clearly that a protein is produced in this frame (Fig. 5.3). The protein is given a hydrophobicity index of 16.18 by ANALYSEQ, and its hydrophobicity/charge profile (Fig. 5.4) shows that the protein is highly charged over much of the protein. The secondary structure

FIGURE 5.3

CODING REGIONS IN THE DNA SEQUENCE AS PREDICTED BY SHEPHERDS RNY PREFERENCE.

The plot represents Shepherd's RNY preference method to predict proteins. The three windows in section **A** represent plots of the probability of the three possible reading frames of the sequence encoding a protein. The uppermost window is using the first base of the sequence as the first base of the triplet, the second window uses the second base as the first base of the triplet, and the third window uses the third base as the first base of the triplet. Termination codons are displayed on a line in the middle of each window, as short vertical lines. The extent of proteins predicted on the basis of open reading frames, with possible ribosome binding sites and start codons are shown to the right of **C**. The names by which they are referred to are indicated above that, to the right of **B**. The estimated molecular weights are listed to the right of **D**. **E** shows the distance in kilobases from the sequence start as given in Fig 5.2. **F** shows the restriction map and restriction map positions in the area sequenced.

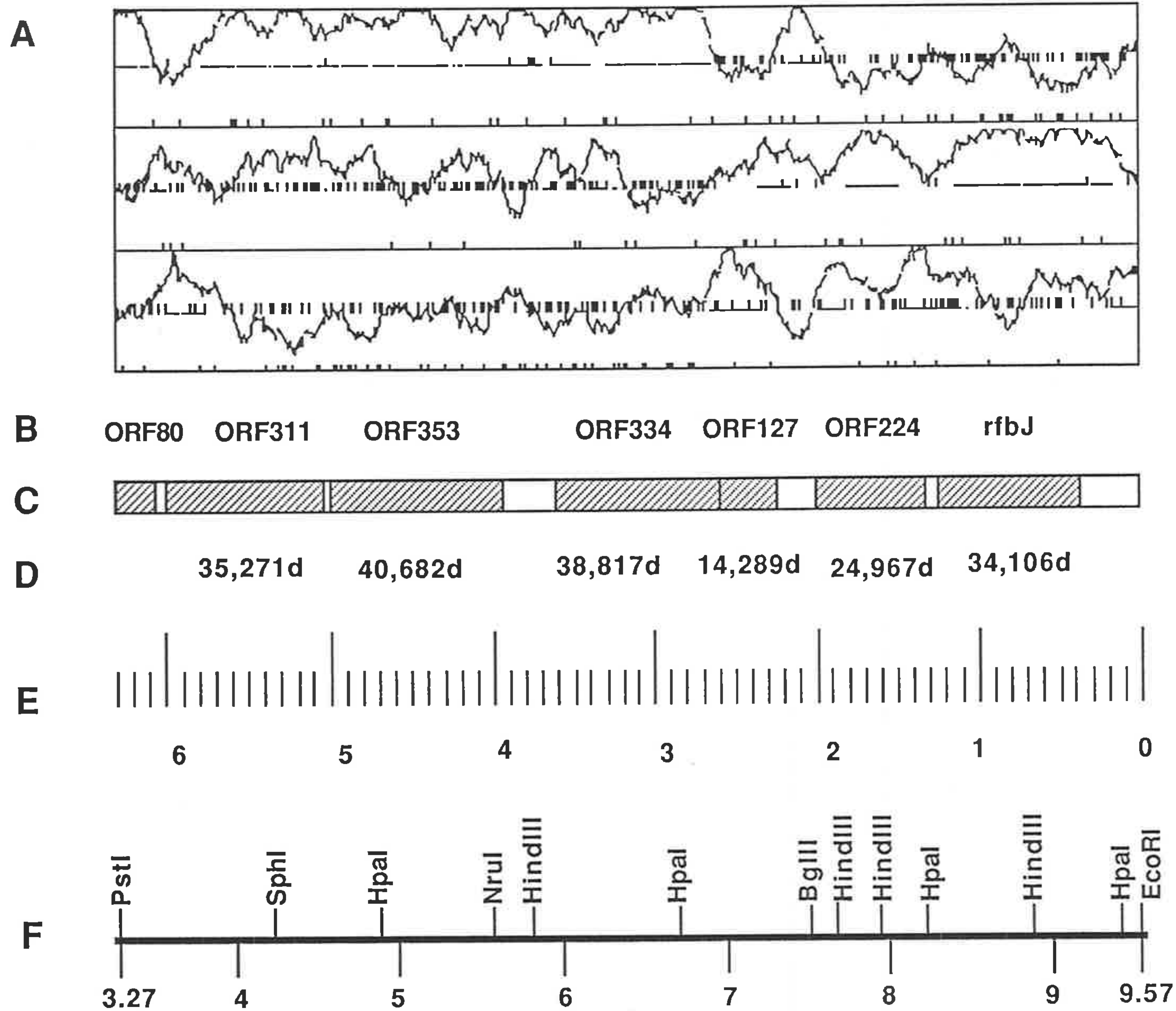
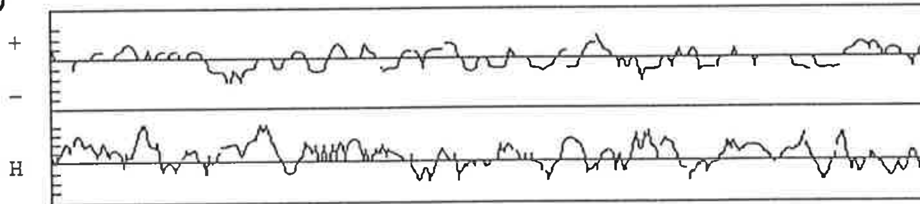
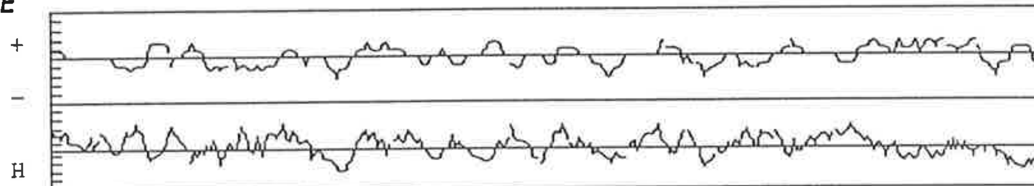

A
B
ORF80
ORF311
ORF353
ORF334
ORF127
ORF224
rfbJ
C
D
35,271d
40,682d
38,817d
14,289d
24,967d
34,106d
E
6
5
4
3
2
1
0
F
PstI
SphI
HpaI
NruI
HindIII
HpaI
BglII
HindIII
HindIII
HpaI
HindIII
HpaI
EcoRI
3.27
4
5
6
7
8
9
9.57

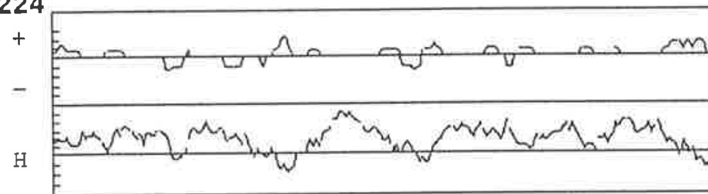
FIGURE 5.4

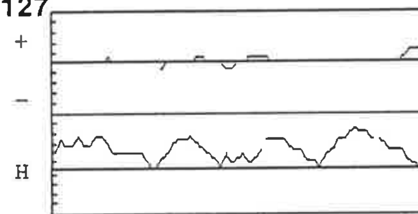
CHARGE AND HYDROPHOBICITY PLOTS OF PREDICTED PROTEINS

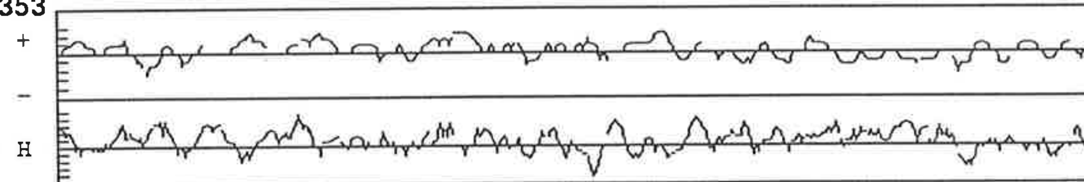
The plots are adaptations of the displays produced by ANALYSEQ which uses the hydophobicity values of Kyte and Doolittle (1982) to calculate the charge and hydrophobicity of a given group of amino acids. The lengths of the predicted proteins are drawn to scale. Each protein sequence is represented in two windows; the top window represents the charge distribution throughout the protein, any value above the dividing line represents a positive charge, any value plotted below the line representing a negative charge, the top of the window represents the maximum charge possible for the span of amino acids screened at any particular point. This top window is divided into 10 equal parts in the scale along the left hand vertical line. The bottom window for each diagram represents the hydrophobicity plot, the line in the middle of this window represents half the hydrophobicity possible for the span of amino acids sampled by the program. The diagram is divided into 10 equal parts along the left hand vertical line. The diagrams are then on the same scale as Fig 6 of the Kyte Doolittle paper and values of + or - 50 could be assigned to the top and bottom of the diagram with corresponding values in between. The amino acid sequences examined are *rfbJ*, *galE*, ORF224, ORF127, ORF353, ORF311 and ORF334. A scale representing the number of amino acids from the beginnings of the predicted proteins is drawn underneath these.

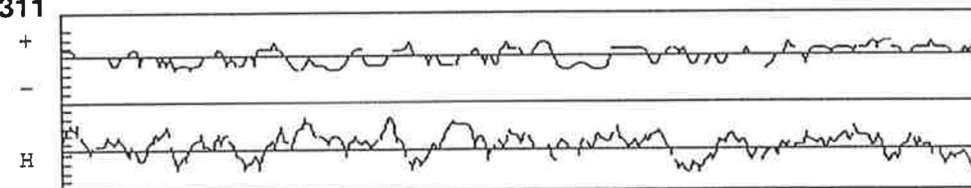

rfbJ
+
−
H

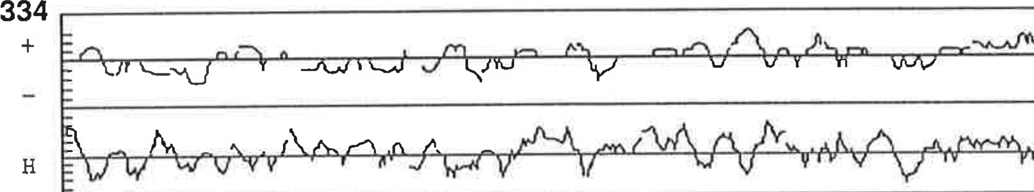

galE
+
−
H


ORF224
+
−
H


ORF127
+
−
H


ORF353
+
−
H


ORF311
+
−
H


ORF334
+
−
H

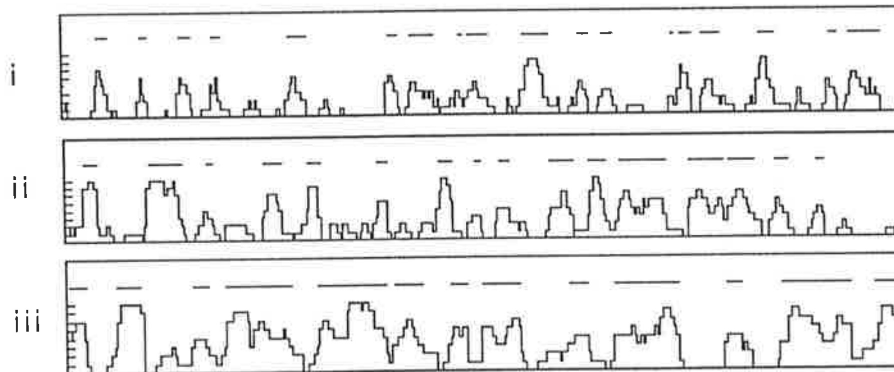
0
100
200
300

FIGURE 5.5

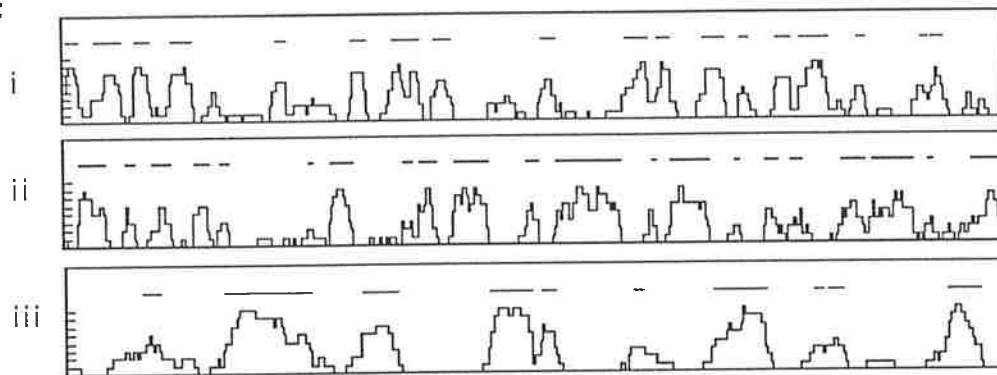
POSSIBLE SECONDARY STRUCTURES IN THE PREDICTED PROTEINS

Shown here are predictions of secondary structures in the proteins encoded by the DNA that was sequenced predicted by the program. PREDICT. This program uses eight different methods to predict protein secondary structure. These methods are based on the observation, of known structures, of the relative frequency with which different types of amino acids are found in different types of secondary structure, in particular helices, extended or beta-structure and turns. Prediction of protein structure by any one of these methods is not all that reliable, a consensus of the eight different methods to give a joint prediction gives a more reliable estimate (Eliopoulos et al., 1982). The plots are of the number of predictions that give a positive result for a given structure at any given amino acid. The joint prediction is in the form of the broken line across the top of each window, the extent of the line is the extent of the predicted structure of that window. Each protein has three plots, labelled **i**, turns or coils, **ii** alpha coils and **iii** beta strands. Each plot has 8 horizontal lines on the left vertical line giving an indication of the number of predictions. At the bottom of each page is a scale of the length of the protein, in amino acid residues. Predictions are made for *rfbJ* , *galE* , ORF224, ORF127, ORF353, ORF311 and ORF334.

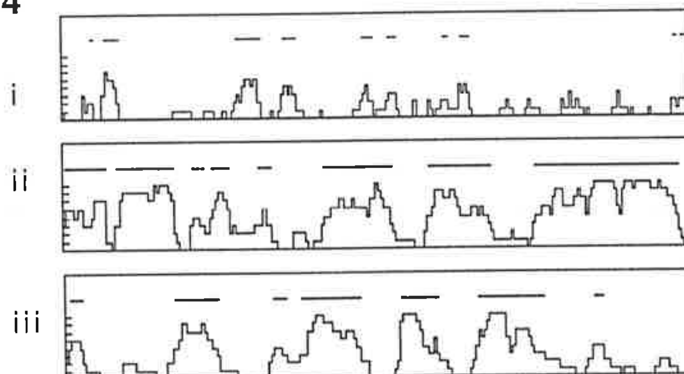
rfbJ



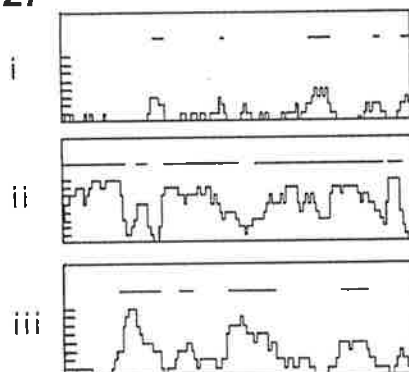
galE



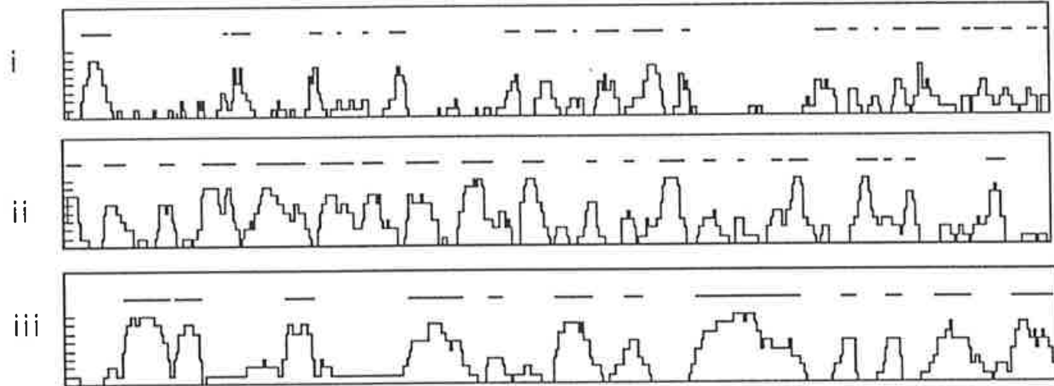
ORF224



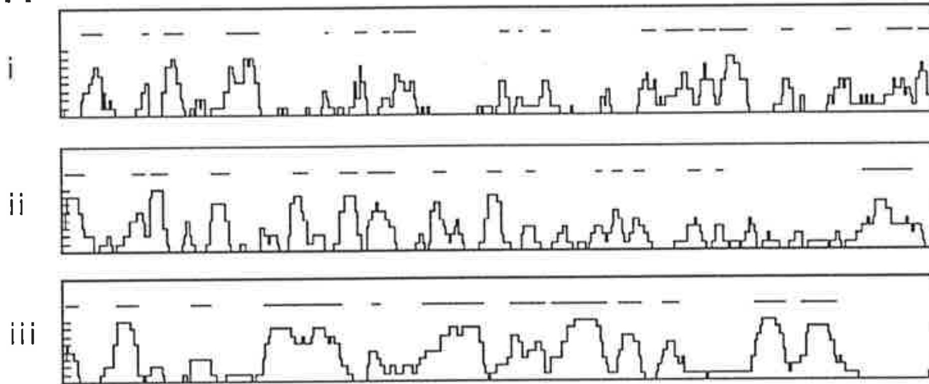
ORF127



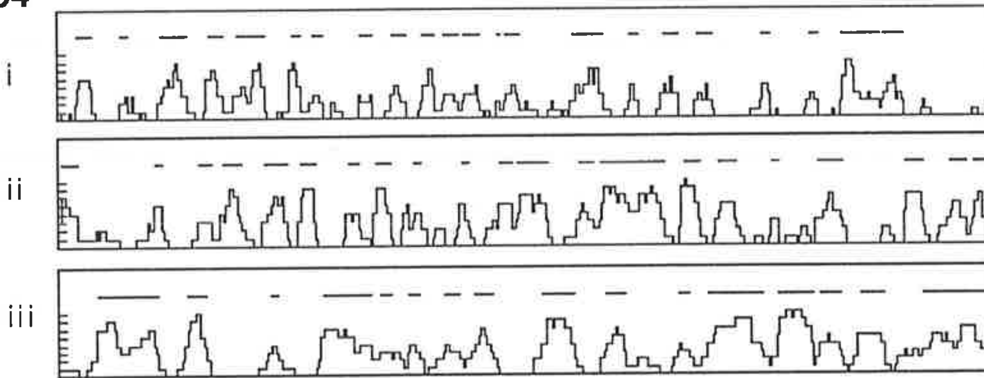
ORF353



ORF311



ORF334



prediction (Fig. 5.5) suggests the presence of both alpha helices and beta pleated sheets.

ORF224, has GUG as a transcription initiation codon, encoding a valine residue, instead of the more common AUG triplet start codon, encoding methionine. The probability of a protein is still quite high according to Shepherd's RNY preference analysis, (Fig. 5.3) although a quite high probability is also seen for the third frame which has several stop codons and therefore could not encode a protein of any significant length. This protein has a predicted molecular weight of 24,967 comprising 224 amino acids. ANALYSEQ gives this protein a hydrophobicity index of 81.74. The hydrophobicity plot (Fig. 5.4) shows large regions of hydrophobicity bracketed by small hydrophilic regions. These hydrophobic regions in the protein correspond to regions of the protein predicted to have beta strands (Fig. 5.5). These discrete regions of hydrophobicity suggest a possible transmembrane locality, and hence a membrane location for this protein. The lack of a signal sequence suggests an inner membrane location (Oliver, 1985). Indeed the program ALOM predicts several transmembrane segments indicative that ORF224 is an integral membrane protein (Table 5.3).

ORF 127 is not predicted in its suggested form by Shepherd's RNY prediction (Fig. 5.3). A much higher probability of being a coding region is given to frame three. However frame three does not have an open reading frame of any significant length. ORF127 has 127 amino acids and a predicted molecular weight of 14,289. It has quite a high hydrophobicity index of 59.8, and hydrophobic profile showing the distribution of charged amino acids along its length (Fig. 5.4). As with ORF224, this protein is predicted to have transmembrane segments, when analysed by ALOM (Table 5.3), but has no signal sequence, again

TABLE 5.3

TRANS-MEMBRANE SEGMENT PREDICTIONS FOR PROTEINS OF THE REGION SEQUENCED

Shown are the transmembrane predictions for the two proteins that are predicted to have transmembrane segments by the program ALOM. Shown for both proteins are the amino acid number range of each of the amino acid strings predicted to have trans-membrane location and the probability calculated by the algorith of Kanehisa, (1982).

```
=============================================================
ORF224
_____________________________________________________________
INTEGRAL    P:I   odds = 2.0E-05   Transmembrane   96 - 112    (85-117)
INTEGRAL    P:I   odds = 8.1E-04   Transmembrane  192 - 208   (188-209)
INTEGRAL    P;I   odds = 2.2E-02   Transmembrane  131 - 147   (130-149)
INTEGRAL    P:I   odds = 6.0E-02   Transmembrane   23 - 39     (23-40)
INTEGRAL    P:I   odds = 6.3E-02   Transmembrane   45 - 61     (45-63)
INTEGRAL    P:I   odds = 2.3E-01   Transmembrane  163 - 179   (163-179)

ORF127
_____________________________________________________________
INTEGRAL    P:I   odds = 4.5E-04   Transmembrane   97 - 113    (90-122)
INTEGRAL    P:I   odds = 3.6E-03   Transmembrane    6 - 22      (1-24)
INTEGRAL    P:I   odds = 7.2E-03   Transmembrane   39 - 55     (37-57)
INTEGRAL    P:I   odds = 3.9E-02   Transmembrane   70 - 86     (69-88)
=============================================================
```

implying an inner membrane location. As with ORF224, we therefore suggest that the protein produced by this open reading frame has an inner membrane location.

The Shine-Dalgarno sequence of ORF334 overlaps with the stop signal of ORF127, it is not uncommon to find a one base-pair overlap between the terminator codon of one cistron and the initiator codon of the next. There are examples in which translation of the downstream member depends on prior translation of the preceding region, but this dependence is not always found with overlapping cistrons and is sometimes found with non-overlapping cistrons. This sort of overlap is taken to mean that these two open reading frames are co-transcribed (Kozak, 1983) and therefore these two proteins are most probably co-transcribed. ORF334 has 334 amino acids with a predicted molecular weight of 38,817 daltons, and has an hydrophobicity index of 26.2. The charge profile shows that the majority of charged amino acids in the first 150 amino acids have a negative charge, and the amino acids in the remainder of the protein are largely positvely charged. This is a distinctly assymetrical distribution. Shepherd's RNY prediction gives a high probability of this region encoding a protein. No transmembrane segments are predicted by ALOM and both beta sheet and alpha helix structures are predicted (Fig. 5.5).

ORF353 comprises 353 amino acids and has a predicted molecular weight of 40,682 daltons, and a hydrophobicity index of 34.7. The charge/hydrophobicity profile shows that the molecule is quite highly charged and that the assymetrical distribution of charged amino acids is even more distinct than in ORF334; the first 220 amino acids are predominantly positively charged, whereas the remainder of the

molecule has predominantly negatively charged residues. Again this protein is predicted to have both alpha helical and beta sheet secondary structures and no transmembrane structures or signal sequence is predicted.

ORF311 has 311 amino acids with a predicted molecular weight of 35,271 dlatons, and a hydrophobicity index of 20.4. Its charge profile can be divided into three sections, the first 130 amino acids are predominantly negatively charged, the section up to about residue 240 has both sorts of charges, and the remainder of the protein has almost exclusively positive charges. The protein is predicted to have both alpha helices and beta strands but no transmembrane segments or signal sequences are predicted.

The final 239 nucleotides of our sequence has the N-terminus of another protein. 80 amino acids are predicted, all quite hydrophilic and for convenience this will be called ORF80 (Fig. 5.3).

5.3.3. Sequence Analysis - Homologies with other proteins.

Three different data bases were searched for homology with our DNA sequence, and three databases were searched for homologies with the amino acid sequences of the individual proteins predicted. Homology in nucleic acid sequence was searched for in the following databases: GENBANK nucleic acids Release 45.0 22 Sept 1986, NBRF (National Biomedical Research Foundation) Nucleic Acids Release 29, 18 Sept 1986, EMBL (European Molecular Biology Laboratory) Nucleic acids database release 9, Sept 1986. Searches for homologies between the amino acid sequences of the predicted proteins and known amino acid sequences of other proteins were done for amino acid sequences in

the NBRF (National Biomedical Research Foundation) Protein database Release 10, 13 Aug 1986, NEWAT Protein database April 1985, and KYOTO UNIVERSITY (PROSEQ) Protein database September 1986.

We made use of programs in the MBIS suite of programs (Bucholtz and Reisner, 1986) to search these databases using the quasi correlation coefficient (Qr) (Lipmann and Pearson, 1985) for high speed searching of the amino-acid residue databases, and a program written by Kanehisa *et al*, (1984) based on the algorith developed by Wilber and Lipman (1983). A third method used for searching the PIR and KYOTO databases used the algorithm and software of Lipman and Pearson, FASTN and FASTP (1985). However we found no homologies that appeared significant.

5.3.4. Sequence Analysis - Homologies between predicted proteins.

We used the program SEQHP (Kanehisa, 1982) to search for homologies between the predicted open reading frames. SEQHP searches locally homologous regions of two protein sequences according to the algorithm described by Goad and Kanehisa, (1982) based on the data of Dayhoff *et al.*, (1978). Homologies between the proteins may give an indication of the overlap in function of a portion of the predicted proteins, or perhaps, an idea of which proteins belong to the same biosynthetic pathway.

Table 5.4 sets out the closest homology scores of the predicted proteins compared with each other; there were other regions of homology between some of the other proteins which are not displayed. In an attempt to see if these homologies could tell us

TABLE 5.5

AMINO ACID HOMOLOGIES BETWEEN THE PREDICTED PROTEINS

The program SEQHP was used to search for homologies between the proteins predicted in the sequence. This program searches locally homologous region of two protein sequences according to the algorith described by Goad and Kanehisa, (1982). The measurement of similarity is taken from the amino acid mutation data among related proteins compiled by Dayhoff and her colleagues (Dayhoff *et al.*, 1978). Those sequences with the lowest distance number are the most closely related. This table only shows the lowest distance score between pairs of proteins to give an indication of their evolutionary relatedness, if no relationship was found that is indicated by the word "None".

	rfbJ	ORF224	ORF127	ORF334	ORF353	ORF311	ORF80
rfbJ		-78	-60	-59	-72	-74	None
ORF224			-89	-89	-73	-64	None
ORF127				None	-61	-53	None
ORF334					-68	-102	None
ORF353						-56	None
ORF311							None

FIGURE 5.6

ALIGNMENT OF AMINO ACIDS IN A REGION THAT SHOWS HOMOLOGIES IN ALL PROTEINS

The alignment calculated by SEQHP (Kanehisa 1982) for amino acids that lie within a region that shows homologies in all proteins that are fully sequenced, are shown. Amino acid residues that are highly concerved are boxed in by dotted lines.

```
                          10                        20
ORF224  V K V Q L L K I P S H L I V A G S S W L S K I I I
ORF127  L K I L W I L V P L Q A I I G G I A Q W Y F S S T L G I
ORF311  M P I I L I I V Q N L L L S K A D I
ORF334  L S R C L S I G G V F S Y L S S L I V
rfbJ    N I E L L V E E L S I D S A L I G I I H
ORF353  L R N Q T I

           30                 40                    50
ORF224  G V Q L A S I S Y L I S M L G E E K Y A I F S L L T
ORF127  G V L L G L I I S F A L T V F A L T V F W G L P L T
ORF311  K L G I K T V F M S N S F A A Y R R S V F E E L S
ORF334  K E R W D A I D F D A S Y I G T S Y P H V F I M M T
rfbJ    A T E Y G H K T S S L I N I E D A N V I K P L K L L
ORF353  V H D I R P F Y Y P D S F I Q K V Y F R F L L K

                       60                 70
ORF224  L L V W C S A V D F G I G T G L Q N Y I S E C
ORF127  I I
ORF311  F P E H T I L A E D M F M A A K M I Q A G Y K V A V C A
ORF334  L I R H G C C L L P L V I K T T R N M P T E I
rfbJ    L A I K Y R R A D I F L N T D S F F A K K
ORF353  S V K R C K H V L T V S V T V K D S S I R K T

            80                  90                    100
ORF224  A K N K S Y D A Y I K S A L H L S F I A I I F F I A L F
ORF311  A V V R S H N Y T P R E E F Q R Y F D T G V F
ORF334  I V S R T K G K R R R I L I D F I A Y L K L A N D E
rfbJ    F N Y Q H M R P Y I I T K R H F D E I G H Y Y A N M H
ORF353  N V P S E K I S V I Y N S V N K D F I O K K E K E N Y F

                  110          120                130
ORF224  I F S G V I S A K Y L S S F H E V L Q D K T R M L F F T
ORF334  S K N I S L K R A F E N V L L K E R P W L Y T
rfbJ    I S F I N M R L E H V Y G P G D G E N K F I P Y I I
ORF353  A V G A S W P H K F G I P P I E A M A S N T P V I V S

                  140                150                160
ORF224  C L V F S S I G I G A I A Y K I L F A E L V G W K A N L
ORF334  L A M A C Y G N S D E K R D L S E F Y A B V G C N K N M
rfbJ    C L N K K Q S C V K C T T G E Q I R D T I F V D D V
ORF353  I P V F H E V L T N G A L Y V N P D D E K S W Q S A I

ORF224  N A
ORF334  N T
rfbJ    N A
ORF353  N I
```

anything about the function of these proteins the relationship of each of the proteins with each other was examined.

The most extensive relationship found is the relationship that ORF224 has with *rfbJ* (distance -78), ORF127 (distance -89), ORF334 (distance -89), ORF353 (-73) and with ORF311 (distance -64). An analysis of the homologies in this group of predicted proteins shows one region of common amino acid homology (Fig. 5.6). As can be seen the amino acid sequence of each of the proteins has homology with all the others and this homology is quite extensive. Conceivably this homologous region of amino acid residues is a domain with similar functions in all of the proteins; the conserved amino acids playing an important structural or catalytic function. It is not possible to determine what that function might be at this stage, it does seems unlikely that these function in the same pathway; *rfbJ* functions in the abequose pathway, however it is most unlikely that ORF311 does, because a deletion that disrupts the fuction of *rfbM* (the abequose pathway is complete in this deletion) lies within ORF311. However it is conceivable that dinucleotide-sugar binding sites have some similarities. This data could provide valuable information in analysing the functional units of these proteins.

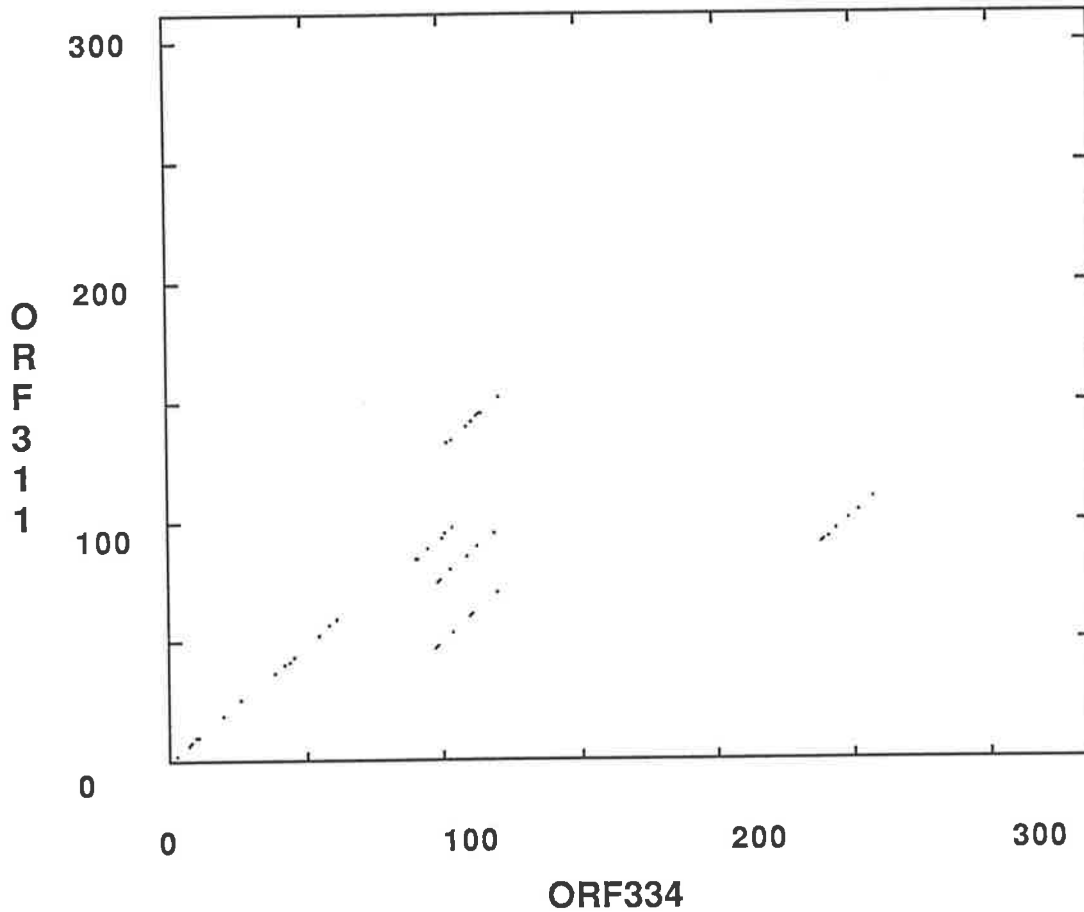
The closest relationship between any two proteins is the one between ORF334 and ORF 311 with a distance of -102. Figure 5.7 produced with the aid of the MTX suite of programs (Reisner and Bucholtz, 1986) using Swanson's algorith, (1984) illustrates the regions of most significant homology between the two predicted proteins. The most significant region of homology is between amino acids 1-122 of ORF311 and amino acids 1 to 118 of ORF334. The diagonal plot represents areas of homology between the two proteins. Figure 5.7

FIGURE 5.7

RELATIONSHIP OF AMINO ACID SEQUENCE OF TWO PREDICTED PROTEINS; ORF311 AND ORF334

The relationship between the amino acid sequences of the proteins ORF311 and ORF334 is shown here. **A** shows the output from the program MTX when the alogarith of Swanson *et al.*, (1982) is used to show the similarity of the amino acid sequences of these two proteins. The amino acids of ORF311, along the x-axis, are compared with the amino acids of ORF334, along the y-axis, dots are used to indicate homology. **B** shows the amino acid homologies found by SEQHP that lie close to the central row of diagonal dots seen in the MTX output. Distance refers to the relationship between the proteins, "1-208 (7)" indicates that the amino acids 1-208 of ORF311 are shown and that 7 spaces have been added to get a better alignment with the amino acid sequence of ORF344, similarly 1-208 amino acids of ORF344 are shown, and 7 spaces have beeen inserted into this sequences to get a better alignment with ORF311.

A



B

```
DISTANCE   -102,    ORF311 - 1-208(7), ORF334 - 1-208(7).

         10        20        30          40        50        60
MKITLIIPTYNAGSLWPNVLDAIKQQTIYP  DKLIVI DSGSKDETVPLASDLKNISIFNID SKD
: :   :: ::         :  :  :     :  : : :  : : :        :   : :
MLISFCIPSYNRKQYLEELLNSINNQEKFNLDIEICISDNASTDGTEEMIDVWRNNYNFPIIYRRN
         10        20        30        40        50        60
     70        80        90        100       110       120
  FNHGGTRNLAVAKTL  DADVI  IFLTQDAILADSDAIKNLVYYFSDPLIAAVCGRQLPHKDANPLQ
  : :  ::       :   :   ::   :: ::  :    :  :          : :     :
SVNLGPDRNFLASVSLANGDYCWIFGSDDA  LAK  DSLAILQTYLDSQADIYLCDRKETGCDLVEIR
  70        80        90         100        110       120       130
130       140       150       160       170       180       180
CMPEILIIVQNLLLSKADIEKLGIKTVFMSNSFAAYRRSVFEELSGFPEHTILAEDMFMAAKMIQAG
            :       :       :         ::  ::                :  :
NPHRSWLRTDDELYV  FN  NNLD  REIYLSRCLSI  G  GVFSYLSSLIVKKERWDAIDFDASYIGTS
      140         150        160         170        180       190
   200
YKVAYCAEA  VVRHS
:          ::
YPHVFIMMTYLIRHG
    200
```

also illustrates the amino acid homology between the two proteins in this region as configured by SEQHP. The charge/hydrophobicity profiles of the two predicted proteins also show some resemblance, in that both are negatively charged in their N terminus. ORF334 is largely negatively charged up to residue 160, and ORF311 is largely negatively charge up to residue 140. The secondary structure profile is similar, in that a large number of short beta sheets are predicted for both protein segments. Perhaps the N-terminal segments of these proteins have related functions; it is too early to say what function that might be.

5.3.5. Sequence Analysis - comparison of *rfbJ* with *galE*.

The function of the *rfbJ* gene product is known and, chemically, well characterized. The enzyme is a dehydrogenase converting CDP-4-keto-3,6-dideoxy-D-glucose to CDP-3,6,-dideoxy-D-glucose, transferring two hydrogen atoms from NADPH to the keto moiety on the carbon 4 (Matsuhashi and Strominger, 1967). Enzyme UDP-galactose-4-epimerase encoded by the gene *galE* epimerises UDP-galactose to UDP-glucose and the reverse reaction. This reaction involves as part of its mechanism a NAD(P) dependent reduction of a keto group on carbon 4 (Walsh, 1979) so in some ways its reaction is similar to that of *rfbJ*. For that reason we chose to compare the amino acid sequence predicted from our sequence and the amino acid sequence predicted from the recently sequenced *E. coli* K12 *galE* gene (Lemaire and Müller-Hill, 1986).

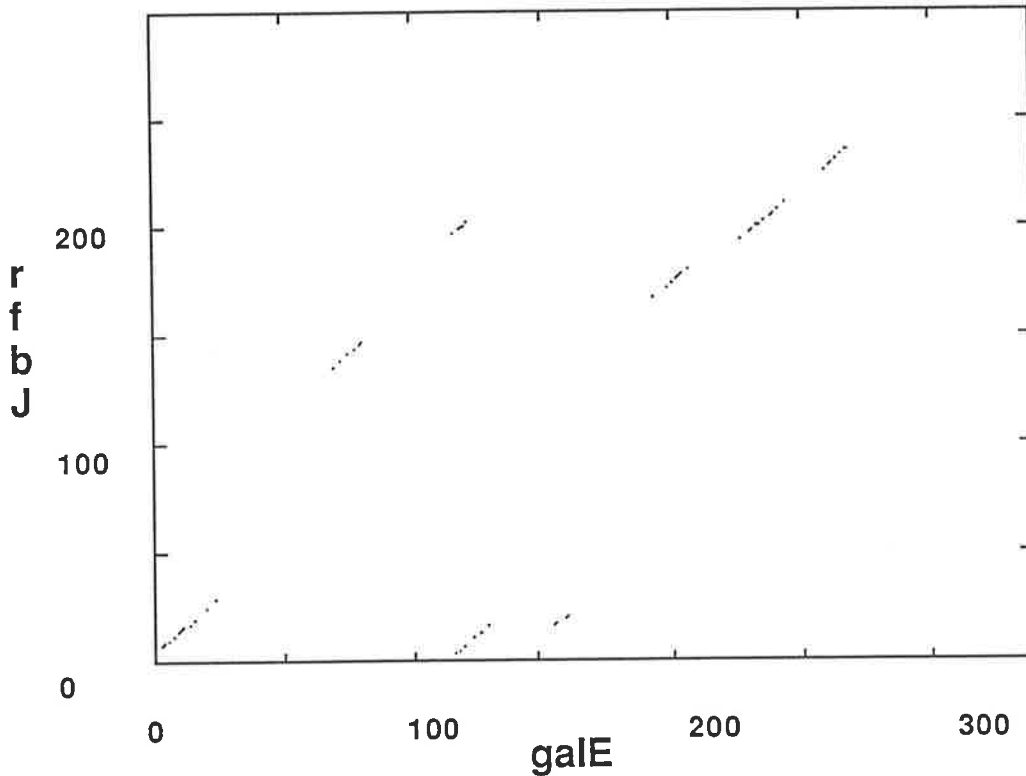
Using the MTX suite of programs (Reisner and Bucholtz, 1986) homology in the amino acid sequence of the two proteins was assessed using Dayhoff's algorith (Dayhoff *et al.*, 1978). A small region in the N-terminus is highly conserved, a longer region of homology is

observed toward the carboxy terminus, as shown by the diagonal plot representing homologies between the two proteins (Fig 5.8). There are also offset homologies between the two proteins but these homologies do not appear to contribute to the overall pattern of homology between the two proteins. Similarities of specific amino acids strings of the two proteins were ascertained using the program SEQHP (Kanehisa, 1982). The amino acids are highly conserved in the N-terminal region; of the 13 amino acids in this region (*rfbJ*, 8-20 and *galE*, 3-15) 8 are identical, and 4 are similar, one is different. Moderate homology is seen in the latter portion of the two proteins (*rfbJ*, 192-283 and *galE*, 224-313). The *galE* protein is more highly charged than is *rfbJ* as seen by their hydrophobicity indices; ANALYSEQ gives *rfbJ* a value of 16.18 whereas *galE* has an index of -26.9. The hydrophobicity plots produced by ANALYSEQ have no distinctive features, consequently it is difficult to find convincing similarities, even in the areas of amino acid similarity. The first 20 amino acids of both proteins are hydrophobic; that these two regions have similar charge characteristics is not surprising considering the extent of amino acid similarity. The secondary structure prediction shows regions of similarity in other parts of the two proteins. The highly conserved region at the N-terminus, shows some similarity, although not complete similarity; *galE* shows a prediction of beta, turn, beta, whereas *rfbJ* shows a beta turn, alpha prediction over that sequence. In the region of moderate amino acid similarity common secondary structural features are also found, reinforcing the validity of the moderate amino acid homology prediction. This region extends in *rfbJ* from amino acid 175 to 255, and in *galE* from amino acid residue 200 to 280. This region predicts several beta strands, and three alpha helices, one lengthy alpha helix is predicted (*rfbJ*; 195-210 and *galE*; 232-250) and two flanking alpha helices that are

FIGURE 5.8

RELATIONSHIP OF THE AMINO ACID SEQUENCE OF TWO PROTEINS; *rfbJ* AND *galE*

The relationship between the amino acid sequences of the proteins producd by *rfbJ* and *galE* is shown here. **A** shows the output from the program MTX when the alogarith of Dayhoff *et al.*, (1978) is used to show the similarity of the amino acid sequence of these two proteins. The amino acids of *rfbJ* along the x-axis, is compared with the amino acid sequence of *galE*, along the y-axis dots are used to indicate homology. **B** and **C** show the amino acids that are common to the two sequences aligned by the program SEQHP. **B** shows the similarities extending from the N-terminal region of both proteins, and **C** shows a second alignment. Distance refers to the relationship between the proteins. In **B** and **C** the numbers outside of parentheses, after the distance refer to the amino acids used in the comparison, the number in parenthesis refers to the number of spaces added to the sequence to obtains the displayed alignment.



B

```
DISTANCE  -104,    galE  -  3-238(8),    rfbJ  -  8-244(7)

       10             20              30             40             50             60
VLVTGGSGYIGSHTCVQLLQNGHDVI  ILDNLCNSKRSVLPVIERLGGKHPTFVEGDIRNEALMT
: : : :: :: :    :   :  :  :           :
VIVSGASGFIGKHLLEALKKSGISVVAITRDVIKNNSNALANVRWCSCDNIELLVEELSIDSALI
 10             20              30             40             50             60             70
 70             80             90             100            110            120            130
EILHDHAIDTVIHFAGLKAVGES  VQKPLEYYDNNDN  GTLRLISAMRAANVKNFIFSSSATVYG
 : :  :       :       : :::     :                   : :
GIIH  LATEYGHKTSSLINIEDANVIKPLKLLDLAIKYRADIFLNTDSFFAKKDFNYQHMRPYII
          80             90             100            110            120            130
          140            150            160            170            180            190
DQPKIPYVESFPTGTPQSPYGKSKLMVEQILTDLQKAQPDWSIALLRYFNPVGAHPSGDMGEDPQ
                          :  :                      :        ::
TKRHFDEIGHYYANMHDISFVNMRL   EHVYGP  GDGENKFIPYIIDCLNKKQSCVKCTTGEQIR
 140            150            160                170            180            190
      200            210            220            230
G  I  PNNLM  PYIAQVAVGRRDSLAIFGNDYPTEDGTGV   RDYIHVMD
  :          :        :         :    : :    :
DFIFVDDVVNAYLTILE  NRKE  VPSYT  EYQVGTGAGVSLKDFLVYLQ
200          210                220          230          240
```

C

```
DISTANCE   -87,      galE  -  225-313(4),      rfbJ  -  192-283(1)

   230            240            250            260            270            280
EDGTGVRDYIHVMDLADGHVVAMEKLANKPGVHMYNLGAGVGNSVLD  VVN  AFSKACGKPVNYHF
  :   :: : : :          :     :    :  : : : :   :   :         :      :
TTGEQIRDFIFVDDVVNAYLTILENRKEVPSYTEYQVGTGAGVSLKDFLVYLQNTMMPGSSSIFEF
        200            210            220            230            240            250
290            300            310
AP    RREGAFRPTGRTPAKPTVNLNWR
       :           :
GAIEQRDNEIM   FSVANNKNLKAMGWK
260            270            280
```

shorter and predicted by fewer of the methods employed by PREDICT. Another common structure predicted is the lengthy alpha helix predicted early on in the sequence (*rfbJ*, 55-80, *galE*, 55-87). This lies in a region with no predicted amino acid homologies; the similarity of secondary structure prediction in this region suggests that similarities between the two proteins lies throughout their entire length, and not just the areas defined by amino acid homologies.

5.3.6. Sequence Analysis - Dehydrogenase structure.

The protein encoded by *rfbJ* is a nicotamide linked dehydrogenase. NAD^+-linked dehydrogenases represent a class of dinucleotide binding proteins that have been extensively studied by X-ray crystallographic techniques. One domain common to all the NAD^+-linked dehydrogenases, the NAD binding domain, consists of a six-stranded parallel beta sheet and four connecting alpha helices, two on each side of the sheet. The 6 beta strands are designated A, B, C, D, E and F, and the alpha helices are designated B, C, D, E and F depending on which beta strand they precede. Adenosine 5'-monophosphate binds in the same way to all four known dehydrogenase structures, and in each case to the last residue in beta A which is a glycine (Rossman *et al.*, 1975).

All this work has been done on NAD binding dehydrogenases, but not much work has been done with NADP binding domains, and what has been done suggests a slightly different structure (Matthews *et al.*, 1977). These workers suggest a structure with eight beta strands, again the two inner most strands containing active sites at least in DHFR (dihydrofolate reductase), however it would be difficult to draw analogies with this enzyme because the substrate and coenzyme binding functions of DHFR are carried out by overlapping portions of

the amino acid sequence rather than by separate domains (Matthews *et al.*, 1978).

The PREDICT program (Eliopoulos *et al.*, 1982) predicts several beta strands within the first 140 amino acids of the *rfbJ* protein. The amino acids of the first three beta strands show a very good fit with regard to the amino acids thought to be necessary for structural stability and for function in the NADH binding domain (Fig. 5.9). Although there are likely contenders for beta sheets, D, E, and F, lying between 83-90, 110- 117 and 132-139, we could not find amino acids that correspond to those predicted to have functions in other NAD binding domains. The amino acids of these three beta strands may be able to act to keep the structural integrity of the active site and to bind the nicotinamide moiety of NAD, but we could not say that with any confidence. It does, however, seem likely that they are a part of such a domain given their proximity to the first three beta sheets and the very good fit that these have with the proposed NAD^+ binding domains of several other dehydrogenases (Fig. 5.9).

The protein encoded by the *galE* gene also has beta strand and alpha helix predictions that fit with the model for the NAD^+ binding domain. In the case of the galE protein, however, similarities in amino acid composition can only be found for the first two functional segments of that domain (Fig. 5.9). As with *rfbJ* beta strands that could correspond to those of the other functional units of the model NAD^+ binding domain are predicted, but amino acids that correspond to the amino acids thought to have function are not found.

It seems clear that the first 80 amino acids of *rfb J* play a role in the NADPH binding, and one could reasonably suggest that the first 140

FIGURE 5.9

AMINO ACID ALIGNMENT OF *rfbJ* AND *galE* IN RELATION TO AMINO ACIDS OF NAD+ BINDING DOMAINS.

The alignment of amino acids in three of the predicted beta strands of proteins encoded by *rfbJ* and *galE* is shown in relation to amino acids with specific functions in published NAD binding domains. The international single letter code is used to designate individual amino aicds. The alignment of amino acid sequences of Dogfish LDH (lactate dehydrogenase), Pig GAPDH (glyceraldehyde-3-phosphate dehydrogenase), Lobster GAPDH, Yeast GADPH, Horse LADH (liver alcohol dehydrogenase), Rat LADH and Bovine GluDH (glutamate dehydrogenase) is adapted from Table IV of Rossman et al., (1975). Where amino acids are similar to those of proposed function in the table of Rossman et al., (1975) homologies are shown by enclosing them in rectangles. The amino acids with proposed function have, either H, denoting a location opposite an alpha helix, S, denoting a location opposite a beta strand, or F, denoting a substrate binding function beneath them.

	beta-A												
Dogfish LDH	N	K	I	T	V	V	G		C	B	A	V	G
Pig GAPDH	V	K	V	G	V	D	G		F	G	R	I	G
Lobster GAPDH	S	K	I	G	I	D	G		F	G	R	I	G
Yeast GAPDH	V	R	V	A	I	D	G		F	G	R	I	G
Horse LADH	S	T	C	A	V	F	G		L	G	G	V	G
Rat LADH	S	T	C	A	V	F	G		L	G	G	V	G
rfbJ	E	Y	V	I	V	S	G	A	S	G	F	I	G
galE	M	R	V	L	V	T	G	G	S	G	Y	I	G
			H		H		F						S

	alpha-B													beta-B												
Dogfish LDH	M	A	D	A	I	S	V	L	M	K	D	L	A	D	E	V	A	L	V	-	D	V	M	E	D	K
Pig GAPDH	R	L	V	T	R	A	A	F	N	S	G	K	V	D	I	V	A	I	N	-	D	P	F	I	D	L
Lobster GAPDH	R	L	V	L	R	A	A	L	S	C	G	-	V	Q	V	V	A	V	N	-	D	P	F	I	A	L
Yeast GAPDH	R	L	V	M	R	I	A	L	S	R	P	B	A	Z	V	V	A	S	B		B	P	F	I	B	L
Horse LADH	L	S	V	I	M	G	C	K	A	A	G	-	A	A	R	I	I	G	V	-	D	I	N	K	D	K
Rat LADH	L	S	V	V	I	G	C	K	T	A	G	-	A	A	K	I	I	A	V	-	D	I	N	K	D	K
rfbJ		K	H	L	L	E	A	I	K	K	S	G	I	S	V	V	A	I	T	R	D	V	I	K	N	
galE		S	H	T	C	V	Q	L	L	Q	N	G	-	H	D	V	I	I	L	-	D	N	L	C	N	S
			S	S			S	S								H		H			F					

	beta-C							
Dogfish LDH	A	K	I	V	S	G	K	D
Pig GAPDH	K	A	I	T	I	F	Q	E
Lobster GAPDH	K	K	I	T	Y	F	N	E
Yeast GAPDH	K	K	I	A	T	Y	Q	E
Horse LADH	G	A	T	E	C	V	N	P
Rat LADH	G	A	T	D	C	I	N	P
rfbJ	A	L	I	G	I	I	H	L
galE	L	M	T	E	I	L	H	D
			H		H		H	

amino acids are involved in a 6 beta strand structure. The latter portion of *rfbJ*, i.e., between amino acids 140 and 299 is, by default, probably involved in binding and modifying the substrate. The homology found between *rfbJ* and *galE* that lies between 192 and 283 of rfbJ becomes significant; perhaps narrowing down the CDP-4-keto-3,6-dideoxy-D-glucose binding site to lie within the 50 or so amino acids between 140 and 192, and the catalytic function in the last 107 amino acids of *rfb J*. However one could not rule out the possibility that in fact a separate NADP+ binding domain exists, and that overlapping amino acids are involved in both coenzyme and substrate binding. The homology found in the very highly conserved segment of 13 amino acids close to the N termini of both these proteins would go beyond the homology expected on the basis that both of these proteins are dehydrogenases, they indicate another, shared factor in the mechanisms of action of both of these enzymes over and above the shared dehydrogenase function. Perhaps these 13 amino acids are involved in specifying the interaction of NADPH with the keto group on carbon 4.

5.3.7. Sequence Analysis - accuracy of the sequence data.

Figure 5.10 gives a summary of restriction endonuclease sites predicted using the program ANALYSEQ (Staden, 1984b). These tally with the restriction map published by myself and H. Brahmbhatt (Brahmbhatt *et al.*, 1987) constructed on the basis of restriction enzyme analysis of cloned DNA.

5.3.8. Sequence Analysis - Special properties; low G+C content

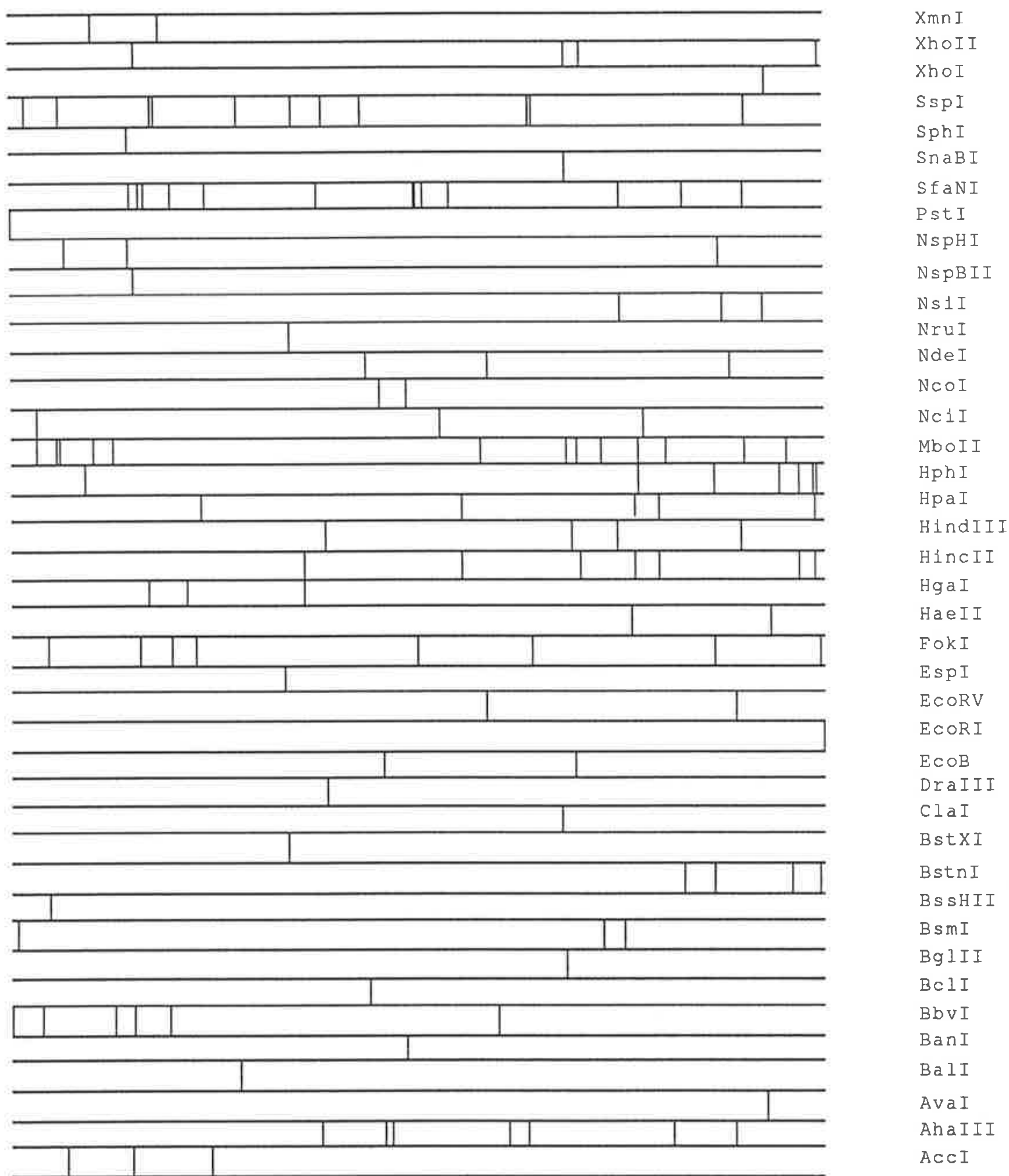
One very striking feature of the sequence of this region is the low G+C content of the sequence; G + C bases account for only 33.6% of all

FIGURE 5.10

RESTRICTION ENZYME SITES PREDICTED FOR THE SEQUENCED REGION.

Restriction enzyme sites were predicted using the program ANALYSEQ. **A** shows a plot of each enzyme that was found by ANALYSEQ plotted next to its name shown in relation to sequence position as defined in fig. 5.2 shown below. **B** shows those enzymes that were not present.

A



B

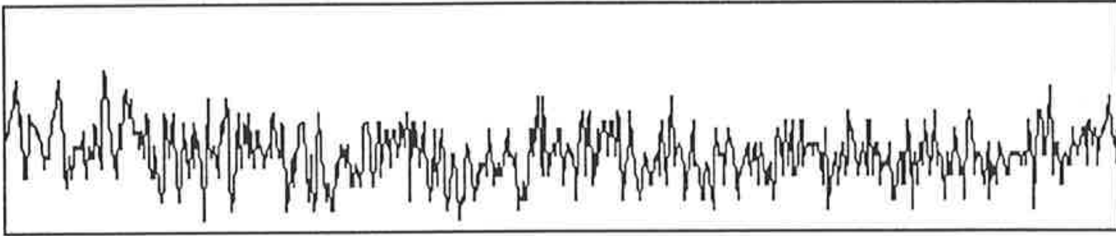
AatII, AhaII, ApaI, AsuII, AvaII, AvrII, BamHI, BanII, BglI, Bsp1286, BstEII, DraII, EcoK, HgiAI, KpnI, MluI, MstI, MstII, NaeI, NarI, NotI, PvuI, PvuII, SacI, SacII, SalI, ScaI, SfiI, SmaI, SpeI, StuI, Tth111I, XbaI and XmaIII.

FIGURE 5.11

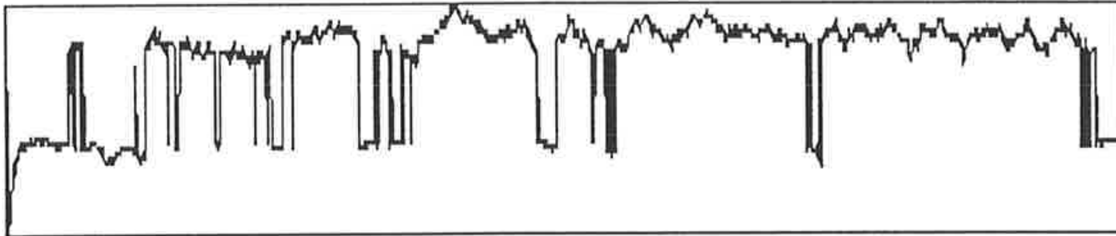
VARIATION IN G+C CONTENT OF THE SEQUENCED DNA AND ASSOCIATED PARAMETERS.

This figure is a compilation of five different plots, arrived at by ANALYSEQ, based on the base composition of the sequenced DNA. **A** shows the variation in G+C content. **B** plots the negentropy. This is a measure of the variation of the compostion of each base, the program compares the base composition of the whole region (T,C,A,G = 0.346, 0.146, 0.318, 0.190) with the base composition of span in question, the program assigns a value based on the similarity, to the whole sequence composition, the values in this instance range from 0.006199 to 0.001984. **C** is a plot in local deviation in base compostion, the base composition of a defined window length is compared to the base composition of the overall sequence. Plots are drawn to fill the window space rather than to follow a particular scale, therefore maxima and minima are indicated to show the range plotted. The maximum value derived here is 35.126, the minimum is 0.029. **D** shows the plot of local deviation in dinucleotide compostion, maximum value is 161.314, the minimum is 10.494. **E** is a plot of local deviation in trinucleotide compostion, maximum value arrived at is 144.73175, the minimum is 28.04539. Beneath this are shown the predicted proteins, and a scale of the number of nucleotides of sequence.

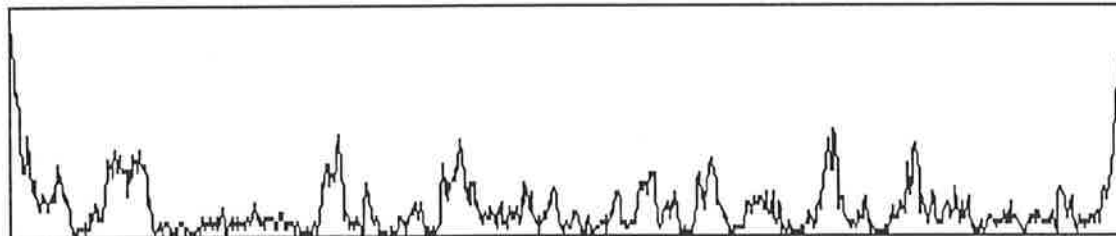
A



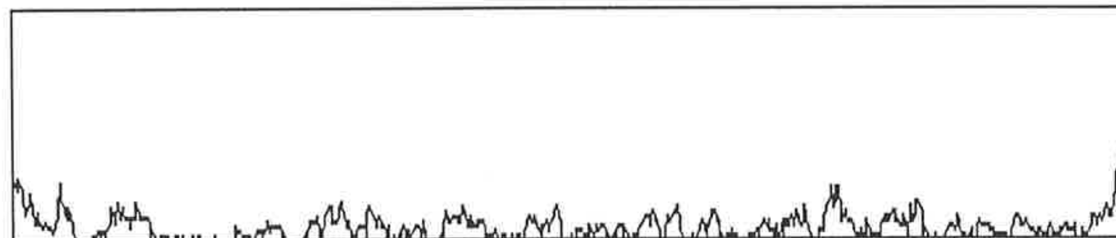
B



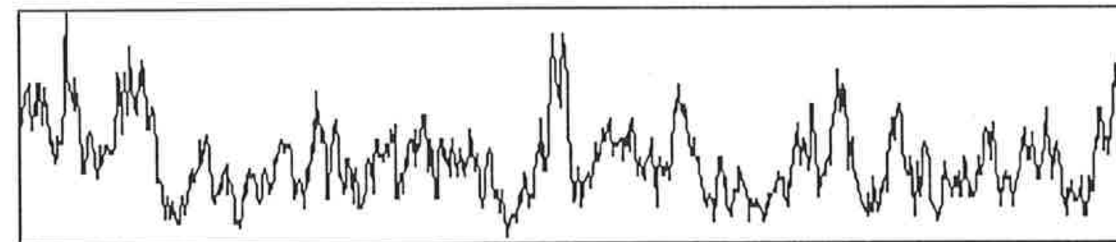
C

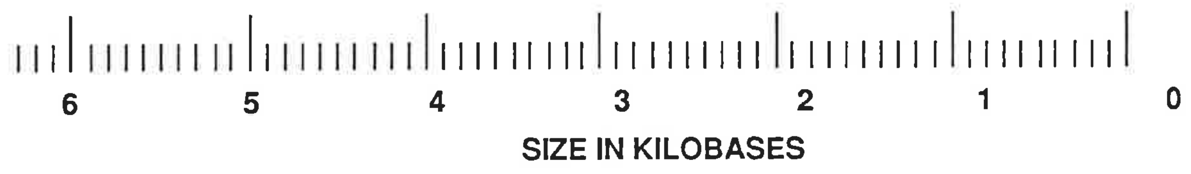


D



E





bases. This is highly atypical of *S. typhimurium* which has an overall G+C content of 51% (Nomore, 1973)

The variation in G+C content of the region sequenced was plotted, using ANALYSEQ, to see if any patterns emerged in relation to G+C content. Figure 5.11, part A shows that although there are variations in G+C content throughout the region, the overall G+C content is low throughout. There appears to be a higher G+C content to the left hand end of the sequence, approximately 720bp from the left hand end but this is not very clear because of the large local variation in G+C content. When the sequence is broken down to segments, of genes and intergenic regions (Table 5.5), the pattern becomes clearer. *rfb J*, ORF224, ORF127, ORF334 and ORF353 have G+C contents ranging from 30.2% to 33.6%. The sequence that codes for the N-terminal portion of a protein to the left hand side of the sequence has a G+C content of 43.7%, clearly much higher. ORF311 has an intermediate G+C content of 37.8%, but if we assume that, indeed, there are two regions of distinct G+C content then this protein is most peculiar in that it is encoded in part on both sorts of DNA and an intermediate G+C content would be expected.

The G+C content of intergenic spaces usually diverges more from 50% than that of coding DNA (Muto and Osawa, 1987) and we find that the G+C content of the intergenic spaces of the postulated low G+C content region follow that pattern; they are 26.2, 29.9 and 24.7%, lower than the coding regions. The one intergenic region that clearly has high G+C content, namely that region between sequence positions 5961 and 6074, has a content of 44.7%. The region to the extreme right hand side has a G+C content of 38.9%, and although this is clearly higher than the low G+C content region it is not as high as the higher G+C

TABLE 5.5

BASE COMPOSITION OF CODING AND NON CODING REGIONS IN THE SEQUENCED DNA.

The percentage content of the four bases in the coding and non-coding region of the DNA sequence are tabulated, as are their G+C content. The regions are listed in the first column, the regions are ordered from sequence position 0, such that the space referred to first refers to the intergenic space starting from position 1 and ending at position 355.

	%T	%C	%A	%G	%G+C
space	32.49	17.51	28.81	21.19	38.9
rfbJ	32.11	12.15	35.90	19.84	32.0
space	39.51	11.11	35.80	13.58	24.7
ORF224	38.12	14.2	30.04	17.64	31.8
space	40.08	13.92	29.96	16.03	29.9
ORF127	40.7	12.1	29.1	18.1	30.2
ORF334	33.7	13.1	33.9	19.3	33.6
space	42.99	10.59	30.84	15.58	26.2
ORF353	32.8	14.7	33.6	18.9	32.6
ORF311	32.7	17.4	29.5	20.4	37.9
space	29.82	21.93	25.44	22.81	44.7
ORF80	30.4	23.3	25.8	20.4	43.7

content region on the left hand side. Presumably the change in G+C content occurs within this region, and it therefore has an intermediate value. Although there do appear to be regions of higher G+C content to either side of the bulk of the DNA that has been sequenced, no clear boundaries can be discerned from the plot of G+C content in figure 5.11 A.

The base content of the sense strand of the various segments of sequence as shown in table 5.5 indicates that the content of cytidine is what is most deficient in the low G+C content region. The value of 17.51% cytidine content of the space to the right of *rfbJ* is very close to the value of 17.4% for ORF311, another region which, presumably, has two distinct parts as far as G+C content is concerned. These values are significantly higher than the values found within the region with low G+C content. This supports the proposition that the higher G+C content at the right end of the sequence represents a boundary beyond which only higher G+C content DNA is found.

Figure 5.11 parts B, C, D and E give a better indication of the boundaries of the low G+C content region. The parameters plotted have a marked change at sequence position 736, corresponding to restriction map position 4.006. The negentropy plot gives an extremely sharp cutoff point, as do the plots for local deviation in base composition and plot for local deviation in trinucleotide composition. The plot for local deviations in dinucleotide composition is less convincing, but overall there seems little doubt about the difference between the DNA to left of the boundary at position 736 and the DNA to the right of that boundary. Less convincing is the right boundary, a marked cut off point can be drawn at postion 220 from the right of the EcoRI (9.57) site on the basis of the negentropy plot. The length of

the region that encodes the lower negentropy value on the right side of the plot is not much longer than some of the regions within the rest of the lower G+C content region. The plot of local deviations in trinucleotide composition shows a similar cut off point, again there are places within the rest of the low G+C content region that have only slightly shorter tracts with high values. The plots of local deviation in base composition and dinucleotide composition are not at all convincing with regards to this boundary. Thus it seems that there are two different sorts of DNA in the region that has been sequenced, one characterised by a G+C content of 32.3% and another characterised by a G+C content of about 44%. One boundary between these two types of DNA lies at position 4.006, but we do not have sufficient data to draw a firm conclusion about a second possible boundary at position 9.35.

5.3.9. Sequence Analysis - Codon usage and amino acid composition.

Table 5.6 shows the codon usage tables for the predicted proteins as well as a codon usage table that is a compilation of all the codons used in genes encoded by *S. typhimurium* that have been sequenced and entered in the GenBank sequence database (Marayama *et al.*, 1986).

There is a large difference between the codon usage usually found in *S. typhimurium* and the codon usage found in the predicted open reading frames. In general codons with a lower G+C content are present in greater numbers in genes encoded in the region that we have sequenced than they are in *S. typhimurium* generally, clearly a reflection of the G+C content of the DNA encoding the open reading frames.

TABLE 5.6

CODON USAGE TABLES FOR EACH OF THE PREDICTED PROTEINS.

Shown are the codon usage tables of the predicted proteins. Amino acids are denoted by their international one-letter code. The number of codons of each sort used are shown to the right of the codons as listed. Also included in this table, for comparison, is a codon usage table derived from sequences of genes of *S. typhimurium* tabulated from the GenBank sequence database (Release 38.0 (Nov 1985)), (Maruyama *et al.*, 1986). The codon frequencies are shown as codon usage per thousand codons in this instance only.

rfbJ

F	TTT	14.	S	TCT	3.	Y	TAT	13.	C	TGT	1.
F	TTC	2.	S	TCC	1.	Y	TAC	1.	C	TGC	3.
L	TTA	14.	S	TCA	6.	*	TAA	0.	*	TGA	0.
L	TTG	5.	S	TCG	2.	*	TAG	0.	W	TGG	3.
L	CTT	2.	P	CCT	3.	H	CAT	7.	R	CGT	0.
L	CTC	2.	P	CCC	0.	H	CAC	1.	R	CGC	0.
L	CTA	2.	P	CCA	4.	Q	CAA	4.	R	CGA	2.
L	CTG	2.	P	CCG	0.	Q	CAG	2.	R	CGG	3.
I	ATT	14.	T	ACT	6.	N	AAT	20.	S	AGT	7.
I	ATC	5.	T	ACC	1.	N	AAC	1.	S	AGC	0.
I	ATA	10.	T	ACA	5.	K	AAA	21.	R	AGA	5.
M	ATG	8.	T	ACG	0.	K	AAG	3.	R	AGG	0.
V	GTT	5.	A	GCT	5.	D	GAT	15.	G	GGT	6.
V	GTC	3.	A	GCC	1.	D	GAC	3.	G	GGC	3.
V	GTA	9.	A	GCA	8.	E	GAA	14.	G	GGA	2.
V	GTG	2.	A	GCG	3.	E	GAG	5.	G	GGG	7.

ORF224

F	TTT	11.	S	TCT	7.	Y	TAT	15.	C	TGT	2.
F	TTC	1.	S	TCC	2.	Y	TAC	2.	C	TGC	2.
L	TTA	13.	S	TCA	9.	*	TAA	0.	*	TGA	0.
L	TTG	5.	S	TCG	0.	*	TAG	0.	W	TGG	3.
L	CTT	5.	P	CCT	0.	H	CAT	5.	R	CGT	0.
L	CTC	2.	P	CCC	0.	H	CAC	0.	R	CGC	0.
L	CTA	5.	P	CCA	1.	Q	CAA	2.	R	CGA	0.
L	CTG	3.	P	CCG	1.	Q	CAG	2.	R	CGG	0.
I	ATT	15.	T	ACT	1.	N	AAT	2.	S	AGT	5.
I	ATC	3.	T	ACC	2.	N	AAC	2.	S	AGC	2.
I	ATA	11.	T	ACA	3.	K	AAA	14.	R	AGA	3.
M	ATG	5.	T	ACG	0.	K	AAG	1.	R	AGG	1.
V	GTT	6.	A	GCT	7.	D	GAT	2.	G	GGT	6.
V	GTC	2.	A	GCC	3.	D	GAC	2.	G	GGC	2.
V	GTA	4.	A	GCA	6.	E	GAA	3.	G	GGA	3.
V	GTG	3.	A	GCG	0.	E	GAG	2.	G	GGG	4.

ORF127

F	TTT	7.	S	TCT	4.	Y	TAT	7.	C	TGT	3.
F	TTC	1.	S	TCC	0.	Y	TAC	1.	C	TGC	0.
L	TTA	8.	S	TCA	2.	*	TAA	0.	*	TGA	0.
L	TTG	3.	S	TCG	0.	*	TAG	0.	W	TGG	4.
L	CTT	5.	P	CCT	0.	H	CAT	0.	R	CGT	0.
L	CTC	0.	P	CCC	0.	H	CAC	0.	R	CGC	1.
L	CTA	3.	P	CCA	2.	Q	CAA	4.	R	CGA	0.
L	CTG	1.	P	CCG	0.	Q	CAG	1.	R	CGG	0.
I	ATT	9.	T	ACT	3.	N	AAT	4.	S	AGT	3.
I	ATC	1.	T	ACC	0.	N	AAC	0.	S	AGC	0.
I	ATA	10.	T	ACA	1.	K	AAA	3.	R	AGA	0.
M	ATG	4.	T	ACG	1.	K	AAG	2.	R	AGG	0.
V	GTT	5.	A	GCT	1.	D	GAT	1.	G	GGT	2.
V	GTC	1.	A	GCC	1.	D	GAC	1.	G	GGC	3.
V	GTA	2.	A	GCA	4.	E	GAA	1.	G	GGA	5.
V	GTG	1.	A	GCG	0.	E	GAG	0.	G	GGG	1.

ORF334

F	TTT	14.	S	TCT	7.	Y	TAT	19.	C	TGT	7.
F	TTC	1.	S	TCC	3.	Y	TAC	1.	C	TGC	2.
L	TTA	16.	S	TCA	3.	*	TAA	0.	*	TGA	0.
L	TTG	5.	S	TCG	2.	*	TAG	0.	W	TGG	5.
L	CTT	10.	P	CCT	2.	H	CAT	3.	R	CGT	3.
L	CTC	3.	P	CCC	0.	H	CAC	0.	R	CGC	1.
L	CTA	3.	P	CCA	5.	Q	CAA	3.	R	CGA	4.
L	CTG	0.	P	CCG	1.	Q	CAG	1.	R	CGG	2.
I	ATT	12.	T	ACT	7.	N	AAT	23.	S	AGT	7.
I	ATC	4.	T	ACC	3.	N	AAC	5.	S	AGC	2.
I	ATA	15.	T	ACA	3.	K	AAA	18.	R	AGA	9.
M	ATG	8.	T	ACG	3.	K	AAG	2.	R	AGG	3.
V	GTT	5.	A	GCT	4.	D	GAT	21.	G	GGT	4.
V	GTC	0.	A	GCC	3.	D	GAC	2.	G	GGC	3.
V	GTA	6.	A	GCA	7.	E	GAA	9.	G	GGA	1.
V	GTG	3.	A	GCG	4.	E	GAG	7.	G	GGG	5.

ORF353

F	TTT	15.	S	TCT	7.	Y	TAT	17.	C	TGT	4
F	TTC	1.	S	TCC	3.	Y	TAC	4.	C	TGC	0
L	TTA	15.	S	TCA	7.	*	TAA	0.	*	TGA	0
L	TTG	3.	S	TCG	2.	*	TAG	0.	W	TGG	7
L	CTT	3.	P	CCT	4.	H	CAT	7.	R	CGT	2
L	CTC	2.	P	CCC	1.	H	CAC	2.	R	CGC	1
L	CTA	2.	P	CCA	6.	Q	CAA	5.	R	CGA	4
L	CTG	4.	P	CCG	2.	Q	CAG	4.	R	CGG	1
I	ATT	14.	T	ACT	7.	N	AAT	16.	S	AGT	7
I	ATC	3.	T	ACC	3.	N	AAC	5.	S	AGC	4
I	ATA	11.	T	ACA	3.	K	AAA	24.	R	AGA	3
M	ATG	8.	T	ACG	4.	K	AAG	7.	R	AGG	1
V	GTT	15.	A	GCT	4.	D	GAT	15.	G	GGT	9
V	GTC	3.	A	GCC	1.	D	GAC	4.	G	GGC	1
V	GTA	11.	A	GCA	10.	E	GAA	10.	G	GGA	1
V	GTG	6.	A	GCG	4.	E	GAG	4.	G	GGG	0

ORF311

F	TTT	15.	S	TCT	7.	Y	TAT	11.	C	TGT	3.
F	TTC	5.	S	TCC	3.	Y	TAC	3.	C	TGC	2.
L	TTA	10.	S	TCA	6.	*	TAA	0.	*	TGA	0.
L	TTG	8.	S	TCG	3.	*	TAG	0.	W	TGG	5.
L	CTT	4.	P	CCT	3.	H	CAT	6.	R	CGT	2.
L	CTC	1.	P	CCC	2.	H	CAC	1.	R	CGC	3.
L	CTA	2.	P	CCA	3.	Q	CAA	8.	R	CGA	2.
L	CTG	4.	P	CCG	6.	Q	CAG	4.	R	CGG	0.
I	ATT	16.	T	ACT	4.	N	AAT	14.	S	AGT	3.
I	ATC	1.	T	ACC	3.	N	AAC	0.	S	AGC	1.
I	ATA	9.	T	ACA	5.	K	AAA	15.	R	AGA	5.
M	ATG	7.	T	ACG	2.	K	AAG	6.	R	AGG	0.
V	GTT	9.	A	GCT	8.	D	GAT	13.	G	GGT	9.
V	GTC	1.	A	GCC	4.	D	GAC	5.	G	GGC	3.
V	GTA	4.	A	GCA	8.	E	GAA	7.	G	GGA	2.
V	GTG	1.	A	GCG	6.	E	GAG	6.	G	GGG	2.

ORF80

F	TTT	3.	S	TCT	0.	Y	TAT	0.	C	TGT	2.
F	TTC	4.	S	TCC	0.	Y	TAC	1.	C	TGC	1.
L	TTA	3.	S	TCA	0.	*	TAA	0.	*	TGA	0.
L	TTG	1.	S	TCG	1.	*	TAG	0.	W	TGG	1.
L	CTT	1.	P	CCT	2.	H	CAT	0.	R	CGT	3.
L	CTC	4.	P	CCC	0.	H	CAC	1.	R	CGC	0.
L	CTA	0.	P	CCA	0.	Q	CAA	2.	R	CGA	1.
L	CTG	2.	P	CCG	1.	Q	CAG	2.	R	CGG	1.
I	ATT	4.	T	ACT	3.	N	AAT	7.	S	AGT	0.
I	ATC	1.	T	ACC	1.	N	AAC	0.	S	AGC	1.
I	ATA	1.	T	ACA	0.	K	AAA	0.	R	AGA	3.
M	ATG	1.	T	ACG	0.	K	AAG	2.	R	AGG	0.
V	GTT	0.	A	GCT	2.	D	GAT	2.	G	GGT	0.
V	GTC	1.	A	GCC	1.	D	GAC	2.	G	GGC	2.
V	GTA	1.	A	GCA	2.	E	GAA	3.	G	GGA	0.
V	GTG	0.	A	GCG	3.	E	GAG	0.	G	GGG	0.

S. typhimurium

F	TTT	24.7	S	TCT	9.1	Y	TAT	17.3	C	TGT	5.1
F	TTC	18.0	S	TCC	12.5	Y	TAC	14.2	C	TGC	4.4
L	TTA	16.3	S	TCA	8.8	*	TAA	0	*	TGA	0
L	TTG	12.5	S	TCG	10.2	*	TAG	0	W	TGG	10.7
L	CTT	12.5	P	CCT	8.5	H	CAT	15.6	R	CGT	13.9
L	CTC	9.5	P	CCC	6.1	H	CAC	9.8	R	CGC	25.4
L	CTA	5.4	P	CCA	2.4	Q	CAA	12.2	R	CGA	4.4
L	CTG	49.5	P	CCG	22.7	Q	CAG	30.8	R	CGG	10.2
I	ATT	34.6	T	ACT	5.1	N	AAT	16.3	S	AGT	6.8
I	ATC	24.7	T	ACC	24.1	N	AAC	14.6	S	AGC	14.2
I	ATA	5.4	T	ACA	5.4	K	AAA	36.9	R	AGA	3.0
M	ATG	27.1	T	ACG	15.9	K	AAG	10.2	R	AGG	2.0
V	GTT	9.5	A	GCT	12.2	D	GAT	33.9	G	GGT	17.6
V	GTC	19.7	A	GCC	31.4	D	GAC	14.2	G	GGC	38.3
V	GTA	9.5	A	GCA	9.8	E	GAA	31.2	G	GGA	11.5
V	GTG	24.4	A	GCG	48.1	E	GAG	22.7	G	GGG	16.6

Table 5.7

AMINO ACID CONTENT OF THE PROTEINS PREDICTED FROM THE SEQUENCE DATA.

The amino acids that each of the predicted proteins is made of are shown in the accompanying table. In addition a measure of the variation from the norm is given by the O-E% value; the variation from the average amino acid composition of proteins found in the EMBL sequence library.

rfb J

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	17	4	19	19	16	18	9	29	24	27	9	21	7	6	0	9	12	19	3	14
O-E%	-34	-54	9	6	49	-28	34	116	22	22	57	63	-55	-49	-32	9	-34	-4	-23	38

ORF224

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	16	4	4	5	12	15	5	29	15	33	5	4	2	4	4	25	6	15	3	17
O-E%	-17	-39	-67	-69	49	-20	12	189	2	100	32	-58	-83	-54	-63	60	-56	2	3	124

ORF127

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	6	3	2	1	8	11	0	20	5	20	4	4	2	5	1	9	5	9	4	8
O-E%	-45	-19	-71	-87	75	3	-100	250	-40	113	85	-27	-70	1	-94	1	-35	7	142	85

ORF334

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	18	9	23	16	15	13	3	31	20	37	8	29	8	4	22	24	16	14	5	20
O-E%	-37	-7	25	-20	25	-54	-55	106	-9	50	41	95	-54	-69	34	3	-21	-36	15	76

ORF353

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	19	4	19	14	16	11	9	28	30	30	8	21	13	9	12	30	17	35	7	21
O-E%	-37	-61	-2	-34	26	-63	28	77	29	13	34	39	-29	-34	-30	22	-21	51	53	75

ORF311

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	26	5	18	13	20	16	7	26	21	29	7	14	14	12	12	23	14	15	5	14
O-E%	-2	-44	6	-30	79	-39	13	86	3	26	33	5	-13	-1	-21	6	-26	-27	24	33

ORF80

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	8	3	4	3	7	2	1	6	2	11	1	7	3	4	9	2	4	2	1	1
O-E%	16	29	-9	-39	143	-70	-38	67	-62	86	-26	103	-29	29	104	-64	-18	-62	-4	-53

TABLE 5.8

G+C CONTENT OF THE 1st, 2nd and 3rd BASES IN CODING TRIPLETS OF THE SEQUENCED REGION.

The G+C content in the first second and third position of the triplet codons of each of the predicted proteins is shown here. The headings 1, 2 and 3 under each of the names of the proteins the position of the G+C content under consideration, in the triplet codon. The values for each position are indicated as percentage base content. Numbers to the right of the letter O represent observed percentage of triplets, and numbers to the right of letter E represent the percentage expected on the basis of Muto and Osawa's (1987) findings.

	rfbJ			ORF224		
	1	2	3	1	2	3
O	41.8	30.1	24.0	36.3	33.6	25.6
E	45	35	20	45	35	20

	0RF127			ORF334		
	1	2	3	1	2	3
O	36.2	32.3	30.2	37.4	34.4	25.15
E	43	34	14	45	34	20

	ORF353			ORF311			ORF80		
	1	2	3	1	2	3	1	2	3
O	41.9	32.0	26.9	44.8	37.0	31.5	48.8	38.8	43.8
E	46	35	21	48	36	29	53	41	38

Indeed there is a shift in the amino acid composition of the proteins encoded here to favour amino acids encoded by A+T rich triplets. Amino acids encoded by triplets that have three bases of G and C exclusively give a good indication of the general trend. The observed amino acid composition of each of the proteins predicted in our sequence is compared to the average amino acid composition of all coding sequences of the EMBL sequence library (Staden, 1984c) by the program ANALYSEQ. Table 5.7 shows a comparison of the amino acids used in the predicted open reading frames. Four amino acids have codons with G and C exclusively i) Alanine can be encoded by GCT, GCC, GCA or GCG, and occurs in decreased amounts in all proteins except in ORF80. ii) Glycine can be encoded by GGT, GGC, GGA, or GGG, and all proteins except for ORF127 have a lower than expected glycine content. iii) Proline can be encoded by CCT, CCC, CCA, or CCG, and occurs in lower than expected numbers in all proteins. iv) Arginine can be encoded by AGA, AGG, GGG, GGA, GGC or GGT, arginine is also under-represented in all proteins except for ORF80. Conversely amino acids that are encoded by triplets with all three bases being either an A or a T occur in greater amounts in the proteins, this is particularly noticeable in the amino acid isoleucine encoded by the triplets ATT or ATC. Thus the amino acids used reflects the G+C content of the region of DNA.

ORF80 stands in contrast to the other proteins; the codons used are not entirely typical of a *S. typhimurium* gene, but they are closer than those used in the other open reading frames; the deviation from the norm may reflect two factors, one being the low number of codons read, and the second being that its G+C content is also lower than the norm being 43.7% instead of 51% as is usual for *S. typhimurium*.

5.3.10. SEQUENCE ANALYSIS - G+C content of the 1st, 2nd and 3rd bases.

Muto and Osawa (1987) observed that a linear plot is achieved when the G+C content of the 1st, 2nd and 3rd bases of triplet codons is graphed against the total G+C content of any organism. The slopes of the three different plots varies; the third base of the triplet has more variation than the first and second bases and its G+C content deviates more from 50% than the first two bases. Table 5.8 shows the G+C content of the first, second and third bases of the predicted proteins of the region sequenced, compared to the expected G+C content of the first second and third bases of codons as calculated from the empirical data of Muto and Osawa (1987).

The degree of difference between the G+C content of the first, second and third bases of the triplets used in our sequence is not as great as expected on the basis of Muto and Osawa's data (1987). The trends of the G+C contents of the the first second and third bases of the coding triplets do follow the trends suggested by these workers, including the trend that ORF80 follows in which the G+C content of the second base is lower than that of the third.

One other aspect should be brought into the discussion here, that being the difference in the content of G and the content of C in the sense strand. As mentioned earlier the C content is radically diferent in the low G+C content region, while the G content is only marginally altered in the sense strand. The generally accepted thesis is that the large variation in G+C contents seen in the various bacteria is due to the gradual of change of the G+C content either up or down from a more uniform state to G+C contents that are more extreme. The drift

that has occured in this segment of low G+C content DNA has been to greatly decrease the C content in the sense strand, and only marginally decrease the G content. As more sequence data comes available on the low G+C content bacteria it would be interesting to note how many of them have this characteristic, it may implicate a particular environment, e.g. a repair mechanism that is less efficient at repairing C base substitutions. Certainly one instance of such a balance is seen in the DNA of the single stranded cubic phages phi-R and phi-X-174. The former has DNA with a base content of 23.5% A, 33.0% T, 24.1% G, 19.3% C, and the latter with 24.6% A, 31.5% T, 24.9% G and 19.0% T (Abshire *et al.*, 1973). This DNA does not have as low a G+C content but the C content is much lower than the G content.

5.3.11. Sequence Analysis - Special structures

The entire sequence was searched for repeated sequences or evidence for transpositions in relation to either the region of non-homology with *S. typhi*, and *S. paratyphi*, and in relation to the low G+C content region, however none was found. There is no evidence for a transpostion event, although our search does not rule out that possibility, either we did not find the relevant clues, or other rearrangements have occurred since such an event and therefore made it impossible to detect.

The sequence was searched for the presence of promoters using ANALYSEQ (Staden, 1984a) which employs the algorith of Hawley and McClure, (1983), however genuine promoters were obscured by the low G+C content of the region and none were found. The high level of adenine and thymine residues, leads to an increase in the number of predicted promoters that the program detects, these generally being

AT rich. Physical data presented in chapter four would suggest that a promoter exists between 9.54 and the beginning of the *rfbJ* gene, but we have not detected it on the basis of a computer prediction. Other promoters if they exist would have to be shown directly, rather than predicted from sequence data, either using a promoter probe vector or looking at RNA transcripts in *S. typhimurium* or in clones of *rfb* in *E. coli*.

We used ANALYSEQ to look for homologies with published consensus sequences (Trifonov and Brandel, 1986) for regulatory or recombination related sites. The distribution of sequences homologous to the lambda repressor binding site (TATCACCCCCAGTGGTA) appears to be significant; of the six times that such sequences are present; starting at positions, 60, 2051, 3278, 3876, 5336, and 6033 only two are in coding regions (3278 and 5336). This may well be coincidence, however, it is possible that a regulatory protein with a recognition site similar to the lambda repressor binding site is active in the region.

5.4. Summary and Conclusion.

The region from PstI at 3.27 to EcoRI at 9.57 was sequenced using the dideoxy sequencing method (Sanger *et al.*, 1977, 1980). DNA used in sequencing was all derived indirectly from lambda AD6. Several open reading frames have been identified and the properties of the predicted proteins are described. Two of the proteins ORF127 and ORF224 appear to have transmembrane segments and no signal sequence suggesting an inner membrane location.

One of the open reading frames, ORF127, has no reasonable ribosome binding site; the distance between the Shine-Dalgarno sequence and

the start codon, in the most likely site is 14 nucleotides. No known ribosome binding site published so far has such a large distance between these two elements, 13 being the greatest distance published to date (Kozak, 1983). The fact that a polyA sequence of 6 nucleotides in length is present between these two elements may be significant. Homopolymeric tracts of adenine-thymine are implicated in an altered structure of DNA, H-DNA. This sort of DNA when adjacent to the usual form of DNA, B-DNA, can lead to kinking (Koo and Cothers, 1987). Given that kinking may occur in the region between the two elements of the ribosome binding site of ORF127, the physical distance between the two will be shorter than suggested by the number of bases, and thus they may still function as a ribosome binding site. It seem unlikely that the length of open reading frame ORF127 would occur by chance with no information encoded, although ANALYSEQ using the method of Shepherd *et al.*, (1981) shows that the pattern of triplet bases used in this area is not usual for protein-coding DNA. This open reading frame has the lowest G+C content for all the proteins; this may interfere with the pattern of bases used within codons; at least 33% G+C is necessary to maintain the codon form RNY at 100%, a lower value of G+C content would progressively decrease the chances of that being obtainable; perhaps, therefore, the low G+C content has intefered with the Shepherd's RNY prediction.

Proteins predicted for this region appear to have some homology with each other, although this does not extend to ORF80, the part of a protein to the left of the sequenced DNA. One region common to all of these proteins has been identified, with extensive amino acid homology. What is suggested is that this common element delineates a functional domain within each of the proteins that encodes it. The suggested domain is unlikely to have an abequose specific function

because one of the proteins, ORF311, is missing in the chromosomal deletion strain P9033 (*his660*) which still has the abequose biosynthetic pathway complete. The data presented with regards to homology would be valuable in an analysis of the functions of these proteins. The argument that these homologies have been detected because of a bias toward amino acids that are encoded by triplets that are A-T rich, particularly isoleucine, can be discounted because of the extent of the homologies; all proteins have homology, and the number of conserved amino acids within the first 115 amino acids of this region is quite high.

Proteins ORF311 and ORF334 have another region of amino acid homology, one that shows a closer relationship between the two proteins than the relationship of proteins that have the common region of homology. Additional features of these two proteins also bear resemblance to each other; they both have a negatively charged N-terminal regions, and have similar predicted secondary structures. Two thirds of the length of these proteins appear to have some similarity; it is thus tempting to suggest that the overall function of these two proteins may be similar or have a similar nature, but what that function might be cannot be ascertained from the data at hand.

Proteins produced by genes *galE* and *rfbJ* also seem closely related with an evolutionary distance of -104 calculated by SEQHP (Kanehisa, 1982) in one region of the amino acid sequence. A very highly conserved sequence of 13 amino acids is found in the N-terminal region of both proteins, 8 residues being identical and 4 similar. The latter portion of the protein has a more moderate, but extended amino acid homology and similar secondary structures are seen in many regions of these proteins. These proteins both have

nicotinamide linked dehydrogenase activity, and amino acids thought to be important within specific secondary structures of nicotinamide-linked dehydrogenases are seen in the same specific structures of both the *galE* protein and the *rfbJ* protein. Three of six regions thought to be important in the dehydrogenase domain are identified in *rfbJ*, and two of these are seen in *galE*, on the basis of specific amino acids found within predicted alpha helices and beta sheets. The secondary structure prediction suggests that the secondary structures that represent the remaining three regions of function (Rossmann *et al.*, 1975) are present, but specific amino acids to fulfil these functions are not seen. In the case of the *rfbJ* protein, this may be explained by the atypical amino acid composition of the protein, perhaps the amino acids used to fulfil the functions of the dehydrogenase domain in *rfbJ* are different, but similar ones, used because of the coding pressure exerted by the low G+C content of this gene. No such explanation can be offered in the case of *galE*, because its G+C content is normal. In any case the nicotinamide binding domain of these two proteins can, in part, be identified.

One striking feature of the region sequenced is the pattern of G+C content within the region, a region of 736 nucleotides on the left ranging between 3.27 and 4.0 (restriction map units) has an average G+C content of approximately 44%, most of the remainder of the sequence has an average G+C content of approximately 32.3%. The extreme right part of the sequenced region also appears to have 200 nucleotides with a higher G+C content. Plots relating to variations in base, dinucleotide and trinucleotide content as well as the plots of G+C content and negentropy do not convincingly show a boundary at the right side of the sequence, but tabulated data of the base composition of the coding and non-coding regions however is more convincing. If

a boundary is present it would be at postion 9.35, however, given the limited amount of DNA sequence to that side, the possibility that what is observed may be a larger than normal variation in the G+C content of the 32% region must be considered. Sequence data to the right of the EcoRI site would be helpful to confirm that the trend to a higher G+C content continued in that direction.

Additionally, although it seems likley that a boundary exists at map position 4.006, the G+C content to the left of that boundary although higher, is not normal for *S. typhimurium*. The G+C content of this region is approximately 44% compared to the average G+C content of the DNA of *S. typhimurium* of 51%. One would thus expect another boundary further to the left of the sequenced DNA.

In any case it seems probable that this DNA taken either from map position 4.0 to 9.35, or from 4.0 to an unknown position to the right of 9.57 was derived from another organism with a lower G+C content, having integrated into the *S. typhimurium* chromosome at this region at some, evolutionarily, earlier stage. This DNA integrated into a region of DNA with an already slightly lower G+C content. The alternate hypothesis that a larger region, one encompassing both the DNA characterised by a G+C content of 32% and the DNA characteristed by a G+C content of 44%, was recombined into the *S. typhimurium* chromosome seems equally likely. No matter which hypothesis is correct, what seems clear is that this region, and therefore at least a portion of the abequose biosynthesis enzyme complement has been aquired from another organism.

The sequenced region was searched for possible structures that could be involved the integration of either the low G+C content region or the

region of non-homology with *S. typhi* and *S. paratyphi*, but none was found. The region was searched for promoters but none was found, however the low G+C content of region was thought to interfere with the search and promoters may be present.

CHAPTER SIX

PROTEINS IN THE REGION ENCODING ABEQUOSE BIOSYNTHESIS

6.1. Introduction

The first enzyme in the pathway for the biosynthesis of abequose, makes CDP-D-glucose from glucose-1-phosphate. This enzyme, CDP-D-glucose pyrophosphorylase (ABE-1) is encoded by the *rfbF* gene in *S. typhimurium* (Nikaido *et al.*, 1967, Levinthal *et al.*, 1969), and has been quite thoroughly characterized in *Y. pseudotuberculosis* (Rubenstein and Strominger, 1974b). It functions as a monomer and has a relative molecular weight of M_r110,000d in *Y. pseudotuberculosis*. Its relative molecular weight has also been estimated in *S. typhimurium* at M_r100,000d (Chojnacki *et al.*, 1968).

The next stage in the biosynthesis of abequose is the conversion of CDP-D-glucose to CDP-4-keto-6-deoxy-D-glucose by the enzyme CDP-D-glucose oxidoreductase (ABE-2). This enzyme is encoded by the gene *rfbG* in *S. typhimurium* (Nikaido *et al.*, 1967, Levinthal *et al.*, 1969) and has been characterized in *Y. pseudotuberculosis* (Gonzalez-Porque and Strominger, 1972a) to function as a dimer with subunit molecular weight of M_r43,000d in *Y. pseudotuberculosis*. This enzyme has not been purified in *S. typhimurium* and therefore no details are available but presumably it has the same characteristics as the *Y. pseudotuberculosis* enzyme.

The final steps in the biosynthesis of CDP-abequose involve the

conversion of CDP-4-keto-6-deoxy-D-glucose to CDP-4-keto-3,6-dideoxy-D-glucose and then to CDP-3,6-dideoxy-D-hexose. These two stages have been studied in detail in *Y. pseudotuberculosis* and two proteins are found to be involved in the first stage; E_1 and E_3 which have a M_r of 61,000d and 41,000d respectively (Gonzalez-Porque and Strominger 1972b, 1972c, Rubenstein and Strominger 1974a). The last step in the biosynthesis of CDP-3,6-dideoxy-D-hexose is catalysed by the E_2 protein (Matsuhashi and Strominger, 1967) but this protein has not been characterized as extensively as the other two enzymes. The same enzymic steps operate in *S. typhimurium* (Pape and Strominger, 1969). In *S. typhimurium* the genes for the various proteins involved in abequose biosynthesis have not been mapped very precisely. The enzymes encoded by *rfbF* and *rfbG* both disappear in the strain carrying the deletion mutation *his*801, and are both present in strains carrying the deletion mutation *his*515, leaving the order of these two genes uncertain. The overall conversion of CDP-4-keto-6-deoxy-D-glucose to CDP-3,6-dideoxy-D-glucose was assayed as one step in establishing the *rfb* gene order, and the designation *rfbH* given to this "gene". Since then Yuasa *et al.*, (1969) have identified a point mutation being defective in the gene encoding E_1, and used the name *rfbH* for this single function. This function was mapped to the same position as the *rfbH* of the older definition was mapped, i.e., between the endpoints of deletion *his* 660 and *his* 515. We have named the genes encoding enzymes E_3 and E_2 *rfbI* and *rfbJ* respectively on the assumption that all three proteins will be present in *S. typhimurium*. We have localized *rfbJ* and this is described in chapter 4 of this thesis, but the order of this gene in relation to *rfbH* and *rfbI* is by no means established. We have a molecular weight estimate for the *rfbF* gene product in *S. typhimurium* (approximately

M_r100,000d) (Chojnacki *et al.*, 1968), but we do not know the size of the other proteins of the abequose biosynthetic pathway in *S. typhimurium*; the enzymes found in *Y. pseudotuberculosis* may provide an indication. The aim of this section of the work was, therefore, to map the genes which encode enzymes involved in abequose biosynthesis in *S. typhimurium*, and assign specific functions to them where possible.

6.2.1. Detection of enzymes encoded by the original clones.

Our first confirmation that ABE-1 and ABE-2 are present in the cloned DNA was obtained by testing *E. coli* K-12 carrying pPR586 and pPR587 for enzyme activity. The gene products of *rfbF* and *rfbG* were measured in *E. coli* strain JM101 carrying pPR586 and pPR587 (Table 6.1) and in wild type *S. typhimurium* P9003, and some of the deletion mutant strains (P9033, P9031, P9034, P9074, and P9029) used in earlier genetic studies (Nikaido *et al.*, 1967, Levinthal *et al.*, 1969). Although cells were grown to approximately the same turbidity, and cell extracts were treated in the same way, no attempt was made to quantify the amount of protein that was used in these assays, therefore, the readings are only an approximate measure of the relative level of enzyme activity in the different cells. Generally we use this data only as a measure of the presence or absence of an enzyme. Despite the lack of attention to standardisation it seems that the clones in *E. coli* K-12 have approximately five times more activity than the original *S. typhimurium* strains. This is presumably because the level of enzyme is boosted by the high copy number of the vector pUC18(19) (Yanisch-Perron *et al.*, 1985). The two clones have approximately the same level of ABE-1 and ABE-2 activity even though the external *lac* promoter reads into the insert DNA from

TABLE 6.1

ABE-1 AND ABE-2 ENZYME ACTIVITY IN STRAINS CARRYING A VARIETY OF *rfb* CLONES

Shown in this tables are representative Optical Density measurements at a wavelength of 325nm, of the assay supernatants for the ABE-1 (CDP-D-Glucose pyrophosphorylase) enzyme assay, and supernatants for the ABE-2 (CDP-D-glucose oxidoreductase) after adjustment for background for each sample. The first column of results shows readings for ABE-1 and ABE-2 assays for some of the chromosomal deletions shown in Fig 1.1 of *S. typhimurium* LT2, for *E. coli* K12 strain JM101, and JM101 carrying pPR586, and pPR587. The second column of readings shows the results of the ABE-1 and ABE-2 assays for a variety of tn1725 insertion mutants of pPR586, pPR586 in JM101 as well as JM101. The third column of results show readings for ABE-1 and ABE enzyme assays in a variety of subclones from the *rfb* region between map positions 9.57 and 15.28.

	ABE-1	ABE-2		ABE-1	ABE-2		ABE-1	ABE-2
pPR586	0.118	0.180	pPR586	0.231	0.263	JM101	0.001	0.000
pPR587	0.143	0.174	pPR606	0.197	0.133	pPR587	0.225	0.206
P9003	0.029	0.045	pPR605	0.261	0.402	pPR884	0.204	0.260
P9033	0.012	0.036	pPR604	0.083	0.085	pPR885	0.275	0.387
P9031	0.024	0.042	pPR603	0.085	0.128	pPR886	0.250	0.439
P9034	0.001	0.007	pPR602	0.152	0.404	pPR887	0.050	0.149
P9074	0.007	0.000	pPR601	0.133	0.260	pPR888	0.000	0.282
P9029	0.003	0.001	pPR600	0.061	0.124	pPR889	0.000	0.000
			pPR599	0.076	0.203	pPR890	0.000	0.000
			pPR598	0.151	0.070			
			pPR597	0.471	0.913			
			pPR596	0.434	0.740			
			pPR595	0.084	0.098			

opposite directions. One might expect expression in strains carrying pPR587 to be considerably higher because the *lac* promoter in this clone is oriented such that it reads in the direction of transcription of the insert DNA. That this is not the case is evidence that there is an *rfb* promoter in this region.

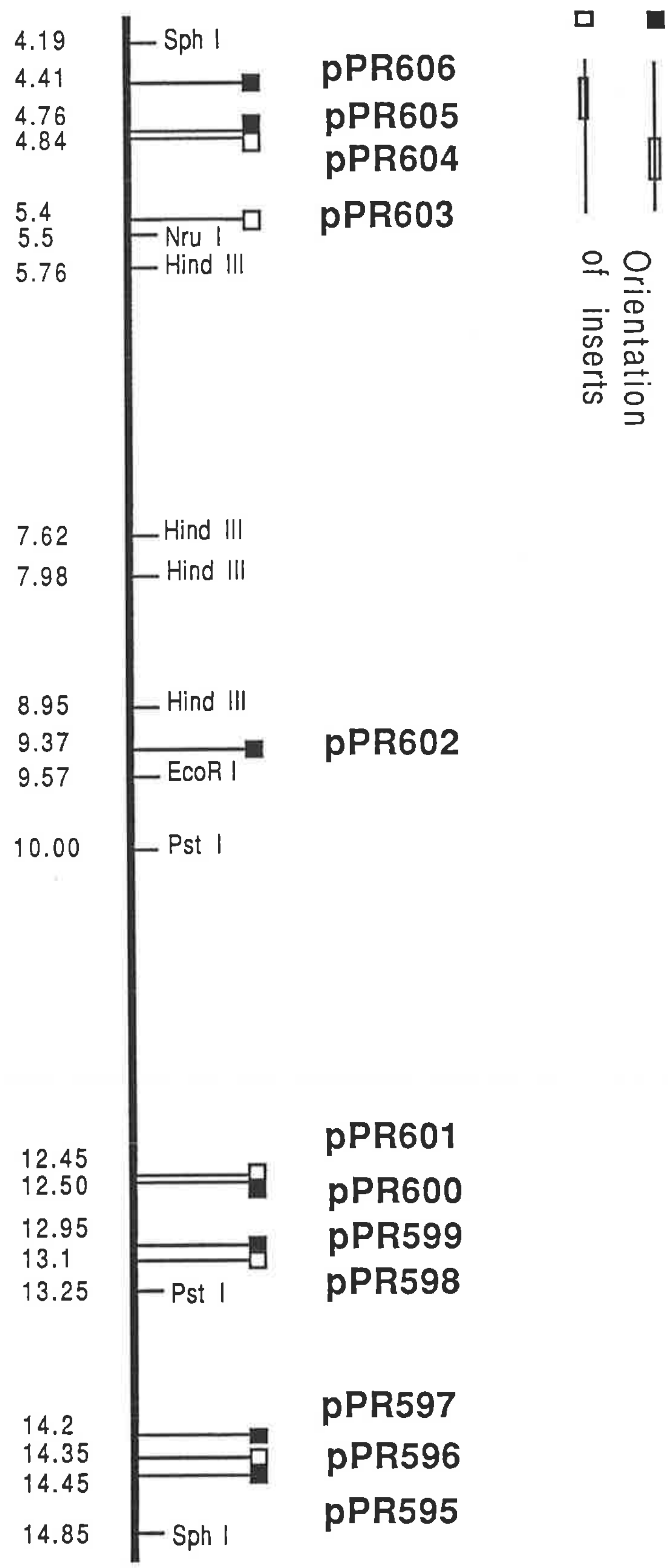
6.2.2. Enzyme activity of tn1725 mutants of pPR586

To further localize these enzymes pPR586 was mutagenized using transposon Tn1725 (Ubben and Schmitt, 1986). This transposon is one of a family of derivatives of Tn1721 all of which have EcoRI sites located 15 bases from either end of the transposon, a feature convenient for mapping the position of the inserted transposon. Most (42 out of 54) of the transposon inserts isolated had generated deletions: the 12 insertion mutant that do not have obvious deletions were mapped using restriction enzymes EcoRI, HindIII and PstI. A very asymmetrical distribution of insertion is seen (Fig. 6.1), with two large regions having no transposon inserts at all. Presumably inserts in these regions are unstable and only deleted versions of these have survived our screening process. The twelve insertion mutants that were used were extremely slow growing, taking 72 hours to get to a stage where cells could be harvested for enzyme assays. The turbidity of the cultures at harvesting was quite variable; all isolates grew differently on repeated attempts at culturing them, indicating some instability in the strains. The levels of enzymes measured in extracts made from repeated attempts at culturing them were variable, and should certainly not be taken as an indication of the relative level of enzymes in these strains (Table 6.1). The presence of ABE-1 and ABE-2 was detected in all the transposon mutants and these two proteins must therefore lie within the region between 5.40

FIGURE 6.1

TRANSPOSON tn1725 INSERTION MUTANTS OF pPR586

This figure shows the position and orientation of the twelve different tn1725 inserts in pPR586 in relation to the position of selected restriction endonuclease sites. To give an indication of the orientation of the inserts a representation of tn1725 is shown with a rectangle representing the chloramphenicol transacetylase gene. Tn1725 inserts are represented with either a small filled square, or a small open square depending on the orientation of the transposon insert.


4.19
Sph I
4.41
pPR606
4.76
pPR605
4.84
pPR604
5.4
pPR603
5.5
Nru I
5.76
Hind III
7.62
Hind III
7.98
Hind III
8.95
Hind III
9.37
pPR602
9.57
EcoR I
10.00
Pst I
pPR601
12.45
12.50
pPR600
12.95
pPR599
13.1
13.25
Pst I
pPR598
pPR597
14.2
14.35
pPR596
14.45
14.85
Sph I
pPR595
Orientation of inserts

and 12.45, and not span 9.37.

These transposon insertion mutant plasmids were transformed into *S. dublin* M6 and screened for 04-specific LPS, using a slide agglutination test. All of them with the exception of pPR602 produced 0-4 specific LPS in *S. dublin* M6. The transposon insert in pPR602 maps 0.20 kb away from the EcoRI site at position 9.57 or 0.155kb from the start of *rfbI*. Thus although we have inactivated the *rfbJ* gene, the reason for loss of expression is due, presumably, to the interruption of transcription from an upstream promoter rather than direct insertion into *rfbJ*.

6.2.3. Localization of the ABE-1 and ABE-2 enzyme activity. Cell extracts from derivatives of strain JM101 carrying the plasmids pPR341, and pPR300 were made and assayed for ABE-1 and ABE-2 activity: the strain carrying pPR341 has both enzymes present and the strain carrying pPR300 has only ABE-2 present. This confirms the order of the genes *rfbF* and *rfbG* as proposed by Nikaido and coworkers (Nikaido *et al.*, 1967). A series of plasmid constructs were made using the restriction sites in the region between 9.57 and 15.28 (Fig. 6.4). The EcoRI to HindIII *rfb* insert in pPR341 was cloned into the EcoRI and HindIII sites of pUC19, to give plasmid pPR884, expression of ABE-1 and ABE-2 was established and pPR884 was cutdown to give a variety of derivative plasmids with varying amounts of *rfb* DNA (Fig. 6.4). Clones carrying the BglII-EcoRI fragment (pPR886), the DraI-EcoRI fragment (pPR889), the HpaI-EcoRI fragment (pPR888), the HindIII-EcoRI fragment (pPR890), the MluI-EcoRI fragment (pPR885), and the XbaI-EcoRI fragment (pPR887) were made. Cell extracts of these clones were prepared and

assayed for the presence of ABE-1 and ABE-2 and results of these experiments can be seen in Table 6.1. All cutdowns including DNA from the EcoRI site at 9.57 to the BglII site at 12.55 had full ABE-1 function, pPR887 (XbaI-EcoRI fragment) has quite low levels of ABE-1 and pPR888 had none. Therefore DNA encoding *rfbF* to lies to the left of the BglII site at 12.55. The low level of activity in pPR887 shows that cutting at the XbaI site interfers with ABE-1 activity, perhaps by truncating the protein resposible for the enzyme activity or by disrupting control elements in the DNA of that region. In any case the start of ABE-1 enzyme is likely to be close to the XbaI site at position 12.38. All pPR884 derivatives including DNA from the EcoRI site to the DraI site had ABE-2 function: therefore the start of ABE-2 is between the DraI site at 11.40 and the HindII site at 10.67.

We have already shown that a fragment of DNA carrying the *rfbJ* gene resides in the region by showing that 04 specific LPS is expressed in salmonellae of groups A and D (described in chapter 4). Sequence analysis has pin-pointed the exact location of this gene (chapter 5) and because we have no direct assay for *rfbJ* no further confirmation of the function of this protein was possible. We were unable to perform the assays that measure activity of the genes *rfbH* and *rfbI* because enzyme substrates were not available, so we could not locate these genes according to their function.

6.3.1. ^{35}S labeled proteins seen in minicells.

Several plasmids were transformed into P2039, a minicell producing strain: minicells were prepared, labeled and electrophoresed on an 11-20% SDS polyacrylamide gel with the resultant autoradiogram shown in figure 6.2.

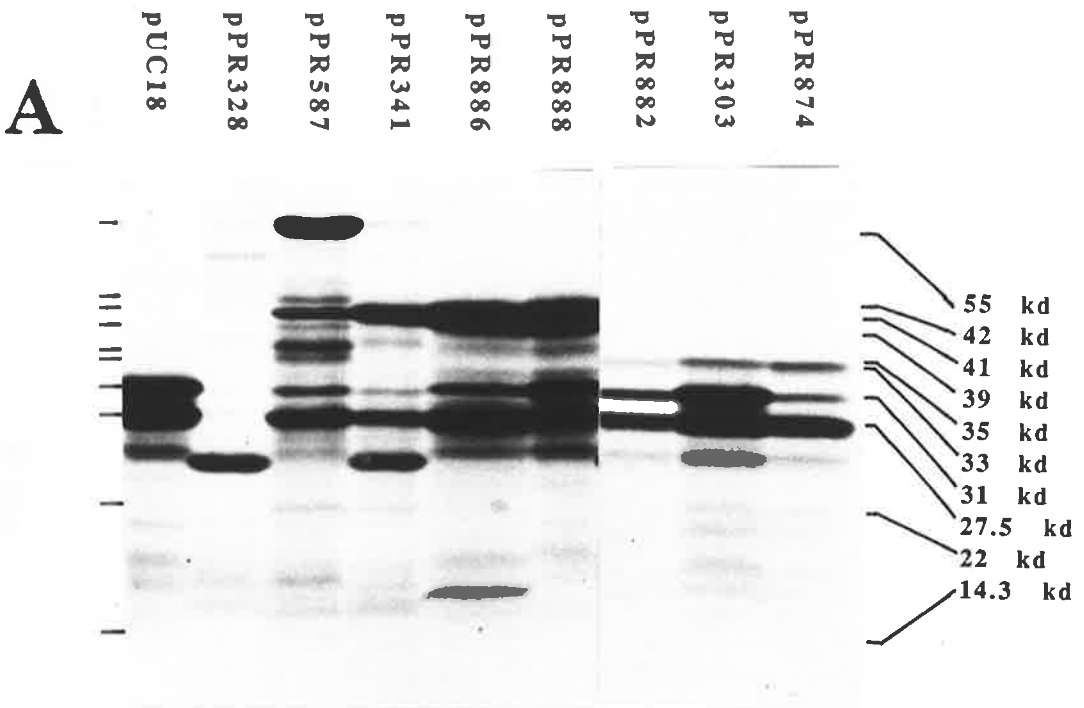
Relative molecular weights of proteins produced in these strains were estimated in an attempt to map the genes that encode these proteins in relation to the restriction data. The plasmid pPR341 produces several proteins in addition to the chloramphenicol transacetylase encoded by its vector, pPR328. The estimated molecular weights of the two most obvious proteins are M_r41,000d and M_r27,500d. Fainter bands correspond to proteins of M_r35,000d and M_r31,000d, two very faint bands, which do not seem to correspond to vector proteins correspond to M_r55,000d and M_r22,000d. Plasmid pPR886, produces two extra bands, in common with pPR341, one of Mr41,000d, and one of Mr27,500d, whereas, pPR888 produces only one extra protein, of Mr41,000d. The difference between pPR886 and pPR888 is not clear in part A of figure 6.2 because the M_r27,500d protein is masked by two TEM beta lactamase related proteins, the active protein of M_r28,000d and a breakdown product of M_r27,000d (Dougan and Kehoe, 1984). The difference between the two strains is seen more clearly in part B of figure 6.2 where the M_r28,000d and the M_r27,000d products of pUC18 are more clearly separated, the *rfb*-specific protein of M_r27,500d lies between the two. Molecular weight estimates for ABE-1 and ABE-2 in *S. typhimurium* are thus M_r27,500d and M_r41,000d respectively. The other proteins produced in the strain carrying pPR341 lie beyond the ABE-1 protein, at least in part beyond the <u>BglII</u> site at position 12.55, and probably entirely beyond the <u>XbaI</u> site at position 12.38. Insufficient coding capacity is available for all of these proteins; a M_r55,000d protein, a M_r22,000d protein a M_r31,000d protein and a M_r35,000d protein if they are not artefacts should be encoded in a space of approximately 3kb of DNA (3.26kb is needed to code for the first three proteins) some of these

FIGURE 6.2

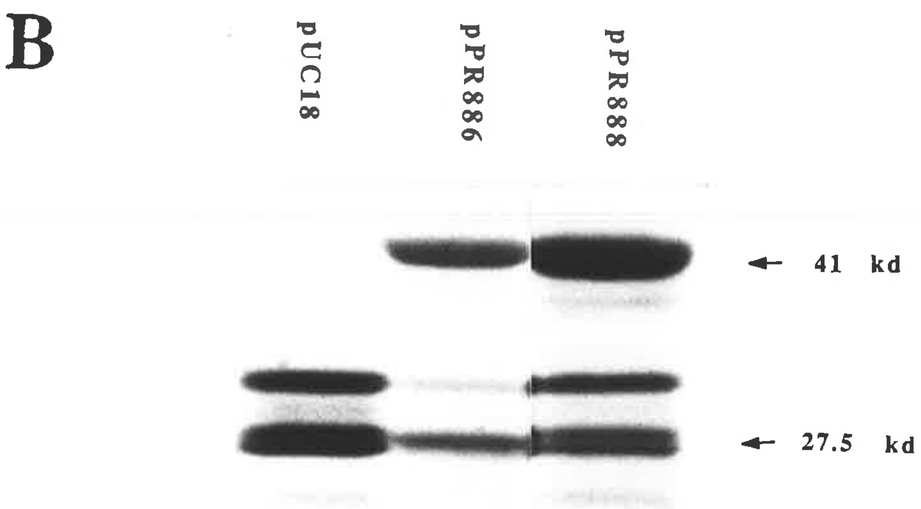
PROTEINS PRODUCED BY MINICELLS IN VARIOUS CLONES CONTAINING *rfb* DNA.

Shown are autoradiographs of ^{35}S labelled proteins, derived from minicell strains, run on 11-20% SDS-polyacrylamide gradient gels. The strains are all derivatives of P2039 that carry pUC18, pPR328 or clones that have *rfb* DNA cloned into these two vectors. pPR587, pPR886, pPR888, pPR883, and pPR874 are derivatives of pUC18, pPR341 is a derivative of pPR328.

B shows more clearly the differences in the protein profiles of pPR886 and pPR888, also shown is the protein profile of pUC18 in this region.



DIRECTION OF ELECTROPHORESIS ▼



are therefore likely to be artefacts. pPR587 probably produces all of the proteins that pPR341 produces; it produces the M_r55,000d protein in much greater quantities, presumably because of the help of the external, *lac* promoter, the M_r31,000d protein may be produced, but if it is produced its production is masked by the beta-lactamase proteins as is the M_r27,000d protein and we have no way of confirming this. One protein that is definitely not produced in the pPR587 clone is the M_r35,000d protein, which may therefore be an artefact.

Clones pPR874, pPR882, and pPR303, produce one extra band that corresponds to a protein of M_r33,000d in minicells. There can be no doubt that this is the protein product of *rfbJ*. We expected pPR883 and pPR303 to produce a protein of M_r24,967d in addition to the M_r34,106d *rfbJ* protein on the basis of a protein (ORF224) predicted by the sequence of the region, but we do not see it. pPR587 produces several extra proteins in minicells that are not in common with what is produced by pPR341, and therefore genes for these could lie to the left of the EcoRI (9.57) site; these correspond to molecular weights of M_r42,000d, M_r39,000d, M_r33,000d. The M_r33,000d protein corresponds to *rfbJ*, the M_r42,000d protein could correspond to the protein predicted as ORF353 by sequence data, the M_r39,000d protein could correspond the to the protein predicted by ORF326, but the M_r35,000kd protein does not fit into any of the predicted open reading frames. A very faint band can be seen at about M_r14,300d, one that is difficult to see in figure 6.2, whether this correspond to ORF127 or not could not be ascertained.

6.3.2. Proteins visualized by Coomassie Brilliant Blue staining in the region 9.57 to 15.28.

Some of the proteins seen after ^{35}S labelling could be seen on Coomassie Brilliant Blue stained gels of the minicells carrying the various *rfb* containing clones and an attempt was made to look at production of these protein in the *E. coli* K12 strain JM101. Denatured whole cells were electrophoresed on SDS polyacrylamide 11-20% gradient gels and stained (Fig. 6.3): a series of proteins seen in these gels that correspond to the bands seen in the minicell producer, albeit at a much lower level. Part A of figure 6.3 shows the disappearance of protein bands as the insert DNA becomes smaller. Bacteria carying pPR884, which has *rfb* DNA from HindIII (15.28) to EcoRI (9.57) as insert, produced four visible proteins, M_r41,000d, M_r31,000d, M_r27,500d, and M_r22,000d. Bacteria carrying pPR885, which has *rfb* DNA from MluI site (14.54) to EcoRI (9.57) loses the M_r31,000d protein indicating that the start of this protein is to the right of the MluI site. The M_r22,000d protein is lost in pPR886 which has *rfb* DNA from BglII (12.55) to EcoRI (9.57), the start of this protein, is somewhere between the end of the M_r31,000d proteins end and the BglII site at position 12.55. The M_r27,500d protein is still present in pPR887, which has *rfb* DNA from XbaI (12.38) to EcoRI (9.57) but is lost in pPR888 which has *rfb* DNA from DraI (11.40) to EcoRI (9.57). The position at which the M_r27,500d protein and thus *rfbF* starts is thus not well defined. Protein expression as visualized by Commassie Brilliant Blue stain whole cell proteins and by expression in minicell producers suggests a start between position 12.38 (XbaI) and 11.40 (DraI). Genetic data using ABE-1 enzyme assay suggests a starting position for *rfbF* past 12.38 or at very least close to 12.38 (see 5.2.3). The possibility that a truncated protein is produced in strains carrying pPR887 could account for these results, if as seems likley, a slightly changed structure resulting from such a

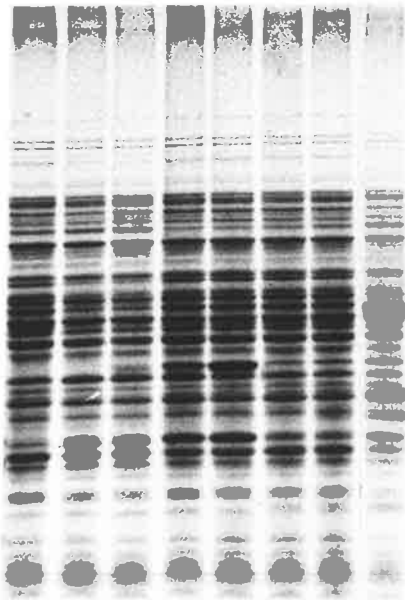
FIGURE 6.3

PROTEINS PRODUCED IN STRAINS CARRYING CLONES WITH *rfb* DNA

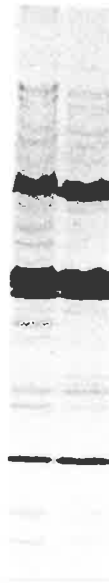
A shows a Coomassie Brilliant Blue stained 11-20% SDS-polyacrylamide gradient gel that has had whole cellular material electrophoresed through it. All strains are derivatives of JM101, with either pUC18 or *rfb* containing derivatives of pUC18 carrying DNA from between positions 9.57 and 15.28. Some of the extra proteins produce bands that are quite faint. Arrows on the side of the gel indicate where extra bands appear in some of the clones.

B shows a 15% SDS-polyacrylamide gel through which denatured whole cells have been electrophoresed. In this case only two tracks are shown, that which has pUC18 in JM101, and a strain which has pPR874. One extra band is visualized as indicated.

A



B



truncation would lead to a marked reduction in enzyme activity of the protein. This truncation if it occurred in strains carrying pPR887 would be very slight, because the protein as seen by Coommassie Brilliant Blue staining is not visibly different to that protein produced in other clones.

The M_r41,000d protein was not lost in the clone, pPR889 that has *rfb* DNA from DraI (11.40) to EcoRI (9.17) but was lost in the clone, pPR890, which has *rfb* DNA from HindII(10.7) to EcoRI (9.57). The start of the M_r41,000d protein coincides with the start of the ABE-2 enzyme being between 11.40 and 10.67.

Proteins that appear in pPR587 tally with the proteins produced in pPR341. The start position of the M_r31,000d protein is therefore further refined to being between SphI (14.85) and MluI (14.54). pPR587 produces an extra protein of M_r55,000d, in addition to those discussed above.

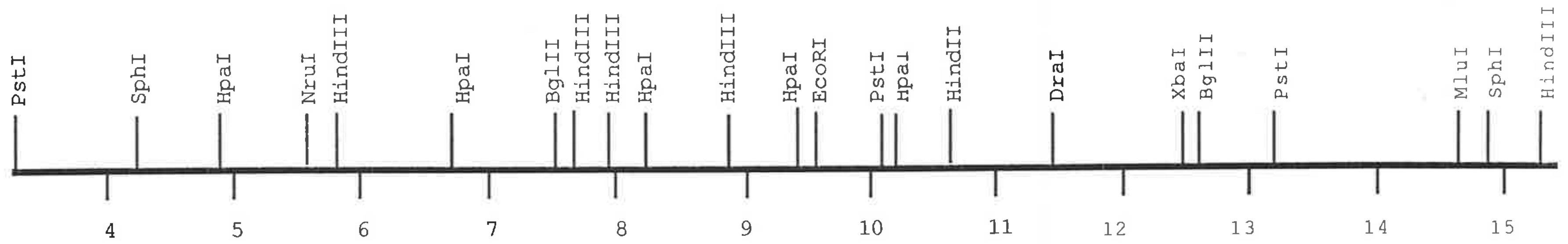
6.4. Summary and conclusions.

We have confirmed the presence of the enzymes ABE-1 and ABE-2 in the cloned region by means of enzyme assays, and we have determined their location in relation to the restriction endonuclease map (Fig. 6.4). In mapping these two enzymes we have determined the gene order of their corresponding genes, *rfbF* and *rfbG*, an order that had been presumed but not proven by the earlier genetic studies (Nikaido *et al.*, 1967, Levinthal *et al.*, 1969). The molecular weights of these two enzymes have been estimated by SDS-polyacrylamide electrophoresis, using ^{35}S labeled minicells and Coomassie Brilliant Blue stained whole cells. The molecular weight of ABE-2 is estimated

FIGURE 6.4

PROTEINS ENCODED IN THAT PART OF THE *rfb* REGION ENCODING ABEQUOSE BIOSYNTHESIS

The extent of *rfb* DNA present in the various clones used to locate proteins are shown in relation to a restriction map. Arrows indicate the direction of transcription from the external *lac* promoter; where two plasmids cover the same region the first plasmid named is oriented such that the external *lac* promoter reads from right to left. At the bottom of the figure estimates of the size and location of the proteins are shown. ORF353, ORF334 and *rfbJ* are localized by sequence data (chapter 5) *rfbG* and *rfbF* are defined by restriction data; the arrows either side of the boxes representing them show the extent of the the region in which they have been localized. The three proteins represented at the right side of the diagram are less certain and they are therefore shown above the other proteins.



pPR587, pPR586

pPR884

pPR341

pPR885

pPR886

pPR887

pPR888

pPR889

pPR890

pPR303

pPR883

pPR874

55,000d

22,000d

31,000d

ORF353 42,000d

ORF334 39,000d

rfbJ 33,000d

rfbG 41,000d

rfbF 27,500d

at M_r41,000d, a value which corresponds quite well with M_r43,000d as seen in *Y. pseudotuberculosis* (Gonzalez-Porque and Strominger, 1972), it is therefore likely that these two forms of the protein would be the same or very similar. Our molecular weight estimate of ABE-1 is M_r27,500d, a value that differs radically from the estimate of M_r110,000d in *Y. pseudotuberculosis* type V (Rubenstein and Strominger, 1974) and from the estimate of M_r100,000d in *S. typhimurium* (Chojnacki *et al.*, 1968). Possible reasons for this difference in the estimated molecular weights are discussed in chapter 7.

The molecular weight of *rfbJ* has been estimated from ^{35}S labeled minicells run of SDS-polyacrylamide gels, as M_r33,000d, this size corresponds quite well to that of M_r34,106d predicted from sequence data.

It appears from Coomassie Blue stained SDS-polyacrylamide gels of clones carrying *rfb* DNA from between EcoRI (9.57) and HindIII (15.28) that two or three other proteins are encoded between map positions 12.38 and 15.28; one of M_r22,000d, and one of M_r31,000d are quite certain. (Fig. 6.4). A M_r55,000d protein appears to be encoded by pPR341, and pPR587, and appears to fit in this region as well, however pPR884 and its derivatives don't seem to encode the M_r55,000d protein. One explanation that can be put forward is that pPR884 and the relevant derivatives of pPR884 have a mutation that stops expression of the M_r55,000d protein and that the M_r22,000d protein as well as the M_r55,000d protein are both produced in the space between the M_r31,000d protein and ABE-1. This explanation seems unlikely because insufficient DNA is available for this, SphI

(14.85) to XbaI (12.38) that is, 2.47kb, is the maximum DNA available to encode these proteins, whereas a total of 3.2kb is necessary to encode them. An alternate explanation is that the M_r22,000d and M_r31,000d proteins are breakdown products of the M_r55,000d protein, the breakdown being more efficient in pPR884 carrying strains than in strains carrying pPR587 or pPR341, an explanation that also seem unlikely given the large variation in expression of the different proteins. The minicells carrying pPR341 do not produce much of the M_r55,000d protein, yet the level of the M_r22,000d and M_r31,000d proteins are the same as in strains carrying pPR587. With no coherent explanation for the M_r55,000d protein we cannot exclude the possibility that the M_r55,000d protein is an artefact.

Molecular weights of two proteins seen in minicells carrying pPR587, that presumably lie to the left of the BglII site at 5.75 correspond to M_r of proteins predicted by sequence analysis: a M_r42,000d protein and a M_r39,000d protein are seen, these estimated molecular weights corresponding quite well with those of the proteins predicted by sequence analysis as ORF353 (M_r40,600d) and ORF326 (M_r38,900d). A very faint band is seen at M_r14,300d which would correspond with the molecular weight of the protein predicted by ORF127 (M_r24,967). An extra protein produced by pPR587 has a M_r35,000d, which cannot be mapped to the right of the EcoRI site or assigned to any of the proteins predicted from the sequence data: it is possibly a breakdown product of a higher molecular weight proteins. One protein that has not been seen but has been predicted from the sequence data is a M_r24,967d protein ORF224: perhaps this protein is not expressed well and cannot be seen in minicells. We would not expect to see a protein corresponding to ORF311, because no clone with DNA extending past

the carboxy terminus of the predicted protein was analysed, the SphI site at 4.19 lies within ORF311.

CHAPTER SEVEN

DISCUSSION

7.1. Introduction.

H. Brahmbhatt, from our group, had cloned some DNA from the *rfb* region of *S. typhimurium* and characterized it. A restriction endonuclease map had been produced and the endpoints of deletions in the strains P9077 (*his1462*) and P9033 (*his660*) had been mapped (Brahmbhatt *et al.*, 1986). The position of these endpoints indicates that the position of genes *rfbP*, *rfbN* and *rfbM* is to the right of map position 4.36 (Brahmbhatt *et al.*, 1986).

7.2.1. Cloning *rfb* DNA.

We describe the cloning of more DNA from the *S. typhimurium rfb* region. A lambda 1059 clone carrying *rfb* DNA was isolated from a gene bank but this clone did not carry sufficient DNA to encode all the abequose functions of this region. DNA derived from this lambda clone, lambda AD6, was used to probe a cosmid bank. DNA from a series of 9 cosmid isolates was cut with EcoRI and an initial restriction map of 40kb of DNA from the region around *rfb* in *S. typhimurium* was constructed. The DNA cloned here was later shown by H. Brahmbhatt to cover the entire *rfb* region, and this data was published jointly (Brahmbhatt *et al.*, 1988).

7.2.2. Localizing the region encoding Abequose biosynthesis genes.

An estimate of the location in this cloned DNA of the genes of the abequose pathway was made by localizing the endpoint of a deletion in the *rfb* region, in strain P9074 (*his*695) (which does not make any of the enzymes involved in abequose biosynthesis (Nikaido *et al.*, 1967, Levinthal *et al.*, 1969)), in relation to a restriction endonuclease map of the region. The endpoint of the deletion mutation in P9074 (*his*695) was found to be to the left of the HindIII site at position 15.28 and therefore all *rfb* genes involved in abequose biosynthesis are likely to lie to the left of that endpoint. The other limit of the region encoding abequose biosynthesis had already been defined by Brahmbhatt *et al.*, (1986) who defined the endpoint of the deletion in strain P9033 (*his*660) (a strain that can synthesize CDP-abequose (Nikaido *et al.*, 1967)), to lie between 4.36 and 4.52. Therefore the genes involved in abequose biosynthesis all lie between 4.36 and 15.28: however that is not to say that all abequose biosynthesis genes must lie in this region (see chapter 3 for a discussion).

7.2.3. Asymmetrical distribution of cosmid clones.

The distribution of *rfb* DNA cloned in our hybrid cosmid clones is asymmetrical (Fig. 3.3): none contain the entire *rfb* gene cluster. An examination of the genes that are likely to be encoded in the various cosmid clones affords an explanation of this observation. The endpoint of the deletion in strain P9077 (*his*1462) lies between position 2.23 and 2.47 (Brahmbhatt *et al.*, 1986) and thus *rfbP* lies to the left of this deletion endpoint. We have defined the deletion endpoint in P9034 (*his*801) to lie between positions 12.38 and 12.48. The deletion endpoint in strain P9074 (*his*695) maps to the left of position 15.28 (*his*695 deletes all the abequose genes and some rhamnose genes *rfbA* and *rfbD*). None of the cosmids pPR398,

pPR403, pPR401, pPR404, pPR400 or pPR405 have any *rfb* DNA beyond position 3, they therefore do not have *rfbP*, the gene encoding UDP-galactosyl transferase. Cosmids pPR402, pPR406 and pPR399 on the other hand don't have *rfb* DNA extending beyond position 15, they therefore would not have DNA encoding all of the rhamnose biosynthesis enzymes. This asymmetrical distribution of clones fits in quite well with the current model proposed to explain lethality of some point mutants in the *rfb* region (Mäkelä and Stocker 1984). The acyl lipid carrier (ACL) upon which the sugars of the O-unit accumulate, is a shared component with peptidoglycan synthesis, and consequently if the pool of ACL is depleted sufficiently, peptidoglycan synthesis decreases to a level where insufficient peptidoglycan is present to maintain structural integrity of the cell wall, and cells lyse. If a point mutant in *rfb* blocks synthesis of mannose, for example, ACL which carries a galactose and a rhamnose residue will accumulate, and prevent reutilization of ACL and the level of peptidoglycan synthesis will diminish. The addition of galactose to ACL is different to the addition of rhamnose, mannose and abequose, in that the addition of galactose is reversible. Mutants in rhamnose biosynthesis should therefore be stable. The distribution of cosmids is such that they either cannot transfer galactose in the case of pPR398, pPR403, pPR401, pPR404, pPR400 and pPR405, or cannot make rhamnose in the case of pPR402, pPR406 and pPR399. Thus in all cases the only possible sugar residue to be transferred to the ACL would be galactose, and this transfer is reversible and would not deplete the cellular pool of ACL. Thus it seems that clones that were able to transfer rhamnose to the ACL have been selected against; *prima facie* evidence that these clones are expressed and interact with the *E. coli* K-12 O-antigen processing mechanisms.

However the absence of a clone that encompasses all of the *rfb* region cannot be explained by this hypothesis. Two possible explanations can be forwarded i) *rfb* DNA of *S. typhimurium* is spread over a larger area than could be cloned into a single cosmid, or ii) the material cloned into *E. coli* K-12 interferes with the cellular processes of the strain, perhaps the multicopy nature of the clones overproduces a by-product, toxic to *E. coli*. The *E. coli* LPS processing system cannot manage to export *S. typhimurium* LPS precursors, with sufficient efficiency, and cosmid clones carrying and overproducing all of the *rfb* region are lethal, through build up of lethal by-products. This lethality could be in the depletion of the pool of ACL or like component leading to a decrease in peptidoglycan synthesis, or perhaps lethality could be due to the build up of a nucleosidic by-product that is toxic to cells. Given that end-product inhibition operates in all of the biosynthetic pathways of the precursor nucleosides (Nikaido and Nikaido, 1966; Mayer and Ginsburg, 1965; Kimata and Suzuki, 1966: Melo and Glaser 1965: Bernstein and Robbins 1965; Kornfeld and Ginsburg 1966) it seems more likely that the problem is at the level of the cell envelope. Thus the products of the *S. typhimurium rfb* region are presumed to be produced in *E. coli*, and seem to interact with its processing machinery.

7.3.1. Localizing *rfbJ* -the gene responsible for 0-4 antigen specificity

We have shown that the 1.23kb HpaI (8.31) to HpaI (9.54) fragment of *S. typhimurium* DNA encodes the E_2 enzyme which catalyzes the last reaction in abequose biosynthesis (Pape and Strominger 1969).

We propose that the gene be called *rfbJ*, and the gene for the E_3 enzyme be known as *rfbI*, because the gene for the E_1 enzyme has already been named *rfbH* (Yuasa *et al.*, 1969): this nomenclature is logical since *rfbI* and *rfbJ* have not yet been used in naming *rfb* genes and the alphabetical order of these names follows the order of enzyme action.

With the localization of *rfbJ* one can now say with certainty that *rfbJ* and *rfbH* lie in the *rfb* region between the deletion endpoint of strains P9033 (*his660*) and P9031 (*his515*). There is still a possibility that *rfbI* may not lie in the group of genes involved in abequose biosynthesis as identified by Nikaido *et al.*, (1967), or indeed the *rfb* region.

The 1.23kb *rfb* DNA insert carried by plasmid pPR874 encodes the *rfbJ* gene as well as 270 nucleotides preceding the start site, and 36 bases after the stop codon of *rfbJ*. It includes DNA from no other open reading frames, and thus the production of 0-4 specific LPS by a pPR874 carrying *S. dublin* M6 is entirely due to the *rfbJ* gene. The abequose residues that are made in *S. dublin* by the *S. typhimurium* E_2 enzyme are handled in the same way as are the tyvelose residues. They are able to transfer to acyl lipid carrier carrying galactose, rhamnose and mannose. Presumably the CDP-abequosyl transferase of *S. typhimurium* is therefore the same as, or similar to the CDP-tyvelosyl transferase of *S. dublin* M6, in that the difference in the two sugars does not prevent transfer. The processing of the 04-like O-unit in *S. dublin* is the same as the 09-like O-unit; they are polymerized and transferred to the core in the same way and the 04-LPS is transported to the outer leaflet of the outer membrane in the

same way as the 09-LPS is, as shown by western blotting. Therefore these processing mechanisms have a sufficiently broad substrate requirement to allow processing of LPS precursors with abequose in place of tyvelose. That the processing is not very specific is expected in view of the lack of specificity found by Shibaev (1978) and Shibaev *et al.*, (1982) in the polymerase system.

pPR874 and its corresponding clone from the group D *Salmonellae* would facilitate immunological experiments relating to the influence of the various dideoxy hexoses on host specificity. Experiments such as those of Valtonen (1970) and the follow-up experiments (Grossman *et al.*, 1986, Liang-Takasaki *et al.*, 1982, 1983) could now be repeated to show, presumably, that the effect that abequose has in increasing lethality of *S. typhimurium* ~~lethality~~ in mice is due to only one enzyme, and not due to an indeterminate length of DNA in the *rfb* region of the *S. typhimurium* chromosome that has been replaced by a corresponding part of the *S. enteritidis* chromosome (tyvelose replaces abequose).

Once the pathways and enzymes of the *rfb* region of *S. typhimurium* are better defined and this definition is extended to related organisms, the potential for dissecting the immunological effects of various sugars of the O-antigen could be realized more effectively.

7.3.2. ORF224 may be an Abequosyl transferase.

There is a difference in the intensity of staining of 04-like LPS produced in *S. dublin* M6 carrying pPR874, and clones carrying *rfb* DNA from BglII (7.57) to EcoRI (9.57), in that *S. dublin* M6 carrying pPR874 stains lighter, and consistently has a weaker reaction in slide

agglutination tests. These tests are not quantitative but the difference is large enough to suggest that the larger clones have something that facilitates processing, or production of abequose. Another protein ORF224 is postulated to exist on the *rfb* DNA of the larger clones, perhaps this has an effect on the processing or production of CDP-abequose. A possible role for ORF224 is that of CDP-abequosyl transferase; one that may transfer abequose more efficiently than the tyvelosyl transferase of *S. dublin* M6. The prediction that ORF224 is a membrane protein, based on sequence data, lends itself to that explanation, given the documented membrane location of the *S. typhimurium* transferases (Osborn and Weiner 1968; Yuasa *et al.* 1969). More direct evidence for a transferase function would be needed for a conclusive answer.

7.4.1. Possible *rfbJ* regulation.

The HpaI fragment cloned in pPR874 gives 0-4 expression in *S. dublin* M6, when cloned in pUC18(19), such that the external *lac* promoter reads in from 8.31 to 9.54, i.e., not in the direction expected to give expression even though the larger EcoRI-BglII (7.57-9.57) fragment gave good expression when cloned in both orientations in relation to an extenal promoter. Expression in clones that do not have external promoters reading in the direction of transcription could be taken as an indication of the presence of a promoter between the HpaI site at 9.54 and the start of the *rfbJ* gene at 9.22.

The inability to clone the HpaI piece in the orientation opposite to that in which it exists in pPR874 could be explained in two ways, one explanation would support the contention that ORF224 is a the CDP-abequosyl transferase. In this explanation a promoter just upstream

to *rfbJ* is either under control of an LPS specific regulator or is much weaker than the *lac* promoter of pUC18, and consequently less E_2 is produced in pPR874, and pPR883. The levels of CDP-abequose in *S. dublin* M6 derivatives carrying these two clones would be lower than in the *S. dublin* M6 carrying pPR303, or pPR882 and these higher levels of CDP-abequose may be toxic to cells. In the larger clones where the two promoters read in the same direction the abequosyl transferase produced is able to clear this toxic level by transporting it more efficiently to the outside of the cell, in the case of the 1.23kb HpaI piece however, no such clearing can take place and therefore it cannot be cloned in that configuration. One other explanation can be offered, that is, that the expression of this gene is under negative regulation and the site at which the regulator acts, perhaps a termination site, lies upstream from the *rfbJ* gene such that it functions in clones with the 2kb EcoRI-BglII fragment but not in the smaller 1.23kb HpaI fragment. Toxic levels of CDP-abequose would only be attained in clones that didn't have the site at which the negative control was exerted and the *lac* promoter reading into the *rfb* DNA such that expression results. Under this hypothesis, lower levels of expression would result in clone pPR883 and in clone pPR874. The latter explanation does not account for the low level of expression in pPR874, as opposed to clones carrying the EcoRI-BglII fragment, the first explanation does not explain the toxicity of the 1.23 HpaI-HpaI fragment in *E. coli.* K-12. If one postulates that ORF224 is the CDP-abequosyl transferase and that a negative controlling element of some sort operates at a site upstream from *rfbJ* such that regulation is not present in the HpaI-HpaI fragment then the observations could be explained.

7.4.2. **Tyvelose expression in hybrid strains.**

Tyvelose is also expressed in these hybrid strains as judged by the strong reaction of the hybrid strains to anti-09 antiserum in a slide agglutination test. This shows that the resident *rfbJ* is not swamped by the multicopy, introduced, *S. typhimurium rfbJ.* Presumably a strong regulatory factor contributes to the maintenance of this balance. Given the likely promoter and postulated regulatory function for the region upstream of the *rfbJ* protein, analysis of the DNA upstream of the start of the *rfbJ* protein and downstream of the EcoRI site at position 9.57 should yield an insight in the regulation of the *rfb* region.

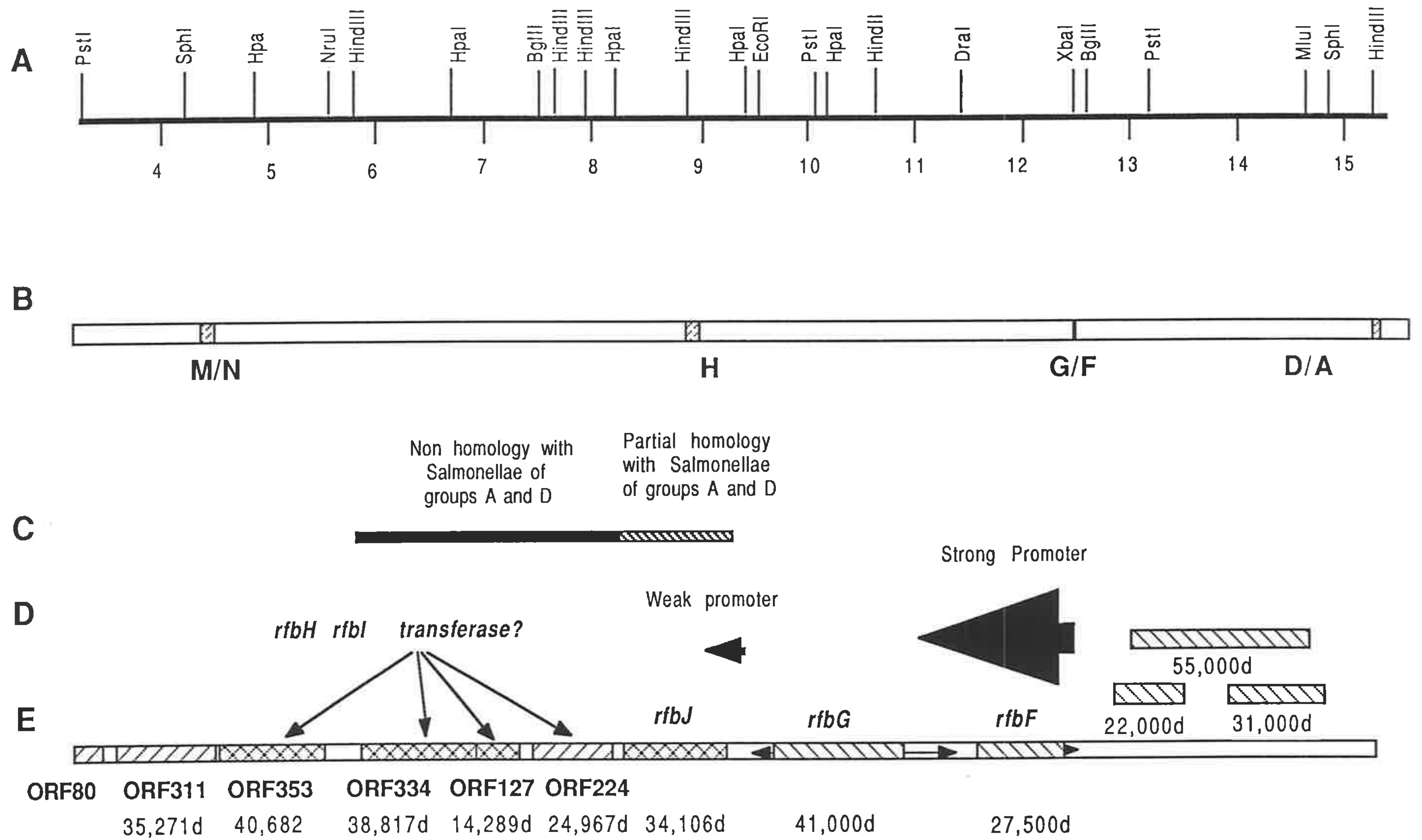
7.5. **Proteins.**

Sequence data predicted several proteins and their predicted properties have been described in chapter five. ORF311, ORF353, ORF334, ORF127 and ORF224 with relative molecular weights of M_r35,271d, M_r40,682d, M_r38,817d, M_r14,289d and M_r24,967d, are described, in addition to *rfbJ* with a relative molecular weight of M_r34,106d (Fig. 7.1). Minicell data, and proteins seen in protein gels stained with Coomassie Brilliant Blue show a series of other proteins to the right of the sequenced region. The gene for a protein of M_r31,000d is the first seen from the right (Fig 7.1), the gene for a protein of M_r22,000d is probably encoded next, and a gene for a protein of M_r55,000d may also be encoded here although the protein is quite possibly an artefact. A protein of M_r27,500d is identified as CDP-D-glucose pyrophosphorylase (ABE-1) and a protein of M_r41,000d is identified as CDP-4-keto-3 deoxy-D-glucose oxidoreductase (ABE-2) the location of genes encoding them is discussed in chapter 6 and is shown in figure 7.1. The protein

FIGURE 7.1

PROTEINS AND POTENTIAL PROMOTERS IN THAT PART OF THE *rfb* REGION ENCODING GENES FOR ABEQUOSE BIOSYNTHESIS

A shows the restriction endonuclease map for the region. **B** shows the location of endpoints of deletion mutations ending in *rfb*. The data shown is largely that of H. Brahmbhatt as jointly published by us (Brahmbhatt *et al.*, 1988). **C** shows the region of non-homology and partial homology found by Verma *et al.*, (1988) when comparing DNA from *S. typhimurium* with *S. typhi* and *S. paratyphi*. **D** shows the positions of the proposed promoters. **E** shows the location and size of proteins in this region. Shading angled down from right to left indicates that a protein has been deduced from sequence data, shading angled down from left to right indicates proteins have been seen on polyacrylamide gels and those proteins with both sorts of shading have been deduced from sequence data and seen on polyacrylamide gels. Proteins sizes are estimates made from polyacrylamide gels unless DNA sequence data is available in which case protein sizes are predicted from sequence data.



A
PstI
SphI
Hpa
NruI
HindIII
HpaI
BglII
HindIII
HindIII
HpaI
HindIII
HpaI
EcoRI
PstI
HpaI
HindIII
DraI
XbaI
BglII
PstI
MluI
SphI
HindIII
4
5
6
7
8
9
10
11
12
13
14
15
B
M/N
H
G/F
D/A
Non homology with Salmonellae of groups A and D
Partial homology with Salmonellae of groups A and D
C
Strong Promoter
D
Weak promoter
rfbH rfbI transferase?
55,000d
22,000d
31,000d
E
rfbJ
rfbG
rfbF
ORF80
ORF311
ORF353
ORF334
ORF127
ORF224
35,271d
40,682
38,817d
14,289d
24,967d
34,106d
41,000d
27,500d

product of *rfbJ* is seen in minicells as a protein of M_r33,000d, corresponding to the size predicted from sequence data of M_r34,106d. When clones carrying DNA covering the region from EcoRI (9.57) down to SphI (4.19) were introduced into minicells strains, a protein corresponding to ORF353 was seen, a protein corresponding to a M_r42,000d and a protein of M_r39,000d could correspond to ORF334, a very faint band of M_r14,300d was also seen which would correspond very well with that expected for ORF127 which has a predicted M_r14,289d. No protein was seen that would correspond to ORF224.

7.5.1. Protein function - *rfbJ*, *rfbG* and *rfbF*.

The function of three of the proteins seen are quite clear. Details of the gene product of *rfbJ* has never been published before and therefore no comparison can be made with earlier results. The gene product of *rfbG* in *Y. pseudotuberculosis* in a purified form (Gonzalez-Porque and Strominger 1972) has been estimated at M_r43,000d, our estimate for the gene product in *S. typhimurium* is M_r41,000d. The gene product for *rfbF* in *Y. pseudotuberculosis* has been estimated at M_r110,000d (Rubenstein and Strominger 1974) and in *S. typhimurium* LT2 has been estimated as M_r100,000d (Chojnacki *et al.*, 1968), our estimate is entirely different; M_r27,500d. Rubenstein and Strominger (1974) treated their purified protein with SDS and boiled it before running it on an SDS Polyacrylamide gel, however Chojnacki *et al.*, (1968) based their estimated on its elution from Sephadex-G-200 columns in 0.145M NaCl and 0.005M Tris HCl. Presumably Rubenstein and Strominger were looking at polypeptides whereas Chojnacki *et al.*, may not have been. At face value therefore, this data would suggest that the protein in *S. typhimurium* is different to the protein in *Y. pseudotuberculosis*, functioning as a tetramer with a

subunit size of M_r27,500d in the former organism instead of a monomer of M_r110,000d in the latter.

Rubenstein and Strominger (1974) used the method of Kamen *et al.*, (1972) in their analysis, a method which differs from the method that we used in the concentrations of the denaturants. Our treatment of samples was to boil them in 2% SDS together with 5% Beta mercaptoethanol prior to electrophoresis, whereas Rubenstein and Strominger used only 0.25% SDS and no beta-mercaptoethanol. Another difference is that samples were electrophoresed in half the concentration of SDS that we used; 0.1% compared to 0.2%. It is thus plausible that our, harsher, treatment was sufficient to dissociate subunits of a multimeric form, whereas the treatment of Rubenstein and Strominger, even though it did involve boiling samples in SDS, was not sufficiently harsh to dissociate subunits and to keep them dissociated. It is interesting to note that some of the ABE-1 enzyme preparations of Rubenstein and Strominger (1974) had a contaminating Mr30,000d protein. This contamination would be explained quite well by our hypothesis that the ABE-1 enzyme is a tetramer; the contamination that Rubenstein and Strominger saw may be due to partial dissociation of the multimeric form of the enzyme.

Although it seems unlikely, we cannot preclude the possibility that the DNA in the *rfb* in our strain of *S. typhimurium* LT2 has had rearrangements such that only a portion of the M_r110,000d ABE-1 protein seen by the earlier workers is functional as an enzyme in our strains.

7.5.2. Protein function - *rfbH* and *rfbI*.

Two other genes (*rfbH* and *rfbI*) whose products we were unable to assay due to unavailability of substrates, encode proteins that have relative molecular weights of M_r61,000d and M_r41,000d respectively in *Y. pseudotuberculosis*. We can attempt to predict which of the open reading frames identified from the sequence data are likely candidates for these two proteins by assuming that they are of a similar relative molecular weight to the proteins found in *Y. pseudotuberculosis*; an unreliable assumption in the light of our findings with regard to the *rfbF* gene product. Abequose genes thus far localized appear to be clustered and on that basis one would expect *rfbH* and *rfbI* to follow on from *rfbJ*. ORF224 is (discussed earlier) suggested to be the abequosyl transferase. That leaves the possibility that *rfbH* and *rfbI* correspond to either ORF127 (M_r14,289d), ORF334 (M_r38,817d), or ORF353 (M_r40,682d); being predicted proteins upstream of the endpoint of the deletion in strain P9033 (*his*660). It seems unlikely that ORF127 corresponds to either gene on the basis of its small size, but which of the other two proteins correspond to *rfbH* or *rfbI* remains to be seen.

The subsequent finding by N. Verma (Verma *et al.*, 1988), based on a comparison with our data, that a region between restriction map positions 5.8 and 8.17 of the *S. typhimurium* restriction map has no

homology with DNA from *S. typhi* and *S. paratyphi* and heteroduplex analysis shows no homology between 8.17 and 9.57 gives us another measure for attempting to assign a position to the *rfbH* and the *rfbI* genes. DNA in the region of non-homology would *prima facie* seem unlikely to encode enzymes related to general 3,6-dideoxyhexose biosynthesis because these functions would presumably be common to all three groups of organisms; and therefore genes encoding *rfbH* and *rfbI* are unlikely to be encoded in that region. Given that *rfbH*, at least, lies between deletion endpoints of strains P9031 (*his*515) and strain P9044 (*his*660) (Yuasa *et al.*, 1969); i.e. between map positions 4.36 and 9.54 (Brahmbhatt et. al., 1988), the only predicted protein that fits both of those criteria is ORF353 and therefore ORF353 is quite likely, under this hypothesis, to be *rfbH*.

Under this hypothesis no open reading frame in the sequenced DNA is likely to be the *rfbI* gene, which most likely lies to the right of the <u>EcoRI</u> (9.57) site. Given the uncertainty of the protein profile encoded by genes to the right of *rfbF* the data presented in this thesis is not sufficient to speculate on the position of the *rfbI* gene, other than to say that it does not lie in the region between the *his* operon and position 12.55 (i.e., the end of the *rfbF* gene). It may well be one of the proteins whose genes map adjacent to the *rfbF* gene, however it may also map almost anywhere else on the chromosome.

If, on the other hand, DNA encoding all three enzymes of the last two steps of abequose biosynthesis (*rfbH*, *rfbI* and *rfbJ*) had been inherited by the Salmonellae of groups A, B and D at the same time, and therefore changes had been occurring for the same length of time for all three, one might expect a similar degree of difference in

homology of the DNA encoding *rfbH* and *rfbI* between the salmonellae ~~*Salmonellae*~~ under consideration as one would for *rfbJ*. If this latter hypothesis is correct then *rfbH* and *rfbI* could both lie in the region of so-called non-homology and could correspond to the predicted proteins ORF334 and ORF353. Which of these two hypotheses is correct can only be confirmed with positive identification of the *rfbH* and *rfbI* gene locations.

7.5. Protein function - proteins of unknown function seen by PAGE.

At the other end of the protein map are probably two proteins of M_r31,000d and M_r22,000d, another protein of M_r55,000d seen on PAGE is possibly an artefact, perhaps a beta-lactamase fusion product.. Three known functions could lie within this region, one or both of the two enzyme fractions that play the role of TDP-rhamnose synthetase, EI and EII (Glaser *et al.*, 1972) may lie within this region, as may the gene product of *rfbA*, the TDP glucose pyrophosphorylase. Any of the two (or three) genes localized to this region could encode any of the functions described above. It is also quite feasible, (whether there are two or whether there are three genes in this region is not important in this assessment) that another function, perhaps *rfbI,* may map here. Chojnacki *et al.*, (1968) give an estimate of approximately M_r50,000d for TDP glucose pyrophosphorylase which is suggestive of the M_r55,000d protein (if it is not an artefact) being TDP glucose pyrophosphorylase. As with *rfbF* however the possibility does exist that these early workers are estimating the molecular weight of a multimeric, native form of the pyrophosphorylase, and it may well be a dimer of either the M_r22,000d protein or of the M_r31,000d protein, or the

pyrophosphorylase may map further to the right. Enzyme assays of cell extracts obtained from strains carrying subclones would give firm evidence of the role of the various proteins. ~~that map here~~.

7.5.5. Protein function - other proteins identified by sequence data.

Three *S. typhimurium*-specific genes lie in the region between *rfbJ* and ORF353, being ORF224, ORF127 and ORF334. The computer program ALOM predicts several transmembrane segments for the protein encoded in ORF224, and there seems to be a hint of more efficient processing of CDP-abequose in *S. dublin* when clones include this open reading frame. These two findings lead~~ing~~ to the only suggestion that we will make for this protein, that being CDP-abequosyl transferase, or perhaps an accessory protein of that transferase, these being functions expected to be localized in a region of *S. typhimurium*-specific DNA. However apart from the suggestion that this is a membrane protein, no other possibilities as far as function is concerned can be gleaned for this protein.

ORF127 appears on the basis of the prediction of the computer program ALOM to have transmembrane segments, and therefore a membrane locality is suggested for this protein, perhaps it is an enzyme involved in processing substitutions to the O-unit.

The deletion endpoint in strain P9033 (*his660*) lies within the coding region of ORF311, so that ORF311 and the protein of which ORF80 ~~is a part~~ are contenders for *rfbN* the gene for dTDP-rhamnosyl transferase, and *rfbM*, the gene for GDP-mannose pyrophosphorylase

(Fig 1.1). These suggested functions could again ~~quite~~ easily be confirmed by looking at levels of enzyme activity in strains carrying subclones encompassing these genes.

Proteins ORF311 and ORF334 have a distant similarity, an observation based on computer analysis of amino acid sequence, charge distribution, and on secondary protein structure predictions. These two proteins would on that basis be contenders for enzymes with similar functions, possibly both pyrophosphorylases or transferases (ORF311 being encoded by either *rfbM* or *rfbN*). No membrane locality is suggested for either of these so perhaps both are pyrophosphorylases, however this suggestion is stretching the data somewhat. ORF344 and ORF353 are candidates for *rfbH* and *rfbI* and certainly one of these two genes is *rfbH*.

7.6.2. An NAD^+ dehydrogenase binding domain in *rfbJ* and *galE*.

It seems clear that the first 80 amino acids of *rfb J* play a role in the NADPH binding, and one could reasonably suggest this for the next 60 amino acids. The latter portion of *rfbJ*, i.e. between amino acids 140 and 299 is, by default, probably involved in binding the substrate. However one could not rule out the possibility that in fact a separate NADP+ binding domain exists, and that overlapping amino acids are involved in both coenzyme and substrate binding. The homology found in the very highly conserved segment of 13 amino acids close to the N termini of both these proteins would go beyond the homology expected on the basis that both of these proteins involve dehydrogenation activity, they indicate another, shared factor in the mechanisms of action of both of these enzymes

over and above the shared dehydrogenase function.

7.7. Much of the sequenced DNA has an extremely low G+C content.

One striking feature of the region sequenced is the pattern of G+C content; a region of 736 nucleotides on the left from 3.27 to 4.0 (restriction map units) has an average G+C content of approximately 44%, most of the remainder of the sequence has an average G+C content of approximately 32.3%. The extreme right part of the sequenced region also appears to have 200 nucleotides with a higher G+C content, although one could not say with confidence that a region of higher G+C content continues past the EcoRI site at 9.57. The G+C content of the 736 nucleotides of DNA to the left of the sequence has a G+C content which is also lower than the average G+C content for *S. typhimurium*; and presumably a boundary between DNA of G+C content of 44% and that of typical (51%) *S. typhimurium* DNA exists still further to the left.

In any case this DNA taken either from map position 4.0 to 9.35, or from 4.0 to an unknown position to the right of 9.57 has a much lower G+C content than DNA to the left of map position 4.0, and indeed the overall G+C content of the *S. typhimurium* chromosome. Clearly whatever the boundaries, the DNA with low G+C has been imported into the *S. typhimurium* chromosome from another organism. Presumably the DNA with a G+C content of 44% is also imported but it remains to be seen whether both sorts of DNA were imported into *S. typhimurium* from the same source. The existance of a protein (ORF311) that sits on the boundary between the 44% and the 32.3% G+C content DNA suggests that these two DNAs have

coexisted for some time and it is therefore likely to have been imported into *S. typhimurium* together.

It seems likely that it is not only abequose biosynthetic genes that are derived from an organism with a G+C content of 32.3%. If one accepts the hypothesis that the region of DNA in *S. typhimurium* that shows non-homology to DNA of salmonellae of groups A and D still encodes abequose biosynthesis enzymes, then one protein encoded here is involved in a non-abequose function; presumably ORF127. If one does not accept this hypothesis then two genes not encoding known abequose biosynthesis functions are located in this region; presumably being ORF127 and ORF334. It is possible that one, or two, proteins are required for abequose related functions in addition to the ones already characterized, however one cannot rule out the possibility that one or both of these predicted proteins encode other functions.

If these proteins do not perform abequose related functions, as seems quite likely, the role suggested for them in biosynthesis and transfer of a substitution is conceptually compatible with being transferred from another organism. If the progenitor organism had these genes present and they added to the variability of the surface carbohydrate, the variablity would be of an advantage to *S. typhimurium* and would be maintained in what seems likely to be one regulon. These similar functions are maintained together, the variation afforded to a related group organisms by the polymorphism in dideoxyhexose content is further expanded by the acquisition of a substitution pathway in *S. typhimurium*.

Another rearrangement which has presumably taken place in this region is that found by N. Verma (Verma *et al.*, 1987) who analysed the *rfb* regions of other organisms in order to compare them with our *S. typhimurium* LT2 restriction map. A region of non-homology is found between map positions 5.8 and 9.57 of the *S. typhimurium* map when comparing this region with the *rfb* region of *S. typhi* and *S. paratyphi*. The sequenced region was searched for regions of homology at the terminal portions of the region of non-homology found between the ~~*Salmonellae*~~ salmonellae examined, and the region immediately to the left of 9.37, however none were present. Searches of sequence data bases and published consensus sequences (Trifonov and Brendel, 1986) as described, found no homologies with insertion elements. Although we found no evidence for possible structures that could be involved in the integration of either the low G+C content region or the region of non-homology with *S. typhi* and *S. paratyphi*, that is not to say that none exist. The region of low G+C content may not be fully sequenced, it may extend beyond 9.37, and some homology may yet exist between the two terminal portions of the low G+C content region.

7.8. Promoters of the *rfb* gene cluster in the region examined.

On the basis of expression of the gene products in this region, good evidence exists for one promoter, and evidence of a much less quantitative nature exists for the presence of a second promoter. Expression of ABE-1 and ABE-2 in strains carrying pPR586 and pPR587 is approximately the same, certainly of the same order based on enzyme assays of extracts of *E. coli* cells carrying them. These two plasmids have approximately 11kb of *rfb* DNA cloned from opposite

ends in pUC18, and therefore the *lac* promoter reads into the cloned DNA in opposite direction. Because there is no strong promoter reading in the orientation opposite to the *lac* promoter in pUC18, one could reasonably suggest the presence of a promoter upstream of *rfbF*. The results of these enzyme assays would place such a promoter between the XbaI site at position 12.38 and the SphI site at 14.85. One other piece of data is relevant here, that being expression of ABE-1 and ABE-2 in transposon inserts in pPR586. Although expression was not quantitated because of the variability in the growth of these cells, reasonable expression of both enzymes was found in all of them including strains carrying pPR586 with transposon inserts upstream from *rfbF*. The transposon insert found closest to the *rfbF* gene, found in plasmid pPR601, is located at position 12.45, i.e., between XbaI at 12.38 and BglII at position 12.55, thus one can suggest that the promoter in this region lies between position 12.45 and the begining of *rfbF*. It should be noted that the orientation of this transposon in pPR601 is such that a promoter known to transcribe from one end of tn1725 (P. Manning *et al.*, personal communication) reads away from the *rfbF* gene and thus the promoter activity seen is not attributable to the tn1725 promoter. To confirm the result found by transposon mutagenesis the EcoRI site generated by the transposon at position 12.45 could be used to advantage, and the EcoRI fragment between position 9.57 to 12.45 could be subcloned and expressed under condition of no external promoters. We have not pursued this matter, but indirect evidence for the location of the promoter can also be seen in the expression of proteins in the minicell system in which the ABE-1 and ABE-2 proteins are the only proteins highly expressed in clones that carry DNA in the region between EcoRI and HindIII (9.57 to 15.28). These

two results suggest the presence of a strong promoter just upstream of *rfbF*.

A second promoter is suggested for a region upstream of *rfbJ* but the evidence for this promoter is based on a less quantitative test, that being intensity of staining of a western blot of whole membrane preparations prepared from hybrid strains of *S. dublin* carrying clones of *S. typhimurium rfb*. Expression is seen in *S. dublin* carrying clones with the EcoRI to BglII (9.57 to 7.57) fragment cloned in both orientations in pUC18(19). Additionally, expression is seen in pPR874 where the insert is cloned in the orientation opposite to that for expression from the external *lac* promoter. The expression in *S. dublin* carrying pPR874 is lower than is seen in *S. dublin* carrying clones with the EcoRI to BglII (9.57 to 7.57) fragment in the same orientation as pPR874, suggesting a promoter of lower strength between the HpaI site at position 9.54 and the beginning of the *rfbJ* gene. Computer analysis of the sequence data in this region shows no promoter that fits the accepted promoter model well. A sequence that fits the accepted promoter model best is at position 9.37, (the -35 region starting at sequence positions 197) but the so-called -35 region and the -10 region of the promoter are 11 nucleotides apart. The two elements of the suggested promoter fit the consensus sequence for promoters exactly but the distance between these two elements is 11 bases whereas generally it is not less than 14. The promoter is one with low expression as judged by our crude measure of expression, perhaps the extremely good fit of the two recognized components of the promoter are offset by the very bad positional fit between the two elements to produce a promoter of low strength. Or perhaps it falls into a category of promoters that does not fit the

consensus model, certainly recent findings of a promoter functional without the so-called -10 region casts doubt on the conformity of all promoters to such a model (Ponnambalam *et al.* 1986). Certainly the promoters suggested for the abequose region should be confirmed, preferably with isolation of the mRNA's of the region and sequence of the initial portion of those mRNAs.

No promoters were found, in the DNA that was sequenced, using the program ANALYSEQ, however the low G+C content of the region may interfere with the computer search and it is felt that more direct means are necessary to screen for promoters in this region.

Transposon inserts into pPR586 were found in non-random locations. A large gap in the middle was found with no inserts apart from one. The suggested explanation of toxic by-product formation may be the answer. Whatever the answer, it is interesting to note that inserts occurred before the suggested promoter between positions 12.38 and 12.45. No inserts are found in the presumptive RNA of this region. A transposon insert is found at position 9.370. ~~great lengths~~ Attempts were ~~taken~~ made to map this transposon precisely, and it is clearly not within the *rfbJ* gene, but the hybrid *S. dublin* strain carrying pPR602, does not express 0-4 specific LPS. The transposon has, it seems, disrupted the expression of *rfbJ* without inserting into the gene, presumably between the gene and the promoter encoding it. Thus in this isolated instance of a transposon in the middle of this region the disruption of the abequose pathway could lead to the same toxic products postulated to build up in other mutants and it is not clear to us why this mutant has survived.

7.9. The genetic structure of the abequose biosynthesis region.

In the region of DNA encoding abequose biosynthesis, it seems that two promoters are active; one upstream of *rfbF*, apparently the first gene in this cluster. The other promoter that we have found is located two genes downstream in front of the *rfbJ* gene. There is no reason to suggest that the transcript read from the first promoter is terminated at or near this point, the suggestion is that this second promoter is a weaker, secondary promoter.

Genes that encode the first two enzymes in the abequose biosynthetic pathway map in the order in which they function, *rfbF* followed by *rfbG*: the genes that encode the remaining three known enzymes are not placed as logically. The last enzyme of the pathway, E_2, is encoded next by *rfbJ*, and after this gene there is some uncertainty; genes with unknown function follow *rfbJ*. One of these genes is known to be *rfbH* (Yuasa *et al.*, 1969), but which one is not clear, we suggest that the gene that follow *rfbJ* is CDP-abequose transferase, although there is no evidence to support that suggestion but we have no firm data relating to the position of *rfbI* and this may be located in a totally different rregion on the *S. typhimurium* chromosome. Quite clearly, there is at least one gene (maybe two) with functions unrelated to abequose biosynthesis interspersed between abequose genes.

If the *rfbI* gene does not map in the sequenced region, as seems possible, and if the boundary of low G+C content is at map position 9.35, as seems likley, an interesting picture of the evolution of the abequose biosynthetic pathway emerges. It would appear that parts

of the pathway have been acquired separately. One piece of DNA now lying between map position 4.006 and 9.35 was introduced from one source and encodes some, perhaps all, of the last steps in abequose biosynthesis. This introduced pathway presumably superceded an earlier pathway that used CDP-4-keto-6-deoxy-D-glucose as a starting substrate. If *rfbI* does not map in this region the pathway becomes even more fragmented, perhaps it represents one of the enzymatic steps of that earlier pathway which is also used in the synthesis of abequose , and maps elsewhere on the chromosome. If either of these two possibilities are correct, then parts of this pathway have been derived from several sources, and not just one source. If this reflects the genetic structure *rfb* regions generally, the acquisition of a new O-antigenic sugar may not be quite as simple as previously thought. Certainly the exchange of partial pathways would increase the potential for the production of new sugars, or novel modifications of existing sugars, and perhaps this is the way that the variability of the surface structures in the *Salmonellae* is further enchanced

BIBLIOGRAPHY

Abshire C.J., Guay R., and Bouley J.P., (1973). The base composition of Bacteriophage nucleic acids. In: Laskin A.I., and Lechevallier H.A. (Eds) "*Handbook of Microbiology*". Vol II (Microbial Composition): pp741-750. CRC press (Cleveland Ohio).

Achtman M., Schwuchow S., Helmuth R., Morelli G., and Manning P.A., (1978). Cell-cell interactions in conjugating *Escherichia coli*: Con$^-$ mutants and stabilization of mating aggregates *Mol. Gen. Genet.,* **164**:171-183.

Anderson M.S., and Raetz C.R.H., (1987). Biosynthesis of Lipid A Precursors in *Escherichia coli*: A cytoplasmic acyltransferase that converts UDP-N-acetylglucosamine to UDP-3-O-(R-3-hydroxymyristoyl)-N-acetylglucosamine. *J. Biol. Chem.,* **262**:5159-5169.

Anderson M.S., Bulawa C.E., Raetz C.R.H., (1985). The biosynthesis of gram-negative endotoxin: Formation of lipid A precursors from UDP-GlcNAc in extracts of *Escherichia coli*. *J. Biol. Chem.,* **260**:15536-15541.

Anderson R.G., Hussey H., and Baddiley J., (1972). Mechanism of wall synthesis in bacteria. The organization of enzymes and isoprenoid phosphates in the membrane. *Biochemical J.,* **127**:11-25.

Barr K., and Rick P.D., (1987). Biosynthesis of Enterobacterial Common Antigen in *Escherichia coli*: *In vitro* synthesis of lipid-linked

intermediates. *J. Biol. Chem.*, **262**:7142-7150.

Bavoil P., and Nikaido N., (1981). Physical interaction between the phage l receptor protein and the carrier immobilized maltose binding protein of *Escherichia coli*. *J. Biol. Chem.*, **256**:11385-11388.

Bayer M.E., (1975). Role of adhesion zones in bacterial cell surface function and biogenesis. In: Tzagoloff A. (Ed), "*Membrane Biogenesis*" pp393-427 Plenum Pbl. Co., New York.

Bayer M.H., Costello G.P., and Bayer M.E., (1982). Isolation and partial characterization of membrane vesicles carrying markers of the membrane adhesion sites. *J. Bacteriol.*, **149**:758-767.

Bayer M.E., (1979). The fusion sites between outer membrane and cytoplasmic membrane of bacteria; their role in membrane assembly and virus infection. In: M. Inouye (Ed.), "*Bacterial outer membranes*". pp167-202. John Wiley & Sons, Inc., New York.

Benton W.D., and Davis R.W., (1977). Screening of recombinant clones by hybridization to single plaques *in situ*. *Science.*, **196**:180-182.

Berst M., Lüderitz O., and Westphal O., (1971). Studies on the structure of T1 lipopolysaccharide. *Eur. J. Biochem.*, **18**:361-368.

Blanden R.V., MacKaness G.B., and Collins F.M., (1966).

Mechanisms of acquired resistance in mouse typhoid. *J. Exp. Med.*, **124**:585-600.

Brahmbhatt H.N., Quigley N.B., and Reeves P.R., (1986). Cloning part of the region encoding biosynthetic enzymes for surface antigens (O-antigen) of *Salmonella typhimurium*. *Mol. Gen. Genet.*, **203**:172-176.

Brahmbhatt H.N., Wyk P.J., Quigley N.B., and Reeves P.R., (1988). A complete physical map of the *rfb* locus encoding biosynthetic enzymes for the O-antigen of *Salmonella typhimurium* LT2. *J. bacteriol.* in press (January).

Braun V., (1975). Covalent lipoprotein from the outer membrane of *Escherichia coli*: Conditions affecting import of maltose binding protein into the periplasm of calcium treated cells. *J. Bacteriol.*, **155**:97-106.

Braun V., and Hantke K., (1974). Biochemistry of the bacterial cell envelope. *Annu. Rev. Biochem.*, **43**:89-121.

Bray D., and Robbins P.W., (1967). The direction of chain growth in *Salmonella anatum* O-antigen biosynthesis. *Biochem. Biophys. Res.,Commun.*, **28**:334-339.

Brozek K.A., Bulawa C.E., and Raetz C.R.H., (1987). Biosynthesis of lipid A precursors in *Escherichia coli*: A membrane-bound enzyme that transfers a palmitoyl residue from a glycerophospholipid to lipid

X. *J. Biol. Chem.,* **262**:5170-5179.

Bruneteau M., Volk A., Singh P.P., and Lüderitz O., (1974). Structural investigation on the *Salmonella* T2 lipopolysaccharide. *Eur. J. Biochem.,* **43**:501-508.

Bucholtz C.A., and Reisner A.H., (1986). MBIS, an integrated system for the retrieval and analyses of sequence data for nucleic acids and proteins. *Nucl. Acids Res.,* **14**:233.

Bullas L.R., Colson C., and Neufeld B., (1980). Deoxyribonucleic acid restriction and modification systems in *Salmonella*: Chromosomally located systems of different serotypes. *J. Bacteriol.,* **141**:275-292.

Bullas L.R., Ryu J-I., (1983). *Salmonella typhimurium* LT2 strains which are $r^- m^+$ for all three chromosomally located systems of DNA restriction and modification. *J. bacteriol.* **156**:471-474.

Burnette S., (1982). "Western Blotting"; electrophoresis of protein from sodium dodecyl sulphate polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radio-iodinated protein A. *Anal. Biochem.,* **112**:195-203.

Chojnacki T., Sawicka, T., and Korzybski T., (1968). Metabolism of nucleoside diphosphate sugars. Separation of nucleoside diphosphate glucose pyrophosphorylases of *S. typhimurium* by gel

filtration on Sephadex G-200. *Acta Biochim. Pol.,* **15**:293.

Clark W.R., McLaughlin J., and Webster M.E., (1958). An aminohexuronic acid as the principal hydrolytic component of the Vi antigen. *J. Biol. Chem.,* **230**:81-89.

Coleman W.G.Jr., and Leive L., (1979). Two mutations which affect the barrier function of the *Escherichia coli* K-12 outer membrane. *J. Bacteriol.,* **139**:899-910.

Coleman W.G.Jr., and Deshpande K.S., (1985). New *cysE-pyrE*-linked *rfa* mutation in *Escherichia coli* K-12 that results in a heptoseless lipopolysaccharide. *J. Bacteriol.,* **161**:1209-1214.

Coleman, W.G.Jr., (1983). The *rfaD* Gene codes for ADP-L-glycero-d-manno heptose epimerase: An enzyme required for lipopolysaccharide core biosynthesis. *J. Biol. Chem.,* **258**:1985-1990.

Collins F.M., (1969). Effect of immune mouse serum on the growth of *Salmonella enteritidis* in non-vaccinated mice challenged by various routes. *J. Bacteriol.,* **97**:667-675.

Coughlin R.T., Tomsager S., and McGroarty E.J., (1983). Quantitation of metal cations bound to membranes and extracted lipopolysaccharide of *Escherichia coli. Biochemistry;* **22**:2002-2007.

Creeger E.S., Schulte T. and Rothfield L.I., (1984). Regulation of

membrane glycosyltransferases by the *sfrB* and *rfaH* genes of *Escherichia coli* and *Salmonella typhimurium*. *J. Biol. Chem.*, **259**:3064-3069.

Creeger, E.S., and Rothfield, L.I., (1982). Cloning genes for bacterial glycosyltranferases. *Methods in Enzymology,* **83**:326-331.

Creeger, E.S., Chen, J.F., and Rothfield, L.I., (1979). Cloning of genes for Bacterial Glycosyltransferases II. Selection of a hybrid plasmid carrying the *rfaH* gene. *J. Biol. Chem.,* **254**:811-815.

Danilov L.L., Maltsev S.G., Shibaev V.N., and Kochetkov N.K., (1981). Synthesis of polyprenyl pyrophosphate sugars from unprotected mono- and oligosaccharide phosphates. *Carbohydr. Res.,* **88**:203-211.

Danilov L.L., Trotzky M.F., Utkina N.S., Shibaev V.N., and Kochetkov N.K., (1980). Synthesis of oligosaccharide phosphates: fragments of the biosynthesis intermediates of *Salmonella* O-specific polysaccharides. *Carbohydr. Res.*, **87**:141-146.

Davis D.B. and Mingoli E.S., (1950). Mutants of *Escherichia coli* requiring methionine or vitamin B_{12}. *J. Bacteriol.,* **60**:17.

Davison J., Heuterspreute M., Merchez M., and Brunel F., (1984). Vectors with retriction-site banks. 1. pJRD158, a 3903bp plasmid containing 28 unique cloning sites. *Gene,* **28**:311-318.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C., (1978). A model of evolutionary changes in proteins. *Atlas of Protein Sequence and Structure.* **5, Supp. 3,** 345-352

De Jongh-Leuvenick J., Vreede, R.W., Marcelis J.H., de Vos M., and Verhoef J., (1985). Detection of antibodies against lipopolysaccharides of *Escherichia coli* and *Salmonella* R and S strains by immunoblotting. *Infect. Immun.,* **50**:716-720.

Dente L., Cesareni G., and Cortese R., (1984). pEMBL: A new family of single stranded plasmids. *Nucl. Acids. Res.,* **11**:1645-1655.

Dougan G., and Kehoe M., (1984). The minicell system as a method for studying expression from plasmid DNA. *Methods in Microbiol.,* **17**:233-258.

Eliopoulos E.E., Geddes A.J., Brett M., Pappin D.J.C., and Findlay J.B.C., (1982). A structural model for the chromophore-binding domain of ovine rhodopsin. *Int. J. Biol. Macromol.,* **4**:263.

Ferro-Luzzi Ames G., (1986). Bacterial periplasmic transport systems: Structure, mechanism and evolution. *Annu. Rev. Biochem.,* **55**: 397.

Ferro-Luzzi Ames, G., and Higgins, C.F., (1983). The organization, mechanism of action, and evolution of periplasmic

transport systems. *Trends in Biochem. Sci.,* **8**:97-100.

Frazer A.C., and Curtis R. III., (1975). Production, properties and utility of bacterial minicells. *Curr. Top. Microbiol.,***69**:1-84.

Funahara, Y., and Nikaido, H., (1980). Asymmetrical localization of lipopolysaccharides on the outer membrane of *Salmonella typhimurium. J. Bacteriol.,* **141**:1463-1465.

Galanos C., and Lüderitz O., (1975). Electrodialysis of lipopolysaccharides and their conversion to uniform salt forms. *Eur. J. Biochem.,* **54**:603-610.

Garger S.J., Griffith O.M., Grill L.K., (1983). Rapid purification of plasmid DNA by a single centrifucation in a two step caesium chloride-ethidium bromide gradient. *Biochem. Biophys. Res. Commun.,* **177**:835-842.

Germanier R., and Furer R., (1975). Isolation and characterization of *galE* mutant Ty21a of *Salmonella typhi*: A candidate strain for a live, oral typhoid vaccine. *J. Infect. Dis.,* **131**:553-558.

Ghuysen J.M., Shockmann G.D., (1973). Biosynthesis of peptidoglycan. In: Leive L., (Ed) "*Bacterial Membranes and Walls*",(Microbiology Series I). pp37-130. Dekker, New York.

Glaser, L., Zarkowsky, H., and Ward, L., (1972). Formation of TDP-L-Rhamnose from TDP-D-Glucose. *Methods in Enzymology,* **8**:446-454.

Goad W.B., and Kanehisa M., (1982). Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucl. Acids Res.,* **10**:247-263.

Gold L., Pribnow D., Schneider T., Shinedling S., Singer B.S., and Stormo G., (1981). Translation initiation in prokaryotes. *Annu. Rev. Microbiol.,* **35**:365-403.

Goldman R., and Strominger J.L., (1972). Purification and properties of C_{55}-isoprenyl-pyrophosphate phosphatase from *Micrococcus lysodeikticus. J. Biol. Chem.,* **247**:5116-5122.

Goldmann R.C. and Leive L., (1980). Heterogeneity of antigenic side-chain length in lipopolysaccharide from *Escherichia coli* 0111 and *Salmonella typhimurium* LT2. *Eur. J. Biochem.,* **107**:145-153.

Gonzalez-Porque P., and Strominger J.L., (1972a). The Biosynthesis of 3,6,-dideoxy sugars: Formation of CDP-4-keto-3,6-dideoxyglucose from CDP-glucose. *Methods in Enzymology,* **18**:461-470.

Gonzalez-Porque P., and Strominger J.L., (1972c). Introduction of the 3-deoxy group in 3,6-dideoxyhexoses; A novel cofactor function

for pyridoxamine 5' phosphate. *Proc. Natl. Acad. Sci. USA.*, **69**:1625-1628.

Gonzalez-Porque P., and Strominger J.L., (1972b). Enzymatic synthesis of cytidine diphosphate-3,6-dideoxyhexoses. VI. Purification to homogeneity and some properties of cytidine diphosphate D-glucose oxidoreductase, enzymes E_1 and E_3. *J. Biol. Chem.*, **247**: 6748-6756.

Grossman N., Joiner K.A., Frank M.M., and Leive L., (1986). C3b binding, but not its breakdown, is affected by the structure of the O-antigen polysaccharide in lipopolysaccharide from *Salmonella*. *J. Immunol.* **136**:2208-2215.

Grossman L., and Leive L., (1984). Complement activation via the alternative pathway by purified *Salmonella* lipopolysaccharide is affected by its structure but not its O-antigen length. *J. Immunol.*, **132**:376-385.

Grunstein M., and Hogness D.S., (1975). Colony hybridization: a method for the isolation of cloned DNAs that contains specific genes. *Proc. Natl. Acad. Sci. USA.*, **72**:3961-3965.

Hanahan D., (1983). Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* ***166***:557-580.

Heath E.C., Mayer R.M., Edström R.D. and Beaudreau C.A.,

(1966). Structure and biosynthesis of the cell-wall lipopolysaccharide of *Escherichia coli*. *Ann. N.Y. Acad. Sci.,* **133**:315-333.

Hellerqvist C.G., Lindberg B., Svensson S., Holme T., and Lindberg A.A., (1969). Structural studies on the O-specific side chains of the cell wall lipopolysaccharides from *Salmonella typhi* and *S. enteriditis*. *Acta Chem. Scand.,* **23**:1588-1592.

Hemming F.W., (1970). Polyprenols. In: Goodwin T.W., (Ed), "*Biochemistry Society Symposia*" **29**:105-117. Academic Press London.

Heuterspreute M., and Davison J., (1984). Restriction site bank vectors. II. DNA sequence analysis of plasmid pJRD158. *DNA.,* **3**:259-268.

Hitchcock P.J., Lieve, I., Mäkelä P.H., Rietschel E.T., Strittmatter W and Morrison D.C., (1986). Lipopolysaccharide nomenclature - past, present and future. *J. Bacteriol.,* **166**:699-705.

Hitchcock, P.J., and Brown, T.M., (1983). Morphological heterogeneity among *Salmonella* lipopolysaccharide chemotypes in silver stained polyacrylamide gels. *J. Bacteriol.,* **154**:269-277.

Hollifield W.G.Jr., and Neilands J.B., (1978). Ferric enterobactin transport system in *Escherichia coli*. Extraction, assay, and specificity of the outer membrane receptor. *Biochemistry* **17**:1922-1928.

Hong G.F., (1982). A systematic sequencing strategy. *J. Mol. Biol.,* **158**:539-549.

Jann B., Reske K., and Jann K., (1975). Heterogeneity of lipopolysaccharides. Analysis of polysaccharide chain length by sodium dodecylsulfate-polyacrylamide gel electrophoresis. *Eur. J. Biochem.,* **60**:239-246.

Jann K., and Westphal O., (1975). Microbial polysaccharides. In: Sela M., (Ed.) "*The antigens*" **3**,:1-125. Academic press, NewYork-San Fransisco-London.

Jann K., and Jann B., (1984). Structure and biosynthesis of O-antigens. In: Rietschel E. Th., (Ed.)*Handbook of Endotoxins* pp138-186

Jann, K., and Jann, B., (1978). Bacterial Polysaccharides. *Methods in Enzymology* **50**:251 - 272.

Johnson E.M., Krauskopf B., and Baron L.S., (1965). Genetic mapping of Vi and somatic antigenic determinants of *Salmonella. J. Bacteriol.,* **90**:302-308.

Johnson E.M., Krauskopf B., and Baron L.S., (1966). Genetic analysis of the *viaA-his* chromosomal region in *Salmonella. J. Bacteriol.,* **92**:1457-1463.

Joiner K.A., Hammer C.H., Brown E.J., Cole R.J., and Frank M.M., (1982a). Studies on the mechanism of bacterial resistance to complement-mediated killing. I. Terminal complement components are deposited and released from *Salmonella minnesota* S218 without causing bacterial death. *J. Exp. Med.,* **155**:797-808.

Joiner K.A., Schmetz M.A., Goldmann R.C., Leive L., Frank M.M., (1984). Mechanism of bacterial resistance to complement-mediated killing: inserted C5b-p correlates with killing for *Escherichia coli* O111/B4 varying in O-antigen capsule and O-polysaccharide coverage of lipid A core oligosaccharide. *Infect. Immun.,* **45**:113-117.

Jones N.C., and Osborn M.J., (1977). Translocation of phospholipids between the outer and inner membranes of *Salmonella typhimurium*. *J. Biol. Chem.,* **252**:7405-7412.

Jousimies H., and Mäkelä P.H., (1974). Genetic analysis of *Salmonella minnesota* R mutants with defects in the biosynthesis of the lipopolysacharide core. *J. Bacteriol.,* **119**:753-759.

Kabsch W., and Sander C., (1983). How good are predictions of protein secondary structure? *FEBS letts.,* **115**:179-182.

Kabsch W., and Sander C., (1984). On the use of sequence homologies to predict protein structure. Identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA.,*

81:1075-1078

Kadam S.K., Rehemtulla A. and Sanderson K.E., (1985b). Cloning of *rfaG*, *B*, *I*, and *J* genes for glycosyltransferase enzymes for synthesis of the lipopolysaccharide core of *Salmonella typhimurium*. *J. Bacteriol.*, **161**:277-284.

Kadam S.K., Peppler M.S., Sanderson K.E., (1985a). Temperature-sensitive mutants of *rfaI* and *rfaJ*, genes for galactosyl transferase I and galactosyl transferase II, for synthesis of lipopolysaccharide in *Salmonella typhimurium*. *Can. J. Microbiol.*, **31**:861.

Kahn M., Kolter R., Thomas C., Figurski D., Meyer R., Remault E., and Helinski D., (1979). Plasmid cloning vehicles derived from plasmid ColE1, R6K and RK2. *Methods Enzymol.*, **68**:268-280.

Kamen R., Kondo M., Römer W., and Weismann C., (1972). Reconstitution Q beta replicase lacking subunit alpha with protein-synthesis-interfering factor i. *Eur. J. Biochem.*, **31**:44-51.

Kanegasaki S., and Wright A., (1970). Mechanism of polymerization of the *Salmonella* O-antigen; utilization of lipid-linked intermediates. *Proc. Natl. Acad. Sci. USA.*, **67**:951-958.

Kanehisa M., Klein P., Greif P., and DeLisi C., (1984). Computer analysis and structure prediction of nucleic acids and proteins. *Nucl.*

Acids Res., **12**:417-428.

Kanehisa M.I., (1982). Los Alamos sequence analysis package for nucleic acids and proteins. *Nucl. Acids Res.,* **10**:183-196.

Karn J., Brenner S., Barnett L., and Cesareni G., (1980). Novel bacteriophage cloning vector. *Proc. Natl. Acad. Sci. USA.,* **77**:5172-5176.

Kauffmann F., (1954). *Enterobacteriaceae.* 2nd Ed, Ejnar Muksgaard, publisher, Copehagen.

Kauffmann F., (1966). *Enterobacteiaceae.* Enjar Munksgaard, Copenhagen.

Kimata K., and Suzuki S., (1966). Studies on cytidine diphosphate glucose pyrophosphorylase and related enzymes of *Azotobacter vinlandii.. J. Biol. Chem.,* **241**:1099.

Kishi K., and Iseki S., (1973b). Genetic analysis of the O antigens in *Salmonella.* II. Heredity of the O antigens 4, 5, and 9 in *Salmonella.. Jpn. J. Genetics,* **48**:133-136.

Kishi K., Iseki S., (1973a). Genetic analysis of the O antigens in *Salmonella* I. Heredity of the blood group-active O-antigens in *Salmonella* groups G, R, and U. *Jpn. J. Genet.,* **48**:89-97.

Klein P., Kanehisa M., and DeLisi C., (1985). The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta,* **815**:468-476.

Knox R.W., Cullen J., and Work E., (1967) An extracellular lipopolysaccharide-phospholipid-protein complex produced by *Escherichia coli* grown under lysine-limiting conditions. *Biochem. J.,* **103**:192-201.

Koch A.L., (1971). The adaptive response of *Escherichia coli* to a feast and famine existence. *Adv. Microb. Physiol.,* **6**:147-217.

Kochibe N., (1970). Immunochemical studies on antigenic substances produced by intergroup hybrids of *Salmonella. J. Biochem. (Tokyo),* **68**:81-89.

Kontrohr T., and Kosics B., (1981). Isolation of adenosine 5'-diphosphate-d-glycero-d-mannoheptose. *J. Biol. Chem.,* **256**:7715-7718.

Koo H-S., and Crothers D.M., (1987). Chemical determinants of DNA bending at adenine-thymine tracts. *Biochemistry,* **26**:3745-3748.

Kornfeld R.H., and Ginsburg V., (1966) Control of synthesis of guanosine 5'-diphosphate-D-mannose and guanosine 5'-diphosphate-L-fucose in bacteria. *Biochim. Biophys. Acta,* **117**:79.

Kozak M., (1983). Comparison of initiation of protein synthesis in prokaryotes, eukaryotes, and organelles. *Microbiol. Revs.*, **47**:1-45.

Kyte J., and Doolittle R.F., (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**:105-132.

Labishinski H., Barnickel G., Bradaczek H., Naumann D., Rietschel E.Th., and Giesbrecht P., (1985). High state of order of isolated bacterial lipopolysaccharide and its possible contribution to the permeation barrier property of the outer membrane. *J. Bacteriol.*, **162**:9-20.

Lederberg E.M., and Cohen S.N., (1974). Transformation of *Salmonella typhimurium* by plasmid deoxyribonucleic acid. *J. Bacteriol.*, **119**:1072-1074.

Lehmann V., Ruprecht E., and Osborn M., (1977). Isolation of mutants conditionally blocked in the biosynthesis of the 3-deoxy-d-manno-octulosonic acid -0 lipid A part of lipopolysacharide derived from *Salmonella typhimurium*. *Eur. J. Biochem.*, **76**:41-49.

Lemaire H-G., and Müller-Hill B., (1986). Nucleotide sequences of the *galE* gene and the *galT* gene in *E. coli*. *Nucl. Acids Res.*, **14**:7705-7711.

Levinthal M., and Nikaido H., (1969). Consequences of deletion

mutations joining two operons of opposite polarity. *J. Mol. Biol.,* **42**:511-520.

Lew H.C., Nikaido H., and Mäkelä P.H., (1978). Biosynthesis of uridine diphosphate *N*-acetylmannosaminurionic acid in *rff* mutants of *Salmonella typhimurium. J. Bacteriol.,* **136**:227-233.

Liang-Takasaki C-J., Mäkelä P.H., Lieve L., (1982). Phagocytosis of bacteria by macrophages: changing the carbohydrate of lipopolysaccharide alters interaction with complement and macrophages. *J. Immunol.,* **128**:1229-1235.

Liang-Takasaki C-J., Mäkelä P.H., Lieve L., (1983). *Salmonellae* activate complement differentially via the alternate pathway depending on the structure of their lipopolysaccharide O-antigen. *J. Immunol.,* **130**:1867-1870.

Lipman D.J., and Pearson W.R., (1985). Rapid and sensitive protein similarity searches. *Science,* **227**:1435-1441.

Losick R., and Robbins P.W., (1967). Mechanism of eta^{15} conversion studied with a bacterial mutant. *J. Mol. Biol.,* **30**:445-456.

Luckey M., Wayne, R., and Neilands J., (1975). *In vitro* competition between ferrichrome and phage for the outer membrane T5 receptor complex of *Escherichia coli. Biochem. Biophys. Res. Commun.,* **64**:687-693.

Lüderitz O., Wesphal O., Staub A.M., and Nikaido H., (1971). Isolation and chemical and immunological characterization of bacterial lipopolysaccharides. In: Weinbaum G. Kadis S., Ajl S.J., (Eds) *Microbial toxins*, **4**;145-233. Academic Press, New York.

Lüderitz O., Staub A.M., and Westphal O., (1966). Immunochemistry of O and R antigens of *Salmonella* and related *Enterobacteriaceae. Bacteriol. Rev.*, **30**:192-225.

Lüderitz, O., Jann, K., Wheat R., (1968). Somatic and capsular antigens of Gram-negative bacteria. In: Florkin M., Stotz E. H., (Eds), *Comprehensive Biochemistry*, **26A**:105-227. Elsevier Publishing company, Amsterdam.

Lugowski C., Romanowska E., Kenne L., and Lindberg B., (1983). Identification of a trisaccharide repeating-unit in the enterobacterial common-antigen. *Carbohydr. Res.*, **118**:173-181.

Lugtenberg B., Meijers J., Peters R., van der Hoek P., and van Alphen L., (1975). Electrophoretic resolution of the major outer membrane proteins of *Escherichia coli* K-12 into four bands. *FEBS letts.*, **58**:254-258.

MacKaness G.B., Blanden R.V., Collins F.M., (1966). Host-parasite relations in mouse typhoid. *J. Exp. Med.*, **124**:573-583.

Mäkelä P.H., (1965). Inheritance of the O-antigens of *Salmonella* groups B and D. *J. Gen. Micrbiol.,* **41**:57-65.

Mäkelä P.H., (1973). Glucosylation of lipopolysaccharide in *Salmonella*: mutants negative for O antigen factor 12/2. *J. Bacteriol.,* **116**:847-856.

Mäkelä P.H., and Mayer H., (1976). Enterobacterial common antigen. *Bacteriol. Rev.,* **40**:591-632.

Mäkelä P.H., and Stocker B.A.D., (1984). Genetics of lipopolysaccharide. In Rietschel E.T., (Ed.) *Handbook of Endotoxin,* **1**:59-137.

Mäkelä P.H., Jahkola M., and Lüderitz O., (1970). A new gene cluster *rfe* concerned with the biosynthesis of *Salmonella* lipopolysaccharide. *J. Gen. Microbiol.,* **60**:91-106.

Mäkelä P.H., (1966). Genetic determination of the O-antigens of *Salmonella* groups B (4,5,12) and C1(6,7). *J. Bacteriol.,* **91**:115-1125.

Mäkelä P.H., Schmidt G., Mayer H., Nikaido H., Whang H.Y., and Noter E., (1976). Enterobacterial common antigen in the *rfb* deletion mutants of *Salmonella typhimurium. J. Bacteriol.,* **127**:1141-1149.

Maniatis T., Fritsch E.F., and Sambrook J., (1982). Molecular

cloning, A laboratory manual. Cold Spring Harbor Laboratories (USA).

Manning P.A., and Achtmann M., (1979). Cell-to-cell interactions in conjugating *Escherichia coli*: The involvement of the cell envelope. In: Inouye M., (Ed) *Bacterial outer membranes* pp409-447. John Wiley and Sons.

Manning P.A., Heuzenroeder M.W., Yeadon J., Leavesley D.I., Reeves, P.R., and Rowley D., (1986). Molecular cloning and expression in *Escherichia coli* K-12 of the O-antigens of the Inaba and Ogawa serotypes of the *Vibrio cholerae* 01 lipopolysaccharides and their potential for vaccine development. *Infect. Immun.*, **53**:272-277.

Marino P.A., Phan K.A., and Osborn M.J., (1985). Energy dependence of lipopolysaccharide translocation in *Salmonella typhimurium*. *J. Biol. Chem.*, **260**:14965-14970.

Markovitz A., (1977). Genetics and regulation of bacterial capsular polysaccharide synthesis and radiation sensitivity. In: Sutherland I., (Ed) *Surface carbohydrates of the prokaryotic cell,* pp415-462. Academic Press Inc. (London), Ltd., London.

Maruyama T., Gojobori T., Aota S-I., and Ikemura T., (1986). Codon usage tabulated from the GenBank genetic sequence data. *Nucl. Acids Res.*, **14**(Suppl):r151-r197. *

Matsuhashi S., and Strominger J.L., (1964). Thymidine

diphosphate 4-acetamido-4,6-dideoxyhexoses. I. Enzymatic synthesis of cytidine diphosphate 3,6-dideoxyhexoses. *J. Biol. Chem.,* **242**:3494-3500.

Matsuhashi, S., (1966). Enzymatic synthesis of cytidine diphosphate-3,6-dideoxyhexoses. II. Reversible, 2-epimerization of cytidine diphosphate paratose. *J. Biol. Chem.,* **241**:4275-4282.

Matsuhashi, S., and Strominger, J.L., (1967). Enzymatic synthesis of cytidine diphosphate-3,6-dideoxyhexoses. IV. Partial resolution of the enzymes required for the reduction of cytidine diphosphate-4-keto-6-deoxy-D-glucose and isolation of a new intermediate. *J. Biol. Chem.,* **242**:3494-3500.

Matsuhashi, S., Matsuhashi, M., Brown, J. G., Strominger, J. L., (1966b). Enzymatic synthesis of cytidine diphosphate-3,6-dideoxyhexoses. III. Cytidine diphosphate D-glucose oxidoreductase. *J. Biol. Chem.,* **241**:4283-4287.

Matthews D.A., Alden R.A., Bolin J.T., Filman D.J., Freer S.T., Hamlin R., Hol. G.J., Kisliuk R.L., Pastore E.J., Plante L.T., Xuong N-H., and Kraut J., (1978). Dihydrofolate reductase from *Lactobacillus casei.* X-ray structure of the enzyme-methotrexate-NADPH complex. *J. Biol. Chem.,* **253**:6946-6954.

Matthews D.A., Alden R.A., Bolin J.T., Filman D.J., Freer S.T., Hamlin R., Xuong N-H., and Kraut J., Poe M., Williams M., and

Hoogsteen K., (1977). Dihydrofolate reductase: X-ray structure of the binary complex with methotrexate. *Science,* **197**:452-455.

Maurer R., Osmond B.C., Shekhtman E., Wong A., Botstein D., (1984). Functional interchangeability of DNA replication genes in *Salmonella typhimurium* and *Escherichia coli* demonstrated by a general complementation procedure. *Genetics,* **123**:108.

Mayer H., and Schmidt G., (1979). Chemistry and biology of the enterobacterial common antigen (ECA). *Curr. top. Microbiol. Immunol.,* **85**:99-153.

Mayer R.M., and Ginsburg V., (1965). Purification and properties of cytidine diphosphate d-glucose pyrophosphorylase from *Salmonella paratyphi* A. *J. Biol. Chem.,* **240**:1900.

Melchior D.L., and Stein J.M., (1976). Thermotropic transitions in biomembranes. *Annu. Rev. Biophys. Bioeng.,* **6**:205-238.

Melo A. and Glaser L., (1965). The nucleotide specificity and feedback control of thymidine diphosphate -glucose pyrophosphorylase. *J. Biol. Chem.,* **240**:398-405

Morona R., and Reeves P., (1982). The *tolC* locus of *Escherichia coli* affects the expression of three major outer membrane proteins. *J. Bacteriol.,* **150**:1016-1023.

Morrison D.C. and Kline L.F., (1977). Activation of the classical and properdin pathways of complement by bacterial lipopolysaccharides (LPS). *J. Immunology.,* **118**:362-368.

Mühlradt P.F., Menzel J., Golecki J.R., and Speth V., (1973). Outer membrane of *Salmonella* sites of export of newly synthesized lipopolysaccharide an the bacterial surface. *Eur. J. Biochem.,* **35**:471-481.

Mühlradt P.F.., and Golecki J.R., (1975). Asymmetrical distribution and artifactual reorientation of lipopolysaccharide in the outer membrane of *Salmonella typhimurium. Eur. J. Biochem.,* **51**:343-352.

Muto A., and Osawa S., (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. U.S.A.,* **84**:166-169.

Nakae T., (1971). Multiple molecular forms of uridine-diphosphate-glucose pyrophosphorylase from *Salmonella typhimurium.* III. Interconversion between various forms. *J. Biol. Chem.,* **246**:4404-4411.

Nakae T., and Nikaido H., (1971a). Multiple molecular forms of uridine-diphosphate-glucose pyrophosphorylase from *Salmonella typhimurium.* I. Catalytic properties of various forms. *J. Biol. Chem.,* **246**:4386-4396.

Nakae T., and Nikaido H., (1971b). Multiple molecular forms of uridine-diphosphate-glucose pyrophosphorylase from *Salmonella typhumurium*. II. Genetic determination of multiple forms. *J. Biol. Chem.,* **246**:4397-4403.

Naumann D., Schultz C., Born J., Labishinski H., Brandenburg K., vn Busse G., Brade H., and Sydel U., (1987). Investigations into the polymorphism of lipidA from lipopolysaccharides of *Escherichia coli* and *Salmonella minnesota* by Fourier-transform infrared spectroscopy. *Eur. J. Biochem.,* **164**:159-169.

Nikaido H., (1970). Structure of cell wall lipopolysaccharide from *Salmonella typhimurium*. Further studies on the linkage between O-side chains and R-core. *Eur. J. Biochem.,* **15**:57-62.

Nikaido H., (1973). Biosynthesis and assembly of lipopolysaccharide and the outer membrane layer of Gram-negative cell wall. In: Leive L (Ed), *Bacterial Membranes and Walls* (Microbiology series 1), **I**:131-208. Dekker, New York.

Nikaido H., Levinthal M., Nikaido K., and Nakune K., (1967). Extended deletions in the histidine-rough-B region of the *Salmonella* chromosome. *Proc Natl. Acad. Sci. USA.,* **57**:1825-1832.

Nikaido H., Nikaido K., Nakae T., and Mäkelä P.H., (1971). Glucosylation of lipopolysaccharide in *Salmonella*: biosynthesis of O

antigenic factor 12_2 .I. Overall reaction. *J. Biol. Chem.*, **246**:3902-3911.

Nikaido K., and Nikaido H., (1971). Glucosylation of lipopolysaccharide in *Salmonalla*: biosynthesis of O antigen factor 12_2. II. Structure of the lipid intermediate. *J. Biol. Chem.*, **246**:3912-3919.

Nikaido N., and Vaara, M., (1985). Molecular basis of bacterial outer membrane permeability. *Microbiol. Revs.*, **49**:1-32.

Nikaido H., (1965). Biosynthesis of cell wall polysaccharide in mutant strains of *Salmonella*. III. Transfer of L-Rhamnose and D-Galactose. *Biochemistry,* **4**:1550-1561.

Nikaido H., Nikaido K., Mäkelä P.H., (1966). Genetic determination of enzymes synthesizing O-specific sugars of *Salmonella* lipopolysaccharides. *J. Bacteriol.*, **91**:1126-1135.

Nikaido N. and Nikaido K., (1966). Enzymatic synthesis of cytidine-diphosphate 3,6-dideoxyhexoses in *Salmonella*. *J. Biol. Chem.*, **241**:1376-1385.

Nishijima M., Nakaike S., Tamori Y., and Nojima S., (1977). Detergent-phospholipase A of *Escherichia coli* K-12. *Eur. J. Biochem.*, **73**:115-124.

Normore W.M., (1973) Guanine-plus-cytosine (GC) composition of

the DNA of bacteria, fungi algae and protozoa. In Laskin A.I., and Lechevalier H.A., (Eds.) *Microbial Composition* (Handbook of Microbiology Vol II) pp585-740. CRC press Cleveland Ohio USA.

Norrander J., Kempe T., and Messing J., (1983). Construction of improved M13 vectors using oligonucleotide-directed mutagenesis. *Gene,* **26**:101-106.

Oliver D., (1985). Protein secretion in *Escherichia coli. Annu. Rev. Microbiol.,* **39**:615-648.

Osborn M.J., and Tze-Yuen R.Y., (1968). Biosynthesis of bacterial lipopolysaccharide. II. Enzymatic formation of the first intermediate in biosynthesis of the O-antigen of *Salmonella typhimurium. J. Biol. Chem.,* **243**:5145-5152.

Osborn M.J., Rick P.D., and Rassmussen N.S., (1980). Mechanism of assembly of the outer membrane of *Salmonella typhimurium.* Translocation and integration of an incomplete mutant lipid A into the outer membrane. *J. Biol. Chem.,* **255**:4246-4251.

Osborn M.J., Gander J.E., and Parisi E., (1972b). Mechanism of assembly of the outer membrane of *Salmonella typhimurium.* Site of synthesis of lipopolysaccharide. *J. Biol. Chem.,* **247**:3968-3972.

Osborne M.J., and Weiner I.M., (1968). Biosynthesis of a bacterial lipopolysaccharide. VI. Mechanism of incorporation of the O-

antigen of *Salmonella typhimurium*. *J. Biol. Chem.*, **243**:2631-2639.

Pape H., and Strominger J.L., (1969). Enzymatic synthesis of cytidine diphosphate 3,6-dideoxyhexoses. V. Partial purification of the two protein components required for introduction of the 3-deoxygroup. *J. Biol. Chem.*, **244**:3598-3604.

Palva T., and Mäkelä P.H., (1980). Lipopolysaccharide heterogeneity in *Salmonella typhimurium* analysed by sodium dodecylsulfate-polyacrylamide gel electrophoresis. *Eur. J. Biochem.*, **107**:137-143.

Ponnambalam S., Webster C., Bingham A., and Busby S., (1986). Transcription initiation at the *Escherichia coli* galactose operon promoters in the abscence of the normal -35 region sequences. *J. Biol. Chem.*, **261**:16043-16048.

Pluschke G., and Achtman M., (1984). Degree of antibody-independent activation of the classical complement pathway by K1. *Escherichia coli* differs with O-antigen type and correlates with virulence of menigitis in newborns. *Infect. Immun.*, **43**:684-692.

Postma P.W., and Langeler J.W., (1985). Phosphoenolpyruvate: Carbohydrate phosphotransferase system of bacteria. *Microbiol. Rev.*, **49**:232-269.

Poutska A., Rackwitz H., Frishauf A., Hohn B., and Lehrbach

H., (1984). Selective isolation of cosmid clones by homologous recombination in *Escherichia coli. Proc. Natl. Acad. Sci. USA.,* **81**:4129-4133.

Prehm P., Strim S., Jann B., and Jann K., (1975). Cell-wall lipopolysaccharides of ampicillin resistant mutants of *Escherichia coli* K12. *Eur. J. Biochem.,* **66**:369-377.

Quigley N.B., and Reeves P.R., (1987). Chloramphenicol resistance cloning vector based on pUC9. *Plasmid,* **17**:672-675.

Raetz C.R.H., Purcell S., Meyer M.V., Qureshi, N., and Takayama, K., (1985). Isolation and characterization of eight Lipid A precursors from a 3-Deoxy-d-manno-occtulosonic acid-deficient mutant of *Salmonella typhimurium. J. Biol. Chem.,* **260**:16080-16088.

Reisner A.H., and Bucholtz C.A., (1986). The MTX package of computer programmes for comparison of sequences of nucleotides and amino acid residues. *Nucl. Acids Res.,* **14**:233-238.

Reynolds P.R., Mottur G.P., and Bradbeer.C., (1980). Transport of vitamin B_{12} in *Escherichia coli.* Some observations on the roles of the gene products of *btuC* and *tonB*. *J. Biol. Chem.,* **255**:4313-4319.

Rick, P.D., Fung L. W.-M., Ho C., and Osborn M.J., (1977). Lipid A mutants of *Salmonella typhimurium.* Purification and characterization of a lipidA precursor produced by a mutant in 3-

deoxy-D-mannooctulosonate-8-phosphate synthetase. *J. Biol. Chem.*, **252**:490-4912.

Rietschel E. Th., Wollenweber H.W., Brade H., Zähringer U., Lindner B., Seydel U., Brdaczek H., Barnickel G., Labishinski H., and Giesbrecht P., (1984). Structure and conformation of the lipid A component of lipopolysaccharides. In: Rietschel. E.Th., (Ed.)*Handbook of endotoxins*, **1**;187-220, Elsevier, Amsterdam.

Rigby P.W.J., Dieckmann, M., Rodes C., and Berg P., (1977). Labelling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *J. Mol. Biol.*, **113**:237-251.

Robbins P.W., and Wright A., (1971). Biosynthesis of O-antigens. In: Weinbaum G., Kadis S., and Ajl S.J., (Eds) *Microbial toxins*, **4**:351-368 Academic Press New York.

Rossmann M.G., Liljas A., Brändén C.I., and Banaszak L.J., (1975). Evolutionary and Structural relationships among dehydrogenases. In: Boyer P.D., (Ed.) *The Enzymes*, pp61-102. (Academic Press, New York ed. 3).

Rubenstein P.A. and Strominger J.L., (1974a). Enzymatic synthesis of cytidine diphosphate-3,6-dideoxyhexoses. VII. Mechanistic roles of enzyme E_1 and pyridoxamine 5'-phosphate in the formation of cytidine diphosphate-4-keto-3,6-dideoxy D-glucose. *J. Biol. Chem.*, **249**:3776-3781.

Rubenstein P.A., and Strominger J.L., (1974b). Enzymatic synthesis of cytidine diphosphate 3,6-dideoxyhexoses. VIII. Studies on the properties of E_3 and its role in the formation of cytidine diphosphate-4-keto 3,6-dideoxyglucose. *J. Biol. Chem.,* **249**:3782-3788.

Rubenstein P.A., and Strominger J.L., (1974c). Enzymatic synthesis of cytidine diphosphate 3,6-dideoxyhexoses. IX. Purification and properties of cytidine diphosphate-D-glucose pyrophosphorylase, from *Pasteurella pseudotuberculosos* type V. *J. Biol. Chem.,* **249**:3789-3796.

Rudén U., and Mäkelä P.H., (1974). O-acetylation and glucosylation of lipopolysaccharide in hybrids between *Salmonella* groups B and C2. *Eur. J. Biochem.,* **48**:11-20.

Samuelsson K., Lindberg K., Brubaker R.R., (1974). Structure of O-specific side chains of lipopolysaccharides from *Yersinia pseudotuberculosis. J. Bacteriol.,* **117**:1010-1016.

Sanderson K.E., and Roth J.R., (1983). Linkage map of *Salmonella typhimurium* edition VI. *Microbiol. Rev.,* **47**:410-453.

Sanger F., Coulson A.R., Barrell B.G., Smith A.J.H., and Roe B.A., (1980). Cloning in single stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.,* **143**:161-178.

Sanger F., Coulson A.R., Hong G.F., Hill D.F., and Peterson G.B., (1982). Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.,* **162**:729-773.

Sanger F., Nicklen S., and Coulson A.R., (1977). DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci. USA.,* **74**:5463-5467.

Sarvas M., (1967). Inheritance of *Salmonella* T1 antigen. *Ann. Med. Exp. Biol. Fenn.,* **45**:447-472.

Saxen H., (1984). Mechanism of the protective action of anti-*Salmonella* IgM in experimental mouse salmonellosis. *J. Gen. Microbiol.,* **130**:2277-2283.

Senior A.E., (1985). The proton-ATPase of *Escherichia coli. Curr. Top. Memb. Trans.,* **23**:135-151.

Shepherd J.C.W., (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and it possible evolutionary justification. *Proc. Natl. Acad. Sci. U.S.A.,* **78**:1596-1600.

Sher M., Lennarz W.J., Sweeley C.C., (1968). The biosynthesis of mannosyl-1-phosphoryl-polyisoprenol in *Micrococcus lysodeikticus* and its role in mannan synthesis. *Proc. Natl. Acad Sci. USA.,* **59**:1313-

1320.

Shibaev V.N., (1978). Biosynthesis of *Salmonella* O antigenic polysaccharides: specificity of glycosyl transferases. *Pure Appl. Chem.*, **50**:1421-1436.

Shibaev V.N., (1986). Biosynthesis of bacterial polysaccharide chains composed of repeating units. *Adv. Carboh. Chem. Biochem.*, **44**:277-339.

Shibaev V.N., Danilov L.L., Druzhinina T.N., Gogolashwili L.M., Maltsev S.D., and Kochetkov N.K., (1982). Enzymatic synthesis of *Salmonella* O-specific polysaccharide analogs from modified polyprenol sugar acceptors. *FEBS Letts.*, **139**:177-180.

Shine J., and Dalgarno L., (1974). The 3'-terminal sequence of *E. coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad Sci. U.S.A.*, **71**:1342-1346.

Sidorczyk Z., Zähringer U., Reitschel E. Th., (1983). Chemical structure of the lipid A component of the lipopolysaccharide from a *Proteus mirabilis* Re-mutant. *Eur. J. Biochem.*, **137**:15-22.

Siewert G., and Strominger J.L., (1967). Bacitracin: an inhibitor of the dephosphorylation of lipid pyrophosphate, an intermediate in biosynthesis of the peptidoglycan of bacterial cell walls. *Proc. Natl. Acad. Sci. USA.*, **57**:767-773.

Skurray R.A., Hancock R.E.W., and Reeves P., (1974). Con⁻ mutants: Class of mutants in *Escherichia coli* K12 lacking a major cell wall protein and defective in conjugation and adsorption of a bacteriophage. *J. Bacteriol.*, **119**:726-735.

Southern E., (1975). Detection of specific sequence among DNA fragment separated by gel electrophoresis. *J. Mol. Biol.*, **98**:503-517.

Staden R., (1980). A new computer method for the storage and manipulation of DNA rel reading data. *Nucl. Acids Res.*, **8**:3673-3694.

Staden R., (1982). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucl. Acids Res.*, **10**:4731-4751.

Staden R., (1984). Computer methods to locate signals in nucleic acid sequences. *Nucl. Acids Res.*, **12**:505-519.

Staden R., (1986). The current status and portability of our sequence handling software. *Nucl. Acids Res.*, **14**:217-231.

Stocker B.A.D, and Mäkelä P.H., (1986). Genetic determination of bacterial virulence with special reference to *Salmonella*. *Curr. Top. Microbiol. Immunol.*, **124**:149-173.

Stocker B.A.D., and Mäkelä P.H., (1980). Bacterial surface

genetics. *Royal Soc. of London Proceed. series B.*, **202**:5-30.

Stone K.J., and Strominger J.L., (1971). Mechanism of action of bacitracin: complexation with metal ion and C_{55}-isoprenyl pyrophophate. *Proc. Natl. Acad. Sci. USA.*, **68**:3223-3227.

Strain S.M., Fesik S.W., Armitage I.M., (1983b). Structure and metal binding propoerties of lipopolysaccharides from heptoseless mutants of *Escherichia coli* studied by ^{14}C and ^{31}P nucelar magnetic resonance. *J. Biol. Chem.*, **258**:13466-13477.

Strain S.M., Fesik S.W., Armitage I.M., (1983a). Characterization of lipopolysaccharide from a heptoseless mutants of *Escherichia coli* by carbon 13 nuclear magnetic resonsance. *J. Biol. Chem.*, **258**:2906-2910.

Sutherland I.W., (1977). Bacterial exopolysaccharides - their nature and production. In: Sutherland I., (Ed) *Surface Carbohydrates of the Prokaryotic Cell*, pp27-96 London Academic Press.

Swanson R., (1984). A unifying concept for the amino acid code. *Bull. Math. Biol.*, **46**:187-203.

Takayama K., Qureshi N., Mascagni P., (1983). Complete structure of lipid A obtained from the lipopolysaccharide of the heptoseless mutant of *Salmonella typhimurium*. *J. Biol. Chem.*, **258**:12801-12803.

Tacken A., Rietschel E.T., and Brade H., (1986)., Methylation analysis of the heptose/3-deoxy-d-*manno*-2-octulosomic acid region (inner core) of lipopolysaccharide from *Salmonella minnesota* rough mutants. *Carbohydr. Res.* **149**:279-291.

Takeshita M., Mäkelä P.H., (1971). Glucosylation of lipopolysaccharide in *Salmonella*: biosynthesis of O antigen factor 12_2 .III. The presence of 12_2 determinants in haptenic polysaccharides. *J. Biol. Chem.,* **246**:3920-3927.

Taylor W.R., and Thornton J.M., (1983). Prediction of super secondary structure in proteins. *Nature* **301**:340-342.

Tenner A.J., Ziccardi R.J., and Cooper N.R., (1984). Antibody-independent C1 activation by *E. coli. J. Immunol.,* **133**:886-891.

Trifonov E.N. and Brendel V., (1986). GNOMIC: A dictionary of genetic codes. VCH (Israel)

Torres-Cabassa A.S., and Gottesman S., (1987). Capsule synthesis in *Escherichia coli* K-12 is regulated by proteolysis. *J. Bacteriol.,* **169**:981-989.

Troy F.A. Frerman F.E., Heath E.C., (1971). The biosynthesis of capsular polysaccharide in *Aerobacter aerogenes*. J. Biol. Chem., **246**:118-133.

Troy F.A., Visay I.K., Tesche N., (1975). Role of undecaprenol phosphate in synthesis of polymers containing sialic acid in *Escherichia coli. J. Biol. Chem.,* **250**:156-163.

Ubben D., and Schmitt R., (1986). Tn1721 derivatives for transposon mutagenesis, restriction mapping and nucleotide sequence analysis. *Gene,* **41**:145-152.

Valtonen V.V., (1970). Mouse virulence of *Salmonella* strains: effect of different smooth-type O side chains. *J. Gen. Microbiol.,* **64**:255-268.

Valtonen V.V., (1977). Role of phagocytosis in mouse virulence of *Salmonella typhimurium* recombinants with O-antgen 6,7 or 4,12. *Infect. Immun.,* **18**:574-578.

Valtonen V.V., Sarvas M., and Mäkelä P.H., (1976b). T2 lipopolysacharide antigen of *Salmonella*: Genetic determination of *Salmonella typhimurium. J. Bacteriol.,* **139**:664-667.

Valtonen V.V., Sarvas M., and Mäkelä P.H., (1976a). T2 lipopolysaccharide antigen of *Salmonella* genetic determination of T2 and properties of T2, T2S and T2SR forms. *Infect. Immun.,* **13**:1647-1653.

Van Alphen L., Lugtenberg B., Rietschel, E. Th., and Mombers C., (1979). Architecture of the outer membrane of *Escherichia coli*

K12. Phase transitions of the bacteriophage K3 receptor complex. *Eur. J. Biochem.*, **101**:571-579.

Verma N.K., Quigley N.B., and Reeves P.R., (1988). Genetic basis of O-antigen variation in *Salmonella*: identification of the non-homologous regions of the *rfb* gene cluster of 3 strains with related O-antigens. *J. Bacteriol.*, (in press, January).

Walenga R.W., and Osborn M.J., (1980). Biosynthesis of Lipid A. Formation of acyl-deficient lipopolysaccharides in *Salmonella typhimurium* and *Escherichia coli*. *J. Biol. Chem.*, **255**:4257-4263.

Walsh C., (1979). Enzymatic reaction mechanisms. pp345-351 and 550-553. W.H. Freeman and Co (San Fransisco)

Wang S.-F., and Gabriel O., (1969). Biological mechanisms involved in the formation of deoxy sugars. V. Isolation and crystallization of thymidine diphosphate-d-glucose oxidoreductase from *Escherichia coli* B. *J. Biol. Chem.*, **244**:3430.

Wayne R., and Neilands J., (1975). Evidence for common binding sites for ferrichrome compounds and bacteriophage phi80 in the cell envelope of *Escherichia coli*. *J. Bacteriol.*, **121**:497-503.

Weiner I.M. Higushi T., Osborn M.J., Rothfield L., Saltmarsh-Andrew Horecker B.L., (1965). Biosynthesis of bacterial lipopolysaccharide. V. Lipid-linked intermediates in the biosynthesis

of the O-antigen groups of *Salmonella typhimurium*. *Proc. Natl. Acad. Sci. USA.*, **54**:228-235.

Weisgerber C., and Jann K., (1982). Glucosyl diphospho-undecaprenol, the mannose acceptor in the synthesis of the 09 antigen of *Escherichia coli*. *Eur. J. Biochem.*, **127**:165-168.

Wilbur W.J., and Lipman D.J., (1983). Rapid similarity searches of nucleic acid and protein databanks. *Proc Natl. Acad. Sci. U.S.A.*, **80**:726-730.

Wollin R., Stocker B.A.D., and Lindberg A.A., (1987). Lysogenic Conversion of *Salmonella typhimurium* bacteriophages A3 and A4 consists of O-acetylation of rhamnose of the repeating unit of the O-antigenic polysaccharide chain. *J. Bacteriol.*, **169**:1003-1009.

Wright A., Dankert M., Fennessey P., and Robbins P.W., (1967). Characterization of a polyisoprenoid compound functional in O-antigen biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.*, **57**:1798-1803.

Yanisch-Perron C., Vieira J., and Messing J., (1985). Improved M13 phage cloning vectors and host strains: Nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene*, **33**:103-119.

Young R., Way J., Way S., Yin J., and Syvanen?(1979). Transposition mutagenesis of bacteriophage lambda. A new gene affecting cell lysis. *J. Mol. Biol.*, 132:307-322.

Yuasa R., Nakae K., and Nikaido H., (1970). Structure of cell wall lipopolysaccharide from *Salmonella typhimurium*. Structure of lipopolysaccharide from a semirough mutant. *Eur. J. Biochem.,* **15**:63-71.

Yuasa R., Levinthal M., and Nikaido N., (1969). Biosynthesis of cell wall lipopolysaccharide in mutants of *Salmonella*. V. A mutant of *Salmonella typhimurium* defective in the synthesis of cytidine diphospho-abequose. *J. Bacteriol.,* **100**:433-444.

Zarkowsky H., Lipkin E., and Glaser L., (1970). The mechanism of 6-deoxyhexose synthesis: The relation of pyridine nucleotide to the structure of the deoxythymidine diphosphate-glucose oxidoreductase. *J. Biol. Chem.,* **245**:6599-6606.