An Archaebacterial Ribosomal Protein Gene Cluster

by

LAWRENCE CHARLES SHIMMIN

B.Sc., The University of Victoria, 1981

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF BIOCHEMISTRY

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April, 1990

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of BIOCHEMISTRY

The University of British Columbia
Vancouver, Canada

Date 900418

Abstract

The eubacteria, archaebacteria and eucaryota evolved from a common ancestral state, the progenote, approximately 4,000 million years ago. The archaebacteria flourish in extreme environments, exhibiting unusual macromolecular structures and metabolism of which much has recently been elucidated. Less, however, is known of the genetics of archaebacteria. In order to investigate gene structure, organization, regulation and evolution in the archaebacteria a gene cluster encoding the ribosomal proteins of the GTPase domain was cloned from the extremely halophilic archaebacterium *Halobacterium cutirubrum*, characterized and compared with the homologous genes and proteins from eubacteria and eucaryota.

A clone containing a 5146 basepair insert of genomic *Halobacterium cutirubrum* NRCC 34001 DNA encoding the GTPase domain ribosomal proteins was characterized and discovered to retain the identical gene order (i.e. L11e, L1e, L10e and L12e) as the homologous *Escherichia coli* genes and in addition two transcribed upstream open reading frames encoding the potential proteins ORF, of unknown function and NAB, bearing sequence similarity to nucleic acid binding proteins.

The predominant transcripts are monocistronic L11e and tricistronic L1e - L10e - L12e transcripts; monocistronic NAB and bicistronic NAB - L11e transcripts are present at reduced levels and the ORF is present as a very rare transcript. Common elements upstream of the transcription initiation sites include the motif TTCGA ... 4-15 bp ... TTAA ... 20-26 bp ... A or G transcription start. The NAB and some of the ORF transcripts are divergently transcribed from a single TTAA promotor element. The NAB and some of the ORF transcripts initiate 1 nucleotide before the coding region; the L11e monocistronic transcript initiates precisely at the first A of the initiator methionine ATG codon. The L1e - L10e - L12e tricistronic transcript has a 75 nucleotide leader that is probably involved in the autogenous regulation of the transcript at the translational level by the L1e protein. Termination of transcription occurs, with a single exception, within T tracts after GC rich regions. Although classic Shine-Dalgarno (eubacterial) type ribosome binding sites are present upstream of the L1e and L10e genes, the mechanism of translation initiation for transcripts with nil or negligible 5' leaders remains to be elucidated.

Alignments between the deduced amino acid sequences of the L11e, L1e, L10e and L12e ribosomal proteins and other available homologous proteins of archaebacteria, eubacteria and eucaryota have been made and show that the L11e, L1e and L10e proteins are colinear whereas the L12e protein has suffered a rearrangement through what appears to be gene fusion events. The L11e proteins exhibit (i) sequence conservation in the region interacting with release factor 1, (ii) conserved proline residues (probably contributing to the elongated shape of the molecule) and (iii) sites of methylation in Eco L11 are not conserved in the archaebacterial L11e proteins. The L1e proteins have regions of very high sequence similarity near the center and carboxy termini of the proteins but the relationships between protein structure and function remain unknown. Intraspecies comparisons between L10e and L12e sequences indicate the archaebacterial and eucaryotic L10e proteins contain a partial copy of the L12e protein fused to their carboxy terminus. In the eubacteria most of this fusion has been removed by a carboxy terminal deletion. Within the L12e derived region a 26 amino acid long internal modular sequence reiterated thrice in the archaebacterial L10e, twice in the eucaryotic L10e and once in the eubacterial L10e was discovered. This modular sequence also appears to be present in single copy in all L12e proteins and may play a role in L12e dimerization, L10e - L12e complex formation and the function of L10e - L12e complex in translation. From these sequence comparisons a model depicting the evolutionary progression gene cluster and proteins from the primordial state to the contemporary archaebacterial, eucaryotic and eubacterial states is presented.

Table Of Contents

Page

Abstract ii

Table of Contents iv

List of Tables vi

List of Figures vii

Abbreviations viii

Acknowledgements ix

Dedication x

Part 1: Introduction 1

1.1 Early Evolution and the Archaebacteria 1

1.2 Halophilic Archaebacteria 4

1.3 Organization of Ribosome Components 5

1.4 GTPase Domain of *Escherichia coli* 6

1.5 Ribosomal Proteins of *Halobacterium cutirubrum* 11

1.6 The Present Investigation 12

Part 2: Materials and Methods 13

2.1 Materials 13

2.2 Enzymes 13

2.3 Nucleotides and Oligonucleotides 13

2.4 Bacterial Strains, Plasmids and Phage Vectors 14

2.5 Media and Culture Conditions 16

2.6 General Molecular Biology Techniques 16

2.7 Isolation of *Halobacterium cutirubrum* Nucleic Acids 16

2.8 Preparation of Radioactive Probes 17

2.9 Southern Blots 17

2.10 Construction and Screening of Libraries 18

2.11 DNA Sequence Analysis 20

2.12 Northern Blots 22

2.13 Analysis of *in vivo* RNA Transcripts 22

2.14 Nomenclature 23

2.14 Statistical Analysis 24

Part 3: Organization and Expression of the GTPase Domain Genes 28

3.1 Isolation of the Genomic Clones pLW99 and pLW173 28

3.2 Nucleotide Sequence Analysis of pLW173 30

3.3 Identification of Proteins Encoded by pLW173 35

3.4 Amino Acid Composition and Codon Utilization within pLW173 38

3.5 Transcription Analysis of pLW173 43

3.5.1 Transcription of the ORF Gene 43

3.5.2 Transcription of the NAB and L11e Genes 46

3.5.3 Transcription of the L1e, L10e and L12e Genes 50

3.6 Concensus Signal Structures 53

3.6.1 Transcription Initiation Regions 53

3.6.2 Transcription Termination Regions 56

3.6.3 Translation Initiation Regions 56

3.7 Comparative Gene Organization and Expression 57

3.8 Summary 64

Part 4: Structure and Function of the GTPase Domain Proteins 66

4.1 Amino Acid Composition 66

4.2 The L11e Proteins 66

4.3 The L1e Proteins 72

4.4 The L10e Proteins 76

4.5 The L12e Proteins 79

4.6 Intra and Interprotein Relationships 86

4.7 Summary 91

Part 5: Evolution of the GTPase Domain 95

5.1 Evolution of the GTPase Domain 95

5.2 Model of the Evolution of the L10e and L12e Genes and Proteins 96

5.3 Future Prospects 103

References 105

List Of Tables

		Page
Table 1	The Characteristics of the Urkingdoms	3
Table 2	Oligonucleotides	15
Table 3	Nomenclature for the GTPase Domain Proteins	25
Table 4	Amino Acid Composition of Proteins Encoded By pLW173	40
Table 5	Codon Utilization of Proteins Encoded By pLW173	41
Table 6	Interkingdom Comparison of Amino Acid Composition of the L11e, L1e, L10e and L12e Ribosomal Proteins	67
Table 7	Intra and Interkingdom Sequence Similarity of the L11e, L1e and L10e Ribosomal Proteins	70

List Of Figures

Page

Figure 1 The Phylogeny of the Urkingdoms 2

Figure 2 Structure of the GTPase Domain 8

Figure 3 Summary of the Organization and Expression of the L11, L1, L10 and L12 Genes in *Escherichia coli* 10

Figure 4 Southern Blot Analysis and Restriction Maps of pLW99 and pLW173 29

Figure 5 The Nucleotide Sequence of pLW173 31

Figure 6 Identification of the Ribosomal Proteins Encoded by pLW173 36

Figure 7 Alignment of the Putative NAB Protein with EcoRI and PstI 39

Figure 8 Transcription of the ORF gene 44

Figure 9 Transcription of the NAB and L11e genes 47

Figure 10 Transcription of the L1e, L10e, and L12e genes 51

Figure 11 Summary of Gene Expression in pLW173 54

Figure 12 Summary of the Organization and Expression of the L11e, L1e, L10e and L12e Genes within the Urkingdoms 58

Figure 13 Autogenous Regulation of the L1e - L10e - L12e Operon 61

Figure 14 Archaebacterial Transcription Initiation Sequences 63

Figure 15 Alignment of the L11e Proteins 68

Figure 16 Alignment of the L1e Proteins 73

Figure 17 Alignment of the L10e Proteins 77

Figure 18 Alignment of the L12e Proteins 80

Figure 19 Line Diagram of the L12e Proteins 83

Figure 20 Intra and Interprotein Relationships between the L10e and L12e Proteins 87

Figure 21 Summary of the Structure and Function of the L11e, L1e, L10e and L12e Ribosomal Proteins 92

Figure 22 Model of the Evolution of the L10e and L12e Genes and Proteins 97

Abbreviations

AMV	Avian myeloblastosis virus
ATP	Adenosine 5' triphosphate
bp	Base pair
BPB	Bromophenol blue
BSA	Bovine serum albumin
dATP	2' deoxyadenosine 5' triphosphate
dCTP	2' deoxycytidine 5' triphosphate
ddATP	2',3' dideoxyadenosine 5' triphosphate
ddCTP	2',3' dideoxycytidine 5' triphosphate
ddGTP	2',3' dideoxyguanosine 5' triphosphate
ddTTP	2',3' dideoxythymidine 5' triphosphate
dGTP	2' deoxyguanosine 5' triphosphate
dITP	2' deoxyinosine 5' triphosphate
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
dTTP	2' deoxythymidine 5' triphosphate
EDTA	Ethylenediamine tetracetic acid
GTP	Guanosine 5' triphosphate
IPTG	Isopropylthio β Dgalactoside
Kbp	Kilobase pair
MOPS	Morpholinopropane sulfonic acid
mRNA	Messenger RNA
PIPES	Piperazine N,N' bis[2 ethane sulfonic acid]
RNA	Ribonucleic acid
RNase	Ribonuclease
rRNA	Ribosomal RNA
SDS	Sodium dodecylsulphate
SSC	standard saline citrate
Tris	Tris hydroxymethylaminomethane
tRNA	Transfer RNA
XC	Xylene cyanol
Xgal	5 bromo 4 chloro 3 indolyl β D galactoside

Acknowledgements

Sincere thanks to Pat Dennis for providing opportunity, advice and encouragement during the eons, fellow workers Sylvia Evans, Celia Ramirez, Bruce May, Phalgun Joshi, Peter Durovic, Janet Yee, Ivy Hui and Joan MacPherson for thoughts and discussions and Deidre de Jong Wong for large scale plasmid preps, the lemon pie and the meticulously crafted Opus on my much used eraser. Special thanks to Jeff Leung, Willa Downing and Craig Newton for tempting me (however occasionally) into the light of (day and) civilization, to Ross MacGillivray for technical expertise in the early days and Al Matheson for inspiration. I also wish to thank the security folks for understanding the need for perspective during those frequent rooftop sojourns and The Anteater for being.

Moat Reef

Part 1: Introduction

1.1 Early Evolution and the Archaebacteria

Life has had a long residence on Earth; the most ancient indisputable microfossils of mat forming bacteria preserved as stromatolites occur in the Warrawoona group in Australia, with an age of 3500 million years (Walter, 1983). The most ancient geologic facies, the Isua Supracrustal Belt in Greenland which is too metamorphosized to preserve microfossil structures, retains carbon isotope ratios indicative of the presence of biological metabolism (Schidlowski, 1988). Because of a lack of geologic facies, *bona fide* fossils and a dearth of distinguishing fossiliferous structures in ancient organisms, the record provides little information on the primordial evolution of life.

Throughout evolution organisms have carried their ancestry with them in their genes, thus molecular chronometers, proteins or nucleic acids maintaining constant functionality and wide distribution, can shed light on the evolution of extant organisms (Zuckerkandl and Pauling, 1965). Prior to 1977, extant life was viewed as dichotomous, the more ancient procaryotes giving rise, only recently through symbiosis, to the eucaryota (Margulis, 1970; Stanier, 1970; Schwartz and Dayhoff, 1978). In 1977 Woese and Fox (1977) proposed, based on comparative analysis of the slowly evolving, universally distributed, easily isolated and functionally constant small subunit ribosomal RNA, that extant life forms could be grouped into 3 urkingdoms, the eubacteria, the urcaryota (the nucleus of the modern eucaryote) and a novel group of organisms, the archaebacteria (Figure 1). Although the ensuing decade has seen vigorous debate on primordial evolution with various alternative phylogenies being proposed (the most recent being Cavalier - Smith, 1987 and Lake, 1988) the original archaebacterial conception remains the most promising description (Woese and Fox, 1978; Steitz, 1978; Fox *et al.*, 1980; Van Valen and Maiorara, 1980; Hori *et al.*, 1982; Lake *et al.*, 1984; Stackebrandt, 1985; Garrett, 1985; Cavalier - Smith, 1986; Lederer, 1986; Lake *et al.*, 1986; Lake, 1986a; Lake, 1986b; Stoffler and Stoffler-Meilicke, 1986; Woese *et al.*, 1986; Zillig, 1986; Gouy and Li, 1989; Olsen and Woese, 1989). Archaebacteria are distinguished from eubacteria and eucaryota by a suite of unique and shared characteristics (Table 1; reviews: Jones *et al.*, 1987; Woese, 1987; Danson, 1989). Archaebacteria display 3 distinct phenotypes derived from an anaerobic, thermophilic ancestor: methanogenic - anaerobic production of methane while reducing CO2;

Figure 1 The Phylogeny of the Urkingdoms

The evolution of the archaebacteria, eubacteria and eucaryota from the ancestral cellular state, i.e. the progenote, is illustrated (from Woese, 1987). The distances are derived from comparative analysis of the 16S and 18S rRNA sequences. The chloroplast and mitochondria are descendents of symbiotic cyanobacteria and purple sulfur eubacteria respectively. The basal bodies and lysosomes of the eucaryota may also represent remnants of ancient symbiotic acquisitions of eubacterial cells by eucaryota.

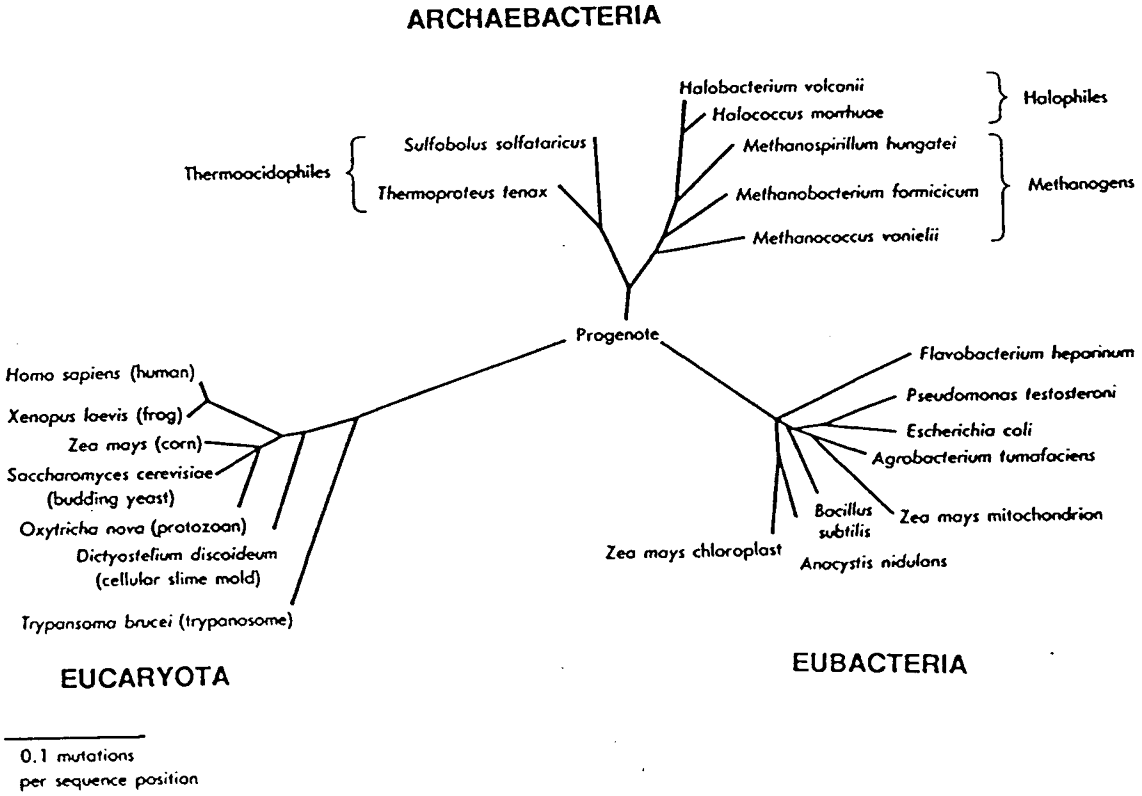


Table1 The Characteristics of the Urkingdoms

Distinguishing features of the eubacteria, archaebacteria and eucaryota are listed. Abbreviations are chloramphenicol (CM), anisomysin (Ani), kanamycin (Kan), pseudouracil (ψ), α - amanitin (Ama) and rifampin (Rif).

Characteristic	Eubacteria	Archaebacteria	Eucaryota
Cellular Organization	anucleate	anucleate	nucleated with organelles
Genome Size (bp)	5x10^5 - 5x10^6	5x10^5 - 10^7	1.5x10^7 - 3x10^{11}
Membrane Lipids	ester linked straight chain	ether linked branched chain	ester linked straight chain
Cell Walls	peptidoglycan	various but not peptidoglycan	various or none
Ribosomes			
rRNA	5S, 16S, 23S	5S, 16S, 23S	5S, 5.8S, 18S, 28S
diptheria toxin	insensitive	sensitive	sensitive
antibiotic sensitivity	CM^S Ani^R Kan^S	CM^R Ani^S Kan^R	CM^R Ani^S Kan^R
Transfer RNA			
TψC loop	TψCG	1 - methyl$\psi\psi$CG	TψCG
1 - methyl adenine	absent	present	present
initiator tRNA	5' monophosphate	5' triphosphate	5' monophosphate
initiator amino acid	N - formyl methionine	methionine	methionine
RNA Polymerase			
number of types	1	1	3
subunits	5	6 - 13	12 or greater
antibiotic sensitivity	Ama^R Rif^S	Ama^R Rif^R	Ama (Pol II)^S (Pol I+III)^R Rif^R
mRNA	uncapped	uncapped	7 - methyl G cap and polyadenylation

thermoacidophilic - sulfur dependent oxidation or respiration at obligately high temperatures to 110°C and halophilic - requirement for extreme salt concentrations to the point of saturation.

1.2 Halophilic Archaebacteria

Magrum *et al.* (1978) discovered that the extreme halophiles were members of the archaebacteria, having descended from the anaerobic methanogens. They have secured a place of paramount importance in the endeavors of humanity by rotting salted fish and turning salt pans a really neat red. Archaebacterial halophiles display a variety of morphologies (rod, coccus, disk and pleomorph), generate energy from the aerobic metabolism of carbohydrates and amino acids, and have optimal growth conditions of 30°C to 50°C, pH neutral (*Halobacterium*) or alkaline (*Natronobacterium*) and 1.7 M to 4.5 M NaCl.

The best characterized halophiles are the closely related species *Halobacterium halobium*, *H. cutirubrum* and *H. salinarium* (now classified as strains of the single species *H. salinarium*; Larsen, 1984). Their genomes contain approximately 4000 Kilobasepairs of DNA and exhibit a high frequency of spontaneous rearrangement by means of deletion, transposition and recombination events caused by more than 50 families of insertion elements and repetitive sequences (Pfeifer *et al.*, 1981; Sapienza *et al.*, 1982; Charlebois and Doolittle, 1988; Pfeifer *et al.*, 1988; Pfeifer *et al.*, 1989). Their genomic DNA can be fractionated into a GC rich fraction containing stable unique single copy chromosomal genes and an AT rich fraction, which undergoes frequent recombinational events, derived from plasmids and a 70 Kilobasepair AT rich island of chromosomal DNA, both enriched for the insertion elements and repetitive sequences (Pfeifer and Betlach, 1985; Kushner, 1985).

H. halobium produces a unique purple membrane, composed of the transmembrane protein bacterio -opsin, a proton pump that generates ATP photosynthetically through establishment of an electrochemical gradient (Stockenius *et al.*, 1979; Stockenius and Bogomolni, 1982). Buoyancy in salt brines required for maintainance of optimal efficiency of the purple membrane is achieved by gas filled proteinaceous vesicles composed, almost exclusively, of a single vacuolar protein. The complex genetics of the purple membrane and vacuole systems have been well studied (Horne *et al.*, 1988; Leong *et al.*, 1988a; Leong *et al.*, 1988b; Betlach *et al.*, 1989; Pfeifer *et al.*, 1989). Both of these systems are inactivated by insertion elements at remarkably high frequencies: 10^{-4} for purple membrane production and 10^{-2} for gas vesicle production (Pfeifer *et al.*, 1981).

Various consensus sequences putatively responsible for expression of halophilic genes have been derived from the approximately 25 cloned genes, however, lack of a transformation system has severely limited functional analysis of these sequences. This lack should be alleviated by the recently described transformation system and shuttle vector for *Haloferax volcanii* (Charlebois *et al.*, 1987; Cline *et al.*, 1989; Lam and Doolittle, 1989).

1.3 Organization of Ribosome Components

The central component of the translation apparatus in all contemporary organisms is a ribonucleoprotein particle, the ribosome. This complex and essential subcellular organelle universally functions by utilizing an mRNA template to align and polymerize amino acids (carried on a set of adaptor tRNAs) into protein. The eubacterial ribosome is comprised of 16S, 23S, and 5S rRNAs and approximately 50 proteins; their eucaryotic counterpart consists of 18S, 5.8S, 28S, and 5S rRNAs and approximately 75 proteins and in archaebacteria the ribosome is comprised of 16S, 23S, and 5S rRNAs and 50 to 65 proteins.

In the eubacteria the organization, transcription and genetic regulation of the 16S - 23S - 5S rRNA transcription units and the ribosomal protein genes have been extensively studied (review: Lindahl and Zengel, 1986). In *E. coli* the rRNAs are encoded on seven operons; the 52 different genes encoding the ribosomal proteins are all single copy, organised into approximately 20 operons distributed throughout the genome and most are located in clusters of one or more transcription units that often contain additional genes encoding protein elements involved in replication (e.g. DNA primase), transcription (e.g. a subunit of RNA polymerase), translation (e.g. EFTu and EFG extrinsic translation factors) or other essential cellular processes. The major gene clusters are the 'RIF' (encoding 4 ribosomal proteins and 2 RNA polymerase subunits), 'STR' (encoding 2 ribosomal proteins and 2 translation factors), 'S10' (encoding 10 ribosomal proteins) and 'SPC' (encoding 14 ribosomal proteins, an RNA polymerase subunit and 2 secretion proteins). Translation of the separate ribosomal protein mRNAs and transcription of the rRNA transcripts are balanced by autogenous translational regulatory mechanisms; assembly of ribosomal particles occurs on nascent rRNA transcripts and neither free rRNAs nor free ribosomal proteins accumulate (review: Nomura *et al.*, 1984).

In eucaryotic cells three separate RNA polymerases are used for transcription of the 18S - 5.8S - 28S rRNA genes, for ribosomal protein encoding genes and for the 5S rRNA genes and tRNA genes. Ribosomal protein genes are encoded on multiple copies of monocistronic transcription units which are rarely clustered within the genome (reviews: Planta *et al.*, 1986; Warner, 1989). Translation of ribosomal protein mRNAs occurs in the cytoplasm and the ribosomal proteins produced are imported into the nucleus where they are assembled into particles at the sites of rRNA transcription. The ribosomal subunits are then exported to the cytoplasm.

Archaebacterial genes encoding the rRNA moieties of the ribosome have been well characterized (Mankin *et al.*, 1984; Hui and Dennis, 1985; Dennis, 1985; Mankin and Kagramanova, 1986; Kjems and Garrett, 1987; Ree *et al.*, 1989; Wolters and Erdmann, 1989). At the outset of this work the structural organization and expression of archaebacterial ribosomal protein genes was unknown. Recently, from this work and others, it has been found that genes encoding ribosomal proteins appear to be clustered as in the eubacteria and facets of their transcriptional organization and regulation have been elucidated (Dennis *et al.*, 1985; Chant *et al.*, 1986; Itoh *et al.*, 1988; Itoh, 1988; Kopke and Wittman-Liebold, 1988; Shimmin *et al.*, 1989a; Shimmin and Dennis, 1989; Auer *et al.*, 1989; Spiridinova *et al.*, 1989; Kopke and Wittmann-Liebold, 1989; Zillig *et al.*, 1989; Ramirez *et al.*, 1990a).

1.4 GTPase Domain of *Escherichia coli*

The binding of extrinsic factors to the ribosome and the concomitant hydrolysis of GTP propagates structural rearrangements in the ribosome during the cyclic amino acid addition process (review: Liljas, 1982). From electron microscopic and biochemical observations it is apparent that the general structure and function of the ribosome factor binding domain (GTPase domain) was fixed prior to the divergence of the archaebacteria, eubacteria, and eucaryotes (Lake, 1983a; Lake, 1983b; Beauclerk *et al.*, 1985; Moller and Maassen, 1986; Oakes *et al.*, 1986; Hanauz *et al.*, 1987; Shimmin and Dennis, 1989; Shimmin *et al.*, 1989a; Shimmin *et al.*, 1989b; Ramirez *et al.*, 1989; Ramirez *et al.*, 1990a; Ramirez *et al.*, 1990b). Subsequently, the proteins forming this domain (i.e. in *E. coli* L11, L1, L10 and L12) have been subject to evolutionary tinkering to refine efficiency and accuracy of protein synthesis in the three separate lineages. The essential features of each protein are, however, expected to be conserved. The GTPase domain forms the stalk structure on the large ribosomal subunit (Figure 2A; Strycharz *et al.*, 1978; Kastner *et al.*, 1981; Marquis *et al.*, 1981; Moller *et al.*, 1983; Traut *et al.*, 1986) and is comprised of a complex of four copies (a pair of dimers) of the L12e protein bound to a single copy of the L10e protein through which the complex binds to the 23S rRNA (Figure 2B; Osterberg *et al.*, 1976; Osterberg *et al.*, 1977; Gudkov *et al.*, 1978; Pettersson and Liljas, 1979; Petterson, 1979; Dijk *et al.*, 1979; Beauclerk *et al.*, 1984). The L11 protein is located at the base of the stalk and is known to bind 23S rRNA at residues 1052 - 1112 (Schmidt *et al.*, 1981; Stoffler-Melicke *et al.*, 1983; Deng *et al.*, 1986; El-Baradi *et al.*, 1987). The L1 protein of *E. coli* has been localized to the ridge region on the 50S subunit opposite the L12 stalk and binds to and protects nucleotides 2100 - 2200 of 23S rRNA (Branlant *et al.*, 1981; Lake and Strycharz, 1981; Oakes *et al.*, 1986).

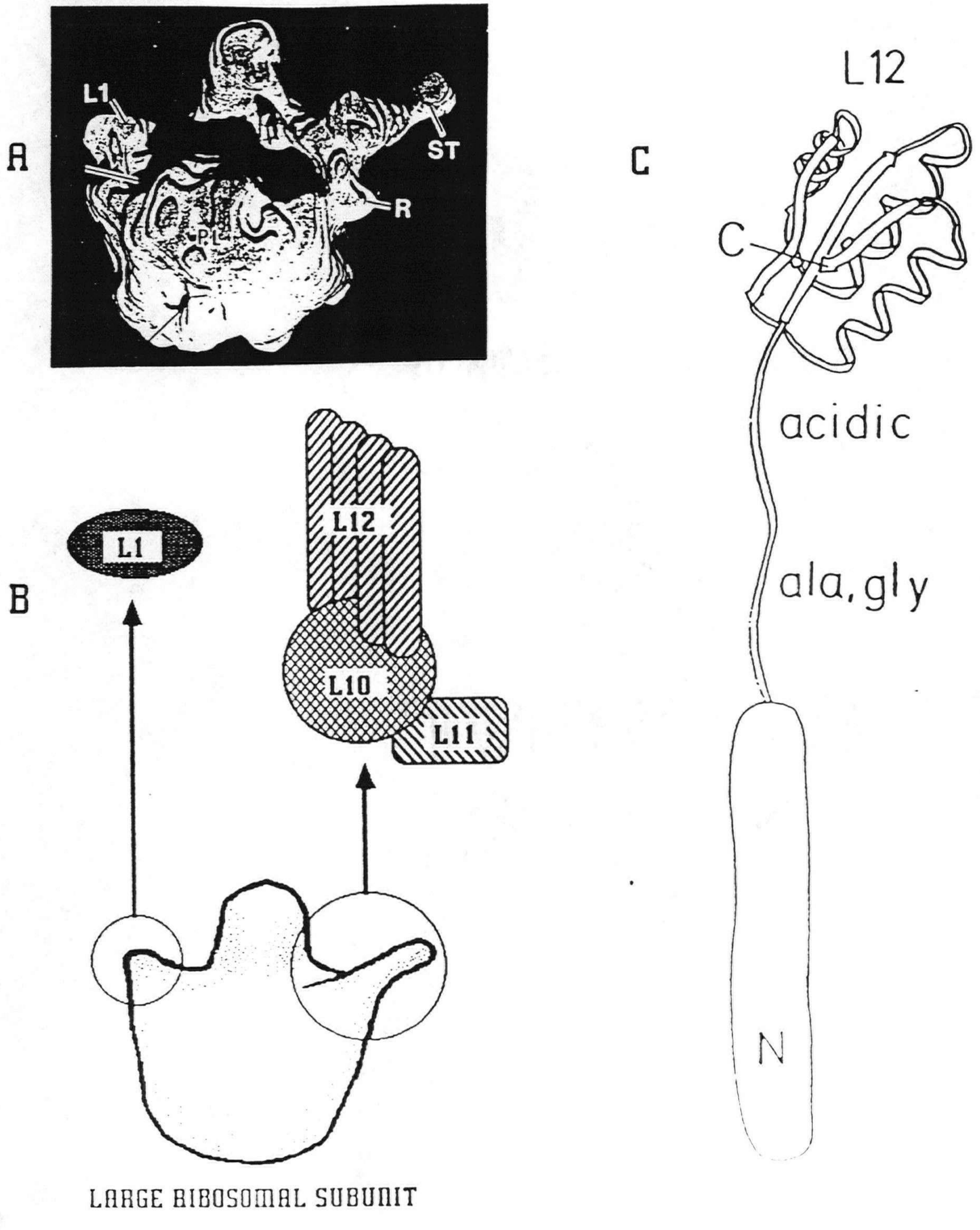
A number of structural and functional features of the *E. coli* L11 protein have been reported. The molecule is highly elongated with an axial ratio of 5.5:1, is rich in proline residues and is the most extensively methylated ribosomal protein, containing nine methyl groups that are added to the protein after translation (Dognin and Wittmann-Liebold, 1977; Giri *et al.*, 1978). The amino terminal domain has been implicated in the interaction of the ribosome with translation release factor 1 (Tate *et al.*, 1984). The protein is involved in the synthesis of guanosine 5' diphosphate, 3' diphosphate (ppGpp) during the stringent response (Friesen *et al.*, 1974; Parker *et al.*, 1976; review: Cundliffe, 1986).

The L1 protein functions to (i) maximize binding of peptidyl - tRNA to the P site, (ii) maximize the GTPase activity associated with EFG - mediated translation and (iii) autogenously regulate the translation of the L11 - L1 mRNA; excess L1 protein not incorporated into ribosomes can bind to a sequence within the 5' untranslated leader of the L11 - L1 mRNA that exhibits both primary and secondary structural similarity to the L1 binding site on 23S rRNA, thereby preventing translation (Dean and Nomura, 1980; Subramanian and Dabbs, 1980; Baughman and Nomura, 1981; Yates and Nomura, 1981; Sander, 1983; Kearney and Nomura,1987; Thomas and Nomura, 1987).

The L10 protein has binding sites for 23S rRNA and four L12 proteins but an active role within the ribosome in the translation process has yet to be demonstrated. Both L10 and the L10 - L12 protein

Figure 2 Structure of the GTPase Domain

(A) A model of the 50S subunit of *Escherichia coli* derived from a computer composite of electron micrographs; 'ST', 'R' and 'L1' indicate the stalk, L11 ridge and L1 shoulder of the subunit respectively (adapted from Radermacher *et al.*, 1987). (B) A schematic illustrating the composition and location of the GTPase domain within the 50S subunit. (C) The structure of the *E. coli* L12 protein (taken from Liljas *et al.*, 1986). The carboxy terminal globular domain of L12 is located at the tip of the stalk structure in diagrams A and B.



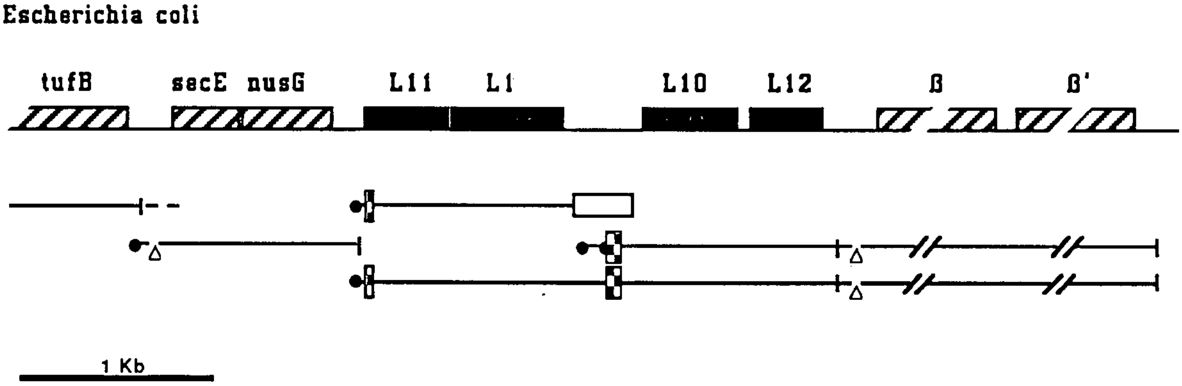
complex function as autogenous translational regulators of the L10 - L12 transcript (Johnsen *et al.*, 1982; Nomura *et al.*, 1984)

L12 is a highly elongated molecule composed of amino and carboxy terminal globular domains connected by an alanine - proline rich region (Figure 2C; Osterberg *et al.*, 1976; Leijonmark *et al.*, 1981). The functions of five translation factors are known to depend upon L12 : IF-2, EF-Tu, EF-G, RF-1 and RF-2 (review: Liljas, 1982). The first three of these factors associate with the ribosome in complex with a GTP molecule to promote a structural rearrangement before GTP is hydrolysed and the factor is released from the ribosome. Biophysical studies on the L12 protein indicate that the amino terminal domain spontaneously dimerizes and contains the site for binding the L12 protein dimers to the L10 protein (Gudkov and Behlke, 1978; Koteliansky *et al.*, 1978). The carboxy terminal domain forms a compact structure of alternating α helices and β sheets that crystallizes as a dimer, contains an anion (potential GTP) binding site, a putative dimerization site, undergoes a conformational change upon interaction with extrinsic translation factors during the protein synthesis cycle and may interact with those factors through a conserved face (Gudkov and Gongadze, 1984; Burma *et al.*, 1985; Leijonmarck and Liljas, 1987). The two domains are separated by an alanine - proline rich region believed to be unstructured and to function as a flexible hinge between domains of the L12 proteins, accounting for the observed high mobility of the carboxy terminal domain (Tritton, 1978; Leijonmarck *et al.*, 1981; Cowgill *et al.*, 1984).

The genes encoding the four proteins of the GTPase domain are genetically linked with two genes encoding RNA polymerase subunit proteins (β and β') in the order L11 - L1 - L10 - L12 - β - β' (Figure 3; Lindahl *et al.*, 1975). The three kilobasepair region of genomic DNA encoding the L11, L1, L10 and L12 proteins has been sequenced and the regulation and expression of the genes within it have been extensively characterized (Post *et al.*, 1979; reviews: Lindahl and Zengel, 1986; Jinks-Robertson and Nomura, 1987). Regulation of the gene cluster is complex; two major promotors upstream of the L11 and L10 cistrons yield both bicistronic L11 - L1 and L10 - L12 transcripts and tetracistronic L11 - L1 - L10 - L12 transcripts. The RNA polymerase subunits encoded by the downstream β and β' genes are produced by elongation of a fraction (20%) of the transcripts exiting the L12 gene, the remainder of the transcripts are terminated by a transcription attenuator in the L12 - β intergenic space (Dennis, 1977; Barry *et al.*, 1979, Barry *et al.*, 1980; Dennis, 1984). Sequences surrounding the RNaseIII processing site

Figure 3 Summary of the Gene Structure of the L11, L1, L10 and L12 Genes in *Escherichia coli*

The organization and transcription of the L11, L1, L10, and L12 ribosomal protein gene cluster of *Escherichia coli* is depicted. Ribosomal protein encoding genes are solid boxes and other protein encoding genes are striped boxes. All genes are oriented and transcribed rightwards. The filled circles (●) represent 5' transcript ends and the vertical lines (**I**) represent 3' transcript ends. The open boxes (□) at the ends of trancripts indicate regions of multiple 3' trancript ends. Only the 3' end of the tufB gene is indicated. Triangular interruptions (△) represent RNaseIII processing sites and the vertical line on the transcripts running through the L12 - β intergenic space represents a transcription attenuator. The checkered boxes (▦) represent sites of autogenous regulation by the L1 and L10 proteins and the L10 - L12 complex by translational inhibition of their respective mRNAs.



downstream from the transcription attenuator appear to be essential for efficient translation of the β and β' genes although processing at this site has no effect on expression of these genes (Barry *et al.*, 1980; Dennis, 1984). The L1 protein autogenously regulates the translation of the L11 and L1 proteins through the leader region of the L11 - L1 mRNA; the regulatory site is well characterized and overlaps with the L11 translation initiation codon (Baughman and Nomura, 1983; Thomas and Nomura, 1987; Said *et al.*, 1988). The L10 protein and L10 - L12 complex autogenously regulate the translation of the L10 - L12 mRNA. The regulatory site is located 140 nucleotides upstream of the L10 translation initiation codon; a model for translational regulation by alternative secondary structures has been proposed to account for the long range effect of the regulatory protein(s) (Fiil *et al.*, 1980; Johnsen *et al.*, 1982; Christensen *et al.*, 1984; Petersen, 1989). The mechanism of the translational enhancement required to produce four copies of the L12 protein versus single copies for all other ribosomal proteins remains unknown although it has been suggested that a promotor exists in the L10 - L12 intergenic space (Newman *et al.*, 1979; Ma *et al.*, 1981).

1.5 Ribosomal Proteins of *Halobacterium cutirubrum*

Early studies of the extreme halophile *Halobacterium cutirubrum* showed that their ribosomes exhibit several unique characteristics that distinguish them from the ribosomes of eubacteria, eucaryota and the thermoacidophilic and methanogenic archaebacteria, suggesting that their structures have undergone substantial alterations in adapting to their extreme environment. Ribosomes of the extreme halophiles require at least 3.4 M K^+ and 0.1 M Mg^{2+} to stabilize their structures; conditions capable of disintegrating and denaturing ribosomes from other organisms (Bayley and Kushner, 1964; Rauser and Bayley, 1968; Kushner, 1985). Most if not all of the 53 ribosomal proteins are acidic with an average isoelectric point of 3.9 (Bayley and Kushner, 1964; Bayley, 1966; Strom and Visentin, 1973; Strom *et al.*, 1975). Acidic residues bind more water molecules per side chain than other amino acids, allowing them to maintain a hydration shell around the proteins in a high salt milieu (Bayley and Morton, 1978; Saenger, 1987; Eisenberg and Wachtel, 1987). Approximately 25 of the proteins have had their amino terminal sequences determined; due to sequence divergence and a lack of assayable functions and comparable sequences, only 6 of the proteins were identified as potential homologues to eubacterial, archaebacterial or eucaryotic ribosomal proteins (Oda *et al.*, 1974; Duggleby *et al.*, 1975; Yaguchi *et al.*, 1982; Matheson *et al.*, 1984). The *H. cutirubrum* L20 protein was demonstrated to have homologues in eubacteria (*E. coli* L12), in eucaryota (*Artemia salina* eL12) and in archaebacteria (*Methanobacterium thermoautotrophicum* 'A' protein) by sequence comparison and by virtue of sharing the characteristics of being (i) the most acidic ribosomal protein, (ii) extremely alanine rich, (iii) the only protein present in multiple copies per ribosome and (iv) part of the protein complex forming the stalk structure of the 50S subunit (Oda *et al.*, 1974; Amons *et al.*, 1979; Matheson *et al.*, 1979; Visentin *et al.*, 1979; Yaguchi *et al.*, 1980; Gudkov *et al.*, 1984). The homology of the *H. cutirubrum* L11 protein and *E. coli* L11 proteins was based on the available short proline rich amino terminal sequence similarity and the presence of the protein in the GTPase domain (Matheson *et al.*, 1984). Other putative homologies were based solely upon short sequence similarities.

1.6 The Present Investigation

The present investigation concerning the GTPase domain of *H. cutirubrum* had 3 objectives: (i) provision of basic data on structure and organization of translated genes in archaebacteria, (ii) investigation of the expression and regulation of those genes and (iii) to provide perspectives on both the early evolution of the translation apparatus and its later adaptation to high concentrations of salt within the extreme halophiles. Proteins comprising the GTPase domain (specifically L20 and L11) were chosen as cloning would be facilitated by the available partial amino acid sequences, the expression and regulation might prove as complex as that of the GTPase domain of *E. coli* and homologues from the extant urkingdoms were available for evolutionary comparison studies. This thesis presents the cloning, sequencing, transcriptional characterization and evolutionary analysis of the *H. cutirubrum* homologues of the L11, L1, L10 and L12 ribosomal proteins comprising the GTPase domain of *E. coli*.

Part 2: Materials and Methods

2.1 Materials

Culture media components were obtained from: agar, yeast extract, tryptone and casamino acids from Difco Laboratories, ampicillin from Sigma Chemical Co., IPTG and Xgal from BRL and D glucose from BDH. Phenol was obtained from Mallinkrodt, redistilled and stored under water at 4°C. Acrylamide was from Eastman Kodak, BioRad Laboratories and Serva and N N' methylene bisacrylamide was from Eastman Kodak. Agaroses were obtained from: analytical agarose from Sigma Chemical Co., preparative agarose from Schwarz/Mann Biotech, low melting point agarose from BRL and Nusieve agarose from FMC Corporation. Formamide was deionized with BioRad AG501 - X8 - D mixed bed resin for one hour (10 mL resin per 100 mL formamide) and stored at -20°C. Formaldehyde was from BDH. Eastman Kodak XRP-1, XAR-5, Amersham Hyperfilm β Max films and Dupont Cronex Lightning Plus intensifying screens were used for autoradiography of labelled nucleic acids.

2.2 Enzymes

Restriction enzymes and DNA modifying enzymes were purchased from the following commercial suppliers and used according to the manufacturers recommendations: Boehringer Mannheim Canada (BMC), Bethesda Research Laboratories (BRL), International Biotechnologies Incorporated (IBI), New England Biolabs (NEB), New England Nuclear (NEN) and Pharmacia (P). Other enzymes were purchased from: T4 polynucleotide kinase - PL Biochemicals, P; DNA polymerase I - BMC; calf intestinal alkaline phosphatase - BMC, P; ribonuclease A , lysozyme - Sigma; T4 DNA ligase and S1 nuclease - P; Klenow fragment - BMC, P, BRL, Promega; AMV reverse transcriptase - BMC, IBI; Sequenase - United States Biochemical Corp.; T7 DNA polymerase - P; exonuclease III - BMC and Promega; terminal transferase - BRL.

2.3 Nucleotides and Oligonucleotides

Ribonucleoside triphoshates, deoxyribonucleoside triphosphates, dideoxyribonucleoside triphosphates and (1) - phosphorothioate deoxynucleotide triphosphates were obtained from Pharmacia.

$\alpha^{32}P$, $\gamma^{32}P$ and $\alpha^{35}S$ labelled nucleotides were obtained from New England Nuclear and Amersham.

The universal 17 nucleotide forward, 5' d[GTAAAACGACGGCCAGT] 3' and 17 nucleotide reverse, 5' d[AACAGCTATGACCATG] 3' sequencing primers were obtained from New England Biolabs. Oligonucleotides used as probes for genes, for progressive deletion of inserts in M13 phage by the procedure of Dale *et al.* (1985) and for primer extension analysis were synthesized by T. Atkinson (University of British Columbia) on an Applied Biosystems 380B DNA Synthesizer and supplied as lyophilized crude powders (Table 2). Crude oligonucleotides were purified by polyacrylamide gel electrophoresis and C18 Sep-Pak reverse phase chromatography and quantified by measuring the A_{260} (Atkinson and Smith, 1984).

2.4 Bacterial Strains, Plasmids and Phage Vectors

Halobacterium cutirubrum NRCC 34001 was obtained from A.T. Matheson at the University of Victoria. *Escherichia coli* strain JM83 (ara, Δ(lac-proAB), rpsL, Φ80dlacZΔM15) was used for propagation of plasmids and construction of the library libLW22 (Viera and Messing, 1982; Messing, 1983). Strains DH1 (F-, recA1, endA1, gyrA96, thi-1, hsdR17, supE44, relA1, λ-) and DH5α (F-, Φ80dlacZΔM15, Δ(lacZYA-argF)U169, recA1, endA1, gyrA96, thi-1, hsdR17, supE44, relA1, λ-) were used for maintainance of plasmids (Hanahan, 1983). Strains JM101 (supE, thi, Δ(lac-proAB), [F', traD36, proAB, lacI^{q}ZΔM15]) and JM109 (recA1, endA1, gyrA96, thi-1, hsdR17, supE44, relA1, λ-, [F', traD36, proAB, lacI^{q}ZΔM15]) were used for propagation of M13 phage (Messing, 1983). Strain JC8111 (recB21, recC22, recF143, sbcB15, argE3, his-4, leu-6, proA2, thr-1, rpsL31, galK2, lacY1, ara-14, xyl-5, mtl-1, supE44) was used for construction of the libraries libLW37 and libLW38 (Boissy and Astell, 1985). The plasmid vectors pUC8, pUC12, pUC13, pEMBL8+, pEMBL8-, pTZ18R, pTZ19R and pBR322 (Bolivar *et al.*, 1977; Viera and Messing, 1982; Messing, 1983; Dente *et al.*, 1983) and the phage vectors M13mp10, M13mp11, M13mp18 and M13mp19 (Messing, 1983; Yanisch-Perron *et al.*, 1985) were used for cloning, subcloning and sequencing. Plasmids pUC12 and pUC13 were obtained from M. Zoller (University of British Columbia), pEMBL8+ and pEMBL8- from J. Leung (University of British Columbia) and pTZ18R and pTZ19R were purchased from Pharmacia.

Table 2 Oligonucleotides

Designation	Sequence 5' - 3'[1]	Length	Degeneracy	Position[2]	Protein or Transcript[3]	Strand[4]
(A) OLIGONUCLEOTIDES FOR GENE PROBES						
oLW9	[6] AlaTyrValTyrGluMet [1] GCGTAGACGTATTCCAT A A A C	17	16	4034 - 4018	L20 {L12e}	antisense
oLW17	[2] AlaGluThrIleGluVal [7] GCGGAGACGATAGAGGT A A A T A T T C C C	17	192	1625 - 1641	L11 {L11e}	sense
oLW35	[95] AsnAspAsnProPheGly [100] AATGATAATCCGTTTGG C C C A C T C	17	64	3236 - 3252	L3 / 4 {L10e}	sense
(B) OLIGONUCLEOTIDES FOR PRIMER EXTENSION						
oLW36	ATGTGGGCTTCTGTCGA	17	1	1165 - 1181	ORF	
oLW38	CGATCTGCGTCTCCTGT	17	1	2494 - 2478	L1e - L10e - L12e	
oLW51	TACGTCGACCGGCGTGGGAC	20	1	1714 - 1695	NAB - L11e	
oLW52	CTTCGAGGTCCACCTCGATG	20	1	1429 - 1410	NAB	
oLW54	CGTTGTCTGCCATCTTTCAC	20	1	2326 - 2307	L1e - L10e - L12e	

(1) The oligonucleotide sequence is written below the peptide sequence from which it was derived where applicable. Numbers preceding and following the amino acid sequence indicate the position of the peptide in the protein.

(2) The position corresponds to that in Figure 5.

(3) The oligonucleotide hybridizes to the gene encoding the indicated equivalent protein for the gene probes and to the indicated transcript for the primer extension oligonucleotides.

(4) Antisense indicates that the oligonucleotide is complementary to the mRNA. All oligonucleotides for primer extension analysis are complementary to the mRNA.

2.5 Media and Culture Conditions

Halobacterium cutirubrum was grown at 42°C in a rich medium as described by Bayley (1971). The medium contained 4.28 M NaCl, 81 mM $MgSO_4$, 27 mM KCl, 10 mM Na citrate, 180 μM $FeSO_4$, 1% w/v yeast extract and 0.75% w/v casamino acids, was adjusted to pH7.4 - 7.6, autoclaved 10 minutes, filtered through Whatmann No.1 filter paper, acidified with HCl to pH6.2 and autoclaved for 20 minutes. *H. cutirubrum* stocks were stored in broth culture at 4°C and maintained by subculturing at 6 month intervals. *Escherichia coli* was grown at 37°C in M9 minimal medium (50 mM Na_2HPO_4, 25 mM KH_2PO_4, 8.5 mM NaCl, 20 mM NH_4Cl, 1 mM $MgSO_4$, 0.1 mM $CaCl_2$ and 0.2% glucose; Miller, 1972) or YT rich medium (86 mM NaCl, 0.8% w/v tryptone and 0.5% w/v yeast extract; Maniatis *et al.*, 1982). Agar plates contained 1.5% w/v agar and soft agar overlays for phage growth contained 0.75% w/v agar. Ampicillin was added to a concentration of 20 μgmL^{-1} to 200 μgmL^{-1} when required for plasmid selection and 50 μM IPTG and 0.005% w/v Xgal were added to cooled (45°C) agar for visualization of β galactosidase activity.

2.6 General Molecular Biology Techniques

General techniques of molecular biology were done essentially as described in Maniatis *et al.* (1982). Small and large scale plasmid preparations were done by the methods of Birnboim and Doly (1980) and Maniatis *et al.* (1982). Phage preparations were done as described by Sanger *et al.* (1977), Messing (1983) and Dente *et al.* (1983)

2.7 Isolation of *Halobacterium cutirubrum* Nucleic Acids

For isolation of *H. cutirubrum* DNA , a 2 L culture was grown to an A_{600} of 1.5 and pelleted, yielding 12 g of cells. The cells were resuspended in 30 mL of 4.0 M NaCl, 120 mM $MgSO_4$, 10 mM Na citrate, 30 mM KCl, lysed at 0°C for 40 minutes by addition of 6 mL of 10% w/v Na deoxycholate and diluted with 70 mL of 100 mM Tris-HCl pH8.0, 10 mM EDTA. The lysed cells were extracted twice with phenol, twice with a 1:1 mixture of n-octanol:chloroform and dialysed against 10 mM Tris-HCl pH8.0, 1 mM EDTA. CsCl was added and the DNA purified through ultracentifugation in a Beckman Ti60 rotor for 70 hours at 36 Krpm. Recovered DNA was precipitated twice with ethanol and resuspended in 10 mM Tris-HCl pH8.0, 1 mM EDTA and stored at minus 20°C. The yield of pure genomic DNA was 6.2 mg.

Glass and plastic wares used for isolation of total cellular *H. cutirubrum* RNA were treated for 1 hour at 121°C with 0.1% v/v diethylpyrocarbonate. Log phase *H. cutirubrum* cells (A_{600} of 0.5 - 0.7) were cooled rapidly with frozen media, treated for 5 minutes with 10 mM NaN_3 and pelleted. The pellet was resuspended in 1 mL of 37 mM NH_4Cl, 2 mM Na_2HPO_4, 2 mM KH_2PO_4 and 5 mM NaCl, lysed by heating at 100°C for 30 seconds after addition of 1 mL of 100 mM NaCl, 10 mM EDTA and 0.5% w/v SDS, extracted thrice with phenol, once with chloroform and precipitated twice with ethanol. DNA contaminants were removed by ultracentrifugation though a 5.7 M CsCl block gradient or by selective precipitation of RNA by 1 M LiCl (Chirgwin *et al.*, 1979; Auffray and Rougesn, 1980). The RNA was resuspended in 10mM Tris-HCl pH7.5, 1 mM EDTA and stored at -70°C. Yields were approximately 250 µg from a 10 mL culture.

2.8 Preparation of Radioactive Probes

Oligonucleotides were 5' end labelled with 5 units of T4 polynucleotide kinase in a 10 µL mixture containing 50 µCi of 3000 Ci $mmol^{-1}$ $\gamma^{32}P$ ATP, 5 pmol oligonucleotide, 50 mM Tris-HCl pH8.0, 10 mM $MgCl_2$, 10 mM DTT. The labelled oligonucleotides were purified by exclusion chromatography on Sephadex G25 fine or by ethanol precipitation. Restriction fragments containing recessed 3' ends were 3' end labelled using Klenow enzyme or 5' end labelled using T4 polynucleotide kinase after dephosphorylation with calf intestinal alkaline phosphatase as described by Maniatis *et al.* (1982). High specific activity double stranded DNA probes were obtained by nick translation using DNA polymerase I and two radioactive deoxynucleotide triphosphates (Rigby *et al.*, 1977) or by random priming from mixed oligonucleotides with Klenow enzyme (Fienberg and Vogelstein, 1982). High specific activity single stranded DNA probes were generated by extension of oligonucleotide primers on M13 templates in the presence of two radioactive deoxynucleotide triphosphates.

2.9 Southern Blots

Southern blots of genomic DNA hybridized to radioactive oligonucleotide mixtures were used to identify the genomic restriction fragments containing the GTPase domain ribosomal protein genes (Southern, 1975). Genomic DNA was digested with a variety of restriction enzymes, loaded onto horizontal agarose gels containing 0.5% to 2% w/v agarose, 40 mM Tris-HCl pH8.0, 20 mM Na acetate, 5 mM EDTA and electrophoresed at 3 Vcm^{-1}. The DNA was then transferred to nitrocellulose by blotting with 20x SSC and dried at 80°C for 2 hours. Dried filters were prehybridized in 6x SSC (900 mM NaCl, 90 mM Na citrate), 10x Denhardt's (2% w/v bovine serum albumin, 2% w/v polyvinyl-pyrollidine, 2% w/v ficoll) at 60°C for 1 hour before addition of radioactive DNA probe and hyridization for 12 hours at reduced temperature according to the characteristics of the probe DNA (i.e. oLW9, 49°C; oLW17, 45°C; oLW35, 41°C). Blots were washed briefly twice at room temperature in 2x SSC, twice at the hybridization temperature for 10 minutes in 1x SSC, once at the hybridization temperature for 10 minutes in 0.2x SSC, dried and exposed to Kodak XRP-1 film.

Southern blot experiments utilizing probes longer than 40 nucleotides (i.e. determination of copy number of the GTPase domain genes and checking the sequence fidelity of the pLW173 and pLW180 clones) were performed by transferring the restricted and electrophoresed DNA onto Genescreen nylon filters (New England Nuclear) with 2x SSC, drying at 80°C for 2 hours, prehybridization at 60°C in 2x SSC, 5x Denhardt's, 50 μgmL^{-1} denatured calf thymus DNA, 0.5% SDS for 6 hours before addition of the radioactive DNA probe (concentration of less than 20 $ngmL^{-1}$) and hybridization at 60°C for 12 hours. The filters were then washed briefly twice at room temperature in 2x SSC, twice at 60°C for 10 minutes in 1x SSC, 0.5% SDS, dried and exposed to Kodak XAR-5 film. For the experiment determining the copy number of the GTPase domain genes decreased stringency was achieved by coordinated reduction of the hybridization and wash temperatures to a minimum of 45°C and for the sequence fidelity Southern blot restricted DNA was electrophoresed in 3% Nusieve agarose for superior resolution of low molecular mass fragments.

2.10 Construction and Screening of Libraries

The 1.3 Kilobasepair BamHI - PstI fragment of *H. cutirubrum* genomic DNA hybridizing to the Hcu L20 specific oligonucleotide probe oLW9 was cloned by the following procedure. Genomic DNA (100 µg) was doubly restricted with BamHI and PstI, electrophoresed on a 4% polyacrylamide gel, stained with ethidium bromide and 6 fractions containing DNA in the size range of 1.0 to 1.6 Kilobasepairs were recovered by excision from the gel, electroelution and ethanol precipitation. The fraction containing the 1.3

Kilobasepair fragment was identified by electrophoresing an aliquot of each fraction on a horizontal agarose slab gel, transferring the DNA to nitrocellulose by blotting with 20x SSC, hybridization with 5' end labelled oLW9, washing nonstringently with 2x SSC at 40°C for 20 minutes and exposure to Kodak XRP-1 film. The library libLW22 was constructed by ligating the DNA size fraction hybridizing to oLW9 into BamHI - PstI restricted pUC8, transformation into JM83 and plating at a density of 500 colonies per plate onto 15 cm agar plates containing ampicillin, IPTG and Xgal. Colonies were lifted from the plates with Colony-PlaqueScreen (New England Nuclear), denatured twice for 2 minutes with 0.5 M NaOH, neutralized twice for 2 minutes with 1.0 M Tris-HCl pH7.5 and dried at room temperature. Each disk was prehybridized in 5 mL of 5x SSC, 0.5% SDS, 10x Denhardt's for 6 hours, hybridized with 5' end labelled oLW9 (2 Mcpm per disk) for 12 hours at 49°C, washed briefly twice at room temperature in 2x SSC, twice at 49°C for 10 minutes in 1x SSC, 0.5% SDS, once at 49°C for 10 minutes in 0.2x SSC, dried and exposed to Kodak XRP-1 film. Positive colonies were picked, streak purified and checked for the correct insert DNA by Southern blotting. This yielded multiple independent clones, two of which, pLW99 and pLW102, were chosen for further study.

The 5.7 and 5.1 Kilobasepair BamHI - ClaI fragments of *H. cutirubrum* genomic DNA hybridizing to the Hcu L11 specific oligonucleotide probe oLW17 were cloned by the following procedure. Genomic DNA (100 µg) was doubly restricted with BamHI and ClaI, electrophoresed on a 0.5% low melting point agarose gel, stained with ethidium bromide and 8 fractions containing DNA in the size range of 4 to 7 Kilobasepairs were recovered by excision from the gel, melting of the gel matrix at 55°C, extraction with phenol twice and twice precipitated with ethanol. Separate fractions containing the 5.7 and 5.1 Kilobasepair fragments were identified by Southern analysis using oLW17 as probe as described above for the Hcu L20 specific fragment. The library libLW38 was constructed by ligating the DNA size fraction containing the 5.1 Kilobasepair fragment hybridizing to oLW17 into pBR322 that was doubly restricted with BamHI and ClaI and dephosphorylated with calf intestinal alkaline phosphatase. The ligated mixture was transformed into JC8111 and then treated as described for libLW22 except that 5' end labelled oLW17 (25 Mcpm per disk) was used as probe and the hybridization and stringent wash temperatures were 45°C. This yielded two independent clones, pLW173 and pLW180, containing the 5.1 Kilobasepair fragment which were used for further studies. Two clones, pLW155 and pLW156, containing the 5.7 Kilobasepair fragment were isolated from a library, libLW37, constructed as for libLW38 except for using the 5.7 Kilobasepair DNA size fraction as insert DNA. Subcloning and sequencing of a 1 Kilobasepair EcoRI fragment contained within the insert of pLW155 that hybridized to oLW17 indicated that the match to oLW17 was fortuitous and thus characterization of these clones was not persued.

2.11 DNA Sequence Analysis

DNA fragments for chemical sequencing were prepared by 3' or 5' end labelling, followed either by digestion with a restriction enzyme yielding size differentiated fragments or by denaturation. The restricted double stranded or strand separated single stranded uniquely end labelled fragments were then purified through polyacrylamide. Chemical sequences of DNA and oligonucleotides were performed essentially as described by Maxam and Gilbert (1977) and by a modified procedure featuring the immobilization of DNA fragments on treated paper substrates (Rosenthal *et al.*, 1985; Rosenthal *et al.*, 1986). For this method 200 Kcpm of a uniquely 3' or 5' endlabelled DNA fragment was denatured by heating at 100°C for 3 minutes, chilled rapidly on ice and immobilized on strips of Hybond M & G paper (Amersham). The strips were washed twice in distilled water, once in ethanol, air dried and transferred to 500 µL Eppendorf tubes where the chemical modification reactions were performed. Reactions were T > C (380 µM K_2MnO_4 for 20 minutes), C (4M hydroxylamine hydrochloride pH6.0 for 10 minutes), A + G (88% v/v formic acid for 10 minutes) and G (50 mM ammonium formate pH3.5, 0.7% v/v dimethyl sulphate for 45 seconds). Reactions were terminated by washing the strips twice in distilled water and once in ethanol. Cleavage was accomplished by addition of 75 µL of 10% v/v piperidine and heating to 90°C for 30 minutes. The products were then recovered by two lyophilizations, redissolved at 1 - 20 Kcpm μL^{-1} in FDM (98% formamide, 10 mM EDTA, 2 $mgmL^{-1}$ XC and 2 $mgmL^{-1}$ BPB), denatured by heating at 90°C for 3 minutes and electrophoresed on polyacrylamide sequencing gels.

For enzymatic sequencing subclones were generated by shotgun and fragment specific subcloning into pUC, pEMBL, pTZ and M13 vectors. Deletion of some of these derivatives was performed on double stranded DNA with exonuclease III as per Henikoff (1984) and on single stranded DNA by the method of Dale *et al.* (1985). Double stranded DNA templates were prepared by the alkaline denaturation and ethanol precipitation in the presence of the appropriate oligonucleotide primer as described by Hattori and

Sakaki (1986). If required the plasmid DNAs were purified through CsCl centrifigation prior to alkaline denaturation. Single stranded DNA templates were prepared as described from M13 derivatives (Sanger *et al.*, 1977; Messing, 1983), from pEMBL derivatives utilizing the IR1 helper phage (Dente *et al.*, 1983) and from pTZ derivatives utilizing the M13KO7 helper phage as described by the manufacturer (Pharmacia). Enzymatic sequencing with Klenow fragment, AMV reverse transcriptase and Sequenase was performed essentially as described by Sanger *et al.* (1977) or by the manufacturers recommendations. For resolution of secondary structure the analogues 7 deaza 2' deoxyguanosine 5' triphosphate and deoxyinosine 5' triphoshate (dITP) were sometimes substituted for dGTP and both dATP and dCTP were used (separately) as the labelled nucleotide (Mills and Kramer, 1979; Mizusawa *et al.*, 1986). For sequencing with T7 DNA polymerase the primer was annealed to the template in 7 μL of 40 mM Tris-HCl pH7.5, 15 mM $MgCl_2$, 50 mM NaCl, 2 to 10 ng oligonucleotide primer and 0.5 - 2 μg template by heating to 60°C for 10 minutes and cooling slowly to 35°C. To the annealed template - primer was added 0.5 μL $\alpha^{32}P$ dCTP (5 μCi of 3000 Ci $mmol^{-1}$), 0.5 μL of 300 mM DTT, 1 μL of nucleotide elongation mix (2 μM dATP, 2 μM dGTP, 2 μM dTTP) and 1 μL of diluted T7 DNA polymerase (0.4 - 1.5 units). This labelling reaction was allowed to proceed for 5 minutes at room temperature before 2 μL was added to1 μL of termination mix specific for each nucleotide (40 mM Tris-HCl 7.5, 10 mM $MgCl_2$, 50 mM NaCl, 150 μM dATP,150 μM dGTP,150 μM dCTP,150 μM dTTP and 3.5 μM ddATP or ddGTP or ddCTP or ddTTP) and the mixes incubated at 37°C for 5 minutes. Reactions were stopped by addition of 3 μL of FDM and heat denatured at 90°C for 2 minutes prior to loading onto polyacrylamide sequencing gels.

Polyacrylamide sequencing gels (ratio of acrylamide to N N' methylene bisacrylamide 39:2) were composed of and treated to various combinations of the following characteristics: (i) gel length 38 cm or 65 cm; (ii) gel thickness 0.17 mm, 0.25 mm, 0.35 mm or variable wedge; (iii) acrylamide concentration 4% to 20%; (iv) buffers 1x TBE (90 mM Tris-HCl pH8.0, 90 mM boric acid, 2.5 mM EDTA), 0.5x TBE, modified TBE (135 mM Tris-HCl pH8.0, 45 mM boric acid, 2.5 mM EDTA) or buffer gradient as described by Biggin *et al.* (1983) and (v) gel temperature while electrophoresing 40°C to 80°C. After electrophoresis gels were dried onto Whatmann No.1 (for 0.17 mm thick gels) or Whatmann 3MM (for all thicker gels) and exposed to Kodak XRP-1, XAR-5 or Amersham Hyperfilm β Max.

2.12 Northern Blots

Twenty µg of total cellular RNA was denatured at 70°C for 10 minutes in 50 mM MOPS, 1 mM EDTA, 0.66 M formaldehyde and 40% v/v formamide, electrophoresed through a 0.66 M formaldehyde, 50 mM MOPS, 1 mM EDTA, 1.5% or 3% w/v agarose horizontal slab gel and transferred to nitrocellulose filters by blotting with 20x SSC. The filters were hybridized with radioactive DNA probes (generated by Klenow extension of appropriate M13 subclones of pLW173) in 5x SSC, 40% v/v formamide and 2x Denhardt's at 42°C and washed stringently in 0.2x SSC at 52°C - 59°C for 30 minutes, dried and exposed to Kodak XAR-5 film with an intensifiing screen.

2.13 Analysis of *in vivo* RNA Transcripts

The *in vivo* RNA transcripts were analysed by both S1 nuclease protection and primer extension experiments (Favaloro *et al.*, 1980; Newman, 1987). For the S1 nuclease protection of DNA by RNA, 100 Kcpm - 250 Kcpm of a 5' or 3' end labelled DNA probe fragment (500 Kcpm pmol^{-1}) was ethanol precipitated with 5 - 20 µg of total cellular RNA obtained from log phase cells, resuspended in 20 µL of 40mM PIPES pH6.8, 400 mM NaCl, 1 mM EDTA and 80% v/v formamide, denatured at 80°C for 15 minutes and hybridized for 3 hours at 64°C - 71°C. Digestion of unprotected single stranded nucleic acids by S1 nuclease was accomplished by addition of 300 µL of ice cold 280 mM NaCl, 30 mM Na acetate pH 4.4, 4.5 mM $ZnCl_2$, 20 µgmL^{-1} single stranded M13 DNA and 90 units of S1 nuclease, and incubation at 37°C for 30 minutes. Products were precipitated twice with isopropanol, resuspended in 5 µL FDM and denatured at 90°C for 2 minutes before loading alongside appropriate size standards (either 3' end labelled and MspI restricted pBR322 DNA or a chemical sequence derived from the DNA probe fragment) on polyacrylamide sequencing gels.

In addition to S1 nuclease experiments, primer extension,by AMV reverse transcriptase of DNA oligonucleotides annealed to RNA was used to localize the 5' ends of *in vivo* RNA transcripts. One ng of 5' end labelled oligonucleotide (100 Kcpm pmol^{-1}) was annealed to 5 - 20 µg of total cellular RNA in 10 µL of 80 mM KCl, 20 mM Tris-HCl pH8.5 and 0.5 mM EDTA by heating to 65°C for 5 minutes, cooling slowly to 37°C and incubating at 37°C for 1 hour. Extension was accomplished by addition of 10 µL of 10 mM $MgCl_2$, 1 mM of each deoxyribonucleotide triphosphate, 10 mM 2 mercaptoethanol, 5 units RNase inhibitor and 5 units AMV reverse transcriptase and incubation for a further hour at 37°C. The products were ethanol precipitated twice, redissolved in 5 µL FDM and denatured by heating at 90°C for 2 minutes before loading on a polyacrylamide sequencing gel alongside a sequencing ladder generated from extension of the labelled oligonucleotide and an appropriate single stranded template.

2.14 Nomenclature

The literature on the ribosomal proteins presents a chaotic set of conflicting systems of nomenclature and with the advent of rapid sequencing of both proteins and nucleic acids the situation will likely deteriorate further. The system of nomenclature used in this work is based upon each protein having a three letter organism identifier followed by an alphanumeric protein identifier as follows:

1) Organism identifier - organisms are identified by the first letter of the genus and first two letters of the species: *Spinacea oleracea* is identified as Sol. Chloroplasts and mitochondria are identified by a {c} or {m} following the species identifier, thus: *Spinacea oleracea* chloroplast is identified as Sol{c}.

2) Protein identifiers - proteins may have any or all of A) experimental, B) homology and C) phylogenetic identifiers:

   A) Experimental - proteins can be identified by a designation based on a defined (published) isolation procedure. Usually the standard alphanumeric system based on two dimensional gel electrophoresis of the large and small subunit ribosomal proteins used for *E. coli* is utilized, i.e. 'L' or 'S' indicating large or small ribosomal subunit respectively followed by a number indicating the relative position in order of decreasing molecular mass. Thus the twentieth largest protein of the large ribosomal subunit of *Halobacterium cutirubrum* is designated Hcu L20.

   B) Homology - proteins from archaebacteria and eucaryota that have been found to be homologous to proteins from *E. coli* are identified with their *E. coli* homologue protein identifier with an 'e' (for equivalent) appended: Hcu L20 is homologous to Eco L12 and thus Hcu L20 = Hcu L12e.

   C) Phylogenetic - proteins found to be present in all urkingdoms and thus present in the progenote carry the identifier 'P' (for progenote) and an (arbitrary) numerical identifier. Thus the *E. coli* proteins L11, L1, L10 and L12 which have extant homologues in all kingdoms would be identified as Eco P1, Eco P2, Eco P3 and Eco P4 respectively. Similarly Hcu L20 = Hcu L12e = Hcu P4. The proteins unique to a single or a pair of urkingdoms would be identified with 'U', 'A' or 'E' (or pairs of letters, e.g. 'UA') as a eucaryotic, archaebacterial or eubacterial designator followed by an (arbitrary) numerical identifier.

In this work the species designations and the correspondence between the experimental and homology protein identifiers for the GTPase domain proteins appears in Table 3 and the phylogenetic status identifiers are not used as complete sets of archaebacterial and eucaryotic ribosomal proteins have yet to be sequenced.

2.15 Statistical Analysis

Sequence similarity searches of the NBRF PIR protein and GENBANK data bases for proteins homologous to the peptides encoded by the clone pLW173 were done by the FASTP and FASTA protein alignment programs (Lipman and Pearson, 1985; Pearson, 1990). The alignments of the ribosomal proteins were based on all available sequences (Table 3) although not all sequences are illustrated in the alignment figures. The alignments for the L11e and L1e ribosomal proteins (Figures 15, 16) were based on the sequence similarity alignment given by the FASTP protein alignment program and optimized by manually maximizing the amino acid identities. Precise placement of gaps was decided by maximizing conservative substitutions at the amino acid level. The alignment of the L10e proteins from positions 1 - 218 (Figure 17) was based solely on sequence similarity. In addition to sequence similarity, in some cases known or hypothesized structure - function relationships were utilized for the alignment of the L10e proteins for position 219 - 372 (Figure 17) and for the L12e proteins over their entire length (Figure 18). It is impossible to state explicitly the relative importance of sequence similarity versus structure - function for these alignments. For example, the alanine - proline rich region in the *E.coli* L12 protein is believed to function as a flexible hinge between the amino terminus (which binds L12 to L10) and the carboxy domain (which binds translation factors); the alanine - proline rich regions in the archaebacterial and eucaryotic L10e proteins have therefore been aligned on the hypothesis that they serve a similar function. The merit of the alignment must therefore be considered both within a structure - function as well as a sequence similarity context.

Table 3 Nomenclature of the GTPase Domain Proteins

Protein	Organism	Urkingdom[2]	Original	Reference
Designation[1]			Nomenclature	
L11e				
Eco L11	*Escherichia coli*	E	L11	Post *et al.*, 1979
Pvu L11e	*Proteus vulgaris*	E	L11	Sor and Nomura, 1987
Sma L11e	*Serratia marscescens*	E	L11	Sor and Nomura, 1987
Hcu L11e	*Halobacterium cutirubrum*	A		Shimmin and Dennis, 1989
Sso L11e	*Sulfolobus solfataricus*	A		Shimmin *et al.*, 1989
Sce L11e	*Saccharomyces cerevisiae*	U	L15	Otaka *et al.*, 1984
L1e				
Bst L1e	*Bacillus stearothermophilis*	E	L1	Kimura *et al.*, 1985
Eco L1	*Escherichia coli*	E	L1	Post *et al.*, 1979
Pvu L1e	*Proteus vulgaris*	E	L1	Sor and Nomura, 1987
Sma L1e	*Serratia marscescens*	E	L1	Sor and Nomura, 1987
Hcu L1e	*Halobacterium cutirubrum*	A		Shimmin and Dennis, 1989
Hha L1e	*Halobacterium halobium*	A	ORF A	Itoh, 1988
Sso L1e	*Sulfolobus solfataricus*	A		Shimmin *et al.*, 1989
L10e				
Eco L10	*Escherichia coli*	E	L10	Post *et al.*, 1979
Hcu L10e	*Halobacterium cutirubrum*	A		Shimmin and Dennis, 1989
Hha L10e	*Halobacterium halobium*	A	ORF B	Itoh, 1988
Sso L10e	*Sulfolobus solfataricus*	A		Shimmin *et al.*, 1989
Hsa L10e	*Homo sapiens*	U	P0	Rich and Steitz, 1987
Sce L10e	*Saccharomyces cerevisiae*	U	A0	Mitsui and Tsurugi, 1988a
			L10e	Newton *et al.*, 1990
L12e				
Bst L12e	*Bacillus stearothermophilis*	E	BL13	Garland *et al.*, 1987
Bsu L12e	*Bacillus subtilis*	E	BL9	Itoh and Wittmann, 1978
Eco L12	*Escherichia coli*	E	L7/12	Terhorst *et al.*,1973
Han L12e	*Haloanaerobium prevalens*	E	A-protein	Matheson *et al.*, 1987
Mly L12e	*Micrococcus lysodeikticus*	E	MA1/2	Itoh, 1981a
Rsp L12e	*Rhodopseudomonas spheroides*	E	RA1	Itoh and Higo, 1983
Sgr L12e	*Streptomyces griseus*	E	SA1	Itoh *et al.*, 1982
Sol{c} L12e	*Spinacea oleracea* {chloroplast}	E	L12	Bartsch *et al.*, 1982
41227 L12e	NRCC 41227	E	L12	Falkenberg *et al.*, 1985
Hcu L12e	*Halobacterium cutirubrum*	A		Shimmin and Dennis, 1989
Hha L12e	*Halobacterium halobium*	A	'A' protein	Itoh, 1988
Mva L12e	*Methanococcus vannielli*	A	L12	Strobel *et al.*, 1988
Sac L12e	*Sulfolobus acidocaldarius*	A	L12	Matheson *et al.*, 1988
Sso L12e	*Sulfolobus solfataricus*	A		Shimmin *et al.*, 1989

Table 3 (Continued)

Protein Designation[1]	Organism	Urkingdom[2]	Original Nomenclature	Reference
Asa L12eI	*Artemia salina*	U	eL12	Amons *et al.*, 1979
Dme L12eI	*Drosophila melanogaster*	U	rp1	Qain *et al.*, 1987
Hsa L12eI	*Homo sapiens*	U	P2	Rich and Steitz, 1987
Rra L12eI	*Rattus rattus*	U	P2	Lin *et al.*, 1982
Spo L12eI	*Shizosaccharomyces pombe*	U	SP-40C	Beltrame and Bianche, 1987
Sce L12eIA	*Saccharomyces cerevisiae*	U	L45	Remacha *et al.*, 1988
			YPA1	Itoh, 1981b
			L12eIA	Newton *et al.*, 1990
Sce L12eIB	*Saccharomyces cerevisiae*	U	L44	Remacha *et al.*, 1988
			L12eIB	Newton *et al.*, 1990
			A2	Mitsui and Tsurugi, 1988c
Asa L12eII	*Artemia salina*	U	eL12'	Amons *et al.*, 1982
Dme L12eII	*Drosophila melanogaster*	U	rp21C	Wigboldus, 1987
Hsa L12eII	*Homo sapiens*	U	P1	Rich and Steitz, 1987
Sce L12eIIA	*Saccharomyces cerevisiae*	U	A1	Mitsui and Tsurugi, 1988b
			L12eIIA	Newton *et al.*, 1990
Sce L12eIIB	*Saccharomyces cerevisiae*	U	L44'	Remacha *et al.*, 1988
			L12eIIB	Newton *et al.*, 1990

(1) The Sce L11e is a partial protein sequence; all others are complete protein and/or nucleotide sequences.

(2) Urkingdom abbreviations are: A, archaebacteria; E, eubacteria and U, eucaryota. Note that the chloroplast of the eucaryote *Spinacea oleracea* is a member of the eubacteria.

The RDF program of Lipman and Pearson (1985) was used to determine the statistical significance of the interkingdom and intrakingdom alignments (as presented in Table 7) and the existence of the 26 amino acid module (Figure 20). Significance values (z) of greater than or equal to 10 indicate homologous proteins, whereas values of 6 to 10, 3 to 6 and less than 3 indicate homology that is probable, possible and unlikely respectively. Although the shortness and great divergence of the modules precludes any single module having a highly significant match to any other module, the modules as a group have a statistically highly significant match. To establish this, two hypothetical proteins of 194 amino acids were constructed from the tandem L10e modules such that a linear comparison of the two proteins yielded all potential intraspecies module pairings, that is:

Protein 1	Eco β'	Hcu α	β	γ	Sso α	β	γ	Sce β
	•	•	•	•	•	•	•	•
Protein 2	Eco γ	Hcu β	γ	α	Sso β	γ	α	Sce γ

The actual match score was manually calculated from the PAM 250 matrix of Dayhoff (1978). Simulated random match scores for each artificial protein versus jumbled versions of the second artificial protein were generated with the RDF program. The significance (z) of the overall module match was calculated by subtracting the random match value from the actual match value and dividing by the standard deviation of the randomized match values.

Part 3: Organization and Expression of the GTPase Domain Genes

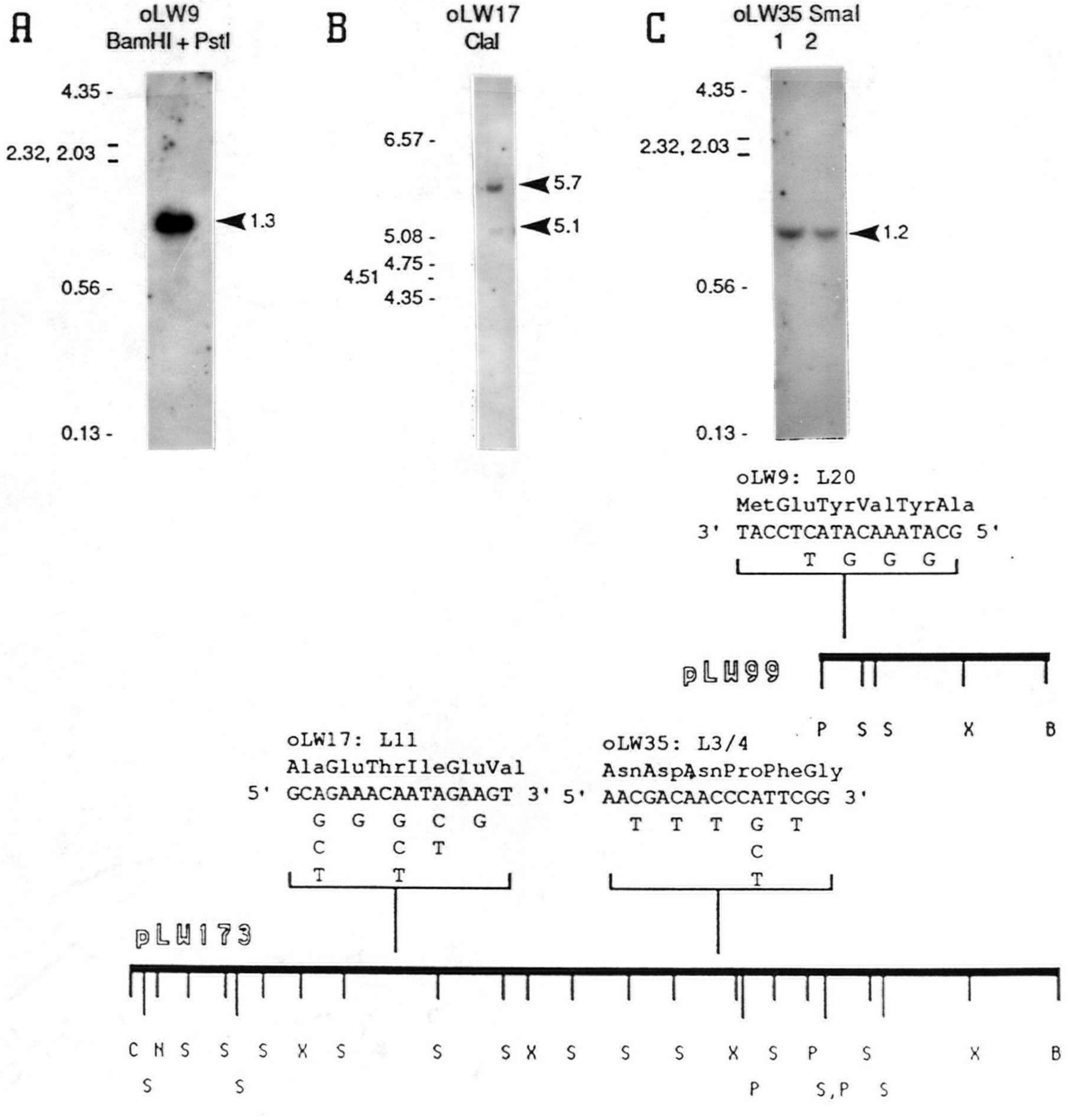
3.1 Isolation of the Genomic Clones pLW99 and pLW173

The amino terminal sequence of the L20 ribosomal protein of *H. cutirubrum* is Met-Glu-Tyr-Val-Tyr-Ala (Oda *et al.*, 1974). A 17 nucleotide long synthetic oligonucleotide mixture complementary to all 16 DNA sequences encoding this hexapeptide was prepared (oLW9) and used to probe restriction enzyme digests of *H. cutirubrum* genomic DNA (Figure 4). The L20 specific oligonucleotide mix (oLW9) hybridized to a 1.3 Kilobasepair PstI - BamHI fragment. Using the oligonucleotide mix as probe, genomic DNA was size fractionated, cloned into the multiple cloning site of plasmid pUC8 and the resulting library was screened by hybridization to oLW9 after being transformed into *E. coli* JM83 for propagation. Two of the positive clones (pLW99 and pLW102) were sequenced and the gene encoding the L20 protein was identified. Sequence analysis indicated an open reading frame extending upstream of the Hcu L20 gene. Northern hybridization of genomic RNA with a plasmid (pLW145) containing the entire Hcu L20 gene indicated a transcript of approximately 2200 nucleotides (data not shown). Nuclease S1 protection experiments using total RNA indicated that the 5' transcript end was located in the 5' flanking region of the Hcu L20 gene upstream of the PstI site and the 3' transcript end was localized to about 20 nucleotides beyond the Hcu L20 coding region (Dennis *et al.*, 1985). Attempts to isolate the upstream region as large fragments in phage (λ1059, EMBL3), cosmid (pJB8) and plasmids (pUC, pBR322, pACYC184) or by short fragment 'walks' utilizing restriction enzymes recognizing 4 basepair sites were unsuccessful. The upstream region was eventually cloned utilizing an oligonucleotide probe to the Hcu L11 gene.

The amino terminal sequence of the L11 ribosomal protein of *H. cutirubrum* is Ala-Glu-Thr-Ile-Glu-Val (Matheson *et al.*, 1984). A mixture of 192 17 nucleotide long synthetic oligonucleotides complementary to all DNA sequences encoding this hexapeptide was prepared (oLW17) and used to probe restriction enzyme digests of *H. cutirubrum* genomic DNA (Figure 4). The L11 specific oligonucleotide mix (oLW17) hybridized strongly to a 5.7 Kilobasepair ClaI - BamHI fragment and weakly to a 5.1 Kilobasepair ClaI - BamHI fragment. Using the oligonucleotide mix as probe, genomic DNA was fractionated (separating the 5.7 and 5.1 Kilobasepair fragments) and the DNA fractions were ligated between the ClaI and BamHI sites

Figure 4 Southern Blot Analysis and Restriction Maps of pLW99 and pLW173

Genomic *Halobacterium cutirubrum* DNA was digested with various restriction enzymes as indicated and hybridized to oligonucleotides (A) oLW9, specific for Hcu L20; (B) oLW17, specific for Hcu L11 and (C) oLW 35, specific for the Hcu L3 / 4 gene. The arrows indicate in (A) a 1.3 Kilobasepair PstI - BamHI fragment that was subsequently isolated as the clones pLW99 and pLW102; (B) the 5.1 Kilobasepair ClaI - BamHI fragment subsequently isolated in the vector pBR322 as the clones pLW173 and pLW180 (the 5.7 Kilobasepair fragment has a fortuitous match to oLW17) and (C) a 1.2 Kilobasepair SmaI (= XmaI) fragment (lane 1: pLW173, lane 2: genomic DNA). The size standards for A and C are HindIII restricted λ DNA, for B a mixture of HindIII and PstI restricted λ DNA. Illustrated below are the oligonucleotide sequences and orientations with the amino acid sequence from which they were derived and their positions of hybridization on restriction maps of pLW99 and pLW173. The restriction sites indicated are: B, BamHI; C, ClaI; N, NheI; P, PstI; S, SalI; X, XmaI (= SmaI).



of the plasmid pBR322. The ligated libraries (libLW37 and libLW38) were transformed and propagated in *E. coli* JC8111. Two positive clones (pLW155 and pLW156) containing the 5.7 Kilobasepair ClaI - BamHI fragment were isolated from libLW37, the location of a perfect match to oLW17 identified and a 1 Kilobasepair region containing the oligonucleotide hybridization site was sequenced. The perfect match to oLW17 within these clones was fortuitous as there was no similarity within the sequenced region to the Hcu L11 protein other than the oligonucleotide match. Two positive clones (pLW173 and pLW180) containing the 5.1 Kilobasepair ClaI - BamHI fragment were isolated from libLW38 and the location of a perfect match to oLW17 was identified. Restriction enzyme and Southern hybridization analysis demonstrated that the smaller 1.3 Kilobasepair PstI - BamHI fragment was derived from the right hand end of the longer 5.1 Kilobasepair ClaI - BamHI fragment. In addition, a mixture of 64 17 base long synthetic oligonucleotides complementary to all DNA sequences encoding the amino acid sequence Asn-Asp-Asn-Pro-Phe-Gly contained within a tryptic peptide fragment of the Hcu L3 / 4 protein (oLW35; A.T. Matheson, personal communication) hybridized to the 5.1 Kilobasepair ClaI - BamHI fragment, indicating the presence of the Hcu L3 / 4 gene (Figure 4).

3.2 Nucleotide Sequence Analysis of pLW173

The nucleotide sequence of the entire 5146 basepair ClaI - BamHI fragment of *H. cutirubrum* genomic DNA was determined and is illustrated in Figure 5. Two clones were sequenced: pLW173, of which at least two determinations of each nucleotide of each strand, all of which agreed and pLW180, of which 95% was sequenced on both strands and was in perfect agreement with the pLW173 sequence. The fragment contained unique sequences complementary to the oLW9 (ATGGAATACGTCTACGC, positions 4018 - 4034), oLW17 (ACTTCGATCGTCTCAGC, positions 1641 - 1625) and oLW35 (CCGAAGGGGTTGTCGTT, positions 3252 - 3236) oligonucleotide probe mixtures.

The sequence fidelity of the cloned fragment contained in pLW173 were checked using Southern hybridization. The 5.1 Kilobasepair ClaI - BamHI inserts of pLW173 were radioactively labelled by nick translation and hybridized to (i) the ClaI - BamHI insert of pLW173 restricted with SalI, (ii) genomic DNA restricted with ClaI + BamHI + SalI and (iii) a mixture of the previous two restriction digests. No differences between the genomic and plasmid restriction patterns were seen (data not shown). To check the copy

Figure 5 The Nucleotide Sequence of pLW173

The complete 5' to 3' nucleotide sequence (upper line) of the 5146 basepair ClaI - BamHI fragment of *H. cutirubrum* genomic DNA is shown. The BamHI, ClaI, SalI, NheI, PstI, XmaI and two of the AvaII restriction sites are indicated. The amino acid sequence of the putative ORF protein is written in single letter code in leftward orientation below the DNA sequence (positions 1244 - 135). The deduced amino acid sequences of the rightward oriented putative NAB protein (positions 1345 - 1548) and the L11 (L11e; positions 1622 - 2110), L8 (L1e; positions 2314 - 2949), L3 / 4 (L10e; positions 2954 - 4009), and L20 (L12e; positions 4018 - 4359) ribosomal proteins are written in single letter code above the DNA sequence. The sizes (number of amino acid residues and molecular weight) and calculated isoelectric points of the proteins and sizes (nucleotide basepairs) of the genes are indicated at the initiation methionine positions.

```
        10        20        30        40        50        60        70        80        90       100       110       120
ATCGATTCCGGATACACGAACCCCACGGAAAACAGCGGATAGACGCCGTTCGACCAGTCGACGATGACCCACGGCACCACTCCCGCGGCCAGCGCCACCGCCACCACCCGGCGGCGCGCC
TAGCTAAGGCCTATGTGCTTGGGGTGCCTTTTGTCGCCTATCTGCGGCAAGCTGGTCAGCTGCTACTGGGTGCCGTGGTGAGGGCGCCGGTCGCGGTGGCGGTGGTGGGCCGCCGCGCGG
ClaI                                                     SalI

       130       140       150       160       170       180       190       200       210       220       230       240
GTCGGCATCGGCTAGCGCCGCGCAACCACGCGCAGGACGTCCTCGTCTTCGAGGACGTGATCGCGGCCCACCTGCTGGTCGTCGTGTTGGGCCGACGGGCCGGTGACCCGCGCGAACCGG
CAGCCGTAGCCGATCGCGGCGCGTTGGTGCGCGTCCTGCAGGAGCAGAAGCTCCTGCACTAGCGCCGGGTGGACGACCAGCAGCACAACCCGGCTGGCCGCCCACTGGGCGCGCTTGGCC
               R  R  A  V  V  R  L  V  D  E  D  E  L  V  H  D  R  G  V  Q  Q  D  D  H  Q  A  S  R  R  T  V  R  A  F  R
       NheI

       250       260       270       280       290       300       310       320       330       340       350       360
AACCGCTCGTCGAGGGTGCCGCCCAGCTTCTGCACGGCGTCGTCCACGGTCTCGCCGCGCCGGATGATGAGCGGCTCCTCGCGGTCGACGCCCCGCCCCGGCTTGTCCATGTAGATCCGG
TTGGCGAGCAGCTCCCACGGCGGGTCGAAGACGTGCCGCAGCAGGTGCCAGAGCGGCGCGGCCTACTACTCGCCGAGGAGCGCCAGCTGCGGGGCGGGGCCGAACAGGTACATCTAGGCC
F  R  E  D  L  T  G  G  L  K  Q  V  A  D  D  V  T  E  G  R  R  I  I  L  P  E  E  R  D  V  G  R  G  P  K  D  M  Y  I  R
                                                                                     SalI

       370       380       390       400       410       420       430       440       450       460       470       480
ATGAGCCCGAGCGCCCGCCACATCCGCTCTTTCAGCACGTCCAATCCCTTCTCCTCGGCGGCCGAAATGAAGATCGCGTCGTCGGGGCTGACGCCGTGATCGCGGAGGGCGTCCTTCATT
TACTCGGGCTCGCGGGCGGTGTAGGCGAGAAAGTCGTGCAGGTTAGGGAAGAGGAGCCGCCGGCTTTACTTCTAGCGCAGCAGCCCCGACTGCGGCACTAGCGCCTCCCGCAGGAAGTAA
I  L  G  L  A  R  W  M  R  E  K  L  V  D  L  G  K  E  E  A  A  S  I  F  I  A  D  D  P  S  V  G  H  D  R  L  A  D  K  M

       490       500       510       520       530       540       550       560       570       580       590       600
GTCCCCGCGTACGACGGCTCGATGAGGTCGACCTTGTTGACGGTGACCAGCGACGGCATGTAGACGCGGTTGTCCATCACCCCGTCGATCAGCCGGTCGACAGACGGGTTGCCACGGATC
CAGGGGCGCATGCTGCCGAGCTACTCCAGCTGGAACAACTGCCACTGGTCGCTGCCGTACATCTGCGCCAACAGGTAGTGGGGCAGCTAGTCGGCCAGCTGTCTGCCCAACGGTGCCTAG
T  G  A  Y  S  P  E  I  L  D  V  K  N  V  T  V  L  S  P  M  Y  V  R  N  D  M  V  G  D  I  L  R  D  V  S  P  N  G  R  I
                               SalI                                                              SalI

       610       620       630       640       650       660       670       680       690       700       710       720
GTCACGTTGGCGTTGATGAACCCGCGCTCGCGCAGGATTCCCTTGACCGTGTCGCTGTCCAACTCCAGCTCCCCGGACGTGTTCACGTCGATGCCGTCCTTGCCTTTCCGGCGCACGGTC
CAGTGCAACCGCAACTACTTGGGCGCGAGCGCGTCCTAAGGGAACTGGCACAGCGACAGGTTGAGGTCGAGGGGCCTGCACAAGTGCAGCTACGGCAGGAACGGAAAGGCCGCGTGCCAG
T  V  N  A  N  I  F  G  R  E  R  L  I  G  K  V  T  D  S  D  L  E  L  E  G  S  T  N  V  D  I  G  D  K  G  K  R  R  V  T

       730       740       750       760       770       780       790       800       810       820       830       840
ACCGACGGTGGCTCGGCGTCGACCCGGATGTTGACGTTGTACAGCTCCTCGGCGAGACGGTCGTACTGCTCGATCTCGAACGCCGACAGCACGAAGATCACCAGATCCGCCCCACGGATC
TGGCTGCCACCGAGCCGCAGCTGGGCCTACAACTGCAACATGTCGAGGAGCCGCTCTGCCAGCATGACGAGCTAGAGCTTGCGGCTGTCGTGCTTCTAGTGGTCTAGGCGGGGTGCCTAG
V  S  P  P  E  A  D  V  R  I  N  V  N  Y  L  E  E  A  L  R  D  Y  Q  E  I  E  F  A  S  L  V  F  I  V  L  D  A  G  R  I
                   SalI

       850       860       870       880       890       900       910       920       930       940       950       960
ACCGAGAGGATCTCTTTCCCGCCGCCGCGTCCCCCGCCGCACCCTCGATGAGCCCCGGCACGTCCAGGAGTTGGATGTTCGCGCCGCGGTACTCCAACATCCCCGGGTTCACGTTGAGG
TGGCTCTCCTAGAGAAAGGGCGGCGGCGCAGGGGGCGGCGTGGGAGCTACTCGGGGCCGTGCAGGTCCTCAACCTACAAGCGCGGCGCCATGAGGTTGTAGGGGCCCAAGTGCAACTCC
V  S  L  I  E  K  G  G  G  R  G  G  A  A  G  E  I  L  G  P  V  D  L  L  Q  I  N  A  G  R  Y  E  L  M  G  P  N  V  N  L
                                                                                                    XmaI

       970       980       990      1000      1010      1020      1030      1040      1050      1060      1070      1080
GTGGTGAACTCGTAGGCGCCGACCTCGCTGTCGGCGTTGGTCATCGCGTTGATCAGCGACGACTTCCCCACGCTGGGGAATCCAACCAGGGCCACGGTCGCGTCGCCGTGCTGTTCGACT
CACCACTTGAGCATCCGCGGCTGGAGCGACAGCCGCAACCAGTAGCGCAACTAGTCGCTGCTGAAGGGGTGCGACCCCTTAGGTTGGTCCCGGTGCCAGCGCAGCGGCACGACAAGCTGA
T  T  F  E  Y  A  G  V  E  S  D  A  N  T  M  A  N  I  L  S  S  K  G  V  S  P  F  G  V  L  A  V  T  A  D  G  H  Q  E  V

      1090      1100      1110      1120      1130      1140      1150      1160      1170      1180      1190      1200
GCGTACCCGCCGCCACCGCCGCTGCCGGACTGCTGGGCTTCGAGCTTCTCCTTTTGCTCCGCGAGCTTCGCCTTCAAGCGGCCGATGTGGGCTTCTGTCGACTTGTTGTACGGCGTGTTC
CGCATGGGCGGCGGTGGCGGCGACGGCCTGACGACCCGAAGCTCGAAGAGGAAAACGAGGCGCTCGAAGCGGAAGTTCGCCGGCTACACCCGAAGACAGCTGAACAACATGCCGCACAAG
A  Y  G  G  G  G  G  S  G  S  Q  Q  A  E  L  K  E  K  Q  E  A  V  N  A  K  L  R  G  I  H  A  E  T  S  K  N  Y  P  T  N
                                                                                                 SalI

      1210      1220      1230      1240      1250      1260      1270      1280      1290      1300      1310      1320
GCGATTTCTTCTTCGAGCGATTCGATGTCCTCCTCGAGCCCCATCACGTGCCCGTAGACAGCAAGCGCCGAAAAGCGCTTTCCTCCGCAGTCCGGTGGGATCGTTTCGACACGTTAATAC
CGCTAAAGAAGAAGCTCGCTAAGCTACAGGAGGAGCTCGGGGTAGTGCACGGGCATCTGTCGTTCGCGGCTTTTCGCGAAAGGAGGCGTCAGGCCACCCTAGCAAAGCTGTGCAATTATG
A  I  E  E  E  L  S  E  I  D  E  E  L  G  M
     ORF  370 aa / 1110 bp / MW 40499 dal / pI 6.8

                        MAB  68 aa / 204 bp / MW 7530 dal / pI 3.8                                      AvaII
                        M  G  D  P  A  A  Y  R  D  S  T  Q  I  V  L  P  V  G  T  L  E  D  I  E  V  D  L  E  A  E  F  M
GCCGAGTGAAGCCATCGCATAGTGATGGGTGACCCTGCTGCGTACCGCGACAGCACGCAGATCGTGCTCCCAGTGGGGACGCTGGAGGACATCGAGGTGGACCTCGAAGCCGAGTTCATG
CGGCTCACTTCGGTAGCGTATCACTACCCACTGGGACGACGCATGGCGCTGTCGTGCGTCTAGCACGAGGGTCACCCCTGCGACCTCCTGTAGCTCCACCTGGAGCTTCGGCTCAAGTAC
      1330      1340      1350      1360      1370      1380      1390      1400      1410      1420      1430      1440

 V  S  V  F  R  P  T  D  A  E  I  V  R  I  I  G  S  P  V  V  I  K  E  V  T  E  F  L  T  R  H  G  V  H  M  P
GTCTCAGTGTTCCGGCCGACCGACGCGGAGATCGTCCGCATCATCGGGAGTCCGGTCGTGATCAAGGAGGTCACCGAGTTCCTCACGCGCCACGGCGTCCACATGCCGTGAGCGATTCGA
CAGAGTCACAAGGCCGGCTGGCTGCGCCTCTAGCAGGCGTAGTAGCCCTCAGGCCAGCACTAGTTCCTCCAGTGGCTCAAGGAGTGCGCGGTGCCGCAGGTGTACGGCACTCGCTAAGCT
      1450      1460      1470      1480      1490      1500      1510      1520      1530      1540      1550      1560

                                                                              Llle  163 aa / 489 bp / MW 17020 dal / pI 2.7
                                                              M  A  E  T  I  E  V  L  V  A  G  G  Q  A  D  P  G  P  P  L
TCCGCGGCGGCGCCCCGCTCGAAAGACAAGGGTTAAACCCGCGGCGGCGGTTTCTCGGAGTATGGCTGAGACGATCGAAGTGCTCGTTGCCGGTGGGCAGGCCGACCCTGGCCCGCCCCT
AGGCGCCGCCGCGGGGCGAGCTTTCTGTTCCCAATTTGGGCGCCGCCGCCAAAGAGCCTCATACCGACTCTGCTAGCTTCACGAGCAACGGCCACCCGTCCGGCTGGGACCGGGCGGGGA
      1570      1580      1590      1600      1610      1620      1630      1640      1650      1660      1670      1680
```

```
AvaII               SalI
 G  P  E  L  G  P  T  P  V  D  V  Q  A  V  V  Q  E  I  N  D  Q  T  E  A  F  D  G  T  E  V  P  V  T  I  E  Y  E  D  D  G
CGGTCCCGAGCTCGGTCCCACGCCGGTCGACGTACAGGCGGTCGTCCAGGAGATCAACGACCAGACCGAAGCGTTCGACGGGACGGAGGTCCCGGTCACCATCGAATACGAGGACGACGG
GCCAGGGCTCGAGCCAGGGTGCGGCCAGCTGCATGTCCGCCAGCAGGTCCTCTAGTTGCTGGTCTGGCTTCGCAAGCTGCCCTGCCTCCAGGGCCAGTGGTAGCTTATGCTCCTGCTGCC
      1690      1700      1710      1720      1730      1740      1750      1760      1770      1780      1790      1800

 S  F  S  I  E  V  G  V  P  P  T  A  A  L  V  K  D  E  A  G  F  D  T  G  S  G  E  P  Q  E  N  F  V  A  D  L  S  I  E  Q
CTCGTTCTCCATCGAAGTCGGTGTTCCGCCGACGGCCGCGCTGGTGAAAGACGAAGCTGGCTTCGACACGGGCTCCGGAGAGCCCCAGGAGAACTTCGTCGCGGACCTCTCCATCGAACA
GAGCAAGAGGTAGCTTCAGCCACAAGGCGGCTGCCGGCGCGACCACTTTCTGCTTCGACCGAAGCTGTGCCCGAGGCCTCTCGGGGTCCTCTTGAAGCAGCGCCTGGAGAGGTAGCTTGT
      1810      1820      1830      1840      1850      1860      1870      1880      1890      1900      1910      1920

 L  K  T  I  A  E  Q  K  K  P  D  L  L  A  Y  D  A  R  N  A  A  K  E  V  A  G  T  C  A  S  L  G  V  T  I  E  G  E  D  A
GCTCAAAACCATCGCCGAGCAGAAGAAACCCGACCTCCTGGCGTACGACGCGCGGAACGCCGCCAAAGAGGTCGCGGGGACGTGTGCGTCCCTCGGCGTCACCATCGAAGGCGAGGACGC
CGAGTTTTGGTAGCGGCTCGTCTTCTTTGGGCTGGAGGACCGCATGCTGCGCGCCTTGCGGCGGTTTCTCCAGCGCCCCTGCACACGCAGGGAGCCGCAGTGGTAGCTTCCGCTCCTGCG
      1930      1940      1950      1960      1970      1980      1990      2000      2010      2020      2030      2040

                  SalI
 A  T  F  N  E  R  V  D  D  G  D  Y  D  D  V  L  G  D  E  L  A  A  A
CCGCACGTTCAACGAGCGCGTCGACGACGGCGACTACGACGACGTGCTCGGCGACGAACTCGCGGCCGCGTAACGCCGCCCGAGGAGTTTCTGCGCCGTTCGGTTCGCGTACTCGATAGC
GGCGTGCAAGTTGCTCGCGCAGCTGCTGCCGCTGATGCTGCTGCACGAGCCGCTGCTTGAGCGCCGGCGCATTGCGGCGGGCTCCTCAAAGACGCGGCAAGCCAAGCGCATGAGCTATCG
      2050      2060      2070      2080      2090      2100      2110      2120      2130      2140      2150      2160

                                                      XmaI
GGCGTGTGTCCGCGGGTCGCGCTCCCACGCTTGCTTCGCTTCGACGCTTTTAAGCCCGGGATCACCGTCTGTAGAACCGAGACAGGCTTCGCCTGTTTCACTGACCCGTAGGAGATCCGA
CCGCACACAGGCGCCCAGCGCGAGGGTGCGAACGAAGCGAAGCTGCGAAAATTCGGGCCCTAGTGGCAGACATCTTGGCTCTGTCCGAAGCGGACCAAAGTGACTGGGCATCTCTAGGCT
      2170      2180      2190      2200      2210      2220      2230      2240      2250      2260      2270      2280

                                  L1e  212 aa / 636 bp / MW 23095 dal / pI 6.7
                                   M  A  D  N  D  I  E  E  A  V  A  R  A  L  E  D  A  P  Q  R  N  F  R  E  T  V  D  L  A
CCTTCAGAGGGTCACCCACTACGGAGGTGAAAGATGGCAGACAACGATATAGAAGAGGCCGTAGCTCGCGCACTTGAGGATGCCCCACAGCGGAACTTCCGTGAGACGGTAGACCTCGCA
GGAAGTCTCCCAGTGGGTGATGCCTCCACTTTCTACCGTCTGTTGCTATATCTTCTCCGGCATCGAGCGCGTGAACTCCTACGGGGTGTCGCCTTGAAGGCACTCTGCCATCTGGAGCGT
      2290      2300      2310      2320      2330      2340      2350      2360      2370      2380      2390      2400

                                           SalI
 V  N  L  R  D  L  D  L  N  D  P  S  Q  R  V  D  E  G  V  V  L  P  S  G  T  G  Q  E  T  Q  I  V  V  F  A  D  G  E  T  A
GTCAACCTGCGCGACCTCGACCTCAACGACCCGTCGCAACGAGTCGACGAGGGCGTCGTGCTGCCGTCGGGCACCGGACAGGAGACGCAGATCGTGGTTTTCGCAGACGGCGAAACCGCG
CAGTTGGACGCGCTGGAGCTGGAGTTGCTGGGCAGCGTTGCTCAGCTGCTCCCGCACCACGACGGCAGCCCGTGGCCTGTCCTCTGCGTCTAGCACCAAAAGCGTCTGCCGCTTTGGCGC
      2410      2420      2430      2440      2450      2460      2470      2480      2490      2500      2510      2520

 V  R  A  D  D  V  A  D  D  V  L  D  E  D  D  L  S  D  L  A  D  D  T  D  A  A  K  D  L  A  D  E  T  D  F  F  V  A  E  A
GTTCGCGCGGACGACGTCGCTGACGACGTCCTCGACGAGGACGACCTCAGCGACCTCGCAGACGACACCGACGCCGCGAAGGATCTCGCAGACGAGACGGACTTCTTCGTGGCGGAAGCA
CAAGCGCGCCTGCTGCAGCGACTGCTGCAGGAGCTGCTCCTGCTGGAGTCGCTGGAGCGTCTGCTGTGGCTGCGGCGCTTCCTAGAGCGTCTGCTCTGCCTGAAGAAGCACCGCCTTCGT
      2530      2540      2550      2560      2570      2580      2590      2600      2610      2620      2630      2640

                                                                                                     SalI
 P  M  M  Q  D  I  V  G  A  L  G  Q  V  L  G  P  A  G  K  M  P  T  P  L  Q  P  D  D  D  V  V  D  T  V  N  R  M  K  N  T
CCCATGATGCAGGACATCGTGGGTGCGCTCGGTCAAGTGCTTGGTCCGCGCGGGAAAATGCCGACCCCGCTCCAGCCCGACGACGACGTCGTCGACACAGTCAACCGCATGAAAAACACC
GGGTACTACGTCCTGTAGCACCCACGCGAGCCAGTTCACGAACCAGGCGCGCCCTTTTACGGCTGGGGCGAGGTCGGGCTGCTGCTGCAGCAGCTGTGTCAGTTGGCGTACTTTTTGTGG
      2650      2660      2670      2680      2590      2700      2710      2720      2730      2740      2750      2760

 V  Q  I  R  S  R  D  R  A  T  F  H  T  R  V  G  A  E  D  M  S  A  E  D  I  A  S  N  I  D  V  I  M  R  A  L  H  A  N  L
GTGCAGATCCGCAGCCGCGACCGCCGCACGTTCCACACGCGCGTCGGCGCGGAGGACATGTCCGCCGAGGACATCGCCAGCAACATCGACGTCATCATGCGTCGGCTGCACGCGAACCTC
CACGTCTAGGCGTCGGCGCTGGCGGCGTGCAAGGTGTGCGCGCAGCCGCGCCTCCTGTACAGGCGGCTCCTGTAGCGGTCGTTGTAGCTGCAGTAGTACGCAGCCGACGTGCGCTTGGAG
      2770      2780      2790      2800      2810      2820      2830      2840      2850      2860      2870      2880

                                                                          L10e  352 aa / 1056 bp / MW 37159 dal / pI 2.9
 E  K  G  P  L  N  V  D  S  V  Y  V  K  T  T  M  G  P  A  V  E  V  A     M  S  A  E  E  Q  R  T  T  E  E  V  P  E·W  K
GAAAAAGGCCCGCTGAACGTGGACTCCGTCTACGTGAAGACAACGATGGGGCCTGCCGTGGAGGTTGCCTAGGATGTCCGCCGAAGAACAACGCACCACCGAGGAGGTTCCCGAGTGGAA
CTTTTTCCGGGCGACTTGCACCTGAGGCAGATGCACTTCTGTTGCTACCCCGGACGGCACCTCCAACGGATCCTACAGGCGGCTTCTTGTTGCGTGGTGGCTCCTCCAAGGGCTCACCTT
      2890      2900      2910      2920      2930      2940      2950      2960      2970      2980      2990      3000

                     SalI
 R  Q  E  V  A  E  L  V  D  L  L  E  T  Y  D  S  V  G  V  V  N  V  T  G  I  P  S  K  Q  L  Q  D  M  A  A  G  L  H  G  Q
GCGACAAGAGGTCGCCGAACTCGTCGACCTCCTGGAGACGTACGACAGCGTCGGCGTGGTGAACGTCACGGGCATTCCGAGCAAGCAGCTCCAGGACATGGCCGCGGCCTGCACGGGCA
CGCTGTTCTCCAGCGGCTTGAGCAGCTGGAGGACCTCTGCATGCTGTCGCAGCCGCACCACTTGCAGTGCCCGTAAGGCTCGTTCGTCGAGGTCCTGTACGCGGCGCCGGACGTGCCCGT
      3010      3020      3030      3040      3050      3060      3070      3080      3090      3100      3110      3120

 A  A  V  A  M  S  A  N  T  L  L  V  R  A  L  E  E  A  G  D  G  L  D  T  L  T  E  Y  V  E  G  E  V  G  L  V  A  T  N  D
GGCGGCCGTGCGCATGAGCCGGAACACCCTGTTGGTTCGCGCGCTCGAAGAAGCAGGAGACGGCCTCGACACGCTCACCGAGTACGTCGAGGGCGAAGTCGGCCTGGTCGCGACCAACGA
CCGCCGGCACGCGTACTCGGCCTTGTGGGACAACCAAGCGCGCGAGCTTCTTCGTCCTCTGCCGGAGCTGTGCGAGTGGCTCATGCAGCTCCCGCTTCAGCCGGACCAGCGCTGGTTGCT
      3130      3140      3150      3160      3170      3180      3190      3200      3210      3220      3230      3240

                                                                                                                  XmaI
 N  P  F  G  L  Y  Q  Q  L  E  N  S  K  T  P  A  P  I  N  A  G  E  V  A  P  N  D  I  V  V  P  E  G  D  T  G  I  D  P  G
CAACCCCTTCGGGCTCTACCAGCAGCTTGAGAACTCGAAGACGCCGGCCCCGATCAACGCCGGCGAGGTCGCGCCCAACGACATCGTCGTGCCGGAAGGTGACACGGGCATCGACCCGGG
GTTGGGGAAGCCCGAGATGGTCGTCGAACTCTTGAGCTTCTCGCGCCGGGGCTAGTTGCGGCCGCTCCAGCGCGGGTTGCTGTAGCAGCACGGCCTTCCACTGTGCCCGTAGCTGGGCCC
      3250      3260      3270      3280      3290      3300      3310      3320      3330      3340      3350      3360
```

```
                PstI
  P  F  V  G  E  L  Q  T  I  G  A  N  A  R  I  Q  E  G  S  I  Q  V  L  D  D  S  V  V  T  E  E  G  E  T  V  S  D  D  V  S
GCCGTTCGTCGGCGAACTGCAGACCATCGGCGCGAACGCGCGCATCCAGGAGGGCTCCATCCAGGTGCTCGATGACTCCGTCGTCACCGAGGAAGGTGAGACGGTCTCCGACGACGTCTC
CGGCAAGCAGCCGCTTGACGTCTGGTAGCCGCGCTTGCGCGCGTAGGTCCTCCCGAGGTAGGTCCACGAGCTACTGAGGCAGCAGTGGCTCCTTCCACTCTGCCAGAGGCTGCTGCAGAG
      3370      3380      3390      3400      3410      3420      3430      3440      3450      3460      3470      3480

                                                                                                         SalI
  N  V  L  S  E  L  G  I  E  P  K  E  V  G  L  D  L  R  G  V  F  S  E  G  V  L  F  T  P  E  E  L  E  I  D  V  D  E  Y  R
CAACGTCCTCTCGGAGCTCGGCATCGAGCCCAAGGAGGTCGGCCTGGACCTGCGCGGCGTGTTCTCCGAGGGCGTGCTGTTCACGCCCGAGGAGCTGGAGATCGACGTCGACGAGTACCG
GTTGCAGGAGAGCCTCGAGCCGTAGCTCGGGTTCCTCCAGCCGGACCTGGACGCGCCGCACAAGAGGCTCCCGCACGACAAGTGCGGGCTCCTCGACCTCTAGCTGCAGCTGCTCATGGC
      3490      3500      3510      3520      3530      3540      3550      3560      3570      3580      3590      3600

  A  D  I  Q  S  A  A  A  S  A  R  N  L  S  V  N  A  A  Y  P  T  E  R  T  A  P  D  L  I  A  K  G  R  G  E  A  K  S  L  G
CGCGGACATCCAGTCCGCGGCGGCGTCGGCGCGCAACCTCTCGGTCAACGCAGCGTACCCGACCGAGCGGACCGCACCGGACCTCATCGCGAAGGGCCGCGGCGAGGCGAAGAGCCTCGG
GCGCCTGTAGGTCAGGCGCCGCCGCAGCCGCGCGTTGGAGAGCCAGTTGCGTCGCATGGGCTGGCTCGCCTGGCGTGGCCTGGAGTAGCGCTTCCCGGCGCCGCTCCGCTTCTCGGAGCC
      3610      3620      3630      3640      3650      3660      3670      3680      3690      3700      3710      3720

PstI                                                                                                                PstI
  L  Q  A  S  V  E  S  P  D  L  A  D  D  L  V  S  K  A  D  A  Q  V  R  A  L  A  A  Q  I  D  D  E  D  A  L  P  E  E  L  Q
CCTGCAGGCCAGCGTCGAGAGTCCGGACCTCGCGGACGATCTCGTGAGCAAGGCCGACGCCCAGGTGCGGGCGCTCGCCGCGCAGATCGACGACGAGGACGCCCTCCCGGAGGAACTGCA
GGACGTCCGGTCGCAGCTCTCAGGCCTGGAGCGCCTGCTAGAGCACTCGTTCCGGCTGCGGGTCCACGCCCGCGAGCGGCGCGTCTAGCTGCTGCTCCTGCGGGAGGGCCTCCTTGACGT
      3730      3740      3750      3760      3770      3780      3790      3800      3810      3820      3830      3840

   SalI
  D  V  D  A  P  A  A  P  A  G  G  E  A  D  T  T  A  D  E  Q  S  D  E  T  Q  A  S  E  A  D  D  A  D  D  S  D  D  D  D  D
GGACGTCGACGCGCCTGCGGCGCCTGCCGGCGGCGAAGCGGACACCACTGCGGACGAACAGAGCGACGAAACACAAGCGTCCGAGGCTGACGACGCCGACGATTCCGACGACGATGACGA
CCTGCAGCTGCGCGGACGCCGCGGACGGCCGCCGCTTCGCCTGTGGTGACGCCTGCTTGTCTCGCTGCTTTGTGTTCGCAGGCTCCGACTGCTGCGGCTGCTAAGGCTGCTGCTACTGCT
      3850      3860      3870      3880      3890      3900      3910      3920      3930      3940      3950      3960

                                                          L12e  114 aa / 346 bp / MW 11550 dal / pI 2.1
  D  D  D  G  N  A  G  A  E  G  L  G  E  M  F  G          M  E  Y  V  Y  A  A  L  I  L  N  E  A  D  E  E  L  T  E  D  N
CGACGACGACGGGAACGCTGGCGCCGAGGGCCTCGGGGAGATGTTCGGATAATAACAATGGAATACGTCTACGCAGCACTCATCCTGAACGAGGCTGACGAAGAGCTGACCGAAGACAAC
GCTGCTGCTGCCCTTGCGACCGCGGCTCCCGGAGCCCCTCTACAAGCCTATTATTGTTACCTTATGCAGATGCGTCGTGAGTAGGACTTGCTCCGACTGCTTCTCGACTGGCTTCTGTTG
      3970      3980      3990      4000      4010      4020      4030      4040      4050      4060      4070      4080

                            SalI                                                        SalI
 I  T  G  V  L  E  A  A  G  V  D  V  E  E  S  R  A  K  A  L  V  A  A  L  E  D  V  D  I  E  E  A  V  E  E  A  A  A  A  P
ATCACCGGCGTCCTGGAGGCCGCCGGCGTCGACGTCGAGGAATCCCGCGCGAAGGCCCTCGTGGCCGCGCTGGAGGACGTCGACATCGAGGAAGCCGTCGAAGAGGCCGCCGCCGCGCCT
TAGTGGCCGCAGGACCTCCGGCGGCCGCAGCTGCAGCTCCTTAGGGCGCGCTTCCGGGAGCACCGGCGCGACCTCCTGCAGCTGTAGCTCCTTCGGCAGCTTCTCCGGCGGCGGCGCGGA
      4090      4100      4110      4120      4130      4140      4150      4160      4170      4180      4190      4200

 A  A  A  P  A  A  S  G  S  D  D  E  A  A  A  D  D  G  D  D  D  E  E  A  D  A  D  E  A  A  E  A  E  D  A  G  D  D  D  D
GCCGCCGCACCTGCGGCGTCCGGCAGCGACGACGAGGCAGCCGCCGACGACGGCGACGACGACGAGGAAGCCGACGCTGACGAGGCCGCCGAGGCCGAGGACGCTGGCGACGACGACGAC
CGGCGGCGTGGACGCCGCAGGCCGTCGCTGCTGCTCCGTCGGCGGCTGCTGCCGCTGCTGCTGCTCCTTCGGCTGCGACTGCTCCGGCGGCTCCGGCTCCTGCGACCGCTGCTGCTGCTG
      4210      4220      4230      4240      4250      4260      4270      4280      4290      4300      4310      4320

 E  E  P  S  G  E  G  L  G  D  L  F  G
GAGGAGCCCAGCGGCGAAGGCCTGGGCGACCTCTTCGGGTAACCCGGTCGCGTCGCGCGCCGACAGCCACGATCACATCGTTTTTTAGCCGCGTGCCACTCGGGAAGCCACGGCGCTGCG
CTCCTCGGGTCGCCGCTTCCGGACCCGCTGGAGAAGCCCATTGGGCCAGCGCAGCGCGCGGCTGTCGGTGCTAGTGTAGCAAAAAATCGGCGCACGGTGAGCCCTTCGGTGCCGCGACGC
      4330      4340      4350      4360      4370      4380      4390      4400      4410      4420      4430      4440

CACGCGTAGATTGAAGACGGGGAGCGCGGGCAGCTGGGCGTGATGGATGTCGCTGGTGTGCGCGTGTTGGTGACGCCGATGGGCGCGCTTGCAGTGCTCGCTGCGGTCGCGGCCGGCGTG
GTGCGCATCTAACTTCTGCCCCTCGCGCCCGTCGACCCGCACTACCTACAGCGACCACACGCGCACAACCACTGCGGCTACCCGCGCGAACGTCACGAGCGACGCCAGCGCCGGCCGCAC
      4450      4460      4470      4480      4490      4500      4510      4520      4530      4540      4550      4560

                                                                         XmaI
GCGCTGGGGACGTGTTCGGGGCTGGTGCCCGGCGTGCACGTGAACACGCTGGCGTTGCTGCTGGCTGCGGCCGCGCCGCTGGCCCCGGGACCGCCACATCTCGTCGGTGCGGCGCTACTG
CGCGACCCCTGCACAAGCCCCGACCACGGGCCGCACGTGCACTTGTGCGACCGCAACGACGACCGACGCCGGCGCGGCGACCGGGGCCCTGGCGGTGTAGAGCAGCCACGCCGCGATGAC
      4570      4580      4590      4600      4610      4620      4630      4640      4650      4660      4670      4680

GCGGCGGGTGTCACGCATTCCATCCTGGACGTGGTCCCGATGCTCGCGCTCGGGGTGCCGGACGCGGCGCTGGCGGTGAGCGTGCTGCCCGGCCATCGGCTGGTGCTTGGCGGTCGCGGC
CGCCGCCCACAGTGCGTAAGGTAGGACCTGCACCAGGGCTACGAGCGCGAGCCCCACGGCCTGCGCCGCGACCGCCACTCGCACGACGGGCCGGTAGCCGACCACGAACCGCCAGCGCCG
      4690      4700      4710      4720      4730      4740      4750      4760      4770      4780      4790      4800

CGGGAGGCGCTGCGGGTGTCTGCCCTCGGGAGCGCGACCGCCGTCGTGGTCGCCGTCGCATTCGGGGTGCCGGCGACGTGGCTGGCGGTTCGGGCGGCACCAGTGGTGTACGCTCACCTC
GCCCTCCGCGACGCCCACAGACGGGAGCCCTCGCGCTGGCGGCAGCACCAGCGGCAGCGTAAGCCCCACGGCCGCTGCACCGACCGCCAAGCCCGCCGTGGTCACCACATGCGAGTGGAG
      4810      4820      4830      4840      4850      4860      4870      4880      4890      4900      4910      4920

CCGGTCGTGCTCGCCGTGCTCGTCGTGGTGCTCGTCGCACGCGAGCGGTCCCGACGCGCGCGGCTCGGCGCGGTGCTCGCGGTGGCTGCCAGCGGCACACTGGGCAGCGTTGCGCTGCCG
GGCCAGCACGAGCGGCACGAGCAGCACCACGAGCAGCGTGCGCTCGCCAGGGCTGCGCGCGCCGAGCCGCGCCACGAGCGCCACCGACGGTCGCCGTGTGACCCGTCGCAACGCGACGGC
      4930      4940      4950      4960      4970      4980      4990      5000      5010      5020      5030      5040

                                                                                                    BamHI
ATGCAGCCCGACGCTGTCGTACCGACCGGTGACGTGTTGTCGCCGCTGTTCGCGGGGTTGTTCGGTGCGCCCGTGTTGTTGGCGGCGTTCCGGGGGGATGGGATCC
TACGTCGGGCTGCGACAGCATGGCTGGCCACTGCACAACAGCGGCGACAAGCCCCAACAAGCCACGCGGGCACAACAACCGCCGCAAGGCCCCCCTACCCTAGG
      5050      5060      5070      5080      5090      5100      5110      5120      5130      5140
```

number of the ribosomal protein genes labelled pLW173 was hybridized to genomic DNA by Southern blot analysis under progressively less stringent conditions. No other hybridizing bands were seen on such Southerns (data not shown).

3.3 Identification of Proteins Encoded by pLW173

Four of the open reading frames encoded on pLW173 were identified as encoding expressed proteins by matching the putative translated gene products to amino acid sequences of known *H. cutirubrum* proteins (Figure 6). Homologous proteins from the eubacterium *E. coli* were also identified (Figure 6).

The open reading frame located at nucleotide positions 1622 - 2110 encodes a protein that has 163 amino acids, a molecular mass of 17020 daltons and an isoelectric point of 2.6. The amino acid residues 2 - 35 of the translated open reading frame match the amino terminal sequence of the Hcu L11 protein perfectly; the amino terminal methionine is apparently removed by post-translational modification. The single internal peptide available matches the protein gene product from residues 77 to 101 exactly (A.T. Matheson, personal communication). Thus the open reading frame located at nucleotide positions 1622 - 2110 encodes the Hcu L11 protein. The Hcu L11 protein is homologous to the Eco L11 protein: they retain 33% amino acid identities over 138 residues (significance z = 35 by the RDF program) requiring only a single gap of one amino acid residue to maintain alignment.

The open reading frame located at nucleotide positions 2314 - 2949 encodes a protein that has 212 amino acids, a molecular mass of 23095 daltons and an isoelectric point of 4.5. The amino acid residues 2 - 37 of the translated open reading frame match the amino terminal sequence of the Hcu L8 protein with the single exception of a glutamic acid residue versus a glycine residue at position 15 in the translated open reading frame and protein sequences respectively. The amino terminal methionine of the protein is removed by post-translational modification. Thus the open reading frame located at nucleotide positions 2314 - 2949 apparently encodes the Hcu L8 protein. The deduced Hcu L8 protein is homologous to the Eco L1 protein: they have a linear correspondence yielding 32% amino acid identities over 213 positions (z = 36 by the RDF program) but require 10 gaps in the alignment.

The open reading frame located at nucleotide positions 2954 - 4009 encodes a protein that has 352

Figure 6 Identification of the Ribosomal Proteins Encoded by pLW173

The sequences of the *H. cutirubrum* L11, L8, L3 / 4 and L20 ribosomal proteins as translated from their respective genes are shown with the matches to peptide sequence data underlined. The Hcu L11, L8, L3 / 4 and L20 ribosomal protein sequences are also aligned with the *E. coli* L11, L1, L10 and L12 ribosomal protein sequences respectively. Amino acid identities are indicated by solid circles (•) and gaps (-) have been inserted where neccesary to maintain alignment. The Eco L12 protein has undergone rearrangement and thus identities are indicated between the Hcu L20 and Eco L12 C domain (•), the Eco L12 N terminus and Eco L12 C domain (◊), and the Hcu L20 and Eco L12 N terminus (o).

```
Eco L11     MAKKVQAYVKLQVAAGMANPSPPVGPALGQQGVNIMEFCKAFNAKTDSIEKGLPIPVVITVYADRSFTFVTKTPPAAVLLKKAAGIKSGSGKPNKDKVGKISRAQLQEIA
                        ●● ● ● ● ●● ●● ●●   ●         ●  ●     ●   ●● ●    ● ●●      ●● ● ● ●  ●●   ●●● ●    ●   ●  ●●  ●●
Hcu L11         MAETIEVLVAGGQADPGPPLGPELGPTPVDVQAVVQEINDQTEAF-DGTEVPVTIEVEDDGSFSIEVGVPPTAALVKDEAGFDTGSGEPQEMFVADLSIEQLKTIA
                    10        20        30        40        50        60        70        80        90       100       110

Eco L11     QTKAADMTGADIEAMTRSIEGTARSMGLVVED
              ●  ●    ●         ●●  ● ●   ●
Hcu L11     EQKKPDLLAYDARNAAKEVAGTCASLGVTIEGEDARTFNERVDDGDYDDVLGDELAAA
                   120       130       140       150       160

Eco L1      MAKLTKRMRVIREKVDATKQYDINEAIALLKELAT-AKFVESVDVAVNL-GIDARKSDQNVRGATVLPHGTGRSVRVAVFTQGANAEAAKAAGAELVGMEDLADQI----
                           ●         ●● ●  ● ●    ● ● ●● ●●●●   ●       ●    ●●● ●●●      ●●  ●  ●  ●    ●  ●  ●●
Hcu L8                   MADND-IEE-AVARAL--EDAPQANFRETVDLAVNLADLDLNDPSQAVDEGVVLPSGTGQETQIVVFADGETAVAADDV-ADDVLDEDDLSDLADDT
                    10        20        30        40        50        60        70        80        90       100       110

Eco L1      ---KKGEMMFDVVIASPDAMRV-VGQLGQVLGPRGLMPNPKVGTVTPNVAEAVKNAKAGQVRYRNDKNGIIHTTIGKVDFDADKLKENLEALLVALKKAKPTQAKG-VYI
               ●      ●   ●    ●   ●● ●●●●●●●●● ●● ●        ●   ●   ●   ●  ●       ●●  ●  ●  ●     ●                ●●
Hcu L8      DAAKDLADETDFFVAEAPMMQDIVGALGQVLGPAGKMPTPLQP--DDDVVDTVNAMK-NTVQIRSRDAATFHTAVGAEDMSAEDIASNIDVI---MAALHANLEKGPLNV
                   120       130       140       150       160       170       180       190       200       210       220

Eco L1      KKVSISTTMGAGVAVDQAGLSASVN
              ●   ●●●●  ● ●
Hcu L8      DSVYVKTTMGPAVEVA
                   230       240

Eco L10     MALNLQDKQAIVAE--VSEVAKGA------LSAVVADSRGVTVDKMTELRKAGREAGVYMRVVRNTLLRRAVE--GTPFECLKDAFVGPTLIAYSMEHPGAAARLFKEFA
            ●    ●     ● ●    ●●●          ●  ●    ●         ●          ●  ●●●●● ●● ●  ●     ●     ●          ●        ●
Hcu L3/4    MSAEEQRTTEEVPEWKRQEVAELVDLLETVDSVGVVNVTGIPSKQLQDMRRGLH-GQAAVRMSRNTLLVRALEEAGDGLDTLTEVVEGEVGLVATMDMPFGL-VQQLENS
                    10        20        30        40        50        60        70        80        90       100       110

Eco L10     KANAKFEVKAAAFEGELIPASQIDRLA-----------------------------------------------------------------------------------
            ●  ●       ●      ●
Hcu L3/4    KTPAPINAGEVAPNDIVVPEGDTGIDPGPFVGELQTIGANAAIQEGSIQVLDDSVVTEEGETVSDDVSNVLSELGIEPKEVGLDLRGVFSEGVLFTPEELEIQVDEYRAD
                   120       130       140       150       160       170       180       190       200       210       220

Eco L10     -----------------------------------TLPTYEEAIARLMATMKEASAGKLVRTLAAVRDAKEAA
                                                ●           ●      ●   ●● ●●●  ●   ●
Hcu L3/4    IQSAAASARNLSVNAAVPTERTAPDLIAKGRGEAKSLGLQASVESPDLADDLVSKADAQVRALAAQIDDEDALPEELQDVDAPAAPAGGEADTTADEQSDETQASEADDA
                   230       240       250       260       270       280       290       300       310       320       330

Hcu L3/4    DDSDDDDDDDDGNAGAEGLGEMFG
                   340       350

Eco L12 N TERMINUS                                           MSITKDQIIE-AVAAM----SVMDVVELISAMEEKFGVSAAAAVA--VAAGPVEAA...
                                                                 ●●   ● ● ●●     ●  ●   ●  ● ●●
Eco L12 C DOMAIN       ...EEKTEFDV--ILKAAGANKVAVIKAVRGATGL--GLKEAKD-LVESAPAALKEGVSKDDAEALKKALEEAGAEVEVK
                          ●● ●●      ● ●●●     ●    ● ●  ●   ●    ●   ● ●                           ooo  o   o o    o
Hcu L20     MEVVYAALILNEADEELTEDNITGVLEAAGVD---VEE-SR-AKALVAALED-VD-IEE-AVEE-----------------------AAAAPAAAPAASGSDDEAAADDG
                    10        20        30        40        50        60        70        80        90       100       110

Hcu L20     DDDEEADADEAAEAEDAGDDDDEEPSGEGLGDLFG
                   120       130       140
```

amino acids, a molecular mass of 37159 daltons and an isoelectric point of 2.6. Amino acid residues 17 - 41, 70 - 109, 151 - 187, 195 - 216, 254 - 273 and 280 - 295 of the translated open reading frame are virtually identical to sequences of internal tryptic peptide fragments of the Hcu L3 / 4 protein thus identifying this open reading frame as encoding the Hcu L3 / 4 protein (A.T. Matheson, personal communication). The Hcu L3 / 4 and Eco L10 proteins are homologous: the Eco L10 protein has a large internal deletion of 118 residues and a carboxy terminal truncation of 61 residues with respect to the Hcu L3/4 protein but retains 23% amino acid identity over 169 residues (z = 10 by the RDF program).

The open reading frame located at nucleotide positions 4018 - 4359 encodes a protein that has 114 amino acids, a molecular mass of 11550 daltons and an isoelectric point of 2.1. The translated open reading frame matched the partial sequence of 77 amino acids of the Hcu L20 protein available and was later confirmed in its entirety with the complete sequencing of the 114 residues of the protein (Liljas *et al.*, 1986). The Hcu L20 protein has previously been shown to be the homologue of the Eco L12 protein by means of statistical analysis (3.5 standard deviations from the mean by Monte Carlo simulation) and by virtue of the conservation of structure, dimerization and function (Yaguchi *et al.*, 1980; Matheson, 1985).

In summary, the Hcu L11, Hcu L8, Hcu L3 / 4 and Hcu L20 proteins are encoded by the open reading frames located at nucleotide positions 1622 - 2110, 2314 - 2949, 2954 - 4009 and 4018 - 4359 respectively, are homologous (or equivalent 'e') to the Eco L11, Eco L1, Eco L10 and Eco L12 proteins respectively and are henceforth referred to as Hcu L11e, Hcu L1e, Hcu L10e and Hcu L12e (for both genes and proteins) respectively.

Analysis of the 1621 nucleotides in front of the Hcu L11e gene on the 5.1 Kilobasepair fragment revealed two potential coding regions (designated ORF and NAB). The 784 nucleotides distal to the L12e gene bears an extreme paucity of restriction sites and is devoid of coding potential.

The ORF gene, initiating at nucleotide 1244 and terminating at nucleotide 135, encodes a potential protein of 370 amino acids with a molecular mass of 40499 daltons, an isoelectric point of 5.5 and is oriented opposite to and divergently transcribed from the ribosomal protein genes on the genomic fragment. This potential protein shows no similarity to any known protein sequence in the NBRF PIR and GENBANK databases.

The NAB gene, initiating at nucleotide position 1245 and terminating at position 1548, has the same orientation as and is located immediately upstream of the four ribosomal proteins. The NAB - L11e intergenic space is 73 nucleotides in length. The gene potentially encodes a short 68 amino acid protein with a molecular mass of 7530 daltons and an isoelectric point of 3.8. The potential gene product exhibits amino acid sequence similarity to proteins binding nucleic acids (thus the name: Nucleic Acid Binder) and especially to the restriction endonucleases EcoRI and PstI: 32% and 30% amino acid identity respectively over the central region (positions 6 - 46) of the aligned proteins (Figure 7).

3.4 Amino Acid Composition and Codon Utilization within pLW173

The amino acid compositions of the six *H. cutirubrum* proteins encoded by the 5.1 Kilobasepair fragment are listed in Table 4. The composition of the four *H. cutirubrum* ribosomal proteins differs from the equivalent *E. coli* proteins in that they have about twice the content of acidic (aspartic + glutamic acids) residues and half the content of basic (arginine + lysine) residues. It is believed that the high content of acidic residues in the halophilic proteins aids in preserving their structure and function in the high intracellular ionic strength environments in which they exist (Bayley and Morton, 1978; Eisenberg and Wachtel, 1987; Saenger, 1987). The putative proteins encoded by NAB and ORF are also rich in acidic residues.

The codon utilization of the Hcu ORF, NAB, L11e, L1e, L10e and L12e genes representing 3837 nucleotides or 75% of the 5146 bp fragment are shown in Table 5. Four points are apparent. First, the TTT codon for phenylalanine is never utilized, there is a strong preference to avoid codons with adjacent T residues and in only two cases do codons with T at the third position precede codons with T at the first position. There are unique TTT and TTTT tracts within the coding regions at positions 427 - 425 and 2498 - 2501 respectively, and no T tract longer than four. In both the (-) strand of coding regions and within non-coding regions T tracts are much more prevalent. The paucity of T tracts on the (+) strand within coding regions is probably related to their participation in transcription termination (see section 3.6.2). Second, G or C occurs 87% of the time at the third position of the codon, the reason for the enhancement above the high G plus C context of *H. cutirubrum* genomic DNA (68%) remains unknown. Third, arginine is entirely encoded by the CGN codons and never by the AGR codons (68 - 0; N = any nucleotide, R = purine). A similar bias exists for arginine codons in *E. coli*. In *Saccharomyces cerevisiae* ribosomal protein

Figure 7 Alignment of the Putative NAB Protein with EcoRI and PstI

The 68 amino acid putative protein encoded by the NAB gene is aligned with amino acid residues 50 - 117 and residues 151 - 218 of the eubacterial restriction endonucleases EcoRI and PstI respectively. Amino acid identities (•) and conservative substitutions (o) between the NAB and EcoRI proteins and the NAB and PstI proteins are indicated above the EcoRI and PstI sequences respectively. Conservative substitutions are defined as substitutions within the amino acid groups: D - E - Q - N - R - K - H; L - I - V - M - F; A - S; S - T; A - G.

```
                         10                  20                  30                    40                  50                  60
NAB     M G D P A A Y R D S T Q I V L P V G T L E D I E V D L E A E F M V S V F R P T - D A E I V R I I G S P V V I K E V T E F L T R H G V H M P

                    ● ● ● ●   o               ● o o ● o   ● ●   o   o o ● ●             ● o   ● ● o o               o   ●           o ●     o
EcoRI   P Q L S F A Y R D S I K K T E I N E A L K K I D P D L G G T L F V S N S S I K P D G G I V E V K D D Y G E W A V V L V A E A K H Q G K D I
       50                  60                  70                  80                  90                  100                  110

              ●         ●   ● ● ●   ●     ●     o o o o   ● o     ● ●         ●           o o ● o   ●     o         o ●     o o o     o       o
PstI    D A I P I K L L D G T Q I S L S P G E H N Q L H A D I V H E F - C P A F V G D - M G K I L Y I G D T A S S A N E G G K L M V L D S E Y L K
                        160                 170                 180                   190                 200                 210
```

Table 4 Amino Acid Composition of Proteins Encoded by pLW173

Amino acid composition is shown in absolute and, in parentheses, mole percent values. The single letter amino acid code is indicated in the second column.

		L11e	L1e	L10e	L12e	NAB	ORF
		163 AA	212 AA	352 AA	114 AA	68 AA	370 AA
		MW 17020	MW 23095	MW 37159	MW 11550	MW 7530	MW 40499
		pI 2.7	pI 6.7	pI 2.9	pI 2.1	pI 3.8	pI 6.8
ALANINE	A	20 (12.3%)	23 (10.8%)	41 (11.6%)	28 (24.6%)	4 (5.9%)	28 (7.6%)
ARGININE	R	3 (1.8%)	15 (7.1%)	14 (4.0%)	1 (0.9%)	4 (5.9%)	31 (8.4%)
ASPARAGINE	N	4 (2.5%)	9 (4.2%)	12 (3.4%)	2 (1.8%)	0 (0.0%)	16 (4.3%)
ASPARTIC ACID	D	18 (11.0%)	33 (15.6%)	42 (11.9%)	20 (17.5%)	5 (7.4%)	30 (8.1%)
CYSTEINE	C	1 (0.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
GLUTAMINE	Q	7 (4.3%)	8 (3.8%)	17 (4.8%)	0 (0.0%)	1 (1.5%)	10 (2.7%)
GLUTAMIC ACID	E	19 (11.7%)	14 (6.6%)	41 (11.6%)	22 (19.3%)	7 (10.3%)	34 (9.2%)
GLYCINE	G	16 (9.8%)	11 (5.2%)	31 (8.8%)	9 (7.9%)	4 (5.9%)	39 (10.5%)
HISTIDINE	H	0 (0.0%)	2 (0.9%)	1 (0.3%)	0 (0.0%)	2 (2.9%)	5 (1.4%)
ISOLEUCINE	I	7 (4.3%)	7 (3.3%)	12 (3.4%)	3 (2.6%)	6 (8.8%)	23 (6.2%)
LEUCINE	L	11 (6.7%)	16 (7.5%)	31 (8.8%)	8 (7.0%)	4 (5.9%)	32 (8.6%)
LYSINE	K	5 (3.1%)	5 (2.4%)	7 (2.0%)	1 (0.9%)	1 (1.5%)	15 (4.1%)
METHIONINE	M	1 (0.6%)	8 (3.8%)	4 (1.1%)	1 (0.9%)	3 (4.4%)	8 (2.2%)
PHENYLALANINE	F	5 (3.1%)	5 (2.4%)	5 (1.4%)	1 (0.9%)	3 (4.4%)	8 (2.2%)
PROLINE	P	11 (6.7%)	10 (4.7%)	17 (4.8%)	3 (2.6%)	5 (7.4%)	12 (3.2%)
SERINE	S	5 (3.1%)	7 (3.3%)	21 (6.0%)	4 (3.5%)	3 (4.4%)	19 (5.1%)
THREONINE	T	11 (6.7%)	13 (6.1%)	19 (5.4%)	2 (1.8%)	5 (7.4%)	15 (4.1%)
TRYPTOPHAN	W	0 (0.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	1 (0.3%)
TYROSINE	Y	3 (1.8%)	1 (0.5%)	5 (1.4%)	2 (1.8%)	1 (1.5%)	9 (2.4%)
VALINE	V	16 (9.8%)	25 (11.8%)	31 (8.8%)	7 (6.1%)	10 (14.7%)	35 (9.5%)

Table 5 Codon Utilization of Proteins Encoded by pLW173

Codon usage is shown in absolute and, in parentheses, percentage values. 'rPRO' includes the L11e, L1e, L10e and L12e ribosomal proteins. 'ALL' includes all six proteins encoded by pLW173.

		L11e	L1e	L10e	L12e	NAB	ORF	rPRO	ALL
		163 AA	212 AA	352 AA	114 AA	68 AA	370 AA	841 AA	1279 AA
PHE	UUU	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	UUC	5 (3.1%)	5 (2.4%)	5 (1.4%)	1 (0.9%)	3 (4.4%)	8 (2.2%)	16 (1.9%)	27 (2.1%)
LEU	UUA	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	UUG	0 (0.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	4 (1.1%)	1 (0.1%)	5 (0.4%)
LEU	CUU	0 (0.0%)	2 (0.9%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (0.4%)	3 (0.2%)
	CUC	9 (5.5%)	10 (4.7%)	18 (5.1%)	3 (2.6%)	3 (4.4%)	15 (4.1%)	40 (4.8%)	58 (4.5%)
	CUA	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	CUG	2 (1.2%)	4 (1.9%)	11 (3.1%)	5 (4.4%)	1 (1.5%)	13 (3.5%)	22 (2.6%)	36 (2.8%)
ILE	AUU	0 (0.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	1 (0.3%)	1 (0.1%)	2 (0.2%)
	AUC	7 (4.3%)	6 (2.8%)	11 (3.1%)	3 (2.6%)	6 (8.8%)	22 (5.9%)	27 (3.2%)	55 (4.3%)
	AUA	0 (0.0%)	1 (0.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.1%)	1 (0.1%)
MET	AUG	1 (0.6%)	8 (3.8%)	4 (1.1%)	1 (0.9%)	3 (4.4%)	8 (2.2%)	14 (1.7%)	25 (2.0%)
VAL	GUU	2 (1.2%)	3 (1.4%)	2 (0.6%)	0 (0.0%)	0 (0.0%)	2 (0.5%)	7 (0.8%)	9 (0.7%)
	GUC	10 (6.1%)	10 (4.7%)	20 (5.7%)	6 (5.3%)	5 (7.4%)	16 (4.3%)	46 (5.5%)	67 (5.2%)
	GUA	1 (0.6%)	2 (0.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (0.4%)	3 (0.2%)
	GUG	3 (1.8%)	10 (4.7%)	9 (2.6%)	1 (0.9%)	5 (7.4%)	17 (4.6%)	23 (2.7%)	45 (3.5%)
SER	UCU	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	1 (0.0%)
	UCC	4 (2.5%)	2 (0.9%)	9 (2.6%)	2 (1.8%)	0 (0.0%)	2 (0.5%)	17 (2.0%)	19 (1.5%)
	UCA	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (1.5%)	0 (0.0%)	0 (0.0%)	1 (0.1%)
	UCG	1 (0.6%)	2 (0.9%)	4 (1.1%)	0 (0.0%)	0 (0.0%)	11 (3.0%)	7 (0.8%)	18 (1.4%)
PRO	CCU	1 (0.6%)	1 (0.5%)	2 (0.6%)	2 (1.8%)	1 (1.5%)	0 (0.0%)	6 (0.7%)	7 (0.5%)
	CCC	5 (3.1%)	2 (0.9%)	5 (1.4%)	1 (0.9%)	0 (0.0%)	2 (0.5%)	13 (1.5%)	15 (1.2%)
	CCA	0 (0.0%)	1 (0.5%)	0 (0.0%)	0 (0.0%)	1 (1.5%)	1 (0.3%)	1 (0.1%)	3 (0.2%)
	CCG	5 (3.1%)	6 (2.8%)	10 (2.8%)	0 (0.0%)	3 (4.4%)	9 (2.4%)	21 (2.5%)	33 (2.6%)
THR	ACU	0 (0.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.1%)	1 (0.1%)
	ACC	4 (2.5%)	5 (2.4%)	10 (2.8%)	2 (1.8%)	2 (2.9%)	9 (2.4%)	21 (2.5%)	32 (2.5%)
	ACA	0 (0.0%)	2 (0.9%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	2 (0.5%)	3 (0.4%)	5 (0.4%)
	ACG	7 (4.3%)	6 (2.8%)	7 (2.0%)	0 (0.0%)	3 (4.4%)	4 (1.1%)	20 (2.4%)	27 (2.1%)
ALA	GCU	2 (1.2%)	2 (0.9%)	2 (0.6%)	3 (2.6%)	1 (1.5%)	0 (0.0%)	9 (1.1%)	10 (0.8%)
	GCC	8 (4.9%)	7 (3.3%)	13 (3.7%)	16 (14.0%)	1 (1.5%)	13 (3.5%)	44 (5.2%)	58 (4.5%)
	GCA	0 (0.0%)	7 (3.3%)	3 (0.9%)	4 (3.5%)	0 (0.0%)	1 (0.3%)	14 (1.7%)	15 (1.2%)
	GCG	10 (6.1%)	7 (3.3%)	23 (6.5%)	5 (4.4%)	2 (2.9%)	14 (3.8%)	45 (5.4%)	61 (4.8%)

Table 5 (Continued)

		L11e	L1e	L10e	L12e	NAB	ORF	rPRO	ALL
		163 AA	212 AA	352 AA	114 AA	68 AA	370 AA	841 AA	1279 AA
TYR	UAU	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	UAC	3 (1.8%)	1 (0.5%)	5 (1.4%)	2 (1.8%)	1 (1.5%)	9 (2.4%)	11 (1.3%)	21 (1.6%)
OCH	UAA	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
AMB	UAG	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
HIS	CAU	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	CAC	0 (0.0%)	2 (0.9%)	1 (0.3%)	0 (0.0%)	2 (2.9%)	5 (1.4%)	3 (0.4%)	10 (0.8%)
GLN	CAA	0 (0.0%)	2 (0.9%)	3 (0.9%)	0 (0.0%)	0 (0.0%)	3 (0.8%)	5 (0.6%)	8 (0.6%)
	CAG	7 (4.3%)	6 (2.8%)	14 (4.0%)	0 (0.0%)	1 (1.5%)	7 (1.9%)	27 (3.2%)	35 (2.7%)
ASN	AAU	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	AAC	4 (2.5%)	9 (4.2%)	12 (3.4%)	2 (1.8%)	0 (0.0%)	16 (4.3%)	27 (3.2%)	43 (3.4%)
LYS	AAA	4 (2.5%)	3 (1.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (0.8%)	7 (0.8%)	10 (0.8%)
	AAG	1 (0.6%)	2 (0.9%)	7 (2.0%)	1 (0.9%)	1 (1.5%)	12 (3.2%)	11 (1.3%)	24 (1.9%)
ASP	GAU	0 (0.0%)	3 (1.4%)	4 (1.1%)	0 (0.0%)	0 (0.0%)	3 (0.8%)	7 (0.8%)	10 (0.8%)
	GAC	18 (11.0%)	30 (14.2%)	38 (10.8%)	20 (17.5%)	5 (7.4%)	27 (7.3%)	106 (12.6%)	138 (10.8%)
GLU	GAA	8 (4.9%)	4 (1.9%)	13 (3.7%)	8 (7.0%)	1 (1.5%)	8 (2.2%)	33 (3.9%)	42 (3.3%)
	GAG	11 (6.7%)	10 (4.7%)	28 (8.0%)	14 (12.3%)	6 (8.8%)	26 (7.0%)	63 (7.5%)	95 (7.4%)
CYS	UGU	1 (0.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.1%)	1 (0.1%)
	UGC	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
OPA	UGA	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
TRP	UGG	0 (0.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	1 (0.3%)	1 (0.1%)	2 (0.2%)
ARG	CGU	0 (0.0%)	2 (0.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (0.8%)	2 (0.2%)	5 (0.4%)
	CGC	2 (1.2%)	10 (4.7%)	10 (2.8%)	1 (0.9%)	3 (4.4%)	14 (3.8%)	23 (2.7%)	40 (3.1%)
	CGA	0 (0.0%)	1 (0.5%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (0.2%)	2 (0.2%)
	CGG	1 (0.6%)	2 (0.9%)	3 (0.9%)	0 (0.0%)	1 (1.5%)	14 (3.8%)	6 (0.7%)	21 (1.6%)
SER	AGU	0 (0.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	1 (1.5%)	0 (0.0%)	1 (0.1%)	2 (0.2%)
	AGC	0 (0.0%)	3 (1.4%)	7 (2.0%)	2 (1.8%)	1 (1.5%)	5 (1.4%)	12 (1.4%)	18 (1.4%)
ARG	AGA	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	AGG	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
GLY	GGU	4 (2.5%)	3 (1.4%)	2 (0.6%)	0 (0.0%)	1 (1.5%)	2 (0.5%)	9 (1.1%)	12 (0.9%)
	GGC	8 (4.9%)	5 (2.4%)	22 (6.3%)	8 (7.0%)	1 (1.5%)	19 (5.1%)	43 (5.1%)	63 (4.9%)
	GGA	1 (0.6%)	1 (0.5%)	2 (0.6%)	0 (0.0%)	0 (0.0%)	4 (1.1%)	4 (0.5%)	8 (0.6%)
	GGG	3 (1.8%)	2 (0.9%)	5 (1.4%)	1 (0.9%)	2 (2.9%)	14 (3.8%)	11 (1.3%)	27 (2.1%)

genes both the AGR and CGN codons are used whereas in the archaebacterium *S. solfataricus* the arginine codon bias of the GTPase domain ribosomal protein genes is inverted from that of *H. cutirubrum*, arginine being encoded exclusively by the AGR and never by the CGN codons (Ramirez *et al.*, 1990a). Fourth, genes encoding proteins in eubacteria and eucaryota usually contain more R / N / Y and G / non-G / N triplets in the functional reading frame than in the other two possible reading frames or intergenic regions (Y = pyrimidine; Shepherd, 1981; Trifonov, 1987). The six genes encoded by pLW173 (and the majority of archaebacterial genes) follow both the R / N / Y and G / non-G / N patterns. These codon biases are believed to be derived from a frame monitoring / maintaining system developed during the early evolution of the translation apparatus, apparently preceding the divergence of the three urkingdoms.

3.5 Transcription Analysis of pLW173

The *in vivo* transcripts produced from the 5.1 Kilobasepair fragment of genomic DNA were detected and analyzed by Northern hybridization, S1 nuclease protection analysis and primer extension analysis utilizing reverse transcriptase. Total RNA was isolated from exponentially growing cells and used in these procedures. The results of these analyses for the ORF gene, the NAB and L11e genes and the L1e, L10e and L12e genes are shown in Figures 8, 9 and 10 respectively and summarized in Figure 11.

3.5.1 Transcription of the ORF Gene

The very rare leftwards transcript of the ORF gene was identified by Northern hybridization utilizing ΦLW703, containing the 189 nucleotide SalI insert from positions 324 - 512 inclusive, as probe (labeled with α^{32}P dCTP by primer extension with Klenow fragment) and was estimated to be about 1200 nucleotides in length (Figure 8A). The 5' ends of the transcript were identified by primer extension analysis with oLW36 on total RNA as template (Figure 8B). The major 5' end site occurs at the G residue at position 1245, one nucleotide in front of the putative ATG initiation codon and minor 5' end sites occur at positions 1248 (C residue) and 1293 (G residue yielding a 49 nucleotide leader). By densitometry of the autoradiogram the relative transcription levels at the minor sites are 11% and 6% of the major transcription initiation site respectively. The 3' transcript end site was identified by S1 nuclease protection of a 906 basepair PstI - NheI fragment (positions 3609 of pBR322 to position 131 of the insert) 3' end labelled with

Figure 8 Transcription of the ORF Gene

Line Diagram

The transcription of the ORF gene of *Halobacterium cutirubrum* is depicted. The gene is oriented towards the left and part of the pBR322 vector is included on the left; the junction between vector and insert is at the ClaI site. The transcripts are indicated above; open circles (O) represent 5' transcript ends and the vertical line (**I**) represents the 3' transcript end. Restriction sites and their positions used to generate probes for transcription analysis are indicated below.

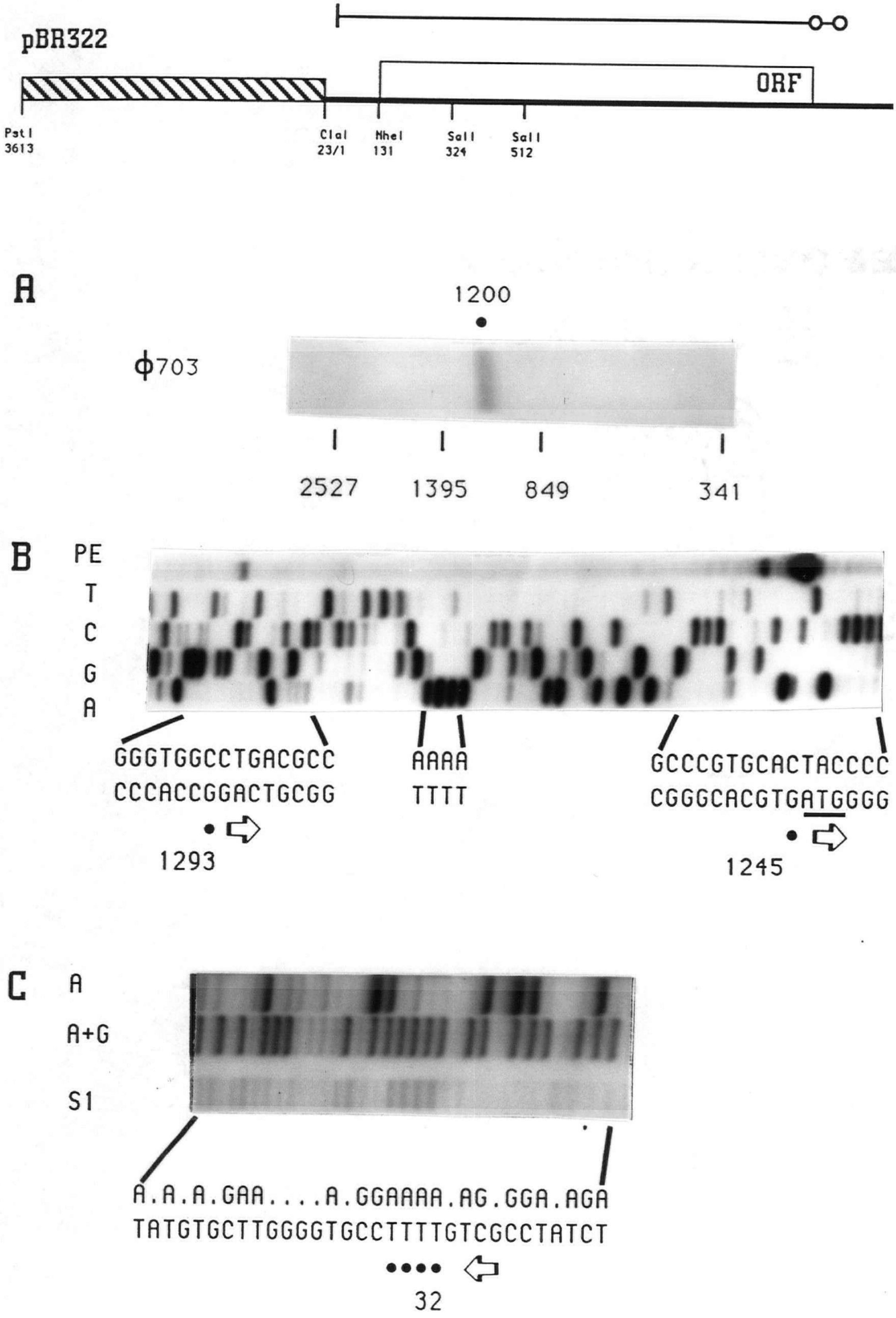
Photographs

In the primer extension and S1 nuclease protection experiments having sequence ladders the sequence is written below the photograph and the positions of the transcription initiation site(s) and termination site(s) are indicated by filled circles and the direction of transcription is indicated by an arrow (or arrows). The translation initiation codon is underlined.

A Northern blot of genomic *H. cutirubrum* RNA probed with phage ΦLW703 containing a 189 nucleotide SalI insert (positions 324 - 512 inclusive). Size standards were HaeIII restricted and 5' end labeled single stranded M13mp18 phage.

B 5' end of the ORF transcript detected by primer extension (PE) using oLW 36. Sequence is derived from priming phage ΦLW706 containing a 563 nucleotide insert (positions 943 - 1505) with oLW36 using T7 DNA polymerase.

C 3' end of the ORF transcript detected by S1 nuclease protection of an PstI - NheI fragment (position 3613 of pBR322 to position 131 of the insert DNA) 3' end labelled with α^{32}P dCTP at position 132. Sequence ladder was generated by chemical sequencing from the NheI site.


pBR322
ORF
PstI
3613
ClaI
23/1
NheI
131
SalI
324
SalI
512
A
1200
ϕ703
2527
1395
849
341
B
PE
T
C
G
A
GGGTGGCCTGACGCC
CCCACCGGACTGCGG
1293
AAAA
TTTT
GCCCGTGCACTACCCC
CGGGCACGTGATGGGG
1245
C
A
A+G
S1
A.A.A.GAA....A.GGAAAA.AG.GGA.AGA
TATGTGCTTGGGGTGCCTTTTGTCGCCTATCT
32

α^{32}P dCTP at the NheI site. Termination occurs primarily within the TTTT sequence at positions 32 - 29 (Figure 8C).

3.5.2 Transcription of the NAB and L11e Genes

Three different rightwards transcripts are detectable from the NAB - L11e region by Northern hybridization using M13 phage subclones (labeled with α^{32}P dCTP by primer extension with Klenow fragment) as probes (Figure 9A). Using ΦLW635 (containing a 94 nucleotide MboI insert; positions 1473 - 1380) as a probe, a 270 nucleotide long monocistronic NAB transcript and an 850 nucleotide long bicistronic NAB - L11e transcript were detected. The 850 nucleotide long bicistronic NAB - L11e transcript and a third transcript, a 600 nucleotide long L11e monocistronic transcript, were detected using the phage ΦLW670 (containing a 115 nucleotide TaqI insert; positions 2028 - 1914) as a probe. The phage ΦLW607 containing a 363 nucleotide insert of 5 noncontiguous MboI fragments (positions 1562 - 1299 and 1735 - 1633) detects all three transcripts and densitometry of the autoradiogram revealed relative transcriptional expression levels of 1% and 10% of the monocistronic L11e transcript for the NAB and bicistronic NAB - L11e transcripts respectively.

Low resolution of the 5' ends of the ORF, NAB - L11e bicistronic and L11e monocistronic transcripts were detected by S1 nuclease protection of a 529 basepair SalI fragment (positions 1177 - 1706) 5' end labelled at positions 1178 and 1710 with γ^{32}P ATP (Figure 9B). Three specific protection products were observed. The first was a very rare product about 65 nucleotides in length resulting from protection by the leftwards ORF transcript with a 5' end at position 1245 and protecting 5' label at position 1178. The second and third products were about 90 and 370 nucleotides in length and correspond to the rightwards 5' transcript ends at approximate positions 1620 and 1340 respectively and protecting label at position 1710. (A fourth band of about 210 nucleotides in length is the result of contamination of the 5' end labeled probe with the 383 basepair SalI fragment (positions 2060 - 2443) and protection by the 5' end of the L1e - L10e - L12e RNA transcript; see section 3.5.3). Densitometry on this autoradiogram indicates that the relative expression levels of the ORF and NAB - L11e bicistronic transcripts are 0.2% and 11% of the L11e monocistronic transcript expression respectively.

Primer extension analysis of transcripts encoding NAB was performed on total RNA with the

Figure 9 Transcription of the NAB and L11e Genes

Line Diagram

The transcription of the NAB and L11e genes of *Halobacterium cutirubrum* are depicted. The genes are oriented to the right and the transcripts are indicated immediately above; open circles (○) represent 5' transcript ends, vertical lines (**I**) represent the 3' transcript ends and open boxes (□) represent multiple 3' transcript ends. Restriction sites and their positions used to generate probes for transcription analysis are indicated below.

Photographs

In the primer extension and S1 nuclease protection experiments having sequence ladders the sequence is written below the photograph and the positions of the transcription initiation site(s) and termination site(s) are indicated by filled circles and the direction of transcription is indicated by an arrow (or arrows). Translation initiation codons are underlined. C and F are composites of different exposures of single autoradiograms.

A Northern blot of genomic *H. cutirubrum* RNA probed with phage containing regions of the NAB, L11e and both of the NAB and L11e genes. Probes are: NAB, phage ΦLW635 containing a 94 nucleotide MboI insert (positions 1473 - 1380 inclusive); L11e, phage ΦLW670 containing a 115 nucleotide TaqI insert (positions 2028 - 1914 inclusive); NAB - L11e, phage ΦLW607 containing a 363 nucleotide insert of 5 noncontiguous MboI fragments (containing positions 1562 - 1299 and 1735 - 1633 inclusive). Size standards were HaeIII restricted and 5' end labeled single stranded M13mp18 phage.

B Low resolution of the 5' ends of the ORF, NAB - L11e bicistronic and L11e monocistronic transcripts detected by S1 nuclease protection of a 529 basepair SalI fragment (positions 1177 - 1706) 5' end labeled with γ^{32}P ATP at positions 1178 and 1710. The very rare transcript from the ORF gene is visible on the original autoradiogram but not on the photographic reproduction. Size standards were MspI restricted and 3' end labeled pBR322.

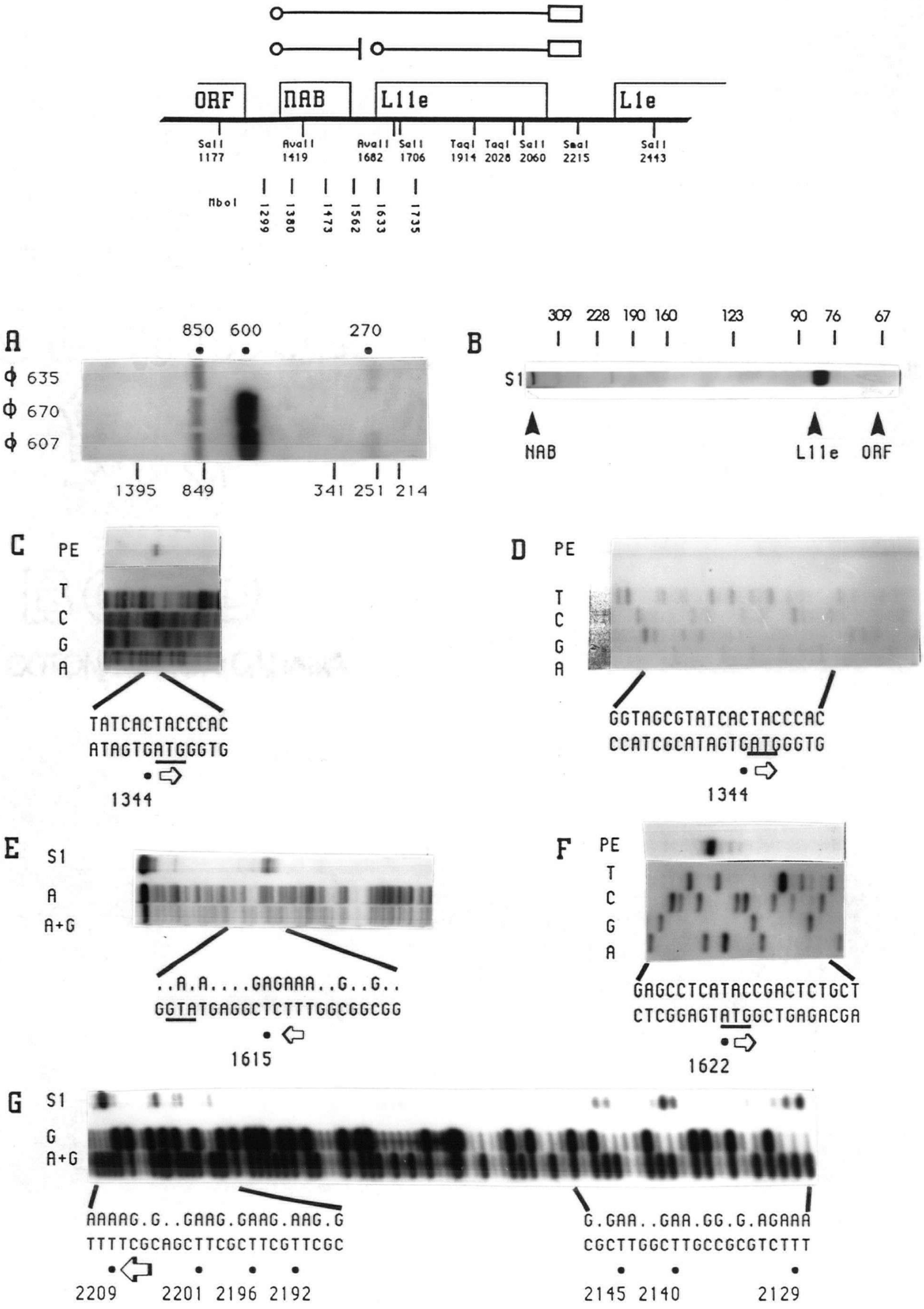
C 5' end of the NAB - L11e bicistronic transcript detected by primer extension (PE) using oLW51. Sequence is derived from priming phage ΦLW566 containing a 3382 nucleotide ClaI - PstI insert (positions 1 - 3382 inclusive) with oLW51 using T7 DNA polymerase.

D 5' end of the NAB transcript detected by primer extension (PE) using oLW52. Sequence is derived from priming phageΦLW566 with oLW52 using Klenow fragment.

E 3' end of the NAB monocistronic transcript detected by S1 nuclease protection of an AvaII fragment (positions 1419 - 1682) 3' end labeled with α^{32}P dTTP at position 1421. The sequence ladder was generated by chemical sequencing of the AvaII fragment.

F 5' end of the L11e transcript detected by primer extension (PE) with oLW51. The sequence is derived from priming phage ΦLW566 with oLW51 using T7 DNA polymerase.

G 3' ends of the L11e transcripts detected by S1 nuclease protection of a 155 nucleotide SalI - XmaI fragment (positions 2060 - 2215) labeled with α^{32}P dTTP at position 2064. The sequence ladder was generated by chemical sequencing of the SalI - XmaI fragment labeled at position 2064.


ORF
NAB
L11e
L1e
SalI 1177
AvaII 1419
AvaII 1682
SalI 1706
TaqI 1914
TaqI 2028
SalI 2060
SmaI 2215
SalI 2443
MboI
1299
1380
1473
1562
1633
1735
A
850
600
270
ϕ 635
ϕ 670
ϕ 607
1395
849
341
251
214
B
309
228
190
160
123
90
76
67
S1
NAB
L11e
ORF
C
PE
T
C
G
A
TATCACTACCCAC
ATAGTGATGGGTG
1344
D
PE
T
C
G
A
GGTAGCGTATCACTACCCAC
CCATCGCATAGTGATGGGTG
1344
E
S1
A
A+G
..A.A....GAGAAA..G..G..
GGTATGAGGCTCTTTGGCGGCGG
1615
F
PE
T
C
G
A
GAGCCTCATACCGACTCTGCT
CTCGGAGTATGGCTGAGACGA
1622
G
S1
G
A+G
AAAAG.G..GAAG.GAAG.AAG.G
TTTTCGCAGCTTCGCTTCGTTCGC
2209
2201
2196
2192
G.GAA..GAA.GG.G.AGAAA
CGCTTGGCTTGCCGCGTCTTT
2145
2140
2129

oligonucleotides oLW51 (positions 1714 - 1695; within the L11e gene; Figure 9C) and oLW52 (positions 1429 - 1410; within the NAB gene; Figure 9D). A unique 5' end site was detected at position 1344, corresponding to the G residue immediately in front of the NAB ATG translation initiation codon.

The position of 3' transcript end sites within and the extent of trancription through the NAB - L11e intergenic space was assessed by S1 nuclease protection of a 263 basepair AvaII fragment (positions 1419 - 1682) 3' end labeled at position 1421 with $\alpha^{32}P$ dTTP (Figure 9E). Two protection products were observed and correspond to full protection by read-through transcripts and partial protection by transcripts with 3' end sites at a T residue immediately after a TTT tract near position 1615. Densitometry of the autoradiagram indicates that readthrough from the NAB gene into the L11e gene constitutes 90% of the transcripts.

The 5' transcript end site for the 600 nucleotide L11e monocistronic transcript was identified at high resolution by primer extension analysis with oligonucleotide oLW51 (Figure 9F). The predominant 5' end site was detected at the A residue at position 1622. This residue is the initial nucleotide of the initiator methionine ATG codon of the L11e coding region; apparently this abundant transcript has no leader. There is a gap of 6 nucleotides between the 3' end of the monocistronic NAB transcript and the 5' end of the monocistronic L11e transcript.

The 3' transcript end sites in the L11e - L1e intergenic space were identified at low resolution by S1 nuclease protection of the 383 basepair SalI fragment (positions 2060 - 2443) 3' end labeled with $\alpha^{32}P$ dTTP at position 2064. Protection of fragments up to approximately 155 nucleotides in length was observed with little or no full length transcription readthrough into the L1e gene (data not shown). High resolution determination of the 3' transcript end sites in the L11e - L1e intergenic space was achieved by S1 nuclease protection of the 155 basepair SalI - XmaI fragment (positions 2060 - 2215). The probe fragment was 3' end labeled with $\alpha^{32}P$ dTTP at position 2064 and seven different sizes of protection products were observed corresponding to 3' end sites within T tracts near positions 2129, 2140, 2145, 2192, 2296, 2201, and 2209 (Figure 9G). It is worth noting here that tracts of two or more T residues within the coding regions of the six genes did not generate 3' ends in nuclease S1 protection experiments and therefore the seven 3' ends observed in the noncoding 3' flanking region of the L11e gene between positions 2129 and 2209 constitute transcription terminations and are not artifacts due to the action of S1 nuclease at regions of DNA / RNA hybrid instability caused by poly T tracts.

3.5.3 Transcription of the L1e, L10e and L12e Genes

A variety of M13 subclones of regions of the L1e, L10e and L12e genes were used as probes in Northern hybridization of total RNA and the results indicated that these three genes are encoded on a single tricistronic transcript of approximately 2150 nucleotides in length. Illustrated in Figure 10A is total RNA probed with ΦLW730, an M13 subclone containing a 294 nucleotide SalI insert (positions 2736 - 2443 inclusive), and labeled with α^{32}P dCTP by primer extension with Klenow fragment.

The 5' end sites of the tricistronic L1e - L10e - L12e transcript were analyzed by nuclease S1 protection of total RNA by the 383 basepair SalI fragment (positions 2060 - 2443) 5' end labeled with γ^{32}P ATP at position 2447 (Figure 10B). This resulted in protection of fragments ranging in size from approximately 120 to 210 nucleotides in length, corresponding to 75 nucleotides upstream and 10 nucleotides downstream of the L1e ATG translation initation codon. The 5' end sites were analyzed at higher resolution by primer extension with oLW38 (positions 2494 - 2478; Figure 10C) and oLW54 (positions 2326 - 2307; Figure 10D). The major transcripts had a 5' end at position 2239, 75 nucleotides in front of the L1e ATG translation initation codon and about 30 nucleotides beyond the last termination site for transcripts exiting the L11e gene. A number of other less abundant 5' ends located between positions 2239 and 2322 were also apparent and coincident both in primer extension and S1 nuclease protection experiments; the shortest of these at position 2322 is just within the coding region of L1e and probably represents an intermediate in the degradation of the tricistronic mRNA. No other major 5' ends were detected by S1 nuclease protection of SalI fragments between nucleotides 2322 and 4360, in agreement with the Northern hybridization analysis indicating a single tricistronic transcript encoding the L1e, L10e and L12e ribosomal proteins.

The 3' end sites of the tricistronic L1e - L10e - L12e transcript were mapped by S1 nuclease protection by total RNA of the SalI - XmaI fragment (positions 4159 - 4644) 3' end labeled with α^{32}P dTTP at position 4163. The 3' transcript end was located near nucleotide 4402 within a tract of six T residues (Figure 10E). All transcripts exiting the L12e gene terminate in this region. Attempts to identify transcripts from either strand of the DNA beyond position 4163 by Northern hybridization and S1 nuclease

Figure 10 Transcription of the L1e, L10e and L12e Genes

Line Diagram

The transcription of the L1e, L10e and L12e genes of *Halobacterium cutirubrum* are depicted. The genes are oriented to the right and the transcripts are indicated above; the open circle (O) represents the 5' transcript end and the vertical line (I) represents the 3' transcript end. Restriction sites and their positions used to generate probes for transcription analysis are indicated below.

Photographs

In the primer extension and S1 nuclease protection experiments having sequence ladders the sequence is written below the photograph and the positions of the transcription initiation site(s) and termination site(s) are indicated by filled circles and the direction of transcription is indicated by an arrow (or arrows). The L1e translation initiation codon is underlined. D is a composite of different exposures of a single autoradiogram.

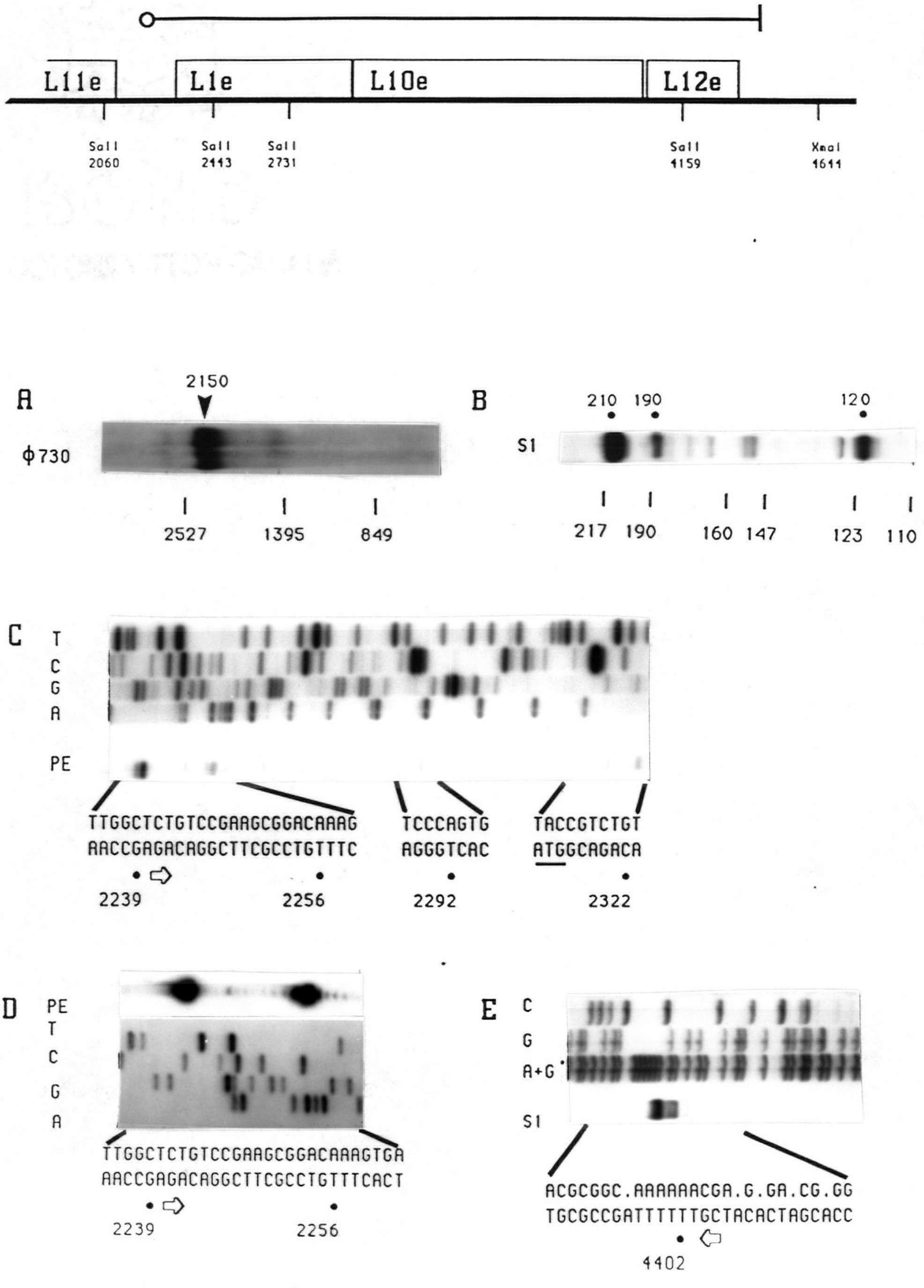
A Northern blot of genomic *H. cutirubrum* RNA probed with phage ΦLW730 containing a 294 nucleotide SalI insert (positions 2736 - 2443 inclusive). The autoradiogram is overexposed and bands at approximately 3 Kilobases and 1.5 Kilobases are due to weak hybridization to the 23S and 16S rRNAs respectively. Size standards were HaeIII restricted and 5' end labeled single stranded M13mp18 phage.

B 5' ends of the L1e - L10e - L12e transcript detected by S1 nuclease protection of a SalI fragment (positions 2060 - 2443) 5' end labeled at position 2447 with γ^{32}P ATP. Size standards were MspI restricted and 3' end labeled pBR 322.

C 5' ends of the L1e - L10e - L12e transcript detected by primer extension (PE) with oLW 38. Sequence is derived from priming phage ΦLW566 with oLW38 using Klenow fragment.

D 5' ends of the L1e - L10e - L12e transcript detected by primer extension (PE) with oLW 54. Sequence is derived from priming phage ΦLW730 containing a 389 nucleotide SalI - SalI insert (positions 2060 - 2448 inclusive) with oLW54 using T7 DNA polymerase.

E 3' end of the L1e - L10e - L12e transcripts detected by S1 nuclease protection of a SalI - XmaI fragment (positions 4159 - 4644) 3' end labeled with α^{32}P dTTP at position 4163. The sequence ladder was generated by chemical sequencing of the SalI - XmaI fragment.


L11e
L1e
L10e
L12e
SalI 2060
SalI 2443
SalI 2731
SalI 4159
XmaI 4644
A
2150
ϕ730
2527
1395
849
B
210
190
120
S1
217
190
160
147
123
110
C
T
C
G
A
PE
TTGGCTCTGTCCGAAGCGGACAAAG
AACCGAGACAGGCTTCGCCTGTTTC
TCCCAGTG
AGGGTCAC
TACCGTCTGT
ATGGCAGACA
2239
2256
2292
2322
D
PE
T
C
G
A
TTGGCTCTGTCCGAAGCGGACAAAGTGA
AACCGAGACAGGCTTCGCCTGTTTCACT
2239
2256
E
C
G
A+G
S1
ACGCGGC.AAAAAACGA.G.GA.CG.GG
TGCGCCGATTTTTTGCTACACTAGCACC
4402

protection experiments were negative, implying that this region probably represents a transcriptionally inactive space.

3.6 Consensus Signal Structures

Illustrated in Figure 11 is a line diagram depicting the genes encoded on pLW173 and summarizing the results of the transcript mapping experiments. Also summarized are the 5 transcriptional promotors (ORF P1, ORF P2, NAB, L11e and L1e), the 4 transcriptional terminators (ORF, NAB, L11e and L12e) and the translation initiation regions of the 6 genes.

3.6.1 Transcription Initiation Regions

Sequences surrounding the 5' transcript end sites are summarized (Figure 11 Transcription Promotion). The two conserved sequences that appear to constitute a part of the *H. cutirubrum* transcriptional promoter are TTCGA and TTAA. The spacing between these two elements is 4 - 15 nucleotides and the distance to the transcription start site is 20 - 26 nucleotides. More than one TTCGA element may be present. It is interesting to note that the very weak ORF P2 promoter exhibits the least conservation to the consensus at the appropriate position and this may be the reason for the low level of expression of transcription. The ORF P1 promotor has well conserved consensus elements but yields the rarest transcript; this may be related to the fact that the TTAA element is used on the complementary strand for promotion of the much more highly expressed NAB and NAB - L11e transcripts. Alternatively the downstream TTTT tract (positions 1274 - 1271) may result in premature termination; this would not be detected by the primer extension or S1 nuclease protection experiments. It is possible that this constitutes a mechanism for regulation of expression of the ORF gene; an antitermination factor could allow readthrough from the strong P1 promotor to elevate expression dramatically above the rare expression from the very poor P2 promotor. Such activation of the ORF P1 promotor could conceivably be coupled to down regulate the NAB promotor; this would have little effect on the level of expression of the L11e protein due to it being primarily present as a monocistronic transcript.

Figure 11 Summary of Gene Expression in pLW173

Line Diagram

The genomic organization of the L11e, L1e, L10e and L12e ribosomal protein gene cluster of *Halobacterium cutirubrum* is depicted. Known ribosomal protein encoding genes are solid boxes and putative protein encoding genes are striped boxes. Genes above the line are oriented and transcribed rightwards and those below the line are oriented and transcribed leftwards. The restriction sites indicated on the 5.1 Kilobasepair fragment (scale at top) are: AvaII, A; BamHI, B; ClaI, C; NheI, N; PstI, P; SalI, S; XmaI, X. Transcripts of the *Halobacterium cutirubrum* genes are indicated. The open circles (O) represent 5' transcript initiation sites and the vertical lines (I) represent 3' transcript termination sites. The open boxes (□) at the ends of trancripts indicate regions of multiple 3' trancript ends. The stippled box (▨) on the L1e - L10e - L12e tricistronic transcript indicates the region of potential regulation by autogenous inhibition of translation by the L1e protein (discussed in section 3.7).

Transcription Promotion

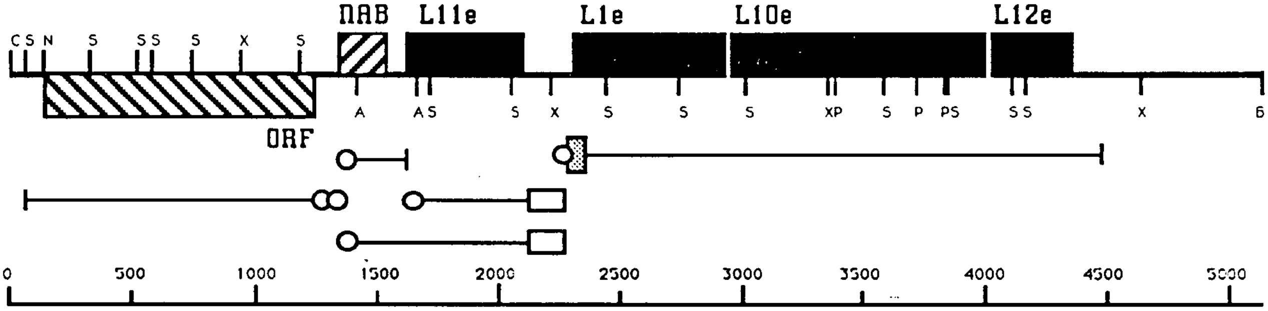
Sequences upstream of putative transcription initiation sites are shown with 5' end sites aligned at position +1 and highlighted with a solid circle (•). The position of the 5' end sites in the sequence presented in Figure 5 are indicated on the right. The ATG translation inititation codons adjacent to the 5' transcript end sites are heavy overlined. Sequences resembling the consensus TTCGA and TTAA motifs are underlined with the conserved bases highlighted. Where the terminator of the upstream gene overlaps with the promotor (L11e and L1e), the termination site(s) are light overlined.

Transcription Termination

Sequences upstream from putative transcription termination sites are aligned with the first base of the termination codon (heavy overline) set at +1. The position of the +1 nucleotide in the sequence presented in Figure 5 is indicated on the left. The GC rich tracts and poly T tracts are underlined and highlighted and the most prominent 3' end site in each T tract is indicated by a solid circle (•). Where the promotor of the downstream gene overlaps with the terminator (NAB and L11e), the promotor sequences (TTCGA and TTAA) are light overlined.

Translation Initiation Sites

Sequences surrounding the AUG translation initiation regions are presented with the initiation methionine codons heavy underlined. The first base in the initätion codons are aligned at +1 and their position in the sequence presented in Figure 5 is indicated on the left. Upstream termination codons are indicated by light underline. The sequence of the 3' end of the 16S ribosomal RNA is indicated at top and sequences complementary to the 3' end of the 16S rRNA are highlighted with solid circles (•). In the L12e initiation sequence an internal AUG codon near the end of the L10e cistron is overlined; if it were recognized as an initiation codon, it would produce a tripeptide.



TRANSCRIPTION INITIATION REGIONS

```
                                                                                 •
ORF P1                       cactaTgCGAtggcttcactcggcgtaTTAAcgtgtcgaaacgatcccaccGGAC...      1293

                                                                                 •
ORF P2    cgtgtcgaaacgatcccaccggacTgCGgaggaaagcgcTTttcggcgcttgctgtctacgggcacgtGATG...      1245

                                                                                 •
NAB                       ctccggtgggatcgtTTCGAcacgTTAAtacgccgagtgaagccatcgcatagtGATG...      1344

                                                                                 •
L11e      cgaTTCGAtccgcggcggcgccccgcTCGAaagacaagggTTAAacccgcggcggcggtttctcggagtATGG...      1622

                                                                                 •
L1e                                tgcTTCGcTTCGAcgcttTTAAgcccgggatcaccgtctgtagaaccGAGA...      2239
               -60       -50       -40       -30       -20       -10       +1
```

TRANSCRIPTION TERMINATION REGIONS

```
                                                                                                                                  ••••
ORF   cgcTAGccgatgccgaCGGCGCGCCGCCGGGtggtGGCGGtGGCGCtGGCCGCGGGagtggtgccgtgggtcatcgtcgactggtcgaacggcgtctatccgctgTTTTccg
       134
                                                                   •
NAB   ccgTGAgcgaTTcgatCCGCGGCGGCGCCCCGCtcgaaagacaagggTTaaaCCCGCGGCGGCGGTTTctc
      1549
                      •          •    •                                                       •   •    ••         •
L11e  gcgTAACGCCGCCCGaggagTTTctgcgccgTTcggTTcgcgtactcgatagcggcgtgtgtCCGCGGGtCGCGCtcccacgcTTgcTTcgcTTcgacgcTTTTaag
      2111
                                         ••
L12e  gggTAACCCGGtCGCGtCGCGCGCCGacagccacgatcacatcgTTTTTagc
      4360
        +1       10        20        30        40        50        60        70        80        90        100
```

```
                         3'  UCCUCCACUAGGUCG...16S rRNA
```

TRANSLATION INITIATION REGIONS

```
                                 • • •••••      ••••••
ORF (P1)                   ...uacggGcAcGUGAUGgggcucgAGGAGGac...

                                                ••••••
ORF (P2)                              gAUGgggcucgAGGAGGac...      1244

                                           ••••• ••
NAB                                   gAUGGGUGAcCCugcugcg...      1345

                                              •••  ••••
L11e                                   AUGgcuGAGacGAUCgaa...      1622

                                  ••••        •••  ••••
NAB/L11e                    ...uucucGGAGuAUGgcuGAGacGAUCgaa...      1622

                            ••••••
L1e                    ...acuacGGAGGUgaaagAUGgcagacaacgauaua...      2314

                        ••••••
L10e                ...gccguGGAGGUugccuaggAUGuccgccgaagaacaa...      2954

                 •••• •• ••
L12e        ...cucggGGAGaUGuUCggauaauaacaAUGgaauacgucuacgca...      4018
                    -20       -10        +1       10
```

3.6.2 Transcription Termination Regions

The positions of 3' transcript end sites are located uniformly within (or in the case of the NAB monocistronic transcript immediately after) tracts of T residues and are often preceded by GC rich tracts but not by inverted repeats capable of forming stem - loop structures (Figure 11 Transcription Termination). Longer T tracts appear to result in more efficient termination. Tracts of Ts within coding regions are statistically much less frequent than expected. For active genes where protein products must be stoichiometrically balanced (i.e. ribosomal proteins), it might be important to minimize the potential for premature transcription termination.

3.6.3 Translation Initiation Regions

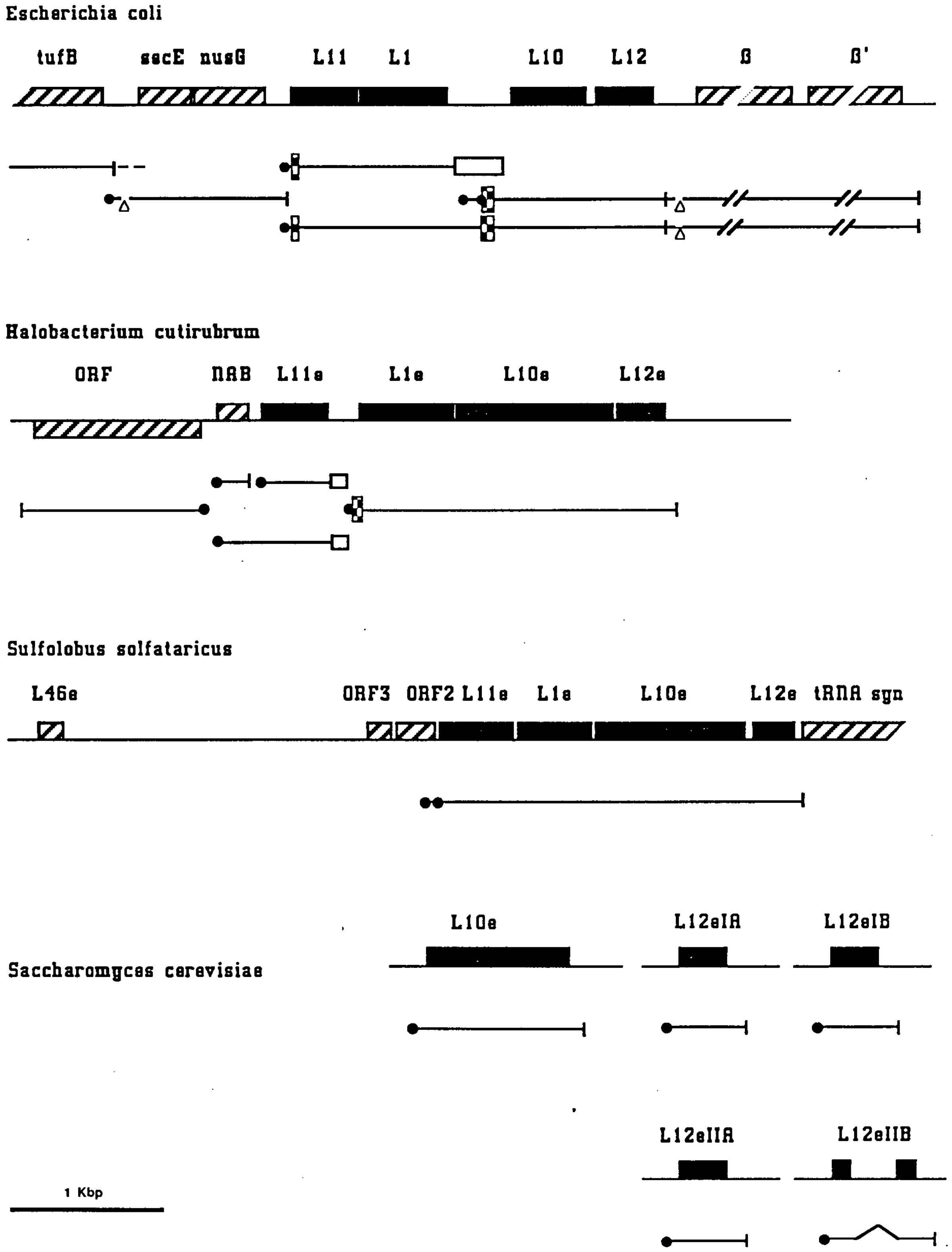
The regions surrounding the translation inititation sites on mRNAs derived from the 5.1 Kilobasepair fragment of genomic DNA are depicted (Figure 11 Translation Initiation Regions). Athough all of these regions exhibit sequences that are complementary to the 3' end of the *H. cutirubrum* 16S rRNA, the location of many of these matches precludes their identification as potential eubacterial Shine - Dalgarno sequences (Shine and Dalgarno, 1972; Hui and Dennis, 1985). The position of the complementary sequence is 3' to the AUG initiation codon on the ORF P2, NAB and L11e monocistronic transcripts that lack a 5' untranslated leader and is 5' to the AUG initiation codons on the ORF P1, NAB - L11e and L1e - L10e - L12e transcripts. The spacing between the last base of the complementary sequence and the first base of the initiation codons varies from -15 for ORF P2 to +11 for L12e (a negative value denotes a complementary sequence located 3' to the AUG initiation codon). It has not yet been demonstrated in halobacteria that these complementary sequences function to position ribosomes at authentic AUG initiation codons and other mechanisms (e.g. eucaryotic type 'thread on') may be involved. In addition to the potential for transcriptional regulation by the ORF P1 promotor as described previously the presence of the 5' Shine - Dalgarno site on the transcript initiated at ORF P1 may serve to increase the efficiency of translation of this transcript. Finally, it should be noted that the L12e cistron is translated about four fold more frequently than the preceding L1e or L10e cistrons. It is not yet clear how this translational enhancement is achieved.

3.7 Comparative Gene Organization and Expression

In the archaebacterium *H. cutirubrum* the genes encoding the four GTPase domain ribosomal proteins are linked in a single copy gene cluster in the order L11e, L1e, L10e and L12e, identical to the order of the homologous genes in the eubacterium *E. coli* (Figure 12; Post *et al.*, 1979; Shimmin and Dennis, 1989). Intergenic spacing between the L11e - L1e, L1e - L10e and L10e - L12e genes of *H. cutirubrum* are 203, 4 and 8 nucleotides respectively and compare to spacing of 6, 415 and 69 nucleotides respectively for the corresponding intergenic regions of *E. coli.* The Hcu ORF and Hcu NAB genes located upstream of the ribosomal protein gene cluster are not homologues of the secE and nusG genes occupying the comparable positions in *E. coli* and the β and β' RNA polymerase subunit genes, although preserved in *Halobacterium*, are not located distal to the L12 gene as in *E. coli* but rather located upstream of the S12e - S7e - EFGe - EFTue gene cluster (Zillig *et al.*, 1989). In the archaebacterium *S. solfataricus* the GTPase domain genes are also conserved in a single copy gene cluster with intergenic spacing of -1, -1 and 44 nucleotides between the L11e - L1e, L1e - L10e and L10e - L12e genes respectively (negative values denote overlapping genes; Shimmin *et al.*, 1989a; Ramirez *et al.*, 1990a). The *S. solfataricus* gene cluster is also flanked by unique genes: a tRNA synthetase located downstream and several open reading frames located upstream. Two Kilobasepairs upstream is a gene encoding a ribosomal protein with no homologue in eubacteria but homologous to L46 of the eucaryote *S. cerevisiae*. Less information is known of the homologous genes in eucaryota; the best studied organism is *S. cerevisiae* where the L10e and L12e genes have been characterized but the entire sequences of the L11e and L1e proteins have yet to be published (Otaka *et al.*, 1984; Remacha *et al.*, 1988; Mitsui and Tsurugi, 1988a; Mitsui and Tsurugi, 1988b; Mitsui and Tsurugi, 1988c; Newton *et al.*, 1990). The single L10e gene of *S. cerevisiae* has no introns and has been localized to chromosome XII (Newton *et al.*, 1990; C. Newton, personal communication). Some eucaryota (*Artemia salina, Drosophila melanogaster, Homo sapiens* and *S. cerevisiae*) are known to have two different types of L12e genes: L12e type I and L12e type II (Amons *et al.*, 1982; Wigboldus, 1987; Rich and Steitz, 1987; Newton *et al.*, 1990). *S. cerevisiae* is unique in having four distinct L12e genes, a pair of each type, named L12eIA, L12eIB, L12eIIA and L12eIIB (Figure 12; Newton *et al.*, 1990). The average identity between the type I and type II proteins is 20%, that between the A and B copies of each type is 54%; thus the duplication of the type I

Figure 12 Summary of the Structure and Expression of the L11e, L1e, L10e and L12e Genes within the Urkingdoms

The organization and transcription of the L11e, L1e, L10e, and L12e ribosomal protein gene clusters of *Escherichia coli*, *Halobacterium cutirubrum*, *Sulfolobus solfataricus* and *Saccharomyces cerevisiae* are depicted. Ribosomal protein encoding genes are solid boxes and other protein encoding genes or open reading frames are striped boxes. Genes above the line are oriented and transcribed rightwards and those below the line are oriented and transcribed leftwards. The open circles (○) represent 5' transcript ends and the vertical lines (|) represent 3' transcript ends. The open boxes (□) at the ends of trancripts indicate regions of multiple 3' trancript ends. The dashed lines (- -) indicate a small amount of readthrough transcription. In the *Escherichia coli* diagram the tufB, secE and nusG genes encode proteins functioning in translation elongation, protein secretion and transcription termination respectively. The β and β' genes encode the β and β' subunits of RNA polymerase respectively. Triangles (Δ) represent ribonuclease processing sites and the vertical line on the transcripts running through the L12 - β intergenic space represents a transcription attenuator. The checkered boxes (⊞) represent sites of autogenous regulation in *Escherichia coli* and putative autogenous regulation in *Halobacterium cutirubrum*. In *Saccharomyces cerevisiae* the Sce L12eIIB gene contains an intron and the Sce L11e and Sce L1e genes have yet to be characterized.


Escherichia coli
tufB
secE
nusG
L11
L1
L10
L12
β
β'
Halobacterium cutirubrum
ORF
NAB
L11e
L1e
L10e
L12e
Sulfolobus solfataricus
L46e
ORF3
ORF2
L11e
L1e
L10e
L12e
tRNA syn
Saccharomyces cerevisiae
L10e
L12eIA
L12eIB
L12eIIA
L12eIIB
1 Kbp

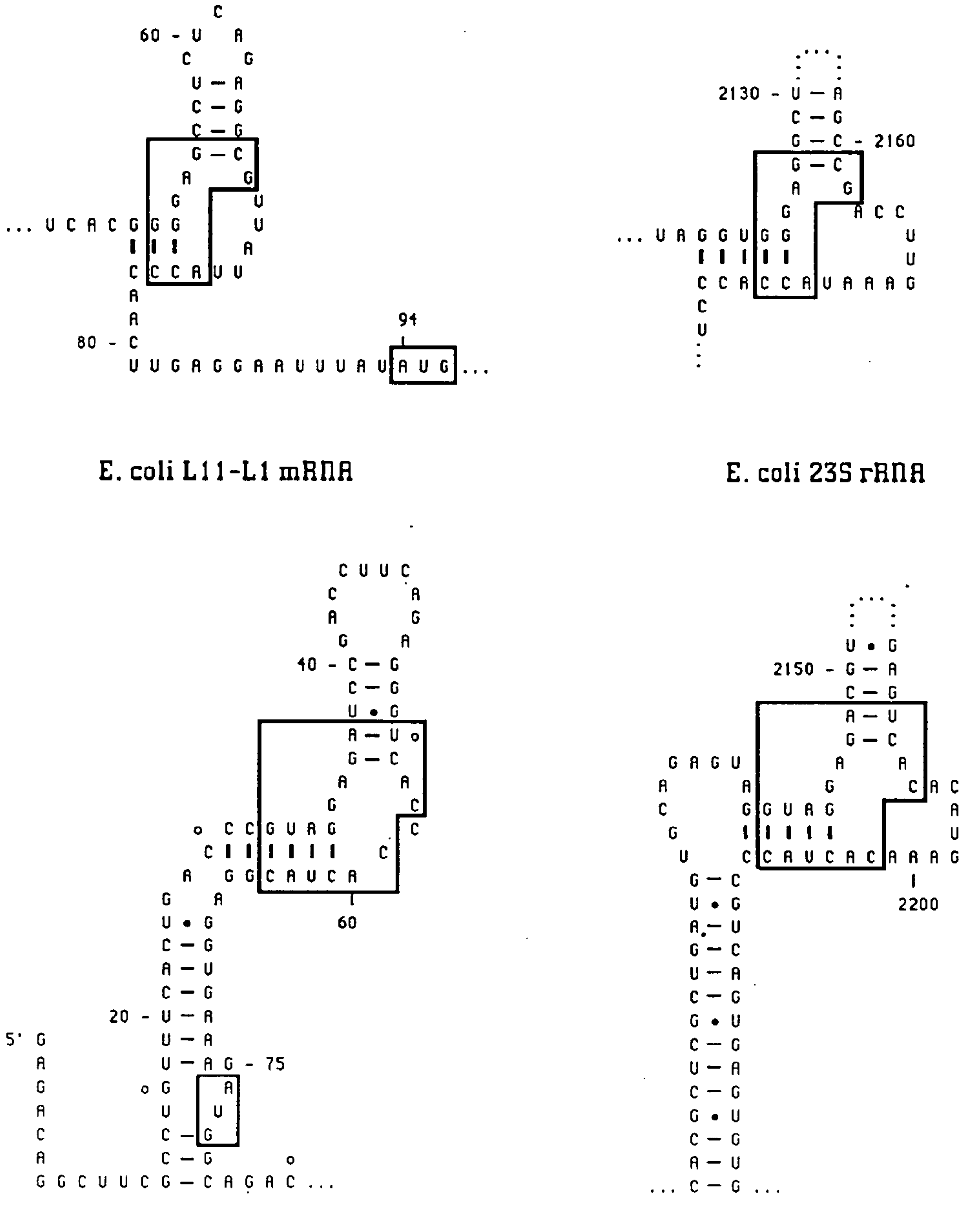
and II genes into A and B copies is apparently a relatively recent event of approximately 1000 million years ago whereas the type I: type II divergence is very ancient (Newton *et al.*, 1990). The three Sce L12eIA, L12eIB and L12eIIA genes are intronless whereas the Sce L12eIIB gene contains a 301 nucleotide long intron between codons 38 and 39 (Remacha *et al.*, 1988; Newton *et al.*, 1990). The genes are not closely linked; although the L12eIA, L12eIB and L12eIIB genes appear to be located on chromosome IV, the L12eIIA gene is located on either chromosome VII or XV (C. Newton, personal communication). The functional significance of two different L12e - like proteins in eucaryotes and apparently only one in archaebacteria and eubacteria remains to be elucidated.

Transcription patterns of the GTPase domain ribosomal protein genes of eubacteria, archaebacteria and eucaryota are each unique (Figure 12). In *E. coli* the genes are divided 2 and 2 ; both bicistronic L11 - L1 and L10 - L12 transcripts and tetracistronic L11 - L1 - L10 - L12 transcripts are produced and transcripts encoding the downstream β and β' genes are produced by elongation of a fraction of the transcripts exiting the L12 gene. The L1 protein and L10 - L12 complex autogenously regulate the expression of the GTPase domain proteins.

In *H. cutirubrum* the ribosomal protein genes are transcribed primarily into monocistronic (L11e) and tricistronic (L1e - L10e - L12e) transcripts; the bicistronic NAB - L11e transcript represents about 10% of the total L11e mRNA. Regions of the 23S rRNA gene of *E. coli* involved in binding of the Eco L11 and Eco L1 proteins (i.e. nucleotides 1052 - 1112 and 2100 - 2200 respectively) are homologous to nucleotides 1142 - 1201 and 2123 - 2222 of the 23S rRNA gene of *H. halobium* indicating the conservation of the L11e and L1e binding domains in the rRNA (Mankin and Kagramanova, 1986). The L1e gene is transcribed as the proximal cistron in the tricistronic L1e - L10e - L12e mRNA and is preceded by a 75 nucleotide long untranslated leader. The leader contains a region that has a sequence and structure almost identical to a region within the L1e binding domain in 23S rRNA (Figure 13). Furthermore, both the primary nucleotide sequence and secondary structure of these sites are highly similar to the *E. coli* L11 - L1 mRNA leader sequence that has been implicated in autogenous translational regulation. In *H. cutirubrum* the L11e gene is transcribed usually as a monocistronic mRNA lacking a 5' untranslated leader; a search in and around the gene for sequences resembling the L11e and L1e binding domains in 23S rRNA has been negative. It is possible that the NAB protein and/or the small

Figure 13 Autogenous Translational Regulation of the L1e - L10e - L12e Transcript

The binding domain of ribosomal proteins L1 and L1e on *Halobacterium halobium* and *Escherichia coli* 23S rRNA (right) and regions in the leader of the *Halobacterium cutirubrum* L1e - L10e - L12e and the *Escherichia coli* L11 - L1 mRNA are illustrated. Regions that exhibit sequence and structural similarity to each other and to the binding domain on 23S rRNA are depicted (boxed). The 5' ends of the mRNAs are nucleotide +1. The L1e AUG initiation codon on the *Halobacterium cutirubrum* L1e - L10e - L12e mRNA (position 75) corresponds to nucleotide 2314 in Figure 5. The open circles (o) next to positions 17, 27, 54 and 84 in the *Halobacterium cutirubrum* L1e - L10e - L12e mRNA show the positions of the minor ends present in the S1 nuclease and primer extension analysis of Figure 10.



E. coli L11-L1 mRNA

E. coli 23S rRNA

H. cutirubrum L1e-L10e-L12e mRNA

H. halobium 23S rRNA

amount of bicistronic NAB - L11e mRNA may have some regulatory significance. A further search within the L1e - L10e - L12e transcript for the homologue of the *E. coli* L10 - L12 autogenous regulation region was also negative.

In the distantly related archaebacterium *S. solfataricus* the four GTPase domain ribosomal proteins are apparently transcribed from twin promoters on a single tetracistronic transcript (Ramirez *et al.*, 1990a). No feature homologous to the autogenous regulatory sites has been identified within or flanking the transcript. The mechanism of enhancement of expression of the L12e proteins within the archaebacteria and eubacteria to maintain a stoichiometry of four protein copies per gene is unknown; it remains to be seen whether this regulatory feature has an origin coincident with the development of the gene cluster.

In eucaryota the L11e and L1e genes have yet to be characterized and the L10e and L12e genes are transcribed in unlinked monocistronic transcripts (Mitsui and Tsurugi, 1988a; Mitsui and Tsurugi, 1988b; Mitsui and Tsurugi, 1988c; Remacha *et al.*, 1989; Newton *et al.*, 1990). In *S. cerevisiae* the four distinct L12e genes are transcribed at greatly differing levels (Newton *et al.*, 1990).

Archaebacterial transcription signals have been identified by alignment of sequences at known transcription initiation and termination sites and identification of conserved consensus sequences. Actual binding of RNA polymerase *in vitro* to regions upstream of transcription initiation sites has been observed for only four methanogen genes; actual initiation *in vitro* at the *in vivo* transcription initiation sites has yet to be demonstrated (Thomm *et al.*, 1987; Brown *et al.*, 1988; Thomm and Wich, 1988; Thomm *et al.*, 1989).

A universal archaebacterial hexanucleotide promotor motif (TTTAAA) located approximately 25 nucleotides upstream of the transcription initiation site and bearing similarity to the eucaryotic RNA polymerase II promotor motif but not to eubacterial promotors has been proposed although with the subsequent increase in available sequences there appear to be substantial variations within the thermoacidophile, methanogen and halophile groups (Reiter *et al.*, 1988). Thermoacidophiles exhibit the greatest fidelity to the universal archaebacterial promotor whereas the methanogens have an extended AT rich sequence (TTTAA/TATA) that is generally well conserved (Figure 14). Within the halophiles the core promotor element appears to be shortened to the tetranucleotide TTAA motif. More extensive consensus sequences that have been proposed based primarily on stable RNA transcripts tend to be

Figure 14 Archaebacterial Transcription Initiation Sequences

Archaebacterial transcription initiation sites are aligned at the initiation nucleotide at position +1 (scale at bottom). Consensus sequences are illustrated on the first line of each group; the A or G transcription initiation sites are in all cases 20 to 25 nucleotides 3' to the consensus sequence motifs. Matches to the consensus sequences are indicated in upper case characters. The species abbreviations are: Mva, *Methanococcus vanielii* ; Tte, *Thermoproteus tenax* ; S B12, *Sulfolobus* strain B12. References appearing at the end of each entry are: 1, Shimmin and Dennis, 1989 and this work; 2, Chant *et al.*, 1986; 3, Dunn *et al.*, 1981; 4, Betlach *et al.*, 1986; 5, Blanck and Oesterhelt, 1987; 6, Cue *et al.*, 1985; 7, Cram *et al.*, 1987; 8, Wich *et al.*, 1986; 9, Wich *et al.*, 1987; 10, Reiter *et al.*, 1988.

```
HALOPHILES   (TTCGA) ... TTAA ... A or G initiation

Hcu ORF P1             TgCGAtggcttcactcggcgtaTTAAcgtgtcgaaacgatcccaccG    1
Hcu ORF P2                TgCGgaggaaagcgcTTtcggcgcttgctgtctacgggcacgtG    1
Hcu NAB                       TTCGAcacgTTAAtacgccgagtgaagccatcgcatagtG    1
Hcu L11e                 cTCGAaagacaagggTTAAacccgcggcggcggtttctcggagtA    1
Hcu L1e                tgcTTCGcTTCGAcgcttTTAAgcccgggatcaccgtctgtagaaccG    1
Hcu rRNA P1            tggTTCGAcggtgttTTAtgtaccccaccactcggatgagatgcgaA    2
Hha bop              atactgattgggtcgTatAgttacacacatatcctcgttaggtactgttG    3
Hha brp                  tagcttgggtcttttTTgAtgctcggtagtgacgtgtgtattcatA    4
Hha hop                  gttgggggaggttatTTAAtggcgtgccgtgtccttccgaacaccA    5

METHANOGENS   TTTAA/TATA ... A or G initiation

Mva hisA                 atttcttaggtaccaaTATATAtgttaaaacctaatttaacataG    6
Mva mcr                      ttaatgaaaacttgaaTATATcttcctttaataatgttatGA    7
Mva tRNA (pMT21)        taataaccgaaataTTTATATActagaatacccttcctatactatG    8
Mva rRNA (pMV1)           tacatacctaaaacaaTAcATAttacaacacgttttcatattatG    8

THERMOACIDOPHILES   TTTAAA ... A or G initiation

Tte rRNA                    agggagcgaaaaatTTTAAtttagggtgttttaggatggtcG    9
                                                       •
Tte tRNA (ala)                ctctagcgaaaaaaTTTAAAtcggtgagtaagtacgctgG    9
S B12 SSV1 T1+2          cagaactggaggggTTTAAAaacgtaagcgggaagccgatattG    10
S B12 SSV1 T3               ttagttaggctcttTTTAAAgtctaccttctttttcgcttacA    10
S B12 SSV1 T4                 agaagatagcccttTTTAAAgccataaatttttatcgcttA    10
                            -40       -30       -20       -10       +1
```

poorly conserved in polypeptide encoding transcripts (Betlach *et al.*, 1984; Mankin *et al.*, 1984; Dennis, 1985; Hui and Dennis, 1985; Chant *et al.*, 1986; Chant and Dennis, 1986; Daniels *et al.*, 1986; Blanck and Oesterhelt, 1987; Das Sarma *et al.*, 1987). Some of the extended features proposed may affect specific transcripts; the TTCGA element 5 to 15 nucleotides upstream of the TTAA motif proposed in this work is not common to all halophile transcripts but appears in a sufficient number of transcripts encoding both stable RNAs and polypeptides to suggest its participation in some aspect of transcription initiation.

Archaebacterial transcription termination appears to occur within poly T tracts that are sometimes preceded by GC rich sequences and/or short inverted repeats, reminiscent of rho independent termination in eubacteria. The paucity of poly T tracts displayed within the coding strands of genes within the extreme halophiles is probably due to the apparent efficacy of T residues to facilitate transcription termination within the GC rich DNA of these organisms. This is best illustrated by the Hcu L11e transcript where all seven tracts of two to four Ts in the L11e - L1e intergenic space result in termination (Figure 9 and Figure 11).

Classic Shine - Dalgarno sequences facilitating translation initiation are usually evident immediately upstream of the initiator methionine codon in the thermoacidophile and methanogen transcripts. Translation initiation in halophiles is enigmatic; some transcripts such as Hcu L11e, Hha S12e and Hha brp have no leader nucleotides at all, others have very short leaders. Various mechanisms for positioning of the ribosome at the authentic methionine initiator codon have been suggested, including a eucaryotic type 'thread on' mechanism, ribosomal recognition of short hairpin structures formed at the 5' end of leaderless transcripts and classic Shine - Dalgarno type positioning when sequences complementary to the 3' end of the 16S rRNA are present either in an appropriate position 5 to 10 nucleotides preceding the initiator methionine codon or even within the coding region in transcripts with negligible leaders (Dunn *et al.*, 1981; Betlach *et al.*, 1984; Betlach *et al.*, 1986; Blanck and Osterhelt, 1987; Shimmin and Dennis, 1989). Elucidation of ribosome binding to mRNA awaits further study.

3.8 Summary

The organization of the GTPase domain genes is conserved in the eubacterium *E. coli*, the archaebacteria *H. cutirubrum* and *S. solfataricus* but not in the eucaryote *S. cerevisiae*. Most features of the transcription and regulation patterns differ between these diverse organisms. The autogenous regulation of translation by the L1 protein in *E. coli* is conserved in *H. cutirubrum*, whether it is conserved in *S. cerevisiae* is unknown. The mechanisms by which the Hcu L11e protein is regulated and the L12e protein is amplified remain to be elucidated.

The cloned 5146 basepair fragment of *H. cutirubrum* genomic DNA encodes two potential proteins of unknown function (ORF and NAB) and the four GTPase domain ribosomal proteins; the ORF gene is oriented opposite to the remaining genes. Transcription analysis demonstrates that the ORF is encoded on a very rare monocistronic transcript initiated at two distinct promotors, rare initiation occurs on a G residue yielding a 49 nucleotide leader and the majority transcript initiates on a G residue 1 nucleotide before the putative translation initiation codon. Termination occurs in a poly T tract. The NAB and L11e proteins are encoded on both individual monocistronic and a bicistronic transcript. The NAB monocistronic and NAB - L11e bicistronic transcripts share an initiation site at a G residue 1 nucleotide before the translation initiation codon. The monocistronic L11e transcript has no 5' leader sequence, initiating precisely on the A residue of the ATG translation initation codon. All three transcripts terminate in T tracts. A tricistronic transcript with a 75 nucleotide 5' leader sequence containing the site for autogenous regulation of translation by the L1e protein and terminating in a tract of 6 T residues encodes the three ribosomal proteins L1e, L10e and L12e. The level of transcription of the ribosomal proteins appears similar from Northern and S1 nuclease protection experiments. The NAB monocistronic transcript appears at a level approximately 1% of the monocistronic L11e transcript and the bicistronic NAB - L11e transcript is approximately 10% of the L11e monocistronic transcript level. The ORF transcript is expressed at a level approximately 500 fold lower than the ribosomal protein transcripts. Common elements upstream of the transcription initiation sites include the motif TTCGA ... 4 - 15 bp ... TTAA ... 20 - 26 bp ... A or G transcription start. The TTAA motif is conserved within the archaebacteria whereas the TTCGA motif is apparently unique to the extreme halophiles. Transcription termination occurs within poly T tracts that may be preceded by a GC rich region, a motif characteristic of most archaebacterial transcripts. The sequences or structures facilitating ribosome binding are not obvious; although the L1e and L10e cistrons are preceded by classic Shine - Dalgarno sequences, transcripts having negligible or nil 5' leaders (e.g. the L11e monocistronic transcript) must utilize some other mechanism to initiate translation.

Part 4: Structure and Function of the GTPase Domain Proteins

4.1 Amino Acid Composition

The amino acid compositions of the *E. coli* L11, L1, L10, and L12 ribosomal proteins and the equivalent proteins from two archaebacteria (*H. cutirubrum* and *S. solfataricus*) and a eucaryote (*S. cerevisiae*) are presented in Table 6. The composition of the proteins in the three urkingdoms are similar. The proteins have primordial amino acid compositions, with few or no cysteine or tryptophan residues, as is the case with ribosomal proteins in general (Wittmann-Liebold, 1986; Taylor, 1989). The *H. cutirubrum* proteins are, however, more acidic than those from other organisms; the aspartate and glutamate content is increased about two - fold while the lysine and arginine content is greatly reduced. The L12e proteins are alanine rich (average 21.3%) and acidic (average 22.6% acidic versus 7.0% basic residues). The Hcu L12e protein is the most extreme, with 36.8% acidic to 1.8% basic residues, partially due to the adaptation to their high salt enviroments. The eubacterial, archaebacterial and most eucaryotic type I L12e proteins contain a unique arginine residue whereas the eucaryotic type II contain a unique tryptophan residue.

4.2 The L11e proteins

The alignment of the Hcu L11e and Sso L11e proteins is shown in Figure 15. Both proteins can be aligned end to end with 40% amino acid sequence identity and only a single gap at position 65 in the Hcu L11e sequence (Table 7). The Sso L11e protein has an extra amino acid at its amino terminus, an unusual carboxy terminus containing (i) a tryptophan residue at alignment position 169 and (ii) a five residue extension. Alignment of the two archaebacterial L11e proteins with the eubacterial Eco L11 and the available amino terminal portion of the eucaryotic Sce L11e protein is also shown in Figure 15 (Post *et al.*, 1979; Otaka *et al.*, 1984). Identification of the products of the archaebacterial genes as the homologues of the Eco L11 ribosomal proteins is based on amino acid sequence similarity and also on physicochemical and genetic characterization (Matheson *et al.*, 1984; Matheson, 1985; Shimmin *et al.*, 1989; Shimmin and Dennis, 1989; Ramirez *et al.*, 1990a; Ramirez *et al.*, 1990b). The Eco L11 sequence aligns end to end with the two archaebacterial sequences requiring only one gap in the archaebacterial proteins at alignment position 50 and exhibit 33% and 32% amino acid identity over the common aligned region (Table 7). The

Table 6 Interkingdom Comparison of the Amino Acid Compositions of the L11e, L1e, L10e and L12e Ribosomal Proteins

Polypeptide length is in amino acid residues. Composition values are in mole percent.

		L11e			L1e			L10e				L12e						
		Eco	Hcu	Sso	Eco	Hcu	Sso	Eco	Hcu	Sso	Sce	Eco	Hcu	Sso	Sce	Sce	Sce	Sce
															IA	IB	IIA	IIB
POLYPEPTIDE LENGTH		142	163	170	234	212	221	165	352	335	312	121	114	105	110	106	106	106
ALANINE	A	14.1	12.3	7.1	14.1	10.8	8.1	20.0	11.6	10.7	11.5	23.1	24.6	17.1	20.9	20.8	18.9	23.6
ARGININE	R	2.8	1.8	0.6	4.7	7.1	5.0	7.3	4.0	1.5	4.2	0.8	0.9	1.0	0.9	-	-	-
ASPARAGINE	N	2.8	2.5	4.7	5.1	4.2	5.4	1.8	3.4	3.9	4.2	0.8	1.8	2.9	1.8	2.8	4.7	3.8
ASPARTATE	D	4.2	11.0	5.9	6.0	15.6	3.2	3.6	11.9	5.1	5.4	5.0	17.5	1.9	6.4	7.5	7.5	10.4
CYSTEINE	C	0.7	0.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.9
GLUTAMINE	Q	4.9	4.3	4.7	3.8	3.8	5:0	1.8	4.8	3.3	1.9	0.8	-	3.8	1.8	-	0.9	0.9
GLUTAMATE	E	4.2	11.6	5.9	4.7	6.6	7.7	9.1	11.6	7.8	8.7	13.2	19.3	15.2	14.6	15.1	14.2	10.4
GLYCINE	G	8.5 .	9.8	8.2	8.5	5.2	4.1	5.5	8.8	7.2	7.1	6.6	7.9	6.7	12.8	9.4	9.4	13.2
HISTIDINE	H	-	-	0.6	0.9	0.9	-	0.6	0.3	1.2	1.3	1.0	-	-	-	-	-	0.9
ISOLEUCINE	I	6.3	4.3	9.4	4.7	3.3	7.2	3.0	3.4	9.3	5.8	5.0	2.6	6.7	3.6	4.7	4.7	5.7
LEUCINE	L	4.9	6.7	8.8	7.3	7.5	10.0	9.1	8.8	8.4	8.0	6.6	7.0	8.6	8.2	9.4	10.4	7.5
LYSINE	K	10.6	3.1	12.9	9.8	2.4	13.1	7.3	2.0	12.2	4.8	10.7	0.9	11.4	6.4	6.6	4.7	4.7
METHIONINE	M	4.2	0.6	1.8	3.0	3.8	2.3	3.6	1.1	1.5	1.9	3.3	0.9	1.9	1.8	1.9	1.9	1.9
PHENYLALANINE	F	2.8	3.1	0.6	1.7	2.4	3.2	3.6	1.4	3.9	4.8	1.7	0.9	1.0	2.7	1.9	3.8	4.7
PROLINE	P	6.3	6.7	6.5	3.0	4.7	5.9	3.0	4.8	4.5	4.8	1.7	2.6	2.9	1.8	2.8	1.9	0.9
SERINE	S	5.6	3.1	5.9	3.4	3.3	4.1	3.6	6.0	3.6	7.1	5.0	3.5	5.7	7.3	7.5	7.5	6.6
THREONINE	T	6.3	6.7	8.2	5.6	6.1	5.0	5.5	5.4	6.6	4.5	2.5	1.8	3.8	1.8	2.8	2.8	2.8
TRYPTOPHAN	W	-	-	0.6	-	-	0.5	-	0.3	0.3	0.3	-	-	-	-	-	0.9	0.9
TYROSINE	Y	1.4	1.8	1.8	1:3	0.5	1.8	1.8	1.4	2.7	3.2	-	1.8	1.9	1.8	1.9	0.9	0.9
VALINE	V	9.2	9.8	6.5	12.4	11.8	8.6	9.1	8.8	6.6	10.6	13.2	6.1	6.7	5.5	4.7	4.7	3.8

Figure 15 Alignment of the L11e Proteins

Upper Panel

The amino acid sequences of the L11e ribosomal proteins from the archaebacteria *Halobacterium cutirubrum* (Hcu L11e) and *Sulfolobus solfataricus* (Sso L11e), predicted from the nucleotide sequences of the respective genes and partially confirmed by amino acid sequencing, are aligned with the L11e ribosomal proteins from the eubacterium *Escherichia coli* (Eco L11) and the eucaryote *Saccharomyces cerevisiae* (Sce L11e). Only the amino terminal 21 amino acid residues of the Sce L11e protein have been determined. Gaps required for alignment are indicated as dashes (-). The amino acid position scale is below the Sce L11e sequence and the proteins are aligned over 170 amino acid positions. Similarities are indicated between the eubacteria and eucaryota (above the Eco L11 sequence), between the eubacteria and archaebacteria (between the eubacterial and archaebacterial sequences), within the archaebacteria (between the Hcu L11e and Sso L11e sequences) and between the archaebacteria and eucaryota (between the archaebacterial and eucaryotic sequences) by the following symbols:

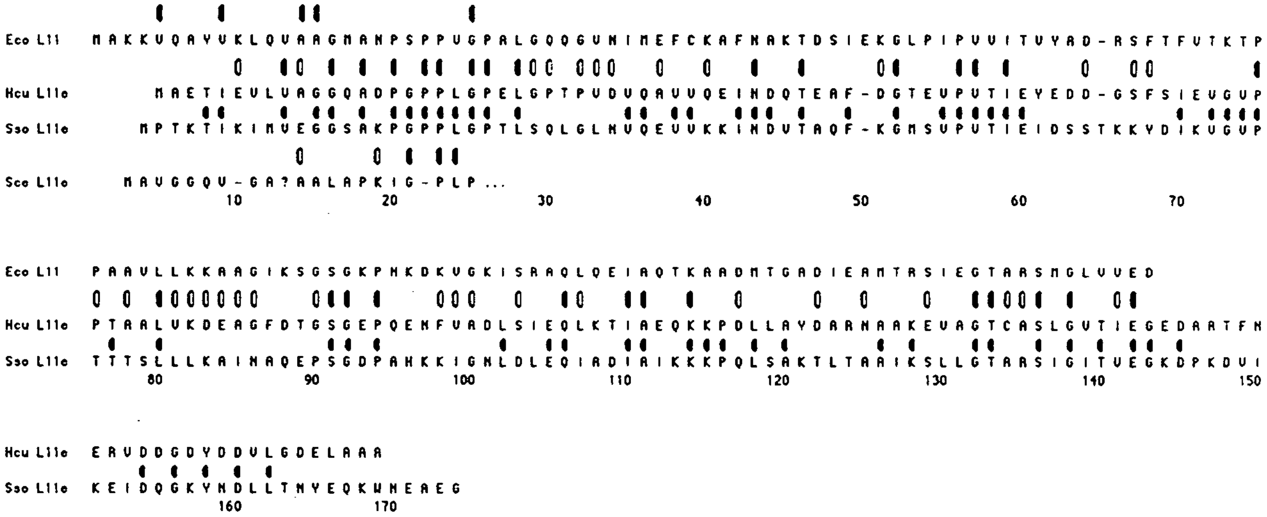
▪ Intrakingdom identities within the archaebacteria

▮ Interkingdom identities where all positions within 2 urkingdoms are identical

▯ Interkingdom identities where 2 out of 3 positions within 2 urkingdoms are identical

Lower Panel

Line diagrams illustrate the end to end interkingdom alignment of the archaebacterial Hcu L11e and Sso L11e proteins to the complete eubacterial Eco L11 and the partial eucaryotic Sce L11e sequences. Gaps required to maintain maximum identity in the alignment are indicated by white bars. The common scale is in amino acid residue positions along the linear interkingdom alignment.



```
Eco L11   M A K K V Q A Y V K L Q V A A G M A N P S P P V G P A L G Q Q G V N I M E F C K A F N A K T D S I E K G L P I P V V I T V Y A D - R S F T F V T K T P
Hcu L11e        M A E T I E V L V A G G Q A D P G P P L G P E L G P T P V D V Q A V V Q E I N D Q T E A F - D G T E V P V T I E Y E D D - G S F S I E V G V P
Sso L11e      M P T K T I K I M V E G G S A K P G P P L G P T L S Q L G L N V Q E V V K K I N D V T A Q F - K G M S V P V T I E I D S S T K K Y D I K V G V P
Sce L11e      M A V G G Q V - G A ? A A L A P K I G - P L P ...
                        10                  20                  30                  40                  50                  60                  70

Eco L11   P A A V L L K K A A G I K S G S G K P N K D K V G K I S A A Q L Q E I A Q T K A A D M T G A D I E A M T A S I E G T A A S M G L V V E D
Hcu L11e  P T A A L V K D E A G F D T G S G E P Q E N F V A D L S I E Q L K T I A E Q K K P D L L A Y D A R N A A K E V A G T C A S L G V T I E G E D A A T F N
Sso L11e  T T T S L L L K A I N A Q E P S G D P A H K K I G N L D L E Q I A D I A I K K K P Q L S A K T L T A A I K S L L G T A A S I G I T V E G K D P K D V I
              80                  90                  100                 110                 120                 130                 140                 150

Hcu L11e  E A V D D G D Y D D V L G D E L A A A
Sso L11e  K E I D Q G K Y N D L L T N Y E Q K W N E A E G
                        160                 170
```

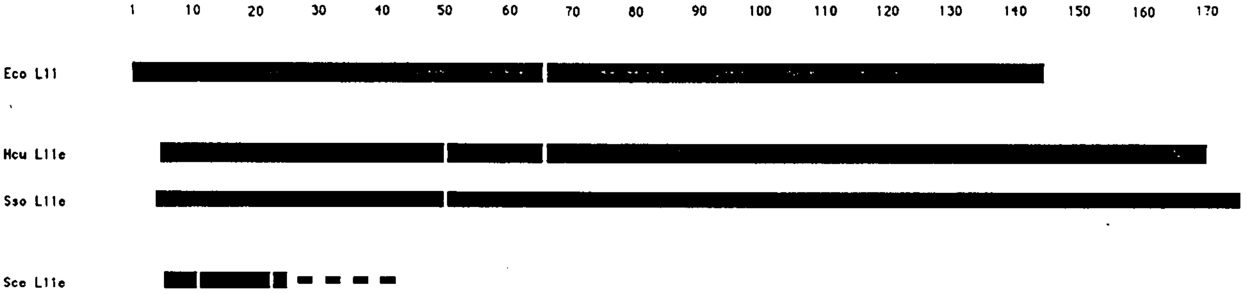


Table 7 **Intra and Interkingdom Sequence Similarity of the L11e, L1e and L10e Ribosomal Proteins**

Proteins Compared	% Amino Acid Identity	Deletion / Insertion[1] Index	Significance[2] (Z)
L11e			
Hcu/Sso	40	1	45
Eco/Hcu	33	1	35
Eco/Sso	33	2	32
L1e			
Bst/Eco	50	0	62
Hcu/Sso	31	4	40
Bst/Hcu	29	11	30
Bst/Sso	28	9	21
Eco/Hcu	32	11	36
Eco/Sso	26	9	19
L10e			
Hcu/Sso	27	1	45
Hsa/Sce	53	4	78
Eco/Hcu	24	6	10
Eco/Sso	21	6	10
Eco/Hsa	15	6	2
Eco/Sce	17	6	3
Hcu/Hsa	25	8	42
Hcu/Sce	24	9	35
Sso/Hsa	24	8	25
Sso/Sce	24	8	27

[1] The number of insertions and deletions introduced into the aligned pair of proteins

[2] Significance value is from the RDF program of Lipman and Pearson (1985).

Hcu L11e and the Sso L11e proteins have carboxy terminal extensions of 26 and 31 residues respectively compared to the Eco L11 protein. The Sso L11e protein has an extra threonine residue at position 65 which is not present in the Hcu L11e or Eco L11 protein. The Eco L11 protein has a short amino terminal extension of 3 or 4 residues that is not present in the archaebacterial proteins.

The Eco L11, Hcu L11e and Sso L11e proteins contain 9, 11 and 10 proline residues respectively within the aligned region. Conservation of the proline residues in all three sequences occurs at seven positions (positions 20, 22, 23, 26, 56, 75 and 94 of Figure 15). The unique cysteine residue of the Eco L11e protein (position 39) is not present in the Sso L11e protein and present but not positionally conserved in the Hcu L11e protein (position 134). Between the eubacterial and archaebacterial L11e proteins there exist two short regions of high sequence conservation. The amino terminal region contains the highest concentration of conserved residues (positions 13 - 28) and is characterized by four conserved proline residues at positions 20, 22, 23 and 26. The second region is located in the carboxy terminus of Eco L11 (positions 132 - 142) and contains conserved threonine and serine residues at positions 133 and 136 respectively. Regions of high amino acid sequence similarity in homologous proteins that are separated by large distances in evolutionary time are usually indicative of a conserved structural or functional domain in the protein.

A number of structural and functional features of the Eco L11e protein have been reported. The axial ratio of the Eco L11e protein is 5.5:1; the conserved proline residues probably contributing to the highly elongated shape of the molecule (Giri *et al.*, 1978). The amino terminal domain of the Eco L11 protein (residues 1 - 64, positions 2 - 66) has been implicated in the interaction of the ribosome with translation release factor 1 (Tate *et al.*, 1984). Thus the highly conserved amino terminal region of the L11e proteins between positions 13 - 28 may be responsible for this interaction.

The Eco L11 protein binds to a conserved region of the *E. coli* 23S rRNA located between nucleotides 1052 and 1112 (Schmidt *et al.*, 1981). Heterologous binding studies have demonstrated that the Eco L11 protein can bind to a conserved domain in the large subunit rRNA from both archaebacteria and eucaryotes (Beauclerk *et al.*, 1985; El-Baradi *et al.*, 1987). Both the nucleotide sequence and secondary structure of the homologous region within the *H. halobium* 23S rRNA (i.e. nucleotides 1142 - 1201) are remarkably similar to the L11 binding domain in *E. coli* 23S rRNA (Mankin *et al.*, 1986; Shimmin and Dennis, 1989). The domain within the L11e proteins that binds to the conserved target within the large subunit rRNA has yet to be elucidated.

In *E. coli* the L11 protein is the most highly methylated ribosomal protein, containing nine methyl groups that are added to the protein after translation. The modifications are a trimethyl-alanine at the amino terminal residue (the initiator methionine residue is removed) and two trimethyl-lysines at residues 3 and 39 (alignment positions 2, 4 and 40 respectively in Figure 17; Dognin and Wittman-Liebold, 1977). The first two positions (2 and 4) are within the amino terminal region unique to the Eco L11 protein and not part of the archaebacterial or eucaryotic proteins. The *E. coli* lysine (39) residue is conserved in the Sso L11e protein but not in the lysine depleted Hcu L11e protein. Like their eubacterial homologue, the initiator methionine residue is post-translationally removed from the amino terminus of the L11e proteins of *H. cutirubrum*, *S. solfataricus* and *S. cerevisiae* ; however, none of these proteins contains the amino terminal methylation modification (Matheson *et al.*, 1984; Otaka *et al.*, 1984; Matheson, 1985). Thus sites of methylation appear not to be conserved between eubacteria and archaebacteria.

The Eco L11e protein is an important component of the GTPase domain of the 50S subunit and is involved in the synthesis of guanosine 5' diphosphate, 3' diphosphate (ppGpp) during the stringent response (Friesen *et al.*, 1974; Parker *et al.*, 1976; review: Cundliffe, 1986). Methanogens and halophiles apparently lack a stringent response (Beauclerck *et al.*, 1985; Chant, J and Dennis P.P, unpublished results).

4.3 The L1e proteins

The Hcu and Sso L1e amino acid sequences can be aligned end to end with the introduction of four small gaps at alignment positions 19, 30 - 31, 93 and 207 and exhibit 31% amino acid identity (Figure 16, Table 7). The Sso L1e sequence has extensions of four and two residues at its amino and carboxy terminal ends respectively. There are two highly conserved regions between the archaebacterial proteins, the first from positions 137 - 151 exhibiting 12 out of 15 identical residues and the second between positions 224 - 235 exhibiting 10 out of 12 identical residues.

The alignment of the two archaebacterial proteins with their somewhat longer eubacterial counterparts is also shown in Figure 16. Optimal eubacterial to archaebacterial alignment required introduction of eight

Figure 16 Alignment of the L1e Proteins

Upper Panel

The amino acid sequences of the L1e ribosomal proteins from the archaebacteria *Halobacterium cutirubrum* (Hcu L1e) and *Sulfolobus solfataricus* (Sso L1e), predicted from the nucleotide sequences of the respective genes and for the Hcu L1e protein partially confirmed by amino acid sequencing, are aligned with the L1e ribosomal proteins from the eubacteria *Bacillus stearothermophilus* (Bst L1e) and *Escherichia coli* (Eco L1). There is no available sequence representing the eucaryota. Gaps required for alignment are indicated as dashes (-). The amino acid position scale is below the Sso L1e sequence and the proteins are aligned over 230 amino acid positions. Similarities are indicated within the eubacteria (between the Bst L1e and Eco L1 sequences), between the eubacteria and archaebacteria (between the eubacterial and archaebacterial sequences) and within the archaebacteria (between the Hcu L11e and Sso L11e sequences) by the following symbols:

▮ Intrakingdom identities within the eubacteria and the archaebacteria

▮ Interkingdom identities where all positions within the 2 urkingdoms are identical

▯ Interkingdom identities where 3 out of 4 positions within the 2 urkingdoms are identical

Lower Panel

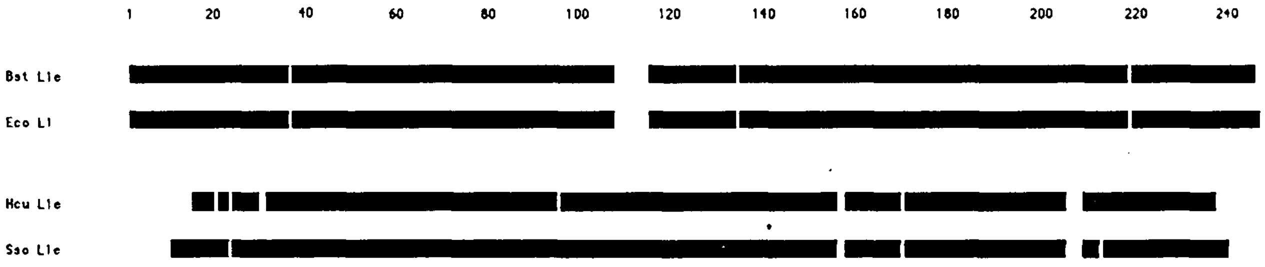
Line diagrams illustrate the end to end interkingdom alignment of the archaebacterial Hcu L1e and Sso L1e proteins to the complete eubacterial Bst L1e and Eco L1 sequences. Gaps required to maintain maximum identity in the alignment are indicated by white bars. The common scale is in amino acid residue positions along the linear interkingdom alignment.

```
Bst L1e  MPKVDKKYLEALKLVDASKAYPIAQAIEIVKKTNV-AKFDATVEVAFAL-GVDPKKADQQIRGAVVLPHGTGKVA
Eco L1   MAKLTKRMRVIREKVDATKQYDINEAIALLKELAT-AKFVESVDVAVNL-GIDARKSDQNVRGATVLPHGTGRSV
Hcu L1e               MADND-IEE-AVARAL--EDAPQANFAETVDLAVNLADLDLNDPSQRVDEGVVLPSGTGQET
Sso L1e           MKKVLADKESLIE-ALKLALSTEYNVKANFTQSVEIILTFKGIDMKKGDLKLAEIVPLPKQPSKAK
                 10        20        30        40        50        60        70

Bst L1e  RVLVFAKGEKAKEAEAAGADYVGDTEYIMKI-------QQGWFDFDVVVATPDMMGE-VGKLGAIIGPKGLMPNP
Eco L1   RVAVFTQGANAEAAKAAGAELVGMEDLADQI-------KKGEMNFDVVIASPDAMRV-VGQLGQVLGPRGLMPNP
Hcu L1e  QIVVFADGETAVAADDV-ADDVLDEDDLSDLADDTDAAKDLADETDFFVAEAPMMQDIVGALGQVLGPAGKMPTP
Sso L1e  AVLVVPSFEQLEYAKKASPNVVITAEELQKLQGQKAPVKKLAIQHEWFLINQESMALAGAILGPALGPAGKFPTP
            80        90       100       110       120       130       140       150

Bst L1e  KTGTVTFDVAKAVQEIKAGKVEYAVDKAGNIHVPIGKVSFDNEKLAENFAAVYEAIIKAKPAAAKG-TYVKNVTI
Eco L1   KVGTVTPNVAEAVKNAKAGQVRYRNDKNGIIHTTIGKVDFDADKLKENLEALLVALKKAKPTQAKG-VYIKKVSI
Hcu L1e  LQP--DDDVVDTVNAMK-NTVQIASADAATFNTAVGAEDMSAEDIASNIDVI---MAALHANLEKGPLNVDSVYV
Sso L1e  LPN--TADISEYINAFK-ASVIVKTKDQPQVQVFIGTEDMKPEDLAENAIAV---L-NAIENKAKVETNLANIVV
                160       170       180       190       200       210       220

Bst L1e  TSTMGPGIKVDPTTVAVAQ
Eco L1   STTMGAGVAVDQAGLSASVN
Hcu L1e  KTTMGPAVEVA
Sso L1e  KTTMGKAVKVKAA
           230       240
```



additional gaps at alignment positions: 23, 36, 50, 107 - 113, 133, 154 - 155, 168 and 203 - 205. The interkingdom amino acid sequence similarity of the L1e proteins from the two urkingdoms ranges from 26% to 32% amino acid identity (Table 7). The two highly conserved regions within the archaebacterial L1e proteins are also conserved in their eubacterial counterparts. The conserved region of 14 residues centrally located within the L1e proteins (positions 137 - 150) is characterized by three conserved proline residues at positions 143, 148 and 150. The second region is found in the carboxy terminus of the L1e proteins (alignment positions 227 - 235) and is characterized by conserved hydroxyl (threonine or serine) residues at positions 227 and 228.

The Hcu L1e protein contains a nearly perfect heptapeptide repeat DLAD (D/E) TD at positions 105 - 111 and 115 - 121 that is present but less well conserved in the Sso L1e sequence (KLQGGKR and KLAIQNE). Since the eubacterial Eco L1 and Bst L1e sequences contain a corresponding heptapeptide gap in this region, two evolutionary scenarios are possible. The heptapeptide duplication occurred before divergence and was followed by a deletion in the ancestral L1e gene of the eubacterial lineage after divergence from archaebacteria or the repeat may have arisen through a partial duplication event in the archaebacterial ancestral gene after divergence from eubacteria; the latter possibility is the most parsimonious.

The L1 protein of *E. coli* has been localized to the ridge region on the 50S subunit opposite the L12 stalk and binds to and protects nucleotides 2100 - 2200 of 23S rRNA (Branlant *et al.*, 1981; Lake and Strycharz, 1981; Oakes *et al.*, 1986). The protein functions to (i) maximize binding of peptidyl - tRNA to the P site, (ii) maximize the GTPase activity associated with EFG mediated translation and (iii) autogenously regulate the translation of the L11 - L1 mRNA (Subramanian and Dabbs, 1980; Sander, 1983; Dean and Nomura, 1980; Baughman and Nomura, 1981; Yates and Nomura, 1981; Thomas and Nomura, 1987; Kearney and Nomura, 1987). The binding domain of the L1e protein in the large subunit RNA of archaebacteria and eukaryota has been sufficiently conserved such that it can still be recognized and protected by Eco L1 *in vitro* implying that the domain within the L1e protein is probably also highly conserved (Zimmerman *et al.*, 1980; Gourse *et al.*, 1981). This conserved domain, responsible for the autogenous control of the L11 - L1 mRNA in *E. coli*, probably autogenously regulates the tricistronic L1e - L10e - L12e mRNA of *H. cutirubrum* (Shimmin and Dennis, 1989). At the present time correlations between the conserved primary structure regions of the L1e protein and its functional role in binding rRNA, autogenous regulation, peptidyl - tRNA binding to the P site and the GTPase activity associated with EFG mediated translation cannot be made.

4.4 The L10e proteins

The end to end alignment of the 352 and 337 amino acid long archaebacterial L10e proteins from *H. cutirubrum* and *S. solfataricus* is illustrated and exhibits 27% amino acid sequence identity (Figure 17, Table 7). The proteins exhibit greater than 30% amino acid sequence identity distributed relatively uniformly with no deletions or insertions through the first 302 positions. Identities beyond position 302 are negligible except for the extreme carboxy terminus (positions 365 - 371). Although no identities occur between positions 335 and 362, the lack of identity is partially the result of modification required for the adaptation to a high salt environment in the halophilic archaebacteria. This region in both proteins is rich in charged amino acids; the Hcu L10e protein containing twelve aspartic acid and two glutamic acid residues and the Sso L10e protein containing nine glutamic acid and six lysine residues. An alanine - proline rich region (positions 318 - 331) precedes the charged region in the Hcu L10e protein but is absent from the Sso L10e protein.

The *H. cutirubrum* and *S. solfataricus* L10e proteins can be aligned to the shorter eubacterial *E. coli* L10e protein yielding 24% and 21% amino acid sequence identity respectively (Figure 17 and Table 7). During the course of evolution following the divergence of the archaebacterial and eubacterial lineages the 165 amino acid long eubacterial *E. coli* protein has suffered a large internal deletion (positions 141 - 258, Figure 17), a 3' terminal truncation (position 297 and beyond) and five shorter deletion or insertion events, one of which (positions 15 - 16) removed the unique and conserved tryptophan residue. The eubacterial *E. coli* L10 protein is homologous to the *H. cutirubrum* L10e and *S. solfataricus* L10e proteins by virtue of both sequence similarity and genetic linkage. The significance of the archaebacterial to eubacterial protein sequence comparisons, i.e. z = 10 for *H. cutirubrum* versus *E. coli* and z = 10 for *S. solfataricus* versus *E. coli*, are just within the range regarded as indicative of certain homology. In all three organisms the L10e genes exhibit positional conservation within the conserved L11e, L1e, L10e, L12e tetragenic cluster. The probability of fortuitous clustering of nonhomologous genes performing

Figure 17 Alignment of the L10e Proteins

Upper Panel

The amino acid sequences of the L10e ribosomal proteins from the archaebacteria *Halobacterium cutirubrum* (Hcu L10e) and *Sulfolobus solfataricus* (Sso L10e), predicted from the nucleotide sequences of the respective genes and for the Hcu L10e protein partially confirmed by amino acid sequence, are aligned with the L10e ribosomal proteins from the eubacterium *Escherichia coli* (Eco L10) and the eucaryotes *Homo sapiens* (Hsa L10e) and *Saccharomyces cerevisiae* (Sce L10e). Partial amino acid sequence of the Bst L10e protein indicates that it shares the features of the Eco L10 protein (A.T. Matheson, personal communication). Gaps required for alignment are indicated as dashes (-). The amino acid position scale is below the Sce L10e sequence and the proteins are aligned over 372 amino acid positions. Similarities are indicated between the eubacteria and eucaryota (above the Eco L10 sequence), between the eubacteria and archaebacteria (between the eubacterial and archaebacterial sequences), within the archaebacteria (between the Hcu L10e and Sso L10e sequences), between the archaebacteria and eucaryota (between the archaebacterial and eucaryotic sequences) and within the eucaryota (between the Hsa L10e and Sce L10e sequences) by the following symbols:

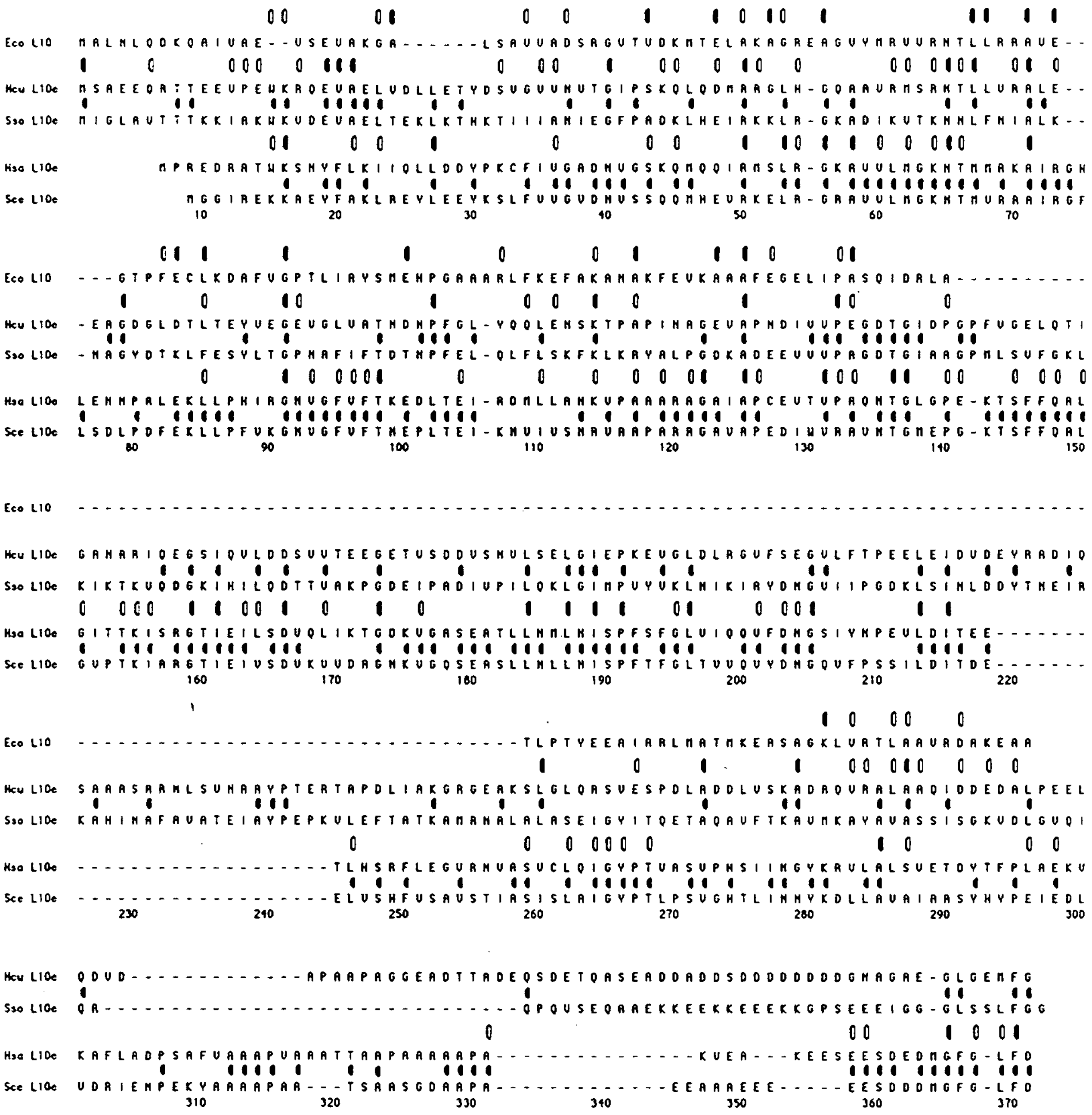
▮ Intrakingdom identities within the archaebacteria and the eucaryota

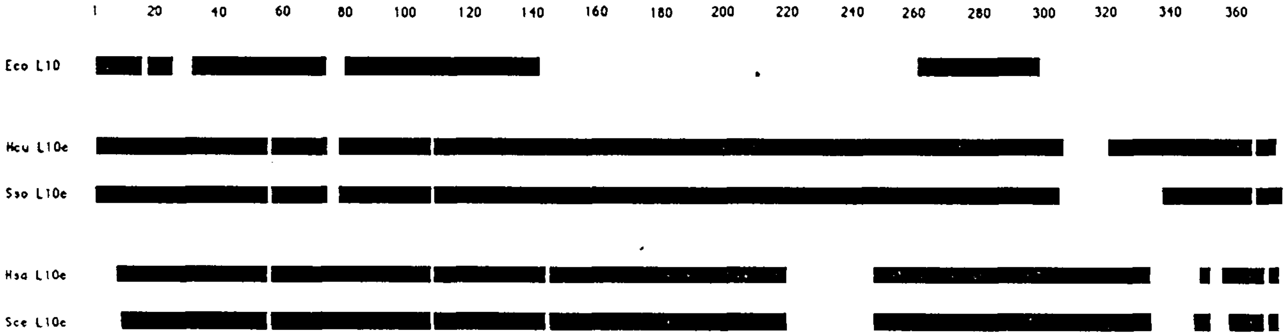
█ Interkingdom identities where all positions within 2 urkingdoms are identical

▯ Interkingdom identities where 2 out of 3 positions between the eubacterial and archaebacterial urkingdoms or 2 out of 3 positions between the eubacterial and eucaryotic urkingdoms or 3 out of 4 positions between the archaebacterial and eucaryotic urkingdoms are identical.

Lower Panel

Line diagrams illustrate the end to end interkingdom alignment of the archaebacterial Hcu L10e and Sso L10e proteins to the complete eubacterial Eco L10 and the eucaryotic Hsa L10e and Sce L10e sequences. Gaps required to maintain maximum identity in the alignment are indicated by white bars. The common scale is in amino acid residue positions along the linear interkingdom alignment.


Eco L10
Hcu L10e
Sso L10e
Hsa L10e
Sce L10e
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370


1
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
Eco L10
Hcu L10e
Sso L10e
Hsa L10e
Sce L10e

analogous functions (i.e. factor binding GTPase domain) in two separate lineages is exceedingly remote.

The very high statistical significance of the archaebacterial versus eucaryotic protein sequences (Hsa L10e and Sce L10e) ranging from z = 25 for Sso L10e versus Hsa L10e to z = 42 for Hcu L10e versus Hsa L10e unequivocally demonstrate that these proteins are homologous. The gene encoding the ancestral eucaryotic L10e protein has an insertion preceding the alanine - proline rich region (positions 305 - 319), two internal deletions (positions 219 - 244 and 332 - 344), and five short deletion - insertion events with respect to its archaebacterial homologue. The deletion at position 332 - 344 follows the alanine - proline rich sequence and extends into the region of high amino acid charge density that is also present in the *H. cutirubrum* and *S. solfataricus* proteins but truncated from the *E. coli* protein.

The sequence similarity and structural features (discussed in section 4.6) of the alignment of eucaryotic L10e proteins to the archaebacterial L10e proteins unequivocally indicate that these proteins are homologous and thus, despite the very low similarity of 15% - 17% identity at the amino acid level and statistical significance z = 2 - 3, the eucaryotic L10e protein must be the homologue of the eubacterial Eco L10. The Hsa L10e and L12e proteins are known to form a complex analogous to the L10 - L12 complex of *E. coli* but because of the low statistical significance of the eubacterial - eucaryotic match Rich and Steitz (1987) were unable to identify P0 as the L10e protein. Thus slowly evolving archaebacterial proteins may serve as a link in identifying proteins present in the progenote that have diverged too greatly between the eubacterial and eucaryotic urkingdoms to be demonstrably homologous by sequence similarity methods.

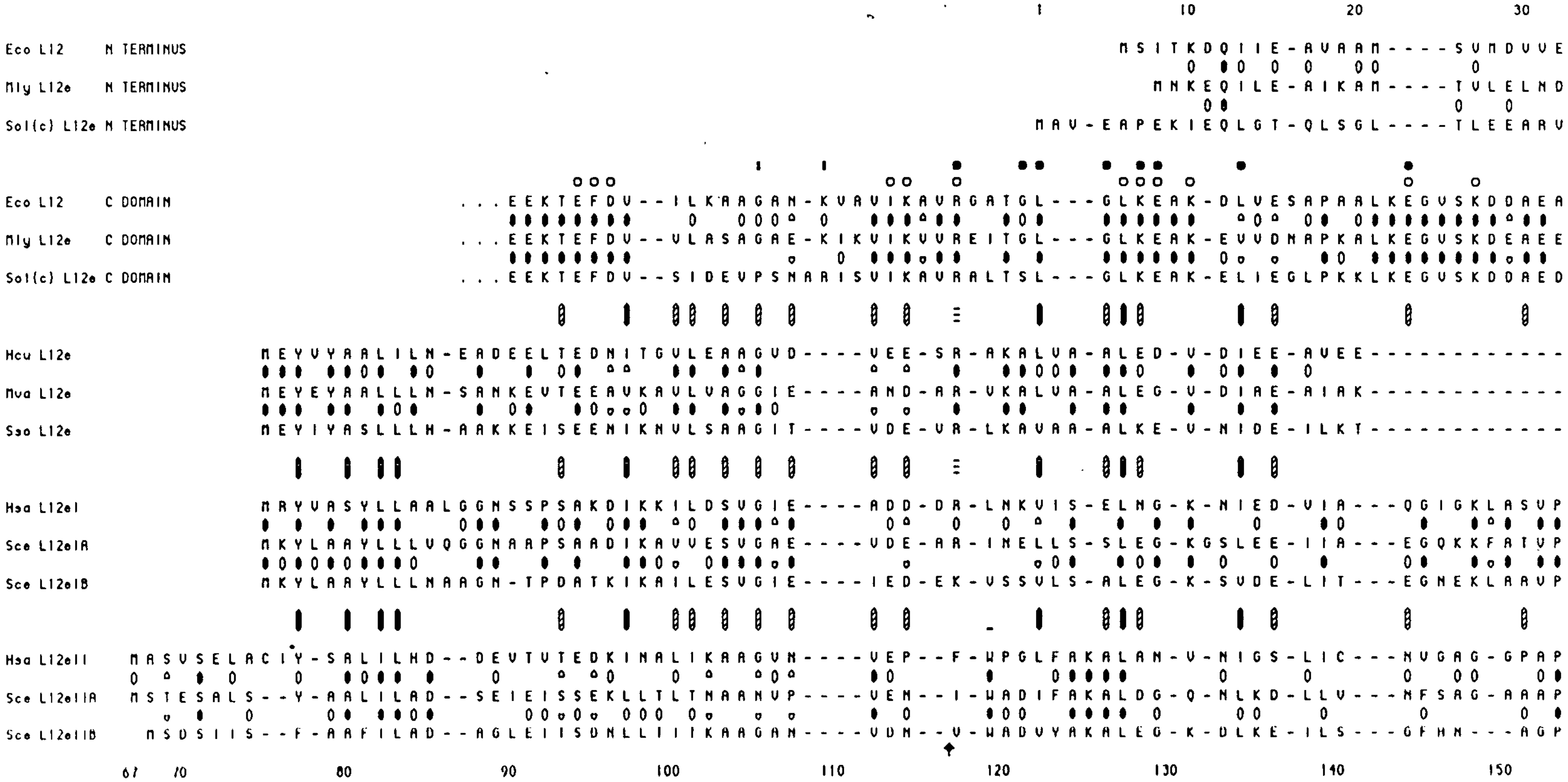
4.5 The L12e proteins

The complete amino acid sequences of 26 L12e proteins are presently available; 9 eubacterial, 5 archaebacterial, 7 eucaryotic type I and 5 eucaryotic type II (Table 3). Figures 20 and 21 illustrate the alignment of three typical but distantly related eubacterial (*E. coli*, *Micrococcus lysodeikticus* and the chloroplast of *Spinacea oleracea*), three archaebacterial (*H. cutirubrum*, *Methanococcus vanielii* and *S. solfataricus*), three eucaryotic type I (*H. sapiens* and 2 from *S. cerevisiae*) and three eucaryotic type II (*H. sapiens* and 2 from *S.cerevisiae*) L12e amino acid sequences. Intrakingdom (and within eucaryota, intratype) alignments and comparisons are readily made; the eubacteria, archaebacteria, eucaryotic type I

Figure 18 Alignment of the L12e Proteins

The amino acid sequences of the L12e ribosomal proteins from three archaebacteria *Halobacterium cutirubrum* (Hcu L12e), *Methanococcus vanielli* (Mva L12e) and *Sulfolobus solfataricus* (Sso L12e), predicted from the nucleotide sequences of the respective genes and confirmed by amino acid sequence, are aligned with the L12e ribosomal proteins from three eubacteria *Escherichia coli* (Eco L10), *Micrococcus lysodeikticus* (Mly L12e) and *Spinacea oleracea* chloroplast (Sol{c} L12e) and two eucaryotes *Homo sapiens* (Hsa L12eI and Hsa L12eII) and *Saccharomyces cerevisiae* (Sce L12eIA, Sce L12eIB, Sce L12eIIA, and Sce L12eIIB). The amino terminal 66 amino acid positions of the eubacterial proteins have no direct counterpart in the archaebacterial or eucaryotic proteins; rather it exhibits a degree of sequence similarity with its own carboxy domain (positions 1 - 49 align to positions 122 - 170). In the eubacterial protein position 66 is fused to position 90 and divides the protein into the amino terminus and the carboxy domain. The intervening positions, beginning at position 67, form the unique amino termini of the archaebacterial and eucaryotic proteins. The region of positions 46 - 66 within the amino terminus of the eubacterial protein and approximate positions 161 - 188 of the archaebacterial and eucaryotic proteins are homologous alanine - proline rich hinge regions. Gaps required for alignment are indicated as dashes (-). The amino acid position scale is indicated at top for the eubacterial amino terminal sequences and at bottom for the eubacterial carboxy domain, archaebacterial and eucaryotic sequences; the proteins are aligned over 228 amino acid positions.

Similarities within and between urkingdoms and features of the sequence and structure of the proteins are indicated by the symbols: Intrakingdom identities where 2 out of 3 residues are identical (0 , ▵▿); Intrakingdom identities where 3 out of 3 residues are identical (▲▼); Interkingdom identities where at least 2 out of 3 positions within each kingdom are inclusively identical or conserved (▯); Interkingdom identities where all positions in urkingdoms are identical or conserved (▮); Residues within the Eco L12 carboxy domain involved in the conserved face (□); Residues within the Eco L12 carboxy domain involved in the dimerization site (■); Residues within the Eco L12 carboxy domain involved in the anion (putative GTP) binding site (▮); Position of the intron within the Sce L12eIIB gene (⬆); Position of the unique tryptophan residue within the eucaryotic L12eII proteins (−). Position of the usually unique arginine residue (rarely substituted by lysine) within the eubacterial carboxy domain, archaebacterial and eucaryotic L12eI proteins (≡). Symbols are not shown for the alignment of the eubacterial amino terminus with the eubacterial carboxy domain and archaebacterial and eucaryotic proteins. Symbols are also not shown for the hinge region (approximate positions 45 - 66 for the eubacterial amino terminus and 160 - 190 for the archaebacterial - eucaryotic proteins) due to the relaxed constraint on amino acid sequence conservation in this region. Conserved amino acid substitutions are within the groups: D - E - Q - N - R - K - H, L - I - V - M - F, Y - F, A - G, S - T, A - S.


1 10 20 30
Eco L12 N TERMINUS MSITKDQIIE-AVAAM----SVMDVVE
Mly L12e N TERMINUS MNKEQILE-AIKAM----TVLELND
Sol(c) L12e N TERMINUS MAV-EAPEKIEQLGT-QLSGL----TLEEARV
Eco L12 C DOMAIN ...EEKTEFDV--ILKAAGAN-KVAVIKAVRGATGL---GLKEAK-DLVESAPAALKEGVSKDDAEA
Mly L12e C DOMAIN ...EEKTEFDV--VLASAGAE-KIKVIKVVREITGL---GLKEAK-EVVDNAPKALKEGVSKDEAEE
Sol(c) L12e C DOMAIN ...EEKTEFDV--SIDEVPSNARISVIKAVRALTSL---GLKEAK-ELIEGLPKKLKEGVSKDDAED
Hcu L12e MEYVYAALILN-EADEELTEDNITGVLEAAGVD----VEE-SR-AKALVA-ALED-V-DIEE-AVEE------------
Mva L12e MEYEYAALLLN-SANKEVTEEAVKAVLVAGGIE----AND-AR-VKALVA-ALEG-V-DIAE-AIAK------------
Sso L12e MEYIYASLLLN-AAKKEISEENIKNVLSAAGIT----VDE-VA-LKAVAA-ALKE-V-NIDE-ILKT------------
Hsa L12eI MRYVASYLLAALGGNSSPSAKDIKKILDSVGIE----ADD-DR-LNKVIS-ELNG-K-NIED-VIA---QGIGKLASVP
Sce L12eIA MKYLAAYLLLVQGGNAAPSAADIKAVVESVGAE----VDE-AR-INELLS-SLEG-KGSLEE-IIA---EGQKKFATVP
Sce L12eIB MKYLAAYLLLNAAGN-TPDATKIKAILESVGIE----IED-EK-VSSVLS-ALEG-K-SVDE-LIT---EGNEKLAAVP
Hsa L12eII MASVSELACIY-SALILHD--DEVTVTEDKINALIKAAGVN----VEP--F-WPGLFAKALAN-V-NIGS-LIC---NVGAG-GPAP
Sce L12eIIA MSTESALS--Y-AALILAD--SEIEISSEKLLTLTNAANVP----VEN--I-WADIFAKALDG-Q-NLKD-LLV---NFSAG-AAAP
Sce L12eIIB MSDSIIS--F-AAFILAD--AGLEIISDNLLIITKAAGAN----VDN--V-WADVYAKALEG-K-DLKE-ILS---GFHN---AGP
67 70 80 90 100 110 120 130 140 150

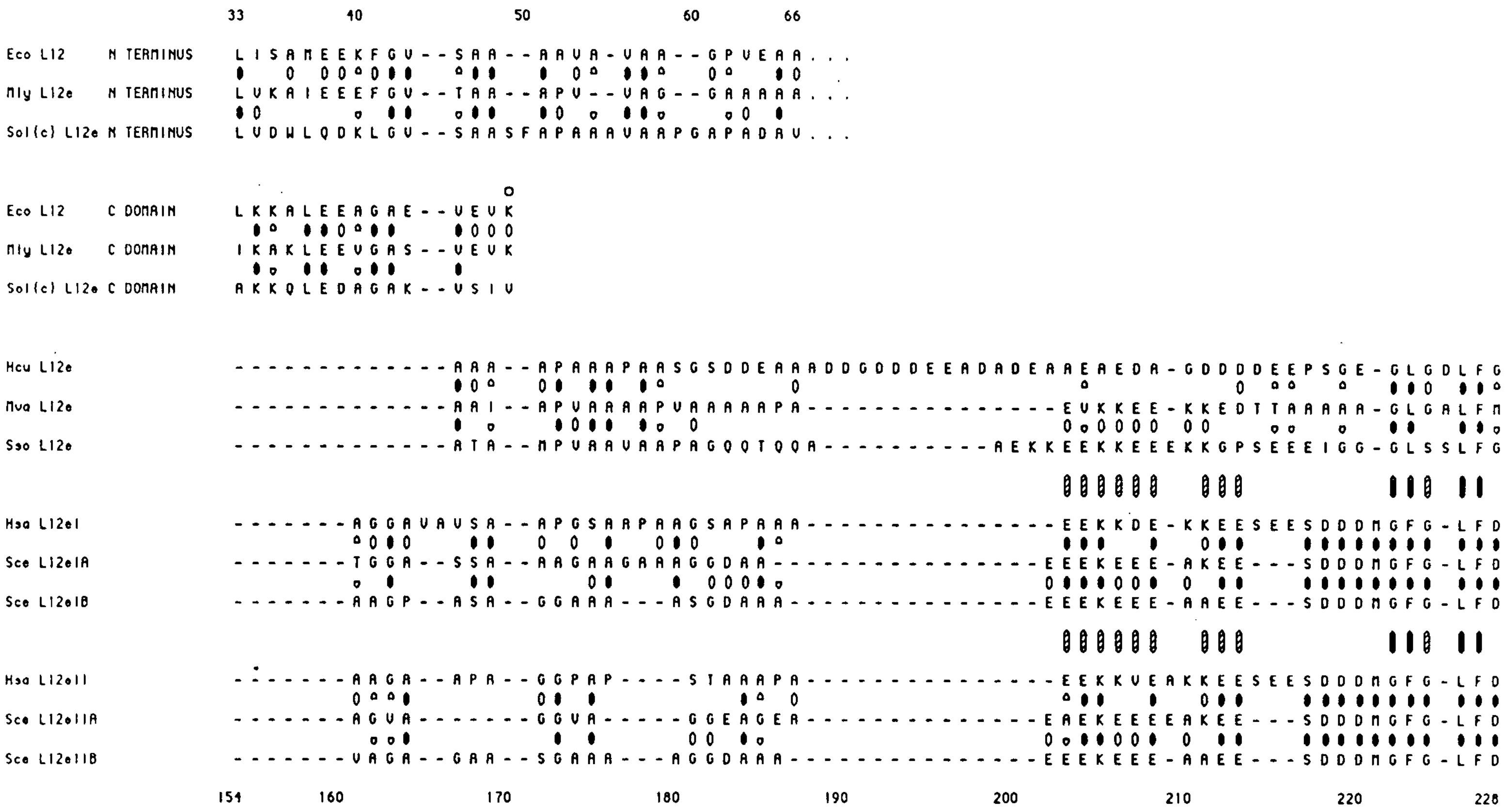
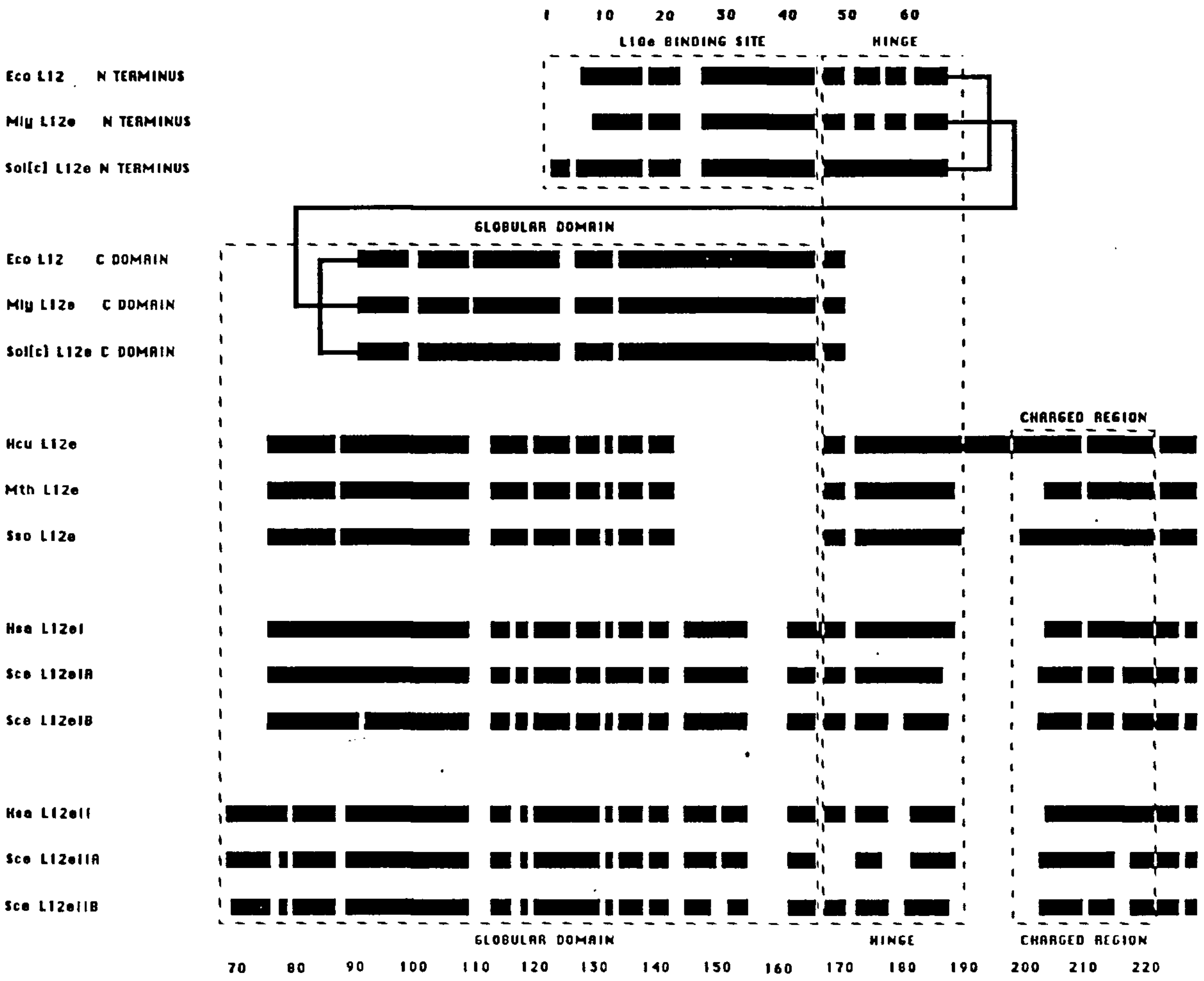

33 40 50 60 66
Eco L12 N TERMINUS L I S A M E E K F G V - - S A A - - A A V A - V A A - - G P V E A A . . .
Mly L12e N TERMINUS L V K A I E E E F G V - - T A A - - A P V - - V A G - - G A A A A A . . .
Sol(c) L12e N TERMINUS L V D W L Q D K L G V - - S A A S F A P A A A V A A P G A P A D A V . . .
Eco L12 C DOMAIN L K K A L E E A G A E - - V E V K
Mly L12e C DOMAIN I K A K L E E V G A S - - V E V K
Sol(c) L12e C DOMAIN A K K Q L E D A G A K - - V S I V
Hcu L12e - - - - - - - - - - - - - - A A A - - A P A A A P A A S G S D D E A A A D D G D D D E E A D A D E A A E A E D A - G D D D D E E P S G E - G L G D L F G
Mva L12e - - - - - - - - - - - - - - A A I - - A P V A A A A P V A A A A P A - - - - - - - - - - - - - - - - - - E V K K E E - K K E D T T A A A A A - G L G A L F M
Sso L12e - - - - - - - - - - - - - - A T A - - M P V A A V A A P A G Q Q T Q Q A - - - - - - - - - - - - - A E K K E E K K E E E K K G P S E E E I G G - G L S S L F G
Hsa L12eI - - - - - - - - A G G A V A V S A - - A P G S A A P A A G S A P A A A - - - - - - - - - - - - - - - - - E E K K D E - K K E E S E E S D D D M G F G - L F D
Sce L12eIA - - - - - - - - T G G A - - S S A - - A A G A A G A A A G G D A A - - - - - - - - - - - - - - - - - E E E K E E E - A K E E - - - S D D D M G F G - L F D
Sce L12eIB - - - - - - - - A A G P - - A S A - - G G A A A - - - A S G D A A A - - - - - - - - - - - - - - - - - E E E K E E E - A A E E - - - S D D D M G F G - L F D
Hsa L12eII - - - - - - - - A A G A - - A P A - - G G P A P - - - - S T A A A P A - - - - - - - - - - - - - - - - - E E K K V E A K K E E S E E S D D D M G F G - L F D
Sce L12eIIA - - - - - - - - A G V A - - - - - - - - G G V A - - - - - - G G E A G E A - - - - - - - - - - - - - - - - - E A E K E E E E A K E E - - - S D D D M G F G - L F D
Sce L12eIIB - - - - - - - - V A G A - - G A A - - S G A A A - - - A G G D A A A - - - - - - - - - - - - - - - - - E E E K E E E - A A E E - - - S D D D M G F G - L F D
154 160 170 180 190 200 210 220 228

Figure 19 Line Diagram of the L12e Protein Alignment

Line diagrams illustrate the end to end interkingdom alignment of the L12e ribosomal proteins from three archaebacteria *Halobacterium cutirubrum* (Hcu L12e), *Methanococcus vanielli* (Mva L12e) and *Sulfolobus solfataricus* (Sso L12e), three eubacteria *Escherichia coli* (Eco L10), *Micrococcus lysodeikticus* (Mly L12e) and *Spinacea oleracea* chloroplast (Sol{c} L12e) and two eucaryotes *Homo sapiens* (Hsa L12eI and Hsa L12eII) and *Saccharomyces cerevisiae* (Sce L12eIA, Sce L12eIB, Sce L12eIIA, and Sce L12eIIB). Gaps required to maintain maximum identity in the alignment are indicated by white bars. The eubacterial protein is shown divided into the amino terminus and carboxy domain. The globular domain, hinge region, L10 binding site of Eco L12 and the unique highly charged carboxy terminal region of the archaebacterial and eucaryotic proteins are highlighted by dashed boxes. The amino acid position scale is indicated at top for the eubacterial amino terminal sequences and at bottom for the eubacterial carboxy domain, archaebacterial and eucaryotic sequences.



and eucaryotic type II proteins averaging 52%, 42%, 57% and 53% amino acid sequence identity respectively.

The archaebacterial and the eucaryotic L12eI proteins can be linearly aligned to each other end to end with the initiation methionine at position 75. The unique conserved arginine residue of the archaebacterial L12e proteins (position 117) is conserved in the Hsa L12eI and Sce L12eIA proteins and conservatively substituted with a charged basic lysine residue in the Sce L12eIB protein. The archaebacterial - eucaryotic type I interkingdom similarity averages 30% amino acid identity over the amino terminal region (positions 75 - 140; reasons for the exclusion of the carboxy terminal region for calculation of interkingdom sequence similarity are discussed in section 4.6). The eucaryotic L12eII proteins align with their archaebacterial and eucaryotic type I homologues but have an extended amino terminus, lack the conserved arginine residue (position 117) and contain, immediately adjacent, a unique tryptophan residue (position 119). Archaebacterial - eucaryotic type II interkingdom similarity averages 26% amino acid identity over the amino terminal region (positions 75 - 140). Thus the archaebacterial L12e protein share greater structural, compositional and slightly higher sequence similarity with the eucaryotic L12eI protein than with the eucaryotic L12eII proteins. The eucaryotic L12eI and L12eII proteins share only 20% amino acid identity over the amino terminal region (positions 75 - 140). This suggests that either the type I and type II proteins diverged from each other prior to the divergence of eucaryota and archaebacteria and the archaebacterial and eubacterial lineage either lost (possibly during the development of the L11e - L1e - L10e - L12e gene cluster) or never contained the type II gene, or the rate of divergence of the eucaryotic L12e proteins is significantly greater than that of the archaebacterial L12e proteins and the sequence similarity difference between the archaebacterial versus eucaryotic type I proteins (30% identity) and archaebacterial versus eucaryotic type II proteins (26% identity) is fortuitous or insignificant.

Although the eubacterial L12e protein cannot be linearly aligned with its archaebacterial or eucaryotic counterparts two homologous domains are common to the proteins from all three urkingdoms. The first domain is located near the amino terminus of the archaebacterial and eucaryotic proteins and in the middle of the eubacterial protein (position 90 - 140); interkingdom similarities for this domain average 28% amino acid sequence identity and range from a minimum of 17 percent identity between the Sce L12eIIB and Eco L12 proteins to a maximum of 36 percent between the Eco L12 and Hcu L12e proteins. The second common region is the alanine - proline rich sequence located between positions 46 - 66 in the eubacterial protein and between positions 167 - 187 in the archaebacterial and eucaryotic proteins (Figure 18). The length and sequence of these alanine - proline rich sequences is highly variable even within urkingdoms, with substitutions occurring between alanine, proline, serine, threonine and glycine.

The amino terminus of the eubacterial L12e protein (positions 1 - 43, Figure 18), exhibits greater sequence identity (12 of 31 amino acid identities for Eco L12) to its own carboxy terminus (position 122 - 164) than to any sequence within the archaebacterial or eucaryotic proteins (Figure 19). The second half of this intramolecular complementarity in the eubacterial L12e protein appears to align to regions interrupted by deletion within the archaebacterial (positions 142 - 164) and eucaryotic (positions 153 - 160) L12e proteins (Figure 18).

The eubacterial L12e protein has no homologue to the carboxy terminal region of the archaebacterial and eucaryotic L12e proteins. This region consists of a very highly charged region (approximate positions 200 - 220) followed by a mostly hydrophobic extreme carboxy terminus and is very highly conserved both within and between urkingdoms. The Hcu L12e protein has the least sequence similarity in this region due to a complete absence of basic (i.e. lysine) residues and the predominance of aspartate over glutamate residues; this likely occurred during adaptation to a high salt environment.

Biophysical studies on the Eco L12 protein indicate that the amino terminal domain spontaneously dimerizes and binds to the Eco L10 protein (Koteliansky *et al.*, 1978). The carboxy terminal domain forms a compact structure that crystallizes as a dimer, contains an anion binding site, and may interact through a conserved face with extrinsic translation factors during the protein synthesis cycle (Leijonmarck and Liljas, 1987; Figure 18). The two domains are separated by an alanine - proline rich region believed to be unstructured and to function as a flexible hinge between domains of the L12e proteins, accounting for the observed high mobility of the carboxy terminal domain of Eco L12 (Tritton, 1978; Leijonmarck *et al.*, 1981; Cowgill *et al.*, 1984).

The alignment of the L12e proteins illustrated in Figure 18 implies that the amino terminal end of the archaebacterial - eucaryotic proteins contains the factor binding domain (extending over the region 94 - 148), the dimerization site (positions 117 - 144) and the putative anion binding site (positions 105 and 109). The alignment suggests that the ancestral globular domain comprised 75 to 80 amino acids wherefrom the eubacteria have lost approximately 15 amino acids, including the unique two conserved tyrosine residues (positions 77 and 79 in archaebacteria, 77 and 81 in eucaryotic type I and position 77 in eucaryotic type II), on the amino side of the region of the conserved face. The archaebacteria and eucaryota have lost approximately 25 and 10 amino acids respectively from the carboxy side of the region of the conserved face. However the dimerization, anion and factor binding domains have been retained in all three urkingdoms. The alanine - proline rich hinge region is evident on the amino and carboxy sides of the globular domain of the eubacterial and archaebacterial - eucaryotic proteins respectively (Figure 19).

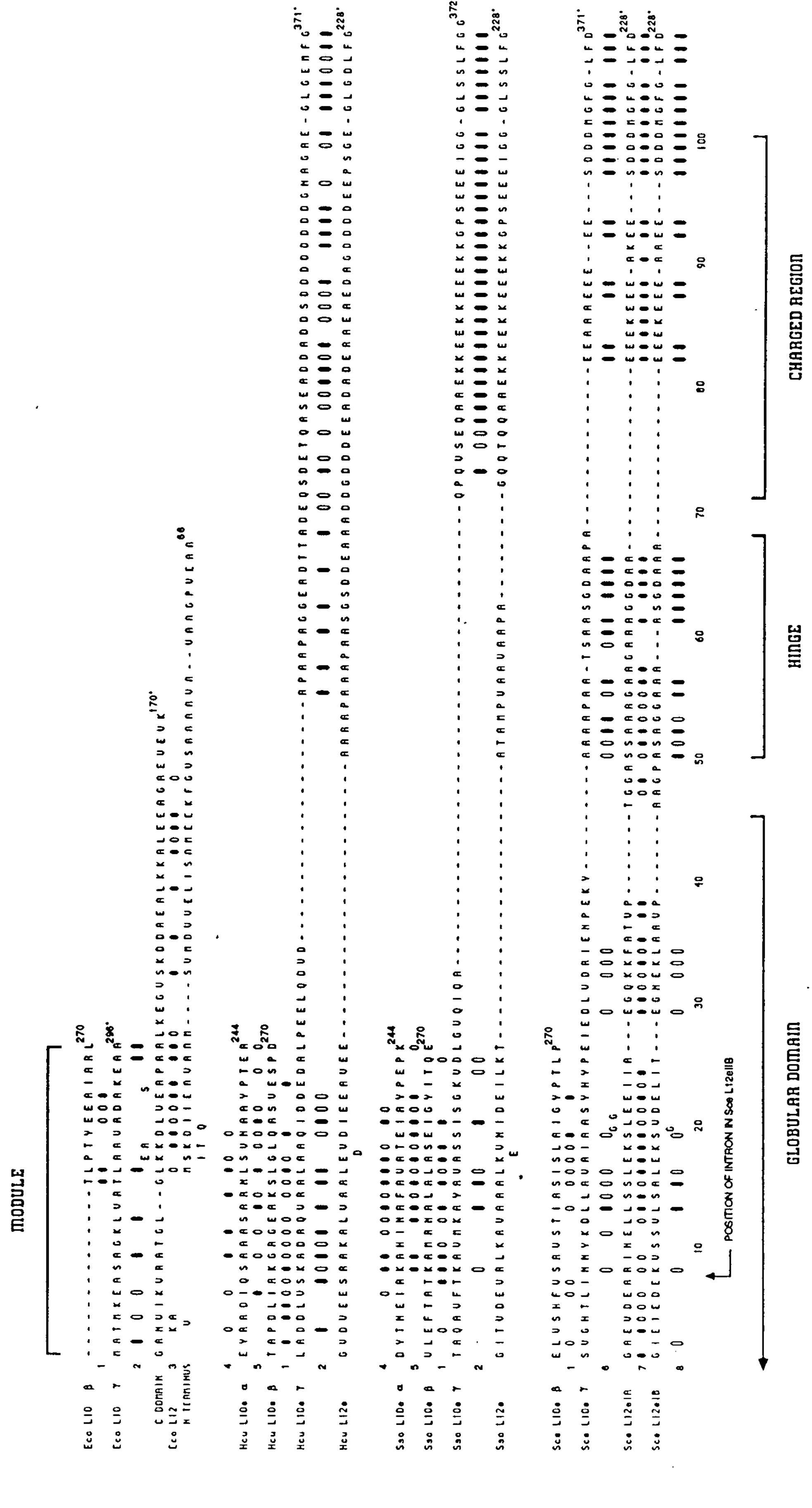
4.6 Intra and Interprotein Relationships

The archaebacterial and eucaryotic L10e and L12e proteins contain a region of high amino acid charge density near their carboxy terminal ends (Rich and Steitz, 1987; Ramirez *et al.*, 1989; Shimmin *et al.*, 1989b; Newton *et al.*, 1990). The L10e and L12e proteins of *H. cutirubrum* and *S. solfataricus* exhibit a high degree of sequence similarity at their carboxy terminal ends (Figure 20). For the two *S. solfataricus* proteins, the 31 carboxy terminal residues which contain the region of high charge density were found to be identical (and, remarkably, the nucleotide sequence was also identical) except for an extra glycine at the end of the L10e sequence. The degree of amino acid sequence identity was less pronounced although highly significant for the *H. cutirubrum* proteins (45% over 29 residues) and the *S. cerevisiae* proteins (75% over 23 residues).

Extension of the archaebacterial L10e - L12e alignments into the central regions of the proteins resulted in the discovery of a modular sequence of length 26 amino acids. The module, tandemly reiterated three times in the L10e proteins, was present in single copy in the L12e proteins. The three L10e module copies were designated α, β and γ. In the Hcu L10e protein short sequences flanking the triple modules are strikingly similar; positions 210 - 218, Figure 17, (PEELEIDVD) compared with positions 297 - 304 (PEELQDVD) have, excluding a one amino acid gap, 7 out of 8 amino acid residues and 21 out of 24 nucleotides identical. It remains unknown whether this nearly perfect direct repeat in the DNA was involved in the module duplication process. These modular sequence domains are separated from the high charge density domain by the alanine - proline rich hinges in the Hcu L10e, HcuL12e and Sso L12e proteins but not in the alanine - proline rich region deficient SsoL10e protein.

Figure 20_ Intra and Interprotein Relationships between the L10e and L12e proteins

Four intraspecies comparisons of L10e and L12e amino acid sequences are presented from top to bottom: *Escherichia coli*, *Halobacterium cutirubrum*, *Sulfolobus solfataricus* and *Saccharomyces cerevisiae.* The regions of the modules, hinge and charged region are indicated. The L10e α, β and γ sequences are the three 26 residue long module repeats. For Eco L12 where the protein has undergone major rearrangements and alterations during eubacterial evolution, a complete and a partial copy of the module appear to be present in the carboxy domain and amino terminus respectively. In Sce L10e one copy of the module is not present. Amino acid comparisons of identities (▮,▮) and conservative substitutions (▯,▯) are as follows: line 1, L10e β with γ; line 2, the γ module and carboxy terminus of L10e with the module and carboxy terminus of L12e; line 3, the carboxy domain and amino terminus of Eco L12; line 4, L10e α with γ; line 5, L10e α with β; line 6, Sce L10e to Sce L12eIA; line 7, Sce L12eIA to Sce L12eIB; line 8, Sce L10e to Sce L12eIB. The relative position of the intron within the Sce L12eIIB gene is indicated by the arrow. The numbers at the end of the sequences designates the position number of the terminal amino acid of the modules and proteins (from Figures 17 and 18). Residues representing the carboxy termini of the respective proteins are identified (*). Conservative substitutions are defined as being within the groups: D - E - Q - N - R - K - H, L - I - V - M - F, Y - F, A - G, S - T, A - S.


MODULE
GLOBULAR DOMAIN
HINGE
CHARGED REGION
POSITION OF INTRON IN Sce L12eIIB
Eco L10 β
Eco L10 γ
C DOMAIN Eco L12
N TERMINUS
Hcu L10e α
Hcu L10e β
Hcu L10e γ
Hcu L12e
Sso L10e α
Sso L10e β
Sso L10e γ
Sso L12e
Sce L10e β
Sce L10e γ
Sce L12eIA
Sce L12eIB
10
20
30
40
50
60
70
80
90
100
270
296'
170'
66
244
270
371'
228'
372'
228'

Rich and Steitz (1987) noted sequence similarity at the carboxy terminal ends of the eucaryotic Hsa L10e and L12e proteins, including the alanine - proline rich and high charge density domains. Extension of the eucaryotic alignment resulted in the discovery of a single copy of the module domain in the eucaryotic L12e protein sequences and two tandemly reiterated copies in the eucaryotic L10e protein sequences (Shimmin *et al.*, 1989b). One of the L10e protein's modules was not generated in the ancestral eucaryotic gene or was removed by a deletion (Figure 17). Which of the three modules is actually missing cannot be ascertained, although the alignment given in Figure 17 (i.e. missing the α module, positions 219 - 244) yields the highest degree of sequence similarity.

Deletions / insertions within the carboxy terminal region of the archaebacterial - eucaryotic L12e protein appear in similar, and sometimes identical, positions in the L10e proteins, suggesting that these sequences may still be functional homologues and that selection preserves the similarity of the sequences (Figure 20). The only anomaly in this pattern is the lack of the alanine - proline rich hinge feature in the Sso L10e protein. The absolute identity of the *S. solfataricus* L10e and L12e proteins from positions 74 - 105 (Figure 20) at both the amino acid and nucleic acid level suggests a very recent recombinational restoration event and if this event removed the hinge region leaving only one proline at position 69 (Figure 20) then other thermoacidophilic archaebacteria should retain the extended hinge region. The virtually identical alanine - proline rich hinge and high charge density domains of the four Sce L12e proteins, and to a lesser extent the Sce L10e protein, indicates that restoration or conservation of this region is not confined to the archaebacteria (Figure 18 and Figure 20).

Neither of the eubacterial L10e and L12e proteins contain the carboxy terminal region of high amino acid charge density and thus they do not exhibit sequence similarity at their carboxy terminal ends. However, regions corresponding to potential single intact and partial copies of the residual modular sequence were located when the Eco L10 and Eco L12 sequences were aligned to the corresponding archaebacterial and eucaryotic proteins (Figure 20). The Eco L10 complete module matches well with the Hcu L10e γ module (9 identical amino acid residues) and because of this is aligned at the γ position. The 12 amino acids preceding the intact Eco L10 module may represent a second partial module; equal sequence similarity of these 12 residues with the other L10e proteins is evident whether they are aligned at positions 141 - 152 or as a partial module at positions 259 - 270 (Figure 17).

When treated as a group the modules were highly significant (z = 17) although the statistical significance of matches between individual modules is low or nil. The evidence for the existence of the modules is strengthened by the presence of a conserved gap of precisely 26 residues eliminating the α module within both of the eucaryotic Hsa L10e and Sce L10e proteins and the termination of the eubacterial Eco L10 protein precisely at the end of its intact module.

Thus the existence of a statistically significant module of 26 amino acids in the L10e proteins, repeated thrice in archaebacteria, twice in eucaryotes, and an intact plus a possible partial copy in eubacteria has been demonstrated (Figure 19 and Figure 20). Sequence similarity indicates that a single copy of the module also exists in the L12e protein. Several features of the module region are noteworthy. First, the central section (positions 8 - 21, Figure 20) of the modules within L10e is the most highly conserved (31% amino acid identities and 29% conservative amino acid substitutions); the flanking regions (positions 1 - 7 and 22 - 26, Figure 20) have similarity that is close to random for sequences of this amino acid composition (10% identical and 15% conservative amino acid substitutions). Second, the conserved arginine of the eubacterial, archaebacterial and eucaryotic L12el proteins (position 117, Figure 18) generally aligns with positively charged (five lysine) or hydrophilic (one asparagine, three serine) residues (position 8, Figure 20). Third, the hydrophobic residue in position 16 appears to be the most highly conserved residue (eleven leucine, three valine, one methionine, one isoleucine). Attribution of the 12 amino acid residues preceding the Eco L10 γ module to a partial β module preserves this leucine residue. Fourth, alanine is also highly conserved in positions 9, 13, 15, 17 (Figure 20); this being best exhibited by the Sso L10e protein where of the sixteen alanines present within the three modules, twelve align perfectly at these positions.

The putative copy of the module present in Eco L12 (positions 105 - 141, Figure 18) contains the majority of the L12 dimerization site of the globular domain (primarily positions 117 - 134). Alignment of the L12e proteins to the L10e proteins (Figure 20) revealed that the dimerization site in the carboxy terminal domain of Eco L12 (positions 8 - 20) was aligned with the region of highest conservation in the L10e modules (approximate positions 8 - 21). This would suggest that these modules may be reiterative L12 dimerization sites. Furthermore, the amino terminal end of the eubacterial L12e protein appears to be a duplication in part of this same dimerization site and this may explain the tendency of the Eco L12 amino terminal domain to spontaneously dimerize (Figure 17). The carboxy terminal region of the Eco L10 protein is thought to be responsible for binding the L12 dimers; this may be facilitated by the presence of the protein's terminal intact module (Liljas, 1982).

The presence of these putative dimerization modules in all L10e proteins suggests a mechanism for interaction with the L12e proteins. The *E. coli* L12 globular domain undergoes a conformational change upon interaction with extrinsic translation factors and if this conformational change exposes the L12 dimerization site then the L12 protein could conceivably fold about the hinge and bring the L12 dimerization site into interaction with the dimerization site of the L10 module (Gudkov and Gongadze, 1984). Thus the bound extrinsic translation factor would be brought to the ribosome surface in a specific orientation. It is possible that the multiple modules in the L10e proteins also serve as multiple interaction sites for the L12e protein and if other ribosomal proteins contain the module sequence then the L12e protein (with bound translation factor) could be targeted to various sites on the ribosome surface. This mechanism of action would be possible for all types of L10e - L12e protein complexes.

4.7 Summary

The structural features of the eubacterial, archaebacterial and eucaryotic L11e, L1e, L10e and L12e ribosomal proteins are illustrated in Figure 21. Complete amino acid sequences of the eucaryotic Sce L11e and Sce L1e proteins have been determined but not published; only a short amino terminal region of the Sce L11e protein is available.

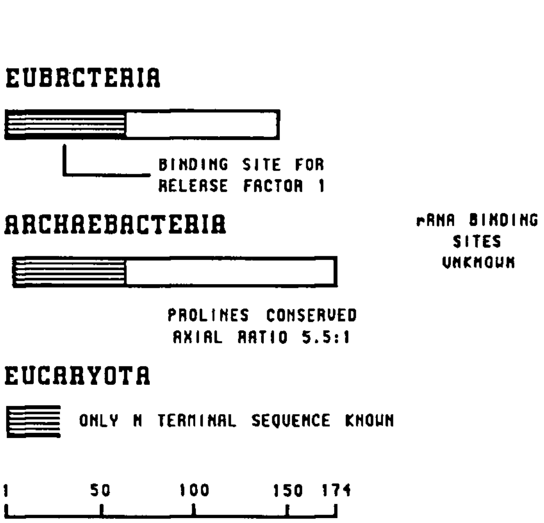
The L11e proteins are colinear, rich in conserved proline residues (which may contribute to the highly elongated structure of the protein) and exhibit two regions of high amino acid sequence conservation in the amino and carboxy terminal domains. The amino terminal 64 amino acid residues of the Eco L11 protein is known to interact with translation release factor 1 and this region is the best conserved region of the protein; a second conserved region exists in the carboxy terminal domain. The binding site within the L11e protein for rRNA remains indeterminate. A structural feature, i.e. methylation patterns, and a functional feature, i.e. synthesis of ppGpp during the stringent response, of the Eco L11 protein appear not to be conserved in the archaebacteria.

The L1e proteins of eubacteria and archaebacteria are colinear and preserve two regions with very

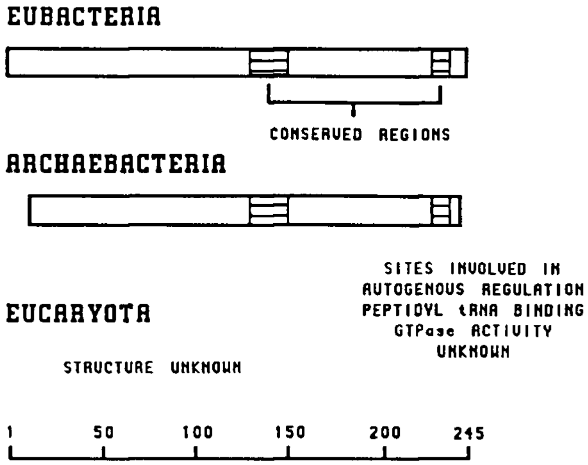
Figure 21 Summary of the Structure and Function of the L11e, L1e, L10e and L12e Ribosomal Proteins

The structural features of the eubacterial, archaebacterial and eucaryotic L11e proteins are illustrated at the upper left with an amino acid scale corresponding to the sequence scale utilized in the L11e alignment of Figure 15. The conserved amino terminal region responsible for the binding of release factor 1 during translation termination is shaded. The L1e proteins of eubacteria and archaebacteria are illustrated at upper right with an amino acid scale corresponding to the sequence scale utilized in the L1e alignment of Figure 16 and the highly conserved regions indicated. The structural features of the L10e and L12e proteins from the eubacteria, archaebacteria and the eucaryota are illustrated at lower left and lower right respectively with amino acid scales corresponding to the sequence scales utilized in the L10e and L12e alignments of Figures 17 and 18 respectively. The upper amino acid scale is unique to the eubacterial L12e protein and has a fusion of position 66 with position 90. The archaebacterial and eucaryotic L12e proteins correspond with the lower L12e scale. The archaebacterial L12e protein is composed of a globular domain containing a single copy of the module, a hinge and the charged carboxy terminus. The archaebacterial L10e contains a fusion of three quarters of a copy of the L12e protein and a triplication of the modular sequence present in the L12e part of the fusion. The eucaryotic proteins are very similar to their archaebacterial counterparts: there exist two types of L12e protein (i.e. L12e type I which is similar to the archaebacterial L12e and the L12e type II which differs in the globular domain) and the L10e protein has only two modules. The eubacterial proteins have undergone substantial alterations. The L10e protein has a large internal deletion, only one complete and one partial module, and the carboxy terminal sequences containing the hinge and highly charged regions are truncated. The L12e protein retains the globular domain with the internal module and dimerization site but the hinge has been relocated to the amino terminal side of the globular domain. The amino terminal end, responsible for dimerization of the L12e proteins and binding to the L10e protein is partially derived from a module.

L11e

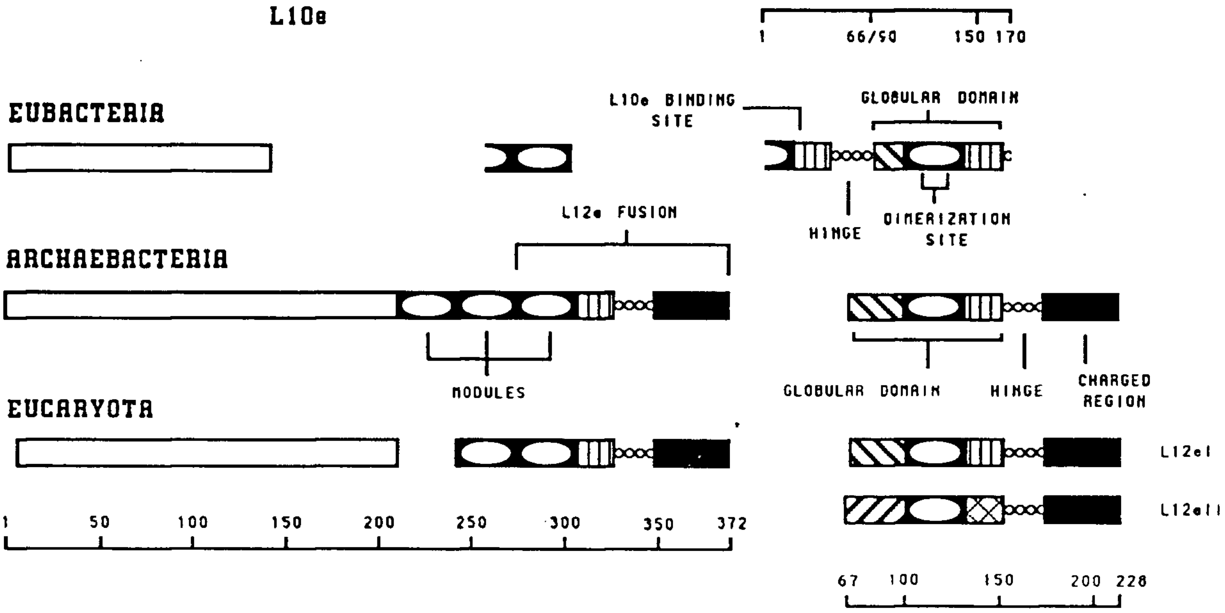

EUBACTERIA
BINDING SITE FOR RELEASE FACTOR 1
ARCHAEBACTERIA
rRNA BINDING SITES UNKNOWN
PROLINES CONSERVED
AXIAL RATIO 5.5:1
EUCARYOTA
ONLY N TERMINAL SEQUENCE KNOWN
1
50
100
150
174

L1e



L10e

L12e



high amino acid sequence conservation in the center and the carboxy terminal regions of the protein. The correspondence between the highly conserved regions and the functional aspects of the L1e protein, i.e. rRNA binding, autogenous control, interaction with peptidyl - tRNA at the P site and indirect interaction with the GTPase domain, is unknown.

The L10e proteins from all three urkingdoms are colinear, although the eubacterial protein is half the size of its archaebacterial and eucaryotic homologues. The archaebacterial and eucaryotic proteins contain approximately three fourths of a copy of the archaebacterial - eucaryotic type L12e protein (including the module, the flexible hinge region and the high charge density region) fused to their carboxy termini. The module which contains a putative dimerization site has been duplicated in the eucaryotic and triplicated in the archaebacterial proteins. The eubacterial protein has suffered a large internal deletion, retains one intact and one partial copy of the module and the flexible hinge and high charge density domains are absent. The regions of the L10e protein responsible for binding to rRNA is unknown but likely to be in the amino terminal domain common to all the proteins.

The L12e proteins are not colinear, the eubacterial protein having suffered substantial alterations. The archaebacterial L12e protein is composed of a globular domain containing a copy of the 26 residue long module, a flexible alanine - proline rich hinge and the charged carboxy terminus. The eucaryotic proteins are very similar to their archaebacterial homologues: there exist two types of L12e protein; an L12e type I which is similar to the archaebacterial L12e and a type II which differs in the globular domain. The eubacterial L12e protein has a globular domain responsible for translation factor interactions and containing sites utilized for dimerization and anion binding. In contrast to the archaebacterial and eucaryotic proteins the hinge has been relocated to the amino terminal side of the globular domain and the charged domain is absent. The amino terminal end, responsible for dimerization of the L12e proteins and binding to the L10e protein, is partially derived from a module;

Part 5: Evolution of the GTPase Domain

5.1 Evolution of the GTPase Domain

Although all extant organisms display clustering (and cotranscription) of the 16S / 18S and 23S / 28S rRNA genes, ancient clustering of translated genes is evidenced only by the archaebacteria and eubacteria, where gene clusters corresponding to the 'RIF', 'STR', 'S10' and 'SPC' ribosomal protein and the β / β' RNA polymerase subunit operons of *E. coli* have now been identified (Auer *et al.*, 1989; Leffers *et al.*, 1989; Puhler *et al.*, 1989; Shimmin and Dennis, 1989; Ramirez *et al.*, 1990a). The L11e, L1e, L10e and L12e ribosomal proteins are conserved in all three extant urkingdoms and thus must have evolved before the existence of the last common ancestral state (the progenote) and more probably during the initial development of the translation apparatus.

The functional, structural and sequence similarity data evidenced by the L11e, L1e, L10e and L12e proteins presented in this thesis suggest that the archaebacteria are a coherent phylogenetic group more similar to each other than to either eubacteria or eucaryotes. Amino acid sequence identity in the protein alignments indicate that there is always higher identity within the deep divergence of the thermoacidophilic (represented by *S. solfataricus*) and methanogenic / halophilic (represented by *H. cutirubrum*) branches of the archaebacteria than between the archaebacteria and either eucaryota or eubacteria. Furthermore, as illustrated by the line diagrams of Figures 15, 16, 17 and 19, the interruptions in the two archaebacterial proteins (required to maintain amino acid alignment with the eubacterial and eucaryotic proteins) are almost always at identical positions. In many cases the positions of these archaebacterial interruptions are unique to archaebacteria and therefore probably represent deletion or insertion events that took place after divergence of the archaebacteria from eubacteria and eucaryota. A number of the positions of archaebacterial interruptions are shared with eucaryota and few if any are shared with eubacteria. This is exemplified by the highly similar structure of the L10e and L12e proteins of eucaryota and archaebacteria as compared to eubacteria (Figure 21). This may mean either that the archaebacteria and eucaryota are more closely related to each other than to eubacteria or that a plethora of rearrangements in the genes encoding these proteins occurred in the very early eubacterial lineage following its divergence from the archaebacteria and eucaryota.

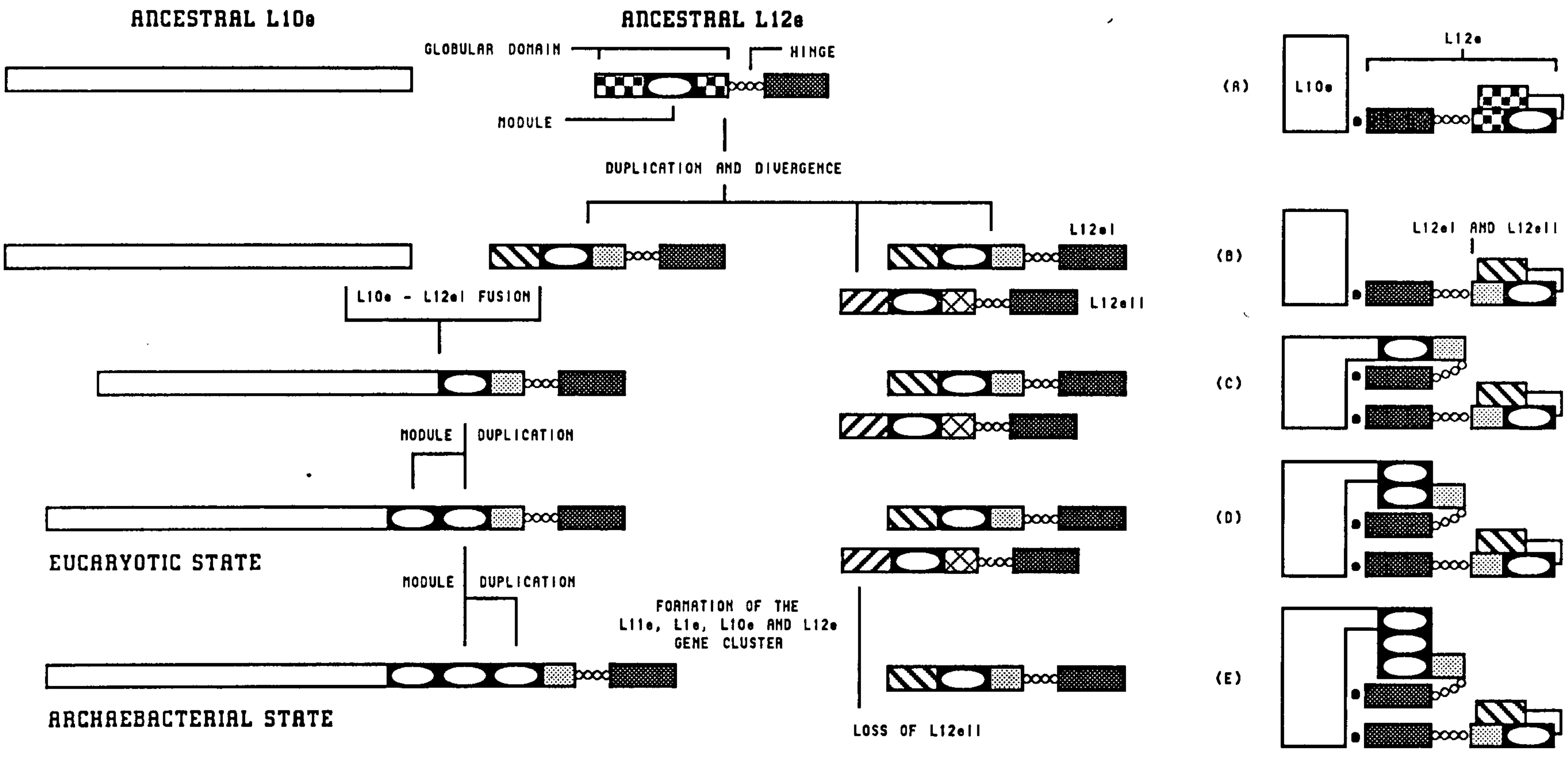
5.2 Model of the Evolution of the L10e and L12e Genes and Proteins

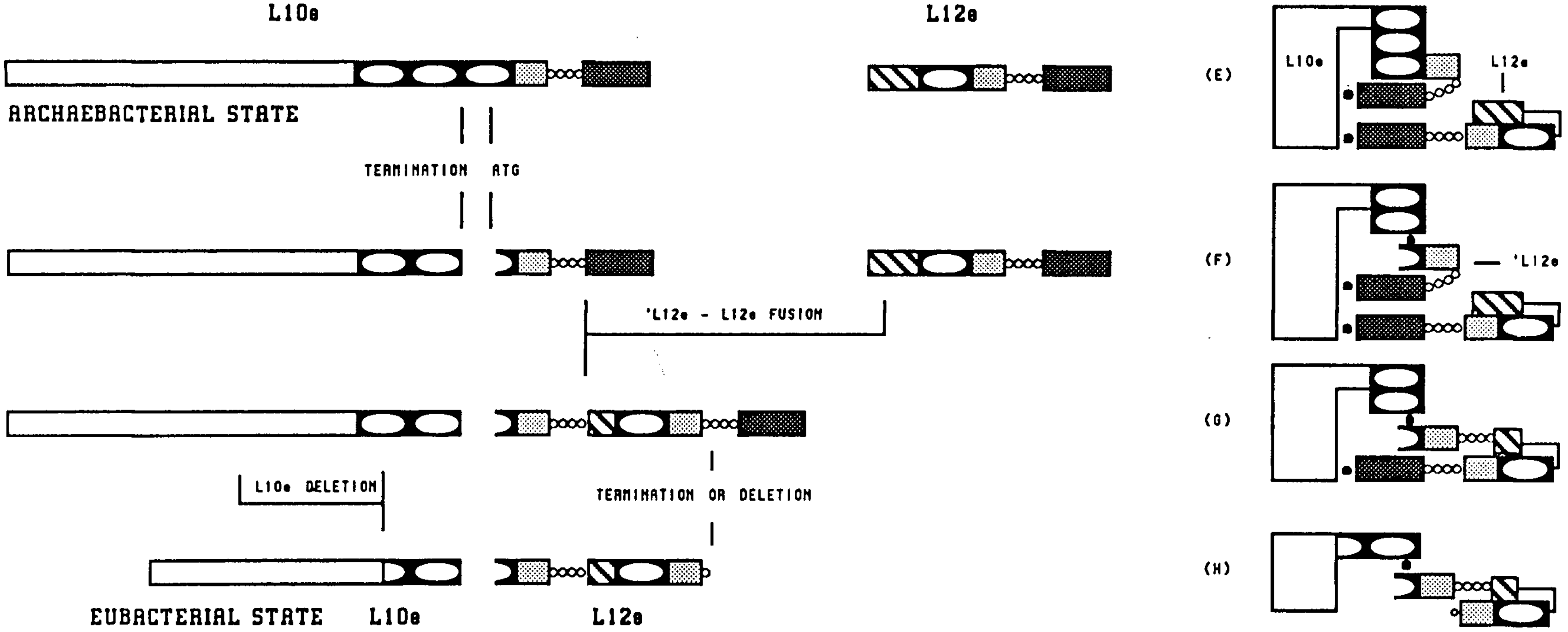
Functional, structural and sequence similarity information all indicate that the genes encoding the contemporary L10e and L12e proteins were derived from single ancestral genes present in the common primordial ancestor. During evolution these genes have undergone numerous alterations and rearrangements within the progenote and in the independent lines of descent to produce a variety of products (Figure 21). By recognizing common and conserved features in the proteins encoded by these genes, it is possible not only to suggest a structure for the ancestral genes but also to construct a model that integrates the hypothetical genetic and structural evolution of the L10e and L12e genes and proteins in eucaryota, archaebacteria, and eubacteria (Figure 22).

It is postulated that initially (in the preprogenote or progenote state) there existed a single ancestral L10e protein at least 210 amino acids long that lacked the modular sequence and the highly charged carboxy terminus and a single ancestral L12e gene composed of an amino terminal globular domain and highly charged carboxy terminus separated by a flexible hinge (i.e. structurally similar to the archaebacterial and eucaryotic L12e proteins). The highly charged moiety and conserved hydrophobic residues of the carboxy terminus of the L12e type proteins may have been responsible for or facilitated the binding to the L10e protein and / or the dimerization of the L12e protein. Duplication of the L12e gene (presumably to ensure the elevated stoichiometry of the L12e dimer(s) in the ribosome) and subsequent divergence produced the type I and type II genes found in contemporary eucaryotes; the type II gene was possibly lost during the formation of the L11e - L1e - L10e - L12e gene cluster and development of translational enhancement of the L12e gene in the archaebacterial and eubacterial lineages. A further duplication of the L12e type I gene provided an extra copy for the gene fusion event which created a splice between L10e and one of the copies of L12e; the fusion junction was possibly immediately preceding the conserved basic residue at position 8 of Figure 20 where the intron occurs in the Sce L12eIIB gene. This fusion would have facilitated the autoassociation of the L10e - L12e complex. In addition, if the module contains a dimerization site as has been suggested, this would allow specific targeting of the globular domain of the L12e protein to the fusion protein. Duplication of the module within the L10e gene results in the present eucaryotic state and a second module duplication produced the contemporary archaebacterial state.

Figure 22 Model of the Evolution of the L10e and L12e Genes and Proteins

Illustrated is a model to explain how simple rearrangements might explain the evolutionary divergence and contemporary relationships between eubacterial, archaebacterial and eucaryotic L10e and L12e genes and proteins. The genes and proteins are illustrated sinister and dexter respectively. The stages and intermediates are illustrated from A to H: (A), L10e and L12e gene structures in the progenote. The binding of the L12e carboxy terminus to the L10e protein is illustrated with a filled circle (•). For clarity only a single L12e protein is illustrated and the globular domain is shown folded to indicate the compactness of this domain. (B), Duplication of the L12e and divergence resulting in the L12e type I and II genes. A further duplication results in 2 copies of L12e type I. (C), a deletion fuses the 3' portion of an L12e type I gene copy to the L10e gene. (D), duplication within the L10e - L12e fusion gene of a 26 codon long module originally from the L12e gene sequence results in the contemporary eucaryotic state. (E), a second duplication of the module within L10e, loss of the L12e type II gene results in the contemporary archaebacterial state. (F), the eubacterial state may have arisen from either the eucaryotic or archaebacterial states; for simplicity only descent from the archaebacterial state is illustrated. Within the eubacterial line translation stop and start codons are generated within the fusion gene to produce two separate and nonoverlapping open reading frames. Part of the distal module is lost. The smaller ORF, designated 'L12e, remains bound to the truncated L10e through the carboxy terminus and also the partial module. (G), 'L12e is fused by deletion with the second copy of the ancestral L12e gene to produce a 'L12e - L12e hybrid containing two L10e binding sites and two hinges. (H), generation of a translation termination codon or a carboxy terminal deletion truncates the 'L12e - L12e fusion to the contemporary eubacterial L12e state. In this gene the proximal and distal regions encoding the amino terminus and the carboxy domain exhibit a degree of sequence similarity and the hinge is relocated. Binding of the L12e protein to the L10e protein is now only through the amino terminus. A deletion within the L10e gene generates the contemporary eubacterial L10e gene.


ANCESTRAL L10e
ANCESTRAL L12e
GLOBULAR DOMAIN
HINGE
MODULE
DUPLICATION AND DIVERGENCE
L12eI
L10e - L12eI FUSION
L12eII
MODULE
DUPLICATION
EUCARYOTIC STATE
MODULE
DUPLICATION
FORMATION OF THE
L11e, L1e, L10e AND L12e
GENE CLUSTER
ARCHAEBACTERIAL STATE
LOSS OF L12eII
(A)
L10e
L12e
(B)
L12eI AND L12eII
(C)
(D)
(E)


L10e
L12e
ARCHAEBACTERIAL STATE
TERMINATION
ATG
'L12e - L12e FUSION
L10e DELETION
TERMINATION OR DELETION
EUBACTERIAL STATE
L10e
L12e
(E)
L10e
L12e
(F)
'L12e
(G)
(H)

At this point the eubacteria diverged from the archaebacteria and eucaryotes. If the 12 amino acid partial Eco L10 β module is real then the eubacterial L10e protein probably evolved from a three module ancestor, if the match is fortuitous then evolution may have occurred from a two module ancestor. The eucaryotic L10e state of two modules may have arisen from deletion of a module from the three module archaebacterial L10e state. Thus it is impossible to determine the branching order of eubacteria, archaebacteria and eucaryota from the present data. For simplicity the derivation of the eubacterial state is described as if arising from a three module L10e ancestor.

Four additional steps are required to achieve the contemporary eubacterial state. First, within the L10e fusion gene a translation start site is generated in the γ module and a translation stop is generated upstream at the 3' end of the β module resulting in production of a short 'L12e peptide. This short peptide remains associated with the L10e protein through its carboxy terminus and through the partial module binding to the remaining two modules of the L10e protein. Second, fusion of the 'L12e gene to the L12e gene removing the unique carboxy terminal end of the 'L12e protein and preserving the functional factor binding globular domain in the L12e protein. This results in a 'L12e - L12e fusion protein associated with the L10e protein through two flexible hinge regions. Third, the deletion or termination of the unique carboxy terminal sequence of the 'L12e - L12e protein leaving the present eubacterial L12e state. The L12e protein binds to L10e through the partial module of its amino terminus. The modern L12e of eubacteria have ragged amino terminal ends, starting between positions 1 and 8 on Figure 18. This may represent fine tuning of the amino terminal binding domain during the evolution of the primary eubacterial lineages. Finally, an internal deletion in L10e (possibly removing the now redundant carboxy terminal binding site) resulting in the shortened contemporary eubacterial L10e gene.

A number of previously published models of the interkingdom relationships of the structural and functional features of the L12e proteins have considered the evolution of the L12e genes (and proteins) in isolation. This has resulted in an enigma: a perplexing series of alignments derived from a variety of sequence and structural criteria, some of which are equally meritorious but apparently mutually exclusive. Alignments based on duplications (Amons *et al.*, 1979; Jue *et al.*,1980), linear correspondence (Yaguchi *et al.*, 1980; Wittmann-Liebold, 1985), transpositions (Lin *et al.*, 1982; Otaka *et al.*, 1985; Matheson, 1985), and conservation of structural features (Liljas *et al.*, 1986) have been proposed. All of these alignments consider the evolution of the L12e gene (and protein) in isolation.

The interkingdom alignments, structural and functional features and evolution of the L12e genes (and proteins) have been considered in concert with those of the related L10e genes (and proteins). The most likely evolutionary events are those which preserve the structure and function of the L10e - L12e complex and the alignments presented for the L10e, L12e and interprotein relationship between the L10e and L12e proteins permits preservation of the structure and function of the complex and resolves some of the enigmatic features of the previous models presented for evolution of the L12e protein.

The sheer variety of different alignments yielding approximately equivalent sequence similarities suggests that the L12e protein originally arose from duplications of a shorter peptide sequence; Jue *et al.* (1980) have suggested that the eubacterial L12e protein is derived from a quadruplication of a 30 amino acid long peptide. Although this does not take into account the now known structural features of the L12e protein (domains and hinge) it is interesting to note that the third peptide corresponds fairly well to the module.

Amons *et al.* (1979) have suggested duplication of the eubacterial gene giving rise to the archaebacterial - eucaryotic gene. However, the presence of the unique highly charged carboxy terminus of the archaebacterial - eucaryotic protein makes this derivation unlikely in my view. An end to end linear alignment of all L12e proteins suggested by Wittmann-Liebold (1985) results in very low sequence similarity (typically 18% interkingdom similarity) and fails to conserve structural features (e.g. the hinge).

Yaguchi *et al.* (1980) first proposed the alignment that conserves the unique arginine residue (position 117, Figure 18) and preserves the factor interacting globular domain. However, their alignment of the globular domain is based only on the *E. coli* and *H. cutirubrum* sequences and differs from that presented in Figure 18 in the region immediately after the arginine residue, where they introduce a 9 amino acid gap in the *H. cutirubrum* archaebacterial - eucaryotic L12e type protein. This gap would eliminate most of the putative dimerization site in the archaebacterial and eucaryotic L10e and L12e modules.

Lin *et al.* (1982) and Otaka *et al.* (1985) have aligned the amino terminus of Eco L12 to the region of the 28th to 34th residue (positions 103 - 113, Figure 18) of the archaebacterial and eucaryotic type I L12e proteins. Their alignment puts the entire amino terminus of the eubacterial L12e protein within the highly conserved region of the module. They have considered the evolution of the L12e protein in isolation and thus have forced a one to one correspondence between the eubacterial and archaebacterial - eucaryotic proteins by transposing the 36 residues of the amino terminal end of the archaebacterial - eucaryotic protein to the carboxy terminus of the eubacterial protein. This would mean that the globular factor binding domain of the eubacterial protein would have to be derived from the fusion of the unique highly charged carboxy terminal domain of the archaebacterial - eucaryotic protein and its amino terminus. The generation of this extremely compact functional domain by such means appears unlikely.

Matheson (1985) has proposed an alignment conserving the unique arginine but transposing the central 35 amino acids of the archaebacterial - eucaryotic protein to the amino terminus to yield the eubacterial protein. The globular factor binding domain of eubacterial L12e would have to be derived from the fusion of the amino terminus and the unique highly charged carboxy terminus of the archaebacterial - eucaryotic protein. As for the previous transposition model this is unlikely.

Liljas *et al.* (1986) has suggested that the amino terminal end of the archaebacterial - eucaryotic protein represents the factor binding domain and the hydrophobic extreme carboxy terminus represents the binding site to the L10e protein; the two domains being separated by the alanine - proline rich hinge structure. The structure of the archaebacterial - eucaryotic L12e would correspond to an inverted eubacterial L12e structure. The crystal structure of the Eco L12 globular domain has its ends in close spatial proximity; thus the conversion of the L10e binding site from the archaebacterial - eucaryotic carboxy terminus to the eubacterial amino terminus necessitates only a small shift in the joining of the hinge from one end to the other end of the globular domain. The eucaryotic L12e proteins have more sequence on the carboxy end of the common globular domain (approximate positions 144 - 153, Figure 18) than the archaebacterial L12e proteins. The alignment would suggest that this is part of the ancestral globular domain and thus the actual spatial proximity of the amino and carboxy termini of the globular domains is unknown. The alignment of Liljas *et al.* was based on secondary structure prediction and has 22% identities at the amino acid level. The amino acid secondary structure of L12e proteins tends to be difficult to accurately predict, the proposed alignment of Figure 18 fits such predictions as well as their previous model, particularly over the module region. In their model the dimerization site of Eco L12 aligns to the beginning and therefore the unconserved portion of the L10e modules; the highly conserved central region of the modules aligns to a region in eubacterial L12e proteins that is less well conserved. The alignment illustrated in Figure 18 shifts the start of the globular domain alignment by 15 amino acid residues (from position 75 to position 90, Figure 18) and achieves both higher sequence similarity (28%) and alignment of the dimerization site of L12e with the conserved region of the L10e protein modules. The structural proposal of Liljas *et al.* (1986) appears to be fundamentally correct and is preserved in the alignment illustrated in FIgure 18.

Although virtually all of these derivations of the eubacterial L12e protein are possible by varying the positions of the fusion, deletion and termination events in the proposed model (Figure 22), the most likely evolutionary events are those which preserve the structure and function of the L10e - L12e complex. The proposed model always retains an L12e protein that has a potential L10e binding site and a preserved functional globular domain separated by a hinge. The previous models do not retain all of these features.

5.3 Future Prospects

A number of archaebacterial genes involved in transcription, translation and metabolism have now been characterized and the forefront of research on the molecular biology of archaebacteria is now focussing on cellular processes, especially by analysis *in vivo.* Recently the development of a transformation system for *Haloferax volcanii* utilizing a shuttle vector capable of propagation and selection (mevinolin resistance) in both *E. coli* and *H. volcanii* and the demonstration of homologous recombination in H.volcanii has made such *in vivo* genetic studies possible for the halophilic archaebacteria (Charlebois *et al.*, 1987; Cline *et al.*, 1989; Lam and Doolittle, 1989).

As nothing is known in archaebacteria concerning the functional importance of specific nucleotides within (or surrounding) consensus signal sequences, the first objective of future research on the molecular biology of the GTPase domain will be directed toward elucidation of i) the regions (and bases) important in the promotion and termination of transcription, ii) the interaction of the ribosome at the translation initiation site and iii) the autogenous regulation within a ribosomal protein gene cluster. The gene cluster characterized in this thesis serves as a source of five promotors of widely differing (i.e. 500 fold) efficiency, four terminators, eight translational initiation sites and the L1e autogenous regulatory site, all of which can be modified by deletion or site directed mutagenesis and analysed *in vivo* by transformation into *H. volcanii.*

The second objective is investigation of the structure to function relationships within and between the GTPase domain proteins, especially of the various unique and duplicated domains present within the L10e and L12e proteins of all urkingdoms. The domains of these proteins can be converted into cassettes bounded by restriction sites and hybrid proteins composed of various wild type and mutagenized domains from all urkingdoms can be constructed and transformed into various hosts to elucidate the functionality of the domains of the recombinant proteins.

The composition, structure and evolution of the progenote ribosome and the subsequent evolutionary divergence into the urkingdoms will be better understood upon completion of the projects characterizing the entire repertory of ribosomal proteins for the archaebacterium *Halobacterium marismortui* and the eucaryote *Saccharomyces cerevisiae* (Otaka *et al.*, 1984; Kimura *et al.*, 1989). Characterization of the GTPase domain genes of the organisms representing the earliest eubacterial branch (*Thermotoga maritima*), archaebacterial branch (*Thermococcus celer*) and eucaryotic branch (*Giardia lamblia*), also presently underway, should yield insights into the early evolution of the GTPase domain. The GTPase domain genes are extremely ancient, their evolution occurring in a series of discrete steps over the interval from the preprogenote state, through the primary speciation event giving rise to the urkingdoms and extending well into the main eubacterial lineages. The evolution of the L10e and L12e proteins exhibits a series of discrete alterations over the interval of contention between the phylogenetic trees of Woese, Cavalier - Smith and Lake, that is, during the primary speciation event giving rise to the urkingdoms (Woese and Olsen, 1986; Cavalier - Smith, 1987; Lake, 1988). A discrete phylogenetic tree of the evolution of the translation apparatus may eventually be constructed over the contentious time span if during this time a sufficient number of ribosomal proteins either first appeared or share the complex alterations exhibited by the GTPase domain proteins.

References

Amons, R., Pluijms, W. and Moller, W. (1979) The primary structure of ribosomal protein eL12 / eL12 - P from *Artemia salina* 80S ribosomes. *FEBS Letters* **104:** 85 - 89.

Amons, R., Pluijms, W., Kriek, T. and Moller, W. (1982) The primary structure of protein eL12' / eL12' - P from the large subunit of *Artemia salina* ribosomes. *FEBS Letters* **146:** 143 - 147.

Atkinson, T. and Smith, M. (1984) Solid phase synthesis of oligodeoxyribonucleotides by the phosphite - triester method. *In* Oligonucleotide Synthesis. A Practical Approach. *Edited by* Gait, M.J. IRL Press, Oxford / Washington, pp. 35 - 81.

Auer, J., Lechner, K. and Bock, A. (1989) Gene organization and structure of two transcriptional units from *Methanococcus* coding for ribosomal proteins and elongation factors. *Can. J. Microbiol.* **35:** 200 - 204.

Auffray, C. and Rougesn, F. (1980) Purification of mouse immunoglobulin heavy - chain messenger RNAs from total myeloma tumor RNA. *Eur. J. Biochem.* **107:** 303 - 314.

Barry, G., Squires, C. and Squires, C.L. (1979) Control features within the rplKAJL-rpoBC transcription unit of *Escherichia coli. Proc. Natl. Acad. Sci. USA* **76:** 4922 - 4926.

Barry, G., Squires, C.L. and Squires, C. (1980) Attenuation and processing of RNA from the rplJL-rpoBC transcription unit of *Escherichia coli. Proc. Natl. Acad. Sci. USA* **77:** 3331 - 3335.

Bartsch, M., Kimura, M. and Subramanian, A.R. (1982) Purification, primary structure, and homology relationships of a chloroplast ribosomal protein. *Proc. Natl. Acad. Sci. USA* **79:** 6871 - 6875.

Baughman, G. and Nomura, M. (1983) Localization of the target site for translational regulation of the L11 operon and direct evidence for translational coupling in *Escherichia coli. Cell* **34:** 979 - 988.

Bayley, S.T. (1966) Composition of ribosomes of an extremely halophilic bacterium. *J. Mol. Biol.* **15:** 420 - 427.

Bayley, S.T. (1971) Protein synthesis systems from halophilic bacteria. *In* Protein Biosynthesis in Bacterial Systems. *Edited by* Last, J.A. and Laskin, A.I. Marcel Dekker, New York, pp. 89 - 110.

Bayley, S.T. and Kushner, D.J. (1964) The ribosomes of the extremely halophilic bacterium, *Halobacterium cutirubrum. J. Mol. Biol.* **9:** 654 - 669.

Bayley, S.T. and Morton, R.A. (1978) Recent developments in the molecular biology of extremely halophilic bacteria. *CRC Crit. Rev. Microbiol.* **6:** 151 - 205.

Beauclerk, A.A.D., Cundliffe, E. and Dijk, J. (1984) The binding site for ribosomal protein complex L8 within 23S ribosomal RNA of *Escherichia coli. J. Biol. Chem.* **259:** 6559 - 6563.

Beauclerk, A.A.D., Hummel, H., Holmes, D.J., Bock, A. and Cundliffe, E. (1985) Studies of the GTPase domain of archaebacterial ribosomes. *Eur. J. Biochem.* **151:** 245 - 255.

Beltrame, M. and Bianche, M.E. (1987) Sequence of the cDNA for one acidic ribosomal protein of *Schizosaccharomyces pombe. Nucleic Acids Res.* **15:** 9089.

Betlach, M., Friedman, J., Boyer, H.W. and Pfeifer, F. (1984) Characterization of a halobacterial gene affecting bacterio - opsin gene expression. *Nucleic Acids Res.* **12:** 7949 - 7959.

Betlach, M., Shand, R. and Leong, D. (1989) Regulation of the bacterio - opsin gene of a halophilic archaebacterium. *Can. J. Micro.* **35:** 134 - 140.

Betlach, M.C., Leong, D. and Boyer, H.W. (1986) Bacterio - opsin gene expression in *Halobacterium halobium. Syst. Appl. Microbiol.* **7:** 83 - 91.

Biggin, M.D., Gison, T.J. and Hong, G.F. (1983) Buffer gradient gels and 35S label as an aid to rapid DNA sequence determination. *Proc. Natl. Acad. Sci. USA* **80:** 3963 - 3965.

Birnboim, H.C. and Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* **7:** 1513 - 1523.

Blanck, A. and Oesterhelt, D. (1987) The halo - opsin gene. 2. Sequence, primary structure of halorhodopsin and comparison with bacteriorhodopsin. *EMBO Journal* **6:** 265 - 273.

Boissy, R. and Astell, C.R. (1985) An *Escherichia coli* recBCsbcBrecF host permits the deletion - resistant propagation of plasmid clones containing the 5'-terminal palindrome of minute virus of mice. *Gene* **35:** 179-185.

Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.C., Heynecker, H.L., Boyer, H.W., Crosa, J.H. and Falkow, S. (1977) Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. *Gene* **2:** 95 - 113.

Branlant, C., Krol, A., Machatt, A. and Ebel, J.P. (1981) The secondary structure of the protein L1 binding region of the ribosomal 23S RNA: homologies with putative secondary structures of the L11 mRNA and of a region of mitochondrial 16S rRNA. *Nucleic Acids Res.* **9:** 293 - 307.

Brown, J.W., Thomm, M., Beckler, G.S., Frey, G., Stetter, K.O. and Reeve, J.N. (1988) An archaebacterial RNA polymerase binding site and transcription initiation of the hisA gene in *Methanococcus vanielii. Nucleic Acids Res.* **16:** 135 - 150.

Burma, D.P., Srivastava, S., Srivastava, A.K., Mahanti, S. and Dash, D. (1986) Conformational change of 50S ribosomes during protein synthesis. *In* Structure, Function and Genetics of Ribosomes. *Edited by* Hardesty, B. and Kramer, G. Springer Verlag, New York, pp. 438 - 453.

Cavalier - Smith, T. (1986) The kingdoms of organisms. *Nature* **324:** 416 - 417.

Cavalier - Smith, T. (1987) The origin of the eucaryote and archaebacterial cells. *Ann. NY Acad. Sci.* **503:** 17 - 54.

Chant, J. and Dennis, P.P. (1986) Archaebacteria: transcription and processing of ribosomal RNA sequences in *Halobacterium cutirubrum. EMBO Journal* **5:** 1091 - 1097.

Chant, J.S., Hui, I., de Jong-Wong, D., Shimmin, L.C. and Dennis, P.P. (1986) The protein synthesizing machinery of the archaebacterium *Halobacterium cutirubrum* : molecular characterization. *System. Appl. Microbiol.* **7:** 106 - 114.

Charlebois, R.L. and Doolittle, W.F. (1988) Transposable elements and genomic structure in Halobacteria. *In* Mobile DNA. *Edited by* Howe, M. and Beng, D. American Society for Microbiology, Washington, D.C., pp. 297 - 307.

Charlebois, R.L., Lam, W.L., Cline, S.W. and Doolittle, W.F. (1987) Characterization of pHV2 from *Halobacterium volcanii* and its use in demonstrating transformation of an archaebacterium. *Proc. Natl. Acad. Sci. USA* **84:** 8530 - 8534.

Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* **18:** 5294 - 5299.

Christensen, T., Johnsen, M., Fiil, N.P. and Friesen, J.D. (1984) RNA secondary structure and translation inhibition: analysis of mutants in the rplJ leader. *EMBO Journal* **3:** 1609 - 1612.

Cline, S.W., Lam, W.L., Charlebois, R.L., Schalkwyk, L.C. and Doolittle, W.F. (1989) Transformation methods for halophilic archaebacteria. *Can. J. Microbiol.* **35:** 148 - 152.

Cowgill, C.A., Nichols, B.G., Kenny, J.W., Butler, P., Bradbury, E.M. and Traut, R.R. (1984) Mobile domains in ribosomes revealed by proton nuclear magnetic resonance. *J. Biol. Chem.* **259:** 15257 - 15263.

Cram, D.S., Sherf, B.A., Libby, R.T., Mattaliano, R.J., Ramachandran, K.L. and Reeve, J.N. (1987) Structure and expression of the genes, mcrBDCGA, which encode the subunits of component C of methyl coenzyme M reductase in *Methanococcus vanielii*. *Proc. Natl. Acad. Sci. USA* **84:** 3992 - 3996.

Cue, D., Beckler, G.S., Reeves, J.N. and Konisky, J. (1985) Structure and sequence divergence of two archaebacterial genes. *Proc. Natl. Acad. Sci. USA* **82:** 4207 - 4211.

Cundliffe, E. (1986) Involvement of specific portions of ribosomal RNA in defined ribosomal functions: a study utilizing antibiotics. *In* Structure, Function and Genetics of Ribosomes. *Edited by* Hardesty, B. and Kramer, G. Springer Verlag, New York, pp. 586 - 604.

Dale, R.M.K., McClure, B.A. and Houchins, J.P. (1985) A rapid single - stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: application to sequencing the corn mitochondrial 18S rDNA. *Plasmid* **13:** 31 - 40.

Daniels, C.J., Douglas, S.E. and Doolittle, W.F. (1986) Genes for transfer RNAs in *Halobacterium halobium. Syst. Appl. Microbiol.* **7:** 26 - 29.

Danson, M.J. (1989) Central metabolism of the archaebacteria: an overview. *Can. J. Microbiol.* **35:** 58 - 64.

Das Sarma, S., Damerval, T. and de Marsac, N.T. (1987) A plasmid-encoded gas vesicle gene in halophilic archaebacterium. *Mol. Microbiol.* **1:** 365 - 370

Dayhoff, M. (1978) Atlas of Protein Sequence and Structure, 8th Edition. MacGregor and Warner, Silver Springs, Maryland, pp. 352 - 375.

Dean, D. and Nomura, M. (1980) Feedback regulation of ribosomal protein gene expression in *Escherichia coli. Proc. Natl. Acad. Sci. USA* **77:** 3590 - 3594.

Deng, H., Odom, O.W. and Hardesty, B. (1986) Localization of L11 on the *E. coli* ribosome by singlet-singlet energy transfer. *Eur. J. Biochem.* **156:** 497 - 503.

Dennis, P.P. (1977) Transcription patterns of adjacent segments on the chromosome of *E. coli* containing genes coding for four 50S ribosomal proteins and the β and β ' subunits of RNA polymerase. *J. Mol. Biol.* **115:** 603 - 625.

Dennis, P.P. (1984) Site specific deletions of regulatory sequences in a ribosomal protein - RNA polymerase operon in *Escherichia coli. J. Biol. Chem.* **259:** 3202 - 3209.

Dennis, P.P. (1985) Multiple promotors for the transcription of the ribosomal RNA gene cluster in *Halobacterium cutirubrum. J. Mol. Biol.* **186:** 457 - 461.

Dennis, P.P., Hui, I., Shimmin, L.C., McPherson, J., Pao, C.C. and Matheson, A.T. (1985) Archaebacteria: our first look at their ribosome component genes. *In* The Molecular Biology of Bacterial Growth. *Edited by* Schaechter, M., Neidhardt, F.C., Ingraham, J.L. and Kjeldgaard, N.O. Jones and Bartlett Publishers, Inc., Boston, pp.78 - 91.

Dente, L., Cesareni, G. and Cortese, R. (1983) pEMBL: a new family of single stranded plasmids. *Nucleic Acids Res.* **6:** 1645 - 1655.

Dijk, J., Garrett, R.A. and Muller, R. (1979) Studies on the binding of the ribosomal protein complex L7/L12-L10 and protein L11 to the end one third of the 23S RNA: a functional center of the 50S subunit. *Nucleic Acids Res.* **6:** 2717 - 2729.

Dognin, W.L. and Wittman-Liebold, B. (1977) The primary structure of L11, the most heavily methylated protein from *Escherichia coli* ribosomes. *FEBS Letters* **84:** 342 - 346.

Duggleby, R.G., Kaplan, H. and Visentin, L.P. (1975) Carboxy - terminal sequences of procaryotic ribosomal proteins from *Escherichia coli, Bacillus stearothermophilus*, and *Halobacterium cutirubrum. Can J. Biochem.* **53:** 827 - 833.

Dunn, R., McCoy, J., Simsek, M., Majumdar, A., Chang, S.H., RajBhandary, U.L. and Khorana, H.G. (1981) The bacteriorhodopsin gene. *Proc. Natl. Acad. Sci. USA* **78:** 6744 - 6748.

Eisenberg, H. and Wachtel, E.J. (1987) Structural studies of halophilic proteins, ribosomes and organelles of bacteria adapted to extreme salt concentrations. *Ann. Rev. Biophys. Chem.* **16:** 69 - 92.

El-Baradi, T.T.A.L., de Regt, C.H.F., Einerhand, S.W.C., Teixido, J., Planta, R.J., Ballesta, J.P.G. and Raue, H.A. (1987) Ribosomal proteins EL11 from *Escherichia coli* and L15 from *Saccharomyces cerevisiae* bind to the same site in both yeast 26S and mouse 28S rRNA. *J. Mol. Biol.* **195:** 909 - 917.

Falkenberg, P., Yaguchi, M., Roy, C., Zuker, M. and Matheson, A.T. (1985) The primary structure of the ribosomal A-protein (L12) from the moderate halophile NRCC 41227. *Biochem. Cell Biol.* **64:** 675 - 680.

Favaloro, J., Treisman, R. and Kamen, R. (1980) Transcription maps of polyoma virus-specific RNA: analysis by two dimensional nuclease S1 gel mapping. *Methods in Enzymology* **65:** 718 - 749.

Fienberg, A.P. and Vogelstein, B. (1982) A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132:** 6 - 13.

Fiil, N.P., Friesen, J.D., Downing, W.D. and Dennis, P.P. (1980) Post-transcriptional regulatory mutants in a ribosomal protein - RNA polymerase operon of *E. coli. Cell* **19:** 837 - 844.

Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Luersen, K.R., Chan, K.N. and Woese, C.R. (1980) The phylogeny of prokaryotes. *Science* **209:** 457 - 463.

Freisen, J.D., Fiil, N.P., Parker, J.M. and Haseltine, W.A. (1974) A new relaxed mutant of *Escherichia coli* with an altered 50S ribosomal subunit. *Proc. Natl. Acad. Sci. USA* **71:** 3465 - 3469.

Garland, W.G., Louie, K.A., Matheson, A.T. and Liljas, A. (1987) The complete amino acid sequence of the ribosomal 'A' protein (L12) from *Bacillus stearothermophilus. FEBS Letters* **220:** 43 - 46.

Garrett, R.A. (1985) The uniqueness of archaebacteria. *Nature* **318:** 233 - 235.

Giri, L., Dijk, J., Labischinski, H. and Bradaczek, H. (1978) Shape of protein L11 from the 50S ribosomal subunit of *Escherichia coli*. *Biochemistry* **17:** 745 - 749.

Gourse, R.L., Thurlow, D.L., Gerbi, S.A. and Zimmerman, R.A. (1981) Specific binding of a prokaryotic ribosomal protein to a eukaryotic ribosomal RNA: implications for evolution and autoregulation. *Proc. Natl. Acad. Sci. USA* **78:** 2722 - 2726.

Gouy, M. and Li, W. - H. (1989) Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. *Nature* **339:** 145 - 147.

Gudkov, A.T. and Behlke, J. (1978) The N - terminal sequence of L7/L12 is responsible for its dimerization. *Eur. J. Biochem.* **90:** 309 - 312.

Gudkov, A.T. and Gongadze, G.M. (1984) The L7/L12 proteins change their conformation upon interaction of EF-G with ribosomes. *FEBS Letters* **176:** 32 - 36.

Gudkov, A.T., Venyaminov, S.Y. and Matheson, A.T. (1984) Physical studies on the ribosomal "A" protein from two archaebacteria, *Halobacterium cutirubrum* and *Methanobacterium thermoautotrophicum. Can. J. Biochem. Cell Biol.* **62:** 44 - 48.

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

Hanauz, G., Stoffler-Melicke, M. and van Heel, U. (1987) Characteristic views of prokaryotic 50S ribosomal subunits. *J. Mol. Evol.* **28:** 347 - 357.

Hattori, M. and Sakaki, Y. (1986) Dideoxy sequencing method using denatured plasmid templates. *Anal. Biochem.* **152:** 232 - 238.

Henikoff, S. (1984) Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28:** 351 - 359.

Hori, H., Itoh, T. and Osawa, S. (1982) The phylogenetic structure of the metabacteria. *Zbl. Bakt. Hyg., I. Abt. Orig.* **C3:** 18 - 30.

Horne, M., Englest, C. and Pfeifer, F. (1988) Two gas vacuole proteins in *Halobacterium halobium. Mol. Gen. Genet.* **213:** 459 - 464.

Hui, I. and Dennis, P.P. (1985) Characterization of the ribosomal RNA gene clusters in *Halobacterium cutirubrum. J. Biol. Chem.* **260:** 899 - 906.

Itoh, T. (1981a) Primary structure of an acidic ribosomal protein from *Micrococcus lysodeikticus. FEBS Letters* **127:** 67 - 70.

Itoh, T. (1981b) Primary structure of an acidic ribosomal protein YPA1 from *Saccharomyces cerevisiae.* Isolation and characterization of peptides and the complete amino acid sequence. *Biochim. Biophys. Acta* **671:** 16 - 24.

Itoh, T. (1988) Complete nucleotide sequence of the ribosomal 'A' protein operon from the archaebacterium *Halobacterium halobium. Eur. J. Biochem.* **176:** 297 - 303.

Itoh, T. and Higo, K.I. (1983) Complete amino acid sequence of an L7/L12-type ribosomal protein from *Rhodopseudomonas spheroides. Biochim. Biophys. Acta* **744:** 105 - 109.

Itoh, T. and Wittman-Liebold, B. (1978) The primary structure of *Bacillus subtilis* acidic ribosomal protein B-L9 and its comparison with *Escherichia coli* proteins L7/L12. *FEBS Letters* **96:** 392 - 394.

Itoh, T., Kamazaki, T., Sugiyama, M. and Otaka, E. (1988) Molecular cloning and sequence analysis of the ribosomal 'A' protein gene from the archaebacterium, *Halobacterium halobium. Biochim. Biophys. Acta* **949:** 110 - 118.

Itoh, T., Sugiyama, M. and Higo, K.I. (1982) The primary structure of an acidic ribosomal protein from *Streptomyces griseus. Biochim. Biophys. Acta* **701:** 164 - 172.

Jinks-Robertson, S. and Nomura, M. (1987) Ribosomes and tRNA. *In Escherichia coli* and *Salmonella typhimurium* ; Cellular and Molecular Biology, Vol. 2. *Edited by* Neidhardt, F.C. American Society for Microbiology, Washington D.C., pp. 1358 - 1385.

Johnsen, M., Christiensen, T., Dennis, P.P. and Fiil, N.P. (1982) Autogenous control: ribosomal protein L10-L12 complex binds to the leader of its mRNA. *EMBO Journal* **1:** 999 - 1004.

Jones, W.J., Nagle Jr., D.P. and Wittmann, W.B. (1987) Methanogens and the diversity of archaebacteria. *Microbiol. Rev.* **51:** 135 - 177.

Jue, R.A., Woodbury, N.W. and Doolittle, R.F. (1980) Sequence homologies among *E. coli* ribosomal proteins: evidence for evolutionarily related groupings and internal duplications. *J. Mol. Evol.* **15:** 129 - 148.

Kastner, B., Stoffler-Meilicke, M. and Stoffler, G. (1981) Arrangement of the subunits in the ribosome of *Escherichia coli* : demonstration by immunoelectron microscopy. *Proc. Natl. Acad. Sci. USA* **78:** 6652 - 6656.

Kearney, K.R. and Nomura, M. (1987) Secondary structure of the autoregulatory mRNA binding site of ribosomal protein L1. *Mol. Gen. Genet.* **210:** 609 - 618.

Kimura, M., Arndt, E., Hatakeyama, T., Hatekeyama, T. and Kimura, J. (1989) Ribosomal proteins in halobacteria. *Can. J. Microbiol.* **35:** 195 - 199.

Kimura, M., Kimura, J. and Ashman, K. (1985) The complete primary structure of ribosomal proteins L1, L14, L15, L23, L24 and L29 from *Bacillus stearothermophilus. Eur. J. Biochem.* **150:** 491 - 497.

Kjems, J. and Garrett, R.A. (1987) Novel expression of the ribosomal RNA genes in the extreme thermophile and archaebacterium *Desulfurococcus mobilis. EMBO Journal* **6:** 3521 - 3530.

Kopke, A.K.E., and Wittmann - Liebold, B. (1988) DNA sequence of the gene for ribosomal protein L23 from the archaebacterium *Methanococcus vanielii. FEBS Letters* **239:** 313 - 318.

Kopke, A.K.E. and Wittmann - Liebold, B. (1989) Comparative studies of ribosomal proteins and their genes from *Methanococcus vannielii* and other organisms. *Can. J. Microbiol.* **35:** 11 - 20.

Koteliansky, V.E., Domogatsky, S.P., and Gudkov, A.T. (1978) Dimer state of protein L7/L12 and EF-G dependent reactions on ribosomes. *Eur. J. Biochem.* **90:** 319 - 323.

Kushner, D.J. (1985) The Halobacteriaciae. *In* The Bacteria. Vol. VIII. *Edited by* Woese, C.R. and Wolfe, R.S. Academic Press, New York, pp. 171 - 214.

Lake, J.A. (1983a) Evolving ribosome structure: domains in archaebacteria, eubacteria, and eucaryotes. *Cell* **33:** 318 - 319.

Lake, J.A. (1983b) Ribosome evolution: the structural bases of protein synthesis in archaebacteria, eubacteria and eukaryotes. *Prog. in Nucleic Acid Res.* **30:** 163 - 194.

Lake, J.A. (1986a) An alternative to archaebacterial dogma. *Nature* **319:** 626.

Lake, J.A. (1986b) In defense of bacterial phylogeny. *Nature* **321:** 657 - 658

Lake, J.A. (1988) Origin of the eucaryotic nucleus determined by rate-invarient analysis of rRNA sequences. *Nature* **331:** 184 - 186.

Lake, J.A. and Strycharz, W.A. (1981) Ribosomal proteins L1, L17, L27 localized at single sites on the large subunit by immune electron microscopy. *J. Mol. Biol.* **153:** 979 - 992.

Lake, J.A., Henderson, E., Clarke, M.W., Scheinman, A. and Oakes, M.I. (1986) Mapping evolution with three dimensional ribosome structure. *System. Appl. Microbiol.* **7:** 131 - 136.

Lake, J.A., Henderson, E., Oakes, M. and Clark, M.W. (1984) Eocytes : a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl. Acad. Sci. USA* **81:** 3786 - 3790.

Lam, W.L. and Doolittle, W.L. (1989) Shuttle vectors for the archaebacterium *H. volcanii. Proc. Natl. Acad. Sci. USA* **86:** 5478 - 5482.

Larsen, H. (1984) Family V. Halobacteriaceae. *In* Bergey's Manual of Systematic Bacteriology. *Edited by* Krieg, N.R. and Holt, J.G. Williams and Wilkins, Baltimore, pp. 261 - 267.

Lederer, H. (1986) Archaebacterial status quo is defended. *Nature* **320:** 220.

Leffers, H., Gropp, F., Lottspeich, F., Zillig, W. and Garret, R.A. (1989) Sequence, organization, transcription and evolution of RNA polymerase subunit genes from the archaebacterial extreme halophiles *Halobacterium halobium* and *Halococcus morrhuae. J. Mol. Biol.* **206:** 1 - 17.

Leijonmarck, M. and Liljas, A. (1987) Structure of the C - terminal domain of the ribosomal protein L7/L12 from *Escherichia coli* at 1.7Å. *J. Mol. Biol.* **195:** 555 - 580.

Leijonmarck, M., Pettersson, I. and Liljas, A. (1981) Structural studies on the protein L7/L12 from *E. coli* ribosomes. *In* Structural Aspects of Recognition and Assembly in Biological Macromolecules. *Edited by* Balaban, M., Sussmann, J.L., Traub, W. and Yonath, A. ISS, Rehovot and Philadelphia, pp. 761 - 777.

Leong, D., Boyer, H. and Betlach. M. (1988a) Transcription of genes involved in bacterio - opsin gene expression in mutants of a halophilic archaebacterium. *J. Bacteriol.* **170:** 4910 - 4915.

Leong, D., Pfeifer, F., Boyer, H. and Betlach. M. (1988b) Characterization of a second gene involved in bacterio - opsin gene expression in a halophilic archaebacterium. *J. Bacteriol.* **170:** 4903 - 4909.

Liljas, A. (1982) Structural studies of ribosomes. *Prog. Biophys. Mol. Biol.* **40:** 161 - 228.

Liljas, A., Thirup, S. and Matheson, A.T. (1986) Evolutionary aspects of ribosome - factor interactions. *Chemica Scripta* **26B:** 109 - 119.

Lin, A., Wittmann-Liebold, B., McNally, J. and Wool, I.G. (1982) The primary structure of the acidic phosphoprotein P2 from rat liver 60S ribosomal subunits. *J. Biol. Chem.* **257:** 9189 - 9197.

Lindahl, L. and Zengel, J. (1986) Ribosomal genes in *Escherichia coli. Ann. Rev. Genet.* **20:** 297 - 326.

Lindahl, L., Jaskunas, S.R., Dennis, P.P. and Nomura, M. (1975) Cluster of genes in *Escherichia coli* for ribosomal proteins, ribosomal RNA and RNA polymerase. *Proc. Natl. Acad. Sci. USA* **72:** 2743 - 2747.

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

Ma, J.C., Newman, A.J. and Hayward, R.S. (1981) Internal promotors of the rpoBC operon of *Escherichia coli. Mol. Gen. Genet.* **184:** 548 - 550.

Magrum, L.J., Luehrsen, K.R. and Woese, C.R. (1978) Are extreme halophiles actually "bacteria"? *J. Mol. Evol.* **11:** 1 - 8.

Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbour Laboratory, Cold Spring Harbour, New York.

Mankin, A.S. and Kagramanova, V.K. (1986) Complete nucleotide sequence of the single ribosomal RNA operon of *Halobacterium halobium* : secondary structure of the archaebacterial 23S rRNA. *Mol. Gen. Genet.* **202:** 152 - 161.

Mankin, A.S., Teterina, N.L., Rubstov, P.M., Baratova, L.A. and Kagaramanova, V.K. (1984) Putative promotor region of rRNA operon from archaebacteium *Halobacterium halobium. Nucleic Acids Res.* **12:** 6537 - 6546.

Margulis, L. (1970) Origin of eukaryotic cells. Yale University Press, New Haven.

Marquis, D.M., Fahnestock, S.R., Henderson, E., Woo, D., Schwinge, S., Clarke, M.W. and Lake, J.A. (1981) The L7/L12 stalk, a conserved feature of the prokaryotic ribosome, is attached to the large subunit through its N terminus. *J. Mol. Biol.* **150:** 121 - 132.

Matheson, A.T. (1985) Ribosomes of archaebacteria. *In* The Bacteria : Vol. 8, The Archaebacteria. *Edited by* Woese, C.R. and Wolfe, R.S. Academic Press, New York, pp. 761 - 777.

Matheson, A.T., Louie, K.A. and Bock, A. (1988) The complete amino acid sequence of the ribosomal A protein (L12) from the archaebacterium *Sulfolobus acidocaldarius. FEBS Letters* **231:** 331 - 335.

Matheson, A.T., Louie, K.A., Tak, B.D. and Zuker, M. (1987) The primary structure of the ribosomal A - protein (L12) from the halophilic eubacterium *Haloanaerobium praevalens. Biochemie* **69:** 1013 - 1020.

Matheson, A.T., Yaguchi, M., Balch,W.E. and Wolfe, R.S. (1979) Sequence homologies in the N - terminal region of the ribosomal 'A' proteins from *Methanobacterium thermoautotrophicum* and *Halobacterium cutirubrum. Biochim. Biophys. Acta* **626:** 162 - 169.

Matheson, A.T., Yaguchi, M., Christensen, P., Rollin, C.T. and Hasnain, S. (1984) Purification, properties and N - terminal amino acid sequence of certain 50S ribosomal subunit proteins from the archaebacterium *Halobacterium cutirubrum. Can. J. Biochem. Cell Biol.* **62:** 426 - 433.

Maxam, A.M. and Gilbert, W. (1977) Sequencing end - labelled DNA with base specific chemical cleavages. *Proc. Natl. Acad. Sci. USA* **74:** 560 - 564.

Messing, J. (1983) New M13 vectors for cloning. *Methods in Enzymology* **101:** 20 - 79.

Miller, J.H. (1972) Experiments in Molecular Genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, pp. 431 - 433.

Mills, D.R. and Kramer, F.R. (1979) Structure independent nucleotide sequence analysis. *Proc. Natl. Acad. Sci. USA* **76:** 2232 - 2235.

Mitsui, K. and Tsurugi, K. (1988a) cDNA and deduced amino acid sequence of acidic ribosomal protein A0 from *Saccharomyces cerevisiae. Nucleic Acids Res.* **16:** 3573.

Mitsui, K. and Tsurugi, K. (1988b) cDNA and deduced amino acid sequence of acidic ribosomal protein A1 from *Saccharomyces cerevisiae. Nucleic Acids Res. 16:* 3574.

Mitsui, K. and Tsurugi, K. (1988c) cDNA and deduced amino acid sequence of acidic ribosomal protein A2 from *Saccharomyces cerevisiae. Nucleic Acids Res.* **16:** 3575.

Mizusawa, S., Nishimura, S. and Seela, F. (1986) Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy - 7 - deazaguanosine triphosphate in place of dGTP. *Nucleic Acids Res.* **14:** 1319 - 1324.

Moller, W. and Maasen, J.A. (1986) On the structure, function and dynamics of L7/L12 from *Escherichia coli* ribosomes. *In* Structure, Function and Genetics of Ribosomes. *Edited by* Hardesty, B. and Kramer, G. Springer Verlag, New York, pp. 309 - 325.

Möller, W., Schrier, P.I., Maasen, J.A., Zantema, A., Schop, E., Reinalda, H., Cremers, A.F.M. and Mellema, J.E. (1983) Ribosomal proteins L7/L12 of *Escherichia coli.* Localization and possible molecular mechanism in translation. *J. Mol. Biol.* **163:** 553 - 573.

Newman, A. (1987) Specific accessory sequences in *Saccharomyces cerevisiae* introns control assembly of pre - messenger RNAs into spliceosomes. *EMBO Journal* **6:** 3833 - 3839.

Newman, A., Linn, T. and Hayward, R. (1979) Evidence for co-transcription of the RNA polymerase genes rpoBC with a ribosomal protein gene of *Escherichia coli. Mol. Gen. Genet.* **169:** 195 - 204.

Newton, C.H., Shimmin, L.C., Yee, J. and Dennis, P.P. (1990) A family of genes encode the multiple forms of the yeast L12 equivalent ribosomal proteins and a single form of the L10 equivalent ribosomal protein. *J. Bacteriol.* **172:** 579 - 588.

Nomura, M., Gourse, R. and Baughman, G. (1984) Regulation of the synthesis of ribosomes and ribosome components. *Ann. Rev. Biochem.* **53:** 119 - 162.

Oakes, M., Henderson, E., Scheinman, A., Clark, M. and Lake, J.A. (1986) Ribosome structure, function and evolution: mapping ribosomal RNA, proteins and functional sites in three dimensions. *In* Structure, Function and Genetics of Ribosomes. *Edited by* Hardesty, B. and Kramer, G. Springer Verlag, New York, pp. 47 - 67.

Oda, G., Strom, A.R., Visentin, L.P. and Yaguchi, M. (1974) An acidic, alanine - rich 50S ribosomal protein from *Halobacterium cutirubrum* : amino acid sequence homology with *Escherichia coli* proteins L7 and L12. *FEBS Letters* **43:** 127 - 130.

Olsen, G.J. and Woese, C.R. (1989) A brief note concerning archaebacterial phylogeny. *Can. J. Microbiol.* **35:** 119 - 123.

Osterberg, R., Sjoberg, B., Liljas, A. and Pettersson, I. (1976) Small angle X-ray scattering and crosslinking study of the proteins L7/L12 from *Escherichia coli* ribosomes. *FEBS Letters* **66:** 48 - 51.

Osterberg, R., Sjoberg, B., Pettersson, I., Liljas, A. and Kurland, C.J. (1977) Small angle X-ray scattering study of the protein complex of L7/L12 and L10 from *Escherichia coli* ribosomes. *FEBS Letters* **73:** 22 - 24.

Otaka, E., Higo, K. and Itoh, T. (1984) Yeast ribosomal proteins. *Mol. Gen. Genet.* **195:** 544 - 546.

Otaka, E., Ooi, T., Kumazaki, T. and Itoh, T. (1985) Examination of protein sequence homologies: II. six *Escherichia coli* L7/L12 type ribosomal 'A' protein sequences from eukaryotes and metabacteria, contrasted with those from prokaryotes. *J. Mol. Evol.* **22:** 342 - 350.

Parker, J., Watson, R.J., Freisen, J.D. and Fiil, H.P. (1976) A relaxed mutant with an altered ribosomal protein L11. *Mol. Gen. Genet.* **144:** 111 - 114.

Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology* In press.

Petersen, C. (1989) Long - range translational coupling in the rplJL - rpoBC operon of *Escherichia coli. J. Mol. Biol.* **206:** 323 - 332.

Pettersson, I. (1979) Studies on the RNA and protein binding sites of the *E. coli* ribosomal protein L10. *Nucleic Acids Res.* **6:** 2637 - 2646.

Pettersson, I. and Liljas, A. (1979) The stoichiometry and reconstitution of a stable protein complex from *Escherichia coli* ribosomes. *FEBS Letters* **98:** 39 - 144.

Pfeifer, F. and Betlach, M. (1985) Genome organization in *Halobacterium* : a 70 Kb island of more (AT) rich DNA in the chromosome. *Mol. Gen. Genet.* **198:** 449 - 455.

Pfeifer, F., Blasio, U. and Ghorman, P. (1988) Dynamic plasmid populations in *Halobacterium halobium. J. Bacteriol.* **170:** 3718 - 3724.

Pfeifer, F., Blasio, U. and Horne, M. (1989) Genome structure of *Halobacterium halobium* : plasmid dynamics in gas vacuole deficient mutants. *Can. J. Microbiol.* **35:** 96 - 100.

Pfeifer, F., Weidinger, G. and Goebel, W. (1981) Genetic variability in *Halobacterium halobium. J. Bacteriol.* **145:** 375 - 381.

Planta, R., Mager, W., Leer, R., Wondt, L., Raue, H. and El-Baradi, T. (1986) Structure and expression of ribosomal proteins in yeast. *In* Structure, Function and Genetics of Ribosomes. *Edited by* Hardesty, B. and Kramer, G. Springer Verlag, New York, pp. 699 - 718.

Post, L.E., Strycharz, G.D., Nomura, M., Lewis, H. and Dennis, P.P. (1979) Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit in *Escherichia coli. Proc. Natl. Acad. Sci. USA* **76:** 1697 - 1701.

Puhler, G., Lottspei, F. and Zillig, W. (1989) Organization and nucleotide sequence of the genes encoding the large subunit A, subunit B and subunit C of the DNA dependent RNA polymerase of the archaebacterium *Sulfolobus solfataricus. Nucleic Acids Res.* **17:** 4517 - 4534.

Qian, S., Zhang, J., Kay, M.A. and Jacobs-Lorena, M. (1987) Structural analysis of the *Drosophila* rpA1 gene, a member of the eucaryotic 'A' type ribosomal protein family. *Nucleic Acids Res.* **15:** 987 - 1003.

Radermacher, M., Wagenknecht, T., Verschoor, A. and Frank, J. (1987) Three-dimensional structure of the large ribosomal subunit from *Escherichia coli. EMBO Journal* **6:** 1107 - 1114.

Ramirez, C., Shimmin, L.C. and Matheson, A.T. (1990a) Characterization of the GTPase domain gene cluster from *Sulfolobus solfataricus*. Submitted.

Ramirez, C., Shimmin, L.C., Matheson, A.T. and Dennis, P.P. (1990b) Structure of archaebacterial ribosomal proteins equivalent to proteins L11 and L1 from *Escherichia coli* ribosomes. Submitted.

Ramirez, C., Shimmin, L.C., Newton, C.H., Matheson, A.T. and Dennis, P.P. (1989) Structure and evolution of the L11, L1, L10 and L12 equivalent ribosomal proteins in eubacteria, archaebacteria and eucaryotes. *Can. J. Microbiol.* **35:** 234 - 244.

Rauser, W.E. and Bayley, S.T. (1968) Ribosomal complexes from an extremely halophilic bacterium and the role of cations. *J. Bacteriol.* **96:** 1304 - 1313.

Ree, H.K., Cao, K., Thurlow, D.L. and Zimmermann, R.A. (1989) The structure and organization of the 16S ribosomal RNA gene from the archaebacterium *Thermoplasma acidophilum. Can. J. Microbiol.* **35:** 124 - 133.

Reiter, W., Palm, P. and Zillig, W. (1988) Analysis of transcription in the archaebacterium *Sulfolobus* indicates that archaebacterial promotors are homologous to eucaryotic polII promotors. *Nucleic Acids Res.* **16:** 1 - 20.

Remacha, M., Saenz-Robles, M.T., Vilella, M.D and Ballesta, J.P.G. (1988) Independent genes coding for three acidic proteins of the large ribosomal subunit from *Saccharomyces cerevisiae. J. Biol. Chem.* **263:** 9094 - 9101.

Rich, B.E. and Steitz, J.A. (1988) Human acidic ribosomal phosphoproteins P0, P1, and P2: analysis of cDNA clones, *in vitro* synthesis, and assembly. *Mol. Cell. Biol.* **7:** 4065 - 4074.

Rigby, P.W.J., Diekmann, M., Rhodes, 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.

Rosenthal, A., Jung, R. and Hunger, H - D. (1986) Optimized conditions for solid phase sequencing: simultaneous chemical cleavage of a series of long DNA fragments immobilized on CCS anion - exchange paper. *Gene* **42:** 1 - 9.

Rosenthal, A., Schwertner, S., Hahn, V. and Hunger, H - D. (1985) Solid - phase methods for sequencing of nucleic acids. I. Simultaneous sequencing of different oligodeoxyribonucleotides using a mechanically stable anion - exchange paper. *Nucleic Acids Res.* **13:** 1173 - 1184.

Saenger, W. (1987) Structure and dynamics of water surrounding biomolecules. *Ann. Rev. Biophys. Chem.* **16:** 93 - 114.

Said, B., Cole, J.R. and Nomura, M. (1988) Mutational analysis of the L1 binding site of 23S rRNA in *Escherichia coli. Nucleic Acids Res.* **16:** 10529 - 10545.

Sander, G. (1983) Ribosomal protein L1 from *Escherichia coli.* Its role in the binding of tRNA to the ribosome and in elongation factor G - dependent hydrolysis. *J. Biol. Chem.* **258:** 10098 - 10102.

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

Sapienza, C., Rose, M. and Doolittle, W.F. (1982) High-frequency genomic rearrangements involving archaebacterial repeat sequence elements. *Nature* **299:** 182 - 185.

Schidlowski, M. (1988) A 3,800 - million - year isotopic record of life from carbon in sedimentary rocks. *Nature* **333:** 313 - 318.

Schmidt, F.J., Thompson, J., Lee, K., Diik, J. and Cundliffe, E. (1981) The binding site for ribosomal protein L11 within 23S ribosomal RNA of *Escherichia coli. J. Biol. Chem.* **256:** 12301-12305.

Schwartz, R.M. and Dayhoff, M.O. (1978) Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* **199:** 395 - 403.

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

Shimmin, L.C. and Dennis, P.P. (1989) Characterization of the L11, L1, L10 and L12 equivalent ribosomal protein gene cluster of the halophilic archaebacterium *Halobacterium cutirubrum. EMBO Journal* **8:** 1225 - 1235.

Shimmin, L.C., Newton, C.H., Ramirez, C., Yee, J., Downing, W.L., Louie, A., Matheson, A.T. and Dennis, P.P. (1989a) Organization of genes encoding the L11, L1, L10 and L12 equivalent ribosomal proteins in eubacteria, archaebacteria and eucaryotes. *Can. J. Microbiol.* **35:** 164 - 170.

Shimmin, L.C., Ramirez, C., Matheson, A.T. and Dennis, P.P. (1989b) Sequence alignment and evolutionary comparison of the L10 and L12 equivalent ribosomal proteins from archaebacteria, eubacteria and eucaryotes. *J. Mol. Evol.* **29:** 448 - 462.

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

Sor, F. and Nomura, N. (1987) Cloning and DNA sequence determination of the L11 ribosomal protein operon of *Serratia marcescens* and *Proteus vulgaris*. Translational feedback regulation of *Escherichia coli* L11 operon by heterologous L1 proteins. *Mol. Gen. Genet.* **210:** 52 - 59.

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

Spiridonova, V.A., Akhmanova, A.S., Kagramanova, V.K., Kopke, A.K.E. and Mankin, A.S. (1989) Ribosomal protein gene cluster of *Halobacterium halobium* : nucleotide sequence of the genes coding for S3 and L29 equivalent ribosomal proteins. *Can. J. Microbiol.* **35:** 153 - 159.

Stackebrandt, E. (1985) Phylogeny and phylogenetic classification of prokaryotes. *In* Evolution of Prokaryotes. *Edited by* Schleifer, K.H. and Stackebrandt, E. Academic Press, London, pp. 309 - 334.

Stanier, R.Y. (1970) Organization and control in prokaryotic and eukaryotic cells. *Symp. Soc. Gen. Microbiol.* **20:** 1 - 38.

Steitz, J.A. (1978) Methanogenic bacteria. *Nature* **273:** 10.

Stockenius, W. and Bogomolni, R.A. (1982) Bacteriorhodopsin and related pigments of halobacteria. *Ann. Rev. Biochem.* **52:** 587 - 616.

Stockenius, W., Lozier, R.H. and Bogomolni, R.A. (1979) Bacteriorhodopsin and the purple membrane of halobacteria. *Biochim. Biophys. Acta.* **505:** 215 - 298.

Stoffler - Melicke, M., Noah, M. and Stoffler, G. (1983) Location of eight ribosomal proteins on the surface of the 50S subunit from *E. coli. Proc. Natl. Acad. Sci. USA* **80:** 6780 - 6784.

Stoffler, G. and Stoffler - Meilicke, M. (1986) Electron microscopy of archaebacterial ribosomes. *System. Appl. Microbiol.* **7:** 123 - 130.

Strobel, O., Kopke, A.K.E., Kamp, R.M., Bock, A. and Wittmann - Liebold, B. (1988) Primary structure of the archaebacterial *Methanococcus vannielii* ribosomal protein L12. *J. Biol. Chem.* **263:** 6538 - 6546.

Strom, A.R. and Visentin, L.P. (1973) Acidic ribosomal proteins from the extreme halophile, *Halobacterium cutirubrum. FEBS Letters* **37:** 274 - 280.

Strom., A.R., Hasnain, S., Smith, N., Matheson, A.T. and Visentin, L.P. (1975) Ion effects on protein - nucleic acid interactions; the disassembly of the 50S ribosomal subunit from the halophilic bacterium, *Halobacterium cutirubrum. Biochim. Biophys. Acta* **383:** 325 - 337.

Strycharz, W.A., Nomura, A. and Lake, J.A. (1978) Ribosomal proteins L7/L12 localized at a single region of the large subunit by immune electron microscopy. *J. Mol. Biol.* **126:** 123 - 140.

Subramanian, A.R. and Dabbs, E.R. (1980) Functional studies on ribosomes lacking protein L1 from mutant *E. coli. Eur. J. Biochem.* **112:** 425 - 430.

Tabor, S. and Richardson, C.C. (1987) DNA sequence analysis with modified bacteriophage T7 DNA polymerase. *Proc. Natl. Acad. Sci. USA* **84:** 4767 - 4771.

Tate, W.P., Dognin, M.J., Noah, M., Stoffler - Melicke, M. and Stoffler, G. (1984) The N - terminal domain of *Escherichia coli* ribosomal protein L11: its three dimensional location and its role in the binding of release factor 1. *J. Biol. Chem.* **259:** 7317 - 7324.

Taylor, F.J.R. and Coates, D. (1989) The code within the codons. *BioSystems* **22:** 177 - 187.

Terhorst, C., Moller, W., Laursen, R. and Wittman - Liebold, B. (1973) The primary structure of an acidic protein from 50-S ribosomes of *Escherichia coli* which is involved in GTP hydrolysis dependent on elongation factors G and T. *Eur. J. Biochem.* **34:** 138 - 152.

Thomas, M.S. and Nomura, M. (1987) Translational regulation of the L11 ribosomal protein operon of *Escherichia coli* : mutations that define the target site for repression by L1. *Nucleic Acids Res.* **15:** 3085 - 3096.

Thomm, M. and Wich, G. (1988) An archaebacterial promotor element for stable RNA genes with homology to the TATA box of higher eucaryotes. *Nucleic Acids Res.* **16:** 151 - 164.

Thomm, M., Sherf, B.A. and Reeve, J.N. (1988) The RNA polymerase binding and transcription initiation site upstream of the methylreductase operon of *Methanococcus vanielii. J. Bacteriol.* **170:** 1958 - 1961.

Thomm, M., Wich, G., Brown, J.W., Frey, G., Sherf, B.A. and Beckler, G.S. (1989) An archaebacterial promotor sequence assigned by RNA poymerase binding experiments. *Can. J. Microbiol.* **35:** 30 - 35.

Traut, R.R., Tewari, D.S., Sommer, A., Gavino, G.R., Olson, H.M. and Glitz, D.G. (1986) Protein topography of ribosomal functional domains: effects of monoclonal antibodies to different epitopes in *Escherichia coli* protein L7/L12 on ribosome function and structure. *In* Structure, Function and Genetics of Ribosomes. *Edited by* Hardesty, B. and Kramer, G. Springer Verlag, New York, pp. 286 - 308.

Trifonov, E.N. (1987) Translation framing code and frame-monitoring mechanism as suggested by analysis of mRNA and 16S rRNA nucleotide sequences. *J. Mol. Biol.* **194:** 643 - 652.

Tritton, T.R. (1978) Spin - labelled ribosomes. *Biochemistry* **17:** 3959 - 3964.

Van Valen, L.M. and Maiorara, V.C. (1980) The archaebacteria and eukaryotic origins. *Nature* **287:** 248 - 250.

Viera, J. and Messing, J. (1982) The pUC plasmids, an M13mp7 - derived system for insertion mutagenesis and sequencing with the synthetic universal primers. *Gene* **19:** 259 - 268.

Visentin, L.P., Yaguchi, M. and Matheson, A.T. (1979) Structural homologies in alanine - rich acidic ribosomal proteins from procaryotes and eucaryotes. *Can. J. Biochem.* **57:** 719 - 726.

Walter, M.R., Buick, R. and Dunlop, J.S.R. (1980) Stromatolites 3,400 - 3,500 Myr old from the North Pole area, Western Australia. *Nature* **284:** 443 - 445.

Warner, J.R. (1989) Synthesis of ribosomes in *Saccharomyces cerevisiae. Microbiol. Reviews* **53:** 256 - 271.

Wich, G., Hummel, H., Jarsch, M., Bar, U. and Bock, A. (1986) Transcription signals for stable RNA genes in *Methanococcus. Nucleic Acids Res.* **14:** 2459 - 2479.

Wich, G., Leinfelder, W. and Bock, A. (1987) Genes for stable RNA in the extreme thermophile *Thermoproteus tenax* : introns and transcription signals. *EMBO Journal* **6:** 523 - 528.

Wigboldus, J.D. (1987) cDNA and deduced amino acid sequence of *Drosophila* rp21C, another 'A'-type ribosomal protein. *Nucleic Acids Res.* **15:** 10064.

Wittmann-Liebold, B. (1986) Ribosomal proteins: their structure and evolution. *In* Structure, Function and Genetics of Ribosomes. *Edited by* Hardesty, B. and Kramer, G. Springer Verlag, New York, pp. 326 - 361.

Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.* **51:** 221 - 271.

Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74:** 5088 - 5090.

Woese, C.R. and Fox, G.E. (1978) Methanogenic bacteria. *Nature* **273:** 101.

Woese, C.R. and Olsen, G.J. (1986) Archaebacterial phylogeny: perspectives on the urkingdoms. *System. Appl. Microbiol.* **7:** 161 - 177.

Woese, C.R., Pace, N.R. and Olsen, G.J. (1986) Are arguments against archaebacteria valid? *Nature* **320:** 401 - 402.

Wolters, J. and Erdmann, V.A. (1989) The structure and evolution of archaebacterial ribosomal RNAs. *Can. J. Microbiol.* **35:** 43 - 51.

Yaguchi, M., Matheson, A.T., Visentin, L.P. and Zuker, M. (1980) Molecular evolution of the alanine - rich, acidic ribosomal A protein. *In* Genetics and Evolution of RNA Polymerase, tRNA and Ribosomes. *Edited by* Osawa, S., Ozeki, H., Uchida, H. and Yura, Y. University of Tokyo Press, Tokyo, pp. 585 - 599.

Yaguchi, M., Visentin, L.P., Zuker, M., Matheson, A.T., Roy, C. and Strom, A.R. (1982) Amino - terminal sequences of ribosomal proteins from the 30S subunit of the archaebacterium *Halobacterium cutirubrum. Zbl. Bakt. Hyg., I. Abt. Orig.* **C3:** 200 - 208.

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

Yates, J.L. and Nomura, M. (1981) Feedback regulation of ribosomal protein synthesis in *E. coli* : localization of the mRNA target sites for repressor action of ribosomal protein L1. *Cell* **24:** 243 - 249.

Zillig, W. (1986) Archaebacterial status quo is defended. *Nature* **320:** 220.

Zillig, W., Klenk, H., Palm, P., Puhler, G., Gropp, F., Garrett, R.A. and Leffers, H. (1989) The phylogenetic relations of DNA - dependent RNA polymerases of archaebacteria, eukaryotes, and eubacteria. *Can. J. Microbiol.* **35:** 73 - 80.

Zimmerman, R.A., Thurlow, D.L., Finn, R.S., March, T.L. and Ferrett, L.K. (1980) Conservation of specific protein - RNA interactions in ribosome evolution. *In* Genetics and Evolution of RNA Polymerase, tRNA and Ribosomes. *Edited by* Osawa, S., Ozeki, H., Uchida, H. and Yura, Y. University of Tokyo Press, Tokyo, pp. 569 - 584.

Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J. Theor. Biol.* **8:** 357 - 366.