Comparative Genomics of the T4-Like Escherichia coli Phage JS98: Implications for the Evolution of T4 Phages

Sandra Chibani-Chennoufi, Carlos Canchaya, Anne Bruttin, and Harald Brüssow*

Nestlé Research Center, Nestec Ltd., Vers-chez-les-Blanc, Lausanne, Switzerland

Received 17 June 2004/Accepted 13 September 2004

About 130 kb of sequence information was obtained from the coliphage JS98 isolated from the stool of a pediatric diarrhea patient in Bangladesh. The DNA shared up to 81% base pair identity with phage T4. The most conserved regions between JS98 and T4 were the structural genes, but their degree of conservation was not uniform. The head genes showed the highest sequence conservation, followed by the tail, baseplate, and tail fiber genes. Many tail fiber genes shared only protein sequence identity. Except for the insertion of endonuclease genes in T4 and gene 24 duplication in JS98, the structural gene maps of the two phages were colinear. The receptor-recognizing tail fiber proteins gp37 and gp38 were only distantly related to T4, but shared up to 83% amino acid identity to other T6-like phages, suggesting lateral gene transfer. A greater degree of variability was seen between JS98 and T4 over DNA replication and DNA transaction genes. While most of these genes came in the same order and shared up to 76% protein sequence identity, a few rearrangements, insertions, and replacements of genes were observed. Many putative gene insertions in the DNA replication module of T4 were flanked by intron-related endonuclease genes, suggesting mobile DNA elements. A hotspot of genome diversification was located downstream of the DNA polymerase gene 43 and the DNA binding gene 32. Comparative genomics of 100-kb genome sequence revealed that T4-like phages diversify more by the accumulation of point mutations and occasional gene duplication events than by modular exchanges.

Phage T4 is among the best-characterized biological systems (20); for example, a recent review on the T4 phage genome alone quoted 1,240 references (27). In the era of sequence-driven microbiology and comparative genomics, it is therefore surprising that only a single further T4-type phage genome sequence has been analyzed, that of the Vibrio phage KVP40 (26). This phage differed substantially from T4 in genome size (244 versus 169 kb for T4), making direct comparisons difficult. In fact, 65% of the 386 open reading frames predicted for KVP40 had no known functions and lacked any database matches; 26% of the KVP40 open reading frames shared up to 60% amino acid identity with T4 proteins, defining it clearly as a distant relative of phage T4. However, beyond the conserved structural and DNA replication modules, T4 and KVP40 have diverged substantially from each other by the acquisition of large blocks of unrelated genes.

To better investigate the mode of evolution in this phage group, it will be necessary to compare the genomes from more closely related T4-type phages. Early work with heteroduplex mapping of Escherichia coli phages T2, T4, and T6 DNA revealed that their genomes differed at about 30 loci, showing both substitution and deletion loops (21). However, this analysis also suggested a close matching of the phage DNA sequence over about 85% of the genome length. The correlation of the physical and genetic maps allowed the location of the regions of diversity. In the last decade several of these regions were sequenced for different T-even phages, and the results correlated well with biological differences. For example, the distinct host ranges of T2, T4, and T6 were associated with distinct tail fiber gene constellations (the major difference being the role of gp37 and gp38 as phage adhesins) (28, 32).

The distinct glucosylation pattern of phage DNA was linked to unrelated forms of α-glucosyltransferases in T2 and T6 and T4 (34). Sequence analysis identified gene replacements in the DNA replication genes (31). When six variable loci in a set of 52 natural isolates of T-even phages that all grew on enterobacteria were investigated by PCR sequencing (30), four loci showed a dimorphic pattern (only two gene constellations). These were the regions between the DNA replication genes 59 and 32, the immunity locus upstream of the b-gt (β-glucosyltransferase) gene, the topoisomerase, and the tail locus. In contrast, two regions were polymorphic, with up to nine distinct gene constellations (ip3/2 and ip1 loci encoding internal head proteins).

On the other hand, some efforts were focused on the sequencing of particularly well conserved structural genes like the major head gene g23 in T4-like phages infecting an even wider range of host bacteria. Phylogenetic tree analysis of these single gene sequence data led to the distinction of T-even and pseudo-, schizo- and exo-T-even phages: these phages showed increasingly distant relationships with the reference T4 phage, covering first DNA sequence and then decreasing protein sequence identity (33). Graded relatedness that varied with the phylogenetic distance of the bacterial hosts was also reported in temperate dairy phages (5). In contrast, the standard theory of phage evolution developed in the 1970s for lambdoid coliphages (3) is based on abundant horizontal exchange of functional gene clusters (modules) (7) across wide phylogenetic distances of host bacteria (17).

This model therefore does not postulate evolutionary lineages of phages, but at best only an evolutionary history for individual phage modules. However, it is premature to postulate a distinct mode of evolution in lambdoid and T4-like coli-
phages as long as these inferences rely on the sequences of single phage genes. Despite the prominence of phage T4 for the history of molecular biology, only limited comparative whole genome analysis was published for T4-like E. coli phages. The currently most comprehensive study compared random clones representing 20% of the genome from the pseudo-T-even phage RB49 with T4. RB49 belongs to the three of the 52 T-even phages from the Toulouse collection growing on enterobacteria that do not react with the diagnostic g32 (major head gene) PCR test. RB49 shared only weak DNA but clear protein sequence identity with T4 (10).

In the present report we analyzed 100 kb of the genome sequence from phage JS98, a typical T4-like coliphage stool isolate from Bangladeshi children described in the accompanying paper (8), to provide a broader basis for the comparative genomics of T4-like coliphages.

**MATERIALS AND METHODS**

**Phage selection and sequencing strategy.** Phage JS98, whose isolation is described in the accompanying manuscript (8), was chosen for detailed sequence analysis. Its preliminary characterization suggested a moderately divergent T4-like coliphage (8). JS98 phage DNA was sheared to about 1 kb before cloning into the plasmid vector pUC18; the plasmid was propagated in E. coli. The DNA was sequenced to an eightfold coverage. When the sequences were assembled, the JS98 genome was still represented by 20 contigs, suggesting a large number of host-lethal genes impeding cloning in E. coli. However, most contigs could be projected on the T4 map due to DNA sequence identity with T4 (10). This projection predicted that several contigs were nearly adjacent. Indeed, a number of missing JS98 DNA sequences could be obtained by the sequencing of PCR amplicons obtained with primers placed at the ends of the 20 contigs. This approach reduced the number of contigs to 13.

Several JS98 contigs lacked DNA sequence relationship with phage T4 (Fig. 1, contigs 8, 20, 19, and 40). Sequence matches of the predicted proteins with T4 or its closely related phage RB69 (http://phage.bioc.tulane.edu/) (contig 8: 22% and 40% amino acid identity with nrdC.5, 6 from T4, log E value $H11005/H11002$ 8 and $H11002$ 27, respectively; contig 20: 78% amino acid identity with e.6 homologue from RB69,
log E < 25; contig 19: 56 and 76% amino acid identity with mld4 and dens homologue from RB69, log E = -61 and -24, respectively; contig 40: 35% amino acid identity with arr.2-h homologues from RB69) allowed a tentative ordering of all available contigs (Fig. 1). The remaining gaps correspond in T4 to regions containing genes of mainly unknown functions or functions that are likely to interfere with their cloning.

Sequence analysis. Open reading frames have been predicted with the FrameD program (Toulouse Bioinfo INRA) with ATG, GTG, and TTG as possible start codons and a minimum size of 30 amino acids. Nucleotide and predicted amino acid sequences were compared in the databases of NCBI/GenBank; EMBL; PIR-Protein; SWISS-PROT; and PROPOSITE. The NCBI database contains the complete but as yet unpublished genome sequences from two E. coli T4-like phages (RB69 and RB49) and two Aeromonas T4-like phages (Ach1 and 44RR 28.1). Additional database searches have been conducted with BLAST (1) and PSI-BLAST at the NCBI and FASTA (24). A search for tRNA genes was done with the tRNascan-SE program (25). The motif search was performed with Pfam.

Nucleotide sequence accession number. The sequence data were deposited at GenBank under accession numbers AY746495, AY746496, and AY746497.

RESULTS AND DISCUSSION

Overall, 133 kb (80%) of the genome from JS98 was sequenced. A projection of the sequences on the T4 map is shown in Fig. 1. This tentative assembly should be interpreted with caution because we lack experimental evidence (except for the PCR bridging of supposed adjacent contigs, see above) that the order of the contigs in JS98 is the same as the parallel sequences in T4. Three large contigs representing 100 kb of genome information were analyzed in detail: contig 9, a 48.7-kb segment covering the morphogenesis module; contig 38, a 20.4-kb region containing tail fiber and DNA transaction genes; and contig 12, a 25.4-kb section with DNA replication genes.

Structural gene cluster of phage JS98. The structural genes in JS98 are found on two different contigs (contigs 9 and 38, Fig. 1). According to the tentative projection of JS98 on the T4 sequence, both structural gene clusters are separated by a segment of DNA transaction genes (Fig. 1). This separation of tail fiber genes from the rest of the structural genes is probably an ancient character because it was also observed in phage KVP40 (26).

Endonucleases. Figure 2 shows an alignment of the structural genes in JS98 contig 9 and T4 DNA. In addition to the homing endonuclease genes (segB, -C, -D, and -E), this region of T4 has only three small genes of unknown function that lack counterparts in JS98 (genes 5.3, 24.2, and msv Y.1). The homing endonuclease genes are not genuine phage DNA, but belong to selfish DNA elements that are frequently intron-associated (2, 11). These elements have efficiently invaded the T4 genome and are also found in the DNA transaction module opposite the tail fibre cluster (I-Tevl and segG) (Fig. 3) and the DNA replication module (segA, mohB, mohC, and I-TevII) (Fig. 4). They consistently lack a counterpart in the corresponding JS98 DNA. Inconsistent distribution of these elements has been described for T2, T4 and T6 phages (29). No homing endonuclease has as yet been detected in the JS98 genome. Apparently there are ill-defined barriers to intron promiscuity in bacteriophages (12, 13).

JS98 differs from T4 by a longer wac gene (encoding the phage neck-associated whiskers, which facilitate long tail fiber attachment; references for this and all other T4 genes mentioned are provided in reference 27), a shorter hoc gene (coding for an outer capsid protein in T4), and a possible duplication of g24 (encoding a head vertex protein). At the left side of the JS98 morphogenesis module lie genes for deoxynucleoside monophosphate kinase, gp57A and -B chaperons, and several auxiliary genes related to T4, but JS98 diverges from T4 3 kb further downstream (data not shown).

Head and tail. The structural proteins from JS98 shared amino acid sequence identity with T4 ranging from a minimum of 32% for gene 29 (the tail length determinant in T4) to a maximum of 90% for gene 23 (encoding the major head subunit precursor). Overall, the head gene module showed the highest degree of sequence conservation, but the extent of conservation differed substantially from gene to gene (50 to 90% amino acid identity) (Fig. 2). The next most conserved genes were the four tail genes (51 to 72% amino acid identity), which are mixed with head genes. This intermingling of head and tail genes is probably an ancestral character because it was also observed in vibriophage KPV40 (26).

Baseplate. The protein sequence identity between JS98 and T4 was 35 to 72% amino acid identity for the baseplate “wedge” module and 32 to 61% for the baseplate “hub” module (Fig. 2). The separation of baseplate genes into two clusters interrupted by tail and head genes is not a conserved trait because in vibriophage KPV40 all baseplate genes come in a single gene cluster (Fig. 4). We need sequence data from a third group of distantly related T4-like phages like cyanophage S-PM2 (15) to decide what constellation is the ancient configuration.

At the DNA sequence level, the JS98 head module revealed 72 to 81% base pair identity with T4 DNA, showing only very few small gaps. Individual baseplate wedge genes reached up to 75% base pair identity over the entire gene, while baseplate hub genes shared DNA sequence identity only over segments of the individual genes, resulting in many gaps in the alignment (Fig. 1). Possibly, the less conserved baseplate proteins are less bound by tight protein-protein interactions and have more free surface area. The three-dimensional structure of the T4 baseplate was recently determined by cryoelectron microscopy (22), allowing us to test the correlation between the degree of sequence conservation and location in the baseplate structure. In the wedge module gp9, gp11, and gp12 showed the lowest degree of sequence conservation, and these were also the most outside-oriented structures of the baseplate. gp12 is the short tail fiber, making the external and downside hexagonal garland of the baseplate. gp11 is the even more exposed linker protein connecting the gp12 fibers. gp9 is the most outside oriented protein of the upper, tail-oriented part of the baseplate. The central part of the baseplate hub is composed of the well-conserved gp5 and gp27, making the cell-puncturing device (19), and gp54 and gp48, making the contact to the tail sheath (22). The less well conserved gp26, gp51 (assembly catalyst?), gp28, and gp29 (tail length determinant) were not located in the baseplate cryostructure, precluding an analysis of the structure-conservation correlation.

Fibers. The whisker fiber gene wac, although bracketed by the relatively well-conserved baseplate wedge and the head modules, shared only 32% amino acid identity with its T4 homologue. The JS98 tail fiber genes are found in a different genome segment, contig 38, bracketed, as in phage T4 by the RNase-encoding gene mnh and the lysis protein t (Fig. 3). Also, the tail fiber proteins were less well conserved than the rest of
FIG. 2. Comparison of the morphogenesis modules from coliphages JS98 and T4. Alignment of the baseplate, tail, and head genes from T4 (upper line) and contig 9 from JS98 (lower line). Genes sharing protein sequence identity are linked by blue to red shading according to the color key provided at the bottom right. The likely function of the genes is indicated by the color of the arrows according to the color key at the bottom left. All T4 genes are annotated with their gene numbers, and selected genes are marked with their functions. The top line provides a base pair scale and the positions of the first and last depicted T4 genes on the T4 genome.

<table>
<thead>
<tr>
<th>Function</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseplate</td>
<td>Red</td>
</tr>
<tr>
<td>Tail</td>
<td>Yellow</td>
</tr>
<tr>
<td>Head</td>
<td>Blue</td>
</tr>
<tr>
<td>DNA Packaging</td>
<td>Green</td>
</tr>
<tr>
<td>Replication, Recombination</td>
<td>Orange</td>
</tr>
<tr>
<td>Head Protein</td>
<td>Black</td>
</tr>
<tr>
<td>Baseplate Protein</td>
<td>Dark Blue</td>
</tr>
<tr>
<td>Tail Protein</td>
<td>Dark Yellow</td>
</tr>
<tr>
<td>DNA Protein</td>
<td>Dark Red</td>
</tr>
</tbody>
</table>

The bottom line shows the positions of the first and last depicted JS98 genes on the JS98 genome.
FIG. 3. Comparison of the tail fiber and the adjacent DNA transaction modules from phages JS98 and T4. Alignment of the tail fiber genes (blue), DNA transaction and transcription genes (pink), inserted endonuclease (orange), and lysis (green) genes of contig 38 from JS98 (top) and T4 (bottom). The colors of the connecting shadings indicate the percentage of amino acid identity (see Fig. 2 for a key to the color code). The bottom line provides a scale in base pairs. T4 genes are annotated with the T4 gene terminology. JS98 open reading frames lacking a T4 complement are annotated with the closest match to another T4-like phage (strain in capital letters, percentage amino acid identity in parentheses) or a minus sign if they are lacking any database match.
FIG. 4. Comparison of the DNA replication module from phage JS98, T4, and RB69. Alignment of the DNA replication genes from phage RB69 (top), contig 12 from JS98 (center), and T4 (bottom). DNA replication genes are colored in yellow, DNA-modifying genes in violet, nucleotide metabolism genes in orange, endonuclease genes frequently found associated with mobile DNA (bottom), and the immunity gene is in blue. The T4 open reading frames are annotated with the conventional gene numbers. The RB69 open reading frames are numbered and the degree of amino acid identity are connected by shading, and the degree of amino acid identity is color coded (for the key, see Fig. 2).
the phage structural proteins. In addition, a clear gradient of sequence conservation between T4 and JS98 was seen along the tail fiber structure, decreasing from the highest values for the proximal (gp34) over the hinge protein (gp35) to the small (gp36) and the large distal tail fiber subunit (gp37) showing the lowest degree of conservation. The JS98 gene opposite the T4 gene 38 lacks sequence relatedness with T4.

The tail fiber adhesin domains are located in different genes in T-even phages. In phage T4, the adhesin is located in the C-terminal domain of gp37, while in other T-even phages the adhesin is a separate protein, gp38, which comes in at least three different alleles represented by T2, Ac3, and T6/Ox2 (32). The corresponding predicted JS98 proteins showed clear sequence identity with RB69 (75 and 55% amino acid identity with gp37 and gp38, respectively), T6 (47 and 79% amino acid identity for gp37 and gp38, respectively), and phage Ox2 (83% amino acid identity with gp38), but only low or no amino acid identity with the corresponding T4 proteins. Notably, the host ranges of JS98 and RB69 overlap on the pathogenic E. coli strains; this was not the case between JS98 and T4 (data not shown). The adhesin JS98 is thus a variant of the T6 adhesin allele. This region is a hotspot of lateral gene transfer not only between T4-like phages, but also between a much wider group of morphologically distinct coliphages (14).

Evolutionary analysis of structural genes: point mutations and graded relatedness. The accumulation of point mutations is apparently a major driving force in T4-type phage evolution. However, the degree of sequence diversification was not uniform. The head proteins were the most conserved structural proteins, probably due to substantial protein-protein interactions in the phage capsid. Table 1 shows the statistics of point mutations for the major structural components of the head and the tail sheath. It is apparent that the JS98 and T4 tail gene 18 homologues are separated by more base pair changes than the head gene 23 homologues. Two processes could account for this difference. One possibility is that the differences are the result of phages that diversified entirely separately and then recombined parts of their genomes. The nonhomogeneous sequence conservation over the structural genes could be cited as an argument for frequent recombination among this group of phages in nature. Alternatively, the different degrees of sequence diversification could reflect a differential selection pressure.

Support for this interpretation can be derived from the analysis of individual genes. For example, in gene 23, the second and fourth 250-bp segments showed significantly fewer overall base pair changes and a significant preference for the observed base pair changes to occur at the third base pair position, indicative of selection pressure to maintain the amino acid sequence. In addition, from the 128 codons that differed between g23 from T4 and JS98 exclusively at the third base position, all but one were synonymous changes (Table 1). A similar, although less impressive, avoidance of nonsynonymous base pair changes was also observed for the codons that differed only at the third base position of the g18 sequences (Table 1). In contrast, the first segment of gene 23, which encodes a segment that is cleaved off during maturation of the prohead (20) and thus under less structural constraint, showed a greater number of base pair changes and the bias towards changes at the third codon position was less marked. This second interpretation is also backed by the observation of a similar gradient of sequence identity, although at a lower percentage level, in the T4-KVP40 comparison (head > tail > base plate > tail fiber) (Fig. 5). Even the head genes from a T4-like phage isolated from a very distant bacterial host, cyanobacteria (15), still shared some amino acid sequence identity with T4 (Fig. 5). Comparable series of graded relatedness were seen between structural genes from members of the phage family Siphoviridae and were interpreted as evidence for vertical elements in the evolution of phages (5).

### Evolutionary analysis of structural genes: gene duplications and rearrangements

The structural genes from T4-like phages were apparently also shaped by gene duplications and genome rearrangements. For example, T4 gene 24, encoding in T4 a head vertex subunit, was probably duplicated in JS98 (Fig. 2). However, the “duplicated” gp24 homologues in JS98 differ substantially in amino acid sequence and both show high sequence identity with T4 gp24 and gp23, respectively. Both JS98 and T4 gp24 share 29% amino acid identity, suggesting that g23/g24 were already the result of an even earlier duplication event (16). Thus, instead of duplication of an ancestral JS98 g24 gene, the genes could be the consequence of sequential duplication and divergence of gene 23; the branching order is not certain.

One g24 copy was in addition split into two open reading frames and shared 70% amino acid identity with T4 gp24 and 28% amino acid identity with T4 gp23. The other g24 copy was integral and shared 55 and 28% amino acid identity with T4 gp24 and gp23, respectively. Both JS98 g24 copies shared DNA sequence identity (Fig. 1), while no DNA sequence identity was observed between them and JS98 g23, arguing for a more recent g24 duplication event. Duplication of structural genes seems to be a frequent observation in T4-type phage evolution, as demonstrated by the KVP40-T4 comparison (15) (Fig. 5). The KVP40 baseplate gene 12 is tandemly duplicated, while

### Table 1. Base pair exchanges in the major head and tail genes between phages JS98 and T4

<table>
<thead>
<tr>
<th>Region (bp)</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>Total</th>
<th>3rd/syn</th>
<th>3rd/nonsyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>g23</td>
<td>68</td>
<td>49</td>
<td>185</td>
<td>302</td>
<td>128</td>
<td>1</td>
</tr>
<tr>
<td>1–250</td>
<td>18</td>
<td>11</td>
<td>32</td>
<td>61</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>250–500</td>
<td>6</td>
<td>4</td>
<td>28</td>
<td>38</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>500–750</td>
<td>17</td>
<td>15</td>
<td>32</td>
<td>64</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>750–1000</td>
<td>6</td>
<td>3</td>
<td>25</td>
<td>34</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>1000–1250</td>
<td>10</td>
<td>9</td>
<td>27</td>
<td>46</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>1250–1500</td>
<td>11</td>
<td>7</td>
<td>41</td>
<td>59</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>g18</td>
<td>157</td>
<td>106</td>
<td>336</td>
<td>599</td>
<td>196</td>
<td>7</td>
</tr>
<tr>
<td>1–250</td>
<td>16</td>
<td>12</td>
<td>39</td>
<td>67</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>250–500</td>
<td>32</td>
<td>22</td>
<td>45</td>
<td>99</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>500–750</td>
<td>16</td>
<td>10</td>
<td>37</td>
<td>63</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>750–1000</td>
<td>22</td>
<td>17</td>
<td>48</td>
<td>87</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>1000–1250</td>
<td>24</td>
<td>17</td>
<td>43</td>
<td>84</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>1250–1500</td>
<td>15</td>
<td>4</td>
<td>42</td>
<td>61</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>1500–1750</td>
<td>17</td>
<td>12</td>
<td>42</td>
<td>71</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>1750–1980</td>
<td>15</td>
<td>12</td>
<td>40</td>
<td>67</td>
<td>25</td>
<td>0</td>
</tr>
</tbody>
</table>

*The first column describes the investigated region, which was either the entire gene or successive DNA segments of 250 bp; the second to fourth columns give the number of differences at the first, second, and third codon positions, respectively, of the two phages. The fifth column lists the total number of changes observed over the indicated gene regions. The sixth and seventh columns give the number of synonymous and nonsynonymous changes, respectively, in the codons that differ only at the third base position.
FIG. 5. Comparison of the structural gene modules in more distantly related T4-like phages. The T4 line is a copy from Fig. 2; it was aligned with the structural gene clusters from vibriophage KVP40 (top) and cyanophage S-PM2 (bottom). The putative gene functions and the amino acid sequence identity are color coded as specified in Fig. 3 (pale yellow here indicates 20 to 30% amino acid identity). Duplicated KVP40 genes are marked with two black or red arrows; the corresponding T4 gene was also marked. T4 genes located in a different genome position in KVP40 and deleted KVP40 genes are marked with two black or red arrows; the corresponding T4 gene was also marked. T4 genes located in a different genome position in S-PM2 and deleted S-PM2 genes are marked with two black or red arrows; the corresponding T4 gene was also marked. T4 genes located in a different genome position in KVP40 (top) and cyanophage S-PM2 (bottom). The putative gene functions and the amino acid sequence identity are color coded as specified in Fig. 3. (Panel yellow here indicates 20 to 30% amino acid identity.)
the tail gene 19 in KVP40 was duplicated with one copy in the baseplate and another in the head gene cluster (26).

gp24 plays an interesting role in T4 head maturation. The T4 prohead consists of an outer shell (gp23) and an inner scaffold (gp22). When gp24 is added to the prohead, maturation cleavage occurs: gp23 loses its N-terminal part, while gp22 is degraded completely (8, 20). Further functional data point to a link between gp24 and gp23: gp24 bypass mutations map in g23. Specifically, one class (trh) maps in those regions of gp23 that share sequence similarity with gp24 (16). It therefore seems plausible that gp24 evolved from g23 by developing a new accessory function. In fact, duplications liberate one copy of the duplicated genes from the constraints imposed by selection, since a reserve gene is still in place. Genes are therefore free to evolve new functions.

**DNA replication module.** The 25-kb contig 12 contains many of the DNA replication and recombination genes known from T4 (Fig. 4). Specifically, JS98 encodes close homologues of the T4 replicative helicase (gp41), the RecA-like recombination protein UvsX, the DNA polymerase (gp43), its two clamp-loader subunits (gp62 and gp44), and the sliding clamp (gp45), as well as the recombination-nuclease proteins gp46 and gp47 and, at some map distance, the endonuclease gp49, involved in resolution of the recombination junctions. Amino acid identity ranges from 54% (gp45) to 76% (gp49). Furthermore, the relative gene order of the essential replication and recombination genes was conserved between JS98 and T4 (Fig. 4).

Major differences between the corresponding JS98 and T4 regions were observed at the level of nonessential genes not involved in DNA replication and recombination functions, which were found in the same genome region. In fact, three groups of T4 genes lacked counterparts in JS98. Notably, the genes β-gt and α-gt, involved in glucosylation of the hydroxymethylcytosine residues in T4, are accompanied by the segA and mobB homing endonuclease genes, respectively, which suggests a mobile character for these regions (Fig. 4). Like RB69, JS98 contains more open reading frames in this region. Significant matches were observed with genes annotated as thymidylate synthase and methylthioadenine synthetase. Only open reading frame 10 showed a nonsignificant match with a sugar nucleotidyltransferase over part of its sequence (data not shown) Also, the auxiliary T4 genes ndG and ndD, which encode subunits of an anaerobic nucleotide reductase, lack complements in the corresponding genome region from JS98 (Fig. 4). In T4, ndD was associated with the endonuclease gene mobC, again suggesting a mobile DNA element.

For the region located between uvsX and the DNA polymerase gene 43, all but one JS98 gene matched RB69 genes from the corresponding region (Fig. 4). However, RB69 differs in this region from JS98 by rearrangement, insertion, and replacement of genes (Fig. 4), defining this region as a likely hotspot for gene insertions in T4-like coliphages. There are other but much smaller regions where JS98 differed from both RB69 and T4, e.g., near the T4 pin gene, encoding a protease inhibitor.

As in T4, JS98 has genes in this genome segment that are not involved in DNA replication; the superinfection immunity gene imm, the translational repressor gene regA, rpbH (encoding a RNA polymerase binding protein), and the late sigma factor gene 55, and all were sequence related to T4 and RB69. Interestingly, regA- and g55-like genes were also seen at identical position in KVP40, suggesting ancient insertion events in the DNA replication module.

Near the tail fiber gene cluster, the T4 genome contains a number of genes that block or reduce DNA synthesis when mutated (27). To the former group belong the essential genes g32 (the single-stranded-DNA-binding protein which builds the scaffold of DNA replication) and g59 (the loader of the gp41 helicase). Closely related JS98 genes were identified on contig 38 with closely related map order and sequence relatedness (61 to 76% amino acid identity) (Fig. 3). To the latter group belong genes encoding the α subunit of the ribonucleotide reductase (NrdA), thymidylate synthetase, dihydrofolate reductase (Frd), and RNase H, involved in the processing of Okazaki fragments (Rnh). JS98 showed again T4-related proteins (50 to 86% amino acid identity). Differences between JS98 and T4 were two endonuclease genes in T4 (I-TevI and segG), gene replacements (JS98 frd.1 and -2, flanked by genes lacking any database match), single gene indels (e.g., T4 nrdA.2), and a duplication of frd.3 in JS98 (Fig. 3).

**Diversification of T4-like phage genomes: an outlook.** The fact that three independent isolates of T4-like phages (T4, RB69 [http://phage.bioc.tulane.edu/], and JS98) isolated 40 years apart from two different geographical areas (New York and Dhaka) (23) can be aligned at the DNA level essentially over their entire genome suggests that modular exchanges, which are such a dominant feature in the evolution of the lambda-like supergroup of E. coli, are a less prominent process in T4-like phages. However, this observation should be interpreted with caution. Our current analysis of the JS98 sequence is biased towards the most conserved genome segments (structural and DNA replication genes). When extending the analysis to the remainder of JS98, we will probably see more substantial diversity, as already indicated by the worse DNA sequence alignment for the smaller contigs not analyzed in the present report (Fig. 1). In fact, the comparison of the T4-like phage genome maps on the Tulane University website (http://phage.bioc.tulane.edu/) revealed that the diversity of the sequenced phages was highest over those regions of the T4 genome lacking attributed gene functions.

Nevertheless, our current JS98 sequence analysis identified a number of regions where modular exchanges most likely had occurred. One such region encodes the receptor-recognizing fibers (Fig. 3). This region was previously identified as a recombination hotspot (14, 23, 32). Another likely modular exchange was seen at the level of the JS98 frd.1 and -2 genes (Fig. 3) Interestingly, in an independent study involving nineteen T4-like phages, three very different genome patterns have been reported for the frd.1 to frd.3 region, a nonessential T4 region (23). As in our comparison, most of frd.3 was conserved, while half of frd.2 and the frd2/frd.1 region showed strong polymorphism. Similarly polymorphic loci on the genomes from T4-like coliphages were previously identified by heteroduplex mapping (21) or the sequencing of PCR products in six different genome regions (30). However, the size of the modularly exchanged DNA segments seems to be smaller in T4-like phages than in lambdoid coliphages. Also, the number of distinct alleles seems to be lower; one larger survey identified predominantly two allelic forms. With respect to the head, tail, and baseplate genes or the major DNA replication genes, even a single allelic
form was identified. This single structural gene cluster maintained protein sequence identity over substantial evolutionary distances separating the bacterial hosts (proteobacteria and cyanobacteria) (Fig. 5), suggesting a substantial element of vertical evolution for this genome region in T4-like phages.

In the lambda supergroup of Siphoviridae a much lower degree of structural protein sequence conservation was seen when phages were analyzed over a comparable phylogenetic distance of their bacterial hosts. In fact, lambda-like Siphoviridae from gram-positive and gram-negative bacteria shared essentially only a comparable organization of the head and tail gene map (5). In addition, a number of sequence-unrelated structural gene modules were identified even within a single host species like E. coli (18). Diversification by the accumulation of point mutations is clearly the dominant mode of evolution over diversification by modular exchanges for the T4-like structural genes.

We identified two likely candidates for gene duplications in the T4-JS98 comparison (g24 and frd.3) and the T4-KVP40 comparison revealed that this is in fact a common observation (26) (Fig. 5). We are not aware of similar observations in lambda-like Siphoviridae. This difference might be of functional importance. Duplicated genes allow the accumulation of many mutations in one copy of the gene that is liberated from the constraints to maintain a potentially essential function. This process might be necessary for phages that evolve new functions predominantly from genes already existing in the phage genome. Lambda-like phages, in contrast, seem to have opted for the import of new functions by lateral gene transfer from sources outside of the gene pool of lambda-like phages.

In contrast to the classical modular exchange reactions, where blocks of functionally homologous gene are laterally transferred, the term moron (more DNA) was coined in lambda-like phages to identify additional open reading frames endowed with their own promoter and terminator that can integrate in various positions of the phage genome (18). Candidates for extra DNA were also identified in T4-like phages. For example, the region between the uvsX and g43 replication genes, encoding the β-glucosyltransferase and deoxytusine monophosphate hydroxymethylase involved in the T4-specific chemical modification of cytosine bases, looks like an insertion into T4 DNA. The same applies to the α-glucosyltransferase gene. In contrast to the morons from lambdoid phages, several of these extra DNA segments defined by the T4-JS98 comparison are flanked (α-gr and β-gr) or even bracketed (nrnD) by homing endonucleases. Insertion of these selfish DNA elements is a prominent process in T4 (see above). T4 might be quite permissive for this type of DNA because it occasionally introduces extra genes that developed useful functions for the phage, like the DNA-modifying enzymes that protect the phage DNA against destruction by restriction enzymes.

It is still premature or even wrong to conclude that T4- and lambda-like phages evolve with different modes. Such conclusions must be based on the comparison of more than three phage genomes. Completion of the sequencing of four further T4-like phage genomes (NCBI database, http://phage.bioc.tulane.edu/) will soon allow comparative genomics approaches with this workhorse of molecular biology.

Such data are not only of academic interest. T4-like phages are candidates for phage therapy of various E. coli infections (9). It will be important to understand how extra DNA can be introduced into the T4 phage genome and whether these extra genes are only to the benefit of the phage, as expected for an obligately virulent phage. Temperate phages, in contrast, can coexist with their lysogenic bacterial hosts. Some phage morons are therefore also of survival benefit for the bacterial host. In fact, temperate phages from bacterial pathogens frequently encode bacterial virulence factors (4, 6), which make such phages clearly unsuitable for phage therapy.

ACKNOWLEDGMENTS

We thank the Swiss National Foundation for financial support of Sandra Chibani-Chennoufi and Carlos Canchaya (grant 5002-057832). We thank Elizabeth Kutter for critical reading of the manuscript and Caroline Barretto for preparation of the GenBank file.

REFERENCES


