The EMBO Journal vol.8 no.4 pp.1225–1235, 1989

# Characterization of the L11, L1, L10 and L12 equivalent ribosomal protein gene cluster of the halophilic archaebacterium *Halobacterium cutirubrum*

**Lawrence C.Shimmin and Patrick P.Dennis**

Department of Biochemistry, The University of British Columbia, Vancouver, British Columbia, V6T 1W5, Canada

Communicated by W.Zillig

**We have cloned and characterized a 5.2 kb fragment of genomic *Halobacterium cutirubrum* DNA encoding two potential proteins of unknown function (ORF and NAB) and four proteins which are equivalent to the L11, L1, L10 and L12 ribosomal proteins of *Escherichia coli* (L11e, L1e, L10e and L12e). The ribosomal protein genes are clustered in the same order as that in *E.coli* although the transcription pattern differs. Transcripts characterized include (i) abundant monocistronic L11e and tricistronic L1e–L10e–L12e transcripts; (ii) less abundant bicistronic NAB–L11e and monocistronic NAB transcripts and (iii) a very rare ORF monocistronic transcript. The consensus sequence in the promoter region is TTCGA ... 4–10 nucleotides ... TTAA ... 25–26 nucleotides ... initiation site; termination generally occurs on poly(T) tracts following GC-rich regions. Poly(T) tracts in the sense strands within coding regions are notably absent; this is probably related to their participation in transcription termination and to the fact that these ribosomal protein genes are highly expressed and stoichiometrically balanced. In the third position of the codons G or C is utilized 87% of the time. The 74 nt long untranslated leader of the L1e–L10e–L12e transcript contains a region that has a sequence and structure almost identical to a region within the binding domain for the L1e protein in 23S rRNA and highly similar to the *E.coli* L11–L1 mRNA leader sequence that has been implicated in autogenous translational regulation. Other transcripts are initiated at or adjacent to the ATG translation initiation codon.**

*Key words:* transcript mapping/translation/autogenous control

## Introduction

Structurally and functionally conserved throughout evolution, the ribosome is a complex and essential subcellular organelle that utilizes an mRNA template to align and polymerize amino acids into protein. The eubacterial ribosome is comprised of 16S, 23S and 5S rRNAs and ~50 proteins; their eukaryotic counterpart consists of 18S, 5.8S, 28S and 5S rRNAs and ~75 proteins, while in archaebacteria the ribosome is comprised of 16S, 23S and 5S rRNAs and 50–65 proteins.

In the eubacteria the organization, transcription and genetic regulation of the 16S–23S–5S tRNA transcription units and the ribosomal protein genes have been extensively studied (for review see Lindahl and Zengel, 1986). In *Escherichia coli* the genes encoding the ribosomal proteins are all single copy and most are located in clusters of one or more transcription units that often contain additional genes encoding protein elements involved in replication, transcription, translation or other essential cellular processes. 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 rRNA nor free ribosomal proteins accumulate.

In eukaryotic 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 monocistronic transcription units which are rarely clustered within the genome (for review see Planta *et al.*, 1986). 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.

The archaebacteria are believed to represent a third independent line of evolutionary descent (Woese and Fox, 1977). Halobacteria and methanogens appear to contain linked 16S, 23S and 5S rRNAs genes whereas in the sulfur-dependent thermoacidophiles the 5S rRNA gene is separately transcribed and unlinked to the 16S–23S genes (Neumann *et al.*, 1983; Chant and Dennis, 1985; Dennis, 1985; Hui and Dennis, 1985; Mankin and Kagramanova, 1986; Kjems and Garrett, 1987; Kjems *et al.*, 1987). Genes encoding ribosomal proteins appear to be clustered as in the eubacteria but there is little information as yet regarding their transcriptional organization and regulation (Itoh *et al.*, 1988; Auer *et al.*, 1989; Shimmin *et al.*, 1989a; Zillig *et al.*, 1989).

The stalk structure of the 50 S subunit is a distinct structural feature shared by ribosomes from eubacteria, archaebacteria and eukaryotes (Lake, 1983; Oakes *et al.*, 1986; Hanauz *et al.*, 1987; Beauclerk *et al.*, 1985; Shimmin *et al.*, 1989a; Ramirez *et al.*, 1989). In *E.coli* this structure consists of four copies of the L12 protein bound to a single copy of the L10 protein (Strycharz *et al.*, 1978; Petterson and Liljas, 1979). This complex associated with the L11 protein binds to the 23S rRNA (Dijk *et al.*, 1979; Petterson, 1979; Beauclerk *et al.*, 1984). These proteins are part of the GTPase centre on the large subunit and are required for the binding of extrinsic factors (EFTu and EFG) to the ribosome and the concomitant hydrolysis of GTP (for review see Liljas, 1982). The L1 protein is located on the ridge on the opposite side of the ribosome and is involved in the interaction between peptidyl-tRNA and the ribosome at the P site and indirectly with the GTPase centre (Subramanian and Dabbs, 1980; Lake and Strycharz, 1981; Sander, 1983). The genes encoding these four ribosomal proteins are genetically linked with two genes encoding RNA polymerase

©IRL Press

subunit proteins ($\beta$ and $\beta'$) in the order L11−L1−L10−L12−$\beta$−$\beta'$ (Lindahl *et al.*, 1975; Post *et al.*, 1979). The L1 and L10 proteins autogenously regulate the L11−L1 bicistronic and L10−L12 bicistronic transcripts respectively (Johnsen *et al.*, 1982; Nomura *et al.*, 1984; Thomas and Nomura, 1987).

In order to address questions relating to organization and regulation of gene expression, control of synthesis and assembly of ribosomes and evolution of the protein synthesis apparatus in eubacteria, archaebacteria and eukaryotes, we have cloned and characterized the genes encoding the L11e, L1e, L10e and L12e (equivalent to the L11, L1, L10 and

**Table I.** Oligonucleotides

| Probe | Sequence | Length | Sequence complexity | Nucleotide position | Strand identity |
|---|---|---|---|---|---|
| Oligomers | | | | | |
| oLW 9 | GCRTARACRTAYTCCAT | 17 nt | 16 MIX | 4034−4018 (L12e) | (−) |
| oLW 17 | GCNGAAACNATHGARGT | 17 nt | 192 MIX | 1625−1641 (L11e) | (+) |
| Primers | | | | | |
| oLW 36 | ATGTGGGCTTCTGTCGA | 17 nt | UNIQUE | 1165−1181 (ORF) | (−) |
| oLW 52 | CTTCGAGGTCCACCTCGATG | 20 nt | UNIQUE | 1429−1410 (NAB) | (−) |
| oLW 51 | TACGTCGACCGGCGTGGGAC | 20 nt | UNIQUE | 1714−1695 (L11e) | (−) |
| oLW 38 | CGATCTGCGTCTCCTGT | 17 nt | UNIQUE | 2494−2478 (L1e) | (−) |
| oLW 54 | CGTTGTCTGCCATCTTTCAC | 20 nt | UNIQUE | 2326−2307 (L1e) | (−) |

In the oligomer sequence mixtures the letter R represents G and A; Y:C and T; H:A, T and C; and C; N:G, A, T and C. The nucleotide position is that within the sequence of Figure 1. Strand identity indicates identity (+) or complementary (−) with the mRNA.

```
        10        20        30        40        50        60        70        80        90       100       110       120
         |         |         |         |         |         |         |         |         |         |         |         |
ATCGATTCCGGATACACGAACCCCACGGAAAACAGCGGATAGACGCCGTTCGACCAGTCGACGATGACCCACGGCACCACTCCCGCGGCCAGCGCCACCGCCACCACCCGGCGGCGCGCC  120

GTCGGCATCGGCTAGCGCCGCGCAACCACGCGCAGGACGTCCTCGTCTTCGAGGACGTGATCGCGGCCCACCTGCTGGTCGTCGTGTTGGGCCGACGGGCCGGTGACCCGCGCGAACCGG  240
               A  A  A  V  V  A  L  V  D  E  D  E  L  V  H  D  A  G  V  Q  Q  D  D  H  Q  A  S  A  A  T  V  A  A  F  R

AACCGCTCGTCGAGGGTGCCGCCCAGCTTCTGCACGGCGTCGTCCACGGTCTCGCCGCGCCGGATGATGAGCGGCTCCTCGCGGTCGACGCCCCGCCCCGGCTTGTCCATGTAGATCCGG  360
F  R  E  D  L  T  G  G  L  K  Q  V  A  D  D  V  T  E  G  R  A  I  I  L  P  E  E  R  D  V  G  R  G  P  K  D  M  Y  I  R

ATGAGCCCGAGCGCCCGCCACATCCGCTCTTTCAGCACGTCCAATCCCTTCTCCTCGGCGGCCGAAATGAAGATCGCGTCGTCGGGGCTGACGCCGTGATCGCGGAGGGCGTCCTTCATT  480
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

GTCCCCGCGTACGACGGCTCGATGAGGTCGACCTTGTTGACGGTGACCAGCGACGGCATGTAGACGCGGTTGTCCATCACCCCGTCGATCAGCCGGTCGACAGACGGGTTGCCACGGATC  600
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

GTCACGTTGGCGTTGATGAACCCGCGCTCGCGCAGGATTCCCTTGACCGTGTCGCTGTCCAACTCCAGCTCCCCGGACGTGTTCACGTCGATGCCGTCCTTGCCTTTCCGGCGCACGGTC  720
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

ACCGACGGTGGCTCGGCGTCGACCCGGATGTTGACGTTGTACAGCTCCTCGGCGAGACGGTCGTACTGCTCGATCTCGAACGCCGACAGCACGAAGATCACCAGATCCGCCCCACGGATC  840
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

ACCGAGAGGATCTCTTTCCCGCCGCCGCGTCCCCCCGCCGCACCCTCGATGAGCCCCGGCACGTCCAGGAGTTGGATGTTCGCGCCGCGGTACTCCAACATCCCCGGGTTCACGTTGAGG  960
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

GTGGTGAACTCGTAGGCGCCGACCTCGCTGTCGGCGTTGGTCATCGCGTTGATCAGCGACGACTTCCCCACGCTGGGGAATCCAACCAGGGCCACGGTCGCGTCGCCGTGCTGTTCGACT  1080
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

GCGTACCCGCCGCCACCGCCGCTGCCGGACTGCTGGGCTTCGAGCTTCTCCTTTTGCTCCGCGAGCTTCGCCTTCAAGCGGCCGATGTGGGCTTCTGTCGACTTGTTGTACGGCGTGTTC  1200
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

GCGATTTCTTCTTCGAGCGATTCGATGTCCTCCTCGAGCCCCATCACGTGCCCGTAGACAGCAAGCGCCGAAAAGCGCTTTCCTCCGCAGTCCGGTGGGATCGTTTCGACACGTTAATAC  1320
A  I  E  E  E  L  S  E  I  D  E  E  L  G  M
    ORF  370 aa / 1110 bp / MW 40499 dal

                             NAB  68 aa / 204 bp / MW 7530 dal
                             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  1440

 V  S  V  F  A  P  T  D  A  E  I  V  A  I  I  G  S  P  V  V  I  K  E  V  T  E  F  L  T  A  H  G  V  H  M  P
GTCTCAGTGTTCCGGCCGACCGACGCGGAGATCGTCCGCATCATCGGGAGTCCGGTCGTGATCAAGGAGGTCACCGAGTTCCTCACGCGCCACGGCGTCCACATGCCGTGAGCGATTCGA  1560

                                                                L11e  163 aa / 489 bp / MW 17020 dal
                                                                M  A  E  T  I  E  V  L  V  A  G  G  Q  A  D  P  G  P  P  L
TCCGCGGCGGCGCCCCGCTCGAAAGACAAGGGTTAAACCCGCGGCGGCGGTTTCTCGGAGTATGGCTGAGACGATCGAAGTGCTCGTTGCCGGTGGGCAGGCCGACCCTGGCCCGCCCCT  1680

  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  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  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  2040

  A  T  F  N  E  A  V  D  D  G  D  Y  D  D  V  L  G  D  E  L  A  A  A
CCGCACGTTCAACGAGCGCGTCGACGACGGCGACTACGACGACGTGCTCGGCGACGAACTCGCGGCCGCGTAACGCCGCCCGAGGAGTTTCTGCGCCGTTCGGTTCGCGTACTCGATAGC  2160

GGCGTGTGTCCGCGGGTCGCGCTCCCACGCTTGCTTCGCTTCGACGCTTTTAAGCCCGGGATCACCGTCTGTAGAACCGAGACAGGCTTCGCCTGTTTCACTGACCCGTAGGAGATCCGA  2280

                                    L1e  212 aa / 636 bp / MW 23095 dal
                                    M  A  D  N  D  I  E  E  A  V  A  R  A  L  E  D  A  P  Q  R  N  F  A  E  T  V  D  L  A
CCTTCAGAGGGTCACCCACTACGGAGGTGAAAGATGGCAGACAACGATATAGAAGAGGCCGTAGCTCGCGCACTTGAGGATGCCCCACAGCGGAACTTCCGTGAGACGGTAGACCTCGCA  2400

 V  N· L  A  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  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  2640
```

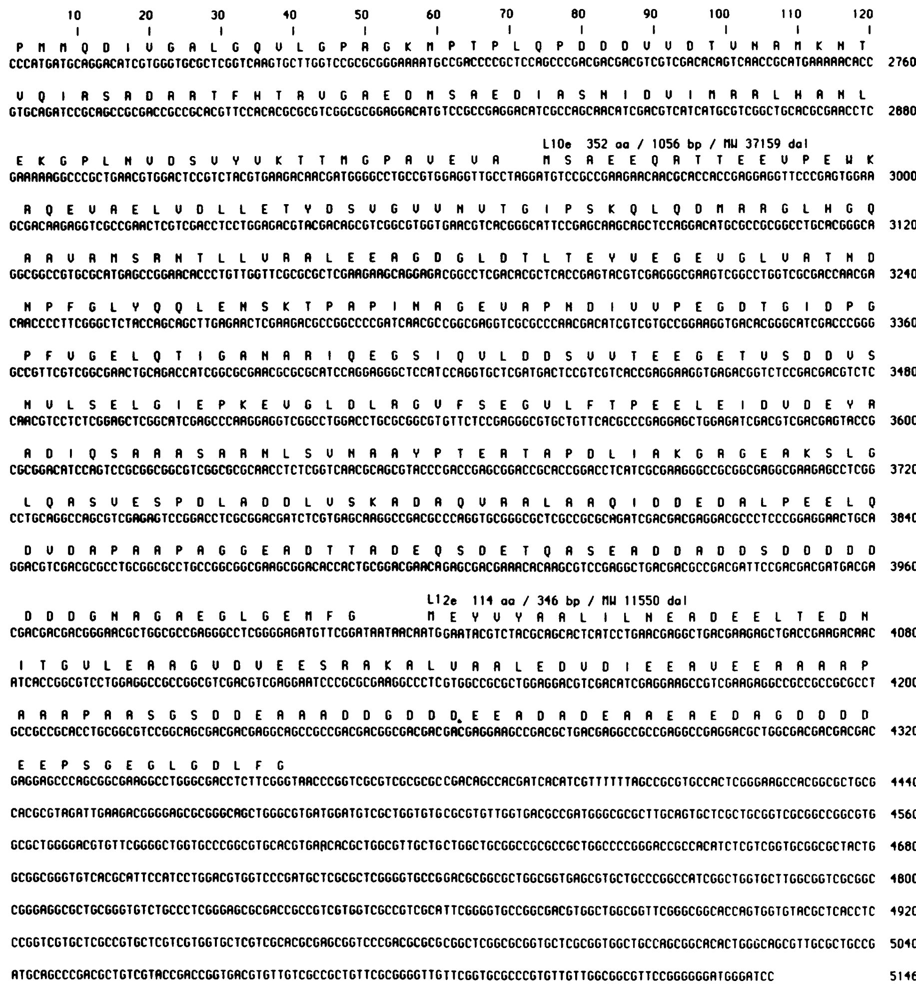


**Fig. 1.** Nucleotide sequence of the 5.2 kb fragment of *H.cutirubrum* genomic DNA. The complete 5′−3′ nucleotide sequence of the 5.2 kb *Cla*I−*Bam*HI fragment of genomic DNA is indicated. The deduced amino acid sequence of the leftward oriented ORF gene is written below the DNA (−) strand sequence (positions 1244−135). The deduced amino acid sequences of the rightward oriented NAB (1345−1548), L11e (1622−2110), L1e (2314−2949), L10e (2954−4009) and L12e (4018−4359) genes are written above the DNA (+) strand sequence. The number of amino acid residues, nucleotide base pairs and mol. wt for each protein are indicated at the initiation methionine positions.

L12 genes of *E.coli*) from two divergent archaebacteria, *Halobacterium cutirubrum* NRCC 34001 and *Sulfolobus solfataricus* P1, and the L10e and four different L12 e genes from the eukaryote *Saccharomyces cerevisiae* (Ramirez *et al.*, 1989; Shimmin *et al.*, 1989a; C.H.Newton, L.C. Shimmin, J.Yee and P.P.Dennis, manuscript submitted). In this paper we report the sequence, transcription and potential regulation of an *H.cutirubrum* genomic clone encoding the L11e, L1e, L10e and L12e proteins and two other proteins of unknown function.

## Results and discussion

The *H.cutirubrum* ribosomal protein L20 is acidic, alanine-rich, present in four copies per ribosome and has previously been shown to be the structural and functional homologue to *E.coli* L12 (Yaguchi *et al.*, 1980).The *H.cutirubrum* L11 protein is associated with the ribosomal 'A' protein complex and similarity between the N-terminal amino acid sequences of this protein and that of *E.coli* L11 suggested that this is the halobacterial L11 homologue (Matheson *et al.*, 1984). In this paper we refer to the *H.cutirubrum* L20 and L11 proteins as Hcu L12e (i.e. L12 equivalent) and Hcu L11e respectively.

### Isolation of genomic clones

The amino terminal sequences of ribosomal protein L12e and L11e of *H.cutirubrum* are MEYVYA and AETIEV respectively. Two 17mer synthetic oligonucleotide mixtures complementary to all DNA sequences encoding these hexapeptides were prepared and used to probe restriction enzyme digests of *H.cutirubrum* genomic DNA (Table I lists the synthetic oligonucleotides used in this work). The L12e specific oligonucleotide mix (oLW 9) hybridized to a 1.3 kb *Pst*I−*Bam*HI fragment. Size fractionated genomic DNA was cloned into the plasmid pUC8, transformed into *E.coli* JM83; transformants were screened using the oligonucleotide mix as probe. The L11e specific oligonucleotide mix (oLW 17) hybridized to 5.2 kb *Cla*I−*Bam*HI fragment. Using the oligonucleotide mix as probe, genomic DNA was fractionated, cloned between the *Cla*I and *Bam*HI sites of the plasmid pBR322 and eventually transformed and propagated in *E.coli* JC8111 (Boissy and Astell, 1985). Restriction enzyme and Southern hybridization analysis demonstrated that the smaller 1.3 kb fragment was derived

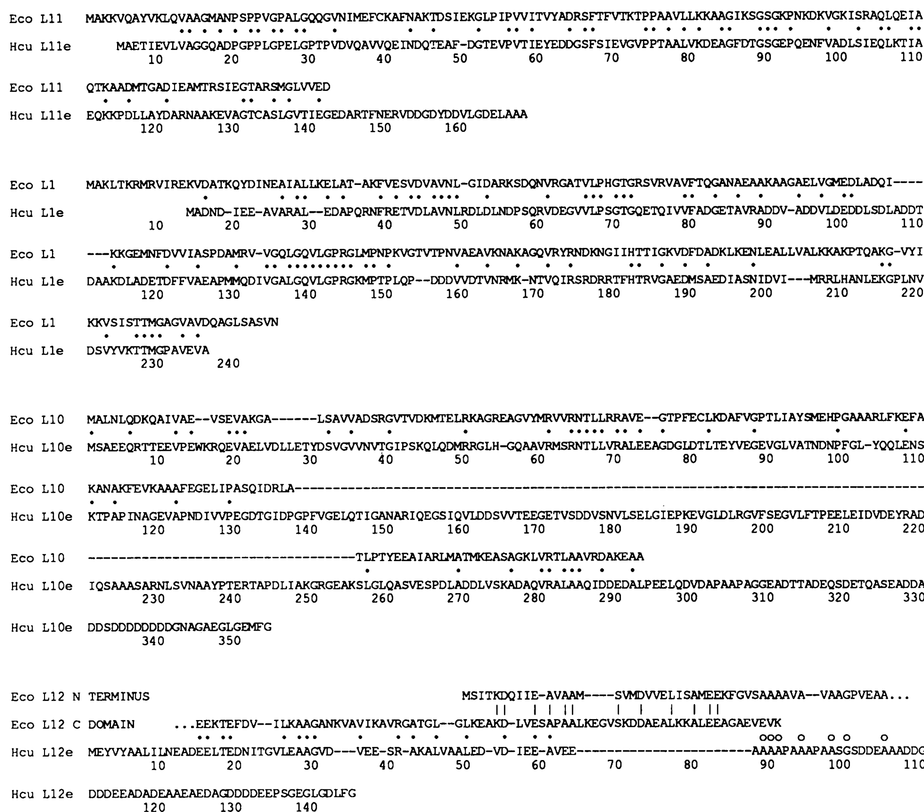


**Fig. 2.** Alignment of the L11e, L1e, L10e and L12e ribosomal proteins. The *H.cutirubrum* L11e, L1e, L10e and L12e ribosomal proteins are aligned with the *E.coli* L11, L1, L10 and L12 ribosomal proteins respectively. Identities are indicated by solid circles (●) and gaps (−) have been inserted where necessary to maintain alignment. The Eco L12 protein has undergone rearrangement and thus identities are indicated between the Hcu L12e and Eco L12 C domain (●), the Eco L12 N terminus and Eco L12 C domain (|), and the Hcu L12e and Eco L12 N terminus (○).

from the right-hand end of the longer 5.2 kb fragment. Clones of the larger fragment were initially difficult to recover because the fragment is unstable in high copy number plasmids and it is efficiently propagated only in the *E.coli* strain JC8111 *recBC sbcB recF*.

### *Nucleotide sequence analysis*

The nucleotide sequence of the entire 5.2 kb fragment of *H.cutirubrum* genomic DNA was determined and is illustrated in Figure 1. The fragment contained unique sequences complementary to the L11e and L12e oligonucleotide probe mixtures (positions 1625−1641 and positions 4018−4034). Further analysis indicated that these two sequences were within the genes encoding the Hcu L11e and Hcu L12e ribosomal proteins. In addition to encoding the L11e and L12e proteins, two other open reading frames encoding proteins homologous to the L1 and L10 proteins of *E.coli* (Hcu L1e and Hcu L10e) were identified by comparison of the predicted amino acid sequence with the sequence of the *E.coli* proteins (Figure 2). Significance of homology was assessed using the jumbling program RDF wherein matches of 10 standard deviations ($z$) from the mean are considered certain homologies (Lipman and Pearson, 1985). The Eco L11 and Hcu L11e proteins have a linear correspondence yielding 33% identities over 138 residues ($z$ = 35) requiring only a single one amino acid gap to maintain alignment. The Eco L1 and Hcu L1e proteins are 32% identical over 213 residues ($z$ = 36) but require 10 gaps in the alignment. The Eco L10 protein has undergone a large internal deletion and c-terminal truncation (23% identical over 169 residues, $z$ = 10) and the Eco L12 protein has been rearranged with respect to its *H.cutirubrum* homologues. The evolution and structure/function of these genes and proteins are discussed in detail elsewhere (Ramirez *et al.*, 1989; Shimmin *et al.*, 1989a; Shimmin *et al.*, 1989b). Regions within the sequences of all four *H.cutirubrum* ribosomal proteins have been confirmed by partial or complete direct protein sequence data (Shimmin *et al.*, 1989a).

The four *H.cutirubrum* ribosomal protein genes on the 5.2 kb genomic fragment are oriented left to right in the order L11e, L1e, L10e and L12e; this is identical to the order of the corresponding genes on the chromosome of *E.coli* (Figure 3). The *H.cutirubrum* intergenic spacing between L11e and L1e, L1e and L10e, and L10e and L12e is 203, 4 and 8 nucleotides (nt) respectively and compare to spacing of 6, 415 and 69 nt for the corresponding intergenic regions of *E.coli*. The 784 nt distal to the L12e gene is devoid of coding potential. In *E.coli* the $\beta$ and $\beta'$ RNA polymerase subunit genes are located distal to the L12 gene; the L12−$\beta$ intergenic space is 323 nt in length and contains a transcription attenuation and an RNase III processing sequence. Recently the genes encoding the large subunits of RNA polymerase have been cloned from *Halobacterium halobium* and found to be located in front of the S12e−S7e−EFGe−EFTe gene cluster (Zillig *et al.*, 1989).

Analysis of the 1621 nt in front of the L11e gene on the 5.2 kb fragment revealed two potential coding regions designated ORF and NAB. The ORF potentially encodes a protein of 370 amino acids and is oriented leftwards on the

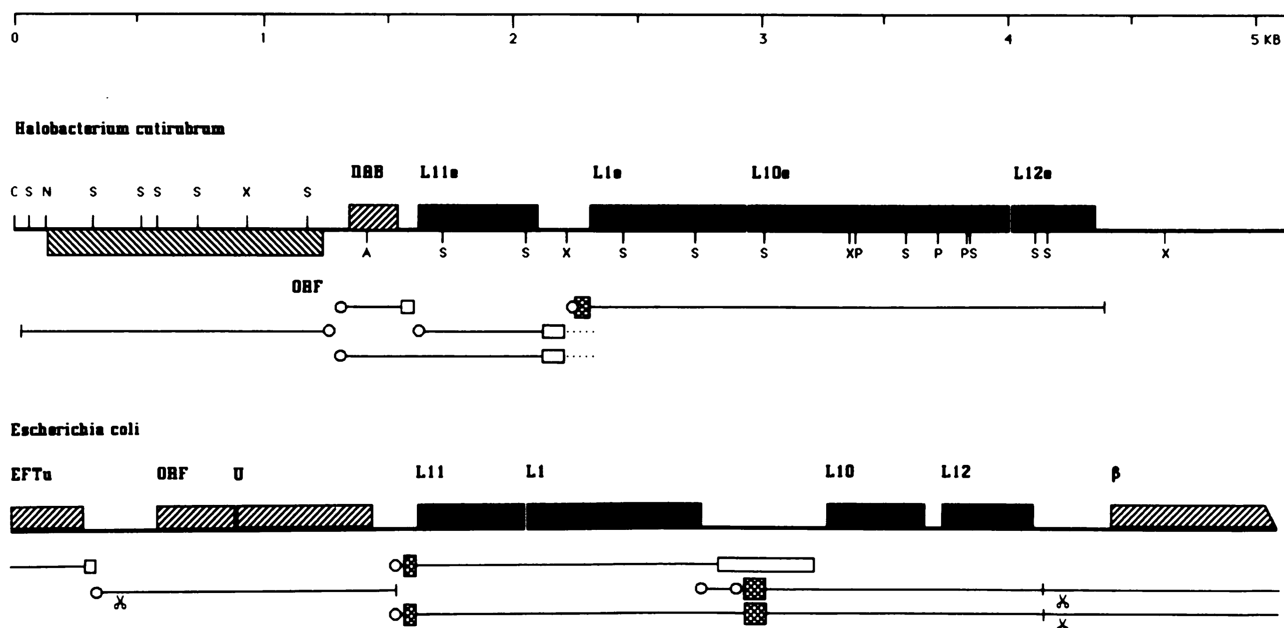



**Fig. 3.** Genomic organization of the L11e, L1e, L10e and L12e ribosomal protein genes. The organization and transcription of the L11e, L1e, L10e and L12e ribosomal protein gene clusters of *H.cutirubrum* (above) and *E.coli* (below) (from Downing and Dennis, 1987) are depicted. Ribosomal protein encoding genes are solid boxes and other proteins encoding genes or open ending frames 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 5.2 kb *H.cutirubrum* fragment are: *Cla*I, C; *Nhe*I, N; *Sal*I, S; *Ava*II, A; *Xma*I (*Sma*I), X; *Pst*I, P; *Bam*HI, B. Transcripts of the *H.cutirubrum* genes are indicated. The open circles (○) represent 5′ transcript ends and the vertical lines (|) represent 3′ transcript ends. The open boxes at the ends of transcripts indicate regions of multiple 3′ transcript ends. In the *H.cutirubrum* diagram the dotted continuation of the monocistronic L11e and bicistronic NAB−L11e transcripts indicate that a very small amount of readthrough may occur. In the *E.coli* diagram only the 3′ end of the EFTu and the 5′ end of the RNA polymerase β subunit genes are indicated. Scissored interruptions represent RNase III 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 *E.coli* and putative autogenous regulation in *H.cutirubrum*.

genomic fragment. This potential protein shows no similarity to any known protein sequence. The NAB potentially encodes a short 68 amino acid protein the exhibits sequence similarity to restriction endonucleases *Eco*RI and *Pst*I. The NAB is oriented rightwards and terminates at position 1548, 73 nt in front of the L11e ATG initiation codon. The NAB sequence appears unrelated to either the ORF or U genes which are located in the corresponding region on the chromosome of *E.coli*. In *E.coli*, the proteins encoded by ORF and U function in protein export and transcription termination respectively.

### *Analysis of transcription products*

The *in vivo* transcripts produced from the 5.2 kb fragment of genomic DNA were detected and analyzed by Northern hybridization, nuclease protection analysis and primer extension analysis. Total RNA was isolated from exponentially growing cells and used in these procedures. The results of the primer extension and S1 nuclease analyses are illustrated in Figure 4 and summarized in Figure 3.

By Northern hybridization the very rare leftwards transcript of the ORF gene was identified and estimated to be ~1200 nt in length (data not shown). Priming with oLW 36 on total RNA as template resulted in a product with a 5′ end site at position 1245, one nucleotide in front of the putative ATG initiation codon (Figure 4A). The 3′ transcript end site was identified by S1 nuclease protection of a 906 bp *Pst*I−*Nhe*I fragment (position 3609 of pBR322 to position 131 within the insert DNA) 3′ end labeled at the *Nhe*I site. Termination primarily occurs within the TTTT sequence at positions 32−29 (Figure 4B).

Four different rightwards transcripts are detectable by Northern hybridization and represent (i) a 250 nt long monocistronic NAB transcript, (ii) an 850 nt long bicistronic NAB−L11e transcript, (iii) a 600 nt long L11e monocistronic transcript and (iv) a 2150 nt long tricistronic L1e−L10e−L12e transcript (data not shown).

By priming with oLW 51 (positions 1714−1695) and oLW 52 (positions 1429−1410) on total RNA, an abundant and a less abundant 5′ end sites were detected at position 1622, corresponding to the A residue of the ATG translation initiation codon of the L11e gene and position 1344, immediately in front of the NAB ATG translation initiation codon (Figure 4D, E and G). This result was confirmed by S1 nuclease protection of the 5′ end labeled *Sal*I fragment (positions 1177−1706) by total RNA (Figure 4C). Three specific protection products were observed. The first was a very rare product ~68 nt 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 ~89 and 367 nt in length and correspond to the 5′ transcript ends at position 1622 and 1344 respectively and protecting label at position 1710. [A fourth band of ~208 nt in length is the result of contamination of the 5′ end labeled probe with the 383 bp *Sal*I fragment (positions 2060−2443) and protection by the 5′ end of the L1e−L10e−L12e RNA transcript; see below.]

The position of 3′ transcript end sites within and the extent of transcription through the NAB−L11e intergenic space was assessed by S1 nuclease protection of a 263 bp *Ava*II fragment (positions 1419−1682) 3′ end labeled at position 1421. Two protection products were observed and

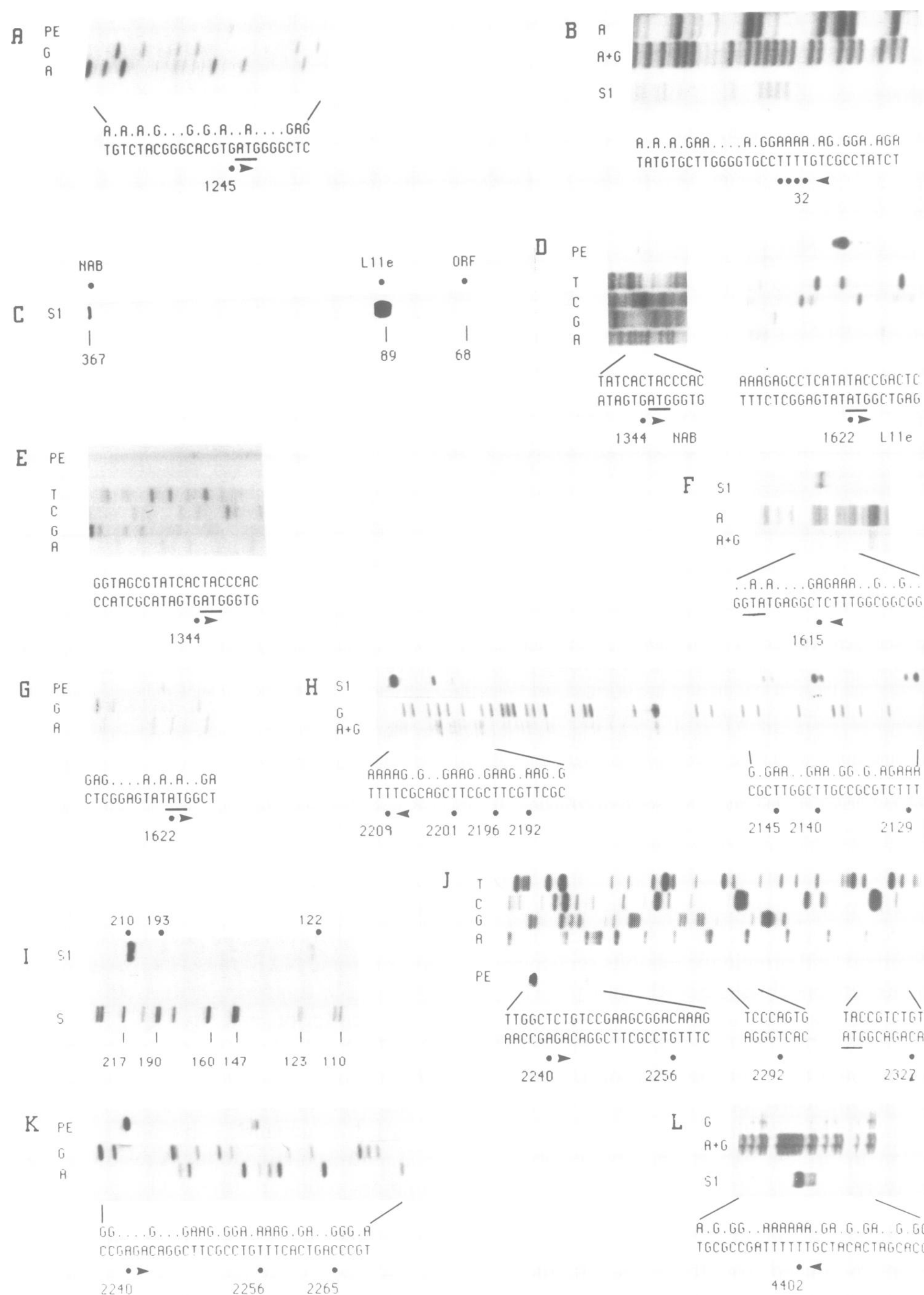


**Fig. 4.** Nuclease protection and primer extension analysis of RNA transcript ends. The products of primer extension and nuclease protection analysis are illustrated alongside sequnce ladders. Where appropriate, the (+) strand DNA sequence is written below the ladder and the nucleotide position of the major 5′ and 3′ transcript ends in the sequence presented in Figure 1 are indicated. Arrows indicate the direction of transcription. The positions of the initiator methionines within the (+) strand are underlined. In D, H and K composites of different exposures of single experiments are illustrated for clarity. (**A**) 5′ end of the ORF transcript detected by primer extension using oLW 36. (**B**) 3′ end of the ORF transcript detected by S1 nuclease protection of an *Pst*I–*Nhe*I fragment (position 3613 of pBR322 to position 131 of the insert DNA) 3′ end labeled at position 132. (**C**) low resolution of the 5′ ends of the ORF, NAB–L11e bicistronic and L11e monocistronic transcripts detected by S1 nuclease protection of a 529 bp *Sal*I fragment (position 1177–1706) 3′ end labeled at positions 1178 and 1706. The very rare transcript from the ORF gene is visible on the original autoradiogram but not on the photographic reproduction. The probe is contaminated by a second *Sal*I fragment (positions 2060–2443) that is also 3′ end labeled at position 2443. This minor component is protected by the 5′ leader at position 2240. Size standards were *Msp*I restricted and 3′ end labeled pBR322. (**D**) The 5′ ends of NAB–L11e bicistronic and L11e monocistronic transcripts detected by primer extension with oLW 51. (**E**) 5′ end of the NAB transcript detected by primer extension with oLW 52. (**F**) 3′ end of the NAB monocistronic transcript detected by S1 nuclease protection of an *Ava*II fragment (positions 1419–1682) 3′ end labeled at position 1421. (**G**) 5′ end of the L11e transcript detected by primer extension with oLW 51. (**H**) 3′ ends of the L11e transcripts detected by S1 nuclease protection of an *Sal*I fragment (2060–2443) labeled at position 2064. The lengths of some of the *Msp*I restricted and 3′ end labeled pBR322 size standards are indicated. (**I**) 5′ ends of the L1e–L10e–L12e transcript detected by S1 nuclease protection of a *Sal*I fragment (positions 2060–2443) 5′ end labeled at position 2442. (**J**) 5′ ends of the L1e–L10e–L12e transcript detected by primer extension with oLW 38. (**K**) 5′ ends of the L1e–L10e–L12e transcript detected by primer extension with oLW 54. (**L**) 3′ end of the L1e–L10e–L12e transcripts detected by S1 nuclease protection of a *Sal*I–*Xma*I fragment (4159–4644) 3′ end labeled at position 4163.

correspond to (i) full protection by read-through transcripts and (ii) partial protection by transcripts with 3′ end sites near position 1615 (Figure 4F).

The 3′ transcript end sites in the L11e–L1e intergenic space were identified by S1 nuclease protection of the 3′ end labeled 383 bp *Sal*I fragment (positions 2060–2443). The probe fragment was labelled at position 2064 and seven different sites of protection products were observed corresponding to 3′ end sites near positions 2129, 2140, 2145, 2192, 2296, 2201 and 2209 (Figure 4H). Each of these sites are within T tracts; there was little or no full length protection of the DNA probe.

To summarize, analysis of band intensities in these experiments suggests that the L11e gene is encoded on an abundant monocistronic transcript. Also detected were a 5-fold less abundant bicistronic NAB–L11e and a 20-fold less abundant monocistronic NAB transcript. There is a gap of 6 nt between the 3′ end of the NAB and the 5′ end of the L11e monocistronic transcripts.

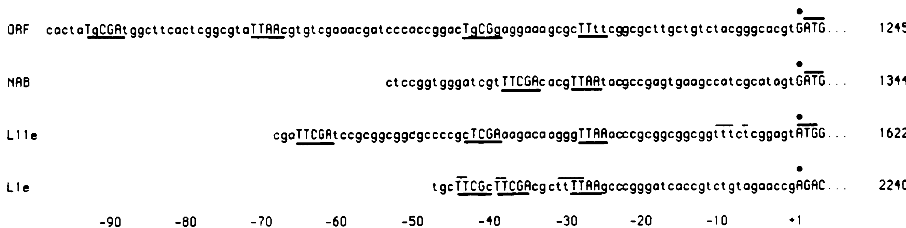
The 5′ end sites of the tricistronic L1e–L10e–L12e transcript detected by Northern hybridization was analyzed by priming with oLW 54 (positions 2326–2307) and oLW 38 (positions 2494–2478; Figure 4J and K). The major transcripts had a 5′ end at position 2240, 74 nt in front of the L1e ATG translation initiation codon and ~30 nt beyond the last termination site for transcripts exiting the L11e gene. A number of other less abundant 5′ ends located between positions 2240 and 2327 were also apparent both in primer extension and S1 nuclease protection experiments and probably represent intermediates in the degradation of the tricistronic mRNA; the shortest of these at position 2327 is just within the coding region of L1e (Figure 4I). No other major 5′ ends were detected between nucleotides 2321 and 4360, the region encoding the L1e, L10e and L12e ribosomal proteins.

The 3′ end sites of the tricistronic L1e–L10e–L12e transcript was mapped by S1 nuclease protection by total RNA of the SalI–SmaI fragment (positions 4159–4644) 3′ end labeled at position 4163. The 3′ transcript end was located near nucleotide 4402 within a run of six T residues (Figure 4L). Virtually all transcripts exiting the L12e gene terminate in this region. Attempts to identify transcripts from either strand of the DNA beyond position 4163 were negative implying that this region probably represents a transcriptionally inactive space.

### Consensus regulatory sequences

Sequences surrounding the 5′ and 3′ transcript end sites are summarized in Figure 5. The conserved sequences that appear to constitute a part of the *H.cutirubrum* transcriptional promoter are TTCGA ... TTAA. The spacing between these elements is 4–10 nt and the distance to the transcription start site is ~25 nt. It is interesting to note that the very weak ORF leftwards promoter exhibits the least conservation to the consensus at the appropriate position but contains a better sequence somewhat further upstream at position −69. These putative promoter elements with appropriate spacing are also present in each of the five tandem promoters of the rRNA transcription unit of *H.cutirubrum* and other archaebacterial promoters (Wich *et al.*, 1986; Brown *et al.*, 1988; Rieter *et al.*, 1988; Thomm and Wich, 1988).

The position of 3′ transcript end sites are located uniformly within (or in the case of the NAB monocistronic transcript immediately after) runs of T residues and are often preceded by GC-rich sequences. Longer T runs appear to result in more efficient termination. Runs of T within coding regions



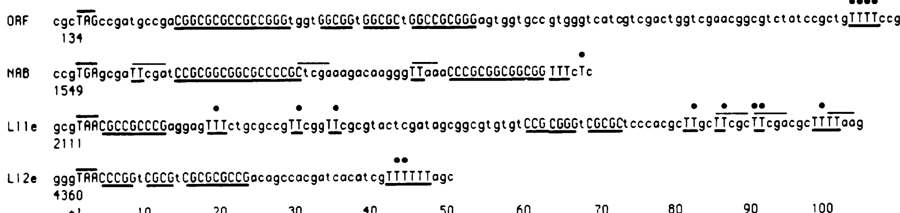




**Fig. 5.** Alignment of sequences at putative transcription initiation and termination sites. Sequences upstream of putative transcription initiation sites are aligned (top) with 5′ end sites at position +1. The position of the 5′ end sites in the sequence presented in Figure 1 are indicated on the right. The ATG translation initiation codons adjacent to the 5′ transcript end sites are heavily overlined. Sequences resembling the concensus TTCGA---TTAA are underlined with the conserved bases highlighted. Where the terminator of the upstream gene overlaps with the promoter (L11e and L1e) the termination site(s) are lightly overlined. Sequences upstream from putative transcription termination sites (bottom) 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 1 is indicated on the left. The GC-rich runs and poly(T) tracts are underlined and highlighted and the most prominent 3′ end site in each T run is indicated. Where the promoter of the downstream gene overlaps with the terminator (NAB and L11e) the promoter sequences are lightly overlined.

are statistically much less frequent than expected (see below). 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.

The regions surrounding the translation initiation sites on mRNAs derived from the 5.2 kb fragment of genomic DNA are depicted in Figure 6. All of these regions exhibit a potential Shine–Dalgarno sequence that is complementary to the 3′ end of the 16S rRNA. The position of the complementary sequence is 3′ to the AUG initiation codon on the ORF, NAB and L11e monocistronic transcripts that lack a 5′ untranslated leader and is 5′ to the AUG initiation codons on the NAB–L11e and L1e–L10e–L12e polycistronic transcripts. The spacing between the last base of the complementary sequence and the first base of the initiation codons varies from −15 for ORF 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. Finally, it should be noted that the L12e cistron is translated ~4-fold more frequently than the preceding L1e or L10e cistrons. It is not yet clear how this translational enhancement is achieved.

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

ORF                                     gAUGgggcucgAGGAGGac...      1244
NAB                                     gAUGGGUGAcCCugcugcg...      1345
L11e                                     AUGgcuGAGacGAUCgaa...      1622
NAB/L11e                  ...uuucucGGAGuAUGgcuGAGacGAUCgaa...      1622
L1e                  ...acuacGGAGGUgaaagAUGgcagacaacgauaua...      2314
L10e              ...gccguGGAGGUugccuaggAUGuccgccgaagaacaa...      2954
L12e     ...cucggGGAGaUGuUCggauaauaacaAUGgaauacgucuacgca...      4018

                  -20        -10         +1        10
```

**Fig. 6.** Translation initiation regions. Sequences surrounding the AUG translation initiation regions are presented. The initiation codon and where present upstream termination codons are indicated by heavy and light underline respectively. The first base in the initiation codons are aligned at +1 and their position in the sequence presented in Figure 1 is indicated on the left. Sequences complementary to the 3′ end of the 16S rRNA are highlighted. 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.

### *Codon utilization and amino acid composition*

The codon utilization in the six *H.cutirubrum* open reading frames representing 3837 nt or 75% of the 5146 bp fragment is summarized in Table II. Three points are apparent. First, there is a strong preference to avoid codons with adjacent T residues and the TTT codon for phenylalanine is never utilized. Similarly, codons with T at the third position only rarely precede codons with T at the first position. There is only a single TTT and a single TTTT within the coding regions at positions 427–425 and 2498–2501 respectively, and no T runs longer than four. In both non-coding regions and on the (−) strand of coding regions T runs are much more prevalent. The avoidance of T runs on the (+) strand within coding regions is probably related to their participation in transcription termination. Premature transcription termination at the TTT and TTTT sequences within the ORF and L1e genes respectively would appear to be inconsequential due to the low level of expression of the ORF and the polar effect of termination in the first cistron of the L1e–L10e–L12e operon. Second, there is a strong preference for G or C rather than A or T at the third position of the codon. This occurs 87% of the time and reflects the high G + C context of *H.cutirubrum* genomic DNA (68%). Third, arginine is entirely encoded by the CGN codons and never by the AGR codons (68:0). A similar bias exists for arginine codons in *E.coli* but not in *S.cerevisiae* ribosomal protein genes (Post *et al.*, 1979; Sharp *et al.*, 1988).

The amino acid compositions of the six *H.cutirubrum* proteins encoded by the 5.2 kb fragment are listed in Table

**Table II.** Codon utilization of genes encoded by pLW173

| Codon | AA | No. | % | Codon | AA | No. | % | Codon | AA | No. | % | Codon | AA | No. | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| UUU | Phe | 0 | (0.0%) | UCU | Ser | 1 | (0.1%) | UAU | Tyr | 0 | (0.0%) | UGU | Cys | 1 | (0.1%) |
| UUC | | 27 | (2.1%) | UCC | | 19 | (1.5%) | UAC | | 21 | (1.6%) | UGC | | 0 | (0.0%) |
| UUA | Leu | 0 | (0.0%) | UCA | | 1 | (0.1%) | UAA | Och | 3 | | UGA | Opa | 1 | |
| UUG | | 5 | (0.4%) | UCG | | 18 | (1.4%) | UAG | Amb | 2 | | UGG | Trp | 2 | (0.2%) |
| CUU | Leu | 3 | (0.2% | CCU | Pro | 7 | (0.5%) | CAU | His | 0 | (0.0%) | CGU | Arg | 5 | (0.4%) |
| CUC | | 58 | (4.5%) | CCC | | 15 | (1.2%) | CAC | | 10 | (0.8%) | CGC | | 40 | (3.1%) |
| CUA | | 0 | (0.0%) | CCA | | 3 | (0.2%) | CAA | Gln | 8 | (0.6%) | CGA | | 2 | (0.2%) |
| CUG | | 36 | (2.8%) | CCG | | 33 | (2.6%) | CAG | | 35 | (2.7%) | CGG | | 21 | (1.6%) |
| AUU | Ile | 2 | (0.2%) | ACU | Thr | 1 | (0.1%) | AAU | Asn | 0 | (0.0%) | AGU | Ser | 2 | (0.2%) |
| AUC | | 55 | (4.3%) | ACC | | 32 | (2.5%) | AAC | | 43 | (3.4%) | AGC | | 18 | (1.4%) |
| AUA | | 1 | (0.1%) | ACA | | 5 | (0.4%) | AAA | Lys | 10 | (0.8%) | AGA | Arg | 0 | (0.0%) |
| AUG | Met | 25 | (2.0%) | ACG | | 27 | (2.1%) | AAG | | 24 | (1.9%) | AGG | | 0 | (0.0%) |
| GUU | Val | 9 | (0.7%) | GCU | Ala | 10 | (0.8%) | GAU | Asp | 10 | (0.8%) | GGU | Gly | 12 | (0.9%) |
| GUC | | 67 | (5.2%) | GCC | | 58 | (4.5%) | GAC | | 138 | (10.8%) | GGC | | 63 | (4.9%) |
| GUA | | 3 | (0.2%) | GCA | | 15 | (1.2%) | GAA | Glu | 42 | (3.3%) | GGC | | 8 | (0.6%) |
| GUG | | 43 | (3.5%) | GCG | | 61 | (4.8%) | GAG | | 95 | (7.4%) | GGG | | 27 | (2.1%) |

Shown is the number and percentage of the 1279 codons utilized by the ORF, NAB, L11e, L1e, L10e and L12e genes combined.

III. 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 aid in preserving their

**Table III.** Amino acid compositions of proteins encoded by pLW173

| | | L11e 163 AA Mol. wt 17020 | L1e 212 AA Mol. wt 23095 | L10e 352 AA Mol. wt 37159 | L12e 114 AA Mol. wt 11550 | NAB 68 AA Mol. wt 7530 | ORF 370 AA Mol. wt 40499 |
|---|---|---|---|---|---|---|---|
| Ala | A | 20 (12.3%) | 23 (10.8%) | 41 (11.6%) | 28 (24.6%) | 4 (5.9%) | 28 (7.6%) |
| Arg | R | 3 (1.8%) | 15 (7.1%) | 14 (4.0%) | 1 (0.9%) | 4 (5.9%) | 31 (8.4%) |
| Asn | N | 4 (2.5%) | 9 (4.2%) | 12 (3.4%) | 2 (1.8%) | 0 (0.0%) | 16 (4.3%) |
| Asp | D | 18 (11.0%) | 33 (15.6%) | 42 (11.9%) | 20 (17.5%) | 5 (7.4%) | 30 (8.1%) |
| Cys | C | 1 (0.6%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| Gln | Q | 7 (4.3%) | 8 (3.8%) | 17 (4.8%) | 0 (0.0%) | 1 (1.5%) | 10 (2.7%) |
| Glu | E | 19 (11.7%) | 14 (6.6%) | 41 (11.6%) | 22 (19.3%) | 7 (10.3%) | 34 (9.2%) |
| Gly | G | 16 (9.8%) | 11 (5.2%) | 31 (8.8%) | 9 (7.9%) | 4 (5.9%) | 39 (10.5%) |
| His | H | 0 (0.0%) | 2 (0.9%) | 1 (0.3%) | 0 (0.0%) | 2 (2.9%) | 5 (1.4%) |
| ILe | I | 7 (4.3%) | 7 (3.3%) | 12 (3.4%) | 3 (2.6%) | 6 (8.8%) | 23 (6.2%) |
| Leu | L | 11 (6.7%) | 16 (7.5%) | 31 (8.8%) | 8 (7.0%) | 4 (5.9%) | 32 (8.6%) |
| Lys | K | 5 (3.1%) | 5 (2.4%) | 7 (2.0%) | 1 (0.9%) | 1 (1.5%) | 15 (4.1%) |
| Met | M | 1 (0.6%) | 8 (3.8%) | 4 (1.1%) | 1 (0.9%) | 3 (4.4%) | 8 (2.2%) |
| Phe | F | 5 (3.1%) | 5 (2.4%) | 5 (1.4%) | 1 (0.9%) | 3 (4.4%) | 8 (2.2%) |
| Pro | P | 11 (6.7%) | 10 (4.7%) | 17 (4.8%) | 3 (2.6%) | 5 (7.4%) | 12 (3.2%) |
| Ser | S | 5 (3.1%) | 7 (3.3%) | 21 (6.0%) | 4 (3.5%) | 3 (4.4%) | 19 (5.1%) |
| Thr | T | 11 (6.7%) | 13 (6.1%) | 19 (5.4%) | 2 (1.8%) | 5 (7.4%) | 15 (4.1%) |
| Trp | W | 0 (0.0%) | 0 (0.0%) | 1 (0.3%) | 0 (0.0%) | 0 (0.0%) | 1 (0.3%) |
| Tyr | Y | 3 (1.8%) | 1 (0.5%) | 5 (1.4%) | 2 (1.8%) | 1 (1.5%) | 9 (2.4%) |
| Val | V | 16 (9.8%) | 25 (11.8%) | 31 (8.8%) | 7 (6.1%) | 10 (14.7%) | 35 (9.5%) |

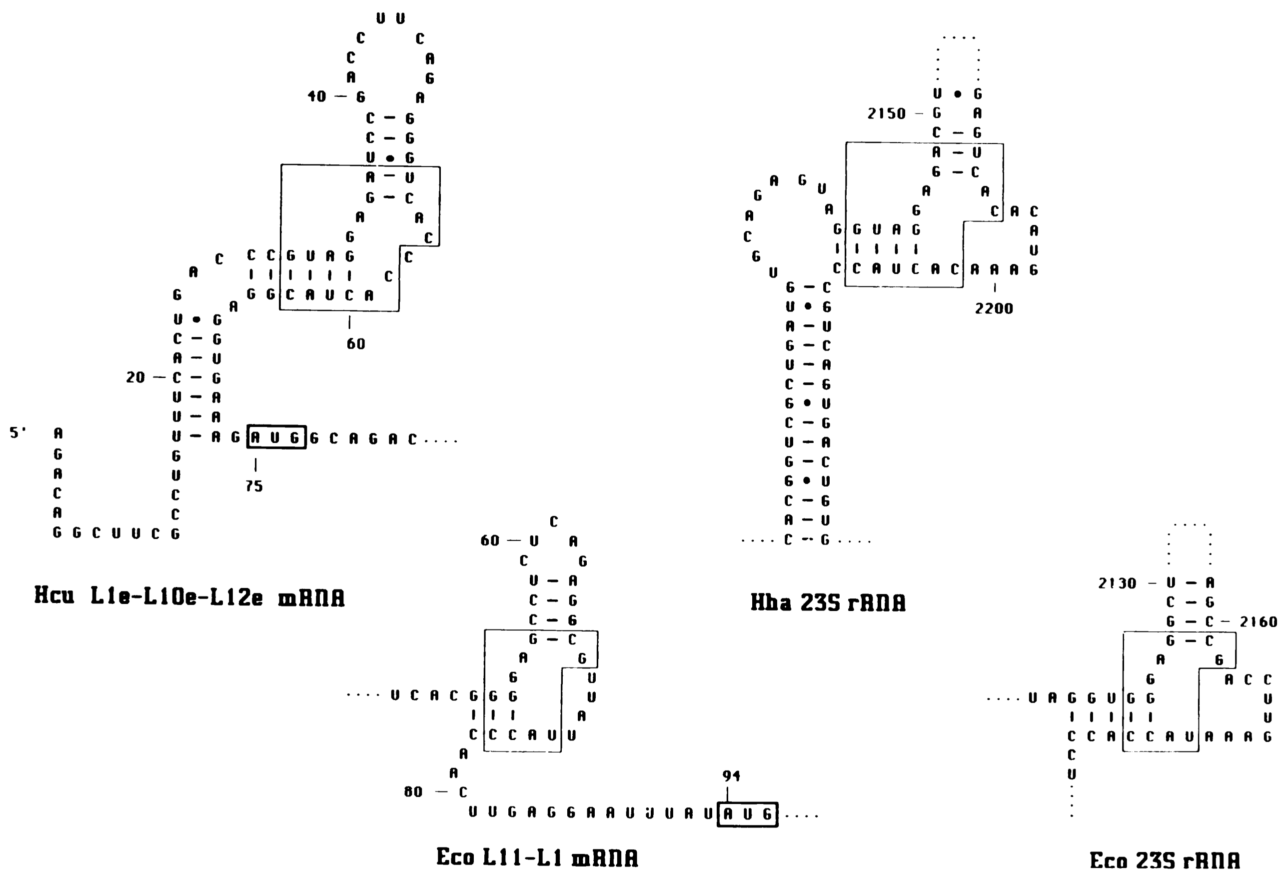



**Fig. 7.** Conserved sequence and structure in mRNA and 23S rRNA. The binding domain of ribosomal proteins L1 and L1e on *H.halobium* (Hha) and *E.coli* (Eco) 23S rRNA (right) and regions in the leader of the *H.cutirubrum* (Hcu) L1e–L10e–L12e and the *E.coli* (Eco) L11–L1 mRNA are illustrated. Regions 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 *H.cutirubrum* L1e–L10e–L12e mRNA (position 75) corresponds to nucleotide 2314 in Figure 1.

structure and function in the high intracellular ionic strength environments in which they function (Bayley and Morton, 1978; Eisenberg and Wachtel, 1987). The putative proteins encoded by NAB and ORF are also rich in acidic residues.

### Autogenous translational regulation

The L11e and L1e ribosomal proteins bind to specific and defined sequences within large subunit rRNA during ribosome assembly. The sequence of the 23S rRNA gene of *H.halobium* (closely related to *H.cutirubrum*) has recently been determined (Mankin and Kagramanova, 1986); the putative L11e and L1e binding domains at positions 1142−1201 and 2123−2222 are clearly identifiable. In *E.coli* the L1 protein plays a crucial role in autogenously regulating translation of the L11−L1 mRNA; excess L1 not immediately assembled into ribosomal particles can presumably bind to a sequence within the 5′ untranslated leader that exhibits both primary and secondary structural similarity to the L1 binding site on 23S rRNA (Figure 7).

In *H.cutirubrum* the L11e gene is transcribed usually as a monocistronic mRNA lacking a 5′ untranslated leader. The L1e gene is transcribed as the proximal cistron in the tricistronic L1e−L10e−L12e mRNA and is preceded by a 74 nt 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 7). 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.

A further search in and around the Hcu L11e and L1e genes for sequences resembling the two respective binding domains in 23S rRNA has been negative. It is possible that the NAB protein and/or the small amount of bicistronic NAB-L11e mRNA may have same regulatory significance.

## Materials and methods

### Materials

Restriction endonucleases were obtained from New England Nuclear, New England Biolabs, Bethesda Research Laboratories and Pharmacia, and were used as recommended by the suppliers. Klenow fragment and AMV reverse transcriptase were from Boehringer Mannheim. Polynucleotide kinase T4 ligase and calf intestinal alkaline phosphatase were from Pharmacia. [$\gamma$-$^{32}$P]ATP (>3000 Ci/mmol) and [$\alpha$-$^{32}$P]dNTPs (>3000 Ci/mmol) were from Amersham and New England Nuclear. Oligonucleotides were synthesized with an automated DNA synthesizer from Applied Biosystems. For resolution of secondary structure in sequencing, 7-deazaguanosine (Boehringer Mannheim) and dITP (Pharmacia) were utilized. Sequenase was from the United States Biochemical Corp.

### Isolation of clones

*Halobacterium cutirubrum* was grown in complex media as described in Sehgal and Gibbons (1960), harvested, resuspended in basal salts (4 M NaCl, 120 mM $MgSO_4$, 10 mM Na citrate, 3 mM KCl) and lysed with 0.2% Na deoxycholate. Crude lysate was extracted with phenol and octanol:chloroform, and banded by CsCl gradient centrifugation. Genomic DNA was restricted with various restriction enzymes, electrophoresed on 0.7% agarose gels, transferred to nitrocellulose and hybridized with the end labeled oligonucleotides specific for the L12e (oLW 9) and the L11e (oLW 17) genes essentially as described in Maniatis *et al.* (1982). The hybridization conditions were 45°C, 6 × SSC and 10 × Denhardt's for oLW 9 (L12e) and 40°C, 6 × SSC, 10 × Denhardt's for oLW 17 (L11e). This resulted in the identification of 1.3 kb *Pst*I−*Bam*HI (L12e) and 5.2 kb *Cla*I−*Bam*HI (L11e) fragments. Genomic DNA was then restricted with the appropriate enzymes, electrophoresed on 0.7% agarose gels and sized fractions recovered. The size fractionated DNA was Southern blotted with the appropriate end labelled oligonucleotide probe to determine which fraction contained the hybridizing fragment. Libraries were then constructed by inserting the fractionated genomic DNA between the *Pst*I and *Bam*HI sites of pUC8 (L12e) or the *Cla*I and *Bam*HI sites of pBR322 (L11e) and transformed. Approximately 1000 colonies were plated onto 147 mm YT ampicillin plates, lifted with a charged nylon membrane (Genescreen Plus from New England Nuclear) and hybridized with the appropriate end labeled oligonucleotide. Positives were picked and subjected to Southern blot analysis to ensure that they contained the 1.3 kb *Pst*I−*Bam*HI (L12e) and 5.2 kb *Cla*I−*Bam*HI fragment. Two independent isolates of each genomic clone were characterized by sequencing; pLW99 and pLW103 for the 1.3 kb *Pst*I−*Bam*HI (L12e) genomic fragment and pLW173 and pLW180 for the 5.2 kb *Cla*I−*Bam*HI genomic fragment.

### Sequencing

The sequence of pLW173 was characterized with two determinations of every nucleotide on both strands. In addition a second independent isolate, pLW180, had 95% of the nucleotides on each strand determined. Subclones were obtained in the vectors M13mp18/19 (Yanisch-Perron *et al.*, 1985), pTZ18R/19R (Pharmacia), pEMBL8+/− (Dente *et al.*, 1983), and the pUC plasmids (Viera and Messing, 1982). Deletion clones were obtained from subcloned fragments by the method of Dale *et al.* (1985) and with exonuclease III (Henikoff, 1984). Enzymatic sequencing was done essentially as previously described with Klenow (Sanger *et al.*, 1977), AMV reverse transcriptase (Biggin *et al.*, 1983) and Sequenase (as per the supplier) on both single stranded templates derived from M13mp18/19 and pEMBL8+/− and double stranded templates derived from the pUC plasmids and pTZ18R/19R. The guanosine analogues 7-deazaguanosine (Mizusawa *et al.*, 1986) and inosine (Mills and Kramer, 1979) were utilized to resolve secondary structure. Chemical sequencing was as previously described (Maxam and Gilbert, 1977).

Potential proteins encoded within sequence were compared with the Protein Sequence Database of the Protein Identification Resource with the program FASTP and significance of matches between potentially homologous proteins was assessed utilizing the RDF program (Lipman and Pearson, 1985). RDF calculates the statistical significance of a match by comparing the actual match to random matches generated by jumbling one of the protein sequences.

### Transcript mapping

RNA was isolated from exponentially growing cells by SDS lysis in the presence of 10 mM azide and extracted with phenol/chloroform. The RNA was subsequently purified by layering the lysate onto a 5.7 M CsCl block gradient and pelleting the RNA by centrifugation.

High resolution determination of the transcript ends was carried out by S1 nuclease protection experiments and primer extension of mRNAs with AMV reverse transcriptase. For S1 nuclease mapping the appropriate DNA fragment was 5′ or 3′ end labeled with T4 polynucleotide kinase or Klenow respectively, hybridized in 80% formamide, 40 mM Pipes, 400 mM NaCl and 1 mM EDTA at 61°C for 3 h and then treated with S1 nuclease. The products were electrophoresed on polyacrylamide gels alongside a chemical sequence ladder derived from the original labeled fragment or *Msp*I cut pBR322 that was 3′ end labeled by Klenow with [$\alpha$-$^{32}$P]dCTP. For primer extension analysis a 5′ end labeled oligonucleotide was hybridized to total RNA and extended with AMV reverse transcriptase and dNTPs (Newman, 1987). The resulting products were electrophoresed on polyacrylamide gels alongside a sequence ladder generated by the extension of the oligonucleotide by Klenow enzyme from an M13 subclone containing the requisite region of genomic DNA.

## Acknowledgements

We thank Tom Atkinson for the synthesis of the oligonucleotides and Celia Ramirez and Alistair Matheson for discussions. This work was supported by grants from the Medical Research Council of Canada and the Office of Naval Research of the USA (to P.P.D.). P.P.D. is a Fellow of the Canadian Institute for Advanced Research. L.C.S. was supported in part by a University of British Columbia Fellowship.

## References

Auer,J., Lechner,K. and Bock,A. (1989) *Can. J. Microbiol.* **35**, 200−204.

Bayley,S.T. and Morton,R.A. (1978) *CRC Crit. Rev. Microbiol.*, **6**, 151−205.

Beauclerk,A.A.D., Cundliffe,E. and Dijk,J. (1984) *J. Biol. Chem.*, **259**, 6559−6563.

Beauclerk,A.A.D., Hummel,H., Holmes,D.J., Bock,A. and Cundliffe,E. (1985) *Eur. J. Biochem.*, **151**, 245−255.

Biggin,M.D., Gison,T.J. and Hong,G.F. (1983) *Proc. Natl. Acad. Sci. USA.*, **80**, 3963–3965.
Boissy,R. and Astell,C.R. (1985) *Gene*, **35**, 179–185.
Brown,J.W., Thomm,M., Beckler,G.S., Frey,G., Stetter,K.O. and Reeve, J.N. (1988) *Nucleic Acids Res.*, **16**, 135–150.
Chant,J. and Dennis,P.P. (1985) *EMBO J.*, **5**, 1091–1097.
Dale,R.M.K., McClure,B.A. and Houchins,J.P. (1985) *Plasmid*, **13**, 31–40.
Dennis,P.P. (1985) *J. Mol. Biol.*, **186**, 457–461.
Dente,L., Cesareni,G. and Cortese,R. (1983) *Nucleic Acids Res.*, **6**, 1645–1655.
Dijk,J., Garret,R.A. and Muller,R. (1979) *Nucleic Acids Res.*, **6**, 2717–2729.
Downing,W.L. and Dennis,P.P. (1987) *J. Mol. Biol.*, **194**, 609–620.
Eisenberg,H. and Wachtel,E.J. (1987) *Annu. Rev. Biophys. Chem.*, **16**, 69–92.
Hanauz,G., Stoffler-Melicke,M. and van Heel,U. (1987) *J. Mol. Evol.*, **28**, 347–357.
Henikoff,S. (1984) *Gene*, **28**, 351–359.
Hui,I. and Dennis,P.P. (1985) *J. Biol. Chem.*, **260**, 899–906.
Itoh,T., Kamazaki,T., Sugitama,M. and Otaka,E. (1988) *Biochim. Biophys. Acta*, **949**, 110–118.
Johnsen,M., Christiensen,T., Dennos,P.P. and Fiil,N.P. (1982) *EMBO J.*, **1**, 999–1004.
Kjems,J. and Garrett,R.A. (1987) *EMBO J.*, **6**, 3521–3530.
Kjems,J., Leffers,H., Garrett,R., Wich,G., Lienfielder,W. and Bock,A. (1987) *Nucleic Acids Res.*, **15**, 4821–4835.
Lake,J.A. (1983) *Prog. Nucleic Acid Res.*, **30**, 163–194.
Lake,J.A. and Strycharz,W.A. (1981) *J. Mol. Biol.*, **153**, 979–992.
Liljas,A. (1982) *Prog. Biophys. Mol. Biol.*, **40**, 161–228.
Lindahl,L. and Zengel,J. (1986) *Annu. Rev. Genet.*, **20**, 297–326.
Lindahl,L., Jaskunas,S.R., Dennis,P.P. and Normura,M. (1975) *Proc. Natl. Acad. Sci. USA.*, **72**, 2743–2747.
Lipman,D.J. and Pearson,W.R. (1985) *Science*, **227**, 1435–1441.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbour Laboratory Press, Cold Spring Harbor, NY.
Mankin,A.S. and Kagramanova,V.K. (1986) *Mol. Gen. Genet.*, **202**, 152–161.
Matheson,A.T., Yaguchi,M., Christensen,P., Rollin,C.F. and Hasnain,S. (1984) *Can. J. Biochem. Cell Biol.*, **62**, 426–433.
Maxam,A.M. and Gilbert,W. (1977) *Proc. Natl. Acad. Sci. USA.*, **74**, 560–564.
Mills,D.R. and Kramer,F.R. (1979) *Proc. Natl. Acad. Sci. USA.*, **76**, 2232–2235.
Mizusawa,S., Nishimura,S. and Seela,F. (1986) *Nucleic Acids Res.*, **14**, 1319–1324.
Neumann,H., Gierl,A., Tu,J., Liebrock,J., Staiger,D. and Zillig,W. (1983) *Mol. Gen. Genet.*, **192**, 66–72.
Newman,A. (1987) *EMBO J.*, **6**, 3833–3839.
Nomura,M., Gourse,R. and Baughman,G. (1984) *Annu. Rev. Biochem.*, **53**, 119–162.
Oakes,M., Henderson,E., Scheinman,A., Clark,M. and Lake,J.A. (1986) In Hardesty,B. and Kramer,G. (eds), *Structure, Function and Genetics of Ribosomes.* Springer-Verlag, New York, pp. 47–67.
Pettersson,I. (1979) *Nucleic Acids Res.*, **6**, 2637–2646.
Pettersson,I. and Liljas,A. (1979) *FEBS Lett.*, **98**, 139–144.
Planta,R., Mager,W., Leer,R., Wondt,L., Raue,H. and El-Baradi,T. (1986) In Hardesty,B. and Kramer,G. (eds), *Structure, Function and Genetics of Ribosomes.* Springer-Verlag, New York, pp. 699–718.
Post,L., Strycharz,G., Nomura,M., Lewis,H. and Dennis,P.P. (1979) *Proc. Natl. Acad. Sci. USA.*, **76**, 1697–1701.
Ramirez,C., Shimmin,L.C., Newton,C.H., Matheson,A.T. and Dennis,P.P. (1989) *Can. J. Microbiol*, **35**, 234–244.
Reiter,W., Palm,P. and Zillig,W. (1988) *Nucleic Acids Res.*, **16**, 1–20.
Sanger,F., Nicklen,S. and Coulsen,A.R. (1977) *Proc. Natl. Acad. Sci. USA.*, **74**, 5463–5467.
Sander,G. (1983) *J. Biol. Chem.*, **258**, 10098–10102.
Sehgal,S.N. and Gibbons,N.E. (1960) *Can. J. Microbiol.*, **6**, 165–169.
Sharp,P.M., Cowe,E., Higgins,D.S., Shields,D.C., Wolfe,K.H. and Wright,F. (1988) *Nucleic Acids Res.*, **17**, 8027–8211.
Shimmin,L.C., Newton,C.H., Ramirez,C., Yee,J., Downing,W., Louie,A., Matheson,A.T. and Dennis,P.P. (1989a) *Can. J. Microbiol.*, **35**, 164–170.
Shimmin,L.C., Ramirez,C., Matheson,A.T. and Dennis,P.P. (1988b) *J. Mol. Evol.*, in press.
Strycharz,W.A., Nomura,M. and Lake,J.A. (1978) *J. Mol. Biol.*, **126**, 123–140.
Subramanian,A.R. and Dabbs,E.R. (1980) *Eur. J. Biochem.*, **112**, 425–430.
Thomas,M.S. and Nomura,M. (1987) *Nucleic Acids Res.*, **15**, 3085–3096.
Thomm,M. and Wich,G. (1988) *Nucleic Acids Res.*, **16**, 151–164.
Viera,J. and Messing,J. (1982) *Gene*, **19**, 259–268.
Wich,G., Hummel,H., Jarsch,M., Bar,U. and Bock,A. (1986) *Nucleic Acids Res.*, **14**, 2459–2479.
Woese,C.R. and Fox,G. (1977) *Proc. Natl. Acad. Sci. USA.*, **74**, 5088–5090.
Yaguchi,M., Matheson,A.T., Visentin,L.P. and Zuker,M. (1980) In Osawa,S., Ozeki,H., Uchida,H. and Yura,Y. (eds), *Genetics and Evolution of RNA Polymerase, tRNA and Ribosomes.* University of Tokyo Press, Tokyo, pp. 585–599.
Yanisch-Perron,C., Viera,J. and Messing,J. (1985) *Gene*, **33**, 103–119.
Zillig,W., Klenk,H., Palm,P., Puhler,G., Gropp,F. Garrett,R.A. and Leffers,H. (1989) *Can. J. Microbiol.*, **35**, 73–80.

*Received on December 20, 1988; revised on January 23, 1989*