6172–6187 *Nucleic Acids Research, 2005, Vol. 33, No. 19*
doi:10.1093/nar/gki915

Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases

Zehua Chen and Thomas D. Schneider*

Center for Cancer Research Nanobiology Program, National Cancer Institute at Frederick, PO Box B, Frederick, MD 21702-1201, USA

Received July 6, 2005; Revised August 7, 2005; Accepted October 4, 2005

ABSTRACT

Molecular information theory was used to create sequence logos and promoter models for eight phages of the T7 group: T7, ϕA1122, T3, ϕYeO3-12, SP6, K1-5, gh-1 and K11. When these models were used to scan the corresponding genomes, a significant gap in the individual information distribution was observed between functional promoter sites and other sequences, suggesting that the models can be used to identify new T7-like promoters. When a combined 76-site model was used to scan the eight phages, 108 of the total 109 promoters were found, while none were found for other T7-like phages, ϕKMV, P60, VpV262, SIO1, PaP3, Xp10, P-SSP7 and Ppu40, indicating that these phages do not belong to the T7 group. We propose that the T7-like transcription system, which consists of a phage-specific RNA polymerase and a set of conserved T7-like promoters, is a hallmark feature of the T7 group and can be used to classify T7-like phages. Phylogenetic trees of the T7-like promoter models and their corresponding RNA polymerases are similar, suggesting that the eight phages of the T7 group can be classified into five subgroups. However the SP6-like polymerases have apparently diverged from other polymerases more than their promoters have diverged from other promoters.

INTRODUCTION

Transcription plays a key role in the expression of genetic information (1). Many double-stranded DNA bacteriophages utilize the transcription system of their hosts. However, a group of phages within the *Podoviridae* family share a common phylogenetic origin with bacteriophage T7 (2); members of this group mainly utilize their own transcription systems, which consist of a phage-specific RNA polymerase (RNAP) and a set of conserved promoters scattered across the phage genome. The phage RNAP is encoded by a single gene in the early region of the phage genome. In the early stage of infection, the host RNAP recognizes promoters located near one end of the bacteriophage DNA and transcribes a segment of ∼20% of the genome, which includes the phage RNAP gene. During the next stage of infection, the phage RNAP specifically recognizes its cognate promoters on the phage genome and carries out all the steps of transcription without any accessory factors. The phage transcription system not only plays an important role in phage gene expression, but also drives the efficient translocation of viral DNA into the host cells. This strategy of infection is characteristic of the T7 group of phages (3). Because the T7 RNA polymerase is simple and T7 promoters are only recognized by their cognate polymerase, a lot of work has been done to understand transcription in T7 and the related phages T3 and SP6 (4–8).

Historically, classification of phages relied on morphology and host range, as well as genome size and type, resulting in the International Committee on the Taxonomy of Viruses (ICTV) taxonomic system (9). Owing to the rapid accumulation of phage genomic data and the absence of morphology and culturing data, a genome-based classification was proposed recently (10). Many phages have been shown to evolve by exchanging modules (11), so a taxonomy based on individual modules was also proposed (12). A group of phylogenetically related phages usually share some unique features, which could also be used for classification. For example, the phage RNAP has been recognized as a hallmark feature for the T7 group of phages (13,14). The T7-like promoters, recognized by the phage-specific RNAP, are also an important feature, but they have not been used for classification. Any phage-specific protein could be obtained by a single modular exchange during evolution and so might not be useful for

*To whom correspondence should be addressed. Tel: +1 301 846 5581; Fax: +1 301 846 5598; Email: toms@ncifcrf.gov

classification by itself. In contrast, a set of T7-like promoters, which are spread across a phage genome, suggests that the genome is organized like type phage T7 or that there was *de novo* appearance of the sites. In this paper we propose that these promoters can be used in conjunction with the phage RNAP as a criterion for classification of T7-like phages.

In recent years, many phage genome sequences have been determined, including more than 10 T7-like phages. The coliphages T7 (15), T3 (16) and K1-5 (17), the yersiniophages ϕYeO3-12 (18) and ϕA1122 (19), the *Salmonella* phage SP6 (17,20) and the *Pseudomonas putida* phage gh-1 (21) are recognized members of the T7 group. Several other phages were also suggested to be related to phage T7 by genome similarity. The *Pseudomonas aeruginosa* phage ϕKMV (22) and cyanophage P60 (23) are thought to be T7-like because they contain >10 T7-like genes (including a T7-like RNAP gene). The *Vibrio parahaemolyticus* phage VpV262 (24) and rosephage SIO1 (25) lack an RNAP gene in their genome, but contain a T7-like head structure module; thus, they have been thought to be members of the T7 supergroup (24). The *Klebsiella* phage K11 (26,27), Kluyveraphage Kvp1 (28), coliphage BA14 (29,30) and several others (3,31,32) were clearly shown to be closely related to phage T7, though no complete genome sequences are available for them.

Predicting promoters in the T7 group of phages is relatively easy, because the promoters are well conserved (33). Some previous predictions were made using a string-based search algorithm (34). However, counting mismatches to a consensus sequence is often misleading (35). Molecular information theory gives more sensitive weights than mismatch counts (35), so it can be used to precisely characterize the sequence conservation at nucleic acid-binding sites, thus providing better models for molecular biologists. Because it provides a universal scale for sequence conservation in bits, molecular information theory has been widely applied to many genetic systems, and it has provided much insight (33,36–38).

In this study, we used information theory (39) to build promoter models for eight phages of the T7 group. Combined models were also built. These models were used to identify new T7-like promoters and to construct phylogenetic trees.

According to the ICTV taxonomic system, the T7-like genus was defined within the *Podoviridae* family as phages that encode a phage-specific RNAP (13,14). Based on this, 10 phages were classified into the T7-like genus, and 16 phages were tentatively assigned to be members of this genus (http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv). There are several features in common for T7 group phages, including terminal redundancy, genome size and morphology. However, the strategy of infection is unique to the T7 group of phages, and this strategy is largely determined by the transcription system, so, following Hausmann (2) we propose that the phage-specific RNAP and a set of conserved T7-like promoters can be jointly used as a hallmark feature to classify a phage as a member of the T7 group or not. In this classification scheme, phages that lack an RNAP gene and/or a set of conserved T7-like promoters, but share at least one genetic module with the members of the T7 group, are assigned to be members of the T7 supergroup but not the T7 group (24). Based on the T7-like promoter models and data from the literature, a classification of more than six subgroups within the T7 group is proposed (Figure 8).

In this study we observed that, although the promoters must coevolve with their polymerases, the relative genetic distance of the DNA-binding sites and their proteins are not the same. That is, neutral drift of the whole proteins is not the same as functional changes of their corresponding DNA-binding sites. We suggest two hypotheses to explain this discrepancy.

MATERIALS AND METHODS

Programs

Most programs used in this work are available at http://www.ccrnp.ncifcrf.gov/~toms/. Other programs used include Blast2 (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html) (40), GAP and PileUp (from the GCG package) (41), dnadist, protdist, neighbor and drawtree (from the PHYLIP 3.6a3 package, which was written by J. Felsenstein and distributed by the University of Washington, http://evolution.gs.washington.edu/phylip.html) (42).

Building promoter models for T7-like phages

The promoter sequences were aligned and sequence logos were made using the programs delila, alist, encode, rseq, dalvec and makelogo (43,44). Information theory based weight matrix models were constructed using the ri program (39).

To create combined models, delila instructions for different phage promoters were concatenated and the promoter sequences were first extracted using delila. Then the combined models were built using the above mentioned programs. The most closely related models were combined, then all models were combined into a single model.

Genome scanning

Genome scanning was carried out by the program scan (39) with each of the phage promoter models or combined models. The programs genhis, genpic and xyplo were used to plot the individual information distribution (39).

Since the T7-like promoters are asymmetric, all genomes were scanned on both strands. However, in all members of the T7 group described so far, all transcripts and encoded proteins originate from only one strand. Therefore, the T7-like phages were also scanned on the noncoding strand alone to give an internal negative control for each scan.

Phylogenetic analysis

The promoters used are listed in Figure 1. How this set was obtained is described in Results.

The base frequency matrices were used to construct a phylogeny for the eight promoter models. The program diffribl was used to calculate the distances among the promoter models (from base position −20 to +5). First, the base frequencies at each position of a promoter model were represented in 3D space, based on Zhang's method (45). The positional distances were calculated in that probability space and then summed across the sites to give a distance between two matrices. An unrooted phylogenetic tree was constructed based on the resulting distance matrix by the program neighbor from the PHYLIP 3.6a3 package (42). The program drawtree was used to plot the tree. In the alternative Euclidean distance method, the 4D distance between individual information weights at a

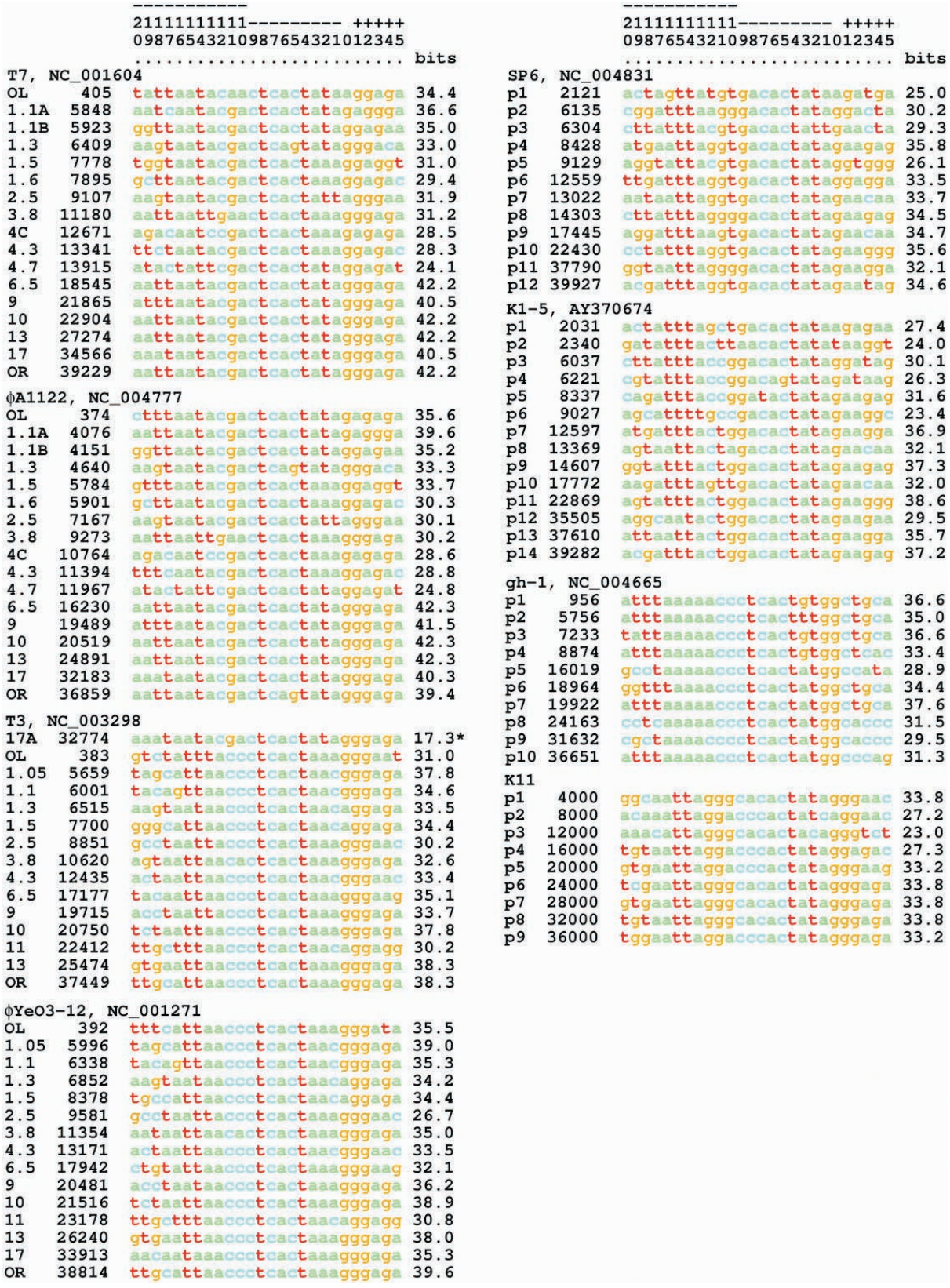

Figure 1. Collection and alignment of 109 T7-like promoters. The promoters were collected from each of the eight phages of the T7 group, and aligned from −20 to +5 (shown on the top of the alignment), relative to the transcription start (at 0). Each section contains promoters from one of the eight phages; the phage names and genome accession numbers are given in the beginning of each section; the promoter names and coordinates are also given. The individual information of each site, given in bits, is based on the promoter model for that phage. 17A (marked by a star) is a ϕA1122-like promoter in the T3 genome. It was not used to build the T3 model and gave only 17.3 bits by the T3 model. It is 40 bits by the ϕA1122 and T7 models (Figure 5).

position were computed, and these were summed across the entire matrix.

Phylogenetic analysis was also performed on the amino acid sequences for single-subunit RNAPs of coliphages T7 (GenBank accession number: NP_041960) (46), T3 (NP_523301) (47) and K1-5 (AAL86891) (48), yersiniophages ϕA1122 (NP_848264) (19) and ϕYeO3-12 (NP_052071) (18), *Salmonella* phage SP6 (NP_853568) (49,50), *P.putida* phage gh-1 (NP_813747) (21,51) and *Klebsiella* phage K11 (CAA37330) (27,52). In addition, three T7-like RNAPs from *Pseudomonas* phage ϕKMV (NP_877465) (22), cyanophage P60 (NP_570316) (23) and *Xanthomonas oryzae* phage Xp10 (NP_858979) (53,54) were included. Three T7-like RNAPs that were found in bacterial genomes, *P.putida* strain KT2440 (NP_744415) (55,20), *Agrobacterium tumefaciens* strain C58 (NP_531879) (56) and *Rhodopseudomonas palustris* strain CGA009 (NP_947869) (57) were also included in the analysis.

The protein sequences were aligned by the program PileUp available in the GCG package (41). The program protdist was used to calculate the distances among these proteins based on the Jones–Taylor–Thornton amino acid replacement model (58), and the resulting distance matrix was used to infer a tree using the neighbor-joining method (59) by the program neighbor. For both protdist and neighbor we used default parameters. The program drawtree was used to plot an unrooted tree. A bootstrap analysis of 1000 replicates was performed to test the statistical significance of branches in the tree (60). The GCG program GAP was used to calculate the per cent similarity and identity among the proteins based on the alignment scoring matrix BLOSUM62 (61), and the result is available in Supplementary Table S1.

RESULTS

Collection of T7-like promoters and building promoter models for T7 group phages

Since T7-like promoters are well conserved, the promoter model should significantly distinguish its own promoters from the background when the model is used to scan its own genome. We used this as a criterion to build T7-like promoter models. Briefly, we first build an initial model by using known promoters or sites that are picked up by a closely related model. We then refine the initial model by incorporating intermediate sites or by removing sites that fall into the background of genomic sequence until we obtain a model that can significantly distinguish the sites in the model from the background. Because of the high information content of T7 promoters (4,33), the sites are significantly separated from the background. As a result, this process is straightforward and we found that it can be applied to other T7-like promoters.

A total of 109 T7-like promoters (including a T7 promoter in the T3 genome) were extracted from eight phages of the T7 group. The promoters were aligned (Figure 1) and respective sequence logos and promoter models were built (Figure 2) as described previously (33,39,43,44). For the promoter alignments in Figure 1 and sequence logos in Figure 2, we used a numbering system with position 0 to indicate the transcription start. Position 0 is always the coordinate of reported binding site locations.

The 17 known T7 promoters from the phage T7 genome (NC_001604) (15) were used to build the T7 promoter model. The average information content at the T7 promoters is 34.9 ± 5.9 bits, where we report the SD of the individual information distribution (39). When the T7 model is scanned over the T7 genome, the highest site, excluding the known promoters, is 13.6 bits. Therefore the background is 3.6 SD below the mean. The probability of inappropriately picking up the background is 1.5×10^{-4}. Therefore the T7 promoters are significantly different from the background.

The T7 model was used to scan the closely-related phage ϕA1122 genome (19). Above 24 bits 17 sites were picked up (the other sequences are lower than 13 bits) at positions equivalent to those in T7. These 17 sites were used to build the ϕA1122 promoter model, which looks almost identical to the T7 model (Figure 2). In fact, 12 T7 promoters are identical to their counterparts of ϕA1122 promoters from −20 to +5. Furthermore, promoters 6.5, 10 and 13 in both genomes and OR in T7 are identical in this range (Figure 1).

The 14 known T3 promoters (16) were used to build the T3 model. The phage ϕYeO3-12 has been shown to be closely related to the phage T3 (18,62), so the T3 model was used to scan the ϕYeO3-12 genome. Above 30 bits 15 sites were picked up (the other sequences are lower than 16 bits) and used to build the ϕYeO3-12 model, which is almost identical to the T3 model (Figure 2). In fact, 10 pairs of the T3 and ϕYeO3-12 promoters are identical from −20 to +5 (Figure 1).

The phage SP6 is a distantly-related member of the T7 group (17,20,50), so the T7, ϕA1122, T3 and ϕYeO3-12 models cannot distinguish SP6 promoters from the background (Figure 5). Four known SP6 promoters (63,64) were aligned against the SP6 genome sequence (17,20) by the program Blast2 (40) to reveal four more SP6 promoters. When a model built with the eight sites was used to scan the genome sequence, four more sites with an individual information above 20 bits were picked up. The model built from the 12 sites can pick up themselves above 25 bits and the background is below 15 bits (Figure 5E), so the 12-site model was used as the SP6 model (Figure 2).

To confirm our SP6 model we started instead from a published frequency table of 11 SP6 promoters (65) to build a temporary model, which was then used to scan the SP6 genome. Above 25 bits 11 sites and one site of 21.9 bits were found. These 12 sites correspond to the sites in the 12-site model mentioned above. A model built with the 11 sites gives the same frequency table as in (65), so we believe these 11 sites are the 11 SP6 promoters described by Lee and Kang (65). Scholl *et al.* (17) predicted 12 promoters in the SP6 genome, 11 of which are the same as predicted in this study, but the other site at position 29410 has only 12 bits. This site contains CGC from −17 to −15, while according to a saturation mutagenesis analysis (8), a positive control had ATT over the same range, so this site may not be functional or is weak. Using a consensus sequence Dobbins *et al.* (20) predicted 10 SP6 promoters, but they missed the one at 6304, which departs from their consensus promoter at position −1 by having a T instead of A. This site should be functional based on the saturation mutagenesis analysis (8). One site at 2121, which was predicted in the present study (Figure 1), departs from all other sites by having an A at position 0 instead of a G,

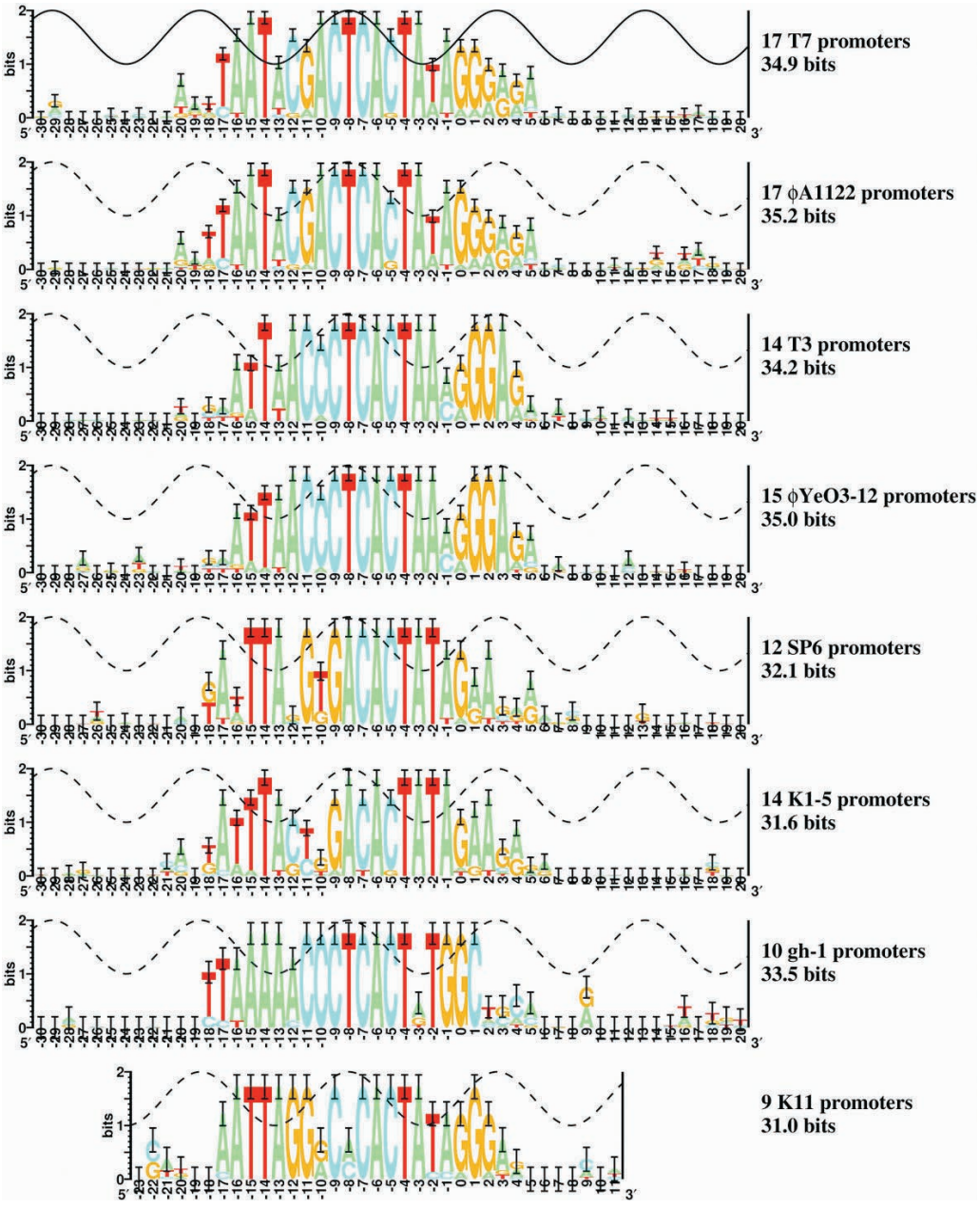

Figure 2. Sequence logos for promoters of eight T7 group phages. The phage genome and the number of sites are given for each logo. In these sequence logos, the height of each letter is proportional to the frequency of that base at each position, and the height of the letter stack is the conservation in bits (43). The sine wave on each logo represents the 10.6 base helical twist of B-form DNA (36,44). Position −8 of T7 promoters faces the polymerase (88) (solid wave) while this has not been shown for other sites (dashed waves). Information content (*Rs*) was calculated from −20 to +5 for the upper six logos, −18 to +5 for the gh-1 logo and −17 to +4 for the K11 logo. Position 0 is the transcription start.

so it might not be functional according to Shin and Kang's analysis (8). However, Shin and Kang's analysis was based on an SP6 consensus promoter that contains an A at +1. The site at 2121 contains a G at +1, so it could be functional, because a G to A variant at position 0 of T7 promoter 10 retains 33% of activity (6). Furthermore, two T7 promoters (OL and 2.5) and three T3 promoters (1.3, 1.5 and 11) also contain an A at 0 and a G at +1 (Figure 1). Finally, the rest of the site at 2121 clearly resembles the other SP6 promoters (Figure 1).

The phage K1-5 has been reported to be closely related to phage SP6 (17,48), so the SP6 model was used to scan the K1-5 genome (17). Above 16 bits 16 sites were picked up and used to build an initial K1-5 model. When the model built from the 16 sites was used to scan the K1-5 genome, 14 sites were picked up above 23 bits; the next best sites, of 17 and 18 bits, were discarded as being too weak. The 14 sites were used to build the K1-5 model, which picks up all 14 sites above 23 bits, while the background is below 14 bits, so the 14-site model was considered as the K1-5 model. The K1-5 model looks highly similar to the SP6 model except for the region from −12 to −10 (Figure 2). The most contradicting base is at position −11. All K1-5 promoters contain a pyrimidine at

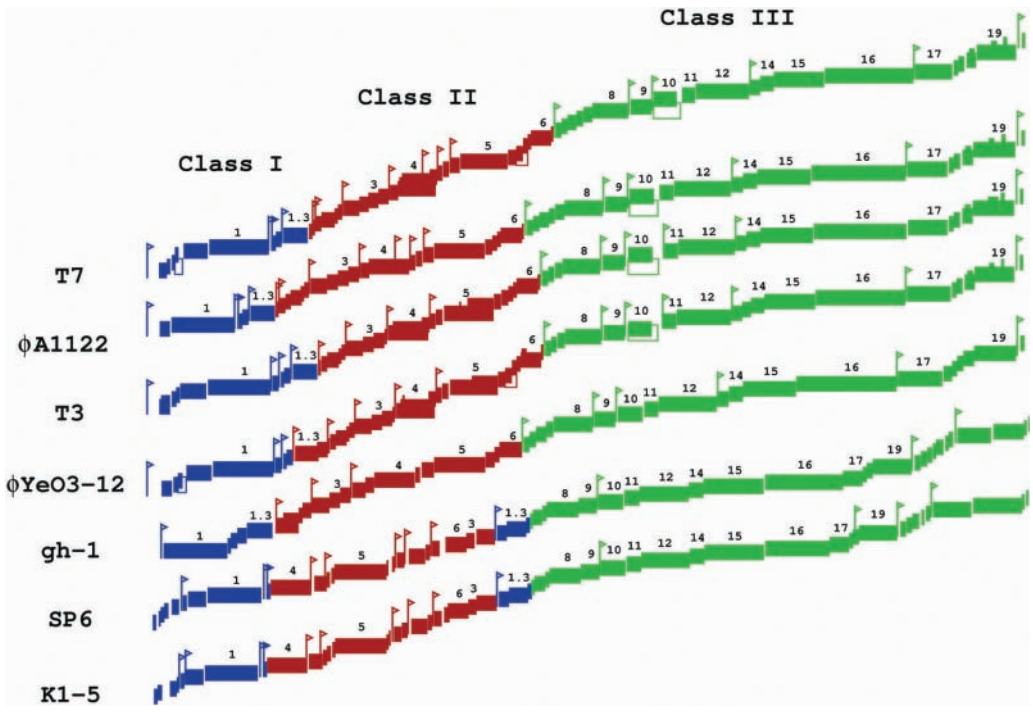

Figure 3. Genomic organization of seven T7 group phages and locations of T7-like promoters used in models. Genes are represented by rectangles that are displaced upwards after each gene to show the genomic structure, and different colors indicate different genomic regions as marked. Gene numbering was given for orthologous genes of these seven phages. Known or predicted promoters are marked as vertical lines with flags. Genome lengths were normalized.

this position, suggesting that SP6 RNAP may not be functional on K1-5 promoters (8), while it remains to be confirmed whether the K1-5 RNAP will recognize SP6 promoters. Besides the 14 sites, Scholl *et al.* (17) predicted one more site at 31054; however, this site has an information of 5 bits by the 14-site model. Even if this site is included in the 14-site model, its information is only 8 bits, so this site was excluded.

Phage gh-1 has been shown to be related to phage T7 (51) and the genome sequence confirmed this (21). However, phage gh-1 is less related to other members of the T7 group (T7, T3 and K11) than they are to each other (Supplementary Figure S4), so the other T7-like promoter models could not be used to identify the gh-1 promoters. In GenBank, 17 putative promoters were annotated for the phage gh-1 genome sequence (NC_004665). These 17 putative promoters were used to build an initial gh-1 promoter model, which was then used to scan the gh-1 genome sequence. Above 24.5 bits 10 sites were picked up, while the other sites are below 14 bits. A model built from the 10 sites can pick up all 10 sites above 28 bits, while the other sites are below 17 bits. Furthermore, these 10 sites are the same as Kovalyova and Kropinski predicted (21). So the 10-site model was viewed as the gh-1 model. It shares a conserved region from −13 to −4 with the T3 and ϕYeO3-12 models (Figure 2).

Though the genome sequence of *Klebsiella* phage K11 is not available, nine phage promoters have been identified experimentally (26,27). The nine published promoters from position −23 to +12 were collected and embedded in 40 kb of random sequence to simulate a genome for subsequent scanning. The nine promoters were extracted and aligned from position −23 to +11, which is exactly the same alignment as shown in reference (26). A logo built from the nine sites is similar to the other seven logos (Figure 2).

In summary, a total of 108 promoters were included in the eight promoter models described above. A genomic map with promoter locations for seven of these phages is shown in Figure 3. By comparing the promoter locations in different genomes, we can see that most promoters have equivalent locations and almost all are located between gene coding regions. This comparison confirmed our predictions and strongly indicated that these phages have a similar strategy of transcriptional control.

Combined models for T7-like promoters

Generally we do not combine binding site models from different organisms because the recognizers may have different DNA site domains and altered binding site affinities. Combined models will often have lower information than either model independently, so that a heterologous model may pick up false sites or miss true sites. However the individual T7-like promoter models are so highly conserved that a combined model can be fruitfully used to identify T7-like promoters (see below). Once such sites are identified, they can be used to produce a model that exclusively represents one recognizer. That is, we use combined models as stepping stones to obtain native models.

Combined models were built from several sets of promoters (Figure 4). First, three pairs of closely related models were combined: T7 and ϕA1122, T3 and ϕYeO3-12, and SP6 and K1-5. Then all eight individual models were combined into a single model, which contains all 109 promoters from the eight phages. When two binding site sets are combined, an increase

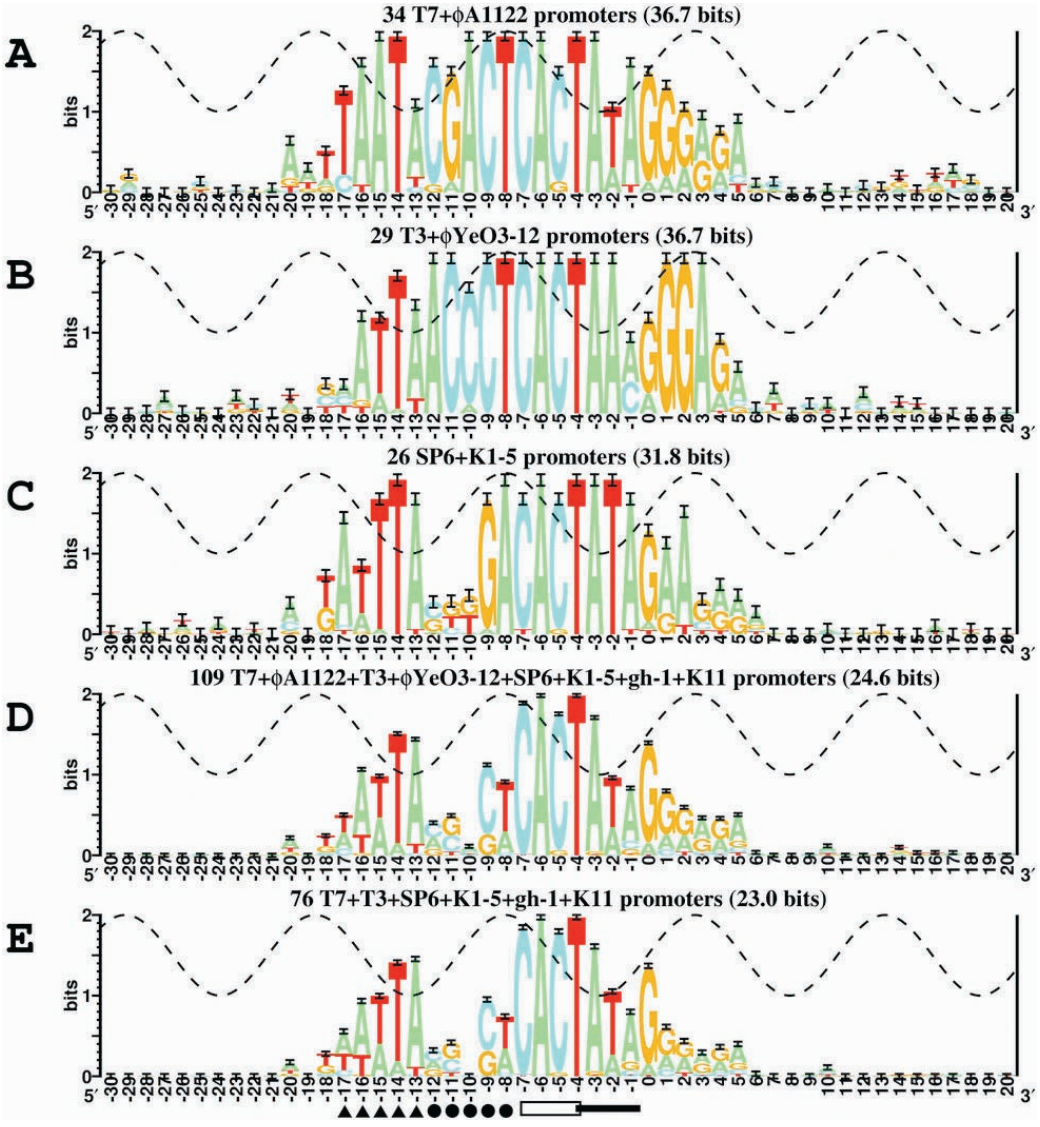

Figure 4. Different combined models of T7-like promoters. Information contents were calculated from −20 to +5 for all combined models. The AT-rich region (−17 to −13, black triangles), 'specificity' region (−12 to −8, black circles), promoter core (−7 to −4, open box) and melting region (−4 to −1, black rectangle) are marked according to (89,90).

in information content in the combined model relative to its parts suggests that the parts are closely related and a decrease clearly indicates that the sets are not so closely related.

When the T7 and ϕA1122 model are combined, the resulting model contains a slightly higher information content (36.7 bits) than each model alone (Figure 4A); the 29-site T3 and ϕYeO3-12 model is 36.7 bits, also higher than each model alone (Figure 4B). For the above two pairs of phages, one phage RNAP should be able to cross-recognize all or most promoters from the other phage.

The SP6 and K1-5 models were combined into a 26-site model, which contains about the same information as each model alone (Figure 4C). Though these two models are highly similar to each other, the RNAPs might not cross-recognize their promoters because these two models contain different bases at position −11 (8).

Finally, all eight models (containing 108 promoters) and a T7 promoter (17A) in the T3 genome (66) were combined into a single 109-site model, the information dropped to 24.6 bits (Figure 4D). Since the 109-site model is biased towards the T7, ϕA1122, T3 and ϕYeO3-12 models, a less biased 76-site model was built by removing the ϕA1122 and ϕYeO3-12 promoters from the 109-site model (Figure 4E).

Comparing different combined models, we can see that trying to give a consensus sequence for all T7-like promoters would be misleading: even the 109-site and 76-site models are biased and cannot be viewed as universal models for identifying all T7-like promoters. If more T7-like promoters are identified and included in a combined model, the conservation (information) would probably be low compared with current combined models.

T7 group genome scanning with the T7-like models

We scanned many genomes with the eight T7-like models and the combined 76-site model. First, each of the nine models was

used to scan the eight phages of the T7 group. Then the nine models were used to scan seven other distantly related T7-like phages and a prophage. These models were also used to scan genome survey sequences from several uncultured viral communities (67–70).

When scanning each of the eight genomes of the T7 group with its own model or a closely related model, a significant gap of 8–14 bits in the individual information (*Ri*) distribution was observed between the promoter sites and the noncoding-strand background or other sequences (Figure 5, double-headed

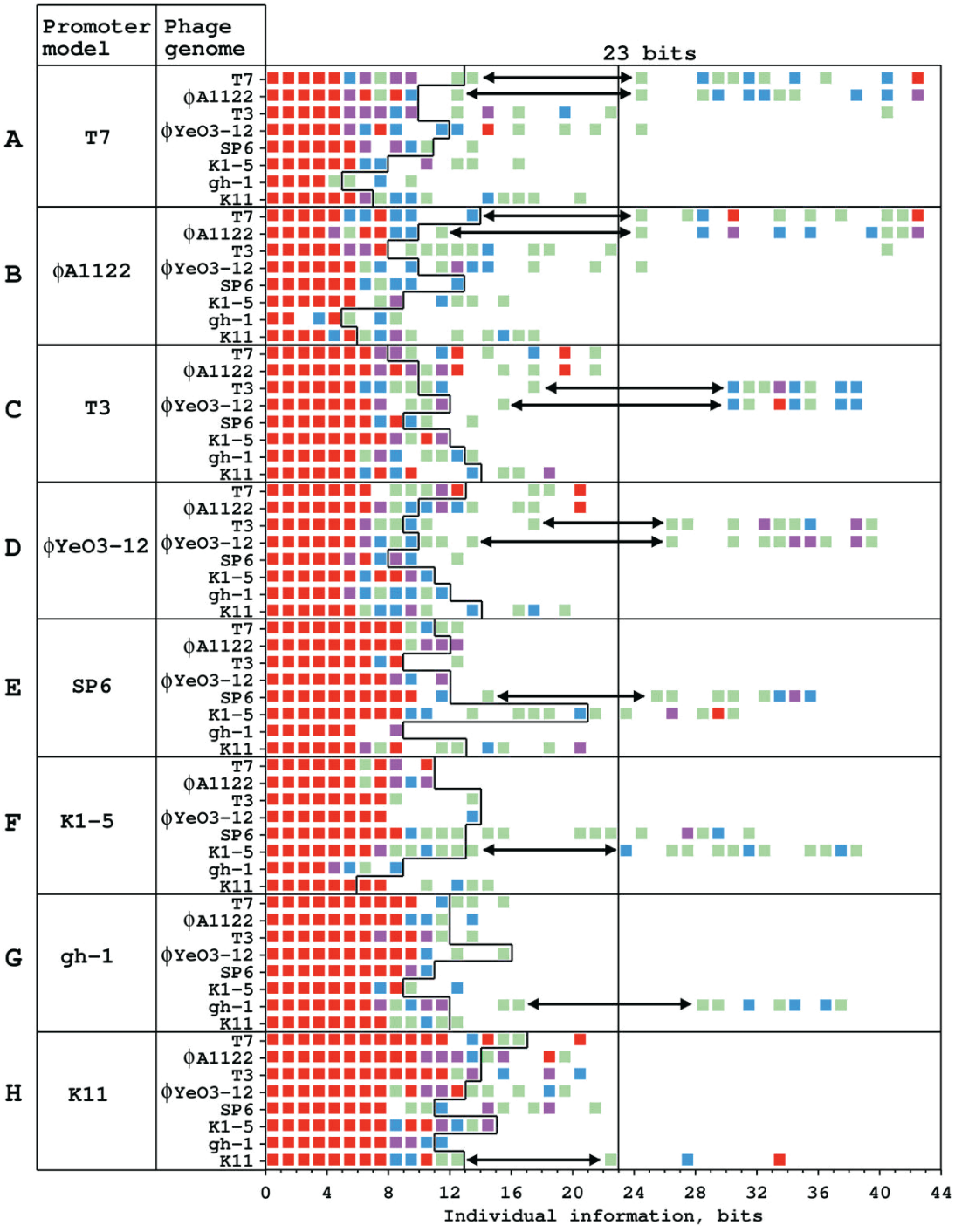

Figure 5. Individual information distribution of genome scanning with the eight T7-like promoter models. The eight models were used to scan both strands of the genomes of T7-like phages and the individual information distributions were plotted for each scan by using colored squares (green means 1 site; blue, 2 sites; purple, 3 sites; red 4 or more sites). When each model was used to scan its own genome or a closely related genome, a significant gap (double-headed arrow) was observed between promoter sites in the model and other sequences. The separating lines around 12 bits mark the highest individual information of the noncoding strand, which acts as a background control, since all experimentally demonstrated T7-like promoters are on the coding strand. For K11 scanned with K11, the site which is plotted at 22 bits is actually 22.99 bits and this is called 23 bits in the text.

arrows), suggesting that the models can be used to identify new T7-like promoters. The higher bound of the gap, which is the lower bound of promoters in the model, ranges from 23 to 30 bits. The lower bound of the gap, which is approximately the higher bound of the background, ranges from 13 to 17 bits (Figure 5). A gap of 8 bits corresponds to a binding constant change of at least $2^8 = 256$-fold (71).

T7 and ϕA1122. The T7 and ϕA1122 models were used to scan the seven genomes of the T7 group and a random sequence into which we inserted the nine K11 promoters. Because these two models are almost identical, similar *Ri* distributions were observed for both models. A significant gap, with a lower bound at 13 or 14 bits and an upper bound at 24 bits, was observed when scanning the T7 and ϕA1122 genomes (Figure 5A and B). When scanning the other six genomes, some T3, ϕYeO3-12 and K11 promoters were as high as 22, 24 and 20 bits, respectively, while the gh-1, SP6 and K1-5 promoters were always lower than 16 bits, suggesting that phages T3, ϕYeO3-12 and K11 are closer to T7 and ϕA1122 than phages gh-1, SP6 and K1-5 are (Figure 5A and B). The so-called T7 promoter (17A) in the T3 genome (66) was picked up at 40 bits by the T7 and ϕA1122 models (Figure 5A and B). This promoter is located in a region covering genes 15 to 18.5, which was probably introduced into a progenitor of T3 by a recombination event between the ancestors of two yersiniophages, ϕA1122 and ϕYeO3-12 (19). So this promoter was originally from a progenitor of ϕA1122 and we named it 17A in this study, since it is identical to promoter 17 in ϕA1122 (Figure 1).

T3 and ϕYeO3-12. Scanning with the T3 and ϕYeO3-12 models also gave similar *Ri* distributions (Figure 5C and D). When scanning the T3 and ϕYeO3-12 genomes, a gap was observed between 14 or 17 to 30 bits. Some T7, ϕA1122 and K11 promoters were picked up from 12 to 22 bits, but no promoters from SP6, K1-5 and gh-1 were found higher than 14 bits.

SP6 and K1-5. When scanning the SP6 genome with the SP6 model, there are no sites between 15 and 25 bits (Figure 5E). When scanning the phage K1-5 genome with the SP6 model, the 14 K1-5 promoters range from 18 to 31 bits and no significant gap can be observed, so no double-headed arrow is shown (Figure 5E). This indicates that, although K1-5 is similar to SP6, it is not so closely related to SP6 as ϕA1122 is to T7, or ϕYeO3-12 is to T3. When scanning the random sequence into which we inserted the 9 K11 promoters, four sites of 18–21 bits were picked up. When scanning the other five phage genomes, no sites above 13 bits were found (Figure 5E). When the K1-5 model was used to scan the K1-5 genome, a gap was observed between 14 and 23 bits (Figure 5F). When scanning the SP6 genome, 11 promoters were above 20 bits, the other one was only 15.3 bits. No sites above 16 bits were found for the other six genomes (Figure 5F).

gh-1. When the gh-1 model was used to scan its own genome, a gap from 17 to 28 bits was observed (Figure 5G). The phage gh-1 is distantly related to other members of the T7 group, so the gh-1 model cannot pick up any of the other seven phage promoters above 16 bits (Figure 5G).

K11. Though no genome sequence is available for phage K11, the K11 model was used to scan other phage genomes and, for consistency, a random sequence into which we inserted the nine K11 promoters. When the random sequence containing embedded K11 promoters was scanned with the K11 model, a gap was observed between 13 and 23 bits (Figure 5H). When the genomes of T7, ϕA1122, T3, ϕYeO3-12 and SP6 were scanned, their promoters gave an *Ri* up to ~22 bits, while the *Ri*s of the K1-5 and gh-1 promoters were much lower (Figure 5H).

Scanning other genomes with the T7-like models

Each of the eight promoter models and the combined 76-site model were also used to scan seven other distantly related T7-like phages, ϕKMV, P60, VpV262, SIO1, PaP3, Xp10 and P-SSP7, and a T7-like prophage, Ppu40 (PHAGE03), which is in the genome of *P.putida* KT2440 (20,55,72). With the 76-site model, 108 of the total 109 promoters in the eight phages of the T7 group were picked up above 12 bits, and smaller gaps of 3–7 bits (compared with individual model scanning, Figure 5) still exist (Figure 6A). In contrast, only three sites (in P-SSP7) above 12 bits were found for the eight distantly related phages (Figure 6B). When these eight phages were scanned with each of the eight models, most sites were below 15 bits (data not shown). Two exceptions, found by scanning with the gh-1 model, are a 17.0-bit site in Ppu40 (NC_002947, 2616401) and a 19.7-bit site in VpV262 (NC_003907, 43750 on the complimentary strand). Since all promoters for the T7 group of phages are higher than 23 bits with their own models (Figure 5), these results show that there are at best weakly conserved T7-like promoters in these eight phages and certainly no full sets of T7-like promoters; similar predictions have been made by other researchers (22,34,53).

ϕKMV. We looked into some other predictions for several of these eight distant T7-like phages. For the phage ϕKMV, three sites were predicted (22). We built a model using these three sites and scanned the ϕKMV genome with this model. The result shows that only the same three sites were found (Supplementary Figure S6). All three sites are located in intergenic regions, which is consistent with them being promoters. However the logo for the three sites is at best weakly T7-like and there are only three sites, while members of the T7 group have at least nine promoters.

Xp10. For the phage Xp10, seven sites were predicted recently and two of them were demonstrated experimentally (73). An initial model (17.8 bits) built from these seven sites picks up the same seven sites above 15 bits, the other sequences are below 13 bits (two are about 12.5 bits and the others are below 11 bits), thus giving a gap of 3 bits. When we included the two 12-bit sites into the model, the information rose to 18.3 bits. When this 9-site model was used to scan the Xp10 genome, the same nine sites were found above 16 bits, while the other sequences are below 12 bits, thus giving a gap of 5 bits (Supplementary Figure S7). This result indicates that the 9-site model is slightly better than the initial 7-site model. The two sites predicted here are located at 15 444 and 42 789 (on the complementary strand), respectively. The site at 42 789 indicates that the phage RNAP could also be responsible for part of the left-oriented transcription.

P-SSP7. For the phage P-SSP7 no promoters have been predicted (74). Genome scanning with the 76-site model found

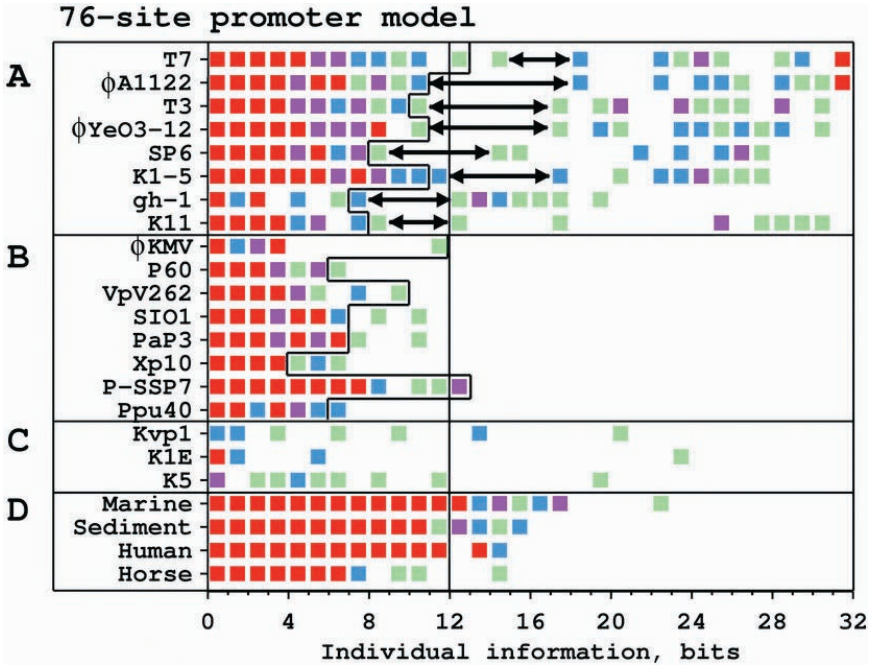

Figure 6. Individual information distribution of genome scanning with the combined 76-site model. (**A**) The eight phages of the T7 group were scanned, small gaps (double-headed arrow) are observed between the promoters and the background. (**B**) Another eight complete T7-like phage genomes were also scanned, and most sites were under 12 bits. (**C**) Scans of partial genomes of three other phages of the T7 group. (**D**) Scans of phage genome survey sequences from four uncultured viral communities, marine (Marine), marine sediment (Sediment), human feces (Human) and horse feces (Horse). The vertical line at 12 bits is the proposed approximate lower bound for functional T7-like promoters. Fragment orientation is unknown for parts (C) and (D), so no background line is shown.

five sites between 10 and 13 bits (Figure 6B). However, a sequence logo built from these five sites shows little conservation (only 11 bits, data not shown), and a model built from these five sites is not distinctly separated with a gap from other sequences. Adding weaker sites into the model did not improve the results. Then we used the program Blast2 to align every 1 kb fragment of the P-SSP7 genome against the rest of the genome. Only two matching sequences of 20 bases were revealed; one is located within the P-SSP7 RNAP coding region (5933) and the other (42 900) right after gene gp19 (coding for DNA maturase). Neither of these is a member of the five sites mentioned above. Since all members of the T7 group contain a promoter after gp19, this result suggests those two sites are likely to be P-SSP7 promoters.

The above analysis shows clearly that the three phages ϕKMV, Xp10 and P-SSP7 do not have a set of conserved T7-like promoters; at best they have several weakly conserved T7-like promoters. The results strongly suggest that these distantly related T7-like phages may have a different strategy of transcriptional control from that of the T7 group phages.

By scanning the available pieces of genome of the phages Kvp1, K1E and K5 with the 76-site model, several promoters of 19–24 bits were found (Figure 6C). One site of 20.3 bits was picked up by the 76-site model from one piece of phage Kvp1 sequence (X96817 at 112) (28). One site of 23.4 bits was picked up by the 76-site model from one piece of phage K1E sequence (X78310 at 692) (75), this site was also picked up by the SP6 model (29 bits) and the K1-5 model (37.2 bits, data not shown). One site of 19.4 bits was picked up from one piece of phage K5 sequence (Y10025 at 352) (76), this site was also picked up by the SP6 model (25.3 bits) and the K1-5 model (33.3 bits, data not shown). The results support the inclusion of these three phages in the T7 group and further indicate that the phages K1E and K5 are related to SP6 and K1-5 (48).

To detect the possible distribution of the T7 group of phages in the environment, genome survey sequences from four uncultured viral communities were scanned with each of the T7-like models and the combined 76-site model. With the 76-site model, a total of 22 sites were picked up above 12 bits from 1200 kb of marine phage sequences (named Marine in Figure 6D) (67). Within these, one site of 22.4 bits was also picked up by the T3 model (23.9 bits) and the ϕYeO3-12 model (24.6 bits) in BH898648 at 276, another site of 17.3 bits was also picked up by the SP6 model (25.8 bits) in AY080583 at 48. These two sites are significant, while the others may not be. For the other three viral communities (68–70), no sites above 16 bits were found with the 76-site model (Sediment, Human and Horse in Figure 6D), and no sites above 18 bits were found with any of the eight individual models (data not shown). These results suggest that the T7 group of phages may not be so widespread as hypothesized previously (67), and so the results support more recent evidence of their rarity (77).

Phylogenetic analysis of the T7-like promoters and the phage RNAPs

Zhang's three dimensional method (45) was used to calculate the distances among the eight promoter models. The resulting base frequency distance matrix was used to infer a phylogenetic tree, which shows five well-defined branches (Figure 7A), indicating that the eight phages can be classified into five subgroups: T7-like, T3-like, K11-like, gh-1-like and

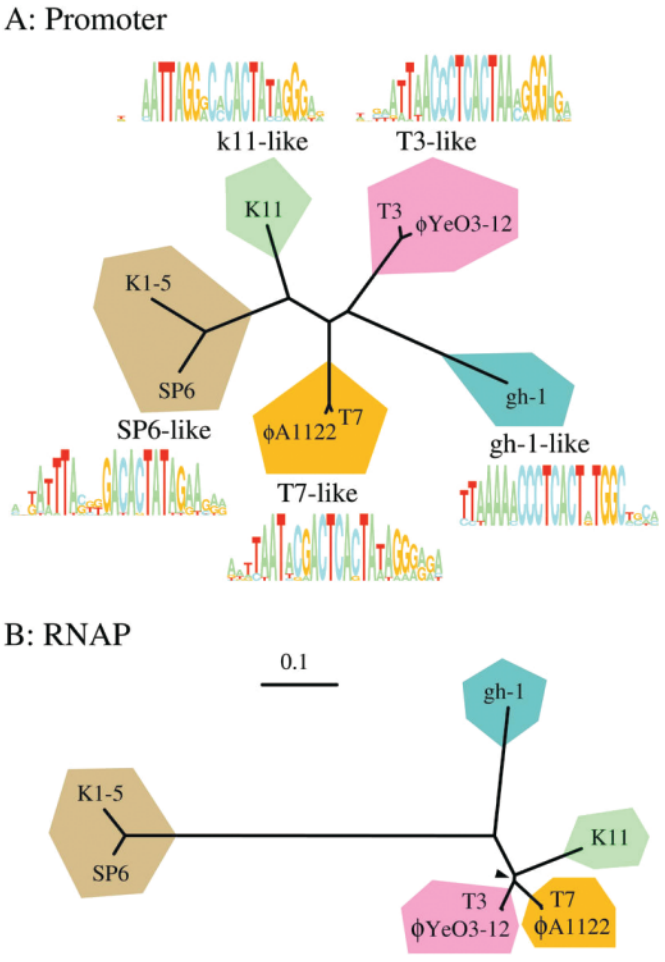

Figure 7. Phylogenetic analysis of the T7-like promoter models and the RNAPs from the eight T7 group phages. (**A**) Distances among different promoter models were calculated by the program diffribl using Zhang's method. The resulting distance matrix was used to generate the tree by the programs neighbor and drawtree. Individual and combined sequence logos are shown for comparison. (**B**) An unrooted tree of the eight RNAP protein sequences was constructed using the programs protdist and neighbor from the PHYLIP 3.6a3 package (42). All nodes have 100% bootstrap support except for one which is 80% (indicated by a black triangle). The two trees in (A) and (B) were normalized, so the scale bar applies to both trees and represents 10% of respective total branch length.

SP6-like. A similar result was obtained with the Euclidean distance method (data not shown). To confirm these results, a phylogeny was also constructed based on the alignment of all 109 promoters by using a standard distance method with the program dnadist (42). We compared the promoters and found that transitions are less frequent than transversions. This is in contrast with protein coding regions which are often assumed to have a transition to transversion ratio of 2.0 (42). When the frequency of substitutions between all four bases is equal, the transition/transversion ratio is 0.5. Dnadist gave infinite distances between some promoters when values of transition/transversion >0.6 were used, suggesting that promoter regions do not favor transitions over transversions. When a ratio of 0.5 was used to estimate the distances among the promoters with the program dnadist, a tree similar to Figure 7A was obtained (Supplementary Figure S1). Furthermore, this tree is consistent with the detailed genome scans described in the previous section. Therefore, the result shown in Figure 7A is method-independent and this validates the Zhang comparison of matrices.

Sequences homologous to phage-type RNAPs are found in many eukaryotes (78). Phylogenetic analysis has shown that three distinct classes of single-subunit RNAP (phage-encoded, plasmid-encoded and nucleus-encoded) are well-defined (79). In this study, we focused on the phage-encoded RNAPs, most of which have been found in T7-like phages. An unrooted tree was generated for the eight RNAPs from the T7 group of phages. As with the promoter models, the protein sequences also show five branches, but with three closely related (T7-like, T3-like and K11-like), one moderately related (gh-1-like) and one distantly related (SP6-like) (Figure 7B).

The DNA site and protein phylogenetic trees were based on different distance computation algorithms, so they are not directly comparable. To address this, both trees were normalized by their respective total branch length, resulting in two trees with the same scale, showing the fraction of total divergence (Figure 7).

A tree was also constructed for all available phage-type RNAPs which are from the eight T7 group of phages, three distantly related T7-like phages and three bacterial genomes (Supplementary Table S1 and Supplementary Figures S2 and S3). The result shows that the three RNAPs (Ppu40, C58 and CGA) found in bacterial genomes are closer to the T7 and T3 RNAPs (average distance, $\bar{x} = 1.32$) than the SP6 and K1-5 RNAPs are ($\bar{x} = 1.50$), while the three RNAPs (Xp10, P60 and ϕKMV) from the three distant T7-like phages are even more distantly related to the T7 and T3 RNAPs ($\bar{x} = 1.75$).

Phylogenetic analysis was also performed for phage RNAPs by using the maximum likelihood method implemented in the program ProML (42). The results are highly similar to Figure 7B and Supplementary Figure S3 (data not shown).

DISCUSSION

The T7-like promoter models

A total of eight T7-like promoter models were built in this study, within which the ϕA1122 model looks almost the same as the T7 model and the ϕYeO3-12 model looks almost the same as the T3 model (Figure 2). For each of these two almost identical pairs, some promoters are identical from −20 to +5 (Figure 1), and the RNAPs are 98.5 and 99.1% identical (Supplementary Table S1), so apparently one phage RNAP should recognize all or most promoters from the other phage. Though the K1-5 model is highly similar to the SP6 model (Figure 2), the SP6 RNAP, which is 85.4% identical to the K1-5 RNAP (Supplementary Table S1), might not recognize K1-5 promoters, because these two models have different bases at position −11 (8). It is not clear whether the K1-5 RNAP recognizes SP6 promoters. The other two models, gh-1 and K11, are much more diversified, except in the region from position −7 to −4, which is the most conserved for all eight models (Figures 2 and 4).

Each model is able to clearly identify the promoters from its own phage genome (Figure 5). Furthermore a leave-one-out test was carried out for all eight models (data not shown). For the T7 and T3 models, which were built from known promoters, the lowest *Ri* for a site which was left out from the model is 19.7 bits (the background, from the untranscribed

complementary strand, is lower than 13 bits). For the other six models, all sites that were left out from respective models are higher than 20 bits, thus the models were further confirmed as being self-consistent.

Using a string-based search algorithm, Lavigne *et al*. (34) predicted promoters for several members of the T7 group. Their method was based on mismatch counting to a consensus and this has been shown to be error prone (35). They missed one T7 promoter (4.7) in the T7 genome because this promoter deviates at six positions from the consensus promoter, and picked up the T7 promoter in the T3 genome when they intended to predict T3 promoters. They also predicted 12 gh-1 promoters, which include the 10 promoters predicted in this study and two more sites at position 2918 (2885–2934) and 11 237 (11 204–11 253). By our 10-site model, these two sites contain only 15.4 and 16.5 bits of information, slightly higher than the background (Figure 5G). Even when we included both of Lavigne's sites into our 10-site model, the information only rose to 18.3 and 18.9 bits, respectively, still too low to be considered as promoters, because they are far lower than the average of 33.5 bits ($P = 4.9 \times 10^{-5}$). One possible explanation is that these two sites are decayed or weak promoters. Therefore the information theory based promoter models appear to be more robust than previous models.

Genome scanning with the T7-like promoter models

Because T7 promoters (4,33) and T7-like promoters (this work) have excess information, when each of the models was used to scan its own genome or a closely related genome, a significant gap ($P < 6 \times 10^{-4}$) was observed in the *Ri* distribution between the promoters in the models and the next weaker site or the background of the opposite strand (Figure 5, double-headed arrows). The gap has a lower bound of 15 bits and an upper bound from 23 to 30 bits. The existence of a gap in the *Ri* distribution allows us to distinguish T7-like promoters from the background when scanning genomes with these models. The combined 76-site model can locate 108 out of the 109 promoters (Figure 6A). So each of the eight individual models can be used to identify closely related promoters, while the combined 76-site model can be used to identify weak or distantly related T7-like promoters. The 76-site model is clearly not a biological model for any particular promoter. However, it is useful for identifying T7-like promoters because it detects the common features among them.

With the nine models, we scanned eight other T7-like genomes, ϕKMV, P60, VpV262, SIO1, PaP3, Xp10, P-SSP7 and Ppu40 (Figure 6B). Unlike scans of the T7 group of phages (Figure 6A) the nine models do not recognize sites in these genomes (Figure 6B and data not shown), so the models distinguish between the T7 group of phages and these eight T7-like phages. The results suggest that these phages use a different strategy of infection by either using the host RNAP or by encoding their own polymerase that recognizes promoters highly distinct from the known T7-like promoters.

Evolution and classification of T7-like phages

Extensive horizontal exchange of genes or groups of genes has resulted in mosaic genomes in temperate phages (11,80). Strictly virulent phages must have few opportunities for recombination with similar phages because the chance for a cell to be coinfected by two different virulent phages is smaller than superinfection of lysogens in natural conditions. However, horizontal exchange has been observed both within the T7 group of phages (16,19,20) and between the members of the T7 group and other phages (17,20,81). During these exchanges, the RNAP gene and the majority of its promoters must be kept together to ensure a functional T7-like transcription system and to retain the strategy of infection for this group of phages. In contrast, none of the other distantly related T7-like phages contain both components of the T7-like transcription system, indicating that the strategy of infection is different from that of the T7 group of phages.

The discovery of an apparent T7-like prophage (Ppu40) in the genome of *P.putida* KT2440 was surprising, since it has homology to most of the essential T7 genes and the gene orders are almost the same (20,55,72). This phage may have evolved from an ancestral phage of the T7 group. During the evolution, one major change for this phage was the insertion of an integration module at the right end of its genome, presumably providing the phage the ability to integrate as a prophage. However the most striking difference between this prophage and the T7 group of phages is that the prophage does not contain a set of well-conserved T7-like promoters (Figure 6B). To investigate this phage we lowered the threshold of genome scanning by the gh-1 promoter model because we had already found a 17-bit gh-1 site in Ppu40. Eleven putative promoters were identified in the prophage genome and used to build a model which gave 19.1 bits (Supplementary Figure S5), significantly lower than that of other T7-like promoter models (Figure 2). This suggests that the promoters have been partially decayed since the phage diverged from the T7 group, and that the strategy of infection has been significantly changed.

The strategy of infection was originally used as a criterion for inferring phylogenetic relationships with the phage T7 (3). To make the concept that a phage has a similar strategy of infection as T7 more precise, in this paper we define the T7 group of phages as those phages that contain both a phage-specific RNAP and a set of conserved T7-like promoters (e.g. as found by the 76-site model). Different phage subgroups can be classified based on the promoter pattern recognized by the phage-encoded RNAP, which is reflected in the sequence logos (Figure 2). Generally, members of the same subgroup can recognize each other's promoters. A phage that shares genomic similarity in any genetic module with the T7 group of phages, but does not contain both components of the T7-like transcription system, can be assigned to the T7 supergroup. These definitions provide a framework for classification of T7 supergroup phages.

Though the phage-encoded RNAP has been recognized as a hallmark feature of the T7 group (13,14), it is not appropriate to classify T7-like phages only based on the phage RNAPs. Several distant T7-like phages (ϕKMV, Xp10 and P60) encode a T7-like RNAP, but they do not contain a set of conserved phage promoters that we could detect, so they do not belong to the T7 group. The RNAP encoded by the prophage Ppu40 is closer to the T7 and T3 RNAPs than SP6 and K1-5 RNAPs are to T7 and T3 (Supplementary Figure S3), but as discussed above, the corresponding promoters have been partially decayed, so this prophage should also not be classified as a member of the T7 group.

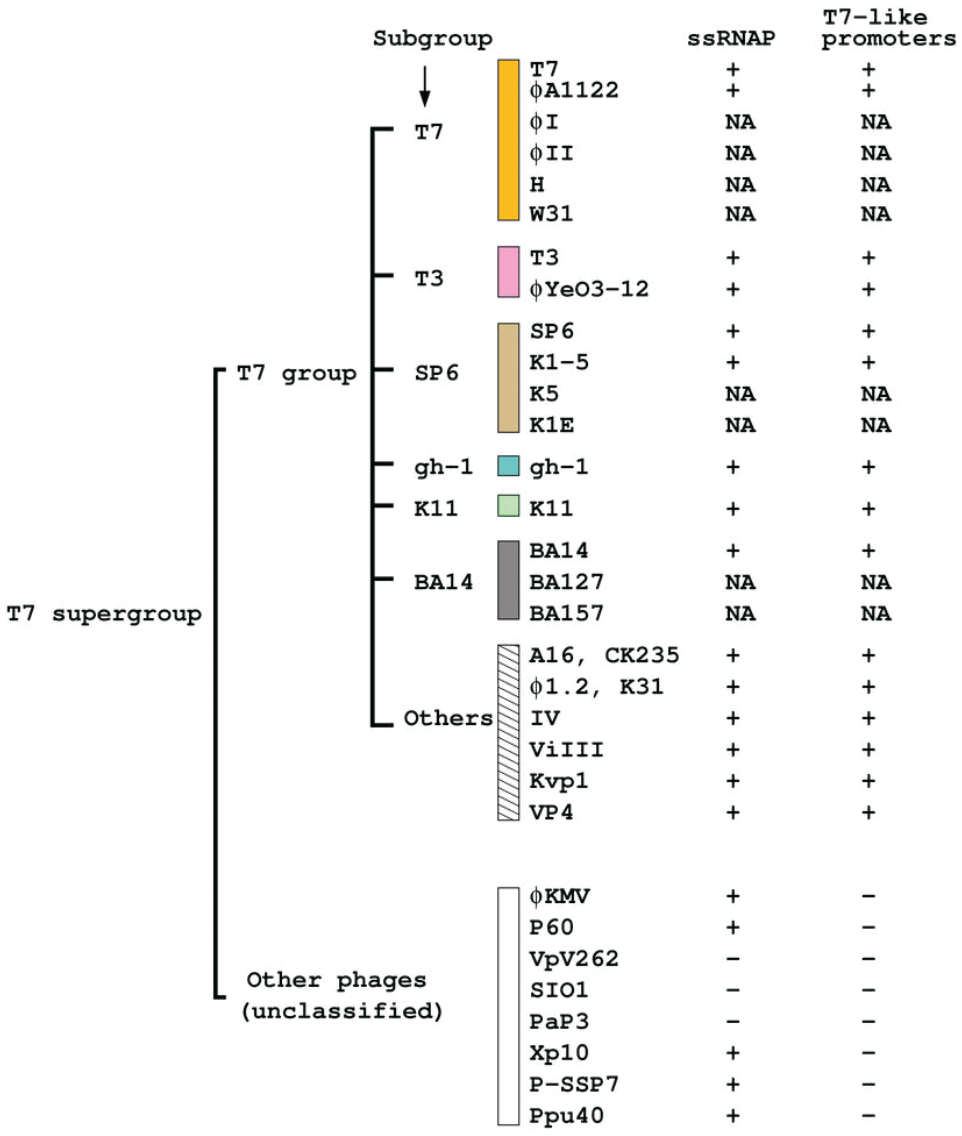

Figure 8. Classification of T7-like phages based in part on phage promoters and single subunit RNAPs (ssRNAPs). NA means data not available, plus means the ssRNAP gene or promoters can be found in the genome, minus means that no ssRNAP gene or promoters can be found.

The T7 group of phages were classified previously into three subgroups (T7-like, T3-like and BA14-like) based on the efficiency of recombination with each other (2). In addition to these three subgroups, we propose at least four more subgroups based on the promoter pattern recognized by their RNAPs (Figure 8: SP6, gh-1, K11 and others).

We characterize these subgroups as follows (Figure 8). T7: the phages ϕI, ϕII, H and W31 have been shown to be more closely related to phage T7 than is T3 (31), and the *Yersinia* phage H and *Escherichia coli* phage ϕII are nearly identical (82), so these phages should belong to the T7 subgroup. T3: the phages T3 and ϕYeO3-12 make the T3 subgroup. SP6: the phages SP6, K1-5, K5 and K1E have been shown to be closely related (17,48), and one K5 promoter and one K1E promoter were picked up above 25 bits by the SP6 and K1-5 models (Figure 6C), suggesting these four phages belong to the SP6 subgroup. gh-1 and K11: so far no closely related members have been found for phages gh-1 and K11, so each of these two phages makes a separate subgroup. BA14: the BA14-like phages include three isolates (29,30). Others: the RNAPs of the phages A16 and CK235 show substantial heterologous transcription on each other's DNA, but much lower activity on several other members of this group (32), so these two phages can be assigned to a separate subgroup. Similarly, the phages ϕ1.2 and K31 have even stronger heterologous transcription, and <10% activity on other members of this group, so they should also belong to a separate subgroup. The *Serratia* phage IV and *Citrobacter* phage ViIII share no detectable similarity with other members of the T7 group (3), so each of these two may belong to different subgroups. The phage Kvp1 has also been shown to be a close relative of the T7 group (28) and, although the one predicted Kvp1 promoter was not found by any of the eight T7-like promoter models, it was identified by the lower information content 76-site model (Figure 6C). So Kvp1 is also in a separate subgroup.

As we were finishing this paper the genome for Vibriophage VP4 was published (NC_007149). When the VP4 genome was

scanned with the 76-site model, 19 sites were picked up above 14 bits while other sequences are below 10 bits. These 19 sites were used to build an initial model. Two of these sites did not fit well, so the final model has 17 sites (Supplementary Figure S8). The VP4 promoters differ from T3 in only a few positions and so can be classified as a new T7 subgroup (Figure 8).

So far there have been only eight (including VP4) genomic sequences available for the T7 group, seven of which were sequenced within the last few years. Sequencing of the 17 other members (listed in Figure 8) should provide further insights into the evolution of this group of phages.

Polymerases and their promoters evolve differently

Phylogenetic analysis confirmed that the promoters coevolve with their RNAPs since the three pairs of most closely related RNAPs recognize promoters that cluster together (Figure 7A and B: T7 and ϕA1122, T3 and ϕYeO3-12, SP6 and K1-5). However, while the promoters have diverged approximately equally (Figure 7A), the polymerases have changed by different amounts relative to each other, since the SP6-like polymerases have diverged away from T7, T3, K11 and gh-1 more than the latter have diverged from each other. The gh-1 polymerases have diverged an intermediate amount (Figure 7B). Apparently the promoters and their corresponding polymerases did not evolve at equal rates.

To understand the difference between the promoter and polymerase phylogenetic trees, we first propose the standard explanation that the polymerase genes diverge by neutral drift. That is, the polymerases all retain common functions of DNA binding and transcription and are mostly altered in irrelevant components, along with some change in DNA pattern recognition. At least 41 amino acids in the polymerase are thought to be in direct contact with the promoters (21), which represents only 5% of the 883 amino acid T7 polymerase protein. This implies that the lengths of the branches should be, for the most part, proportional to evolutionary time as a 'molecular clock'. We tested this hypothesis by creating phylogenetic trees for 17 different genes in seven phages of the T7 group (Supplementary Figure S4). The resulting trees show greater divergence of SP6 and K1-5 proteins than the other phage proteins for 16 of the 17 ortholog trees built. This demonstrates that the polymerase divergence corresponds to phage divergence.

We were concerned that our novel method for comparing the promoters could give anomalous results. We measured divergence of the promoters by using the distance between them in a probability space, but we obtained similar results using conventional methods (Compare Figure 7A with Supplementary Figure S1), so their divergence by similar amounts does not appear to be a function of the method used.

The polymerases and their binding sites must coevolve (83), yet their phylogenetic trees are strikingly different (Figure 7). We propose two hypotheses to explain the different evolution of the promoters and their polymerases: coding saturation and size discrepancy.

In the first hypothesis, coding saturation, when two phages begin to diverge their polymerases may still recognize the promoters of the other phage, so there could be strong selective pressure to become different. For example, inappropriate firing of promoters by a heterologous polymerase might interfere with phage transcription and DNA replication. Once a significant difference in recognition (and the corresponding binding sites) has been achieved, the promoters have no further selective pressure to change. Furthermore, since they coevolve with the polymerase, the promoter-polymerase system may resist change. The requirement for functional coevolution of the binding sites and their similarity in a coding space restricts their change, analogous to a particles trapped in a viscous solution. In this case the divergence proceeds just far enough to be functionally different and no further. In contrast, while remaining functional, the neutral divergence of the proteins is continuous over time, analogous to the Newtonian diffusion of gas spreading out after release from a container. There is no immediate limit to the divergence.

Since all of the protein sequences represent RNAPs, the protein sequence divergence measures mostly neutral and a few functional changes that specify the promoter sequences. This measure reflects the divergence of the strains over time. On the other hand, the DNA sequences of the promoters reflect more functional differences and less neutral drift. For example, if, after divergence, bacteriophage T7 competed with bacteriophage T3 for burst size in bacterial colonies there could be selection for distinct promoter patterns during co-infections.

In this hypothesis the promoter patterns only diverge until they became distinct. Shannon's channel capacity theorem (84), applied to DNA recognition (85,86) states that the promoters can become as distinct as necessary for survival and that a sharp distinction could exist [and indeed does exist (87)] between T7 and T3 promoters. That is, T7 and T3 promoters evolved to form a distinct code in sequence space. Once a code distinguishing the two promoters has been found, there is no further evolutionary pressure or advantage to change, since each phage activates only its own promoters. At this point only slow neutral drift takes place. The drift will be slow because of the constraint that the 17 or so promoters must always correspond to the cognate polymerases. So promoters from different phages should diverge in coding space until they are just distinct and then they stop diverging significantly. This explains the observed uniformly diverged phylogenetic tree of the promoters. The hypothesis depends strongly on the ecology of the phage. If they do not meet and compete in nature, the hypothesis will fail. However, it is well known that T7 phages exchange modular units (16,19,20) and so they must meet at least frequently enough for recombination and hence they have opportunity for promoter competition.

The alternative hypothesis, size discrepancy, is that proteins and DNA-binding sites are vastly different in size and this will affect the rate of divergence. The promoters used in this study (Figure 1) have an average of 33.4 ± 1.7 bits (Figure 2). This is significantly higher than other binding sites (4,33) but it is much smaller than proteins since the alignment for the corresponding eight polymerase proteins gives 1097 bits. Neutral drift of the polymerases could continue for a long time while the promoters, being smaller, may reach limits of change more rapidly. For this hypothesis we propose that the promoters have, for the most part, reached their limits of divergence while the polymerases have not. The main objection to this hypothesis is that the sequence logos still all resemble each other greatly (Figures 2, 4 and 7), so there is clearly more

divergence possible. Therefore we suggest that the coding saturation hypothesis is more likely to explain the difference in evolution of promoters and their polymerases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ilya Lyakhov and Danielle Needle for useful discussions. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. Funding to pay the Open Access publication charges for this article was provided by NIH/NCI.

Conflict of interest statement. None declared.

REFERENCES

1. Browning,D.F. and Busby,S.J. (2004) The regulation of bacterial transcription initiation. *Nature Rev. Microbiol*., **2**, 57–65.
2. Hausmann,R. (1988) The T7 Group. In Calendar,R. (ed.), *The Bacteriophages*. Vol. 1. Plenum Press, NY, pp. 259–289.
3. Korsten,K.H., Tomkiewicz,C. and Hausmann,R. (1979) The strategy of infection as a criterion for phylogenetic relationships of non-coli phages morphologically similar to phage T7. *J. Gen. Virol*., **43**, 57–73.
4. Schneider,T.D. and Stormo,G.D. (1989) Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucleic Acids Res*., **17**, 659–674.
5. Diaz,G.A., Raskin,C.A. and McAllister,W.T. (1993) Hierarchy of base-pair preference in the binding domain of the bacteriophage T7 promoter. *J. Mol. Biol*., **229**, 805–811.
6. Imburgio,D., Rong,M., Ma,K. and McAllister,W.T. (2000) Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. *Biochemistry*, **39**, 10419–10430.
7. Jorgensen,E.D., Durbin,R.K., Risman,S.S. and McAllister,W.T. (1991) Specific contacts between the bacteriophage T3, T7, and SP6 RNA polymerases and their promoters. *J. Biol. Chem*., **266**, 645–651.
8. Shin,I., Kim,J., Cantor,C.R. and Kang,C. (2000) Effects of saturation mutagenesis of the phage SP6 promoter on transcription activity, presented by activity logos. *Proc. Natl Acad. Sci. USA*, **97**, 3890–3895.
9. Ackermann,H.W. (2001) Frequency of morphological phage descriptions in the year 2000. Brief review. *Arch. Virol*., **146**, 843–857.
10. Rohwer,F. and Edwards,R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol*., **184**, 4529–4535.
11. Botstein,D. (1980) A theory of modular evolution for bacteriophages. *Ann. N Y Acad. Sci*., **354**, 484–490.
12. Lawrence,J.G., Hatfull,G.F. and Hendrix,R.W. (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol*., **184**, 4891–4905.
13. Maniloff,J. and Ackermann,H.W. (1998) Taxonomy of bacterial viruses: establishment of tailed virus genera and the order *Caudovirales*. *Arch. Virol*., **143**, 2051–2063.
14. Ackermann,H.W. (2003) Bacteriophage observations and evolution. *Res. Microbiol*., **154**, 245–251.
15. Dunn,J.J. and Studier,F.W. (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol*., **166**, 477–535.
16. Pajunen,M.I., Elizondo,M.R., Skurnik,M., Kieleczawa,J. and Molineux,I.J. (2002) Complete nucleotide sequence and likely recombinatorial origin of bacteriophage T3. *J. Mol. Biol*., **319**, 1115–1132.
17. Scholl,D., Kieleczawa,J., Kemp,P., Rush,J., Richardson,C.C., Merril,C., Adhya,S. and Molineux,I.J. (2004) Genomic analysis of bacteriophages SP6 and K1-5, an estranged subgroup of the T7 supergroup. *J. Mol. Biol*., **335**, 1151–1171.
18. Pajunen,M.I., Kiljunen,S.J., Soderholm,M.E. and Skurnik,M. (2001) Complete genomic sequence of the lytic bacteriophage ϕYeO3-12 of *Yersinia enterocolitica* serotype O:3. *J. Bacteriol*., **183**, 1928–1937.
19. Garcia,E., Elliott,J.M., Ramanculov,E., Chain,P.S., Chu,M.C. and Molineux,I.J. (2003) The genome sequence of *Yersinia pestis* bacteriophage ϕA1122 reveals an intimate history with the Coliphage T3 and T7 genomes. *J. Bacteriol*., **185**, 5248–5262.
20. Dobbins,A.T., George,M.,Jr, Basham,D.A., Ford,M.E., Houtz,J.M., Pedulla,M.L., Lawrence,J.G., Hatfull,G.F. and Hendrix,R.W. (2004) Complete genomic sequence of the virulent *Salmonella* bacteriophage SP6. *J. Bacteriol*., **186**, 1933–1944.
21. Kovalyova,I.V. and Kropinski,A.M. (2003) The complete genomic sequence of lytic bacteriophage gh-1 infecting *Pseudomonas putida*—evidence for close relationship to the T7 group. *Virology*, **311**, 305–315.
22. Lavigne,R., Burkal'tseva,M.V., Robben,J., Sykilinda,N.N., Kurochkina,L.P., Grymonprez,B., Jonckx,B., Krylov,V.N., Mesyanzhinov,V.V. and Volckaert,G. (2003) The genome of bacteriophage phiKMV, a T7-like virus infecting *Pseudomonas aeruginosa*. *Virology*, **312**, 49–59.
23. Chen,F. and Lu,J. (2002) Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl. Environ. Microbiol*., **68**, 2589–2594.
24. Hardies,S.C., Comeau,A.M., Serwer,P. and Suttle,C.A. (2003) The complete sequence of marine bacteriophage VpV262 infecting *Vibrio parahaemolyticus* indicates that an ancestral component of a T7 viral supergroup is widespread in the marine environment. *Virology*, **310**, 359–371.
25. Rohwer,F., Segall,A., Steward,G., Seguritan,V., Breitbart,M., Wolven,F. and Azam,F. (2000) The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr*., **45**, 408–418.
26. Han,K.G., Kim,D.H., Junn,E., Lee,S.S. and Kang,C. (2002) Identification of bacteriophage K11 genomic promoters for K11 RNA polymerase. *J. Biochem. Mol. Biol*., **35**, 637–641.
27. Dietz,A., Weisser,H.J., Kossel,H. and Hausmann,R. (1990) The gene for *Klebsiella* bacteriophage K11 RNA polymerase: sequence and comparison with the homologous genes of phages T7, T3, and SP6. *Mol. Gen. Genet*., **221**, 283–286.
28. Gadaleta,P. and Zorzopulos,J. (1997) *Kluyvera* bacteriophage Kvp1: a new member of the *Podoviridae* family phylogenetically related to the coliphage T7. *Virus Res*., **51**, 43–52.
29. Michalewicz,J., Hsu,E., Larson,J.J. and Nicholson,A.W. (1991) Physical map and genetic early region of the T7-related coliphage, BA14. *Gene*, **98**, 89–93.
30. Mertens,H. and Hausmann,R. (1982) Coliphage BA14: a new relative of phage T7. *J. Gen. Virol*., **62**, 331–341.
31. Hyman,R.W., Brunovskis,I. and Summers,W.C. (1974) A biochemical comparison of the related bacteriophages T7, ϕI, ϕII, W31, H, and T3. *Virology*, **57**, 189–206.
32. Dietz,A., Andrejauskas,E., Messerschmid,M. and Hausmann,R. (1986) Two groups of capsule-specific coliphages coding for RNA polymerases with new promoter specificities. *J. Gen. Virol*., **67**, 831–838.
33. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol*., **188**, 415–431.
34. Lavigne,R., Sun,W.D. and Volckaert,G. (2004) PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. *Bioinformatics*, **20**, 629–635.
35. Schneider,T.D. (2002) Consensus Sequence Zen. *Applied Bioinformatics*, **1**, 111–119.
36. Schneider,T.D. (2001) Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucleic Acids Res*., **29**, 4881–4891.
37. Shultzaberger,R.K., Bucheimer,R.E., Rudd,K.E. and Schneider,T.D. (2001) Anatomy of *Escherichia coli* ribosome binding sites. *J. Mol. Biol*., **313**, 215–228.
38. Rogan,P.K., Faux,B.M. and Schneider,T.D. (1998) Information analysis of human splice site mutations. *Hum. Mutat*., **12**, 153–171.
39. Schneider,T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol*., **189**, 427–441.
40. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett*., **174**, 247–250.
41. Womble,D.D. (2000) GCG: The Wisconsin Package of sequence analysis programs. *Methods Mol. Biol*., **132**, 3–22.

42. Felsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
43. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
44. Schneider,T.D. (1996) Reading of DNA sequence logos: Prediction of major groove binding by information theory. *Methods Enzymol.*, **274**, 445–455.
45. Zhang,C.T. and Zhang,R. (1991) Diagrammatic representation of the distribution of DNA bases and its applications. *Int. J. Biol. Macromol.*, **13**, 45–49.
46. Moffatt,B.A., Dunn,J.J. and Studier,F.W. (1984) Nucleotide sequence of the gene for bacteriophage T7 RNA polymerase. *J. Mol. Biol.*, **173**, 265–269.
47. McGraw,N.J., Bailey,J.N., Cleaves,G.R., Dembinski,D.R., Gocke,C.R., Joliffe,L.K., MacWright,R.S. and McAllister,W.T. (1985) Sequence and analysis of the gene for bacteriophage T3 RNA polymerase. *Nucleic Acids Res.*, **13**, 6753–6766.
48. Scholl,D., Adhya,S. and Merril,C.R. (2002) Bacteriophage SP6 is closely related to phages K1-5, K5, and K1E but encodes a tail protein very similar to that of the distantly related P22. *J. Bacteriol.*, **184**, 2833–2836.
49. Kotani,H., Ishizaki,Y., Hiraoka,N. and Obayashi,A. (1987) Nucleotide sequence and expression of the cloned gene of bacteriophage SP6 RNA polymerase. *Nucleic Acids Res.*, **15**, 2653–2664.
50. Kassavetis,G.A., Butler,E.T., Roulland,D. and Chamberlin,M.J. (1982) Bacteriophage SP6-specific RNA polymerase. II. Mapping of SP6 DNA and selective in vitro transcription. *J. Biol. Chem.*, **257**, 5779–5788.
51. Jolly,J.F. (1979) Program of bacteriophage gh-1 DNA transcription in infected *Pseudomonas putida*. *J. Virol.*, **30**, 771–776.
52. Rong,M., Castagna,R. and McAllister,W.T. (1999) Cloning and purification of bacteriophage K11 RNA polymerase. *Biotechniques*, **27**, 690–692, 694.
53. Yuzenkova,J., Nechaev,S., Berlin,J., Rogulja,D., Kuznedelov,K., Inman,R., Mushegian,A. and Severinov,K. (2003) Genome of *Xanthomonas oryzae* bacteriophage Xp10: an odd T-odd phage. *J. Mol. Biol.*, **330**, 735–748.
54. Liao,Y.D., Tu,J., Feng,T.Y. and Kuo,T.T. (1986) Characterization of phage-Xp10-coded RNA polymerase. *Eur. J. Biochem.*, **157**, 571–577.
55. Nelson,K.E., Weinel,C., Paulsen,I.T., Dodson,R.J., Hilbert,H., Martins dos Santos,V.A., Fouts,D.E., Gill,S.R., Pop,M., Holmes,M. *et al.* (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ. Microbiol.*, **4**, 799–808.
56. Wood,D.W., Setubal,J.C., Kaul,R., Monks,D.E., Kitajima,J.P., Okura,V.K., Zhou,Y., Chen,L., Wood,G.E., Almeida,N.F.,Jr *et al.* (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science*, **294**, 2317–2323.
57. Larimer,F.W., Chain,P., Hauser,L., Lamerdin,J., Malfatti,S., Do,L., Land,M.L., Pelletier,D.A., Beatty,J.T., Lang,A.S. *et al.* (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. *Nat. Biotechnol.*, **22**, 55–61.
58. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.
59. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
60. Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
61. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
62. Pajunen,M., Kiljunen,S. and Skurnik,M. (2000) Bacteriophage ϕYeO3-12, specific for *Yersinia enterocolitica* serotype O:3, is related to coliphages T3 and T7. *J. Bacteriol.*, **182**, 5114–5120.
63. Melton,D.A., Krieg,P.A., Rebagliati,M.R., Maniatis,T., Zinn,K. and Green,M.R. (1984) Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Res.*, **12**, 7035–7056.
64. Brown,J.E., Klement,J.F. and McAllister,W.T. (1986) Sequences of three promoters for the bacteriophage SP6 RNA polymerase. *Nucleic Acids Res.*, **14**, 3521–3526.
65. Lee,S.S. and Kang,C. (1993) Two base pairs at −9 and −8 distinguish between the bacteriophage T7 and SP6 promoters. *J. Biol. Chem.*, **268**, 19299–19304.
66. Rosa,M.D. and Andrews,N.C. (1981) Phage T3 DNA contains an exact copy of the 23 base-pair phage T7 RNA polymerase promoter sequence. *J. Mol. Biol.*, **147**, 41–53.
67. Breitbart,M., Salamon,P., Andresen,B., Mahaffy,J.M., Segall,A.M., Mead,D., Azam,F. and Rohwer,F. (2002) Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA*, **99**, 14250–14255.
68. Breitbart,M., Hewson,I., Felts,B., Mahaffy,J.M., Nulton,J., Salamon,P. and Rohwer,F. (2003) Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.*, **185**, 6220–6223.
69. Breitbart,M., Felts,B., Kelley,S., Mahaffy,J.M., Nulton,J., Salamon,P. and Rohwer,F. (2004) Diversity and population structure of a near-shore marine-sediment viral community. *Proc. R Soc. Lond. B Biol. Sci.*, **271**, 565–574.
70. James Cann,A., Elizabeth Fandrich,S. and Heaphy,S. (2005) Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes*, **30**, 151–156.
71. Schneider,T.D. (1991) Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.*, **148**, 125–137.
72. Weinel,C., Nelson,K.E. and Tummler,B. (2002) Global features of the *Pseudomonas putida* KT2440 genome sequence. *Environ. Microbiol.*, **4**, 809–818.
73. Semenova,E., Djordjevic,M., Shraiman,B. and Severinov,K. (2005) The tale of two RNA polymerases: transcription profiling and gene expression strategy of bacteriophage Xp10. *Mol. Microbiol.*, **55**, 764–777.
74. Sullivan,M.B., Coleman,M.L., Weigele,P., Rohwer,F. and Chisholm,S.W. (2005) Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.*, **3**, e144.
75. Long,G.S., Bryant,J.M., Taylor,P.W. and Luzio,J.P. (1995) Complete nucleotide sequence of the gene encoding bacteriophage E endosialidase: implications for K1E endosialidase structure and function. *Biochem. J.*, **309**, 543–550.
76. Clarke,B.R., Esumeh,F. and Roberts,I.S. (2000) Cloning, expression, and purification of the K5 capsular polysaccharide lyase (KflA) from coliphage K5A: evidence for two distinct K5 lyase enzymes. *J. Bacteriol.*, **182**, 3761–3766.
77. Breitbart,M., Miyake,J.H. and Rohwer,F. (2004) Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett.*, **236**, 249–256.
78. Cermakian,N., Ikeda,T.M., Cedergren,R. and Gray,M.W. (1996) Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. *Nucleic Acids Res.*, **24**, 648–654.
79. Cermakian,N., Ikeda,T.M., Miramontes,P., Lang,B.F., Gray,M.W. and Cedergren,R. (1997) On the evolution of the single-subunit RNA polymerases. *J. Mol. Evol.*, **45**, 671–681.
80. Hendrix,R.W. (2003) Bacteriophage genomics. *Curr. Opin. Microbiol.*, **6**, 506–511.
81. Scholl,D., Rogers,S., Adhya,S. and Merril,C.R. (2001) Bacteriophage K1-5 encodes two different tail fiber proteins, allowing it to infect and replicate on both K1 and K5 strains of *Escherichia coli*. *J. Virol.*, **75**, 2509–2515.
82. Brunovskis,I., Hyman,R.W. and Summers,W.C. (1973) *Pasteurella pestis* bacteriophage H and *Escherichia coli* bacteriophage ϕII are nearly identical. *J. Virol.*, **11**, 306–313.
83. Schneider,T.D. (2000) Evolution of biological information. *Nucleic Acids Res.*, **28**, 2794–2799.
84. Shannon,C.E. (1949) Communication in the presence of noise. *Proc. IRE*, **37**, 10–21.
85. Schneider,T.D. (1991) Theory of molecular machines. I. Channel capacity of molecular machines. *J. Theor. Biol.*, **148**, 83–123.
86. Schneider,T.D. (1994) Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology*, **5**, 1–18.
87. Morris,C.E., Klement,J.F. and McAllister,W.T. (1986) Cloning and expression of the bacteriophage T3 RNA polymerase gene. *Gene*, **41**, 193–200.
88. Muller,D.K., Martin,C.T. and Coleman,J.E. (1989) T7 RNA polymerase interacts with its promoter from one side of the DNA helix. *Biochemistry*, **28**, 3306–3313.
89. Cheetham,G.M.T. and Steitz,T.A. (1999) Structure of a transcribing T7 RNA polymerase initiation complex. *Science*, **286**, 2305–2309.
90. Cheetham,G.M.T., Jeruzalmi,D. and Steitz,T.A. (1999) Structural basis for initiation of transcription from an RNA polymerase–promoter complex. *Nature*, **399**, 80–83.