doi:10.1016/j.jmb.2006.11.050 J. Mol. Biol. (2007) 366, 420–435

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Temporal Regulation of Viral Transcription during Development of *Thermus thermophilus* Bacteriophage ϕYS40

Anastasiya Sevostyanova[1], Marko Djordjevic[2], Konstantin Kuznedelov[3] Tatyana Naryshkina[3], Mikhail S. Gelfand[4,5] Konstantin Severinov[1,3,6]* and Leonid Minakhin[3]*

[1]*Institute of Molecular Genetics Russian Academy of Sciences Moscow 123182, Russia*

[2]*Mathematical Biosciences Institute, The Ohio State University, Columbus OH 43210, USA*

[3]*Waksman Institute for Microbiology, Rutgers the State University of New Jersey, Piscataway NJ 08854, USA*

[4]*Institute for Information Transmission Problems Russian Academy of Sciences Moscow 127994, Russia*

[5]*Faculty of Bioengineering and Bioinformatics Lomonosov Moscow State University, Moscow 119991, Russia*

[6]*Department of Molecular Biology and Biochemistry Rutgers, the State University of New Jersey, Piscataway NJ 08854, USA*

**Corresponding authors*

Regulation of gene expression of lytic bacteriophage ϕYS40 that infects the thermophilic bacterium *Thermus thermophilus* was investigated and three temporal classes of phage genes, early, middle, and late, were revealed. ϕYS40 does not encode a (RNAP) and must rely on host RNAP for transcription of its genes. Bioinformatic analysis using a model of *Thermus* promoters predicted 43 putative σ^A-dependent −10/−35 class phage promoters. A randomly chosen subset of those promoters was shown to be functional *in vivo* and *in vitro* and to belong to the early temporal class. Macroarray analysis, primer extension, and bioinformatic predictions identified 36 viral middle and late promoters. These promoters have a single common consensus element, which resembles host σ^A RNAP holoenzyme –10 promoter consensus element sequence. The mechanism responsible for the temporal control of the three classes of promoters remains unknown, since host σ^A RNAP holoenzyme purified from either infected or uninfected cells efficiently transcribed all ϕYS40 promoters *in vitro*. Interestingly, our data showed that during infection, there is a significant increase and decreaseof transcript amounts of host translation initiation factors IF2 and IF3, respectively. This finding, together with the fact that most middle and late ϕYS40 transcripts were found to be leaderless, suggests that the shift to late viral gene expression may also occur at the level of mRNA translation.

Keywords: Thermus thermophilus; bacteriophage; bioinformatic promoter search; macroarray analysis; leaderless mRNA

Introduction

As of the time of this writing, the complete genomic sequences of more than 380 bacteriophages (NCBI, last modified August 2006) infecting a broad variety of microorganisms have been obtained. During infection, all bacteriophages exploit resources of their hosts to redirect the host gene expression machinery to serve the needs of the virus. While comparative genomics of phages has provided important insights into the process of phage evolution, our understanding of gene expression strategies used by various phages to achieve productive infection is modest at best. Results of biochemical studies of regulation of gene expression

Abbreviations used: RNAP, DNA-dependent RNA polymerase; ORF, open reading frame.

E-mail addresses of the corresponding authors: severik@waksman.rutgers.edu; minakhin@waksman.rutgers.edu

in just a few phages (λ, T4, T7 and, more recently, XP10) have been extremely informative and provided paradigms of genetic regulation of general biological significance. For several other, less-studied phages, recent kinetic analysis of gene transcription patterns and modeling was used to uncover viral regulatory circuits dynamics that suggested the existence of specific regulatory mechanisms.[1–4] Due to the overwhelming diversity of phages, it is clear that further studies of bacteriophage-encoded regulatory mechanisms will reveal novel paradigms of gene regulation.

Previously we presented an approach combining bioinformatics and experimental studies that allowed us to obtain a comprehensive view of temporal gene expression during infection of *Xantomonas oryzae* by phage XP10.[3,5] Here, we extend parts of such analysis to a much larger phage ϕYS40 that infects hyperthermophilic eubacterium *Thermus thermophilus*. Despite recent advances in phage genomics, only a few phages infecting thermophiles have been completely sequenced to date. Most of thermophilic phages whose genomes have been determined infect hyperthermophilic archaeal species and may be of little relevance for understanding phages that infect thermophilic eubacteria.[6–8] The subject of this study, bacteriophage ϕYS40, is similar in its genome size[9] and virion morphology[10] to T4, a prototypical *Escherichia coli* phage whose studies over the years revealed a staggering variety of mechanisms of regulation of gene expression. We hypothesized that like T4, ϕYS40 may also encode a wealth of regulatory mechanisms ensuring coordinated regulation of different temporal classes of viral genes. Uncovering such mechanisms and establishing phage-encoded proteins responsible is of great interest, since proteins from thermophilic organisms are good candidates for crystallization, alone or in complex with their cellular targets. Thus, characterization of regulatory mechanisms encoded by phages infecting thermophilic bacteria will allow us to approach the molecular basis of genetic regulation structurally. With these ideas in mind, we studied host and viral gene expression during ϕYS40 infection. Our results reveal temporal regulation of ϕYS40 transcription and allow identification of early, middle and late phage promoters. Promoters from the last two temporal classes have distinct consensus elements that differ from elements of early viral and housekeeping host promoters and may define a new class of bacterial RNA polymerase (RNAP) promoters. Analysis of early and middle/late phage mRNA strongly suggests that during ϕYS40 infection there occurs a novel regulatory "shift" from host to viral genome expression at the level of translation initiation. Thus, our results show the potential of comprehensive analysis of bacteriophage infection process for identification of novel regulatory mechanisms, and open up several new avenues for experimental investigation of genetic switches in *Thermus*.

Results

Prediction of putative σ^A-dependent –10/–35 promoters in the ϕYS40 genome

Bacteriophage ϕYS40 does not encode a RNAP or any recognizable RNAP σ factor and must therefore rely entirely on host RNAP to transcribe its genes. Transcription from early ϕYS40 promoters is most likely initiated by *T. thermophilus* RNAP holoenzyme containing the primary sigma factor, σ^A. To efficiently compete for RNAP with host promoters, early viral promoters should be strong, i.e. they are expected to have a good match to σ^A consensus promoter elements, which should allow their identification by bioinformatic means. To identify putative ϕYS40 early promoters, we created a bioinformatic model of a *T. thermophilus* σ^A promoter. The model is based on previously reported *T. thermophilus* σ^A promoters (see Accession numbers[32] in Table 1), both those with experimentally verified transcription start points (by primer extension and/or S1 mapping) and those for which such determination was not made. Manual multiple sequence alignment of ten promoters with identified start points revealed, as expected, an unambiguous sequence conservation of the −10 and −35 promoter elements. The SignalX program[11] was applied to this alignment in order to make an initial positional weight matrix (profile) of *T. thermophilus* σ^A promoters. This profile assigns a numerical weight to each nucleotide at each position, so that a total score (*z*-score) of a candidate sequence reflects its similarity to known promoters. Five *T. thermophilus* promoters without experimentally identified start points were analyzed using the initial profile to reveal likely locations of promoter consensus elements and the final profile of a σ^A-dependent *Thermus* promoter was built using a multiple alignment of all 15 known *T. thermophilus* promoters (see Table 1; Supplementary Data, Table S1; Figure 3(a), below). The *z*-score of the consensus *Thermus* promoter was 4.5; the highest and lowest *z*-scores in the training set were 4.42 and 3.02 for the P_{215} promoter and promoter in front of the 4.5 S rRNA gene, respectively (see Table 1).

The promoter profile was used to search the ϕYS40 genome with the GenomeExplorer program.[11] The following search parameters were used: (i) for every ϕYS40 gene, a region from −200 to +75 bp relative to the first nucleotide of the annotated start codon was considered; (ii) the spacer length between the –10 and the −35 promoter elements was allowed to vary between 16 bp–18 bp; (iii) the sequence of the spacer did not influence the search; (iv) irrespective of its direction, a predicted promoter could intersect with an upstream gene by no more than 50 bp; and (v) the search cutoff was set at a *z*-score of no less than 3.5. This cutoff was selected as a tradeoff between search specificity (absence of candidate early promoters upstream of genes coding for previously identified ϕYS40 virion proteins,[9] which should belong to

Table 1. *Thermus thermophilus* promoters

	Promoter sequence[a]		Accession number or reference	Spacer length	Score[b]
	−35	−10			
P31*	CGGCCCAAGGTTTACAAAATCCCCGCCCCCGTCC-	TAGCCTGGGGGCAAGGAGG	D43662	18	3.86
P35*	CCTCTTCTTGTTGACGGGACGGGGAGGAGGGCC--	TATCCTGGGTAAGGCTTGG	D43663	17	4.23
P37*	GGGCCTGGCGTTGACCATCTTCCTCCTTGGCCT--	TATCCTTAGGGTGCGTCCG	D43664	17	4.35
P39*	CGCCTCGCCCTTGACGGGGAGGAGGCAACGGGG--	TAAAACAGGGGCGAGAGCG	D43665	17	3.20
P43*	GAATTCGGTCTTGTCAAGTAAGCTTAGCTATGG--	TAACATAGACCTGGGAGGT	D43666	17	3.78
P214*	TAGGAGGGGCTTGCCAATCCGCCCCTTAGAGTG--	TACCATAGCGATTGCCCGA	D43672	17	4.01
P215*	TGAGTAGCACTTGACATCATAAAGTGCTTGAGG--	TATCATCCGACCTGGGCGC	D43673	17	4.42
Tth slp*	GCCCGACCGCTTGACAAGGGCGCGTGAGGTTTT--	TACGATAGCGCCGGATGCG	X57333	17	3.94
Tth dnaK	CTCAACTCCCTTGACAAAAATGCGGCATGTGCGT-	TAGCCTGGGAGGCGAGGTG	Y07826	18	4.32
Tth rpsT	ACCACCAAGTTTGCCCCTTAGGCGCAGGGTGTGCT	TACACTGGGCCCTGGTTTG	AJ295159	19	3.70
Tth argF*	GCCTAGGCCCTTGCATAAGTTTGCGGGCACGGGG-	TATGCTTAAGGCCTCATGG	Y18353	18	3.13
Tth ORF4 (arg)*	TTTACACCACTTGACAGCTTTTGATTCTGAGTC--	TATCCTCTATTCGGGGAGC	Y18353	17	4.50
Tth 4.5S rRNA	CTCACGCCCTCTAGCCTCAGGGCTTCCATGGGTGC	TATACTACCCGAGCCCCCG	X12643	19	3.02
Tth 16S rRNA	TCGCAAGCCTTGACAAAAAGGAGGGGGATTGA--	TAGCATGGCTTTTCTGCG	Ref. 32	17	4.24
Tth 23S rRNA	TGGGGGCCCTTGACAAAGGCCATGCCTCCTTGG-	TATCTTCCCTTTTGCGCT	Ref. 32	18	4.03

[a] Shown in grey boxes: −35 and −10 sequences.
[b] Scores are computed using the final profile.
* Genes with experimentally mapped transcription start sites. The sites are underlined.

middle or late viral genes) and sensitivity (absence of likely early genes or operons without candidate early promoters in front of them).

Using these parameters, 47 putative −10/−35 promoter sequences were identified. For several candidate promoters, predicted transcription start points were located downstream of annotated translation start codons. Four predicted promoters for which no downstream start codons could be located were excluded from further analysis, leaving a total of 43 promoters listed in Table 2. Putative promoters for which alternative (i.e. different from those reported in the published annotation) downstream translation start codons could be located are marked by an asterisk in Table 2 (two asterisks when alternative start codons are preceded by plausible Shine–Dalgarno sequences). The new open reading frame (ORF) coordinates are also listed in Table 2.

As expected, no promoters were predicted in non-coding regions between ϕYS40 genes in a tail-to-tail arrangement. Among the head-to-head arranged genes (a total of 12 gene pairs), the non-coding region separating genes 27 and 28 contained two divergent predicted promoters, while the rest contained only one (gene pairs 15–16, 32–33, 36–37, 131–132, 136–137, 163–164) or no (55–56, 65–66, 95–96, 140–141, 146–147) predicted −10/−35 promoters. The head-to-head transcribed regions with no predicted promoters likely contain phage promoters that are different from the −10/−35 class promoters (this conjecture was largely confirmed by further analysis, see below).

The ϕYS40 genome contains 170 annotated ORFs and three tRNA genes.[9] Two-thirds of the ϕYS40 genome (114 genes) are transcribed in one direction (leftward in the genome map; see Figure 1, and 56 genes are transcribed in the opposite, rightward, direction. Earlier analysis identified four gene clusters in the ϕYS40 genome.[9] With an exception of rare "intruders", genes within a cluster are transcribed in one direction (leftwards for cluster 1 (genes 1–36) and cluster 3 (genes 62–146), rightwards for cluster 2 (genes 37–61) and cluster 4 (genes 147–170) (Figure 1). While clustering is statistically significant, no inferences about its functional role were made. The distribution of putative early promoters in ϕYS40 gene clusters is highly non-random. Cluster 3 contains 30 predicted promoters; cluster 1, 8; cluster 2, 3; and cluster 4, 2 promoters. In cluster 3, all putative −10/−35 promoters are located upstream of genes 83–137, a group of short genes that code for proteins of unknown function. In other clusters, predicted −10/−35 promoters are located upstream of genes involved in nucleotide metabolism, replication, recombination, and regulation of transcription.[9] Only one predicted −10/−35 promoter-like sequence was found upstream of a ϕYS40 virion structural gene (gene 154), strongly indicating that a separate class of promoters is used for expression of structural (late) ϕYS40 genes.

The logos[12,13] of the −35 and −10 promoter elements of *T. thermophilus* promoters and predicted ϕYS40 early promoters are shown in Figure 2(a) and (b). As can be seen, positions −7, −11, and −12 of the −10 promoter element are the most conserved ones in both the host and predicted viral promoters (the corresponding positions are also highly conserved in the *E. coli* σ^{70}-dependent promoters). Both host and viral promoters have a less conserved extended −10 "TG" motif. The −35 element of predicted phage promoters has a consensus sequence CTTGACa, compared to *T. thermophilus* cTTGACA and *E. coli* TTGACA consensus sequences. Inspection of predicted phage promoter sequences upstream of the −35 element, downstream of the −10 element, or in the spacer between the elements using the SignalX program did not reveal any additional areas of sequence similarities.

Macroarray analysis of gene expression during ϕYS40 infection

To understand the temporal pattern of ϕYS40 gene expression, a macroarray of ϕYS40 genes was

Table 2. Predicted early promoters of the bacteriophage ϕYS40

	Strand	Location[a]	Sequence and spacer[b]	Distance[c]	Score	Gene function
ORF8**	<=	**9399..9920**	TTGACA-17-TATgCT	−11→10	4.20	dUTPase
ORF13	<=	12792..13367	TTGACA-18-TAatCT	13	3.64	Recombination protein
ORF15	<=	14743..15036	TTGACA-18-TAagCT	17	3.96	Unknown
ORF18	=>	16640..17050	TTGACA-17-TATgCT	3	4.20	DNA binding HTH-domain protein, transcription regulator
ORF23	<=	19944..21620	TTGACA-17-TAgCaT	0	4.24	DNA primase bacterial DnaG type
ORF27	<=	23898..25247	TTGACt-17-TAcgaT	0	3.57	DEAD domain helicase
ORF28	=>	25396..26796	TTGACA-18-TATagT	0	3.67	Unknown
ORF33	=>	30387..32498	TTGACA-17-TAcCaT	1	4.24	DNA polymerase, without N-terminal 5′-3′exonuclease domain
ORF37**	=>	**33417..33746**	TTGACA-17-TAataT	−27→10	3.56	Unknown
ORF41	=>	35201..37594	TTGACt-17-TATCaT	1	4.05	Ribonucleotide reductase, alpha subunit, N-terminal portion
ORF61*	=>	**52062..52484**	TTGACt-17-TATgaT	−23→4	3.75	Unknown
ORF83**	<=	**84867..85073**	TTGACA-17-TATtaT	−8→13	3.80	Unknown
ORF84**	<=	**85328..85534**	TTGACA-17-TATgaT	−11→13	4.12	Unknown
ORF85	<=	85767..85919	TTGACA-17-TATCtT	12	4.13	Unknown
ORF86**	<=	**86022..86258**	TTGACg-17-TAagaT	−8→7	3.61	Unknown
ORF87	<=	86382..86618	TTGACt-17-TAagaT	32	3.51	Unknown
ORF88	<=	86909..87154	TTGACA-17-TATagT	30	3.67	Unknown
ORF89	<=	87505..87990	TTGcCA-18-TAaCCT	75	4.03	Unknown
ORF90**	<=	**88074..88439**	TTGACA-17-TaggaT	−56→34	3.94	Unknown
ORF91	<=	88642..89250	TTGACA-17-TAaCCT	0	4.26	Unknown
ORF93	<=	89796..90221	TTGcCt-17-TAgCCT	12	3.72	Unknown
ORF98	<=	91835..92380	TTGACA-17-TAcaCT	13	4.08	Unknown
ORF101	<=	93619..94131	TTGACt-17-TAgCCT	11	3.95	Unknown
ORF103	<=	94885..95373	TTGACc-17-TAaaCT	13	3.87	Unknown
ORF104	<=	95510..96025	TTGACt-17-TAagCT	18	3.59	Unknown
ORF105	<=	96096..96626	TTGACc-17-TAagCT	45	3.81	Unknown
ORF106	<=	96833..97354	TTGACc-17-TAaCCT	17	4.11	Unknown
ORF108	<=	99280..100323	TTGACt-17-TAagCT	29	3.59	ATPase
ORF110	<=	101227..101973	TTGACt-17-TAggCT	35	3.65	Glycosyltransferase
ORF114	<=	103616..104107	TTGACt-17-TATCCT	12	4.13	Unknown
ORF115	<=	104451..104693	TTGACA-17-TAggaT	14	3.94	Unknown
ORF116	<=	104803..105279	TTGACA-17-TAgCtT	0	3.95	Unknown
ORF117*	<=	**105422..105928**	TTGACA-17-TATaCT	−51→0	4.26	Unknown
ORF121	<=	107552..108046	TTGACA-17-TATaaT	18	4.18	Unknown
ORF122	<=	108141..108644	TTGACA-17-TATCCT	16	4.50	Unknown
ORF124	<=	109328..109819	TTGACA-17-TATagT	21	3.67	Unknown
ORF125	<=	109998..110513	TTGACc-18-TATtaT	16	3.65	Unknown
ORF128	<=	111663..112133	TTGACA-17-TAgCaT	0	4.24	Unknown
ORF131	<=	113202..113630	TTGACA-17-TATgaT	104	4.12	Unknown
ORF133	<=	114385..115032	TTGgCA-17-TATaCT	11	3.86	Unknown
ORF137	=>	116815..117474	TTGACA-17-TATagT	65	3.67	Unknown
ORF154	=>	136388..137287	TTGACA-18-TATatT	5	3.89	Unknown
ORF163***	<=	146390..147022	TTGACA-17-TAaggT	6	3.37	Unknown

ORFs preceded by experimentally verified promoters are underlined.
[a] Location: genomic coordinates, re-annotated are shown in bold.
[b] Capitals: consensus nucleotides.
[c] Distance: the distance between the start of transcription and the start codon of the gene.
* Change of the distance by selection of the candidate start codon downstream of the annotated start of ORF.
** Change of the distance by selection of the candidate start codon with a strong Shine–Dalgarno box downstream of the annotated start of ORF.
*** *z*-score of this promoter is below 3.5.

prepared. The array contained spots with equal amounts of PCR-amplified fragments of 29 representative viral genes. One group of spots reported the abundance of mRNA of genes from the predicted "early" region of cluster 3 (genes 83, 91, 94, 116, and 131). Other spots represented genes likely involved in nucleotide metabolism, replication, and recombination (genes 15, 16, 23, 27, 32, 33, 36 and 37), genes coding for structural proteins and DNA packaging enzymes (genes 1, 3, 56, 73, 82, 146, 147, 152, 163 and 164), and genes coding for putative transcription regulators (genes 18 and 71). Since partially overlapping or closely spaced viral genes are likely co-transcribed (transcribed from the same promoter), some spots on the array report abundance of transcripts of multiple genes. For example, gene spots 1, 3, and 15 and 23 and 27 likely report the abundance of polycistronic mRNAs from transcription units comprising genes 1–15 and 19–27, respectively. The array also included pairs of spots corresponding to gene pairs in the "head-to-head" orientation (genes 15-16, 27-28, 32-33, 36-37, 55-56, 163-164; see Figure 1), since these divergently transcribed genes may belong to different temporal classes (see above).

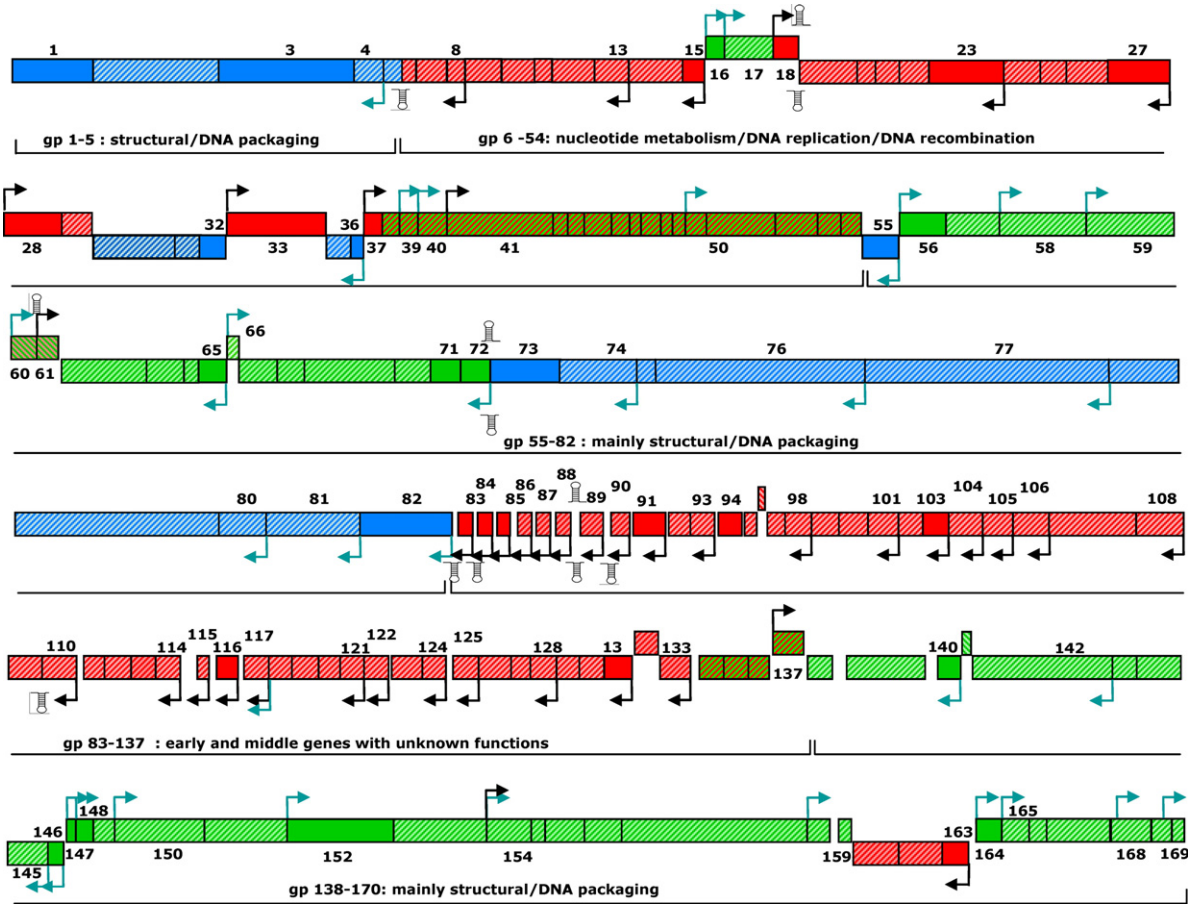

Figure 1. Transcription map of the *T. thermophilus* bacteriophage ϕYS40. Colored boxes on the corresponding strand of the phage DNA represent each gene: upper boxes indicate genes with rightward orientation; lower boxes indicate genes with leftward orientation. The genes belonging to different temporal classes (defined by macroarray analysis and primer extension) are shown in different colors: early, red; middle, green; late, blue. The genes that likely belong to the corresponding classes are represented by shaded boxes of the corresponding color. Double-colored shaded boxes indicate genes with uncertain temporal class. The genes with numbers shown were used in macroarray and/or primer extension analysis or have predicted promoters. The functional modules are indicated by brackets at the bottom of the map. Promoter locations are depicted as bent arrows colored in black or blue to indicate early or middle/late promoters, respectively. Hairpins indicate possible rho-independent terminators.

In order to determine whether ϕYS40 shuts off host gene expression, PCR fragments corresponding to several housekeeping *T. thermophilus* genes, *rpoC* (RNAP β′ subunut), *sigA* (the primary sigma factor σ^A), *dnaK* (protein chaperone), *TTHA0466* (alcohol dehydrogenase), *infB* (translation initiation factor 2, IF2), and *infC* (translation initiation factor 3, IF3), were included in the array. The membrane also contained spots with total genomic DNA of ϕYS40 and its host. As a loading and normalization control, two spots containing a PCR fragment of the *zfrp8* gene from *Drosophila melanogaster* were used. *T. thermophilus* cells were infected with ϕYS40 and total RNA was extracted 0, 25, 50, and 75 min post-infection. The time-points were selected on the basis of a single-burst experiment that indicated that a 25 min time point corresponded to the middle of the eclipse period, the 50 min time point corresponded to its end, while at the 75 min time point progeny phage began to be produced.

Equal amounts of total RNA from each time point were combined with the *zfrp8* probe and used to generate radioactively labeled cDNA by random priming/reverse transcription followed by hybridization to the array. To quantitatively analyze macroarray data, radioactive signals from each spot were corrected for background and normalized based on the relative strength of the *zfrp8* spot signal. Next, the amount of radioactivity in each spot (which corresponds to transcript abundance) was plotted as a function of time post infection. As expected, the total amount of ϕYS40 transcripts increased through infection relative to the control *zfrp8* spot (blue line in Figure 3(a)). In contrast, the total amount of *T. thermophilus* transcripts normalized to the *zfrp8* spot decreased throughout the same period (red line in Figure 3(a)), indicating that ϕYS40 either shuts off host transcription or increases the rate of host transcripts decay. The abundances of some individual transcripts, such as *rpoC*, *sigA*, *infC*, and *dnaK* also decreased between the 25 and 75 min time points (data not shown). Interestingly, the amount of the *infB* transcript, which was relatively low in the beginning of infection, increased rapidly

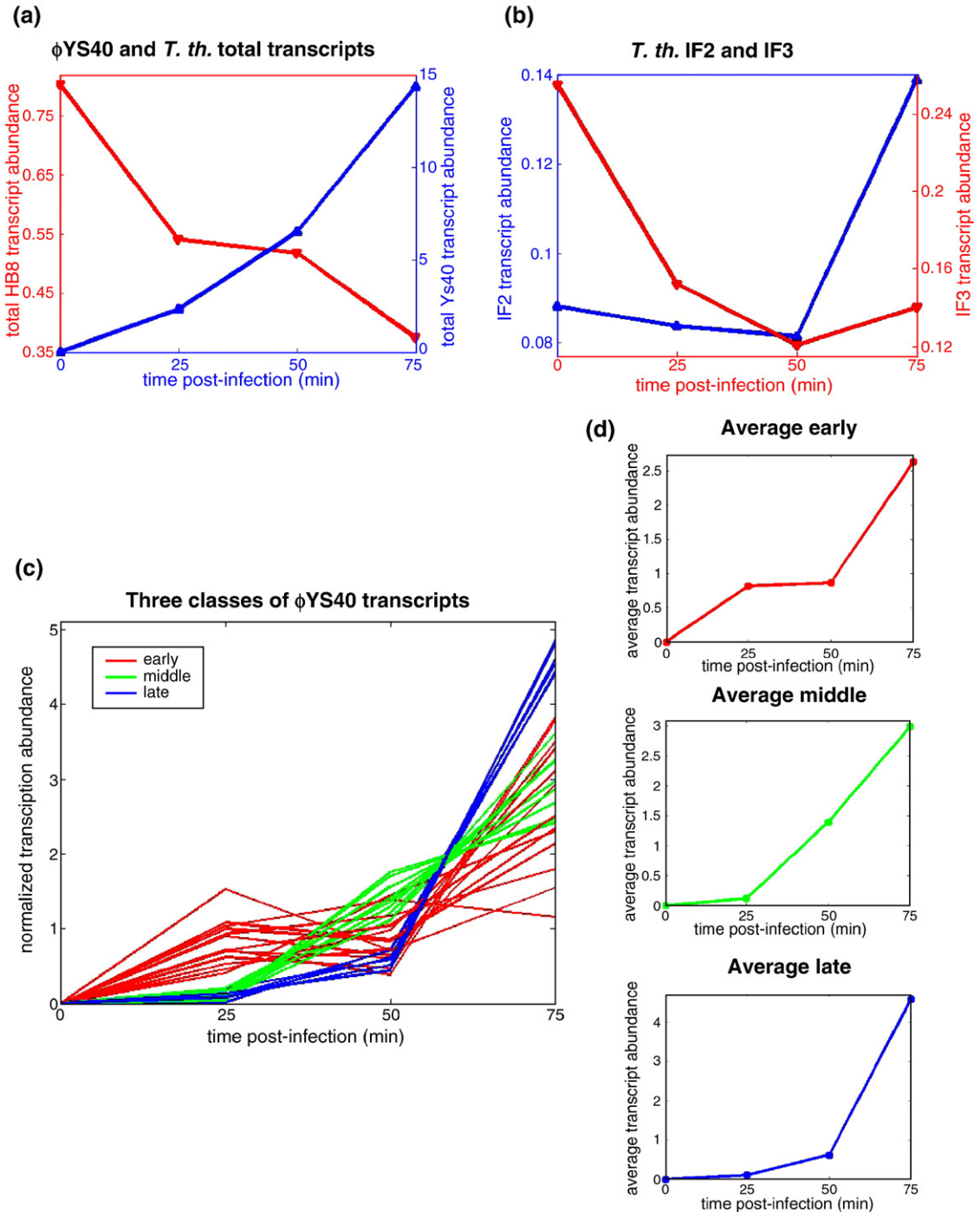

Figure 2. Sequence logo representation of *T. thermophilus* and ϕYS40 promoters. Consensus sequences were plotted with WebLogo.[13] The height of the letter indicates degree of conservation. Positions are done with respect to putative or identified transcription start sites. (a) *T. thermophilus* −10/−35 promoter sequence logo with independently aligned the −35 and the −10 regions. (b)–(e) ϕYS40 predicted early (b), verified early (c), predicted middle/late (d) and verified middle/late (e) independently aligned promoter consensus sequences are plotted.

after the 50 min time point (a blue line in Figure 3(b)), which is contrary to the rapid decrease of *infC* transcript amount during the infection (a red line in Figure 3(b)). This unusual behavior is discussed in more detail in Discussion.

To compare the behavior of individual ϕYS40 transcripts, plots of normalized spot signal intensity *versus* time post infection were scaled to make mean transcript abundances for each spot equal (Figure 3(c)). Systematic clustering analysis of temporal patterns of individual ϕYS40 genes (see Supplementary Data, Figure S1) revealed three different temporal classes. The averages of scaled abundances calculated for each of the three temporal classes are shown as separate panels in Figure 3(d). As can be seen, the three temporal classes are clearly distin-

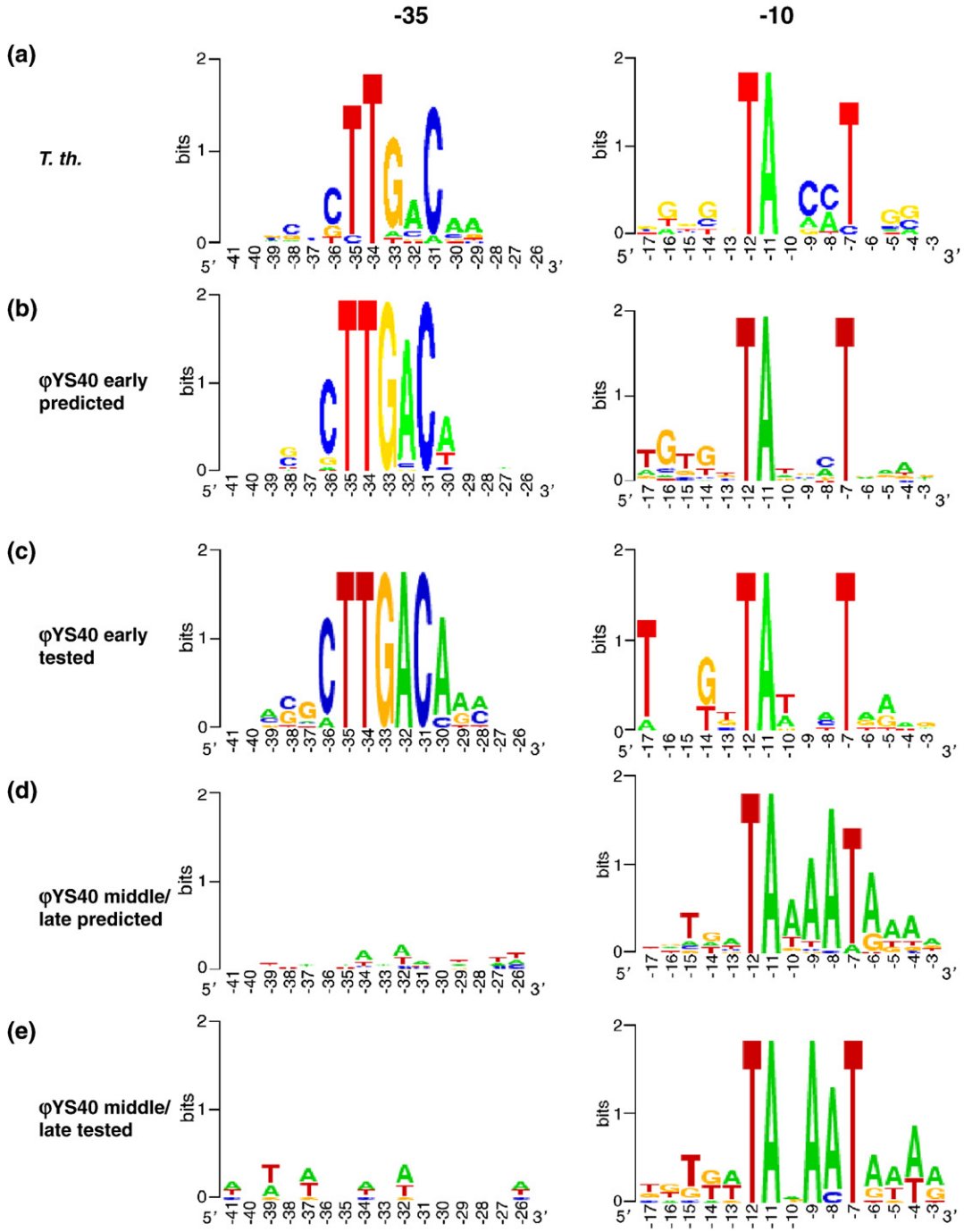

Figure 3. Macroarray data analysis. (a) The abundance of total ϕYS40-encoded transcripts (blue line) is shown together with the abundance of total *T. thermophilus*-encoded transcripts (red line). (b) The transcript abundances of the translation initiation factors IF2 and IF3 (blue line and red line, respectively) are shown together. (c) Normalized transcript abundances are presented for individual ϕYS40 transcripts as a function of time. Transcripts that belong to different temporal classes are shown in different colors. The curves are colored according to Figure 1: early, red; middle, green; late, blue. Classification of individual transcripts into the three temporal classes is performed by the procedure described in Supplementary Data, Appendix 1. (d) The three vertical panels on the right show averaged normalized transcript abundances corresponding to the three temporal classes.

guished by the period of time during which the greatest change in transcript abundance occurs. For the first class, significant amounts of transcripts accumulate during the first 25 min of infection. Genes from this class are classified as ϕYS40 early genes. Transcripts of the second class have very low abundance in the first 25 min of infection but their abundance increases dramatically between 25 min and 50 min post-infection. These transcripts correspond to ϕYS40 middle genes. Finally, the abundance of transcripts from the third temporal class is low during the first 50 min post-infection

but dramatically increases afterwards. These are ϕYS40 late transcripts.

The genomic positions of ϕYS40 genes that belong to different temporal classes are shown in Figure 1. Many genes with unknown function, most notably all of cluster 3 genes located downstream of predicted −10/−35 class promoters, belong to the early class. Genes whose products are involved in DNA replication, recombination and nucleotide metabolism also belong to this class. Every gene (or a group of likely co-transcribed genes) that behaves as early on the macroarray is preceded by a predicted −10/−35 class promoter, independently confirming our promoter prediction results. The only exceptions are co-transcribed genes 163–165. However, this group of genes is preceded by a predicted −10/−35 promoter with a *z*-score of 3.37, just below the cut-off value of 3.5 used for the search. It is therefore likely that this early promoter is functional and we therefore included it in Table 2 (marked by three asterisks).

Most ϕYS40 middle genes encode structural proteins as well as proteins involved in DNA packaging. Late genes with known functions encode exclusively the structural proteins of the phage. There are no predicted early promoters upstream of middle and late genes revealed by macroarray analysis, again suggesting that promoters for genes of these temporal classes differ from the −10/−35 class promoters.

Mapping ϕYS40 promoters *in vivo*

In our initial attempts to identify middle and late promoters of the phage, regions upstream of genes that were found to belong to the middle and late temporal classes were examined bioinformatically for the presence of common sequence motifs that were absent from the early promoters. However, no such motifs could be identified, possible due to the small number of genes examined. To identify ϕYS40 promoters experimentally, primer extension analysis of RNA samples used in macroarray experiments was performed. Overall, 5′ ends of 18 phage transcripts were identified. Primer extension product corresponding to a representative early ϕYS40 promoter (P_{83}; Figure 4(b)) peaked 25 min post-infection and decreased steadily afterwards. Primer extension product corresponding to a representative middle promoter P_{140} appeared between 25 and 50 min post-infection and increased steadily afterwards (Figure 4(b)). Primer extension product corresponding to late transcripts appeared after 50 min and dramatically increased by the end of infection (a representative late transcript from P_{82} is shown in Figure 4(b)). In case of middle and late transcripts, kinetics of primer extension products accumulation during the infection matched that observed in macroarray experiments. However, primer extension products corresponding to RNA transcribed from early promoters decreased between 50 min and 75 min post-infection, while the macroarray data showed continued increase in early transcript abundance. Since primer extension reveals abundance of mRNA transcribed from an individual promoter, it is possible that the increase of macroarray signal later in the infection is due to read-through transcription from middle and/or late promoters located further upstream. In agreement with this idea, for almost half of the early phage genes spotted on the macroarray (6 of 13) there is a predicted middle/late promoter further upstream (Figure 1). Alternative explanations such as (i) preferential degradation of 5′ ends of early phage mRNAs or (ii) transcription antitermination late in infection are also possible.

For each of the nine primer extension reactions designed to reveal the presence of bioinformatically predicted ϕYS40 −10/−35 promoters, expected primer extension products were obtained. Moreover, all nine promoters belonged to the early temporal class (they are underlined in Table 2). The result shows that our bioinformatic analysis identified early ϕYS40 promoters with a high degree of confidence. Transcription start points for six of the early promoters were located in front of annotated genes 18, 85, 91, 103, 116, and 131 (Figures 1 and 4(a)). For three other early genes, 37, 83, and 84, annotated translation start codons were located upstream of experimentally determined (and predicted) transcription start sites. However, additional start codons preceded by plausible Shine–Dalgarno motifs were found downstream of experimentally determined transcription start points, strongly indicating that initial annotations of coding sequences of these genes were incorrect. Interestingly, for three of the nine −10/−35 promoters analyzed, transcription start points coincided with or were very close to (2 bp upstream in the case of P_{18}) translation start points (Figure 4(a)).

In order to identify middle and late viral promoters, regions upstream of genes that were found to belong to the middle and late temporal classes were examined by primer extension. Nine primer extension products corresponding to seven middle (P_{16}, P_{72}, P_{140}, P_{146}, P_{147}, P_{148}, and P_{164}) and two late (P_{55} and P_{82}) transcripts were identified. Sequence alignments of regions upstream of middle and late transcript primer extension products ends revealed a common −10-like element (consensus sequence TAaAATa) with the highest conservation of positions −12, −11, −9, and −7 relative to the transcription start point (Figures 3(e) and 4(a)). Also, a presence of the extended −10 "TG" motif was detected in some middle/late promoters. No additional areas of conservations were apparent. Remarkably, the transcription start sites of eight of the nine experimentally identified middle and late promoters are located 0–2 bp upstream of the first nucleotide of annotated translation start codons (only P_{148} has an obvious upstream Shine–Dalgarno motif; see Figure 4(a)). Barring gross misannotation of ϕYS40 ORF start points, the result suggests that most middle and late viral transcripts (and some early transcripts, see above) are leaderless.

Since no obvious differences between the middle and late promoter sequences could be detected, a

(a)

ϕYs40 tested promoters

	-35	TG	-10	+1	mRNA
early promoters					
P_{18}	TGTTTGCTCTCTAGGGGGC**TTGACA**ATTCTTGTCAATT**TG**TT**TATGCT**AGTCAT**G**CT**atg**				LL
P_{37}	TTATTTTTCCCGGTGCGTA**TTGACA**GCTATGCTGTTTTA**TGG****TAATAT**TGAACA**A**GGGAGAAAC**atg**				
P_{83}	GGTGATTAAGTGTTGGGGC**TTGACA**AAAAGGATTAAGTAGGG**TATTAT**GAAGGT**G**GAGGTGAGTGGG**atg**				
P_{84}	ACGCTATTGGCAGATAGGC**TTGACA**AATAGTCATTAA**TG**GGT**TATGAT**GAAGGT**G**GAGGTGAGTAGG**atg**				
P_{85}	TTTTCTTTTTACTCCCCCC**TTGACA**AAATGTACTAAAAGAGG**TATCTT**AGTGTT**G**GGAGGTGAACC**atg**				
P_{91}	CATCTCAAATATATAACGC**TTGACA**ACGCTTTAGCGCTT**TGCTAACCT**AAAAAC**a**tg				LL
P_{103}	CTGAGCGCTTGCACGCTAC**TTGACC**TAAGGCGTTTTGTCCTT**TAAACT**AAGGC**GA**GGGGGTGAAGAG**atg**				
P_{116}	ATTATTAGTGTACCAACGC**TTGACA**GCGTTTCGCTAT**TGTGTTAGCTT**GAAGAC**a**tg				LL
P_{131}	ATTAATAGTGTACCAACGC**TTGACA**GCCTTAGGTTTG**TG**TTC**TATGAT**TTCTTCT**G**T Š *101bp* Š **gtg**				
middle/late promoters					
P_{16}	CACAACTTTGGGGTGTCTGTCAACAGGTAGCAGAATTGG**TGATAAAAT**AATTCT**G**CT**atg**				LL
P_{55}	TGACTTTATATTATCACAAAATTTTCGTTCAATTAAAGAC**TGTAAACT**AATATT**A**T**atg**				LL
P_{72}	AAAAAATGCCTCCCATTTGTTGGGGGGCATTTTCTTTTT**TGTTATAAT**GTAGCT**A**GC**atg**				LL
P_{82}	AAACGCTTTTTCAAATTTAAATTATTGCTAAAACATTC**TG**AT**TAAAAT**AGTACTT**a**tg				LL
P_{140}	TAAATTTCCATACACACATCATACCACAAGACGACTT**TG**TTA**TAGAAT**ATAATT**a**tg				LL
P_{146}	TTATATTGTAAAACAAAAAGCGAAAAAACAAAAGTTCC**TG**TA**TAAAAT**GAAATT**a**tg				LL
P_{147}	TTTCATTTTATACAGGAACTTTTGTTTTTTCGCTTTT**TG**TTT**TACAAT**ATAAAT**g**tg				LL
P_{148}	ACCTTATTTCGAAAATTTTTAAAAGAACAAGGACTTT**TGTGATAGAAT**TAAGTG**G**AGGAAGTT**atg**				
P_{164}	ACCACACCGTGCCTTTGGTGTCAAGGCGCTGATGGAAGA**TGTTAAAAT**AAAGCT**a**tg				LL

(b)

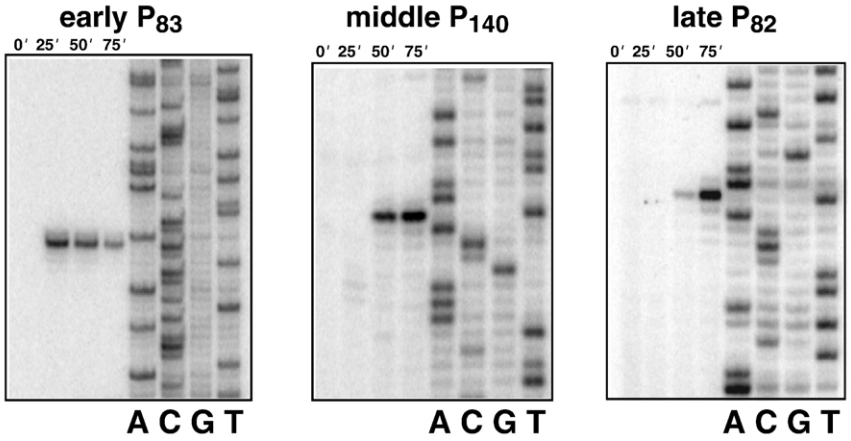

Figure 4. ϕYS40 verified promoters. (a) Alignment of the sequences of verified ϕYS40 promoters is shown. The −35, −10 and TG putative promoter elements are shown in bold. Experimentally determined transcription start sites are both boldface and underlined. The assigned translation initiation codons are shown in bold lower case. Putative leaderless mRNAs transcribed from the corresponding promoters are indicated as LL. (b) The kinetics of accumulation of representative *in vivo* primer extension products obtained with early, middle and late phage transcripts during infection.

profile of a ϕYS40 middle/late promoter was created based on an alignment of eight experimentally confirmed leaderless middle and late promoters. The profile included an ATG/GTG start codon located 5–10 bp downstream of the −10 element (see Supplementary Data, Table S2). The ϕYS40 genome was searched for the presence of middle/late promoters using parameters identical to those used for early phage promoters search (see above) but the search area was limited to positions −75 to +75 relative to the first nucleotide of published annotated start codons.

Thirty-seven additional candidate middle/late promoters were revealed by the search. Several predicted middle/late promoters were located inside of annotated ORFs; however, start codons associated with predicted promoters were in-frame with these ORFs and could therefore be likely used for translation initiation. A few putative middle/late promoters whose ATG/GTG elements were out of frame with annotated ORFs were excluded from further analysis, on the grounds that they were invariably located in areas containing predicted or experimentally confirmed early promoters (data not shown; see also below). The remaining 28 new putative middle/late promoters are listed in Table 3 (new proposed start codons that are in-frame with previously annotated ORFs are marked by asterisks). Table 3 also contains eight experimentally confirmed leaderless middle/late promoters that were also found by the search, as expected.

To assess the quality of bioinformatic predictions of phage middle/late promoters, additional primer extension reactions were performed using primers

Table 3. Predicted middle/late promoters of the bacteriophage ϕYS40

	Strand	Location[a]	Sequence and spacer[b]	Distance[c]	Score	Gene function
ORF4	<=	7412..8068	TAAAATA-(6)-gTG	1	3.92	Unknown
ORF16	=>	15124..15453	TAAAATA-(8)-ATG	3	4.31	Unknown
ORF17	=>	15467..16576	TAAAATA-(9)-ATG	4	4.31	IMP dehydrogenase/GMP reductase
ORF36*	<=	**33031..33318**	TAAAATA-(7)-gTG	11->2	3.92	Unknown
ORF39	=>	34188..34616	TAAAATA-(5)-ATG	0	4.31	Unknown
ORF40	=>	34631..35155	TAAcATA-(7)-ATG	2	3.64	Unknown
ORF50	=>	40558..41013	TAAAATg-(5)-ATG	0	4.10	Unknown
ORF52	=>	42536..43408	TAAtATA-(5)-ATG	0	3.64	*N*-Acyltransferase
ORF55	<=	44426..45127	TAAAcTA-(7)-ATG	2	3.92	Unknown
ORF56	=>	45187..46209	TAAtATA-(9)-ATG	4	3.64	Unknown
ORF58	=>	47564..49414	TAtAATt-(5)-ATG	0	3.69	Unknown
ORF59	=>	49453..51312	TAtAATA-(8)-ATG	3	4.02	Serine kinase
ORF60	=>	51410..51997	TAAgATA-(8)-ATG	3	3.64	dNMP kinase
ORF65*	<=	**55466..56062**	TAtAATA-(7)-ATG	47->2	4.02	Terminal protein in replication
ORF66	=>	56049..56315	TAAAATg-(5)-gTG	0	3.71	Unknown
ORF72	<=	61167..61682	TAtAATg-(8)-ATG	3	3.81	Unknown
ORF74*	<=	**63204..64826**	TAtAATg-(9)-ATG	−8->4	3.81	Unknown
ORF76	<=	65085..69662	TAAAATA-(10)-ATG	5	4.31	Unknown
ORF77	<=	69684..74918	TAgAATA-(5)-ATG	0	4.13	Unknown
ORF80	<=	79880..80743	TAAAATA-(5)-gTG	0	3.92	Unknown
ORF81	<=	80788..82740	TAtAATA-(9)-ATG	4	4.02	Unknown
ORF82	<=	82771..84609	TAAAATA-(6)-ATG	0	4.31	Unknown
ORF117	<=	105422..105979	TAAAAaA-(7)-ATG	2	3.64	Unknown
ORF140	<=	120226..120777	TAgAATA-(5)-ATG	0	4.13	Unknown
ORF142	<=	120953..123997	TAAAAaA-(7)-ATG	2	3.64	Unknown
ORF145	<=	125598..126548	TAAtATA-(5)-ATG	0	3.64	Unknown
ORF146	<=	126553..126813	TAAAATg-(5)-ATG	0	4.10	Unknown
ORF147	=>	126870..127055	TAcAATA-(5)-gTG	0	3.63	Unknown
ORF150	=>	127979..129967	TAAAATA-(7)-ATG	2	4.31	Baseplate assembly protein
ORF152	=>	131870..134260	TAAAAaA-(6)-ATG	1	3.64	wac fibritin neck whisker
ORF154	=>	136388..137287	TAAAATg-(5)-ATG	0	4.10	Unknown
ORF159	=>	143322..143846	TAgAATA-(8)-ATG	3	4.13	Unknown
ORF164	=>	147094..147639	TAAAATA-(5)-ATG	0	4.31	Unknown
ORF165	=>	147677..148306	TAAAATg-(8)-ATG	3	4.10	Unknown
ORF168	=>	150256..151341	TAAAATg-(5)-ATG	0	4.10	Unknown
ORF169*	=>	**151284..151907**	TAtAATA-(5)-ATG	−54->0	4.02	Unknown

ORFs preceded by experimentally verified promoters are underlined.
[a] Location: genomic coordinates.
[b] Capitals: consensus nucleotides.
[c] Distance: the distance between the start of transcription and the start codon of the gene.
* Change of the distance by selection of the candidate start codon downstream of the annotated start codon.

designed to reveal predicted promoters P_{56}, P_{65}, P_{80}, and P_{152}. No primer extension products with P_{80} and P_{152} primers were observed. One should bear in mind, however, that the absence of primer extension products does not necessarily mean that no promoter is located in these regions, since we often find that several primers need to be tested in order to obtain a primer extension product in good yield. Most importantly, primer extension products with P_{56} and P_{65} primers not only matched the predicted start points but also behaved as middle transcripts (data not shown), indicating that our search reveals middle/late promoters of the phage with reasonable confidence.

Analysis of putative middle/late promoter's distribution in the genome revealed the following features. First, with an exception of genes 117 and 154, genes preceded by predicted (or experimentally shown) middle/late promoters did not have predicted −10/−35 promoters in their upstream regions (in contrast, as already mentioned, putative middle/late promoters that were excluded from Table 3 on the grounds that their ATG/GTG elements were out of frame with annotated ORFs were all located in regions harboring early promoters). Second, for most divergently transcribed genes that lacked a predicted −10/−35 promoter in front of them, a putative middle/late promoter was found upstream. Third, predicted middle/late promoters were identified in front of those genes or putatively co-transcribed gene units (operons) that behaved as middle or late on the macroarray but were not tested by primer extension. Overall, the results of middle/late promoter predictions are consistent with experimental data and further extend out understanding of phage transcription. For example, consistent with the macroarray data clustering, a predicted middle-late promoter was identified in front of a rightward-transcribed group of late genes 1–4. In the absence of such a promoter, these genes would have been grouped with early genes transcribed from the P_8-P_{15} promoters (Figure 1).

In vitro transcription from ϕYS40 promoters

The ϕYS40 middle and late promoters resemble late promoters of *E. coli* bacteriophage T4 and other T4-like phages.[14,15] These promoters contain a single

promoter element that is recognized by RNAP holoenzyme containing phage-encoded sigma factor σ^{55}.[16,17] Though ϕYS40 genome does not encode a recognizable sigma factor, it is possible that (i) ϕYS40-encoded σ is so divergent that it is not identified by bioinformatic means or (ii) ϕYS40-encoded regulators allow the σ^A RNAP holoenzyme to transcribe viral middle and late promoters at later stages of infection (or, alternatively, a phage-encoded factor prevents transcription from these promoters early in infection). To investigate this matter further and to independently confirm identification of ϕYS40 promoters, we amplified DNA fragments containing promoters identified by primer extension *in vivo* and performed *in vitro* transcription with host RNAP σ^A holoenzymes affinity purified from ϕYS40-infected or uninfected cells. Representative results are shown in Figure 5. As can be seen, transcripts from both early and middle/late promoters were observed and primer extension reactions showed that in each case *in vitro* transcription start points coincided with those determined *in vivo* (data not shown). No difference in promoter utilization by RNAP purified from infected or uninfected cells was observed. Thus, the σ^A RNAP holoenzyme from uninfected cells efficiently recognized phage middle/late promoters in the absence of added factors (conversely, the σ^A RNAP holoenzyme from ϕYS40-infected cells transcribed from early phage promoters). Likewise, *in vitro* transcription from DNA fragments containing several host promoters did not reveal any difference in transcription efficiency by RNAPs prepared from infected and uninfected *T. thermophilus* cells (data not shown).

Discussion

Here, we report the results of preliminary analysis of gene expression strategy of ϕYS40, a large bacteriophage infecting thermophilic eubacterium *T. thermophilus.* To our knowledge, this is the first time ever such an analysis was undertaken for any bacteriophage infecting a thermophilic bacterium. The approach that we used to identify early viral promoters involved bioinformatic analysis of the phage genome for the presence of sequences with similarities to host housekeeping promoters. Primer extension and *in vitro* transcription analyses showed that our search reveals viral promoters recognized by the host σ^A RNAP holoenzyme with a high degree of confidence, and macroarray and primer extension analyses showed that these promoters belong to the early temporal class of viral genes. The predicted early phage promoters are located in front of ϕYS40 genes that are expected (based on sequence similarities) to be expressed early in the infection. In addition, a large number of putative σ^A promoters were located in front of short genes with unknown function in the ϕYS40 gene cluster 3. The presence of early promoters in front of these genes suggests that the products of at least some of them may be involved in host shut-off.

In general, bioinformatic predictions of bacterial promoter sequences are not highly efficient due to degeneracy of the signal. Our success in prediction of ϕYS40 early promoters could be due to the very tight packaging of genes in the phage genomes (which increases the signal-to-noise ratio by limiting the length of "searchable" DNA in or close to the intergenic regions) and the fact that early phage promoters must be strong to efficiently compete with host promoters, which means that they are more similar to consensus promoters than most host promoters. Despite some differences in the content of promoter consensus elements, predicted ϕYS40 early promoters strongly resemble host −10/−35 promoters, as expected. A total of 86% of predicted phage early promoters have an optimal 17 bp spacer separating basal promoter elements while for host promoters (both predicted and experimentally confirmed) this value is only 59%. The difference is

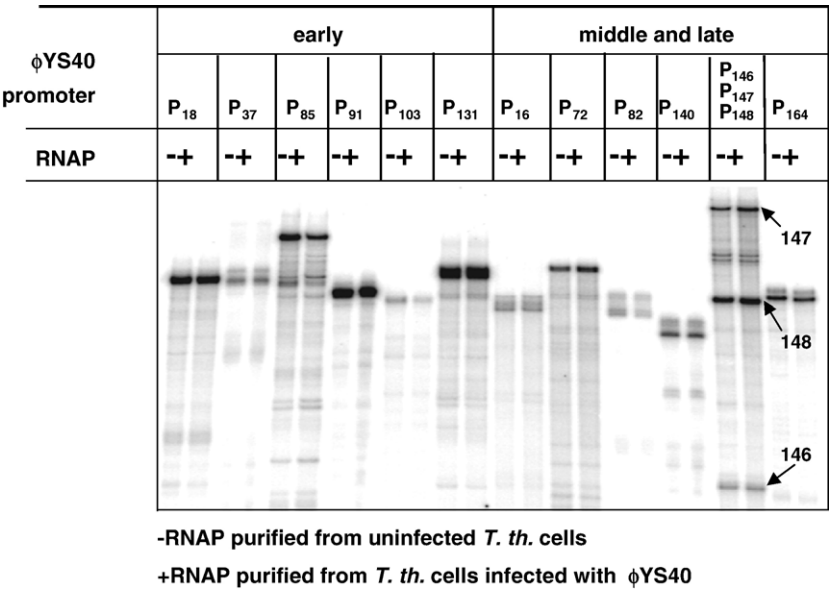

Figure 5. Transcription by *T. thermophilus* RNAP-σ^A holoenzyme. The results of multi-round run-off transcription from representative phage promoters by RNAP–σ^A holoenzymes purified from cells infected with ϕYS40 or uninfected are shown.

statistically significant at least on the level of 0.1%. The optimal spacer length of most putative phage promoters may help them to compete efficiently with host promoters for the σ^A RNAP holoenzyme.

In addition to ϕYS40 early genes, the macroarray analysis revealed the middle and late genes of the phage. By combining the information obtained by primer extension analysis of middle and late genes transcripts and by a bioinformatic search of the ϕYS40 genome, we identified ϕYS40 middle and late promoters. Though the middle and late ϕYS40 genes are clearly distinguished by our clustering analysis, at present we are unable to distinguish the middle and late promoters based on their sequences, and we consequently treated them together. A consensus ϕYS40 middle/late promoter has a single promoter element that is located about ten bases upstream of transcription start point and is similar but clearly distinct from the −10 consensus element of early phage (or housekeeping host) promoters.

The temporal regulation of gene expression of bacteriophage T4, a well-studied *E. coli* phage that is similar in size to ϕYS40, is achieved by sequential interaction of host RNAP with phage-encoded proteins that change its promoter specificity.[14,18] The middle and late T4 promoters differ from early phage promoters and from each other. The middle promoters are recognized by an RNAP holoenzyme containing the primary σ factor of the host, σ^{70}, bound to phage-encoded co-activator AsiA. The middle promoters consist of an extended –10 element (consensus sequence TGnTATAAT) and an upstream MotA box to which phage-encoded co-activator MotA binds. Late T4 promoters contain a single promoter element (consensus sequence TATAAATA), which is only recognized by a holoenzyme containing phage-encoded σ factor gp55. At least *in vitro*, middle/late promoters of ϕYS40 are efficiently recognized by *T. thermophilus* σ^A RNAP holoenzyme without any help from phage-encoded factors. This finding raises questions as to how a change in promoter specificity of host RNAP during ϕYS40 infection is achieved. Clearly, there must exist a mechanism(s) that determines decreased utilization of early promoters late in infection and, conversely, the absence of middle/late promoter utilization early in infection. Identification of ϕYS40 proteins that interact with host RNAP at different stages of infection may help to clarify the issue. However, *T. thermophilus* RNAP purified from ϕYS40-infected cells using a mild single-step affinity purification procedure has unaltered promoter specificity and does not contain any proteins other than the RNAP subunit based on visual inspection of Coomassie-stained gels (unpublished observations). Thus, unlike the straightforward case of T4, which encodes a number of proteins that bind host RNAP tightly, ϕYS40 proteins that control the switch in RNAP promoter specificity may bind host RNAP weakly. Alternatively, a change in promoter specificity could be accomplished by phage-encoded DNA-binding proteins. Since the most apparent difference between host and phage early promoters and the middle/late phage promoters is the absence of the −35 consensus element in the latter, it is possible that a product of an early phage gene shuts off host and early phage promoters by interacting with the –35 element and preventing its recognition by RNAP. A search for such a protein is currently ongoing in our laboratory.

Studies conducted with *E. coli* RNAP identified two classes of promoters, the −10/−35 class and the extended −10 class (consensus sequence TGnTATAAT). For the latter class of promoters, the properly positioned TG motif is strictly required for promoter function.[19,20] Since most ϕYS40 middle/late promoters do not have such a motif, a question arises what determines their highly efficient utilization by the σ^A holoenzyme, since the −10 consensus promoter element, TATAAT, is not sufficient for promoter utilization. Recent analysis identified an additional element recognized by *Thermus* σ^A RNAP, a downstream element GGGA that allows the recognition of the −10 element in the absence of either the −35 element or the TG motif.[21] However, the downstream element is absent from ϕYS40 promoters. Closer analysis of middle/late promoters of ϕYS40 reveals that a TG motif is present in most of them, though its distance from the −10 element varies from 4 to 0 base-pairs. SELEX experiments aimed at determining DNA sequences that strongly bind the *E. coli* σ^{70} RNAP holoenzyme revealed that fragments containing a TGTGnTATAAT sequence bind RNAP most efficiently.[22] On the other hand, analysis of single and double-stranded DNAs that specifically interact with *Thermus* σ^A and σ^A RNAP holoenzyme, respectively, indicated that a TG motif present immediately upstream of the −10 element increases the binding efficiency.[21] Thus, it is possible that the TG dinucleotide located at different distances from the −10 element may make the ϕYS40 middle/late promoters function as an extended −10 element. On the other hand, several predicted (and experimentally verified) ϕYS40 middle/late promoters lack a TG motif. It is therefore possible that the difference in sequence of the −10 element of the middle/late promoters (consensus sequence TAaAATa) and the early promoters (consensus sequence TAtnnT) allows promoter recognition in the absence of additional basal promoter elements. Alternatively, some unrecognized sequence elements may allow the middle/late promoter function and also determine their activation at an appropriate time during infection. Mutational analysis of middle/late promoters coupled with *in vitro* transcription in the presence of extracts of infected cells collected at different times post-infection will be needed to resolve these issues.

The most striking feature revealed by our analysis of middle/late transcripts of ϕYS40 is the fact that most of them appear to be leaderless. In fact, we were only successful in identifying middle/late promoters by including the initiating codon ATG/GTG into the search profile along with the −10 element consensus sequence. Searches using middle/late promoter

profiles in the absence of a requirement for a closely located start codon tended to find phage early promoters as well as many clearly irrelevant sequences (recall that unlike its host, the ϕYS40 genome is AT-rich[9]). The set of promoters revealed by our search likely includes a majority of phage middle/late promoters. However, one should bear in mind that the leaderless model constrain excluded middle promoters like P_{148} from which mRNAs containing canonical Shine–Dalgarno sequences is transcribed (these promoters, however, are in a clear minority of phage middle/late promoters).

It is formally possible that the ATG/GTG motif included in the profile of ϕYS40 middle/late promoters functions as a basal promoter element together with the −10 promoter element. This hypothesis appears unlikely though, since biologically plausible middle/late promoters invariably contained the ATG/GTG sequence in-frame with the downstream ORF. Therefore, it appears that phage middle/late transcripts are truly leaderless. In contrast, the vast majority of host as well as early phage transcripts contain Shine–Dalgarno sequences in front of their start codons and are therefore translated in a conventional way. Thus, a switch from Shine–Dalgarno-dependent to leaderless mRNA translation initiation may occur during ϕYS40 infection.

Translation of most prokaryotic mRNAs is initiated through the 30 S ribosomal subunit, which interacts with the Shine–Dalgrano sequence of the mRNA.[23] Initiation factors IF1, IF2, and IF3 regulate the kinetics of this process. Translation of leaderless mRNAs is initiated through an alternative pathway that involves the recognition of the 5′-terminal AUG codon by 70 S ribosomes.[24] Increased concentrations of IF2 enhance the efficiency of leaderless translation while increase of IF3 concentration decreases it.[25,26] In this regard, it is particularly noteworthy that while abundance of most host transcripts, including the IF3 transcript, decreased during ϕYS40 infection, the IF2 transcript behaved as a late viral gene and its abundance increased dramatically late in infection. Assuming that the change in IF2/IF3 transcript abundance reflects the change in the amount of respective proteins, the difference may provide a mechanism for the hypothetical switch in translational initiation mechanism during ϕYS40 infection.

The activation of IF2 transcription during ϕYS40 infection may occur through the same mechanism as activation of middle/late transcripts. In this regard, it would be of interest to determine if there is a difference between promoters of *T. thermophilus* genes whose transcription is activated or repressed during ϕYS40 infection.

Materials and Methods

Prediction of ϕYS40 promoters

The promoter recognition profiles were constructed using SignalX[11] implementing the formula for positional nucleotide weights.[27] Identification of candidate promoters in the phage genome was done using GenomeExplorer.[11]

Bacterial strains, phage and growth conditions

The *T. thermophilus* HB8 strain and the ϕYS40 phage were generously provided by Dr Tairo Oshima, Tokyo University of Pharmacy and Life Science. The bacterium and the phage were grown in Thermus broth (TB) medium (0.6% (w/v) Tryptone, 0.3% (w/v)yeast extract, 0.4% (w/v) NaCl, 1 mM $MgCl_2$, 0.5 mM $CaCl_2$) at 65 °C with vigorous shaking. To prepare ϕYS40 lysates, a single plaque was resuspended in 100 μl of TB, added to 50 ml of *T. thermophilus* culture (A_{600}=0.2), and cells were allowed to grow until complete lysis occurred (usually 16–20 h). The lysed culture was treated with 0.5 ml of chloroform and cell debris was removed by centrifugation at 10,000 *g* for 10 min. The resultant ϕYS40 stock ($\sim 2\times10^9$–4×10^9 p.f.u./ml) was stored at 4 °C and used to prepare larger amounts of phage lysate by scaling up the procedure described above.

E. coli strains XL-1Blue (New England Biolabs) and BL21(DE3)(Novagen) were used for molecular cloning and protein expression.

Total DNA purification and molecular cloning

ϕYS40 and *T. thermophilus* HB8 total DNA were purified by extraction with phenol/chloroform and subsequent precipitation with ethanol as described.[28]

A *T. thermophilus* HB8rpoC::10H strain containing a 10-histidine affinity tag appended to the 3′ end of the *rpoC* (which encodes the RNAP β′ subunit) was constructed as follows. First, a plasmid pET21tthC_{10H} expressing the *T. thermophilus rpoC* gene with a 3′-terminally located 10-histidine tag was created by re-cloning the corresponding PCR-modified *rpoC*-10His gene from the pET28rpoCZTth plasmid between the NdeI and EcoRI sites of pET21a (Novagen) plasmid. The pET28rpoCZTth plasmid is an expression vector bearing *rpoC* and *rpoZ* genes of *T. thermophilus* HB8 and is an intermediate created during the construction of the multi-gene plasmid co-expressing *T. thermophilus* RNAP core enzyme (K.K., unpublished results). The *T. thermophilus rpoC* gene cloned in pET28-rpoCZTth was obtained through sub-cloning of two PCR fragments, c1tth (2381 bp) and c2tth (2231 bp), in the pT7Blue (Novagen) blunt-end cloning vector. The c1tth and c2tth fragments were joined *via* a unique AvrII restriction site introduced in the primers used for amplification. The sequences of primers used for amplification are available from the authors upon request. The entire *T. thermophilus rpoC* gene was cut from pT7Blue and inserted into the pET28a expression vector between the NdeI and EcoRI restriction sites.

A 750 bp HB8 genomic fragment downstream of *rpoC* sequence with primers containing engineered SalI and HindIII sites. A fragment containing thermostable kanamycin resistance cassette (kat)[29] was amplified using plasmid pMKEβgal[30] as a template with primers containing engineered EcoRI and SalI sites. The two PCR fragments were digested with the appropriate restriction enzymes and simultaneously ligated into EcoRI-HindIII-digested pTZ19R vector, resulting in a plasmid pTZ19kat-f. The EcoRI-HindIII fragment from this plasmid was next cloned into appropriately digested pET21tthC_{10H}. The resultant plasmid, pET21tthC_{10}kat-f, contains a 10-His-tagged gene *rpoC* followed by kat cassette, which in turn is

followed by a 750 bp fragment of *T. thermophilus* chromosome downstream of *rpoC*. In order to increase efficiency of subsequent transformation into *T. thermophilus*, pET21tthC$_{10}$kat-f was transformed into and then purified from *E. coli* K12 ER2925 Dam$^-$ Dcm$^-$ strain (New England Biolabs), followed by digestion with NdeI and HindIII. The restriction digestion reaction was precipitated with ethanol and used for genetic transformation of *T. thermophilus* HB8 following the described procedure.[31] Transformants were plated onto TB plates with 1.5% (w/v) agar and 30 μg/ml of kanamycin. After a 48 h incubation at 65 °C, individual kanamycin–resistant colonies were picked up and grown in liquid TB containing 10 μg/ml of kanamycin, followed by extraction of total genomic DNA. The presence of the required insertion downstream of *rpoC* was confirmed by PCR and DNA sequencing of amplified DNA fragments. ϕYS40 infected the resultant *T. thermophilus* HB8rpoC::10H strains with an efficiency comparable to that of the original HB8 strain.

Plasmid pET28Tthσ^A contains the *T. thermophilus sigA* gene cloned between the NdeI and EcoRI sites of the pET28a expression vector and was a source of N-terminally hexahistidine-tagged σ^A.

Proteins

T. thermophilus RNAP containing C-terminally decahistidine-tagged β′ subunit was purified as follows. Cells were grown in TB medium with 10 μg/ml of kanamycin to A_{600} 0.6–0.9, harvested by centrifugation and disrupted by sonication in buffer A (10 mM Tris–HCl (pH 8.0), 500 mM NaCl, 2 mM imidazole, 5%(v/v) glycerol, 0.2 mg/ml of PMSF, 0.4 mg/ml of pepstatin). After disruption, 0.04 mg/ml of DNase I was added to the cell lysate followed by a 10 min incubation on ice. After centrifugation at 15,000 *g* for 30 min, the cleared lysate was loaded onto a chelating Hi-Trap Sepharose column (Amersham) equilibrated with Ni^{2+}. The column was washed with buffer A containing 40 mM and 80 mM imidazole and bound protein was eluted with buffer A containing 200 mM imidazole, dialyzed against buffer B (20 mM Tris–HCl (pH 8.0), 200 mM KCl, 1 mM DTT, 0.5 mM EDTA and 50% glycerol) and stored at −20 °C. The same procedure was applied for purification of RNAP from HB8rpoC::10H cells infected with ϕYS40.

To purify hexahistidine-tagged *T. thermophilus* σ^A, the pET28Tthσ^A plasmid was transformed in *E. coli* BL21 (DE3) cells and transformants were grown in 1 l of LB medium with kanamycin at 37 °C, induced with 1 mM IPTG, harvested by centrifugation, and disrupted by sonication in buffer A. The cleared cell lysate was loaded onto a chelating H-Trap Sepharose column (Amersham) equilibrated with Ni^{2+}, the column was washed with buffer A containing 20 mM imidazole and hexahistidine-tagged *T. thermophilus* σ^A was eluted with buffer A containing 200 mM imidazole, dialyzed against buffer C (20 mM Tris-HCl (pH 8.0), 200 mM NaCl, 2 mM DTT and 50% glycerol) and stored at −20 °C.

Primer extension

Exponentially growing *T. thermophilus* HB8rpoC::10H cells were infected with ϕYS40 at multiplicity of infection (MOI) of ten and harvested at various time points after infection. At the MOI of 10 used throughout the work, the efficiency of host cell infection was always greater than 95% (i.e. less than 5% of host "survivors" were detected). Total RNA was extracted with RNeasy mini kit (Qiagen) following a procedure recommended by the manufacturer. The absolute amount of total RNA extracted from 1 ml of cell culture infected at an A_{600} of 0.4 was 1.5–5 μg. For primer extension reaction, 8–10 μg of total RNA were reverse-transcribed with 100 units of SuperScript III enzyme from the First-Strand Synthesis kit for RT-PCR (Invitrogen) according to the manufacturer's protocol in the presence of 1 pmol of ^{32}P end-labeled primer. The reactions were treated with RNase H, precipitated with ethanol and dissolved in formamide-containing loading buffer. To identify primer extension products, the sequencing reaction (with the fmol DNA Cycle Sequencing kit from Promega) was performed from a corresponding PCR fragment amplified from the ϕYS40 genome using the same end-labeled primer as that used for primer extension. The reaction products were resolved on 7% (w/v) sequencing gels and visualized using a PhosphorImager (Molecular Dynamics). The sequences of the primers are available from the authors upon request.

In vitro transcription

Multiple-round run-off reactions contained, in 10 μl of standard transcription buffer (40 mM Tris–HCl (pH 8.0), 40 mM KCl, 10 mM $MgCl_2$, 3 mM β-mercaptoethanol), 20 nM of *T. thermophilus* HB8rpoC::10H RNAP core enzyme saturated with 40 nM of *T. thermophilus* σ^A and 2–4 nM of PCR fragments containing ϕYS40 promoters. Reactions were incubated for 10 min at 65 °C, followed by the addition of ATP, CTP, and UTP (0.2 mM each), 20 μM GTP and 3 μCi of [α−^{32}P]GTP (3000 Ci/mmol). Reactions proceeded for 7 min at 65 °C and were terminated by the addition of an equal volume of formamide-containing loading buffer. The reaction products were resolved on a 7% denaturing polyacrylamide gel and visualized using a PhosphorImager.

In vitro transcription reactions for subsequent primer extension analysis contained, in 50 μl of transcription buffer, 40 nM of *T. thermophilus* RNAP core enzyme, 80 nM of *T. thermophilus* σ^A and 6–12 nM of PCR fragments containing ϕYS40 promoters. Reactions were performed as described above, and nucleic acids were precipitated with ethanol and dissolved in Rnase-free water. The reaction products were then used in primer extension reactions as described above.

Macroarray membrane preparation and hybridization

DNA fragments corresponding to each of the selected ϕYS40 genes, *T. thermophilus* HB8 housekeeping genes, and *D. melanogaster zfrp8* gene were amplified from corresponding genomic DNA using gene-specific primer pairs. The sequences of the primers are available from the authors upon request. Membrane preparation, cDNA synthesis and macroarray hybridization were performed as described.[2]

Macroarray data analysis

After hybridization the amount of radioactivity from each spot was quantified using PhosphorImager-generated image files that were analyzed by using the ImageQuant (Molecular Dynamics) software. The background signal was subtracted from signals corresponding to every ORF spot. To allow comparison between the signals on different membranes, the background-corrected

signals were normalized relative to the average of the two *D. melanogaster zfrp8* spot signals. The normalized signals were used in further data analysis.

Acknowledgements

The *Thermus thermophilus* HB8 strain and the ϕYS40 phage were generously provided by Dr Tairo Oshima (Tokyo University of Pharmacy and Life Science). We are grateful to Dr J. Berenguer (Universidad Autonoma de Madrid) for plasmid pMKEβgal, Dr L. Westblade (Rockefeller University) for help with construction of the *T. thermophilus* strain HB8rpoC::10H and Dr S. Minakhina (Waksman Institute for Microbiology) for the *D. melanogaster zfrp8* gene DNA fragment. This work was supported by grants GM59295 and GM64530 from NIH (to K.S.). M.D. acknowledges support from NSF under agreement no. 112050 and NSF grant MCB- 8690418891. M.S.G. was partially supported by grants from the Howard Hughes Medical Institute (55005610), INTAS (05-1000008-8028), and the Russian Academy of Sciences (Program "Molecular and Cellular Biology").

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.11.050

References

1. Ventura, M., Foley, S., Bruttin, A., Chennoufi, S. C., Canchaya, C. & Brussow, H. (2002). Transcription mapping as a tool in phage genomics: the case of the temperate *Streptococcus thermophilus* phage Sfi21. *Virology*, **296**, 62–76.
2. Minakhin, L., Semenova, E., Liu, J., Vasilov, A., Severinova, E., Gabisonia, T. *et al.* (2005). Genome sequence and gene expression of *Bacillus anthracis* bacteriophage Fah. *J. Mol. Biol.* **354**, 1–15.
3. Semenova, E., Djordjevic, M., Shraiman, B. & Severinov, K. (2005). The tale of two RNA polymerases: transcription profiling and gene expression strategy of bacteriophage Xp10. *Mol. Microbiol.* **55**, 764–777.
4. Duplessis, M., Russell, W. M., Romero, D. A. & Moineau, S. (2005). Global gene expression analysis of two *Streptococcus thermophilus* bacteriophages using DNA microarray. *Virology*, **340**, 192–208.
5. Djordjevic, M., Semenova, E., Shraiman, B. & Severinov, K. (2006). Quantitative analysis of a virulent bacteriophage transcription strategy. *Virology*, **354**, 240–251.
6. Peng, X., Blum, H., She, Q., Mallok, S., Brugger, K., Garrett, R. A. *et al.* (2001). Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. *Virology*, **291**, 226–234.
7. Haring, M., Vestergaard, G., Rachel, R., Chen, L., Garrett, R. A. & Prangishvili, D. (2005). Virology: independent virus development outside a host. *Nature*, **436**, 1101–1102.
8. Prangishvili, D., Garrett, R. A. & Koonin, E. V. (2006). Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res.* **117**, 52–67.
9. Naryshkina, T., Liu, J., Florens, L., Swanson, S. K., Pavlov, A. R., Pavlov, N. V. *et al.* (2006). *Thermus thermophilus* bacteriophage ϕYS40 genome and proteomic characterization of virions. *J. Mol. Biol.* **364**, 667–677.
10. Sakaki, Y. & Oshima, T. (1975). Isolation and characterization of a bacteriophage infectious to an extreme thermophile, *Thermus thermophilus* HB8. *J. Virol.* **15**, 1449–1453.
11. Mironov, A. A., Vinokurova, N. P. & Gelfand, M. S. (2000). Software for analysis of bacterial genomes. *Mol. Biol. (Mosk.)*, **34**, 222–231.
12. Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097–6100.
13. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190.
14. Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. & Ruger, W. (2003). Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **67**, 86–156.
15. Nolan, J. M., Petrov, V., Bertrand, C., Krisch, H. M. & Karam, J. D. (2006). Genetic diversity among five T4-like bacteriophages. *Virol. J.* **3**, 30–45.
16. Kassavetis, G. A. & Geiduschek, E. P. (1984). Defining a bacteriophage T4 late promoter: bacteriophage T4 gene 55 protein suffices for directing late promoter recognition. *Proc. Natl Acad. Sci. USA*, **81**, 5101–5105.
17. Kolesky, S. E., Ouhammouch, M. & Geiduschek, E. P. (2002). The mechanism of transcriptional activation by the topologically DNA-linked sliding clamp of bacteriophage T4. *J. Mol. Biol.* **321**, 767–784.
18. Brody, E. N., Kassavetis, G. A., Ouhammouch, M., Sanders, G. M., Tinker, R. L. & Geiduschek, E. P. (1995). Old phage, new insights: two recently recognized mechanisms of transcriptional regulation in bacteriophage T4 development. *FEMS Microbiol. Letters*, **128**, 1–8.
19. Camacho, A. & Salas, M. (1999). Effect of mutations in the "extended −10" motif of three *Bacillus subtilis* sigmaA-RNA polymerase-dependent promoters. *J. Mol. Biol.* **286**, 683–693.
20. Burr, T., Mitchell, J., Kolb, A., Minchin, S. & Busby, S. (2000). DNA sequence elements located immediately upstream of the −10 hexamer in *Escherichia coli* promoters: a systematic study. *Nucl. Acids Res.* **28**, 1864–1870.
21. Feklistov, A., Barinova, N., Sevostyanova, A., Heyduk, E., Bass, I., Vvedenskaya, I. *et al.* (2006). A basal promoter element recognized by free RNA polymerase sigma subunit determines promoter recognition by RNA polymerase holoenzyme. *Mol. Cell*, **23**, 97–107.
22. Gaal, T., Ross, W., Estrem, S. T., Nguyen, L. H., Burgess, R. R. & Gourse, R. L. (2001). Promoter recognition and discrimination by EsigmaS RNA polymerase. *Mol. Microbiol.* **42**, 939–954.
23. Laursen, B. S., Sorensen, H. P., Mortensen, K. K. & Sperling-Petersen, H. U. (2005). Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **69**, 101–123.
24. Moll, I., Grill, S., Gualerzi, C. O. & Blasi, U. (2002). Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol. Microbiol.* **43**, 239–246.
25. Tedin, K., Moll, I., Grill, S., Resch, A., Graschopf, A.,

Gualerzi, C. O. & Blasi, U. (1999). Translation initiation factor 3 antagonizes authentic start codon selection on leaderless mRNAs. *Mol. Microbiol.* **31**, 67–77.
26. Grill, S., Moll, I., Hasenohrl, D., Gualerzi, C. O. & Blasi, U. (2001). Modulation of ribosomal recruitment to 5′-terminal start codons by translation initiation factors IF2 and IF3. *FEBS Letters*, **495**, 167–171.
27. Mironov, A. A., Koonin, E. V., Roytberg, M. A. & Gelfand, M. S. (1999). Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucl. Acids Res.* **27**, 2981–2989.
28. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edit. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
29. Matsumura, M., Katakura, Y., Imanaka, T. & Aiba, S. (1984). Enzymatic and nucleotide sequence studies of a kanamycin-inactivating enzyme encoded by a plasmid from thermophilic bacilli in comparison with that encoded by plasmid pUB110. *J. Bacteriol.* **160**, 413–420.
30. Moreno, R., Zafra, O., Cava, F. & Berenguer, J. (2003). Development of a gene expression vector for *Thermus thermophilus* based on the promoter of the respiratory nitrate reductase. *Plasmid*, **49**, 2–8.
31. Koyama, Y., Hoshino, T., Tomizuka, N. & Furukawa, K. (1986). Genetic transformation of the extreme thermophile *Thermus thermophilus* and of other *Thermus spp*. *J. Bacteriol.* **166**, 338–340.
32. Hartmann, R. K. & Erdmann, V. A. (1989). *Thermus thermophilus* 16S rRNA is transcribed from an isolated transcription unit. *J. Bacteriol.* **171**, 2933–2941.

Edited by M. Gottesman

(Received 7 September 2006; received in revised form 3 November 2006; accepted 14 November 2006)
Available online 18 November 2006