Extremophiles (2000) 4:175–179

ORIGINAL PAPER

Niels Tolstrup · Christoph W. Sensen · Roger A. Garrett
Ib Groth Clausen

Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus solfataricus*

Received: August 20, 1999 / Accepted: January 31, 2000

Abstract The translational starts of 144 *Sulfolobus solfataricus* genes have been determined by database comparison. Half the genes lie inside operons and the other half are at the start of an operon or single genes. A Shine–Dalgarno sequence is found upstream of the genes inside operons, but not for the first gene in an operon or isolated genes; this indicates that two different mechanisms are used for translation initiation in *S. solfataricus*. A box A transcriptional signal is found for the genes starting an operon or isolated genes, but not for the genes inside an operon. The box A signal is located about 27 nt upstream of the start codon, which implies that little or no upstream sequence is available for translation initiation for this group of genes. This finding is discussed.

Key words *Sulfolobus solfataricus* · Translation initiation · Shine–Dalgarno · Transcription · Crenarchaeota

Introduction

The nature of translation initiation in Archaea has been a matter of speculation ever since the discovery of this evolutionarily distinct group of organisms in 1977 (Woese and Fox 1977). With the publication of the genome of *Methanococcus jannaschii* (Bult et al. 1996), additional questions about the nature of translation initiation in Archaea arose from both a biomechanical and an evolutionary perspective.

The mechanism by which translation is initiated in Bacteria, Archaea, and Eucarya has long been considered to be functionally similar, but the components making up the machinery are quite different. Bacteria have three single subunit initiation factors – IF-1, IF-2, and IF-3 (Gualerzi and Pon 1990) – in which IF-3 and the 3′-end of the 16S rRNA bind to the Shine–Dalgarno sequence upstream of the start codon. Because archaeal mRNAs in many ways resemble those of Bacteria in being polycistronic, having Shine–Dalgarno sequences, and lacking polyA tails, it has been suggested that Archaea and Bacteria share common mechanisms of translation initiation (Keeling and Doolittle 1995). Eukaryotic translational initiation is different from that of Bacteria, involving a larger number of factors, sometimes containing multiple subunits (Pain 1996; Clark et al. 1996). A CAP structure (m^7GpppX, where X is any nucleotide) is appended to the 5′-end of the mRNA and facilitates ribosome binding of the mRNA. The start codon is subsequently found by a 5′- to 3′-scanning mechanism (for review, see Sonenberg and Pelletier 1989). With the publication of the *M. jannaschii* genome, it transpired that homologues of most eukaryotic translation factors except for those involved in mRNA CAP recognition are present in this organism (Bult et al. 1996). However, recent analyses have shown that more homology exists between the bacterial and archaeal/eukaryotic processes than was previously thought, leading to the suggestion that some parts of the translation initiation machinery are universally distributed (Kyrpides and Woese 1998).

From an evolutionary point of view, key questions are the evolution and emergence of the translation initiation process and particularly its relationship to the primary branching of the phylogenetic tree. The description of evolutionary patterns related to the position of genetic elements on the chromosome is sparse, but at least the conservation of operon structures in Bacteria and Archaea

Communicated by K. Horikoshi

N. Tolstrup (✉) · I.G. Clausen
Novo Nordisk A/S, Enzyme Research, Bioinformatics and DNA Sequencing, Novo Alle, DK-2880, Bagsværd, Denmark
Tel. +45-44-42-66-86; Fax +45-44-42-30-15
e-mail: ntol@novo.dk

C.W. Sensen
National Research Council Canada, Institute for Marine Biosciences, Halifax, Nova Scotia, Canada, B3H 3Z1

R.A. Garrett
Institute of Molecular Biology, Sølvgade, DK-1307 København K., Denmark

in contrast to eukaryotes defines an evolutionary difference (Woese 1998).

We have analyzed the emerging genomic sequence of *S. solfataricus* (Sensen et al. 1998), which belongs to the distinct archaeal taxon, the Crenarchaeota. This organism is the first genome of a crenarchaeote to be sequenced, and its completion will give important information about the phylogenetic and evolutionary differences between euryarchaeotes and crenarchaeotes. Here we present a statistical analysis of the regions upstream of the start codons of *S. solfataricus* genes.

Materials and methods

A set of good start codons

A subset of the *S. solfataricus* genomic sequences that had few sequencing errors was extracted for the analysis of gene start signals. The dataset from October 16, 1998, contained 14 contigs comprising 1346805 nucleotides with 44 ambiguities. From these sequences all open reading frames (ORFs) spanning at least 150 nucleotides from stop to stop codon were extracted from both the direct sequence and the reverse complement sequence. This step yielded 3453 ORFs comprising 1697756 nucleotides. These ORFs were searched against a database of 143171 selected proteins from SwissProt rel. 35 and TrEMBL rel. 7. The selection criteria for the search database were as follows. Take all sequences that start with an M, but exclude those that have any one of the keywords fragment, partial, incomplete, orf, insertion element, or transposon in their annotation, or come from *S. solfataricus* itself. All hits that had a blast (Altschul et al. 1990) score greater or equal to 80 were extracted (703), and realigned with the program align0 (Myers and Miller 1988) and the scoring matrix pam250. Align0 is a Needleman-Wunch alignment without end-gap penalties. All the alignments where the first amino acid in the sequence aligns to an amino acid inside the protein database sequence were discarded; this can happen when the start of the *S. solfataricus* sequence is missing, or if the start of the database sequence is not correctly annotated. After this step, 511 sequences were left. Only *S. solfataricus* sequences were selected, starting with ATG, GTG, or TTG and where the sequence identity of the first 60 amino acids in the alignment was higher than 25% (145). This set of sequences was run through the n2tool (Parsons et al. 1992) program with a threshold of 45, to remove identical and nearly identical sequences. All sequences were different, so none of the remaining 145 sequences were discarded in this step. The alignments on which the assignment of the start codons depended were investigated manually; 95% of the alignments looked very convincing, and the remaining 5% had insertions at the start of the alignment, thus making an alternative assignment of the start codon conceivable. One sequence had a match to a partial sequence and was discarded, while the remaining 144 genes were kept with their original assignments.

Signal analysis

Sequences were aligned with the annspec program (Workman and Stormo, 2000). This program is initialized with a random weight matrix. It aligns the sequences according to that matrix, and generates a new weight matrix from the alignment. This procedure is repeated until the weight matrix converges.

The information content of the aligned sequences is visualized using sequence logos (Schneider and Stephens 1990), which are based on Shannon's information measure (Shannon 1948; Hamming 1980). The height of each column i of the multiple alignment is the Shannon information $R(i)$, computed as

$$R(i) = 2.0 + \sum_{\alpha=A}^{T} P_i^{\alpha} \log_2\left(P_i^{\alpha}\right)$$

where P_i^{α} is the probability of finding nucleotide α, $\alpha \in \{A,C,G,T\}$ at position i in the alignment. In each column, the four nucleotide letters have heights corresponding to their frequency.

Results

The start codon

The 144 genes have an average length of 895 nt; 79% start with ATG, 11% with GTG, and 10% with TTG. These numbers are similar to earlier reports (Dalgaard and Garrett 1993). In 79% of cases the first possible start codon following an upstream stop codon is used, in 19% the second is used, and in 3% the third or fourth start codon. To get a further characterization of the start codons, the genes were divided according to whether they were in an operon or not.

Operons in *S. solfataricus*

Figure 1 shows the distance between the start or end of the 144 genes and the nearest upstream or downstream ORF. For 135 of the 288 gene ends, an ORF is found within a distance of 40 nt; 47% of these genes overlap, and of the overlapping genes, 52% overlap by four nucleotides, i.e., the last two nucleotides of the ATG of the downstream gene are the first two nucleotides of the stop codon TGA from the upstream gene. The close packing of half the *S. solfataricus* genes indicates that these are part of operons of coregulated genes. In the following we use 40 nt as the threshold for determining whether a gene is part of a putative operon; 73 sequences had an ORF longer than 150 nucleotides within a distance of 40 nucleotides from the start codon, and these genes are most likely inside an operon (42 genes) or at the end of an operon (31 genes). The remaining 71 genes had no ORF within this distance and are either isolated genes (38 genes) or the first gene of an operon (33 genes). These numbers show that 74% of the

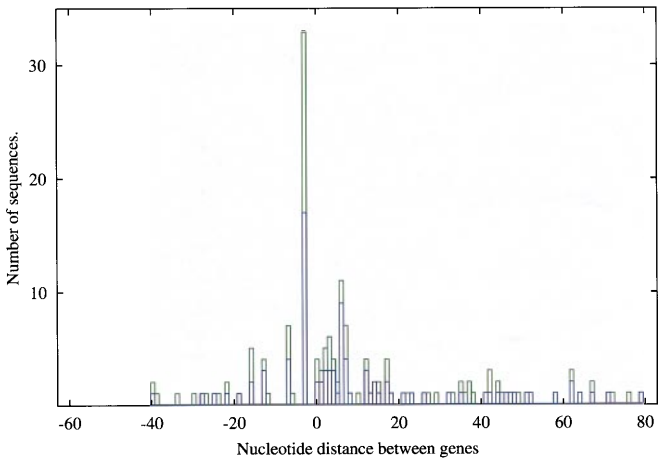

Fig. 1. A histogram of the distance between 144 *Sulfolobus solfataricus* genes and the nearest upstream and downstream ORF with a length longer than 150 nt. Data for upstream ORFs are shown with *blue bars* and downstream ORFs with *green bars*

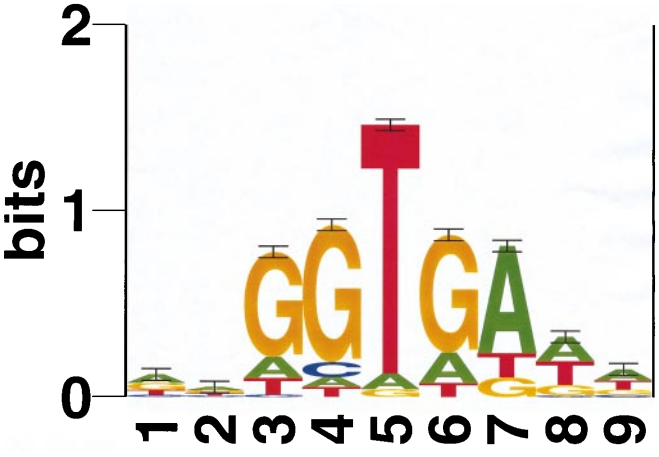

Fig. 3. A sequence logo of the putative SD region upstream from genes found inside an operon. The height of the letters gives the Shannon information

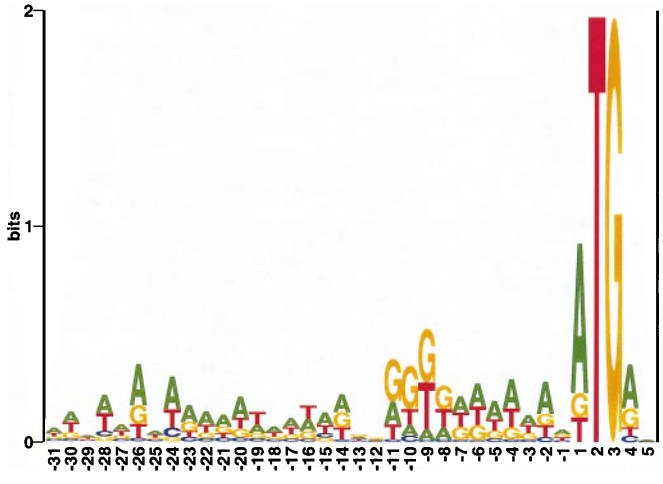

Fig. 2. A sequence logo of the 5′-region upstream from genes found inside an operon. The height of the letters gives the Shannon information

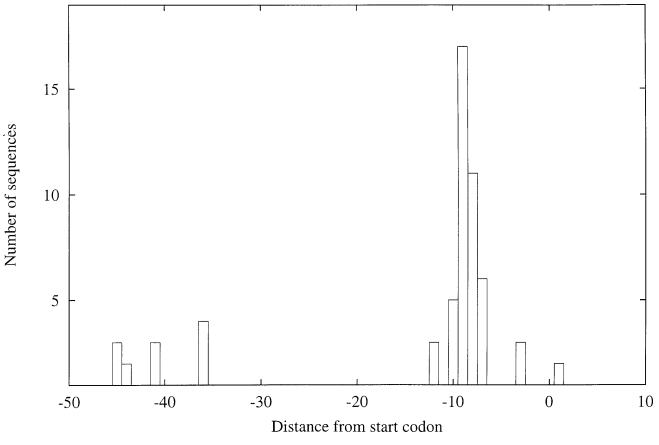

Fig. 4. A histogram of the positions of the Shine–Dalgarno sequences. The position is measured to the T of the GGTGA consensus sequence

genes are found in operons, and that the operons on average contain 3.3 genes.

Genes inside operons

Figure 2 shows a sequence logo of the genes that had an upstream ORF within 40 nt of their start codon. A G-rich region is located between 8 and 11 nucleotides upstream from the start codon. To enhance the signal, the first 15 upstream nucleotides of all sequences were aligned with the annspec program, and the highest scoring of 9 weight matrices generated by the program was used to align the sequences.

Figure 3 shows a sequence logo of the alignment of the G-rich upstream region. A consensus of GGTGA can be seen; a similar consensus was found by the other eight models, and also by alignments done with hidden Markov models (data not shown). This sequence is the reverse complement of a part of the 3′-end of 16S rRNA from *S. solfataricus* (GGAUCACCUCA-3′) (Trevisanato et al. 1996).

The 3′-end of the 16*S* rRNA has been shown to bind to the Shine–Dalgarno sequence of many prokaryotic organisms. The sequence found here presumably has a similar function and is thus best referred to as a Shine–Dalgarno sequence. The position of the Shine–Dalgarno sequences in the operon genes was determined by scoring each position from 50 nt upstream of the start codon to 10 nt downstream with the scoring matrix for the Shine–Dalgarno sequence. The highest scoring position indicates the most likely position of the Shine–Dalgarno sequence. Figure 4 shows the distribution; 55% of the Shine–Dalgarno sequences are located in the region −10 to −7, with a sharp peak at position −9. For 6% of the sequences the highest scoring signal was found downstream from the start codon. These results indicate that translation of genes inside

operons is initiated in a way similar to translation initiation in Bacteria.

Single genes and genes starting an operon

Figure 5 shows a sequence logo of genes with no upstream ORF within 40 nt of their start codon. For these genes no G-rich region is seen upstream from the start codon. The annspec program was used to search for signals in the 15 nt upstream region, but no consistent weight matrix was found. We also looked for sequences similar to the reverse complement of the 3′-end of the 16S rRNA but did not find any overrepresentation of these sequences. There is no indication of a Shine–Dalgarno sequence for this class of genes. Consequently, the mechanism of translation initiation is considered to be different.

Using the annspec program to analyze the region 35–20 upstream of the start codon, we were able to locate a T- and A-rich signal 30–23 nt upstream of the start codon. Two different weight matrices prevailed, one with a consensus sequence AAgTTTAAA and one with ttTATAa as the consensus. This signal is known as the box A signal (Zillig et al. 1988) and is essential for transcription efficiency and start site selection (Reiter et al. 1990; Qureshi et al. 1995). The distance between box A and the transcription start is about 27 nt (Reiter et al. 1990; Zillig et al. 1988; Dalgaard and Garrett 1993); this is very close to the distance found between the box A and the start codon in our analysis. For genes at the start of an operon or single genes there is little or no sequence upstream from the translation start. This finding is unexpected because it leaves little or no space for a ribosome-binding site, but it also explains why no Shine–Dalgarno sequence was seen in this dataset.

This observation needs to be verified experimentally. Table 1 shows experimentally determined transcription start sites for genes from crenarchaeotes and the start of the protein. The transcription start sites are found within 8 nt upstream of the start codon, and in three cases it is the A of the start codon; this leaves no room for an upstream Shine–Dalgarno sequence and confirms the statistical analysis. It further supports the hypothesis that the translational initiation mechanism for genes upstream of an operon is different from that of initiation of genes inside an operon.

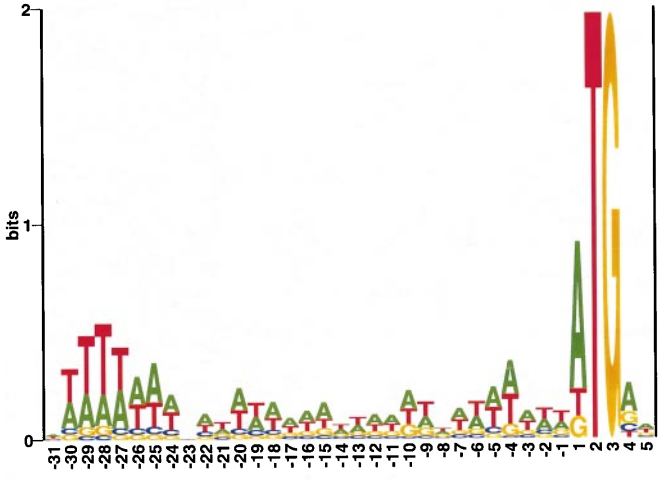

Fig. 5. A sequence logo of the 5′-region upstream from genes with no upstream ORF. The height of the letters gives the Shannon information

Discussion

For genes inside operons, the ribosome binds to the Shine–Dalgarno sequence by a mechanism similar to that used by bacteria. For the nonoperon genes, no Shine–Dalgarno sequence is found, so the ribosome must bind the mRNA by a different mechanism. The initial genes have a box A signal that is not found for the genes inside operons. The distance from the box A signal to the start codon is very close to the distance to the transcription start reported experimentally, and this implies that little or no sequence exists upstream from the start codon and explains why there is no upstream

Table 1. Promoter region

Organism and gene	Sequence	Accession number	Reference
S.s. topR	TAACTCTTTTAAAAGATCTTACCATCATAAAGTTTTTTTCTAGCTATGATT	X98420	Jaxel et al. (1996)
S.a. lrp	ATAAAAATTTTTTAACAGGTCCATATATTTATAATGTAGAAATATATGTCA	Y12821	Charlier et al. (1997)
A.a DNA-Ligase	GTTTTCATCATCAAATATATTATATCGTAATTCAAACTTTTATTTATGGAG	X63438	Kletzin (1992a)
S.a. SOD	TATTTCTAATCAAAATATTTAATCCACCATTGGCTATTTGTTTGTATGACC	X63386	Klenk et al. (1993)
A.a. SOR	AAAGAAAGAATATAAAAGTAGACAGAAATGTATATTTGACCAAAAATGCCG	X56616	Kletzin (1992b)
A.a. *sor*-ORF2	CTTATTTATACATTTGTTAAGGATGGGTTATATTATGCATCTCCCATGTTC	X56616	Kletzin (1992b)
A.a. *sor*-ORF4	TTTATCTCAACTATTTTTAAATACTAGTGCAAGAAAGAATAGCATATGAGT	X56616	Kletzin (1992b)
A.a. *sor*-ORF3	AAAAAAGAAGAATAGATTTATTTCCTGAATAAAGTAGATTAGATTATGGTC	X56616	Kletzin (1992b)

A.a., Acidianus ambivalens; S.a., S. acidocaldarius; S.s., S. shibatae
Experimentally determined transcription start sites are marked in bold; the box A and the start codon are underlined
The EMBL accession numbers are given in column three

Shine–Dalgarno sequence for these genes. The observation that there is little upstream sequence has been confirmed by primer extension analyses made for other crenarchaeotes (Jaxel et al. 1996; Charlier et al. 1997; Kletzin 1992a,b; Klenk et al. 1993). Thus, the ribosome binding works by a different mechanism for these genes.

How is translation initiated for initial or isolated genes? Eukaryotic ribosomes recognize a 5′-capped mRNA. Capping involves three steps: removal of the first phosphate by RNA triphosphatase, capping with RNA guanylyltransferase, and N7 methylation by (guanine-7)methyltransferase. The 40*S* subunit of the ribosome is then bound to the mRNA through the cytoplasmic CAP-binding complex eIF–4F (Varani 1997; Lewis and Izaurralde 1997). However, we could detect none of these proteins: eIF–4E CAP-binding protein, eIF–4F, RNA triphosphatase, RNA guanylyltransferase, or guanine-7 methyltransferase in the partial *S. solfataricus* genome sequence.

We propose, therefore, that unless the genes in question are located in the remaining part of the *S. solfataricus* genomics sequence, another signal could be involved that is downstream to the translation initiation codon, possibly even at the 3′-end of the mRNA rather than processing of the 5′-end by capping and transsplicing, or that some other event initiates translation for these genes. It is also possible that the position of the signal is too variable to be detected by the methods used. This question can only be resolved with more biochemical experiments.

The observations presented in this paper lead on naturally to other analyses that are required to answer two questions: (1) Do euryarchaeotes (e.g., *M. jannaschii*) show the same distribution of Shine–Dalgarno sequences over the chromosome? (2) What are the homologies of initiation factors compared to other Archaea, Bacteria, and eukaryotes? Although our findings at this stage do not lead to final conclusions about the ancient development of the translational initiation machinery, the results suggest that the crenarchaeote *S. solfataricus* has a more specialized initiation machinery than euryarchaeotes.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073

Charlier D, Roovers M, Thia-Toong TL, Durbecq V, Glansdorff N (1997) Cloning and identification of the *Sulfolobus solfataricus* lrp gene encoding an archaeal homologue of the eubacterial leucine-responsive global transcriptional regulator lrp. Gene (Amst) 201:63–68

Clark B, Grunberg-Manago M, Gupta NK, Hershey J, Hinnebusch AG, Jackson RJ, Maitra U, Mathews MB, Merrick WC, Rhoads RE, Sonenberg N, Spremulli LL, Trachsel H, Voorma HO (1996) Prokaryotic and eukaryotic translation factors. Biochimie (Paris) 78:1119–1122

Dalgaard JZ, Garrett RA (1993) Archaeal hyperthermophile genes. In: Kates M, Kusher DJ, Matheson AT (eds) The biochemistry of Archaea. Elsevier, Amsterdam, pp 535–563

Gualerzi CO, Pon CL (1990) Initiation of messenger RNA translation in prokaryotes. Biochemistry 29:5881–5888

Hamming RW (1980) Coding and information theory. Prentice-Hall, Englewood Cliffs, NJ

Jaxel C, de la Tour B, C, D, M, Nadal M (1996) Reverse gyrase gene from *Sulfolobus shibatae* B12: gene structure, transcription unit and comparative sequence analysis of the two domains. Nucleic Acids Res 24:4668–4675

Keeling PJ, Doolittle WF (1995) Archaea – narrowing the gap between prokaryotes and eukaryotes. Proc Natl Acad Sci USA 92:5761–5764

Klenk HP, Schleper C, Schwass V, Brudler R (1993) Nucleotide sequence, transcription and phylogeny of the gene encoding the superoxide dismutase of *Sulfolobus acidocaldarius*. Biochim Biophys Acta 1174:95–98

Kletzin A (1992a) Molecular characterization of a DNA ligase gene of the extremely thermophilic archaeon *Desulfurolobus ambivalens* shows close phylogenetic relationship to eukaryotic ligases. Nucleic Acids Res 20:5389–5396

Kletzin A (1992b) Molecular characterization of the sor gene, which encodes the sulfur oxygenase/reductase of the thermoacidophilic archaeum *Desulfurolobus ambivalens*. J Bacteriol 174:5854–5859

Kyrpides NC, Woese CR (1998) Archael translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2B alpha-beta-delta subunit families. Proc Natl Acad Sci USA 95:3726–3730

Lewis JD, Izaurralde E (1997) The role of the cap structure in RNA processing and nuclear export. Eur J Biochem 247:461–469

Myers E, Miller W (1988) Optimal alignments in linear space. CABIOS 4:11–17

Pain VM (1996) Initiation of protein synthesis in eukaryotic cells. Eur J Biochem 236:747–771

Parsons JD, Brenner S, Bishop MJ (1992) Clusterins cDNA sequences. Comput Appl Biosci 8:461–466

Qureshi SA, Baumann P, Rowlands T, Khoo B, Jackson SP (1995) Cloning and functional analysis of the TATA binding protein from *Sulfolobus shibatae*. Nucleic Acids Res 23:1775–1781

Reiter W, Depohl UH, Zillig W (1990) Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection in vitro. Proc Natl Acad Sci USA 87:9509–9513

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18:6097–6100

Sensen CW, Charlebois RL, Chow C, Clausen I, Curtis B, Doolittle WF, Duguet M, Erauso G, Gaasterland T, Garrett RA, Gordon P, de Jong IH, Jeffries AC, Kozera C, Medina N, De Moors A, van der Oost J, Phan H, Ragan MA, Schenk ME, She Q, Singh RK, Tolstrup N (1998) Completing the sequence of the *Sulfolobus solfataricus* p2 genome. Extremophiles 2:305–312

Shannon CE (1948) A mathematical theory of communication. I. Bell System Technol J 27:379–423/623–656

Sonenberg N, Pelletier J (1989) Poliovirus translation: a paradigm for a novel initiation mechanism. BioEssays 11:128–132

Trevisanato SI, Larsen N, Segerere AH, Stetter KO, Garrett RA (1996) Phylogenetic analysis of the archaeal order of *Sulfolobales* based on sequences of 23S rRNA genes and 16S/23S rDNA spacers. Syst Appl Microbiol 19:61–65

Varani G (1997) A cap for all occasions. Structure (Lond) 5:855–858

Woese CR (1998) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Workman C, Stormo GD (2000) ANN-Spec: A method for discovering transcription factor binding sites with improved speciticity. In: Altman R, Hunter L (eds) Pacific symposium on biocomputing 2000. World Science, Singapore, pp 467–478

Zillig W, Palm P, Reiter W, Gropp F, Pühler G, Klenk H (1988) Comparative evaluation of gene expression in archaebacteria. Eur J Biochem 173:473–482