The Caspid of the T4 Phage Superfamily: The Evolution, Diversity, and Structure of Some of the Most Prevalent Proteins in the Biosphere

André M. Comeau and Henry M. Krisch

Laboratoire de Microbiologie et Génétique Moléculaires, Centre National de la Recherche Scientifique—Université Paul Sabatier-Toulouse III, Toulouse, France

The Escherichia coli bacteriophage T4 has served as a classic system in phage biology for more than 60 years. Only recently have phylogenetic analyses and genomic comparisons demonstrated the existence of a large, diverse, and widespread superfamily of T4-like phages in the environment. We report here on the T4-like major capsid protein (MCP) sequences that were obtained by targeted polymerase chain reaction (PCR) of marine environmental samples. This analysis was then expanded to include 1,000s of new sequences of T4-like capsid genes from the metagenomic data obtained during the Sorcerer II Global Ocean Sampling (GOS) expedition. This data compilation reveals that the diversity of the major and minor capsid proteins from the GOS metagenome follows the same general patterns as the sequences from cultured phage genomes. Interestingly, the new MCP sequences obtained by PCR targeted to MCP sequences in environmental samples are more divergent (deeper branching) than the vast majority of the MCP sequences coming from the other sources. The marine T4-like phage population appears to be largely dominated by the T4-like cyanophages. Using ~1,400 T4-like MCP sequences from various sources, we mapped the degree of sequence conservation on a structural model of the T4-like MCP. The results indicate that within the T4 superfamily there are some clear phylogenetic groups with regard to the more conserved and more variable domains of the MCP. Such differences can be correlated with variations in capsid morphology, the arrangement of the MCP lattice, and the presence of different capsid accessory proteins between the subgroups of the T4 superfamily.

Introduction

Bacteriophage T4 is a remarkably complex nanomachine that parasitizes and proficiently kills Escherichia coli. The T4 virion is assembled from a series of modular components including a large elongated head (~111 × 78 nm), a tail structure (~113 × 16 nm) that can be triggered to contract, and an intricate baseplate that incorporates the tail triggering mechanism and also the 6 long tail fibers that specifically bind to receptors on the surface of the host bacteria (reviewed in Leiman et al. 2003). Over 150 different phages infecting over 30 gram-negative bacterial species have a related morphology (Ackermann and Krisch 1997). The T4 genome is completely sequenced and the functions of the majority of its genes are known (Miller, Kutter, et al. 2003). The structure and function of many of T4’s essential components have been dissected in genetic, biochemical, enzymological (Karam and Konigsberg 2000; Fokine et al. 2004, 2006; Letarov et al. 2005), crystallographic (Shamoo et al. 1995; Karam and Konigsberg 2000; Thomassen et al. 2003; Fokine et al. 2005), and electron micrographic studies (Leiman et al. 2003; Fokine et al. 2004, 2006; Rossmann et al. 2004).

Apart from T4 itself, a number of other distant T4-like phages have been studied in recent years, including Aeromonas spp. phages, Vibrio phages (Matsuzaki et al. 1998, 1999, 2000; Miller, Heidelberg, et al. 2003), and cyanophages (Hambly et al. 2001; Mann et al. 2005; Sullivan et al. 2005; Weigele et al. 2007). These divergent T4-like phages infect hosts that are evolutionarily distant from the Enterobacteriaceae, the classical host of T4, and its closest relatives. They also vary somewhat from the classical T4 morphology, as the length of the head varies from smaller isometric forms (~85 × 85 nm) in the cyanophages (Hambly et al. 2001), to the more elongated forms (~140 × 80 nm) in some vibriophages and Aeromonas phages (Miller, Heidelberg, et al. 2003; Comeau et al. 2007) which contain larger genomes (>230 kb). There is variation in tail length as well, with the cyanophages having tails of up to ~180 nm in length (Hambly et al. 2001; Weigele et al. 2007). One of the criteria used to choose T4-like candidate phages for full genome sequencing was to assure representation of such morphological variants in the T4-like phage database and to include them in a full comparative analysis of the T4 superfamily (Nolan et al. 2006; Petrov et al. 2006; Comeau et al. 2007). Extensive phylogenetic analyses of these genomic sequences (Tétart et al. 2001; Desplats and Krisch 2003; Filée et al. 2006) have confirmed the existence of a large and extremely divergent superfamily of T4-like phages, a situation very similar to that now emerging for the unrelated T7 podovirus supergroup (Rohwer et al. 2000; Chen and Lu 2002; Hardies et al. 2003; Scholl et al. 2004). Until quite recently, the T4 superfamily had been believed to contain only 4 subgroups—the "true" T-evens which are coliphages very closely related to T4; the morphologically similar Pseudo T-evens which are, nonetheless, phylogenetically diverged from T4 and infect a broader range of hosts; the Schizo T-evens which are yet more divergent, morphologically distinguishable from T4, and infect Aeromonas and Vibrio spp.; and finally, the Exo T-evens which are extremely distant from T4, morphologically distinct, and infect cyanobacteria and thermophilic eubacteria (Desplats and Krisch 2003). However, this "simple" scenario had to be expanded as yet more T4-like phages were found in the marine environment (Filée et al. 2005). It became clear that the different subgroups of T4 phages that were previously thought to be narrowly restricted in their environmental range actually have a much more widespread distribution. In addition, the examination of these marine sequences related to
the T4 gp23 major capsid protein (MCP) has resulted in the identification of a much larger set of thus far uncharacterized subgroups of the T4 superfamily.

The environmental MCP sequencing project reported here was targeted on the most divergent of T4-like phages. Simultaneously with this effort, a massive amount of un-targeted, metagenomic data became available from the Sorcerer II Global Ocean Sampling (GOS) expedition (Rusch et al. 2007). This metagenomic data gave us a unique opportunity to look at the diversity and evolution of an enormous and unselected compilation MCP sequences. In addition, we could compare this metagenomic data with sequences obtained from our large-scale cultured phage genome sequencing project (Nolan et al. 2006; Petrov et al. 2006; Comeau et al. 2007) and also with the >100 sequences collected by our earlier targeted analysis of MCP sequences in diverse ocean samples (Filée et al. 2005). The resulting comparative analysis of more than a thousand homologs of the T4 MCP has allowed us to obtain a more complete representation of the vast diversity within the T4 superfamily.

Materials and Methods

Aquatic Samples

Thirteen samples of the concentrated viral size fraction (Suttle et al. 1991) were prepared from seawater originating from the Gulf of Mexico, the western Arctic Ocean, the Chukchi Sea, and the fjords and bays of coastal British Columbia. Ten of these samples were identical to those used in our previous study (Filée et al. 2005); the additional 3 new samples came from the coast of British Columbia (#364—Salmon Arm; #420/481—Jericho Pier). Five freshwater samples were also examined, 2 from lakes Cultus and Chilliwack in British Columbia and one from the Canal du Midi near Toulouse in the southeast of France (43°33’ 39”N 1°28’ 29”E). Additional information on sample sites and sampling techniques are available elsewhere (Short and Suttle 2005; Comeau et al. 2006). Environmental Polymerase Chain Reaction and Sequencing

Degenerate polymerase chain reaction (PCR) primers ScExoT-F (5’-CWC GTC AAY TGA AAG CTC AA-3’; positions 788–807 in T4 g23 NC_000866) and ScExoT-R (5’-AWT TKM AYA CCG TAR CGA GT-3’; positions 1423–1442) were designed on the basis of alignments of cultured T4 superfamily phages to preferentially target Schizo T-even and noncyanophage Exo T-even phages. Control PCR reactions with the resulting primers amplified Schizo T-even and not T-even/Pseudo T-even phages nor cyanophages. PCR amplification, DNA purification, cloning, and sequencing of the amplified environmental sequences were carried out as described previously (Filée et al. 2005), except PCR cycling conditions which were modified as follows: initial denaturation at 94 °C for 1 min; followed by 35 cycles of denaturation at 94 °C for 30 s, annealing at 45 °C for 1 min, and extension at 72 °C for 1 min; followed by final extension at 72 °C for 9 min. GenBank accession numbers for the novel sequences presented in this paper are EU236767–EU236786.

Comparative Genomics and Phylogenetic Analyses

Environmental sequences were compared with complete (http://www.ncbi.nlm.nih.gov/genomes/static/phg.html) and draft (http://phage.bioc.tulane.edu) T4 superfamily genomes. The GOS data set (Rusch et al. 2007) was queried using the built-in Blast function of the CAMERA database (Seshadri et al. 2007). Sequence manipulations were carried out in BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html); and alignments of small data sets were conducted using ClustalW (Thompson et al. 1994). Alignments of large data sets could not be handled by traditional alignment programs and had to be performed using the MUSCLE program (Edgar 2004) using a gap opening penalty of −3 and a gap extension penalty of −0.275. The degree of conservation in alignments (amino acid identity) was calculated using ProtSkin (http://www.mcgmnr.ca/ProtSkin). Protein secondary structure was predicted using NNPREDICT (Kneller et al. 1990) through the STRAP interface (Gille and Frommel 2001). The MCP phylogenetic tree was constrained, given the limitations imposed by the size of the data set, to the protein distance and Neighbor-Joining methods, all carried out using the PHYLIP v3.66 package (http://evolution.genetics.washington.edu/phylip.html). The Jones-Taylor-Thornton (JTT) model was used for protein distance calculations and a nearly full-position alignment was used for the MCP tree construction given the availability of complete, or near full-length, sequences (for both cultured phages and the metagenome) and given the sheer depth of coverage of overlapping fragments for the GOS data. However, 8 common gaps at final alignment positions 1–25 (N-terminus), 156–166, 185–212, 369–380, 390–394, 476–480, 511–520, and 624–651 (C-terminus) were removed before construction of the MCP tree (527 residue final length). During construction of the tree, sequences from 4 cultured phages (MV9a/b, MV12, and MV13) as well as 2 environmental sequences (CS43 and GOS-1131) had to be removed from the analysis due to the creation of unusually long branches, indicating possible PCR chimeras which were verified using Bellerophon (Huber et al. 2004). The MCP tree was rooted using the RM378 sequence, primarily not only for presentational reasons but also because this sequence was the most divergent characterized MCP sequence currently available.

MCP Homology Modeling

The homology model (Dunbrack 2006; Ginalski 2006) of the core of RB49 gp23 was constructed using the SWISS-MODEL server (Arnold et al. 2006) and the DeepView protein modeling/assessment program (Schwede et al. 2003) using the solved T4 gp24 structure (Protein Data Bank: 1YUE) (Fokine et al. 2005) as a template. Portions of the protein not present in the template were modeled using HMMSTR (Bystroff and Shao 2002) (extra
Results and Discussion
Capsid Genes in Cultured T4-Like Phage Genomes

The mature T4 capsid shell (or lattice) is composed essentially of only 4 proteins (fig. 1A)—gp23* ("*" indicates the processed, mature form of the gene 23 protein), the MCP; gp24*, the vertex protein; highly immunogenic outer capsid (Hoc), an accessory protein which protrudes from the surface of the capsid; and small outer capsid (Soc), a small protein (80 aa) which binds to the junctions between gp23* hexamers creating a protein "grid" on the surface of the capsid (reviewed in Leiman et al. 2003). A fifth component, one dodecamer of gp20, the portal protein, is located at the unique vertex that attaches the capsid to the contractile phage tail structure and does not participate in forming the remainder of the capsid lattice.

Only the MCP, which represents 55% of the mass of the capsid (Leiman et al. 2003; Fokine et al. 2004), and the portal protein appear to be required for capsid shell formation. Experiments with T4 mutants in Soc and Hoc (fig. 1A) indicate that these functions are facultative (Miller, Kutter, et al. 2003). Gene 23 can substitute for gene 24 function if it has certain single-amino acid mutations that allow it to occupy the vertex positions normally filled by gp24* (Fokine et al. 2005, 2006). Such a functional substitution is not surprising because the genes encoding for gp23 and gp24 result from an ancient duplication event and they retain nearly 31% protein sequence similarity (17% identity) in T4. The function of the accessory Hoc protein is unclear, although recently Dabrowska et al. (2006) have suggested a role for Hoc in T4’s evasion of the mammalian immune system. Various large recombinant proteins have been fused to Hoc and Soc to create a chimeric production/display system on the T4 capsid without apparent detriment to phage capsid function or phage viability (Ren et al. 1996; Jiang et al. 1997; Ren and Black 1998; Sathaliyawala et al. 2006; Shivachandra et al. 2006). The other accessory capsid protein, Soc, is nonessential, but confers added stability to the phage capsid structure, protecting it from osmotic shock, alkaline pH, and temperature denaturation (Ishii et al. 1978; Iwasaki et al. 2000; Steven et al. 1992). Thus, in particularly harsh conditions, the presence of Soc protein, or its functional analogs, could be advantageous.

A survey (fig. 1B) of the T4 superfamily genomes that are completely sequenced reveals that only 6 of them (T4, T6, RB14, RB32, RB69, and JS98) have homologs of all 5 of the T4 capsid shell proteins. All these coliphages are very closely related to T4. The Soc protein appears to be the shell protein that is the least present in T4-like phages. However, the small size of the Soc protein (~80 aa) makes it difficult to identify distant homologs. There are 10 Soc homologs in the databases, and among these, only the trio of RB14, RB69, and JS98 sequences have modestly diverged from T4 (at both termini). Like Soc, Hoc is a facultative protein that is frequently missing in the T4-like phage genomes (fig. 1B), but it is retained in slightly more phages (12) than the former. With the significant exception of the Exo T-even subgroup, the gp24 capsid vertex protein is found in all T4 superfamily phages for which we have sufficiently complete genomic sequence data to make an evaluation, including the recently completed JS98 (Zuber et al. 2007) which shows a duplication of its g24. The Exo T-even subgroup includes 4 cyanophage genomes (P-SSM2, P-SSM4, S-PM2, and Syn9) (Mann et al. 2005; Sullivan et al. 2005; Weigele et al. 2007) and the genome of phage RM378 which infects Rhodothermus marinus, a thermophilic eubacterium (Hjørleifsdóttir et al. 2002; Blondal et al. 2003). These cyanophages and RM378 have isometric icosahedral capsid structures, unlike the prolate icosahedral form of the other T4-like phages. This important morphological difference may reflect the lattice composition of these phages that apparently have only gp23* and gp20 homologs. Possibly because of a slight structural alteration in their gp23* subunits, they could have obviated the requirement for a distinct vertex protein (see above). Alternatively, these distant isometric T4-like phages may use a vertex protein analog instead of a gp24 homolog, as has been suggested by the cryoEM reconstruction and genomic analysis of Syn9 (Weigele et al. 2007).

All the T4 superfamily phages have both the MCP and the gp20 portal proteins, the latter being involved in DNA packaging also serves as the initiator complex for prehead formation and connects the capsid to the tail structure (Leiman et al. 2003). Sequences in both of these highly conserved genes have been used for the PCR amplification of T4-like sequences in the environment (Zhong et al. 2002; Dorigo et al. 2004; Filée et al. 2005; Jia et al. 2007). The size of the MCPs from the cultured phages for which we have full-length sequences (27; supplementary fig. S1, A65-T84) and ModLoop (Fiser and Sali 2003) (novel loops L135-A151, E160-F171, N200-S213, G235-S250, T332-K336, R374-G381, and K418-A422). A portion of the RB49 sequence (loop L156-I180) was removed from the model building due to its very poor conservation among the ~1400 gp23 sequences. Model quality was assessed within DeepView and using Verify3D (Eisenberg et al. 1997) and VADAR (Willard et al. 2003). The final model had 93.1% of residues within favorable Ramachandran areas (ignoring glycines), compared with 96.9% for the solved 1YUE structure, and had a final root-mean–square (RMS) distance to 1YUE of 2.00 Å across 353 corresponding residues. Final RMS between our model and a previous model (Protein Data Bank: 1Z1U) proposed by Fokine et al. (2005) was 3.55 Å across 304 corresponding residues (the “Insertion” domain was excluded due to slightly different predicted positions). Protein alignment conservation was mapped onto the model structure using ProtSkin (http://www.mcgmnrm.ca/ProtSkin) and visualized in DeepView. Conservation groups were delimited from the phylogenetic tree and were as follows: the “Near T4” group (46 sequences) was composed of all the cultured phages, except the Exo T-evens, plus 13 environmental clones from Filée et al. (2005); the “Far T4” group (277 sequences) was composed of RM378, plus 19/20 of the new clones from this study, clone 37510 from Filée et al. (2005), and the 256 RM378-like GOS hits obtained using either RM378 or our clone ScExo373-21 as queries; finally, the “Cyan T4” group (1,077 sequences) was composed of all cultured cyanophages, plus one of our clones (ScExo420-3), 70 Filée et al. (2005) clones, and the top 1,000 GOS hits against RB49 gp23.
Supplementary Material online) varies somewhat (from 412-560 aa; a 36% difference), but overall the protein is quite well conserved, with ~74% of the protein alignment showing at least 50% similarity. The average identity among the 27 cultured phages over the full length of the protein is ~53% and there is only one major variable segment, located approximately between residues 210–290 (supplementary fig. S1, Supplementary Material online). This variable region is flanked by the primers we normally used to amplify environmental g23 sequences, leading to

Fig. 1.—Organization and comparative genomics of the T4 phage capsid shell lattice proteins. (A) Schematic diagram of the phage capsid lattice in T4 wild type and in the various single and multiple mutants in the capsid genes (Iwasaki et al. 2000; Fokine et al. 2004, 2006). The simplified structure of the triple mutant (g24⁻ hoc⁻ soc⁻) has not yet been experimentally confirmed, but given the presence of only gp23 in this lattice, this seems to be the most plausible configuration. The components of the T4-like phage lattice are the hexamer of gp23*, processed MCP; the pentamer of gp24*, processed vertex protein; Hoc, highly immunogenic outer capsid protein; and Soc, small outer capsid protein. One dodecamer of gp20, the portal protein, replaces gp24* at the unique vertex that attaches the capsid to the contractile phage tail structure. Note that 5 hexamers of gp23* surround each vertex, but 6 hexamers surround each individual gp23* hexamer on the capsid faces (i.e., between vertices). Adapted from Fokine et al. (2006). (B) Comparative genomics of the capsid genes in T4 superfamily genomes, showing their size, orientation and context of the four principle proteins (in black, dark gray or double-lined arrows). Additional T4 genes are in light gray with the genes bordering the capsid genes labeled to indicate their identity. The presence of known/putative mobile elements (e.g., mobile endonucleases) are indicated by hatching and genes not present in the T4 genome (frequently orphans) are in white. The presence of a possible extra gene not included in the original KVP40 analysis (Miller, Heidelberg, et al. 2003) is indicated by the dashed arrow with an “?”. Diagonal lines indicate inversions with respect to the genes/synteny of the preceding phage. Distances between genes/ clusters are indicated in kb, with the distance separating the soc cluster from the rest in the first four phages ranging from 84-91 kb (between the arrows).
the observation of variable-size PCR amplicons (Filée et al. 2005).

Marine Diversity of T4-Like Capsid Genes

In an attempt to “fill in” the T4 superfamily phylogenetic tree, which had an unexplained paucity of culture-independent sequence representatives of the Schizo T-even and noncyanophage Exo T-even subgroups (Filée et al. 2005), which include phages such as RM378 of the thermophilic eubacteria _R. marinus_ (Hjorleifsdottir et al. 2002), we designed a new set of degenerate MCP primers (ScExoT-F and ScExoT-R) with enhanced specificity for these subgroups. This new primer set was used in PCR reactions using as template the 10 marine samples of concentrated viral size fraction from our previous study (Filée et al. 2005) as well as 3 new samples from other marine habitats. Given that some cultured Schizo T-even phage hosts have been isolated from freshwater (_Aeromonas_ spp.) (Ackermann and Krisch 1997), we also attempted PCR amplification on 5 samples from 2 British Columbia lakes (Cultus and Chilliwack) and a canal in the southeast of France (Canal du Midi). Of these 18 samples, only 3 were positive for PCR amplification (all marine; data not shown), with one sample showing a single band of the incorrect size which was later determined to be spurious. The PCR amplified fragments from the remaining 2 positive samples (#373—Salmon Arm; #420—Jericho Pier) were cloned and analyzed. These gave 20 unique _gp23_-like sequences ranging in size from 643 to 694 bp. Surprisingly, almost all of these sequences (19/20) were most closely related to the phage RM378 (51–65% protein similarity), the exceptional sequence being related to the cyanophage P-SSM2 (80% protein similarity). The 19 unique RM378-like nucleotide sequences were 91% identical among themselves, indicating a surprisingly low level of diversity in the 2 samples. Although we cannot exclude the possibility of amplification or sampling artifacts, the low level of positive samples, and the absence of Schizo T-even sequences in these, raises the possibility that Schizo T-even phages may be much rarer in nature than the other T4 superfamily subgroups. Alternatively, the few cultured Schizo T-even sequences used to design the primers may simply not be representative of the subgroup as a whole. This may be the case given that the host bacteria for these phages (_Aeromonas_ and _Vibrio_ spp.) are often extremely abundant in the culturable fraction of the bacterioplankton, yet represent only a minor fraction of the total, culture-independent bacterial biomass (Thompson et al. 2004).

Having completed this focused environmental sequencing, we combined this additional data with the recently released Sorcerer II GOS expedition metagenomic data (Rusch et al. 2007). This massive amount of culture-independent sequence data gave us a unique opportunity to go beyond a relatively modest targeted effort to fill in the denuded branches of our MCP phylogenetic tree by permitting us to analyze the sequence diversity of literally 1,000s of _gp23_ homologs isolated from the global marine sampling. An additional incentive to make this compilation came from the astonishing fact that 5 of the 6 most overrepresented protein families in the GOS metagenome were either T4 structural proteins or enzymes (with _gp23_ actually being ranked first) (Yoosseph et al. 2007). Performing simple Blast searches of the GOS metagenome using various T4 capsid proteins as the query sequences (table 1) confirmed the omnipresence of the marine members of the T4 superfamily. _Gp20_ and _gp23_ homologs are by far the most abundant sequences in the database, with over 3,000 hits each (_E_ < 10^-4_), and are fairly evenly distributed among ~45 sampling sites. The _gp24_ vertex protein is the next most abundant capsid protein, with ~900 homologs but with generally lower _E_ values than those obtained for the first 2 proteins. This presumably reflects a lower level of _gp24_ sequence conservation and the possible replacement of the vertex protein function in many phages (e.g., Exo T-evens) by analogs. The Hoc protein sequence had only ~110 database hits and the Soc protein had no hits in the GOS metagenome, even using a very permissive cutoff (_E_ < 10^-5_). As mentioned previously, the small size of this latter protein may make it difficult to identify homologs or, alternatively, this function may be supplied in distant phages by analogs. In summary, the abundance of the capsid shell proteins in the GOS metagenome follows the same pattern of gene preservation as the cultured phage genomes—with Soc being the sequence least present in the metagenome, followed by Hoc, then _gp24_, and finally by _gp23_ and _gp20_ that are the most frequent sequences having nearly 4 times more hits in the metagenome than all the others. A comparable situation occurs for the proteins that are located internal to the T4 phage capsid—proteins which form part of the scaffold during prohead assembly or serve functions in host takeover/DNA defense once they are injected with the phage DNA during infection (Leiman et al. 2003; Comeau and Krisch 2005; Depping et al. 2005). The _gp21_ prohead protease, responsible for maturing many of the T4 capsid proteins as outlined above, and the essential _gp22_ scaffold protein (Black et al. 1994) both have strikingly different metagenomic profiles than the other internal capsid proteins. These 2 proteins are well conserved within the T4 superfamily, which correlates with their abundance (~1,400–1,800 hits) in the GOS metagenome, but many of the other internal proteins of the capsid have no hits in the GOS metagenome (table 1) which is probably related to their moderate to low conservation even within the known members of the T4 superfamily.

Diversity of the MCP (_gp23_)

The MCP is the foundation of the both the T4-like phage’s capsid structure and also of the operant phylogeny of the T4 superfamily (Desplats and Krisch 2003; Filée et al. 2005). By combining the large amount of very diverse sequence data now available, we hoped to obtain an overview of how the structure of this important protein has evolved in the different branches of the superfamily in the course of evolution. Thus, a comprehensive alignment was made using all available cultured phage _gp23_ sequences (44), all previous marine environmental sequences (85), the new sequences reported in this study (20), along with the top 1,000 GOS hits against RB49 _gp23_ (not T4, for reasons explained below), the top 100 GOS hits against RM378 _gp23_ (to capture very divergent sequences), and
the top 156 GOS hits against one of our new sequences (ScExo373-21) to include additional extremely distant homologs. This ensemble of data represented a total of 1,405 gp23 sequences, but of these, only 1,399 sequences were retained for the final analysis after 6 “problematic” sequences were discarded (see Materials and Methods). The resulting alignment (fig. 2) shows fairly good conservation along most, but not all, of the protein sequence, mirroring the conclusions previously obtained by analysis of the MCPs from cultured phages. The existence of the significant variable region (positions ~150–300 in the current alignment) was fully confirmed and, in addition, there is some sequence plasticity at both extremities of the protein. It is also evident from visual inspection that there are clear-cut sequence subgroups. First, there is separation of the cultured phage sequences (C) from nearly all the others, with the exception of a few of the Filée et al. (2005) environmental sequences (F). This is mostly the result of the sequence in the variable region (centered around position ~225) in the cultured phages being somewhat longer than T4-type phages isolated in the wild. There is also a striking cohesion among the new environmental sequences reported in this study (ScExo) with the 256 GOS hits (GOS lines immediately above ScExo) resulting from using either phage RM378 or the sequence ScExo373-21 as queries. Finally, there is also strong cohesion among the top 1,000 GOS sequences resulting from using RB49 gp23 as the query (GOS).

These clear general trends were reinforced in quantitative detail by a Neighbor-Joining phylogeny of all the ~1,400 gp23 sequences (fig. 3). At the “top” of the tree, all cultured, non-Exo T-even (cyanophages, RM378, ΦSMB14) phages are well grouped together, along with about half of the Filée et al. (2005) environmental sequences (46 total). These sequences form a phylogenetic group representing the Near T4 phages. Close to the root of the tree are the RM378-like sequences, the majority of the new sequences obtained from the present study, and 256 GOS sequences that form several clusters which are very divergent from the rest of the known sequences. These sequences form a Far T4 phylogenetic group, and, except for RM378, there are no cultured representatives in this part of the T4-like phylogenetic space. Clearly, much work remains to be done on these most divergent members of the T4 superfamly phages which include the Exo T-evens subgroup. The only Exo T-even phages for which we do have substantial ecological (Mann 2003) and genomic information (Mann et al. 2005; Sullivan et al. 2005; Weigele et al. 2007) are the cyanophages, and it is clear that these dominate the T4-superfamily tree and, therefore, the GOS metagenome. The Cyano T4 group includes all the cultured cyanophages, the other half of the Filée et al. (2005) environmental sequences, the clone ScExo420-3 from this study, and all of the top 1,000 GOS hits generated using RB49 (a coliphage, not a cyanophage) as the query sequence, whereas 3,243 sequences were obtained with coliphage T4 and only 2,070 sequences using RM378. It should be noted that in this analysis the vast majority of the T4 hits overlap with cyanophage hits and are probably mostly cyanophages, as is suggested by the phylogeny of the top 1,000 GOS hits to RB49 (fig. 3).

Structure and Evolution of the MCP (gp23)

Our further analysis of the MCPs in the T4 superfamily involved an investigation of the structure of the gp23
A structural homology model (Dunbrack 2006; Ginalska 2006) was constructed using the MCP of phage RB49 because, among the available cultured phages, this gp23 sequence has the closest sequence match (∼20/33% identity/similarity vs. ∼17/31% for T4 gp23) to T4 gp24, which served as the template structure because it is currently the only member of the gp23/24 family whose structure is known (Fokine et al. 2005). The structural model resulting from such an analysis (fig. 4; supplementary data set S1, Supplementary Material online) is in excellent agreement with a previous model for the T4 MCP (Fokine et al. 2005). The new prediction for gp23 has
a mixed secondary structure with slightly more β-strand content than the earlier sequence-based prediction (fig. 4A blue strands compared with fig. 2C yellow columns). The MCP structure has a compact central core and an adjacent less ordered Insertion domain (probably mostly random coil) that is likely to be responsible for the interaction between monomers in the hexamer, as it is in gp24 (Fokine et al. 2005). The T4 gp24 structure and the RB49 gp23 model share their structural folds with the MCP of the unrelated phage HK97, even though the two amino acid sequences have diverged considerably from the latter (only ~11% and ~13% identity, respectively), given that structural homologies appear to be conserved much longer than the primary sequence. In fact, Fokine et al. (2005) argue that the HK97 fold may be “universal” in phage capsid proteins.

Our MCP model could then be used to map the location of sequence divergence in the 3-dimensional structure. Such “sequence divergence mapping” permits the easy visualization of the conserved, and hence the potentially most important, motifs within a protein structure. Similarly, such a map highlights the most variable segments within the structure, presumably those that are responsible for variations in capsid morphology and other more subtle kinds of subgroup divergence. Although the N-terminal domain is not well conserved (fig. 4B and supplementary fig. S1, Supplementary Material online), the central core portion of the gp23 protein is. The Insertion domain, which corresponds to the major variable region previously alluded to in the MCP alignment (fig. 2 and supplementary fig. S1, Supplementary Material online), is the least conserved domain. When the gp23 structural model is assembled into a hexamer (the functional unit of the capsid lattice) (Leiman et al. 2003) and this is color coded to reflect the degree of conservation within the different groups of the T4 phylogenetic tree (see above and fig. 3), the sequence conservation-structure relationships become self-evident (fig. 4C). Such a color-coded hexamer structure is shown again for reference in the upper left of the figure. Note that the precise position of the
protruding, weakly conserved N-terminus (α2-3) is somewhat uncertain. The Near T4 group (primarily cultured phages) has good overall conservation of sequence, and strikingly so in the “outer ring” (formed by multiple β-strands); the only exception is the β14 strand at the C-terminus that is poorly conserved. The Cyano T4 group

![Diagram of MCP monomer and assembled hexamer]

**Fig. 4.**—Homology model of the gp23 MCP monomer and assembled hexamer. (A) Protein homology model of the RB49 MCP monomer, using the T4 gp24 vertex protein monomer structure (Fokine et al. 2005) as the template. The α-helix (red) and β-strand (blue) composition of the protein is indicated, along with the N- and C-termini of the molecule. A series of residues are indicated (white font) to show the progression of the polypeptide chain. Note that the model begins at residue 65 because the N-terminus of gp23 is cleaved during phage head maturation. Also indicated are certain key α-helices and β-strands (yellow font) as well as the Insertion domain which is probably responsible for coordinating the interaction between monomers. (B) As in (A), except here the coloring of the molecule is according to the level of amino acid sequence conservation. The degree of conservation, scale at right, is based on amino acid identity from the global alignment in figure 2 (≈1,400 sequences) and was calculated using ProtSkin (www.mcgrnmr.ca/ProtSkin). (C) Protein homology model of the RB49 MCP hexamer, based upon the T4 gp24 vertex protein pentamer structure and a previously proposed gp23 hexamer model (Fokine et al. 2005). The molecules are colored according to amino acid sequence conservation, as in (B), yet subdivided into the different T4 superfamily phylogenetic groups delimited in figure 3. The phylogenetic groups were 1) the complete data set alignment in figure 2 (≈1,400 sequences; Global); 2) an alignment of all cultured phage sequences nearest to T4 (T-evens, Pseudo T-evens, and Schizo T-evens, excluding the Exo T-evens), plus 13 of the Fileè et al. (2005) environmental sequences closest to T4 (46 total; Near T4 group); 3) an alignment of 1,077 sequences including all the cultured cyanophages (Exo T-evens) and the environmental cyanophage-like sequences (70 from Fileè et al. [2005], clone ScExo420-3 from this study, 1,000 GOS hits using RB49 as the query; Cyano T4 group); or 4) an alignment of 277 sequences containing the cultured phage RM378 (an Exo T-even) and the environmental RM378-like sequences (clone 37510 from Fileè et al. [2005], 19 sequences from this study, and 256 GOS hits using RM378/ScExo373-21 as the queries; Far T4 group).
has an outer ring less conserved than the Near T4s, yet has a well-conserved central core (primarily helices α7–8). Finally, the Far T4 group shows the least conserved outer ring and is also the least conserved overall (more blue throughout). Mutant analyses of the T4 MCP indicate that the outer ring residues are responsible, in part, for the morphology of the head, whereas the central core contains residues that allow gp23 to also function as the vertex protein (Black et al. 1994; Fokine et al. 2005). The sequence group shows clear differences with respect to these 2 regions, and it is probable that these lead to morphological changes in the capsid and to different capsid lattice arrangements, and novel interactions with capsid decorating proteins such as Hoc and Soc.

Conclusion and Perspectives

Our analysis shows that the diversity of T4 superfamily capsid proteins from the GOS environmental metagenome follow the same general patterns as do their homologs from cultured phage genomes. Interestingly, the novel environmental gp23 MCP sequences obtained in our targeted sequencing are quite divergent (deeper branching) from most other known sequences, including the overwhelming majority of the GOS metagenome that is dominated by cyanophage-like T4 phages. These results suggest that the T4 superfamily’s diversity has not yet been fully delimited and that there are certainly additional more distant T4-like phages yet to be discovered in the environment. Directly related to this issue, we are using this comprehensive gp23 sequence analysis to iteratively design a new set T4 superfamily PCR primers that should be yet more inclusive in its capacity to amplify distant T4-like phage sequences from the environment.

A homology model of the MCP, combined with the mapping of protein diversity from ~1,400 environmental and cultured sequences, shows that T4 superfamily phylogenetic groups have clear differences with respect to conserved and variable structural areas of their capsid proteins. These differences can be correlated with the diversity of capsid morphology, capsid lattice arrangements, and accessory protein content interactions in this phage family. The conclusions of this structure/diversity analysis would be made much stronger by the availability of the structures for the MCP in several different branches of the T4 superfamily. Given the central role that the MCP plays in coordinating interactions between the different essential and accessory components in the arrangement of the capsid lattice, having a panel of several representative structures could lead to a major advance in our understanding of how virion nanomachines are constructed and have evolved. From this point of view, having the X-ray crystallographic structures of the MCPs of various T4 phage morphotypes, such as the isometric forms of cyanophage S-PM2 and the deep-branching Exo T-even RM378, could be both fascinating and highly informative. Such data would put us in a better position to understand the structural and evolutionary features of the T4-like MCP that have allowed it to become one of the most abundant protein species in the biosphere.

Supplementary Material

Supplementary figure S1 and data set S1 as well as color versions of figures 1 and 3, are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the CNRS. We thank our colleagues Curtis Suttle (University of British Columbia, Vancouver) for the marine samples and Christine Arbiol (CNRS IFR109) for DNA sequencing/figure consultation. We also thank Viknesh Sivanathan for his advice on the STRAP interface and Shibu Yooseph for his help with the CAMERA database and GOS data set. In agreement with the Convention on Biological Diversity (http://www.cbd.int), the genetic information from the CAMERA database analyzed in this publication may be considered the genetic patrimony of the countries from which the samples were procured. A.M.C. gratefully acknowledges the support of the Les Treilles Foundation and H.M.K. that of the Kribu Foundation.

Literature Cited

Depping R, Lotzhaus C, Meyer HE, Ruger W. 2005. The mono-ADP-ribosyltransferases Alt and ModB of bacteriophage T4: target...


Filé J, Bapteste E, Susko E, Krisch HM. 2006. A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. Mol Biol Evol. 23:1688–1696.


Hervé Philippe, Associate Editor

Accepted March 21, 2008