**Nucleic Acids Research**

Analysis of transcription in the archaebacterium *Sulfolobus* indicates that archaebacterial promoters are homologous to eukaryotic pol II promoters

Wolf-Dieter Reiter, Peter Palm and Wolfram Zillig

Max-Planck-Institut für Biochemie, D-8033 Martinsried, FRG

ABSTRACT

The 5'-termini were precisely mapped for five constitutive and one UV-inducible transcript from the *Sulfolobus* virus-like particle SSV1. The comparison of the DNA sequences around these transcriptional initiation sites revealed the presence of two conserved sequence elements: a trinucleotide sequence close to the initiation site itself and an AT-rich hexanucleotide sequence centered about 26 nucleotides upstream of it. Similar DNA sequences were found upstream of the transcriptional start sites for the ribosomal RNA genes in *Sulfolobus* and upstream of transcriptional start sites in other archaebacteria, allowing the derivation of a general consensus sequence for archaebacterial promoters. This consensus sequence is unlike that found in eubacteria but it resembles promoters recognized by eukaryotic RNA polymerase II.

INTRODUCTION

Since the discovery of the archaebacteria as a third major group of organisms besides eubacteria and eukaryotes (1,2) it has become increasingly clear that the study of their molecular biology can significantly contribute to an understanding of the early evolution of life. Though there are a few properties which are unique to archaebacteria, most features are shared either with eubacteria or with eukaryotes. Features which are common to archaebacteria and eubacteria include the prokaryotic cellular organization, the presence of Shine-Dalgarno sequences and the organization of genes in transcriptional units (for review see ref. 3). Features common to archaebacteria and eukaryotes include the presence of introns in stable RNA genes (4,5,6,7), the sequences and properties of translational elongation factors (8,9), the existence of aphidicolin-sensitive DNA polymerases (10,11) and the structure of the

DNA-dependent RNA polymerase (12). As far as transcription in archaebacteria is concerned, it is still an unresolved issue whether the signal structures involved in transcription initiation are shared with eubacteria or eukaryotes or whether they represent a unique feature. This uncertainty is largely due to the fact that there is only a modest number of precisely mapped 5'-ends of transcripts in archaebacteria and in many of these cases it is not clear whether these 5'-ends were due to transcription initiation.

In order to obtain more data on transcription signals in archaebacteria, we determined the transcriptional initiation sites for the ribosomal RNA genes and for protein-encoding genes in the extremely thermophilic archaebacterium *Sulfolobus* sp. strain B12. The transcript mapping data for the ribosomal RNA genes revealed a virtual identity of promoter sequences between *Sulfolobus* (13) and the methanogenic archaebacterium *Methanococcus vannielii* (14) although these two archaebacteria are only distantly related (15). To investigate transcription signals for protein-encoding genes in *Sulfolobus*, we chose the UV-inducible virus-like particle SSV1 (16) as an experimental system. The genome of SSV1 is a 15.5 kb circular double-stranded DNA (17) of known nucleotide sequence (P. Palm and B. Grampp, unpublished data). SSV1 is an ideal system for the study of "standard" *Sulfolobus* promoters since, with the exception of a UV-inducible RNA, all other RNAs are constitutively produced. The approximate map positions on the SSV1 genome have previously been determined for nine transcripts distinguishable by Northern analysis (18). Eight of these RNAs are constitutively expressed and one of them is UV-inducible. Since some of the SSV1 RNAs have common 5'-termini, the nine different transcripts correspond to only six different transcriptional initiation sites. For the overlapping transcripts T1 and T2 a coding function for structural proteins has been established (18,19). The RNAs T3 and T4 appear to be monocistronic mRNAs encoding proteins of 291 and 792 amino acids respectively. The role of the other SSV1 transcripts is not known but an involvement in DNA replication has been suggested for the UV-inducible RNA (18).

We have now precisely mapped the 5'-termini of all SSV1 transcripts by S1 endonuclease and primer extension analysis. By comparing the DNA sequences around these 5'-termini, two conserved sequence elements were identified, one of them being close to the transcriptional initiation site and the other one centered about 26 nucleotides upstream of it. A re-evaluation of published sequence data on transcriptional initiation sites in other archaebacteria revealed the presence of similar sequences allowing the derivation of a consensus sequence for archaebacterial promoters. This consensus sequence which applies both to stable RNA genes and protein-encoding genes, is unlike eubacterial promoters but it is strikingly similar to eukaryotic pol II promoters.

## MATERIALS AND METHODS

### Materials.

T4 polynucleotide kinase, S1 endonuclease and the Klenow fragment of *E. coli* DNA polymerase I were from Pharmacia and restriction endonucleases were obtained from BRL or from Boehringer Mannheim. Vaccinia virus guanylyltransferase and cloned M-MLV reverse transcriptase were from BRL. $[\gamma-^{32}P]ATP$ at 5000 Ci/mMol, $[\alpha-^{32}P]GTP$ at 410 Ci/mMol and $[^{35}S]dATP\alpha S$ at 800 Ci/mMol were obtained from Amersham.

### DNA sequence analysis.

DNA sequences were determined by the dideoxy chain termination method (20) using the M13 cloning and sequencing system (21,22). The general strategy used for the sequence determination of the SSV1 genome has been described previously (19). All DNA sequences were determined from both strands and from overlapping clones.

### Purification of *Sulfolobus* RNA.

Total RNA from UV-irradiated *Sulfolobus* sp. strain B12 was purified as described previously (19). For the mapping of transcript $T_{ind}$, RNA was purified from cells three hours after UV-irradiation. For all other mapping experiments, the RNA was from cells eight hours after UV-irradiation.

### S1 endonuclease mapping of 5'-ends.

Single-stranded 5'-$^{32}$P-labelled DNA probes were used for

the S1 mapping of the 5'-ends of SSV1 RNAs following the
general procedure described previously (18). In short, all
probes were obtained by annealing 5'-labelled primers to
single-stranded M13 (+)-DNA containing appropriate inserts,
followed by the extension of these primers by Klenow enzyme and
cleavage by a restriction enzyme. The purification of the
5'-labelled single-stranded DNA was achieved by electrophoresis
through polyacrylamide sequencing gels containing 7M urea. The
map positions of the oligonucleotide primers, of the M13 insert
DNAs and of the restriction enzyme cleavage sites are shown in
Fig. 1. The sequences of the primers and the lengths of all
probes are listed in Table 1. For the amounts of RNAs and
labelled DNA probes see Fig. 2 legend.

Identification of primary transcripts by in vitro capping.

For the in vitro capping of primary transcripts, 5 µg of
total *Sulfolobus* B12 RNA purified from cells 8 hours after UV
irradiation were incubated with vaccinia virus guanylyl-
transferase and $[\alpha-^{32}P]GTP$ using the conditions described
previously (13). To detect label incorporation into specific
SSV1 RNAs, aliquots of capped RNA were hybridized to unlabelled
single-stranded DNA probes and the resulting hybrids were
digested with S1 endonuclease. For the detection of capped
transcripts T1+T2 probe P4 was used and for the detection of
capped transcript T3 probe P8 was used (Fig. 1 and Table 1).
The DNA probes were obtained by the method described previously
(18) except that unlabelled rather than labelled oligo-
nucleotides were used for primer extension. 10 µg of
M13 (+)-DNA and 1 pMol of primer were annealed for the
synthesis of each probe. In order to localize the single-
stranded DNA after preparative polyacrylamide gel electro-
phoresis, a small amount of labelled DNA probe was run in
parallel and detected by autoradiography. One third of the
total amount of DNA probe was hybridized to 1 µg of capped RNA
and digested with S1 endonuclease using the conditions
specified previously (18). Another third of the DNA probe was
spotted on paper and subjected to long-term autoradiography to
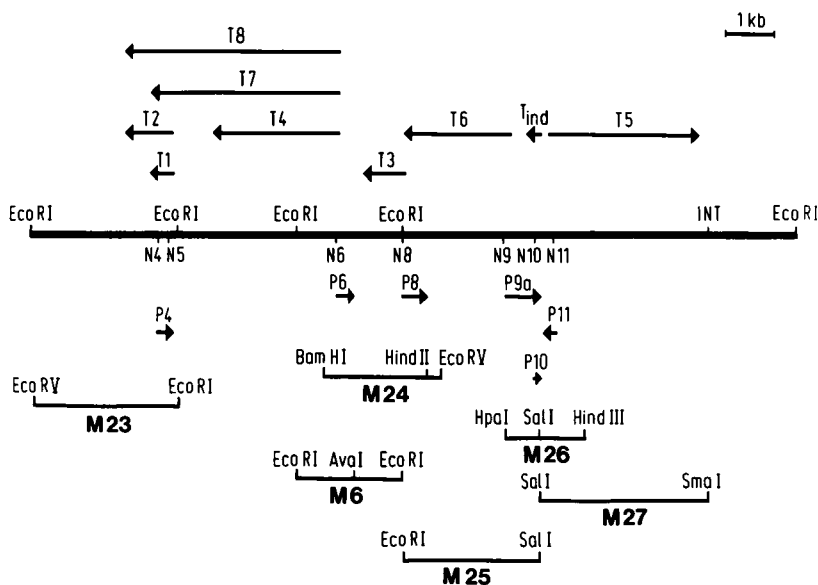check on a possible contamination with the radioactive size

Figure 1. Strategy used for the mapping of SSV1 transcripts.
    The upper part of the figure shows the map positions of all SSV1 transcripts analyzed. The location of the oligonucleotides and of the S1 probes used in this work are given below the linearized EcoRI map of SSV1. The bottom part of the figure shows the M13 clones used for the generation of single-stranded DNA probes. In addition to the flanking restriction sites of the M13 clones, only those restriction sites are shown which correspond to the 3'-ends of S1 probes. A list of all oligonucleotides and M13 clones used for the synthesis of S1 probes is given in Table 1.

Table 1:  Oligonucleotides, templates and restriction enzymes
            used for the synthesis of S1 probes

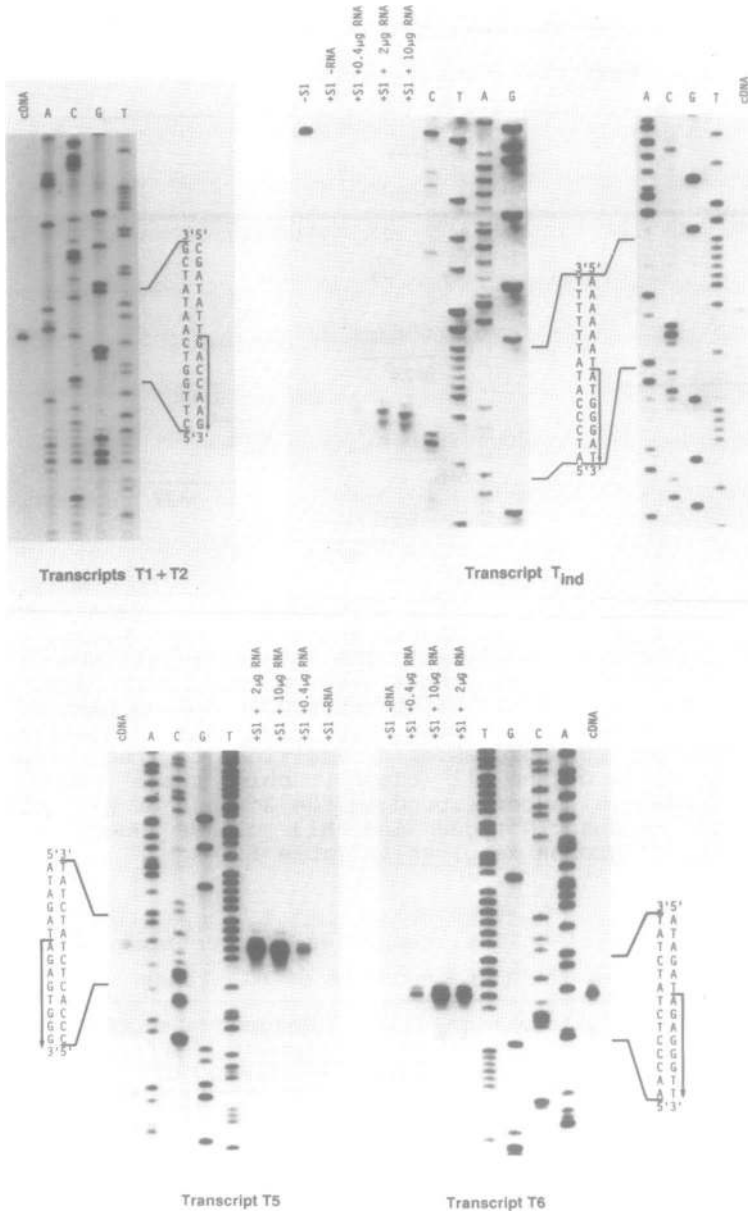| Probe | Oligonucleotide | Enzyme | Length (bases) | Template |
|---|---|---|---|---|
| P4 | N4;  d(TGGGGTGAAGTACCGCT) | EcoRI | 380 | M23 |
| P6 | N6;  d(TGTAAACTCCCGCACCA) | AvaI | 386 | M6 |
| P8 | N8;  d(GTGAACCGCCAGAATTC) | HindII | 501 | M24 |
| P9a | N9;  d(CCATACGGTTTCCTTCC) | SalI | 745 | M25 |
| P10 | N10;  d(TCTTTTACAACTCTCGT) | SalI | 114 | M26 |
| P11 | N11;  d(CGGGACTGAAATCTCAA) | SalI | 293 | M27 |

Figure 2. S1 endonuclease and primer extension mapping of SSV1 transcripts.
The mapping data for transcripts T1+T2, $T_{ind}$, T5 and T6 are shown. The amounts of *Sulfolobus* RNA used for S1 endonuclease mapping are given above the lanes. Lanes marked

"+S1 -RNA" show controls lacking *Sulfolobus* RNA and the lane marked "-S1" shows an undigested DNA probe. The amounts of S1 probes used for DNA-RNA-hybridizations were as follows: probe P9a: 10000 cpm, probe P10: 15000 cpm, probe P11: 10000 cpm. For the primer extension analysis of transcript T5, the whole reaction mixture was applied to the gel. In the other cases, only a fraction was applied, which was 1/400 for transcripts T1+T2, 1/25 for transcript T$_{ind}$ and 1/8 for transcript T6. The lanes marked G,A,T,C show sequencing reactions using the same oligonucleotides as for primer extension analysis and for S1 probe synthesis. For details about the oligonucleotide primers and the S1 probes see Fig. 1, Table 1 and the Materials and Methods section.

marker. Such a contamination could be excluded for both probes synthesized.

Primer extension analysis.

The mapping of 5'-ends of transcripts using 5'-labelled oligonucleotides as primers for reverse transcription was carried out as described previously (13). All reactions contained 0.1 pMol primer and 10 µg RNA. With the exception of oligonucleotide N5, all primers were identical to those used for the synthesis of single-stranded S1 probes. The map positions of all primers are shown in Fig. 1 and their sequences (except N5) are found in Table 1. The sequence of oligonucleotide N5 is d(CAGTCATCTGTTCTTCG).

Electrophoretic analysis of cDNAs and of S1-generated fragments.

All samples were analyzed on 6% polyacrylamide sequencing gels containing 7M urea. Sequencing reactions using the appropriate synthetic oligonucleotides as primers were co-electrophoresed as size standards. Since the DNA fragments of these sequencing reactions lacked a 5'-terminal phosphate, their electrophoretic mobility was usually decreased by about half a nucleotide compared to the S1-generated fragments and to the cDNAs.

RESULTS

Primer extension mapping of transcripts T1 and T2.

The overlapping SSV1 transcripts T1 and T2 encode structural proteins of the virus-like particles. The mapping of their common 5'-end by S1 endonuclease analysis has been

```
                                                                         ↓
T1+T2:   ACAAATTCAGATTCTGAATTCAGAACTGGAGGGGTTTAAAAACGTAAGCGGGAAGCCGATATTGACCAAGGATG


                                                                        ↓ •••••
T3:      CAAAAAGCATAATATCTGCGTTAGTTAGGCTCTTTTTAAAGTCTACCTTCTTTT.TCGCTTACAATGAGGAAGT
            |  | | |             |    || ||||||||| |      |  |||| |||||| |:|| | |
T4:      GCTGGATTAAACTTAAATTTAGAAGATAGCCCTTTTTAAAGCCATAAATTTTTTATCGCTT..AATGAAGTGGG
                                                                        ↓ ••• •••


T5:      GAGAGATAGATAGAGTATAGATAGAGTAAGACTTAAATACTAATTTATACATAGAGTATAGATAGAGTGGGATG
            ||  | ||||||||||||||||||||||||    |||||||| ||    ||| ||||||||||||||||| |
T6:      GATATATAGATAGAGTATAGATAGAGTAAAGTTTAAATACTTAT..ATAGATAGAGTATAGATAGAGGGGTTCAA
                                                                        ↑


                                                                        ↓•• • ••
T_ind:   ATATCTATATACACAAACAGAGTTAGTCGACTCTGTGTATCTTATGTATCTTATACAAAAAATATGGGATGTGC
```
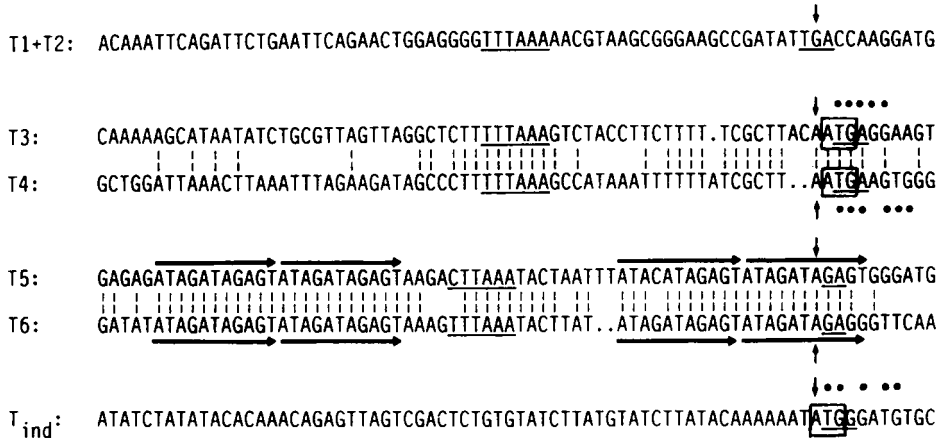
Figure 3. DNA sequences around SSV1 transcriptional initiation sites.
     Regions with an unusually high degree of homology, i.e. in the case of transcripts T3 and T4 and in the case of transcripts T5 and T6, have been aligned. Transcriptional initiation sites are indicated by arrows and the boxes A and B are underlined. Translational initiation codons belonging to open reading frames are boxed and the corresponding putative ribosome binding sites are indicated by closed circles. In the case of transcripts T5 and T6, an 11 base pair repeat sequence is marked by horizontal arrows (see text for details). Note that repeat 3 of transcript T5 and repeat 4 of transcript T6 contain a one-nucleotide mismatch.


reported previously (18). Primer extension analysis of transcripts T1+T2 using oligonucleotide N5 resulted in a single cDNA species of 119 nucleotides (Fig. 2) indicating that the 5'-terminus of both transcripts is the G residue 221 nucleotides upstream of the gene for the major coat protein VP1 (see ref. 19 for the nucleotide sequence of this region). This primer extension mapping differed by 1-2 nucleotides from the result obtained by S1 nuclease protection, a deviation which is within the error margin of the latter method. The exact positions of 5'-ends of transcripts were generally taken from the primer extension assay because of its higher precision.
Mapping of transcript T3.
     As reported earlier (18), the 0.9 kb transcript T3 hybridizes to two adjacent EcoRI-fragments of SSV1, its 5'-end being close to the EcoRI restriction site. For the precise

mapping of T3, the oligonucleotide N8 was chosen, which overlaps this EcoRI site (Fig. 1 and Table 1). S1 endonuclease mapping and primer extension analysis indicated that T3 starts on the first A residue of a sequence AATGA (Fig. 3). The ATG trinucleotide within this sequence is the initiation codon of a 261 amino acid open reading frame which spans the entire length of transcript T3. A pentanucleotide sequence ranging from position +3 to position +7 is perfectly complementary to the 3'-end of *Sulfolobus* B12 16S rRNA and may constitute a ribosome binding site (Fig. 3).

Mapping of transcript T4.

The 2.65 kb transcript T4 has previously been mapped by Northern analysis with an estimated uncertainty of about 200 nucleotides (18). S1 endonuclease mapping using probe P6 and primer extension analysis using oligonucleotide N6 indicated that transcript T4 starts on the first A residue within a sequence AATGA (Fig. 3), a situation already encountered in the case of transcript T3 (see above).

Also similar to the situation with transcript T3, a long open reading frame starts one nucleotide downstream from the 5'-terminus of transcript T4 (Fig. 3). This unassigned open reading frame could encode a protein of 792 amino acids and it spans almost the entire length of transcript T4. A hepta-nucleotide sequence ranging from position +3 to position +9 is almost perfectly complementary to the 3'-end of *Sulfolobus* B12 16S rRNA and could therefore constitute a ribosome binding site (Fig. 3).

Mapping of transcripts T5 and T6.

The SSV1 RNAs T5 and T6 are transcribed in opposite directions with a gap of about 0.8 kb between the map positions of their respective 5'-ends. The entire length of the UV-inducible RNA $T_{ind}$ maps within this gap (ref. 18; see Fig. 1). In initial S1 mapping experiments, probes were chosen that definitely overlapped the transcribed regions as determined by Northern analysis but the protected fragments were to large to achieve a one-nucleotide resolution (data not shown). To improve the precision of this transcript mapping, two additional oligonucleotides (N9 and N11) were selected
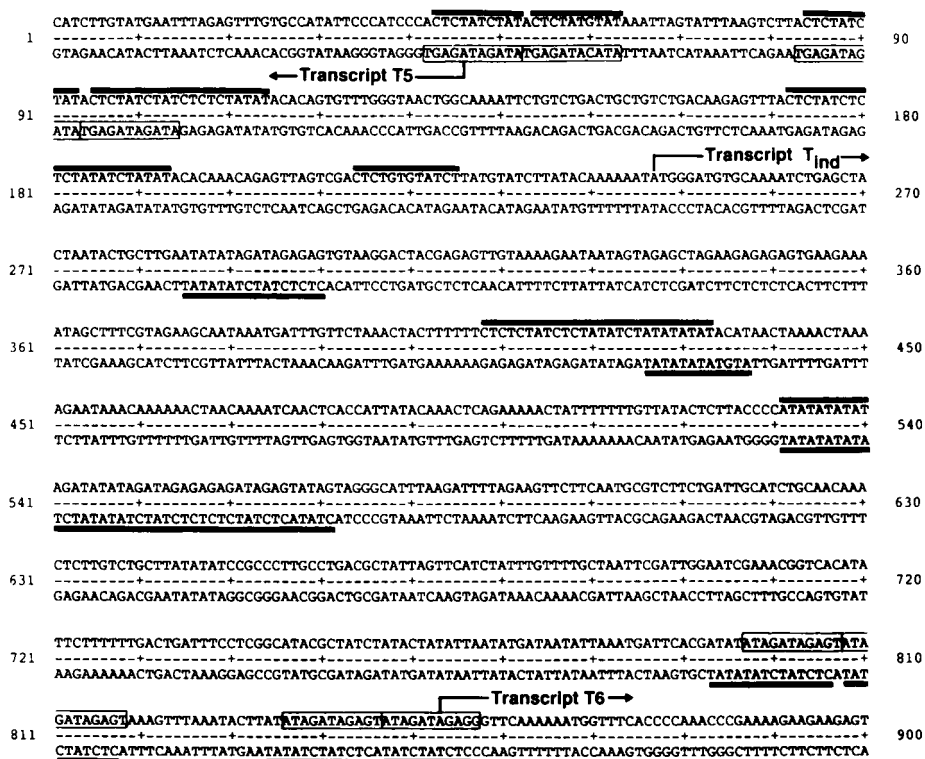
```
       CATCTTGTATGAATTTAGAGTTTGTGCCATATTCCCATCCCACTCTATCTATACTCTATGTATAAATTAGTATTTAAGTCTTACTCTATC
   1   ---------+---------+---------+---------+---------+---------+---------+---------+---------+      90
       GTAGAACATACTTAAATCTCAAACACGGTATAAGGGTAGGGTGAGATAGATATGAGATACATATTTAATCATAAATTCAGAATGAGATAG
                                       ←—Transcript T5——┘

       TATACTCTATCTATCTCTCTATATACACAGTGTTTGGGTAACTGGCAAAATTCTGTCTGACTGCTGTCTGACAAGAGTTTACTCTATCTC
   91  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     180
       ATATGAGATAGATAGAGAGATATATGTGTCACAAACCCATTGACCGTTTTAAGACAGACTGACGACAGACTGTTCTCAAATGAGATAGAG

                                                                          ┌——Transcript T_ind→
       TCTATATCTATATACACAAACAGAGTTAGTCGACTCTGTGTATCTTATGTATCTTATACAAAAAATATGGGATGTGCAAAATCTGAGCTA
  181  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     270
       AGATATAGATATATGTGTTTGTCTCAATCAGCTGAGACACATAGAATACATAGAATATGTTTTTTATACCCTACACGTTTTAGACTCGAT

       CTAATACTGCTTGAATATATAGATAGAGAGTGTAAGGACTACGAGAGTTGTAAAAGAATAATAGTAGAGCTAGAAGAGAGAGTGAAGAAA
  271  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     360
       GATTATGACGAACTTATATATCTATCTCTCACATTCCTGATGCTCTCAACATTTTCTTATTATCATCTCGATCTTCTCTCTCACTTCTTT

       ATAGCTTTCGTAGAAGCAATAAATGATTTGTTCTAAACTACTTTTTTCTCTCTATCTCTATATCTATATATATACATAACTAAAACTAAA
  361  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     450
       TATCGAAAGCATCTTCGTTATTTACTAAACAAGATTTGATGAAAAAAGAGAGATAGAGATATAGATATATATGTATTGATTTTGATTT

       AGAATAAACAAAAAACTAACAAAATCAACTCACCATTATACAAACTCAGAAAAACTATTTTTTTGTTATACTCTTACCCCATATATATAT
  451  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     540
       TCTTATTTGTTTTTTGATTGTTTTAGTTGAGTGGTAATATGTTTGAGTCTTTTTGATAAAAAAACAATATGAGAATGGGGTATATATATA

       AGATATATAGATAGAGAGAGATAGAGTATAGTAGGGCATTTAAGATTTTAGAAGTTCTTCAATGCGTCTTCTGATTGCATCTGCAACAAA
  541  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     630
       TCTATATATCTATCTCTCTATCTCATATCATCCCGTAAATTCTAAAATCTTCAAGAAGTTACGCAGAAGACTAACGTAGACGTTGTTT

       CTCTTGTCTGCTTATATATCCGCCCTTGCCTGACGCTATTAGTTCATCTATTTGTTTTGCTAATTCGATTGGAATCGAAACGGTCACATA
  631  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     720
       GAGAACAGACGAATATATAGGCGGGAACGGACTGCGATAATCAAGTAGATAAACAAAACGATTAAGCTAACCTTAGCTTTGCCAGTGTAT

       TTCTTTTTTGACTGATTTCCTCGGCATACGCTATCTATACTATATTAATATGATAATATTAAATGATTCACGATATATAGATAGAGTATA
  721  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     810
       AAGAAAAAACTGACTAAAGGAGCCGTATGCGATAGATATGATATAATTATACTATTATAATTTACTAAGTGCTATATATCTATCTCATAT

                                              ┌——Transcript T6→
       GATAGAGTAAAGTTTAAATACTTATATAGATAGAGTATAGATAGAGGGTTCAAAAAATGGTTTCACCCCAAACCCGAAAAGAAGAAGAGT
  811  ---------+---------+---------+---------+---------+---------+---------+---------+---------+     900
       CTATCTCATTTCAAATTTATGAATATATCTATCTCATATCTATCTCCCAAGTTTTTTACCAAAGTGGGGTTTGGGCTTTTCTTCTTCTCA
```

Figure 4. DNA sequence of the region containing transcripts T5, T6 and $T_{ind}$.
The eleven base pair repeat elements around the transcriptional initiation sites of T5 and T6 are boxed. DNA sequences of the type $(VT)_n$ with n>4 are indicated by bars above or below the sequence. Note that according to preliminary mapping data, $T_{ind}$ terminates around nucleotide 525.

which mapped closer to the 5'-ends of T5 and T6 (Fig. 1 and Table 1). Using these oligonucleotides, S1 endonuclease mapping and primer extension analysis indicated that transcripts T5 and T6 start at the same relative positions within an imperfect inverted repeat sequence of 64 nucleotides length (Fig. 2 and Fig. 3). Either copy of this inverted repeat sequence contains a directly repeated sequence of 22 nucleotides which is in turn composed of a tandem repeat of 11 nucleotides (Fig. 3 and Fig. 4). Within the region between the 5'-ends of T5 and T6 there were several long stretches of the sequence $(VT)_n$ on
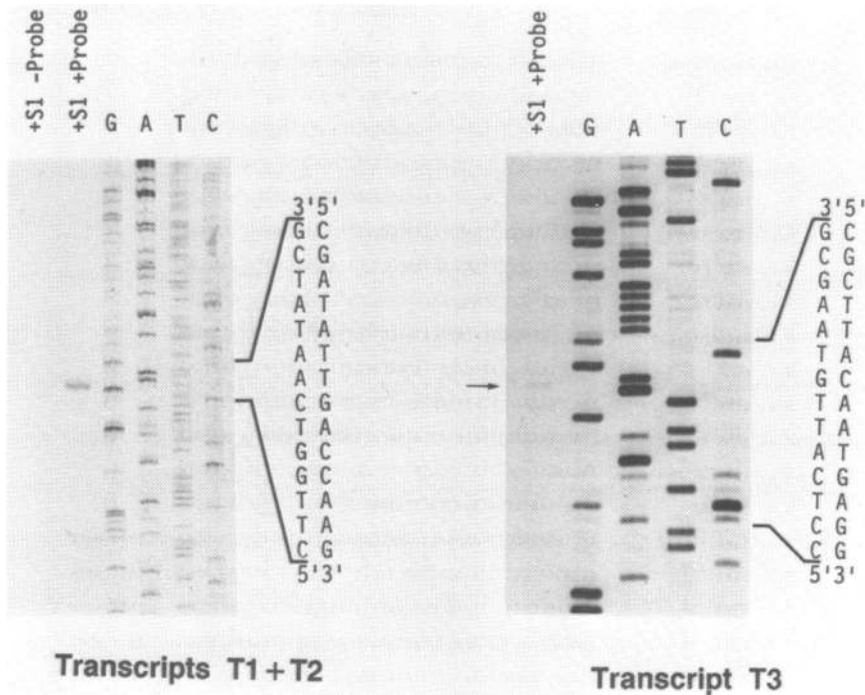
Figure 5. In vitro capping of SSV1 transcripts.

Total capped *Sulfolobus* RNA was hybridized to unlabelled single-stranded DNA probes and digested with S1 endonuclease. Protected capped RNA species are seen in the lanes marked "+S1 +Probe". The lane marked "+S1 -Probe" contains a control not hybridized to a DNA probe. The lanes marked G,A,T,C show the sequences of the DNA probes.

either strand of the DNA with V being A,C or rarely G (Fig. 4). No comparable accumulation of such sequences was found in other parts of the SSV1 genome. These $(VT)_n$-sequences were present within the long inverted repeat containing the start sites for T5 and T6 and they were also found within the DNA region covered by the UV-inducible transcript $T_{ind}$ (Fig. 4).

Mapping of the UV-inducible transcript $T_{ind}$.

Transcript $T_{ind}$ has previously been mapped by Northern analysis with an uncertainty of about 150 nucleotides (18). S1 endonuclease mapping using probe P10 and primer extension analysis using oligonucleotide N10 indicated that the 5'-end of $T_{ind}$ maps on the A residue within the sequence TATGG (Fig. 2

| GENE AND ORGANISM | SEQUENCE | REFERENCE |
|---|---|---|
| T.t. 16S/23S rRNA | AGCGAAAAATT TTTAAT TTAGGGTGTTTTAGGATGGTCGCGCCTTAATTG | 7 |
| T.t. tRNA^Ala | TAGCGAAAAAA TTTAAA TCGGTGAGTAAGTACGCTCGGGGCCGGTAGTCT | 7 |
| T.t. tRNA^Met | GACAAAAGCT TTTTAAA TTCGCGCAAAGCTTAGACCTAGCGGGGTAGGCC | 7 |
| H.c. rRNA P1 | TTCGACGGTGT TTATGT ACCCCACCACTCGGATGAGATGCGAACGACGT | 25 |
| H.c. rRNA P2 | AGTCCGATGCC CTTAAG TACAACAGGGTACTTCGGTGGAATGCGAACGAC | 25 |
| H.c. rRNA P3 | GATTCGATGCC CTTAAG TAATAACGGGTGTTCCGATGAGATGCGAACGAC | 25 |
| H.c. rRNA P4 | GATTCGATGCC CTTAAG TAATAACGGGCGTTACGAGGAATTGCGAACGAC | 25 |
| H.c. rRNA P5 | GATTCGATGCC CTTAAG TAATAACGGGGCGTTCGGGGAAATGCGAACGTC | 25 |
| H.h. Bacterio-opsin | TGGGTCGTAGA GTTACA CACATATCCTCGTTAGGTACTGTTGCATGTTGG | 27 |
| M.v. rRNA | ATACCTAAAAC AATACA TATTACAACACGTTTTCATATTATGCAAATCAA | 14 |
| M.v. tRNA^Arg | TAACCGAAATA TTTATA TACTAGAATACCCTTCCTATACTATGCTCTTGA | 14 |
| M.v. tRNA + 5S rRNA | CTACCGAAAAC TTTATA TATTATAAACACTAGTATTCAGTATGCGAACAAA | 14 |
| S.s. 16S/23S rRNA | TAGAAGTTAGA TTTATA TGGGATTTCAGAACAATATGTATAATGCGGATG | 13 |
| S.s. 5S rRNA | GTAGTTAATTT TTTATA TGTGTTATGAGTACTTAATTTTGCCCACCCGGC | 13 |
| S.s. SSV1, T1+T2 | AACTGGAGGGG TTTAAA ACGTAAGCGGGAAGCCGATATTGACCAAGGAT | this work |
| S.s. SSV1, T3 | GTTAGGCTCTT TTTAAA GTCTACCTTCTTTTTCGCTTACAATGAGGAAGT | this work |
| S.s. SSV1, T4 | AGATAGCCCT TTTTAAA GCCATAAATTTTTTATCGCTTAATGAAGTGGGG | this work |
| S.s. SSV1, T5 | ATAGAGTAAGA CTTAAA TACTAATTTATACATAGAGTATAGATAGAGTGG | this work |
| S.s. SSV1, T6 | ATAGAGTAAAG TTTAAA TACTTATATAGATAGAGTATAGATAGAGGGTTC | this work |

Consensus:                     TTTA A A                  A TG A C

Figure 6. Comparison of archaebacterial promoter sequences.
    This list contains the DNA sequences around the 5'-ends of
all mapped transcripts in archaebacteria except those cases
where the correspondence to transcription initiation appears
questionable (see text for details). The sequences are aligned
for the box A element. Transcriptional initiation sites are
indicated by arrows. Generally conserved sequence elements are
boxed or underlined. Abbreviations used for organisms: T.t.:
*Thermoproteus tenax*, H.h.: *Halobacterium halobium*, H.c.:
*Halobacterium cutirubrum*, M.v.: *Methanococcus vannielii*, S.s.:
*Sulfolobus* sp. strain B12.

and Fig. 3). The first three nucleotides of $T_{ind}$ corresponded
to the initiation codon ATG of a 49 amino acids open reading
frame which is the largest hypothetical protein this transcript
could encode. A heptanucleotide sequence on transcript $T_{ind}$
ranging from positions +2 to +8 may constitute a ribosome
binding site (Fig. 3).

In vitro capping of SSV1 transcripts.

To demonstrate that the mapped 5'-ends of SSV1 transcripts were due to transcription initiation rather than to RNA processing, we used in vitro capping in combination with an S1 nuclease protection assay. In order to avoid any problems with insufficient sensitivity, we confined the application of this method to the two most abundant transcripts, i.e. T1+T2 and T3. In the case of transcripts T1+T2 the protection of total capped RNA with probe P4 gave rise to an RNA fragment about 336 nucleotides long (Fig. 5), the expected size being 337 nucleotides. In case of transcript T3, the protection of total capped RNA by probe P8 resulted in an RNA fragment about 98-99 nucleotides long (Fig. 5), with an expected size of 95 nucleotides. Based on these results we conclude that the 5'-ends of transcripts T1+T2 and T3 were indeed due to transcription initiation.

Sequence comparison around the transcriptional initiation sites.

Comparing the DNA sequences around the 5'-termini of all SSV1 transcripts, two conserved sequence elements were found, one of them ("box B", consensus sequence TGA) being close to the transcriptional initiation site, the other one ("box A", consensus sequence TTTAAA) centered about 26 nucleotides upstream of box B (Fig. 3 and Fig. 6). Since these conserved sequences corresponded to transcription initiation, we consider them promoter elements. A perfect box B without mismatch was found in the case of transcripts T1+T2, T3 and T4. Transcription initiation occurred either on the central G residue of box B (T1+T2) or on an A residue two nucleotides upstream of box B (T3, T4). In the case of transcripts T5 and T6, transcription initiated on an A residue within an 11 bp repeat element (Fig. 3). No perfect box B sequence was found in the vicinity of the initiation sites of T5 and T6, the closest match being a trinucleotide sequence AGA containing the transcriptional start site on the first A residue (Fig. 3). In the case of the UV-inducible RNA $T_{ind}$, transcription initiation occurred at the A residue immediately upstream of a

trinucleotide sequence TGG (Fig. 3), a situation not found for any other SSV1 transcript. As far as box A is concerned, a perfect sequence without mismatch was found for all constitutive transcripts except T5, where a one-nucleotide deviation from the canonical sequence was observed (Fig. 3). The region upstream of the initiation site of the UV-inducible transcript $T_{ind}$ did not contain a box A sequence at the expected distance from the transcript start. This may be related to its special mode of expression.

A high degree of sequence conservation exceeding that due to the boxes A and B is evident for the upstream regions of transcripts T3 and T4 (Fig. 3). This homology is particularly pronounced around the box A element where a decanucleotide sequence is completely conserved (Fig. 3).


## DISCUSSION
### Possible role of transcripts T5, T6 and $T_{ind}$.

The SSV1 transcripts T5 and T6 start at the same relative positions within a long inverted repeat sequence and the entire length of the UV-inducible transcript $T_{ind}$ maps between these two transcriptional initiation sites. The whole region between the transcript starts of T5 and T6 contains stretches of peculiar DNA sequences resulting in a complicated pattern of direct and inverted sequence repeats. Regions of DNA containing such sequence repeats often constitute origins of replication (for review see ref. 23) and it is therefore an attractive hypothesis that the transcripts T5, T6 and $T_{ind}$ might be involved in the initiation of DNA replication. Clearly a detailed investigation of the mode of replication of the SSV1 genome is required to confirm or to dismiss this hypothesis.

### Comparison of promoter sequences for protein-encoding genes and for stable RNA genes in Sulfolobus.

As far as promoter sequences for stable RNA genes in Sulfolobus are concerned, transcript mapping data are available for the 5S rRNA gene and the 16S/23S rRNA gene cluster (13). A comparison between these promoters and the SSV1 promoters indicates a high degree of homology (Fig. 6). It is noteworthy, however, that in the case of the rRNA genes the box B sequence

is TGC whereas it is usually TGA for the SSV1 RNAs. Furthermore the box A sequence conserved between the ribosomal RNA genes is eight nucleotides long; its sequence is TTTATATG (Fig. 6). The heptanucleotide sequence CTTATAT which has been found about 52 nucleotides upstream of the box A elements of both rRNA genes (13) was not present in the case of the SSV1 promoters. This sequence may therefore constitute a promoter element which is specifically involved in the expression of the ribosomal RNA genes. In summary it appears that similar promoter elements are used for the expression of stable RNA genes and protein-encoding genes in *Sulfolobus* though the observed differences may be of functional significance. Clearly this conclusion is based on the assumption that the constitutive SSV1 promoters are representative of promoters for protein-encoding genes in *Sulfolobus*. Preliminary mapping data for two non-viral protein-encoding genes in *Sulfolobus* indicate that this is indeed the case (unpublished data from this laboratory).

Comparison of promoter sequences from different archaebacteria.

The comparison of the *Sulfolobus* promoter sequences with corresponding sequences from other archaebacteria lead to the result that the highest degree of similarity exists between *Sulfolobus* and the methanogenic archaebacterium *Methanococcus*. The promoter sequences for stable RNA genes are virtually identical for these two organisms (ref. 13; see Fig. 6). *Sulfolobus* and *Methanococcus* belong to two different phylogenetic branches of the archaebacteria (15,24) and the striking similarity of promoter sequences between these genera indicates that these sequences have remained essentially unchanged since the separation of the major archaebacterial lineages. Correspondingly one should expect that the promoter sequences in all archaebacteria are similar. In addition to *Methanococcus* and *Sulfolobus*, mapping data for stable RNA genes are available for *Halobacterium cutirubrum*, which belongs to the same branch as *Methanococcus* and for *Thermoproteus tenax*, which belongs to the same branch as *Sulfolobus*. In the case of *Halobacterium*, multiple promoters upstream of the rRNA gene cluster have been characterized (25). These promoters are very similar to each other and have probably arisen by multiplica-

tion events. They fit the general consensus sequence found in *Methanococcus* and *Sulfolobus*, though the box A sequence is slightly different (Fig. 6). In the case of *Thermoproteus tenax*, the 5'-ends have been mapped for the precursor transcripts of three stable RNAs (7). The corresponding upstream sequences also fit the consensus sequence of *Sulfolobus* and *Methanococcus* but the sequence conservation around the transcriptional initiation site is not very pronounced (Fig. 6). In summary, it appears that the conservation of promoter sequences within each of the phylogenetic branches is not as striking as that between the two distantly related archaebacteria *Sulfolobus* and *Methanococcus*.

As far as protein-encoding genes in archaebacteria other than *Sulfolobus* are concerned, almost no data on promoter sequences are available. Though there are several reports on the mapping of mRNAs in methanogens (26) and extreme halophiles (27,28,29,30), a correspondence of the mapped 5'-terminus to transcription initiation has only been demonstrated for the bacterio-opsin mRNA (27). In this case a box B sequence TGC is found with transcription initiating on the central G residue (Fig. 6) and a sequence element resembling box A is found at the expected distance from box B (Fig. 6). It must be born in mind, however, that the bacterio-opsin gene is subject to a complex mode of regulation (28) and that its promoter sequence may therefore be atypical. In the case of the other mapped 5'-termini of mRNAs in methanogens and extreme halophiles, we feel that the correspondence to transcription initiation must be demonstrated before sequence similarities can be discussed in detail.

In summary we conclude that in all those cases where primary transcripts have been mapped, the corresponding upstream sequences fit a general consensus sequence which is exemplified by the sequences found in *Sulfolobus* and *Methanococcus*. Because of insufficient data, however, it is not clear whether archaebacteria other than *Sulfolobus* use essentially the same expression signals for protein-encoding genes and for stable RNA genes.

Comparison of archaebacterial promoters to those in  eubacteria and eukaryotes.

The comparison of the archaebacterial  promoter  sequences with eubacterial promoters does not reveal common features.  In contrast  to  archaebacterial  promoters,  all  eubacterial promoters including those recognized by  secondary  holoenzymes consist of a conserved region around position -10 and a  second conserved region usually around position -35  (for  review  see ref. 31). There is a  striking  similarity,  however,  between archaebacterial  promoters  and  promoters  recognized  by eukaryotic RNA  polymerase II.  The  latter  type  of  promoter contains  a  weakly  conserved  dinucleotide  CA  around  the transcriptional initiation site and an AT-rich "TATA-box" motif centered  about  25-30  nucleotides  upstream  of  it  (32).  A somewhat  different  situation  is  encountered  for  pol II promoters in yeast, where the distance between the TATA-box and the  transcript  start  is  between  40  and  120  nucleotides (reviewed in ref. 33). With this exception, the spacing between conserved sequence elements and their positions relative to the transcriptional  initiation  site  are  the  same  for  archae-bacterial  promoters  and  eukaryotic  pol II  promoters. Furthermore, the sequence of the eukaryotic TATA-box (consensus TATAAA)  and  the  sequence  of  the  archaebacterial  box A (consensus  TTTA$^A/_T$A)  are  similar.  Recently  it  has  been demonstrated that a TATA-box is also required for the efficient expression  of  some  genes  transcribed  by  eukaryotic  RNA polymerase III (34,35), indicating that the  presence  of  this sequence element is not confined to eukaryotic pol II promoters but includes some pol III promoters as well. In most instances, however,  internal  promoter  sequences  are  sufficient  for faithful and efficient pol III transcription  (for  review  see ref. 36) and one may speculate that in these cases the TATA-box has become dispensable during evolution.

The  homology  between  archaebacterial  and  eukaryotic promoter sequences  is  paralleled  by  a  striking  similarity between archaebacterial and  eukaryotic  RNA  polymerases.  The earlier observation by Huet *et al.*  (37)  that  archaebacterial and eukaryotic RNA polymerases display similar subunit patterns

and strong immunological crossreactions has now been corroborated by the amino acid sequences of the large subunits of these enzymes (H. Leffers and G. Pühler, unpublished results). These sequence data indicate a high degree of homology between the archaebacterial and eukaryotic RNA polymerases whereas the homology between the archaebacterial and the eubacterial enzymes is considerably lower. Thus it appears that both the transcribing enzymes and their recognition sequences have been specifically conserved between eukaryotes and archaebacteria.

## REFERENCES

1. Woese, C.R. and Fox, G.E. (1977) Proc. Natl. Acad. Sci. USA 74, 5088-5090.
2. Woese, C.R., Magrum, L.J. and Fox, G.E. (1978) J. Mol. Evol. 11, 245-252.
3. Dennis, P.P. (1986) J. Bact. 168, 471-478.
4. Kaine, B.P., Gupta, R. and Woese, C.R. (1983) Proc. Natl. Acad. Sci. USA 80, 3309-3312
5. Daniels, C.J., Gupta, R. and Doolittle, W.F. (1985) J. Biol. Chem. 260, 3132-3134.
6. Kjems, J. and Garrett, R.A. (1985) Nature 318, 675-677.
7. Wich, G., Leinfelder, W. and Böck, A. (1987) EMBO J. 6, 523-528.
8. Kessel, M. and Klink, F. (1980) Nature 287, 250-251.
9. Lechner, K. and Böck, A. (1987) Mol. Gen. Genet. 208, 523-528.
10. Zabel, H.-P., Fischer, H., Holler, E. and Winter, J. (1985) System. Appl. Microbiol. 6, 111-118.
11. Prangishvili, D. (1986) Molekularnaya Biologiya 20, 477-488.
12. Zillig, W., Schnabel, R. and Stetter, K.O. (1985) Curr. Top. Microbiol. Immunol. 114, 1-18.
13. Reiter, W.-D., Palm, P., Voos, W., Kaniecki, J., Grampp, B., Schulz, W. and Zillig, W. (1987) Nucl. Acids Res. 15, 5581-5595.
14. Wich, G., Hummel, H., Jarsch, M., Bär, U. and Böck, A. (1986) Nucl. Acids Res. 14, 2459-2479.
15. Woese, C.R. and Olsen, G.J. (1986) System. Appl. Microbiol. 7, 161-177
16. Martin, A., Yeats, S., Janekovic, D., Reiter, W.-D., Aicher, W. and Zillig, W. (1984) EMBO J. 3, 2165-2168.
17. Yeats, S., McWilliam, P. and Zillig, W. (1982) EMBO J. 1, 1035-1038.
18. Reiter, W.-D., Palm, P., Yeats, S. and Zillig, W. (1987) Mol. Gen. Genet. 209, 270-275.

19. Reiter, W.-D., Palm, P., Henschen, A., Lottspeich, F., Zillig, W. and Grampp, B. (1987) Mol. Gen. Genet. 206, 144-153.
20. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
21. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980) J. Mol. Biol. 143, 161-178.
22. Messing, J. and Vieira, J. (1982) Gene 19, 269-276.
23. Scott, J.R. (1984) Microbiol. Rev. 48, 1-23.
24. Leffers, H., Kjems, J., Ostergaard, L., Larsen, N. and Garrett, R.A. (1987) J. Mol. Biol. 195, 43-61
25. Dennis, P.P. (1985) J. Mol. Biol. 186, 457-461.
26. Bollschweiler, C., Kühn, R. and Klein, A. (1985) EMBO J. 4, 805-809.
27. DasSarma, S., RajBhandary, U.L. and Khorana, H.G. (1984) Proc. Natl. Acad. Sci. USA 81, 125-129.
28. Betlach, M., Friedman, J., Boyer, H.W. and Pfeifer, F. (1984) Nucl. Acids Res. 12, 7949-7959.
29. Blanck, A. and Oesterhelt, D. (1987) EMBO J. 6, 265-273.
30. Lechner, J. and Sumper, M. (1987) J. Biol. Chem. 262, 9724-9729.
31. Doi, R.H. and Wang, L.-F. (1986) Microbiol. Rev. 50, 227-243.
32. Bucher, P. and Trifonov, E.N. (1986) Nucl. Acids Res. 14, 10009-10026.
33. Struhl, K. (1987) Cell 49, 295-297.
34. Tyler, B.M. (1987) J. Mol. Biol. 196, 801-811.
35. Murphy, S., Di Liegro, C. and Melli, M. (1987) Cell 51, 81-87.
36. Ciliberto, G., Castagnoli, L. and Cortese, R. (1983) Curr. Top. Dev. Biol. 18, 59-88.
37. Huet, J., Schnabel, R., Sentenac, A. and Zillig, W. (1983) EMBO J. 2, 1291-1294.