Gene organization, transcription signals and processing of the single ribosomal RNA operon of the archaebacterium *Thermoproteus tenax* 

Jørgen Kjems\*, Henrik Leffers, Roger A.Garrett, Günter Wich<sup>1</sup>, Walfred Leinfelder<sup>1</sup> and August Böck<sup>1</sup>

Biostructural Chemistry, Kemisk Institut, Aarhus Universitet, DK-8000 Aarhus C, Denmark and <sup>1</sup>Lehrstuhl für Mikrobiologie der Universität, D-8000 München 19, FRG

Received March 13, 1987; Revised and Accepted May 12, 1987

Accession no. Y00346

#### ABSTRACT

The single ribosomal RNA (rRNA) operon from the extreme thermophile and archaebacterium *Thermoproteus tenax* was sequenced. Sites of transcriptional initiation and termination were established and the processing sites on the primary transcript were mapped with nuclease S1. The operon contained genes coding for 16S and 23S RNAs but lacked those coding for tRNA and 5S RNA. Transcription initiates 175 bp upstream from the start of the 16S RNA gene (Wich *et al.*, EMBO J. <u>6</u>, 523-528, 1987 [1]) and terminates 49 bp downstream from the 23S RNA gene within a long pyrimidine sequence. An open reading frame downstream from the rRNA operon is transcribed.

The sequences bordering both 16S and 23S RNA genes can form putative processing stems in the primary transcript that involve the whole of the 16S-23S RNA spacer. The stems contain irregular features that constitute processing signals and are conserved in other archaebacteria. The 16S RNA stem is cut prior to that of the 23S RNA and RNA maturation follows. An unusual 14 bp helix can form between the extremities of the transcript such that the whole transcript is highly structured and a fork-like structure is formed together with the processing stems.

The 23S RNA sequence was aligned with other available 23S-like RNA sequences (Leffers et al., J. Mol. Biol. <u>195</u>, in press [2]): a putative secondary structure exhibiting archaebacterial-specific features was deduced using comparative sequence analyses. A rooted phylogenetic tree was also derived for the archaebacteria that confirms their division into three major subgroups.

#### **INTRODUCTION**

The archaebacteria constitute a diverse group of organisms which includes the extreme thermophiles, the extreme halophiles and the methanogens [3]. Comparative sequence studies of both the 16S RNAs [4] and 23S RNAs [2] are consistent with the archaebacteria having evolved as a separate kingdom. Nevertheless, there are major phenotypic differences amongst the archaebacteria that include the subunit patterns of the RNA polymerases [5] and ribosome shapes [6] which have led Lake [6] to question whether all of the archaebacteria belong to one taxon. The gene organization of the rRNA operons also exhibit differences [8, 9]: thus, tRNA genes are absent from the sequenced extreme thermophile operons, and the *Desulfurococcus mobilis* operon lacks a coupled 5S RNA gene [9] and exhibits an intron in the 23S RNA gene [10].

rRNA genes are convenient systems for studying archaebacterial gene organization and transcription signals because of their relative ease of isolation. The availability of operon

sequences from various archaebacteria including *Halobacteria* [11-14], *Halococcus morrhuae* [2, 9], *Methanococcus vannielii* [15 and references therein] and *Methanobacterium thermo-autotrophicum* [16] provides a basis for comparative studies. Here we develop this approach by determining and analysing the sequence of the single rRNA operon from the extreme thermophile *Thermoproteus tenax;* the 16S RNA leader and gene sequences were presented elsewhere [1, 17].

#### MATERIALS AND METHODS

### Preparation of cells, cellular RNA and DNA; cloning and DNA sequencing

Cells from *T. tenax* were kindly provided by W. Zillig and grown under the conditions described by Zillig *et al.* [18]. Total cellular RNA was prepared using hot phenol extractions (65°C) essentially as described by Aiba *et al.* [19]. The phenol solution was frozen at  $-80^{\circ}$ C for 2 min and thawed before separating the aqueous phase. Two phenol-chloroform extractions were performed before the RNA was precipitated with ethanol and stored at  $-80^{\circ}$ C.

The procedures for DNA preparation and genomic cloning were described earlier (20). Subcloning procedures and the sequencing of the 16S RNA leader sequence and gene have also been described [1, 17]. The remaining sequence was obtained by digesting large DNA clones with four base pair-specific restriction enzymes. The mixtures were fractionated on 5% polyacrylamide gels and cloned into M13mp18 or mp19 in both directions [21]. Single-stranded DNA was prepared and used as a template for DNA sequencing by the dideoxy-nucleotide procedure [22] using [ $\alpha^{35}$ S]dATP as radioactive substrate. The DNA sequence was determined on 4-6% polyacrylamide gels containing salt or wedge-shaped gradients [23]. S1 nuclease mapping

0.1 to 1 µg DNA restriction fragment was labelled at either the 5'- or 3'-end [24]. It was digested with another restriction enzyme and purified on agarose or polyacrylamide gels to ensure that only one labelled end was present. 0.1-0.5 µg labelled DNA fragment (~100,000 cpm) was co-precipitated with 50-100 µg total cellular RNA. It was then denatured at 85°C for 10 min and hybridized in 40 mM PIPES, pH 6.4, 1 mM EDTA, 0.4 M NaCl, 80% formamide at 2-3 different temperatures. Optimal hybridization temperatures were calculated according to Favaloro *et al.* [25] and samples were also hybridized 5°C above and below these optima. After 3 hr, 300 µl ice-cold buffer (50 mM sodium acetate, pH 4.6, 0.28 M NaCl, 4.5 mM ZnSO<sub>4</sub>) was added, containing 20 µg/ml carrier ssDNA and 1 unit/µl S1 nuclease, and the solution was incubated for 1 hr at 30°C. The protected sites were determined by co-electrophoresing the sequence of the DNA fragment [26].

Determination of the 23S RNA secondary structure and phylogenetic analysis

All secondary structure predictions were performed using the sequence comparison approach that is based on evidence for compensating base changes. All of the available 23S-like RNAs were aligned for archaebacteria, eubacteria and the eukaryotic cytoplasm as described in detail by Leffers *et al.* [2]. Phylogenetic analyses for the archaebacterial kingdom were also based on the aligned 23S RNA sequences from *M. vannielii* [15]; *H. halobium* [14]; *H. morrhuae, M. thermoautotrophicum* and *D. mobilis* [2]. Homology values and  $K_{nuc}$  values were calculated as described by Hori and Osawa [27] and the phylogenetic tree was derived using the FITCH program of Felsenstein [28].

### RESULTS

#### Gene organization and primary structure of the operon

Only one rRNA operon was detected in the *T. tenax* chromosome using the Southern blotting procedure [29] and probing with radioactively labelled clones from within both the 16S and the 23S RNA gene (results not shown). This agrees with the conclusion of Neumann *et al.* [8]. The gene organization of this rRNA operon is illustrated in Fig. 1 and the nucleotide sequence is given in Fig. 2. The primary transcript contained neither tRNA genes nor introns. Moreover, the longest transcript terminated 49 bp after the 23S RNA gene (see below) and there was no coupled 5S RNA gene. A similar gene organization has also been reported for the extreme thermophile *D. mobilis* [9].

We sequenced 780 bp upstream and 2380 bp downstream from the primary transcript. An analysis of these sequences revealed neither rRNA nor tRNA genes but there was an open reading frame (ORF) downstream from the 23S RNA gene that coded for 388 amino acids (Fig. 1); the amino acid sequence is given in Fig. 2. Northern blotting experiments [24] were performed to establish whether the putative gene was transcribed. A *DdeI* restriction fragment extending from nucleotides 6397 to 7094 (Fig. 2) was labelled with  $[\alpha^{-32}P]dATP$  on the

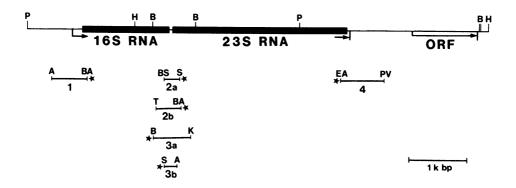


Figure 1: A map of the rRNA operon showing the arrangement of the rRNA genes, the location of the ORF and the major restriction sites. The limits of the rRNA transcript are indicated. The restriction fragments labelled 1 to 4 were used for S1 nuclease mapping. Asterisks denote the position of the end label on the DNA strand that was complementary to the RNA. Restriction enzyme abbreviations: A-AluI, B-BamHI, BA-BanI, BS-BstNI, EA-EagI, H-HindIII, K-KpnI, P-PstI, PV-PvuII and S-Sau96A.

TACCCCTTCC	CCACTACACT	CAAACACACC	CCCAACCTCC	GGGCAGGATT	TATCA ATCCC	COACCONAAA		TACCCTCTTT	TACCATCCTC
-100		-80							-10
			GGCCTCGGCC 40	GAATTGCCTT 50	AATGTTCCCC 60	CGACTAACCC 70	CCCTGGGTTT 80	AGCCCGGGCG 90	GGCGTCGGGG 100
GAGAAAACCC 110	CACGCCGCCG 120	AAACGCGCGG 130	AGTCCGGGGT 140	GGTCCCGCCC 150	CCGAGGGCGG 160	CCTGCGGCCG 170	стстс -1	6S RNA⊶C	TAGACCCTCG 1690
1700	1710	1720	1730	GCCGTCC -		4770	4780	CGAAAGAGCC 4790	4800
GGTACTCGTA 4810	ATACTCCTTT 4820	TTCTCCCGGG 4830	CTCCGTTTGG 4840	CTACAGTGTG 4850	TTCTGGTGGA 4860	GTTTATGCAC 4870	TGCAAGCCTG 4880	CTCATATTCG 4890	TGGAGGAGTC 4900
AGTGACGTAT 4910	GGGACGTTCG 4920	TAGCGATAGA 4930	TACAAATTCA 4940	GAGTCTGACG 4950	GAGGCAAGCC 4960	TATTTGTCTG 4970	GCGTGTTAAG 4980	ATCGCGGCAC 4990	AGCGGCGCGT 5000
CCGTCTGAGA 5010	AGACGACTTG 5020	CTATATGTCT 5030	CGCCCTCTTC 5040	ACATTTTCCG 5050	TTAGCAGTCT 5060	AGCTGTTCCG 5070	CCTCGCGCTA 5080	CAATACATCA 5090	AATACGGCCT 5100
CAGGTGAGCC 5110	GAGCTCCTCC 5120	CTCATGGCCT 5130	CGTAGAGCAA 5140	GTCTTGTCTG 5150	ACGTCGGCGT 5160	ATCCGTATAC 5170	TAAGATATTC 5180	CTCATGCCGG 5190	AGATACCCTT 5200
TGCTATTGGA 5210	CGTAGCTCGG 5220	GCGGAATGAG 5230	GTCTAGGACT 5240	ACCTCCACGG 5250	CGTCTCGATA 5260	TGTGGCGCCC 5270	CTCAGCCTCC 5280	TCCGCCGGAC 5290	CGCCACGAAG 5300
GAGGCTCTAG 5310	GAGAGCCTCT 5320	AGGAAGATGT 5330	GCATGTACCT 5340	CTCGGCGAGG 5350	GCTTGATAGT 5360	CCTCGTTTGA 5370	GCTGTACTCC 5380	TCCCAACTCA 5390	GCTGGGCGAG 5400
CCTCCTCAGC 5410	TTTTCGAGGG 5420	CCCTCTTCGC 5430	CTTAGACGCC 5440	CTATTGGCGA 5450	GACCCGGCGA 5460	GGGCATTTTT 5470	CACCATCTTT 5480	ACGAGGGACT 5490	CGAGGTCCGC 5500
5510	5520	5530	5540	CTGAGGCAGT 5550	5560	5570	5580	5590	5600
GCCGCGATGG 5610	GGCCAGCCCA 5620	GTCTAAAACC 5630	AACAGATCTA 5640	GCTTCCCGGG 5650	AACGCACCTC 5660	TCCAGCCCCC 5670	GTATAGGCGG 5680	CCAAGCTCCG 5690	TAAGGGATAT 5700
CCTTCTCCCC 5710	ACCTTCACCG 5720	CGAGATCCCA 5730	ATCGCTGAAC 5740	TCCGCAGCAT 5750	ATCCCAGAGC 5760	CCGGGAGCCG 5770	AACAGAACTA 5780	CGTACCTTAG 5790 M P	5800
								n r	PRA
GCCGCCACAT 5810 V Y L	5820	5830	5840	CGTAGAGAAA 5850 SLY0	5860	5870	5880	ATCCACATGC 5890	
5810 VYL	5820	5830 VIM	5840 CFN	5850 S L Y Q	5860 Y T W AGTACACTTG	5870 NLL	5880 A P M I	ATCCACATGC 5890 G R A	CACCAAGAGC 5900 M G L
5810 WYL TTGGTATTTG 5910 GVLA	5820 A A G F GCCGCGGGGCT 5920 E A V	5830 V I M TTGTGATTAT 5930 G F T	5840 C F N GTGTTTCAAT 5940 I Y V I	5850 S L Y Q TCCTTATATC 5950 V S T	5860 Y T W AGTACACTTG 5960 V A Q	5870 N L L GAATTTGTTG 5970 P A G G	5880 A P M I GCTCCGATGA 5980 A L A	ATCCACATGC 5890 G R A TAGGGCGCGCC	CACCAAGAGC 5900 M G L CATGGGGCTG 6000 G P R G
5810 WYL TTGGTATTTG 5910 GVLA GGCGTCTTGG 6010	5820 A A G F GCCGCGGGCT 5920 E A V CAGAGGCCGT 6020	5830 V I M TTGTGATTAT 5930 G F T CGGCTTCACA 6030	5840 C F N GTGTTTCAAT 5940 I Y V I ATATACGTTA 6040	5850 S L Y Q TCCTTATATC 5950 V S T TAGTGTCAAC 6050	5860 Y T W AGTACACTTG 5960 V A Q GGTCGCGCAA 6060	5870 N L L GAATTTGTTG 5970 P A G G CCGGCCGGCG 6070	5880 A P M I GCTCCGATGA 5980 A L A GAGCTCTGGC 6080	ATCCACATGC 5890 G R A TAGGGCGCGCC 5990 D L R TGACTTAAGA 6090	CACCAAGAGC 5900 M G L CATGGGGCTG 6000 G P R G GGCCCGCGCG 6100
5810 WYL TTGGTATTTG 5910 GVLA GGCGTCTTGG 6010 VGA GCGTGGGGGGC	5820 A A G F GCCGCGGGGCT 5920 E A V CAGAGGCCGT 6020 L S A CCTCTCGGCC	5830 V I M TTGTGATTAT 5930 G F T CGGCTTCACA 6030 V L S A GTCCTCTCGG	5840 C F N GTGTTTCAAT 5940 I Y V I ATATACGTTA 6040 L G F CGCTGGGCTT	5850 S L Y Q TCCTTATATC 5950 V S T TAGTGTCAAC 6050 I G A CATAGGGGCA	5860 Y T W AGTACACTTG 5960 V A Q GGTCGCGCAA 6060 A L A P GCCCTGGCGC	5870 N L L GAATTTGTTG 5970 P A G G CCGGCCGGCG 6070 G P A CGGGGCCTGC	5880 A P M I GCTCCGATGA 5980 A L A GAGCTCTGGC 6080 L L Y GCTTTTGTAT	ATCCACATGC 5890 G R A TAGGGCGCGC 5990 D L R TGACTTAAGA 6090 L A W G CTGGCGTGGG	CACCAAGAGC 5900 H G L CATGGGGCTG 6000 G P R G GGCCCGCGCG 6100 L G S GGCTCGGCAG
5810 W Y L TTGGTATTG 5910 G V L A GCCGTCTTGG 6010 V G A GCGTCGGGGCG 6110 A G E CGCGGGCGAG	5820 A A G F GCCGCGGGCT 5920 E A V CAGAGGCCGT 6020 L S A CCTCTCGGCC 6120 G V L Y GGAGTGCTCT	5830 V I M TTGTGATTAT 5930 G F T CGGCTTCACA 6030 V L S A GTCCTCTCGG 6130 G I A ACGGAATAGC	5840 C F N GTGTTTCAAT 5940 I Y V I ATATACGTTA 6040 L G F CGCTGGGCTT 6140 F N L GTTCAATCTG	5850 S L Y Q TCCTTATATC 5950 V S T TAGGGTCAAC 6050 I G A CATAGGGGCA 6150 A V K V GCCGTCAAAT	5860 Y T W AGTACACTTG 5960 V A Q GGTGCGCGCAA 6060 A L A P GCCCTGGCGC 6160 Y Q D GGTACCAAGA	5870 N L L GAATTTGTTG 5970 P A G G CCGGCCGCG 6070 G P A CGGGGCCTGC 6170 K L G CAAGCTGGGC	5880 A P M I GCTCCGATGA 5980 A L A GAGCTCTGGC L Y GCTTTTGTAT 6180 L A T G CTCGCCACAG	ATCCACATGC 5890 G R A TAGGCCGCC 5990 D L R TGACTTAAGA 6090 L A W G CTGGCCTGGG 6190 L V S GCCTCGTGTC	CACCAAGAGC 5900 M G L CATGGGGCTG 6000 G P R G GGCCCGCGGG 6100 L G S GGCTCGGCAG 6200 L G F GCTCGGCTTC
5810 V Y L TIGGTATTIG 5910 G V L A GCCTCTIGG 6010 V G A CCCTGGGGCC 6110 A G E CGCGGGCCAG 6210 G L G S GCCTTGGGT	5820 A A G F GCCCCGCGCT 5920 E A V CAGAGGCCT 6020 C S A CCTCTCGCC 6120 G V L Y GGACTGCTCT 6220 A V A CCGCCGTGGC	5830 V I M TTGTGATTAT 5930 G F T CGGCTTCACA 6030 V L S A GTCCTCTCGG 6130 G I A ACGGAATAGC 6230 N P L CAACCCGTC	5840 C F N GTGTTTCAAT 5940 I Y V I ATATACGTTA 6040 L G F CGCTCGCGCTT 6140 F N L GTTCAATCTG 6240 I A S V ATACCCTCCCG	5850 S L Y Q TCCTTATATC 5950 V S T TAGTGTCAAC 6050 I G A CATAGGGCA 6150 A V K V G N Y TGGCCAAAT 6250 G N Y	SB60 Y T W AGTACACTTG 5960 V A Q GGTCCCCCAA GGCCCTGCGCC GCCCTGCGCC GCCCTGCGCC Y Q D GGTACCAAGA 6260 R E A CAGAGAGGCC	5870 N L L GAATTTGTTG 5970 P A G G CCGGCCGGCG 6070 G P A CGGGCCCTGC 6170 K L G CAAGCTGGGC 6270 T L A I ACTTTGGCTA	5880 A P M I GCTCCGATGA 5980 A L A GAGCTCTGGC 6080 L Y GCTTTTGTAT 6180 L A T G CTCCCCACAG G V V TCGGCGTAGT	ATCCACATGC S890 G R A TAGGGCGCGC 5990 D L R TGACTTAAGA 6090 L A V G CTGGCGTGGG 6190 L V S GCCTCGTGTCC 6290 E L L TGAACTCCTA	CACCAAGACC 5900 H G L CATCGGGCTG 6000 G P R G GGCCCGCGCG 6100 L G S GGCCCGGCAG 6200 L G P GCTCGGCATC 6300 V L V P GCTCTAGCTTC
5810 V Y L TTGGTATTTG 5910 V L A GCGTGGGGGC 6010 V C A CGCGGGGCGAG CGCGGGGCGAG CGCGGCGCGAG CGCGCGCG	5820 A A G F GCCCCGGCGCT 5920 E A V CAGAGGCCGT 6020 L S A CCTCTCGGCC 6120 G V L Y GGAGTGCTCT 6220 A V A CCGCCCTGGC 6320 L V D	5830 V I M TTGTGATTAT 5930 G F T CGGCTTCACA 6030 V L S A GTCCTCTCGG 6130 G I A ACGGAATAGC 6230 N P L CAACCCGCTC 6330 Y P R G	5840 C F N GTGTTTCAAT 5940 I Y V I ATATACGTTA 6040 L G F CGCTGGGCTT 6140 F N L GTTCAATCTG 6240 I A S V ATAGCCTCCG 6340 L S G	S850 S L Y Q TCCTTATATC 5950 V S T TAGTGTCAAC 6050 I G A CATAGGGGCA 6150 G N Y TGGGCAACTA 6350 V S P	SB60 Y T W AGTACACTTG 5960 V A Q GGTCGCCCAA 6060 Y Q D GGTACCAAGA 6260 Y Q D GGTACCAAGA 6260 R E A CAGAGAGGCC 6360 R R A L	5870 N L L GAATTGTTG 5970 P A G G CCCGCCCGCCG 6070 G P A CGGGGCCTGC 6170 K L G CAAGCTGGGC 6270 T L A I ACTTTGCTA 6370 L D A	5880 A P M I GCTCCGATGA 5980 A L A GAGCTCTGGC 6080 L L Y GCTTTGTAT 6180 C A T G CTCGCCACAG 6280 G V V TCGGCGTAGT 6380 R P V	ATCCACATGC S890 G R A TAGGCGCGC 5990 D L R TGACTTAGA 6090 L A V G CTGGCCTGGC 6190 L V S 6290 CCTCGTCTC 6290 E L L TGAACTCCTA 6390 T L Y A	CACCAAGAGC 5900 H G L CATGEGGCTG 6000 G P R G GGCCGCGCGG 6100 L G S GGCCGCGCAG 6200 L G P GCCCGCCTC 6300 V L V P GTCTTAGTTC 6400 S Y A
5810 V Y L TTGGTATTTG 5910 V C A GCCTGCGGCGC GCCTGCGCGCGC GCCGCGGCGCGG GL G S GCCCTCGCGT 6310 L S L CCCTGTCCCT 6410	5820 A A G F 5920 E A V CAGAGGCCGT 6020 L S A CCTCTCGCCC G V L Y GGAGTGCTCT 6220 A V A CCGCCGTGGC L V D CCTCGTCGAT 6420	5830 V I M TTGTGATTAT 5930 G F T CGGCTTCACA 6030 G I A ACGGAATAGC 6230 N P L CAACCGGTC 6330 Y P R G TACCCGAGGG 6430	5840 C F N GTGTTTCAAT 5940 I Y V I ATATACGTTA 6040 L G F CGCTGGGCTT 6140 F N L GTTCAATCTG 6240 I A S V ATAGCCTCCG 6340 L S G GGCTCTCCGG 6440	5850 S L Y Q TCCTTATATC 5950 V S T TAGTGTCAAC 6050 I G A CATAGGGCAA 6150 G N Y TGGCCAACTA 6350 V S P GGTCTCTCCG 6450	5860 Y T W AGTACACTTG 5960 V A 0 GGTCGCGCAA 6060 Y 0 D GGTACCAAGA 6260 R E A CAGAGAGCCC 6360 R R A L AGCGCGCCC 6460	5870 N L L GAATTTGTTG 5970 P A G G CCGGCCGGCG 6070 G P A CGGGGCCTGCC 6170 K L G CAAGCTGGGCC T L A I ACTTIGGCTA 6370 L D A TGCTCGACGC 6470	5880 A P M I GCTCCGATGA 5980 A L A GAGCTCTGGC 6080 L Y GCTTTTGTAT 6180 C CCCCCACAG 6280 G V V TCGCGTAGT 6380 R P V GCGCTTCTGG 6480	ATCCACATGC S890 G R A TAGGGCGCGC 5990 D L R TGACTTAAGA 6090 L A V G CTGGCCTGGG CTGGCCTGGG CTGGCCTGGG L V S GCCTCGTGTC 6290 T L Y A ACGCTCTACG 6490	CACCAAGACC 5900 H G L CATGGGCTG 6000 G P R G 6000 L C S GGCCCGCCG 6100 L C S GGCTCGCAC 6200 L G P GTCTGGCAC 6300 V L V P GTCTTAGTC 6400 S Y A CCTCCTACCC 6500
5810 V Y L TTGGTATTTC 5910 C V L A GCGTCTTGG 6010 V C A CCCTGCGGCC 6110 C C C C GCCCGCGCCA 6210 C L G S GCCTTGCT 6310 L S L CCCTGTCCCT 6410 L G A	5820 F GCCCCGCGCT 5920 E A V CAGAGGCCGT 6020 C S A CCTCTCGCCC 6120 G V L Y GGAGTGCTT 6220 A V A CCCCCTCGCC CCCCCTCGCCAT	5830 V I M TTGTGATTAT 5930 G F T CGGCTTCACA 6030 V L S A GTCCTCTCGG 6130 G I A ACGGAATAGC 6230 N P L CAACCCGGTC 6330 V P R G TACCCAGGG 6430 S L A	5840 C F N GTGTTTCAAT 5940 L G F CGCTGGGCTT 6140 F N L GTTCAATCTG 6240 I A S V ATAGCCTCCG 6340 L S G GGCTCTCCGG 6440 S S L	S850 V S L Y Q TCCTTATATC 5950 V S T TAGTGTCAAC 6050 I G A CATAGGGCA A V K V GCCGTCAAAT 6250 G N Y TGGCCAACTA 6350 V S P GGTCTCTCCC 6450 H L L V	SB60 Y T W AGTACACTTG 5960 V A 0 GGTCGCCCAA 6060 A L A P GCCTGGCGCC 6160 Y 0 D GGTACCAAGA 6260 R E A CAGAGAGGCC CAGAGAGGCC R R A L	5870 N L L GAATTTGTTG 5970 CCGGCCGGCC 6070 G P A CGGGGCCTGC 6170 K L G 6270 T L A I 6370 L D A TGCTCGACGC 6470 L V V L V	S880 A P H I GCTCCGATGA S980 L A A GGCTCTGGC GCTTTGTAT 6180 L A T G GCTCCGCCACAG 6280 G V V TCGCGCTACTC 6380 R P V GCGCTCTCGG 6480 L L A S CTCTTGGCT	ATCCACĂTGC 5890 G R A TAGGGCCGCC 5990 D L R GOTTAAGA 6090 L V S GCTCGGCTGGG 6190 L V S GCTCGGTGCG 6290 E L L TGAACTCCTA 6390 T L Y A ACGCTCTACG 6490 L Y P	CACCAAGAGC 5900 H G L CATGEGECTG 6000 G P R G GGCCCCCCGC 6100 L G S GGCTCGGCAG 6200 L G P GCTCGGCATC 6300 V L V P GTCTTAGTTC 6400 S Y A CCTCCTACGC
5810 V Y L TTGGTATTG 5910 G V L A GCGTGGGGC 6010 V G A GCCTGGGGCG G L G S GCCTTGGG G L G S GCCTTGGCT 6310 L S L CGCTGTCCCT 6310 L G A CCTCGGCCGA 610 C A A R GGAGCCGCGA	5820 A A G P C GCCCCGCC 59200 L A V CCAGAGCCCCT 60200 A V A CCCCCTCCGCCA 6220 A V A CCCCCCTCCC 6220 A V A CCCCCCTCC CCC 4 V D CCCCCCCCCC 6420 V P L L GCCCCCTCC 6420 V P L L	5830 V I M TTGTCATTAT 55300 G F T CGGCTTCACA 6130 G I A ACCGATACC 6430 N P L CAACCCGTC 6430 N P L CAACCGATC 6430 S L A TCTCCTTGGC 6530 G A L GGGCCCCTC	5840 C F N GTGTTTCAAT 5940 L C F N 47474CGTTA 6040 L G F CCCTG6GCTT 6140 F N L GTTCAATCTG 6140 F N L GTCAATCTG 6440 S S L CTCCTCGCTA 6440 S S L CTCCTCGCTA 6440 C S S L CTCCTCGCTA CCTCGCCTA CCTCGCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCTA CCTCCCCCCTA CCTCCCCCCTA CCTCCCCCCCA C C CCCCCCCCCCCA C C CCCCCCCCCC	5850 500 V ST TAGTGTCAAC 60500 V S T TAGTGTCAAC 60500 I G A CATAGGGCAA 61500 A V K V GCCGTCAAAT 62500 G N Y GGCGCAACT 64500 H L L V CATCTCTCG 64500 H L L V CATCTCTCG 65500 G P L 66500 G P L 66500 G P L	5860 Y T V AGTACATTG 55960 V A Q GGTGCGCAA A L A P GGCCGCGCCC GCCCGGCCC GCCCGGCCC CAGCAGCACC CAGCAGCACC CAGCAGCACC G G G TCGGCGGAGG G 5500 K A I CAAGCCCATC	5870 N L L GAATTFGTTC GATTGTTC GOV P A G G CCGGCCGGCC 6070 C P A GGGGCCTCC 6170 K L G CAACCTGGCC 6270 L D A TCCTCGACC 6470 E L V GGAACTCGTC 6470 F L Q CGACCTCGCC 6470 C L A L 70 CAACTCGCCACC 6470 C L A L 70 CAACTCGCCACC 6470 C L A L 70 CAACTCGCCACC 70 C L A L 70 C L A L 70 C CGCCCCCACC 70 C L A L 70 C CGCCCCCCCCCCCC 6470 C L A L 70 C L A L 70 C C CGCCCCCCCCC 6470 C L A L 70 C C CACCC 70 C L A L 70 C CACCCCCACCC 70 C C CACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	5880 A P M I GCTCCATCA 59800 L L Y GCTTTGTAT 6180 L A T G CTCCCCACAG 6280 G V V TCCGCCTATCT 6380 C V V CCGCCTTCTGC 6480 L A X S CTCTTGGCT 6480 L A X S	ATCCACATCC 5890 G R A TAGGCCCCC 5990 D L R TGACTTAACA 6090 L A W G CTGCCTGCGC CGCCC CGCCCTGCC E L L 6390 E L L 6390 T L Y A ACCCTCATCC 6490 L Y P CCCCTCATCC	CACCAAGACC 5900 H G L CATGGGCTG 6000 G P R G 6100 L G S GGCCGCCGCG 6200 L G P GTCTGGCAG 6200 L G P GTCTGGCAG 6400 S Y A CCTCCTACCC 6500 L L V CCTCCTACTT 6600 C L L V CCTCCTACTT 6600 L A V CCTCCTACTT 6600 C L A A H
5810 V Y L TTGGTATTTG 5910 V C A GCCTGCGGCC GCCGCGCCAC GCCTGCGCCAC GCCTGCGCCAC GCCTCGCGCAC GCCTCGCCCA GCCTCGCCCA CCCCGCCCAC GAA R GCACCCCCCA GCA A R GCACCCCCCA GCA A C GCACCCCCCA GCA C GCACCCCCCA GCA C GCACCCCCCA GCA CCTCGCCCCAC	5820 A A G P C GCCCCGCC 59200 L A V CCAGAGCCCT 60200 A V A CCCCCTCCGCCA 62200 A V A CCCCCCTGCCA 62200 V L Y D CCCCCCTCCCA 64200 V P L L GCCCCCTCC 66220 V P L L GCCCCCTCC 66200 V P L L GCCCCCTCCC 66200 V C V C GCCCCTCCCCCCCCC 66200 V C V C GCCCCCTCCCCCCCCC 66200 V C L V D C CCCCCCTCCCCCCCCCCC 66200 V L L L C CCCCCCTCCCCCCCCCCCCCCCCCCCCCCCC	5830 V I M TTGTCATTAT CGGCTTCACA 6 6 130 G F T CGGCTTCACA 6 130 G I A ACCGATACC 6 230 N P L 6 330 G A ACCGATACC 6 330 G T A CCCACCCCCTC 6 330 G T A CCCCCCCCCC 6 330 G A CACCCCCCCCCCCCCCCCCC 6 35 C A CACCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	5840 C F N GTGTTTCAAT 59400 L C F N 4 ATATACGTTA 60400 L G F CCCTG6GCTT 6140 F N L GTTCAATCTG 62400 S S L CTCCTCGCG GGCTCCTCCGG GGCTCCTCCGG A D K W GGCTACAAGT 66400 A V G	5850 S L Y Q TCCTTATATC 59500 V S T TAGTGTCAAC 60500 I G A CATAGGGCA CATAGGGCA G N Y GCCGTCAAT 62500 G N Y GGCGCAACT 64500 H L L V CATCTCTCG 64500 G P L 66500 C T G	5860 Y T V AGTACATTG 55960 V A Q GGTGCGCAA A L A P GGCCGCGCCC GCCCTGGCCC GCCCTGGCCC GCCCGGCCC CAGACAGCCC CAGACAGCCC CAGACAGCCC CAGACAGCCC CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCA CAGCCAGCACA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCA CAGCAGCAGCAGCA CAGCAGCAGCAGCA CAGCAGCAGCAGCA CAGCAGCAGCAGCA CAGCAGCAGCAGCA CAGCAGCAGCAGCA CAGCAGCAGCAGCAGCA CAGCAGCAGCAGCAGCA CAGCAGCAGCAGCAGCA CAGCAGCAGCAGCAGCAGCA CAGCCAGCAGCAGCAGCAGCA CAGCCAGCAGCAGCAGCAGCAGCA CAGCCAGCAGCAGCAGCAGCAGCA CAGCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG	5870 N L L GAATTFGTTG GATCATCA 5970 P A G G 6070 P A G G 6070 CGGGCCTGC 6170 K L G 6270 CAACCTGGCC 6270 L D A TCCTCGACG 6470 E L V GGAACTCGTC 6570 Y L A L TATCTGGCC 6670 I L V	5880 A P M I GCTCCATCA 59800 L L Y GGCTTTGGC GCTTTGTA 6180 L A T G CTCCCCACAG G V V TCCGCGTATCT GS00 L A S CTCTGGCC G V V TCCGCGTATCTGG GSCT GSCT	ATCCACATCC S990 G R A TAGGCCCC S990 D L R TAGGCCCC 6090 L V S GCTCGCTCGC 6090 L V S GCTCGCTCGC 6090 L V S GCTCGCTCGC 6390 T L Y A ACCCTCACC 6490 L Y P CCCTCTACC 6490 A A G GCCCCCCCCC 6690 S A A G GCCCCCCCCCCC 6690 S A A G GCCCCCCCCCCCCCC 6690 S A S A G GCCCCCCCCCCCCCCCCCCC 6690 S A S A G GCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCARGACC 5900 H G L CATGEGECTG 6000 G P R G 6100 L G S GCCCCECGECG 6200 L G P GCTCGECATC 6400 V L V P GCTCCTACTTC 6400 C L C V CCTCCTACTTC 6500 L L V CCTCCTACTC 6500 C L L V CCTCCTACTC 6500 C L L V CCTCCTACCC 6500 C L C V CCTCCTACTC 6500 C C C C C C C C C C C C C C C C C C C
5810 V Y L TTGGTATTG 5910 G V L A GCGTGGGGC 6010 V G A GCCTGGGGCG G L G S GCCTTGGG G L G S GCCTTGGG G C C G G C C G G C C G G C C C G G C C C G G C C C G G C C C C	5820 A A G F A GCCGCGCT 5920 E A V CAGAGCCCT 6020 A V A CCTCTCGCGC CCTCTCGCC 6220 A V A CCCCCTCGC 6220 A V A CCCCCTCCA 6220 V L Y CCCCCTCCA 6220 V L V CCCCCTCCA 6220 V L V C CCCCCTCCA 6220 V A C CCCCCTCCA CCCCC C C CCCCC C C C C C	5830 V I M TTCTCATAT 5930 V L SA G P T CGGCTTCACA 6130 G I A ACGCATTAC 6330 N P L CAACCGCTC 6330 N P L CACCCCAC 6330 C A CACCCCAC 6330 C A CACCCCAC 6330 C A CACCCCAC 6330 C A CACCCCAC 6330 C A CACCCCAC 6330 C A C CACCCCCCCC 6330 C A C C C C C C C C C C C C C C C C C C	5840 C F N GTGTTTCAAT 5940 L C F N 6040 L G F 6140 F N L GTTCAATCGT 6140 F N L GTTCAATCGT 6240 S S L CTCCCCGGCT 6340 L S G GGGTCTCCCC GGGCTCCCCC A D K W GGCTCACCAGT 6640 A V G CTCCCCGGGG 6740 Y T A GTACCGTCCC	5850 CTCATTATATC 5950 V S T TAGTGTCAAC 6050 I G A 6050 CATAGGGCAAC 6150 A V K V GCCGTCAAAT 6250 G N Y GCCGTCACTCG 6350 V S P GGTCTCTCCG 6350 V S P GGTCTCTCCG 6550 H L 450 CATCCTCTG 6550 L T G TCTCACGGCC K A V G AGGCCTCCTC	5860 Y T V AGTACATTG 55960 V A Q GGTGCGCAA A L A P GGCCGCGCCC GCCCTGGCCC GCCCTGGCCC GCCCGGCCC CAGACAGCCC CAGACAGCCC CAGACAGCCC CAGCCAGCA GG G G CGGCGGCGC CAGCCGCCC CAGCCAGCA CAGCCAGCAC CAGCCAGCAC CAGCCAGCAC CAGCCAGCAC CAGCCAGCAC CAGCCACACAC CAGCCACACAC CAGCCACACAC CAGCCACACAC CAGCCACACAC CAGCCACACAC CAGCCACACAC CAGCCACACAC CAGCCACACAC CAGCCACACACAC CAGCCACACAC CAGCCACACAC CAGCCACACACAC CAGCCACACACAC CAGCCACACAC CAGCCACACACAC CAGCCACACACACAC CAGCCACACACACACACACACACACACACACACACACAC	5870 N L L GAATTTGTTG G S A G CCGGCCGGCG G P A G CGGGCCTGCG G P A G CGGGCCTGCG G CAAGCTGGGC C CAAGCTGGGC C CAAGCTGGGC G 20 C CAAGCTGGCC G 20 C CAAGCTGGCC G 20 C C CAAGCTGGCC G 20 C C CAAGCTGGCC G 20 C C C C C C C C C C C C C C C C C C C	5880 A P M I GCTCCGATGA 5980 A L A GAGCTCTGGC 6080 C L L Y GCTCTGGCA 6180 C V V TCGCGTAGT 6380 C V V TCGCGTAGT 6480 C CCCTTCGGC 6480 C A V S TCGCTGCGCA 6480 C A V S TCGCCTGCCG 6480 C A V S TCGCCTGCGCA 6480 C A V S TCGCCTGCCG 6480 C A V S TCGCCTGCCG 6480 C A V S TCGCCGCA 6680 C A V S TCGCCGCA 6680 C A V S TCGCCGCCCCCCCC 6680 C A V S TCGCCGCCCCCCCC 6680 C A V S TCGCCGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	ATCCACATCC ATCACCACATCC S990 D L R TAGGCCCCCC 6090 L V S CCTGCCTGGC 6290 C L V S CCTGCCTGGC 6290 T L Y A ACCCTCATCC 6390 C V Y ACCCTATCCAC 6590 C V P CCCCCGCCCC 6590 C S S R I TCCACCACGAC 6790 V Y A	CACCAAGACC 5900 H G L CATGGGGCTG 6000 G P R G 6200 L G S GGCTCGCCAG 6200 L G P GCTCGCCATC 6300 V L V P GTCTTAGTTC 6400 S Y A CCTCCTACTC 6500 L L V CCTCCTACTC 6500 L L V CCTCCTACTC 6600 CTCCCCACGC 6500 L L V CCTCCTACTC 6600 CTCCCCACCC 6500 CTCCCCACCC 6500 L A CCTCCCACCC 6500 L A 6500 CTCCCCCACCC 6500 CTCCCCCACCC 6500 CTCCCCCCCCCCCC 6500 CTCCCCCCCCCCCCCCC 6500 CTCCCCCCCCCCCCCCCCCC 6500 CTCCCCCCCCCCCCCCCC 6500 CTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
5810 V Y L TTGGTATTTC 5910 V C A GCGTCTTGG 6010 V C A GCCTGGGGCC GL G S GCCTGGGCCA GL G S GCCTGCGCCA 6310 L S L CCCTGTCCCT 6410 CCCTGCCCA 6510 G A A R GCACCCCCCA 6510 G A A R CACGCCCCA 6510 C CACGCCCA 6510 C CACGCCCA 6510 C C A C CCCGCCCCA 6510 C C A C CCCGCCCCA 6510 C C A C CCCCGCCCA 6510 C C C GCCCGCCCA 6510 C C C CCCCCCCCCA 6510 C C C CCCCCCCCCA 6510 C C C CCCCCCCCCA 6510 C C C CCCCCCCCCA 6510 C C C CCCCCCCCA 6510 C C C CCCCCCCCCA 6510 C C C CCCCCCCCCCCA 6510 C C C CCCCCCCCCA 6510 C C C CCCCCCCCCCCA 6510 C C C CCCCCCCCCCA 6510 C C C CCCCCCCCCCA 6510 C	5820 A A C F P GCCCCCCCCC E A V CAGAGCCCT 6020 L S A CCTCTCCGCC 6120 C V L Y GGACTGCTCT 6220 A V A CCCCCCTCCC 6520 P L L GCCCCCTCC CCCCCTCC 6520 P L L CTCCCCCCCC 6520 P L L CTCCCCCCTC 6520 P L L CTCCCCCCTC 6520 P L L CTCCCCCTCC 6520 P L L CTCCCCCTCCC 6520 P L L CTCCCCCTCCC 6520 P L L CTCCCCCCTCC 6520 C L D L CTCCCCCTCCC 6520 C L D L CTCCCCCTCCC CCCCCTCCCCCCCCCC CCCCCTCCCCCC	5830 V I M TTGTGATTAT CGGCTTCACA 6030 V L S A 6030 V L S A 6030 V L S A 6130 G I A ACCGATACC 6230 Y P R G 6330 Y P R G 6430 S L A TCTCCTTGC 6630 G A L CGGCCCTCT 6630 G A L CGGCCCTCT 6630 G C I A A A ACCGATACC 6530 C C I C C C C 6530 C C C C C 6530 C C C C C 6630 C C C C C 6630 C C C C C C 6630 C C C C C 6630 C C C C C C C 6630 C C C C C C C 6630 C C C C C C C C C 6630 C C C C C C C C C 6630 C C C C C C C C C C 6630 C C C C C C C C C C C 6630 C C C C C C C C C C C C C C C C C C C	5840 C F N GTGTTCAAT GTGTTCAAT GTGTTCAAT 6040 L G F CCCTGGGCTT 6140 F N L GTTCAATCTC 6140 F N L GTTCAATCTC 6140 C S L CTCCCGTGCA 6140 C S L CTCCCCGCTA 6540 A V G CCCCCGCCC 6440 S S L CTCCCCCCTA 6540 A V G CCCCCCCCCC 6440 S S L CTCCCCCCCCA 6540 A V G CCCCCCCCCC 6440 S S L CTCCCCCCCCC 6440 S S L CTCCCCCCCCC 6440 S S L CTCCCCCCCCCCC 6440 S S L CTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	5850 510 510 510 510 510 510 510 5	5860 Y T V AGTACACTTC Y A Q GGTCGCCCAA 6060 Q A Q GGTCGCCCACACA 6260 R E A CAGAGAGCCC CAGAGAGCCC G G C G G C G A I I CAAGCCCATCC GGCCCATCACA GAGCCCATCT GGCCCACACA G A I CAAGCCCACCC G A I CAAGCCCACCC G G C G A I CAAGCCCACCC G C C G C G C G C G C C CACCCACCACA 6460 G C C C CACCACCACA 6460 G C C C C C C C C C C C C C C	5870 N L L GAATTGTTC GAATTGTTC 6070 P A GCGCCCCCC 6070 C P A CCGGCCCTCC 6170 K L G CAACCTGCCC 6270 L D A TCCTCGCCC 6470 E L V GCAACTCGTC 6470 C L D A CCCCCCCCC 6570 G S A CCCCCCCCCC 6670 G S A CCCCCCCCCC 6770 G S A CCCCCCCCCC 6770 G S A CCCCCCCCCC 6770 C S A CCCCCCCCCCC 6770 C S A CCCCCCCCCCCC 6770 C S A CCCCCCCCCCCCC 6770 C S A CCCCCCCCCCCCCCC 6770 C S A CCCCCCCCCCCCCCC 6770 C S A CCCCCCCCCCCCCCC 6770 C S A CCCCCCCCCCCCCCC 6770 C S A CCCCCCCCCCCCCCCC 6770 C S A CCCCCCCCCCCCCCCC 6770 C S A CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	5880 A P M I GCTCCATCA 580 A C A T G GGCCTCTCGC GG V CCCCCACA GG V CCCCCCACA GG V CCCCCCACA GG V CCCCCCACA GG V CCCCCCACA GG G V CCCCCCACA GG G CCCCCCCACA GG G CCCCCCCC	$\begin{array}{c} {\rm ATCCACATCG} \\ {\rm ATCCACATCG} \\ {\rm G} & {\rm R} & {\rm A} \\ {\rm TAGGGCCCCCC} \\ {\rm S990} \\ {\rm D} & {\rm L} \\ {\rm R} \\ {\rm A} \\ {\rm V} \\ {\rm G} \\ {\rm G} \\ {\rm CTGGCCTGGG} \\ {\rm G} \\ {\rm G} \\ {\rm CTGGCCTGGG} \\ {\rm G} \\ {\rm G} \\ {\rm CTGGCCTGGG} \\ {\rm G} \\ {\rm G} \\ {\rm CTGGCTGGGC} \\ {\rm G} \\ {\rm G} \\ {\rm G} \\ {\rm CTGGCTGGGC} \\ {\rm G} \\ {\rm $	CACCAAGACC 5900 H G L CATGGGGCTG 6000 G P R G 6200 L G S GGCTCGCCAG 6200 L G P GCTCGCCATC 6300 V L V P GTCTTAGTTC 6400 S Y A CCTCCTACTC 6500 L L V CCTCCTACTC 6500 L L V CCTCCTACTC 6600 CTCCCCACGC 6500 L L V CCTCCTACTC 6600 CTCCCCACCC 6500 CTCCCCACCC 6500 L A CCTCCCACCC 6500 L A 6500 CTCCCCCACCC 6500 CTCCCCCACCC 6500 CTCCCCCCCCCCCC 6500 CTCCCCCCCCCCCCCCC 6500 CTCCCCCCCCCCCCCCCCCC 6500 CTCCCCCCCCCCCCCCCC 6500 CTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
5810 V Y L TTGGTATTG 5910 G V L A GGCGTGGGGC 6010 V G A GCCTGGGGGCG G L G S GCCTTGGG G L G S GCCTTGGCTGGC G L G S GCCTTGGCTGGC C GCCTGCCCCA L G A CCTCGGCGCA G A A CCTCGGCGCA G A A GGAGCCGGCA C A GAGCCCGGCA C A CACTACCCC A R A S GCCCCGGCAT G C A CACTACCCC A C A C CACTACCCC C	$\begin{array}{r} 5820 \\ A & G \\ ccccccccccccccccccccccccccccccccc$	5830 V I M TTCTCATTAT 5930 V L SA G F T CGGCTTCACG 6030 V L SA GCCCTCTCAC 6030 N P L CACCCCCTCC 6330 N P L CACCCCCCC 6430 S L A TCTCCTGCC 6630 C A CACCCCCCC CACCCCCC CACCCCCCC 6630 C A CCCCCCCCC 6630 C A CCCCCCCCCC 6630 C A CCCCCCCCCC 6630 C A CCCCCCCCCC 6630 C A CCCCCCCCCCCC 6630 C A CCCCCCCCCCCCC 6630 C A CCCCCCCCCCCCCCC 6630 C A CCCCCCCCCCCCCCC 6630 C A CCCCCCCCCCCCCC 6630 C A CCCCCCCCCCCCCCCC 6630 C C C CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	5840 C F N GTGTTTCAAT 5940 L C F N 6040 L C F 6140 F N L GTCAATACGTT 6140 F N L GTCAATACGT 640 C GGCTCTCCCC 640 A S S CTCCCCCGGC 640 C S S L CTCCCCCCCA A D K V 6540 A V G 6540 A V G 6740 CTCACCGCCC 6640 A V G 6740 A TCACCCCCCC 6840 S A L A TCACCCTCCCC 6840 S A L A CTCACCGCCCCC 6840 C A C C 6840 C C C C C C 6840 C C C C C C C C C C C C 6840 C C C C C C C C C C C C C C C C 640 C C C C C C C C C C C C C C C C C C C	5850 TAGTGAAC 5950 V S T TAGTGCAAC CATAGGCGAC CATAGGCGAC G N Y GCCGTCAAAT 6250 G N Y GCCGTCAAAT 6350 V S P GGTCTCTCCG 6450 C P L GGGCCCCCT C 6550 G P L GGGCCCCCCT 6550 G P L GGGCCCCCCT 6550 G P L GGGCCCCCCT 6550 G A V C CATCCCCGG 6450 A V K V C CATCCCCGG 6450 A V C A C C 6550 C C C C 6550 C C C C C 6550 C C C C C C 6550 C C C C C C 6550 C C C C C C C 6550 C C C C C C C C 6550 C C C C C C C C C C C C C C C C C C C	5860 Y T V AGTACACTTG 5960 V A 0 GGTGCGCAA A L A P GGCGCCGGACG 6160 R E A 6220 G GCCGCGACG CAGAGAGCGCC CAGAGAGCGCC CAGAGAGCGCC CAGAGAGCCC CAGAGCCCCCC CAGAGCCCCCC CAGAGCCCCCCC CAGAGCCCCCCC CAGAGCCCCCCC CAGAGCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCCC CAGAGCCCCCCCCC CAGAGCCCCCCCCC CAGAGCCCCCCCCCC	5870 N L L GAATTTGTTG GAATTTGTTG GO CGCCCCCGGCG G P A GCGGCCTGCC G P A GCGGCCTGCC G270 T L A ACTTGGCTA G370 L D A TGCTGACCG G470 E L V GGAACTGCT G570 G S A CGCGCCGCC G670 G S A CGCGCCCGCC GCCACCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCC GCCACCCCCCC GCCACCCCCC GCCACCCCCCCCC GCCCCCCCCCC	5880 A P M I GCTCCATCA 580 A L A 6080 L L Y 6080 G V V TCGCCACAC 6380 G V V TCGCCACAC 6480 L A T G 6480 L L A 550 CCTCTCGCC 6480 L L A 550 A V 5 CCTCTCGCCA 6680 L N L CCCCCTCCCC 6580 A V S CCCTCCGCA 6680 L A S CCCCTCCCCC 6580 A C CCCCCCCCCC 6680 C C CCCCCCCCCC 6680 C C CCCCCCCCCCC 6680 C C CCCCCCCCCCC 6680 C C CCCCCCCCCCC 6680 C C CCCCCCCCCCCC 6680 C C C CCCCCCCCCCCCCCCC 6680 C C C C CCCCCCCCCCCCCCCCC 6680 C C C C C C C C C C C C C C C C C C C	$\begin{array}{c} {\rm ATCCACATGC} \\ {\rm ATCCACATGC} \\ {\rm G} & {\rm R} & {\rm A} \\ {\rm TAGGGCCCCC} \\ {\rm S990} \\ {\rm D} & {\rm L} \\ {\rm R} \\ {\rm G} \\ {\rm S990} \\ {\rm D} \\ {\rm L} \\ {\rm R} \\ {\rm CTGGCCTGGG} \\ {\rm G} \\ {\rm G} \\ {\rm G} \\ {\rm G} \\ {\rm CTGGCCTGGG} \\ {\rm G} \\ {\rm G} \\ {\rm G} \\ {\rm CTGGCCTGGG} \\ {\rm G} \\ {\rm G} \\ {\rm CTGGCTGGGC} \\ {\rm G} \\ {\rm G} \\ {\rm G} \\ {\rm G} \\ {\rm CCCCCTGTATCG} \\ {\rm G} \\$	CACCAAGACC 5900 H G L CATGGGGCTG 6000 G P R G GGCCCCGCGG 6100 L G S GGCCCGCGCG 6200 L G F GCTCGCCTTC 6300 V L V P GTCTTAGTTC 6400 S Y A CCTCCTACGCA CTCCTACGCC CCTCCGCAG T L A H ACCCTCCACGCC 6600 F G P CTCTCGCGCG 6800 R D P Q AGGGACCCCC 7000
5810 V Y L TTGGTATTG 5910 G V L A GCGTGGGGGC 6010 V G A GCGTGGGGGG G L G S GCGTTGGGT G L G S GCGTTGGGT G L G S CGCGTGCGCT G L G A CCTCGGCGA C A CCTCGGCGA G A A CCTCGGCGCA C A GAGCCGCGA C A S GCGCTGCCCT G A A CCTCGCGCGA C A S CCCCGGCAT G C A C A C A C	5820 A A G F F GCCCGCGCT 59202 E A V CAGAGGCCT G L S Å CCTCTCGGC G V L Y GGAGTGCTCT 6220 A V A CCCCCTGCAT 6220 V P L L CTCCTCGAT 6220 V P L L CTCCTCGAT 6220 V P L L GACCGCCTTCC C GCCCCTCCAT 6220 V A A A CCCCCTTCAT 6220 C CCCCTCCAT 6220 C L V D CTCCTCCAT 6220 C L C C C C C C C C C C C G C C C C C G C C C C C G C C C C C G C C C C	5830 V I M TTCTCATTAT 5930 G P T CGGCTCACCG 6030 V L S A GCCCTCCAC 6230 N P L CACCCCACC 6230 S L A TCTCCTGGC CACCCCCC CACCCCCC CACCCCCC G C A TCCCCACGC G C L A GCCCCCCCC G C L A CCCCCCCCC G C C CACCCCCCC G C C CACCCCCCC G C C C C C C C C C G C C C C C C G C C C C C C	5840 C F N GTGTTTCAAT 5940 L C F N 6040 L C F 6140 F N L GTCAATCGT 6140 F N L GTCAATCGC 6340 L S C GGGTCTCCCC 6340 A D K Y GGCTCACCAGC 640 A V G 640 A V G C A V G	5850 5950 V 57 TAGTGTCAAC 60500 I G A CATAGGGCA CATAGGGCA G N Y GCCGTCAAAT 62500 G N Y GCCGTCAAAT 62500 V S P 63500 V S P 66500 G P L GGGCCCCTCG 64500 H L L V CATCTCTTGG 64500 H L C CATCGCCCCCC 65500 A AGCCTCTCG 66500 A AGCCCTCTCG 66500 A A A CGCCCATCGC 66500 A A A CGCCCATCGC 66500 A A A CGCCCATCGC 66500 A A A CGCCCATCGC 66500 A A A CGCCCATCGC 66500 A A CGCCCATCGC 66500 C A CGCCCATCGC 66500 C A CGCCCATCGC 66500 C A CGCCATCGC 66500 C A CGCCATCGCC 66500 C A CGCCATCGC 66500 C A CGCCATCGCC 66500 C A C C C A C C C A C C C A C A	5860 Y T V AGTACACTTG 5960 V A 0 GGTGCGCAA A L A P GGCGCGCGCAG GGCCGGCAGCA CAGAGAGCCC CAGAGAGCCC CAGAGAGCCC CAGAGAGCCC CAGAGAGCCC CAGAGAGCCC CAGAGCCCCCC CAGAGCCCCCCC CAGAGCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCC CAGAGCCCCCCCCC CAGAGCCCCCCCCC CAGAGCCCCCCCCCC	5870 N L L GANTITETTE 5970 P A G G CCGCGCCGCC G P A G P A G CGGCCCCGC G CGCCCCCCC 6070 K L G CAACTGGCC 6170 K L A I ACTTIGGCTA 6370 L D A TGCTCGACCC 6470 Y L A L 6770 Y L A L 6670 J L Y CGAACTGCTC 6670 G S A CGCCACGCC GCCACCCCCC 6670 G A CCCTCCC 6670 G A CCCCTCC 6670 G A CCCCCCCC 6670 G A CCCCCCCCCC 6670 G A CCCCCCCCCC 6670 G A CCCCCCCCCCCC 6670 G A CCCCCCCCCCCC 6707 G A CCCCCCCCCCCCCC 6707 G C CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	5880 A P M I GCTCCGATCA 5980 A L A GAGCTCTGCC 6080 L L Y GCCTTCTGC 6180 L A T G CTCCCCCACA 6380 R P V CCCCTTCTCC 6380 R V V CCCCTTCTCC 6380 A V S CTCTCGCCTACA 6580 L N L TTTGACCTC 6680 L P N L TTTGACCTC 6680 C CCTCCGCCTACA 6680 C CCTCCGCCTACA 7080 CCCCCACACA 6680 CCCCCCCCCCC 6780 CCCCCCCCCCCCCC 6680 CCCCCCCCCCCCCCC 6780 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	ATCCACGATCC G R A TAGGGCGCC 5990 D L R TGCACTAACA 6090 L A V G CTGCGCGCC 6090 L A V G CTGCGCGTCGC 6090 L V S GCTCGCTCCTAC 6390 T L Y A ACGCTCTACC 6490 L Y P CCCCCCATCCC 6490 CCTCGCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCC 6490 CCTCCCCCCCCC 6490 CCTCCCCCCCCC 6490 CCTCCCCCCCC 6490 CCTCCCCCCCCCCC 6490 CCTCCCCCCCCCC 6490 CCTCCCCCCCCCCC 6490 CCTCCCCCCCCCC 6490 CCTCCCCCCCCCCC 6490 CCTCCCCCCCCCCCCC 6490 CCTCCCCCCCCCCCCCCCC 6490 CCTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCAAGACC 5900 H G L CATGGGCTG 6000 G P R G 6000 L C S GGCCCGCGCG 6100 L C S GGCTCGCCGCA 6200 L C S GGTCTGGCAC 6400 V L V P GCTCGGCAC 6400 L V CCTCCTACGTC 6400 L V CCTCCTACGC 6400 L V CCTCCTACGC 6400 L V CCTCCTACGC 6400 L V CCTCCTACGC 6400 L V CCTCCTACGC 6400 CCTCCGCACA 6700 CCTCCGCACA 6700 CCTCCGCCC 6800 L G G CCTCCGCGCC 6900 CCTCCGCGCC 6900 CCTCCGCGCC 7000 GTCCTCACGCA 7100

<u>Figure 2</u>: The DNA sequence corresponding to the rRNA primary transcript and the downstream ORF. The sequence of the 16S RNA was presented elsewhere [17] and the sequence of the 23S RNA is contained within the secondary structure shown in Fig. 5. The start of the primary transcript is indicated and a putative promoter sequence upstream it is boxed. The end of the transcript is also shown within a pyrimidine sequence. The amino acid sequence corresponding to the ORF sequence is included.

antisense strand and hybridized with total cellular RNA to reveal an band of about 4000 nucleotides (data not shown); we inferred, therefore, that the gene is expressed. The putative gene is preceded by an A-T rich sequences although none was unambiguously identified as a promoter (Fig. 2). Moreover, no Shine-Dalgarno sequence was detected. No homology was detected between this ORF and those flanking other archaebacterial rRNA operons (unpublished work).

#### Initiation and processing of the 16S RNA leader transcript

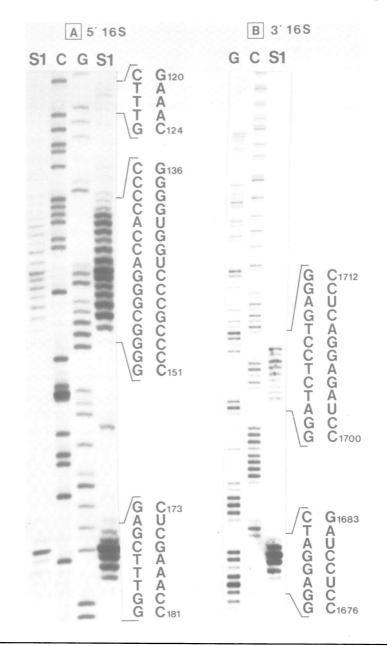
It has been shown previously by S1 nuclease mapping that initiation of rRNA transcription occurs at a single site 175 bp upstream from the 16S RNA gene [1]. A sequence comparison between the upstream region of this site and the upstream regions of tRNA primary transcripts revealed a common conserved A-T-rich sequence within a G-C-rich intercistronic spacer that could be a promoter motif [1]. This sequence, which is boxed in Fig. 2, differs from the putative promoter motifs found in the extreme halophiles [9, 11, 12] and methanogens [30].

The present results reveal how the leader sequence is processed down to the mature 16S RNA. The processing sites were monitored by S1 nuclease mapping using the 5'-end labelled restriction fragment 1 (Fig. 1). The results are illustrated on the autoradiogram in Fig. 3A and presented on a putative secondary structure of the leader and spacer sequences in Fig. 4. A series of strong cuts were detected from position 139 (Fig. 3A), located within a 3-nucleotide bulge in the putative processing stem D, and extending about 9 nucleotides downstream (Fig. 4). This probably reflects an initial endonuclease cut at the bulge followed by exonuclease trimming but it could also result from a pyrimidine-specific endonuclease activity since the cuts occur in helix D where C-152 could migrate along the helix as an unpaired nucleotide (Fig. 4). Fairly strong processing cuts were also observed after A-122 and U-163 which lie in internal and terminal loops, respectively (Fig. 4).

Longer exposures revealed several weak bands in the 16S RNA leader sequence (Fig. 4). They probably result from unspecific degradation of the transcript and, thus, their exclusive occurrence in putative single-stranded regions lends support to the proposed secondary structure. Further support derives from the observation that all of the double helices in Fig. 4 are highly G-C-rich while the interhelical regions are A- and U-rich. One major 5'-end of the 16S RNA was detected (Figs. 3A and 4).

### Processing of the 16S-23S RNA spacer

The spacer between the 16S and 23S RNA genes is only about 60 bp long and its sequence is complementary to those bordering both RNA genes; it is completely involved in the 16S and 23S RNA processing stems D and E (Fig. 4). Continuous transcription through the spacer was demonstrated by hybridizing total cellular RNA to the anti-sense strand of both 5'- and 3'-end labelled restriction fragments covering the entire spacer (fragments 2 and 3 in Fig. 1); the DNA fragments were protected against S1 nuclease digestion by continuous RNA transcripts (results not shown). Evidence for partially protected DNA fragments was also observed and autoradiograms showing these cuts are given in Figs. 3B and C; the cuts are superimposed on the putative secondary structure in Fig. 4. 5'- and 3'-end labelled restriction fragments yielded different sets of data. 5'-end labelled fragments 2a and 2b (Fig. 1) revealed major cuts in the sequences



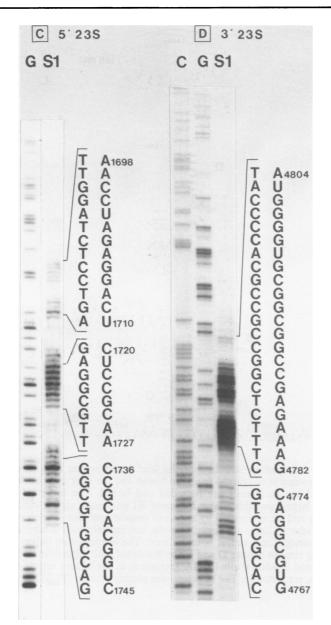


Figure 3: Autoradiograms showing examples of S1 nuclease mapping of the limits, and the processing sites, for the 16S and 23S RNAs. The limits of the 16S RNA were not determined experimentally earlier [17]. Experiments A to D were performed using fragments 1, 3b, 2a and 4, respectively, (see Fig. 1). The complementary RNA sequence is shown alongside the DNA sequence. C and G denote DNA sequencing tracks (Maxam-Gilbert). A correction was made for the one nucleotide difference between the S1 nuclease tracks and the DNA sequencing tracks.

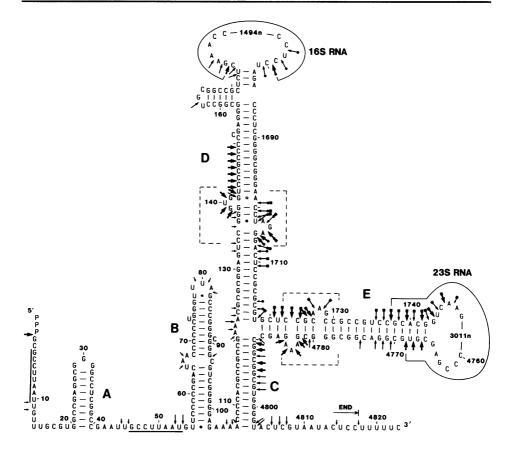


Figure 4: Putative secondary structure of the leader and spacer regions of the primary transcript. The sequence is numbered from the 5'-end and the structures of the 16S and 23S RNAs are omitted but their ends are included and boxed. The limits of the 23S RNA correspond to the largest possible major fragment. S1 nuclease cuts are classified as weak, medium and strong and are designated by arrows of proportional size. The secondary structure of the 16S RNA leader region is supported by weak cuts concentrated in unstructured regions, and an eight nucleotide direct repeat is underlined. The 3-nucleotide bulges that constitute primary processing sites in helices D and E are bracketed with stipled lines. Processing cuts detected with 5'- and 3'-end labelled DNA fragments in the spacer region are denoted by squares and circles, respectively. The structure was plotted using a "PLSTRUC" program (N. Larsen, unpublished).

C-1700 to G-1707, U-1721 to A-1728 and G-1738 to C-1745; these correspond to the 3-nucleotide loops in both 16S and 23S RNA processing stems and to the 5'-ends of mature 23S RNA; the heterogeneity at the 5'-end of 23S RNA was confirmed by primer extension experiments using reverse transcriptase which yielded essentially the same pattern of cuts (results not shown). 3'-end labelled fragments 3a and 3b (Fig. 1) revealed strong processing

cuts after the sequences C-1677 to C-1680 and A-1699 to A-1708 which correspond to the 3'-ends of mature 16S RNA (with a main end at C-1679) and the 3-nucleotide loop in processing stem D, respectively. These differences between the results obtained with 5'- and 3'-end labelled restriction fragments yield important information on the order of processing of the RNA transcript which is considered further in the Discussion.

Primary and secondary structure of the 23S rRNA

The 23S RNA contains approximately 3031 nucleotides and is larger than most archaebacterial 23S RNAs but shorter by about 46 nucleotides than that of *D. mobilis* [2]. The secondary structural model presented in Fig. 5 is based on a recent comparative sequence study of 23S-like RNA sequences from archaebacteria, eubacteria and the eukaryotic cytoplasm [2]. The format corresponds to that of the refined Santa Cruz/Urbana model for eubacteria [31]. Like the other extreme thermophile RNAs, the *T. tenax* RNA exhibits a very high G-C base pair content (79.5%) reflecting its stability at high temperature. The G·U content (9.9%) is slightly higher than the A-U content (8.9%), perhaps revealing the necessity to maintain unstable regions in the RNA structure that participate in cooperative effects or protein binding.

Most aspects of rRNA sequence and secondary structure that are typically archaebacterial [2] are present in the *T. tenax* structure and the secondary structural elements that are exclusive to archaebacteria are encircled in Fig. 5. The terminal helix joining the heterogeneous ends of the molecule is the maximum size (see Fig. 4) and may not form in the free RNA (see Discussion). There is an exceptional sequence in the region associated with peptidyl transfer in domain V where three highly conserved nucleotides are altered only in *T. tenax;* they are indicated by asterisks in Fig. 5. Two of these alterations have been shown to correlate with chloramphenicol resistance in mitochondrial ribosomes [32 and references therein]. Termination and processing of the transcript downstream from the 23S RNA

The termination site of the primary transcript and the processing sites distal to the 23S RNA were determined using the 750 bp restriction fragment 4 (Fig. 1). It contained about 120 bp of the 23S RNA gene and 630 bp downstream from the gene. It was hybridized to total cellular RNA and only very weak protection of the whole DNA fragment was observed; this was comparable to that observed in a control reaction containing total *E. coli* tRNA and probably resulted from DNA-DNA rehybridization. It was inferred, therefore, that the primary transcript terminates within this fragment. No single strong site was observed but a series of bands occurred in the sequences C-4790 to G-4797 and G-4802 to U-4809 downstream from the 23S RNA (see Figs. 3D and 4). The most downstream S1 nuclease cut occurred after C-4817 which lies in the long pyrimidine sequence: 5'-CTCC<sub>4817</sub>TTTTTCTCCC-3'. There is no other  $T_5$  sequence in the whole primary transcript. We could not infer from the S1 nuclease mapping results whether the heterogeneity at the 3'-end of the primary transcript reflects partially unspecific termination or exonuclease activity.

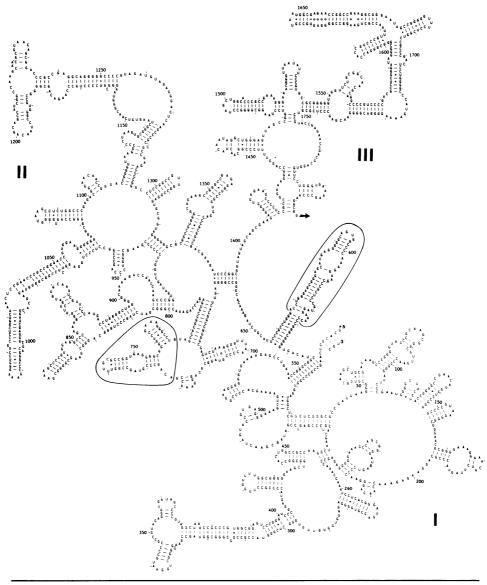
Several strong processing cuts occur between G-4782 and G-4786 which includes the

3-nucleotide loop in the 23S RNA processing stem E, in addition to the three strong cuts at C-4766 to U-4768 (Fig. 3D) which correspond to the 3'-ends of the mature 23S RNA.

# **DISCUSSION**

## Initiation signals: a promoter

A comparative sequence analysis of the 5'-flanking regions of the rRNA operon and four tRNA gene clusters from T. tenax revealed two common sequence features that could be



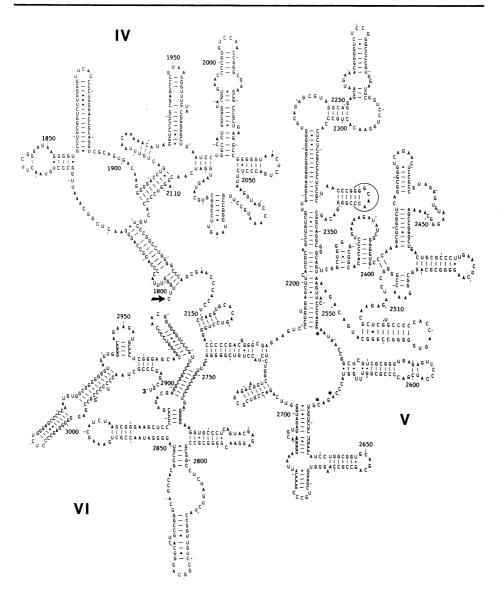


Figure 5: Putative secondary structure of the 23S RNA arranged as domains I to III (5'-half) and domains IV to VI (3'-half) based on a recently revised version of the Santa Cruz/Urbana model for *E. coli* [2]. Base pairs that are proven by sequence comparisons, or for which there is no negative evidence, are indicated by a line or dot between the nucleotides. Those which can form but for which there is conflicting evidence from sequence comparisons are juxtapositioned with no sign between them. Archaebacterial-specific features of the secondary structure are encircled. Three highly conserved nucleotides in the peptidyl transferase region of domain V that are only altered in *T. tenax* are denoted by asterisks. The structure was edited using an "EDSTRUC" program and plotted using a "PLSTRUC" program (N. Larsen, unpublished).

important for transcriptional initiation; one was at the site of transcription initiation and the other was upsteam and a putative promotor [1]. No homology was detected between the sequence at the initiation site and those found in *Methanococcus* and *Halobacteria* [11, 30]. Moreover, the conserved A-T-rich motif, and putative promoter, located about 18 bp upstream of the initiation site (Fig. 2) is absent from the corresponding sequences in other archaebacteria. This absence of common promoter and initiation motifs amongst the archaebacterial rRNA operons may relate to their adaptation to widely differing and extreme environments, and reflect the differing subunit structures of their RNA polymerases [33].

### Secondary structure of the leader and spacer region

The putative secondary structure presented in Fig. 4 reveals an unusually high degree of secondary structure within the leader and spacer regions. Within the leader sequence the G-C-rich helices A and B are interspersed by A- and U-rich regions. Helix B has a counterpart in another extreme thermophile *D. mobilis*, and the base pairing scheme is supported by coordinated base changes (J.K., unpublished); such a helix would be a candidate for involvement in antitermination as has been found in eubacteria [34]. Other eubacterial and archaebacterial operons exhibit the long processing stems but none, so far, including the extreme thermophile *D. mobilis* [9] involve all of their spacer sequence in processing stems. This high level of structuring may yield some insight into the function of the 16S-23S RNA spacer. The structure presented offers two alternative coaxially stacked forms: either helix D on E or helix E on C (and the additional possibility of a functionally important exchange between them). Both forms will result in a large separation of the RNAs such that they can assemble without mutual interference and this may reflect the main function of all 16S-23S-like RNA spacer regions. Helix C could also act, simultaneously, as a termination signal (see below). Processing signals and steps

It was proposed earlier that processing of the large rRNAs in *H. cutirubrum* and *M. vannielii* involved endonuclease cleavage of the primary transcript at or near the short discontinuities on opposite strands of the processing stems [12, 13, 15]. Both the processing stems and the short discontinuities are conserved in all archaebacterial RNA transcripts so far investigated [2 and references therein]. The processing sites may be comparable with the RNAse III recognition sites in eubacterial RNA precursors [35 and references therein]. Our results establish that a putative processing enzyme does, indeed, cut at the bulged loops that are staggered on opposite strands of both processing stems in *T. tenax* (Fig. 4). No conserved sequence motif is present at these sites but the purine content of the loops is generally high.

The S1 nuclease results suggest the following scheme of processing: Endonuclease digestion occurs at or near the 3-nucleotide loop on both strands of the 16S RNA processing stem. The 3'-end of the 16S RNA is then matured either by rapid exonuclease trimming, an endonuclease, or by a combination of both. The main 3'-end corresponds to C-1679 (Fig. 3B). This confirms an earlier estimate based on oligonucleotide analyses [36] and demonstrates that

the final cytidine of the putative Shine-Dalgarno sequence C-C-U-C-C is removed (Fig. 4). Processing of the 23S RNA appears to involve corresponding steps although here the heterogeneity of the mature ends suggests the involvement of an exonuclease. The high degree of heterogeneity contrasts with that found in both the 16S RNA of *T. tenax* and the 23S RNAs of other archaebacteria [2]. This result may be related to the exceptional fact that the terminal helix of the 23S RNA forms part of the processing stem in *T. tenax*. A consequence of these frayed ends, however, is that the terminal helix will exhibit a maximum of three base pairs (Fig. 5) and generally less (Fig. 4). This indicates, as for the 23S RNAs of *D. mobilis* and *M. thermoautotrophicum*, that it will generally be absent in the mature RNA [2].

By using both 5'- and 3'-end labelled DNA fragments covering the 16S-23S RNA spacer we were able to deduce the *order* of processing cuts. While cuts in the bulged loop on the 3'-side of the 16S RNA processing stem were detected using both 5'- and 3'-end labelled fragments those in the bulged loop of the 23S RNA processing stem were only detected using 5'-end labelled fragments (Fig. 4). This demonstrates that processing of the 16S RNA must occur prior to that of the 23S RNA. The S1 nuclease cuts obtained using 5'-end labelled fragments reveal, by a similar rationale, that processing of the 16S RNA occurs at the bulged loop prior to final maturation.

Finally, the S1 nuclease results (Fig. 3) show that many of the processing intermediates occur in high yield relative to the mature RNAs. To an unknown degree, the method gives an overestimate of these yields because DNA fragments preferentially select for larger RNA molecules. It is also likely, however, that the processing enzymes are less active at very high temperatures.

### Termination signal: a terminator

The longest detectable transcript terminates within a 13-pyrimidine sequence immediately

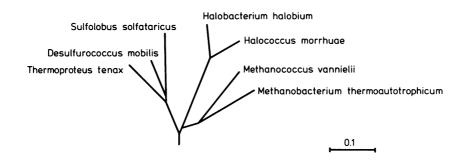


Figure 6: Unrooted phylogenetic tree for the archaebacteria derived from 23S ribosomal DNA sequences as described in Materials and Methods. The S. solfataricus 23S RNA sequence (C.R. Woese, pers. commun.) is included for comparison. The distance measure corresponds to 0.1 mutational events per sequence position.

downstream from helix C (Fig. 4). Other fragments resistant to S1 nuclease were detected with ends within this helix.

Our knowledge of other archaebacterial termination signals is limited. There is some similarity with a possible terminator in the rRNA operon of D. mobilis which consists of a T-rich sequence immediately following the 23S RNA processing stem [9]. On the other hand, although the primary transcripts of rRNA operons from both M. vannielii and M. thermoautotrophicum terminate within T-rich sequences they are followed by a short "hairpin" structure [16, 30] no such structure follows the T-rich sequence in T. tenax.

23S RNA structure and phylogenetic implications

The 23S RNA sequence was aligned with other available 23S-like RNA sequences on the basis of both conserved sequence regions and secondary structural elements as described by Leffers *et al.* [2]. Homology values and  $K_{nuc}$  values (the average number of mutational events per sequence position) were calculated. The latter give only an approximate measure of evolutionary distance [27] because variations occur in the evolutionary clocks (or mutational rates) of different species [37].

A phylogenetic tree was calculated from the  $K_{nuc}$  values for the archaebacteria and was rooted using a eubacterial 23S RNA (see Figure 6). It resembles the tree obtained with 16S RNA sequences [4] and demonstrates the three main subkingdoms of the archaebac-teria. A similar tree was obtained by considering only the unstructured regions of the 23S-like RNAs (see Fig. 5) thus avoiding artefacts that could result from the high G-C content of the extreme thermophiles. As discussed earlier [38] such trees need to be tested by com-parison with those based on other universal macromolecular sequences.

### **ACKNOWLEDGEMENTS**

We thank Niels Larsen for his unstinting assistance with the computing and Heidi Hummel for her advice on culturing *T. tenax* cells. J.K. and the research were supported by the Danish Natural Science Research Council. J.K. also received an EMBO travel fellowship. H.L. was granted a candidate stipend from the Carlsberg Foundation. Lisbeth Heilesen is thanked for helping with the manuscript.

\*To whom correspondence should be addressed

### **REFERENCES**

- 1. Wich, G., Leinfelder, W. and Böck, A. (1987) EMBO J. 6, 523-528.
- 2. Leffers, H., Kjems, J., Østergaard, L., Larsen, N. and Garrett, R.A. (1987) J. Mol. Biol. <u>194</u>, in press.
- Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R., Magrum, L., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Luehrsen, K.R. Chen, K.N. and Woese, C.R. (1980) Science 209, 457-463.

- 4. Woese, C.R. and Olsen, G.J. (1986). System. Appl. Microbiol. 7, 161-177.
- 5. Zillig, W., Schnabel, R. and Stetter, K.O. (1985) Curr. Top. Microbiol. and Immunol. 33, 1-18.
- 6. Lake, J.A., Hendersen, E., Oakes, M. and Clark, M.W., (1984) Proc. Natl. Acad. U.S.A. <u>81</u>, 3786-3790.
- 7. Lake, J.A. (1986) Nature <u>319</u>, 626.
- 8. Neumann, H., Gierl, A., Tu, J., Leibrock, J., Staiger, D. and Zillig, W. (1983) Mol. Gen. Genet. <u>192</u>, 66-72.
- 9. Larsen, N., Leffers, H., Kjems, J. and Garrett, R.A. (1986) System. Appl. Microbiol. 7, 49-57.
- 10. Kjems, J. and Garrett, R.A. (1985) Nature <u>318</u>, 675-677.
- 11. Mankin, A.S., Teterina, N.L., Rubtsov, P.M., Baratova, L.A. and Kagramanova, V.K. (1984) Nucl. Acids Res. <u>12</u>, 6537-6546.
- 12. Hui, I. and Dennis, P. (1985) J. Biol. Chem. 260, 899-906.
- 13. Chant, J. and Dennis, P. (1986) EMBO J. 5, 1091-1097.
- 14. Mankin, A.S. and Kagramanova, V.K. (1986) Mol. Gen. Genet. 202, 152-161.
- 15. Jarsch, M. and Böck, A. (1985) Mol. Gen. Genet. 200, 305-312.
- 16. Østergaard, L., Larsen, N., Leffers, H., Kjems, J. and Garrett, R.A. (1987). System. Appl. Microbiol. in press.
- 17. Leinfelder, W., Jarsch, M. and Böck, A. (1985) System. Appl. Microbiol. 6, 164-175.
- Zillig, W., Stetter, K.O., Schäfer, W., Janekovic, D., Wunderl, S., Holz, I. and Palm, P. (1981) Zbl. Bakt. Hyg. I. Abt. Orig. C2, 205-227.
- 19. Aiba, H., Adhya, S. and de Crombrugghe, B. (1981) J. Biol. Chem. 256, 11905-11910.
- 20. Lechner, K., Wich, G. and Böck, A. (1985) System. Appl. Microbiol. 6, 164-170.
- 21. Norrander, J., Kempe, T. and Messing, J. (1983) Gene 26, 101-114.
- 22. Sanger, F., Nicklen, S. and Coulsen, A.R. (1977) Proc. Natl. Acad. Sci. U.S.A. <u>74</u>, 5463-5467.
- 23. Ansorge, W. and Labeit, S. (1984) J. Biochem. Biophys. Methods 9, 33-47.
- 24. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A laboratory manual. Cold Spring Harbor Press, N.Y.
- 25. Favaloro, J., Treisman, R. and Kamen, R. (1980) Methods in Enzymol. 65, 718-723.
- 26. Maxam, A.M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 560-564.
- 27. Hori, H. and Osawa, S. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 381-385.
- 28. Felsenstein, J. (1978) J. Syst. Zool. 27, 401-410.
- 29. Southern, E.M. (1975) J. Mol. Biol. <u>98</u>, 503-517.
- 30. Wich, G., Hummel, H., Jarsch, M., Bär, U and Böck, A. (1986) Nucl. Acids Res. <u>14</u>, 2459-2479.
- 31. Noller, H.F. (1984) Ann. Rev. Biochem. 53, 119-162.
- 32. Ettayebi, M., Prasad, S.M. and Morgan, E.A. (1985) J. Bact. <u>162</u>, 551-557.
- 33. Gropp, F., Reiter, W.D., Sentenac, A., Zillig, W., Schnabel, R., Thomm, M. and Stetter, K.O. (1986) System. Appl. Microbiol. 7, 95-101.
- 34. Li, S.C., Squires, C.L. and Squires, C. (1984) Cell <u>38</u>, 851-860.
- 35. Brosius, J., Dull, T.J., Sleeter, D.D. and Noller, H.F. (1981) J. Mol. Biol. <u>148</u>, 107-127.
- 36. Woese, C.R., Gupta, R., Hahn, Ch..M., Zillig, W. and Tu, J. (1984) Syst. Appl. Microbiol. 5, 97-105.
- 37. Britten, R.J. (1986) Science 231, 1393-1398.
- 38. Woese, C.R., Pace, N.R. and Olsen, G.J. (1986) Nature 320, 401-402.