Analysis of the phage sequence space

Lima-Mendez, Gipsi; Toussaint, Ariane; Leplae, Raphaël

Published in:
Virology

DOI:
10.1016/j.virol.2007.03.047

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com
ScienceDirect
Virology 365 (2007) 241–249
VIROLOGY
www.elsevier.com/locate/yviro

Minireview

Analysis of the phage sequence space: The benefit of structured information

Gipsi Lima-Mendez *, Ariane Toussaint, Raphaël Leplae

Service de Conformation de Macromolécules Biologiques et de Bioinformatique (SCMBB), Université Libre de Bruxelles, CP 263, Boulevard du Triomphe, 1050, Bruxelles, Belgium

Received 8 January 2007; returned to author for revision 7 March 2007; accepted 28 March 2007
Available online 7 May 2007

Abstract

Phages are the most abundant biological entities on Earth and are central players in the evolution of their bacterial hosts and the emergence of new pathogens. In addition, they bear an enormous potential for the development of new drugs, therapies or nanotechnologies. As a result, interest in phages is reviving. In the genomic era, our perspective on the phage sequence space remains incredibly sparse. The modular and combinatorial structure of phage genomes is largely documented. It is confirmed by new sequence information and it fuels a recurrent debate on the need to revise phage taxonomy. The absence of structured, computer readable information on phages is a major bottleneck for an extensive global analysis of phage genomes and their relationships, but such information is essential to reassess phage classification. Based on the ACLAME database, which is dedicated to the organization and analysis of prokaryotic mobile genetic elements, we discuss here how structured information on phage-encoded proteins helps global *in silico* analysis and allows the prediction of prophages in bacterial genome sequences, providing access to additional phage sequence information.

Keywords: Bacteriophages; Comparative genomics; Prophage detection; Phage classification

Introduction

In the last century, studies on bacteriophages were fundamental to the emerging fields of molecular biology and genomics (Fiers et al., 1976). Then, after several decades of languishing as tools for molecular biology, phages regained the interest they deserve due to their incredible abundance and population dynamics ($>10^{30}$ tailed bacteriophages on Earth and $\sim 10^{25}$ infections every second), their genetic diversity (Pedulla et al., 2003) and the major role they play in bacterial evolution and possibly in planetary biogeochemical cycles (Wilhelm and Suttle, 1999). In addition, phages represent a huge potential for new therapies and nanotechnologies (Levin and Bull, 2004; Zhang, 2003). Along with plasmids and other elements, phages mobilize pathogenicity determinants and hence, are central players in the emergence of new pathogens (Banks et al., 2002; Boyd and Brussow, 2002) and more generally in the divergence between closely related bacterial strains and species (Brussow et al., 2004; Canchaya et al., 2003).

First revealed by electron microscopic analysis of heteroduplex DNA between different lambdoid phages (Westmoreland et al., 1969), the mosaic structure of phage genomes is now more easily assessed by comparison of their DNA sequences (Casjens et al., 1992). Reshuffling of genetic information can occur by recombination between phages growing lytically in a single host cell, between a growing phage and one or more prophages residing in a latent state in the infected cell, or between latent prophages, even if defective, residing in the same host genome (Hendrix, 2003).

Despite the abundance and the limited size of their genomes, the number of phages characterized genetically is amazingly low (<20) and phage genomics is still in its infancy. Only 403 phage genomes were available on the NCBI website on March 7, 2007, a total of a little over 17.5 Mb, i.e. the equivalent of only 3.5 *Escherichia coli* K-12 chromosomes. In addition, the distribution of sequenced phage genomes among bacterial host species is biased towards a small number of hosts, with no significant change in recent years (see Table 1 for a comparison between 2003 and 2006).

This little information and the limited sequence similarity between orthologs make the functional assignment of phage

* Corresponding author. Fax: +32 2 6505425.
E-mail address: gipsi@scmbb.ulb.ac.be (G. Lima-Mendez).

0042-6822/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2007.03.047

Table 1
Evolution of the number of phage genomes sequenced per bacterial genus

From 306 phages 2006		From 184 phages in 2003	
Host	Nb phages	Host	Nb. phages
Escherichia	47	*Escherichia*	39
Staphylococcus	35	*Streptococcus*	17
Pseudomonas	26	*Lactococcus*	15
Streptococcus	19	*Mycobacterium*	14
Vibrio	17	*Pseudomonas*	13
Salmonella	16	*Vibrio*	11
Lactococcus	16	*Staphylococcus*	10
Mycobacterium	15	*Bacillus*	8
Bacillus	14	*Salmonella*	6
Burkholderia	12	*Spiroplasma*	4
Lactobacillus	9	*Chlamydophila*	4
Sulfolobus	9	*Yersinia*	4
Chlamydophila	6	*Burkholderia*	3
Xanthomonas	5	*Lactobacillus*	3
Spiroplasma	4	*Methanothermobacter*	2
Yersinia	4	*Xanthomonas*	2
Streptomyces	3	*Streptomyces*	2
Aeromonas	3	*Mycoplasma*	2
Prochlorococcus	3	*Haemophilus*	2
Mycoplasma	3	*Sulfolobus*	2
Bordetella	3	*Listeria*	2
Listeria	3	*Listeria*	2
	NA	3	
	Methanothermobacter	2	
	Clostridium	2	
	Acholeplasma	2	
	Haemophilus	2	
	Shigella	2	
	Synechococcus	2	

The table is sorted by number of phage genomes per host. Only those bacterial genera for which 2 or more phage genomes have been sequenced are shown (data from http://www.ncbi.nlm.nih.gov/genomes/static/phg.html). NA stands for not assigned.

proteins difficult. In addition, annotations retrieved from GenBank sequence files may be obscured by the widely divergent terminologies used for different phages. Such freestyle annotations are of little use for computer-assisted analysis. The marked tendency for a conserved organization of functions along phage genomes (Brussow et al., 2004; Canchaya et al., 2003; Casjens et al., 1992; Pedulla et al., 2003) does help for function prediction, but overall, full-sized or cryptic prophages are usually not easily identifiable and remain very poorly annotated (Casjens, 2003), if even recognized.

The first global analysis of the phage proteome did not reveal any single gene common to all phages (Rohwer and Edwards, 2002) and confirmed a previous conclusion from more limited comparative analysis that the classical taxonomy system based on viral particle morphology does not always fit a phylogeny based on sequence analysis (Brussow and Hendrix, 2002; Hendrix et al., 2000; Lawrence et al., 2002; Nelson, 2004; Pedulla et al., 2003).

An easily accessible, complete and duly annotated repository of phage and prophage sequences is often called for and would undoubtedly help in setting standards for phage comparative genomics. None is so far available and most existing sequence repositories, e.g., GenBank (Benson et al., 2005), do not provide dedicated data structures and related file formats for such elements. Useful information may not be readily available, and a number of phages have no gene field mentioned in their GenBank genome sequence files (e.g., phage Mu, GenBank accession AF083977).

The ACLAME database (Leplae et al., 2004) aims at providing a reticulate classification of the prokaryotic 'mobilome', i.e. all the proteins encoded by prokaryotic mobile genetic elements (MGEs), whether extrachromosomal or integrated in the host genome (see http://aclame.ulb.ac.be). Based on the modular/combinatorial structure of MGEs (transposons, plasmids, phages and genomic islands), ACLAME exploits the idea that grouping together different types of MGEs sharing identical functions (e.g. replication and site-specific recombination, transposition, conjugation, etc.) within a single repository, optimizes the use of experimental evidence available on any element to support the functional annotation of another element (Toussaint and Merlin, 2002).

ACLAME version 0.2 provides access to 10,446 proteins as defined in 184 sequenced phage genomes (4 dsRNA, 10 ssRNA, 30 ssDNA and 140 dsDNA phages). In this review, we use that and additional more recent phage sequence information to illustrate how the structuring of phage genomic information opens the way to a new type of large-scale analysis that is essential to better understand phage biology and evolution and to conceive an appropriate mode of phage classification.

Clustering the phage proteomic pool

ACLAME version 0.2 allows for visualizing 10,446 annotated proteins extracted from the NCBI GenBank files of 184 phages, individually and in families of functionally related proteins. The ACLAME protein families are built by a clustering procedure based on sequence similarity (Enright et al., 2002) as originally described for a set of proteins encoded by 119 sequenced phage genomes (Leplae et al., 2004). Similar procedures were recently used to generate Phage Orthologous Groups (POGs) of proteins encoded by 164 completely sequenced dsDNA phages (Liu et al., 2006) and "phamilies" of proteins encoded by 30 mycobacteriophages (Hatfull et al., 2006). In ACLAME, each family has an identifier (cluster ID). Fifty-two percent of the families contain 3 or more members (5488 proteins) and about one-third of the proteins analyzed are singletons (3792 proteins, i.e. 36%) or in two-member families (1166 proteins, i.e. 11%). These percentages remain similar if proteins from newly sequenced phage genomes are included in the data set (O. Zekri and R. Leplae, unpublished results). Moreover, despite the different sets of phage proteins considered and the different clustering algorithms and parameters used to build the ACLAME families, POGs and "phamilies", the trend in the distribution of proteins among families is similar. In the three cases, the average number of proteins per family is around 2 (2.0, 1.9 and 2.2, respectively) and about half of the total number of proteins analyzed remain as singletons or in pairs (47%, 48% and 50%, respectively).

ORFan phage proteins

It has been claimed that phages encode more ORFan proteins than bacterial genomes (Edwards and Rohwer, 2005) although, to our knowledge, an analysis of the frequency and distribution of ORFan genes/proteins in phage genomes has not been published. When each of the 10,446 phage proteins was compared with the non-redundant protein sequence database from NCBI (NRDB-NCBI version 11-Sep-2004), using the PSI-BLAST program with a score threshold of 0.01 and four iterations as parameters, the proportion of singleton ORFans (with just one hit) represents ~31% of the proteins, i.e. twice the percentage of ORFans usually found in bacterial genomes (Siew and Fischer, 2003) or in plasmids (Leplae et al., 2006). Counting proteins with up to 4 hits raises the percentage to ~50%, leading to the conclusion that the proportion of ORFans is indeed higher in phages than in plasmids or bacterial chromosomes. The few Archaeal viruses that are in ACLAME version 0.2 have the highest proportion of ORFans (around 80%, Fig. 1). Otherwise, no obvious correlation can be detected between the abundance of ORFans and phage size or host (data not shown).

The most frequently represented phage functions

Each family of 3 or more phage proteins in ACLAME is assigned with a function and its identifier (function ID), based on available experimental evidence and/or reliable sequence similarity for several proteins within the family. This functional annotation deserves some discussion. Annotations retrieved from the GenBank files are not always very informative, even for paradigm phages such as λ, Mu, or ϕ29. Phage specific functions are not available in any of the two most commonly used ontologies, Gene Ontology (Harris et al., 2004) and MultiFun (Serres and Riley, 2000). GO, the only structured ontology available online, has become the *de facto* standard for describing the principal attributes, the molecular function, biological process, and cellular component of knowledge about gene products across many databases. However, it is very much oriented towards eukaryotes and contains only half a dozen phage-related terms. An initial list of functions was developed within ACLAME version 0.2 to annotate phage and plasmid protein families (Leplae et al., 2006). It has subsequently been revised and completed to put together the first version of the PhiGO phage ontology (Toussaint et al., submitted for publication), which will replace the ACLAME list of phage functions in the next version of the database and will be incorporated in GO. PhiGO is accessible at http://aclame.ulb.ac.be/Classification/phage_functions.html. It includes relevant GO terms (usually molecular functions) in which case, the annotation is linked to GO, providing access to any supplementary information available in that database.

The ACLAME web interface displays for each protein (i) the annotation retrieved from the GenBank file (or a revised manual annotation for a few reference genomes), (ii) the result of PSI-BLAST (Altschul et al., 1997) searches (4 iterations, *E* value <0.01) against proteins in NRDB-NCBI, SwissProt (Boeckmann et al., 2003) and SCOP (Andreeva et al., 2004) and (iii) a multiple sequence alignment of all the proteins within a given family generated by ClustalW (Thompson et al., 1994) using the default parameters. The search results made with Hidden Markov Models derived from the multiple alignments are also available for each protein family. Individual phages can be visualized as a succession of proteins organized in the order of their coding genes. Each protein is tagged and linked with the family it belongs to and with the corresponding function when assigned (for an example see http://aclame.ulb.ac.be/perl/Aclame/Genomes/prot_view.cgi?mode=genome&id=mge:70). Thus individual proteins are accessible via the protein families as well as in the context of the phage genome.

The distribution of the main general functional categories in the 184 annotated genomes can be calculated (Fig. 2). It reveals

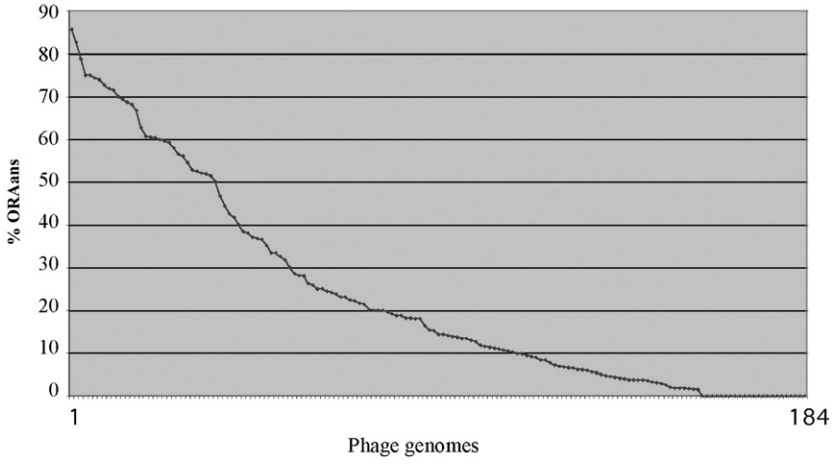

Fig. 1. Distribution of ORFan proteins in phages. Each point on the curve represents the percentage of singleton ORFans on a given genome. The genomes with the largest number of ORFans are Spiroplasma viruses SVTS2, 1-C74 and 1-R8A2B, which are between 6 and 8 kb long, followed by *P. aeruginosa* phage phiKZ (280 kb), mycobacteriophages Barnyard (70 kb) and Bxz1 (156 kb), Halovirus HF2 (78 kb), Vibriophage VpV262 (46 kb), *S. islandicus* filamentous virus (40 kb) and phage RM 378 from *R. marinus* (130 kb). Except for Archaeal viruses, few of which are in ACLAME version 0.2, no obvious correlation could be detected between the abundance of ORFans and phage size or host (data not shown).

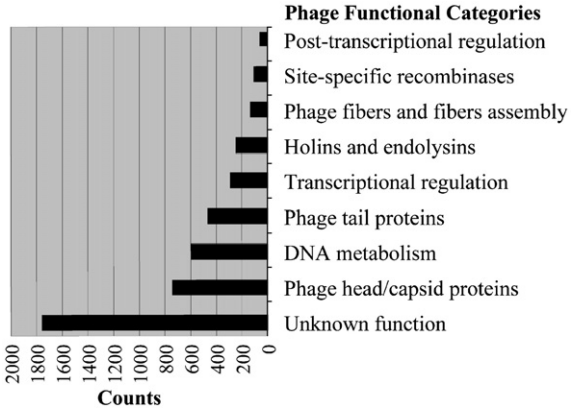

Fig. 2. Distribution of the main functional categories in phages. All functions related to head morphogenesis as portal, terminases, scaffold proteins, capsomeres, etc., were counted into the phage head-capsid category. DNA metabolism includes all functions related to DNA synthesis, DNA replication, recombination and repair except site-specific recombinases, which were counted separately because of their relevance to a possible distinction between temperate and virulent phages.

a large excess of proteins of unknown function, fitting the known genetic variety of phages and the large number of ORFan proteins just discussed. Other main functional categories include structural proteins (heads, tails and fibres), DNA metabolism (DNA replication, recombination and repair), transcriptional and post-transcriptional regulation or lysis of the host, present in all phages, and site-specific recombination, which is restricted to temperate phages.

The largest phage protein families in ACLAME version 0.2 (cluster:vir:0 to cluster:vir:6) contain from 112 to 51 proteins. They include transcriptional repressors (cluster:vir:1, 100 proteins and cluster:vir:5, 51 proteins), tyrosine site-specific recombinases, i.e. integrases (cluster:vir:2, 83 proteins) and proteins with unknown function (cluster:vir:4, 54 proteins). Although direct comparison is impossible due to the different protein sets analyzed and procedures used, the trend is the same as in POGs where the largest families include integrases, tail tape measure proteins (TMP, see further comments below), transcriptional repressors and terminases. As illustrated by cluster:vir:0 (112 proteins) and cluster:vir:3 (80 proteins), several of the largest families contain multidomain proteins, which are brought together by the clustering procedure because they share one or more domains. This local similarity is a source of ambiguity and possible errors for functional annotation. Several proteins in cluster:vir:0 were annotated as tail tape measure protein (TMP) in the corresponding GenBank sequence files. Two other families have that same function, cluster:vir:85 (11 proteins) and cluster:vir:816 (3 proteins). These are homogeneous and respectively include the well-characterized λ gpH (Katsura, 1990) and T4 gp29 (Abuladze et al., 1994) TMP and their relatives. In cluster:vir:0, the GenBank sequence file annotations obviously rely on similarity with phage TP901-1 TMP, which, to our knowledge, is the only one in the family that has been experimentally shown to bear that function (Pedersen et al., 2000). Some conserved domains within the 112 proteins in cluster:vir:0 match COG (Tatusov et al., 2003) or Pfam (Bateman et al., 2004) domains with only one of them mentioning a TMP signature (PF05017). That domain is present on only 50 of the proteins in cluster:vir:0 (data not shown). The "phage tail length tape measure protein" function is presently assigned to the family in ACLAME, although further analysis is needed to determine which of the 112 proteins are actual TMP. Cluster:vir:3 is another group of multidomain proteins, several of which have been experimentally demonstrated to be phage fibres [as gpS-S′ in phage Mu (Grundy and Howe, 1984); gpH and gpG in P2 (Haggard-Ljungquist et al., 1995) and gpTal in TP901 (Vegge et al., 2005)]. The cluster is therefore presently assigned with the phage fibre function.

Phages can have overlapping open reading frames (Pavesi, 2006; Zajanckauskaite et al., 1997), which are not always properly identified/annotated in GenBank files. An example would be the holin/anti-holin couples acting as a clock during phage infection. Holins are regulated by anti-holins (S_{107} in λ), which are often encoded by the same gene, in the same frame but using alternative start codons (for other examples see Wang et al., 2000). Holins and their cognate anti-holins having almost identical sequences are expected to cluster within the same families. However, the cluster:vir:151 family, containing proteins such as the paradigm λS_{105} holin, do not contain any holin/anti-holin couples, indicating that one of the two gene products was probably not annotated in the GenBank sequence files. The 'holin' function (function:58) is used for those ACLAME families but this is an arbitrary choice and further refinements need to be made for such peculiar families.

Cluster and function linkage: conservation of functional organization on phage genomes

Phage morphogenesis is an intricate and well-timed building process, which involves many specific interactions between proteins participating in assembling heads/capsids, tails and fibers. In addition, the corresponding genes tend to remain linked on phage genomes (Casjens et al., 1992). This tendency can be assessed along the genomes in ACLAME. The frequency of each pair of genes (defined as pairs of adjacent cluster IDs) occurring among 140 dsDNA phages can be computed and compared with their expected frequency using the binomial formula. As illustrated in Fig. 3, the most significant pairs of protein families (hence of adjacent genes) are also pairs of the same two ACLAME/GO functions, reflecting the fact that pairs of non-similar proteins fulfilling the same function are often adjacent on their cognate genomes (Overbeek et al., 1999). Terminases are members of a large number of conserved pairs. Formed of large TerL and small TerS subunits, they cleave the phage DNA as part of the maturation process. With portal proteins, they are among the most conserved phage proteins in both their sequence and relative location on the genomes (Casjens, 2003). Out of 128 conserved pairs of functions involving TerL (ACLAME function:6), 41 are TerL-TerS (function:303), 35 are TerL-portal protein (function:179) (Fig. 3A) and 8 are TerL-head scaffold protein (function:69) pairs (not shown). As expected from the intimate interactions

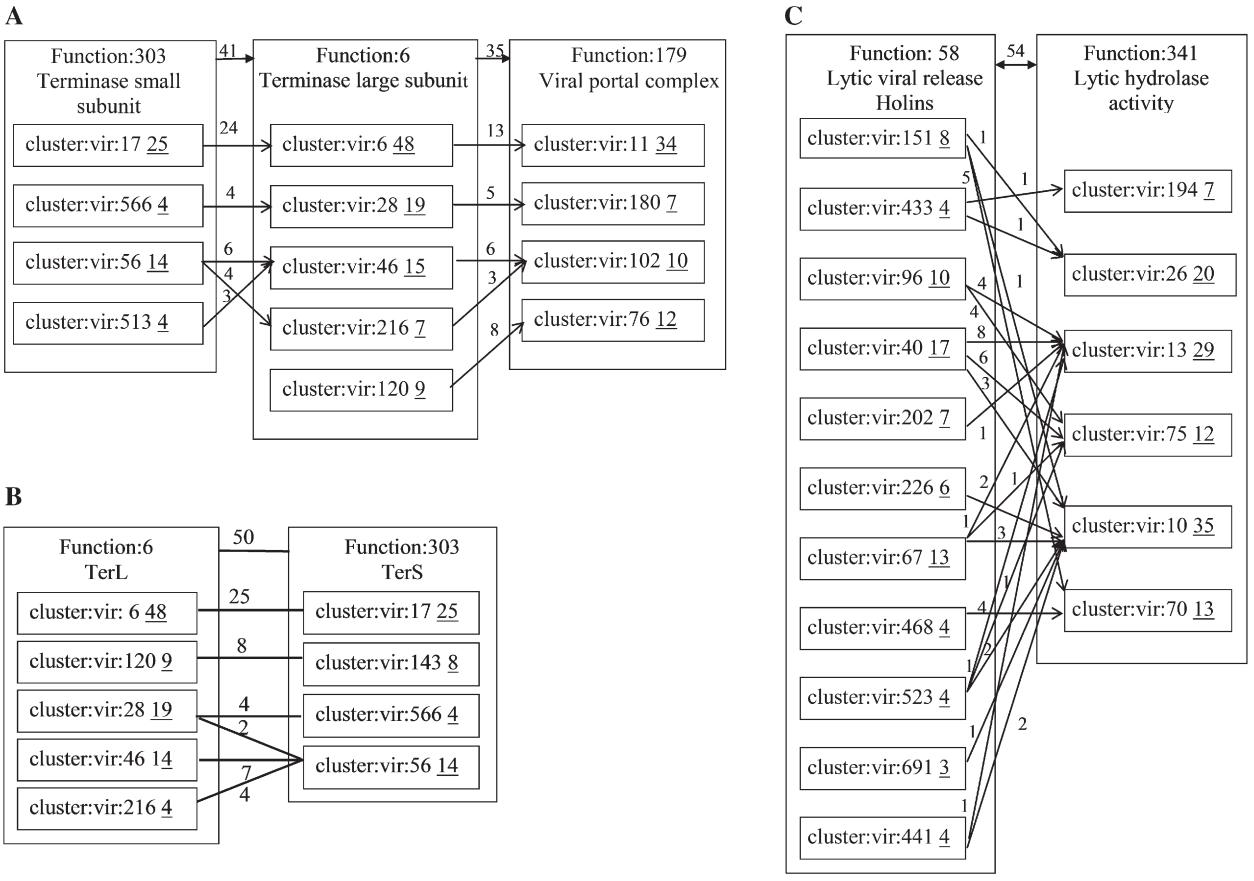

Fig. 3. Conserved gene/protein pairs detected on phage genomes. (A) Pairs involving the large terminase subunit TerL. Proteins in TerL families were associated with small terminases TerS and portal proteins. A near one-to-one association between the three families can be observed. Underlined numbers represent the occurrences of proteins in the family (also in panels B and C). (B) Association between large and small terminase subunits. Because *terL* and *terS* genes are not always adjacent, some of the associations shown here were not parts of the 34 adjacent pairs shown in panel A. Cognate partners could not be identified for all terminases; no TerS partner was found for any member of cluster:vir:152 and cluster:vir:55. (C) Holins and endolysins. The graph shows the association in pairs of proteins belonging to families annotated with the functions holin and endolysin. Proteins in one holin family were associated with depolymerising enzymes of different families and vice-versa.

between TerL and TerS (Casjens et al., 2005), families of TerL subunits are associated with a limited number of TerS families whether adjacent (Fig. 3A) or not (Fig. 3B) in the genome.

Individual members of holin families are associated with partners from different endolysin families, contributing to 54 conserved function pairs (Fig. 3C). This ought to reflect the lack of interaction between these proteins during the lysis process and the observed cross-complementation between endolysins and holins from unrelated phages (Wang et al., 2000).

Besides these well-known conserved pairs, the analysis can reveal new conserved associations, which may provide hints for new functional assignments. Terminases for instance often form conserved pairs with cluster:vir:4 members, which would thus appear as good candidates to participate in the maturation process. Proteins in cluster:vir:4 bear a signature of HNH endonucleases (IPR002711). It has been suggested (Crutz-Le Coq et al., 2002) that protein gp13 from lactococcal bacteriophage bIL170, a member of this cluster, could be part of a functional module with terminase and be involved in phage DNA packaging. These authors propose that the protein could act as a site-specific endonuclease analogous to that of the large terminase subunit, or as a structure-specific endonuclease, clearing branched DNA prior to packaging, as T4 endonuclease VII (Golz and Kemper, 1999).

From protein families to the phage population landscape

Protein families provide a basis for displaying the relationships between all phage genomes. A global view of all phage relationships at the protein sequence level can be derived from a pair-wise similarity score calculated as the number of protein families in common between each pair of genomes divided by the number of proteins of the shorter of the two. The similarity matrices can be represented as heat maps, where the rows and columns represent the phages and the cells represent the similarity score as a colour gradient. The darker the cell the more similar the two phages are and vice-versa. In Fig. 4, all protein families assembled from 306 phage genome sequences (available at NCBI as of Feb, 2006) have been included. Phages cluster according to the type of their genome. Not unexpectedly, ssRNA and ssDNA phages separate from the bulk of dsDNA phages. The latter form interconnected super-groups, supporting the idea that phages may arise from a common gene pool (Hendrix et al., 1999). Virulent phages documented as belonging

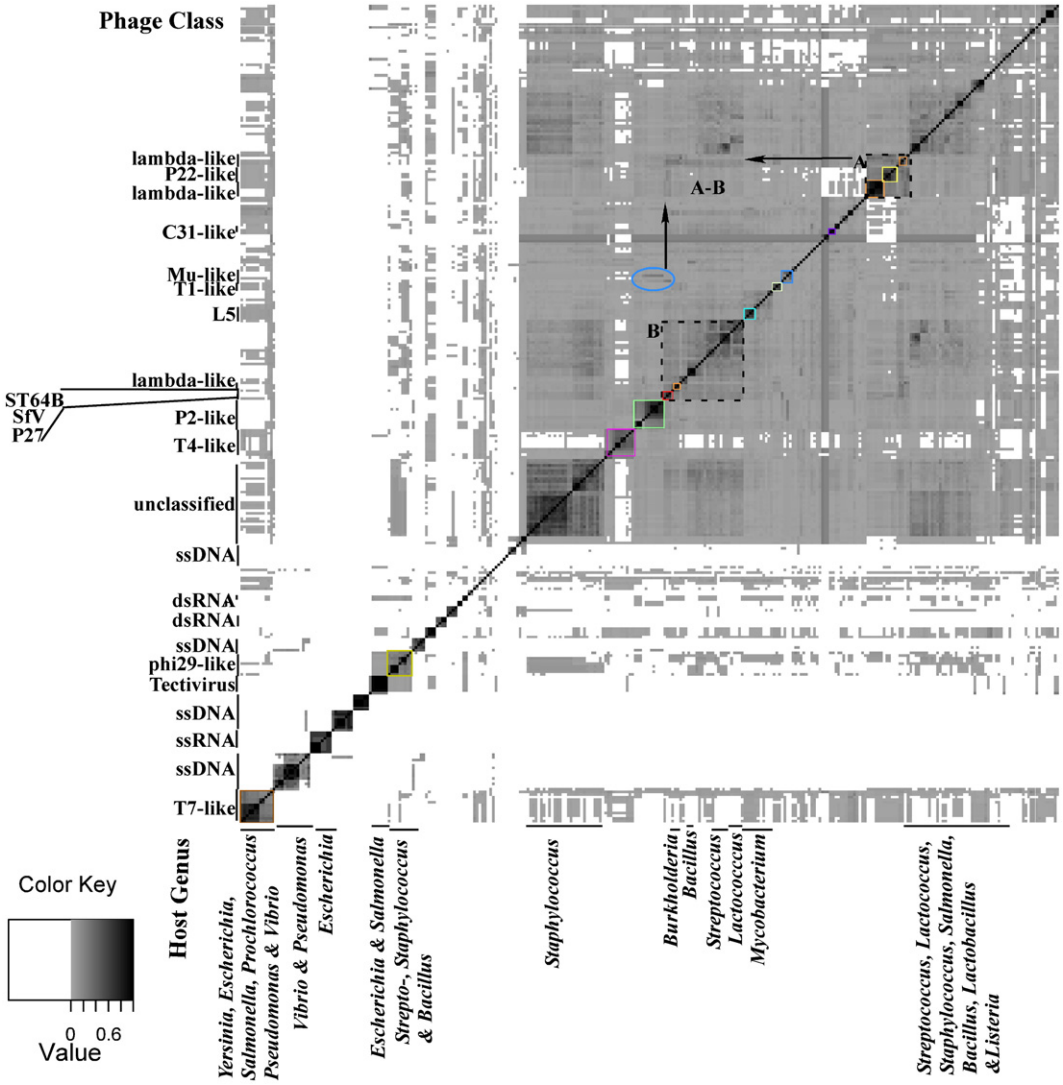

Fig. 4. Phage population landscape. The heat map represents the phage pairwise similarity matrix between 306 phages genomes based on protein content. The 306 genomes were downloaded form the NCBI in Feb 2006 and processed according to the ACLAME procedures described earlier (Leplae et al., 2004) to build protein families. On the bottom, the host genera are indicated for the darker clusters. On the left side, the main phage classes are labelled: ssDNA, ssRNA, dsRNA and the ICTV family and genus for the dsDNA phages. Groups of related phages can be recognized: T7-like (brown), phi-29 (olive green), T4-like (magenta), P2-like (green), L5-like (cyan), T1-like (beige), Mu-like (blue), phiC31 (violet), λ-like (orange), and P22-like (yellow). SfV, P27, ST64B and P2-like phages link Mu-like and lambdoid phages (blue circle). The arrows indicate the relationship between the different subgroups of lambdoid phages. The heat map was built with the 'heatmap' function of the R statistical package (http://www.r-project.org/).

to the T7 and T4 families form separated clusters as expected (brown and magenta squares respectively), but they are, in addition, related to other dsDNA phages.

Within the large dsDNA group, composed mainly by temperate phages, sub-groups such as the P2 (green square), P22 (yellow square), L5 (cyan square) and λ (orange squares) related phages can be recognized. Several areas contain phages infecting firmicutes, mostly lactic acid bacteria but also pathogenic streptococci and staphylococci, the latter being very much interconnected. The relatedness between a large set of those phages has been analyzed previously (Brussow et al., 2004).

The source of some widely discussed inconsistencies associated with the use of a classical taxonomy to represent a reticulate evolutionary scenario is clearly visible in the heat map. λ-like phages split among several clusters. One (dashed square A on Fig. 4) contains 3 subgroups with respectively, shiga-toxin-encoding phages (large orange square), *Podoviridae* P22, ST64T, ST104, HK620 and Sf6, all of which infect enterobacteria (yellow square) and the *Siphoviridae* λ, HK022 and HK097 (small orange square). The λ-related *Myoviridae* P27, ST64B and SfV (red square) are within a second interconnected group (dashed-line square B), and their connection with the groups in A can be seen as a darker zone at the A–B intersection (indicated with arrows). The four transposable phages, enterobacteriophage Mu, *Pseudomonas* phages D3112, B3 and *Burkholderia* phage BcepMu form a clear subgroup (blue square) with connections (blue circle) to the P27, ST64B and SfV group and P2-like phages.

It may be argued that groups of genes coding for interacting head and tail proteins, which tend to move together between genomes, tend to bias this type of global analysis. Such protein families can be filtered out to uncover the contribution of other families. This is illustrated in Fig. 5, where 147 temperate phages (identified through a literature search) have been analyzed in more details. In panel A, all protein families have been included. In panels B to F, the score matrices were calculated by selecting or discarding some protein families, based on their ACLAME annotation. Removing families of structural proteins (panel B) does not drastically change the overall picture. Consistent with this, panel E shows the significant contribution of those non-structural proteins, especially for phages in the top right group. Using only structural protein families (panel C), or excluding all non-structural protein families (panel D) loosens the connections. Panel F illustrates how the same type of analysis can be used for a defined set of protein families, here portal and terminase. The flexibility of the method opens the way to the definition of a bar code to characterize phages based on functional modules as proposed earlier (Lawrence et al., 2002).

Prophage detection

Prophages have been found in many sequenced bacterial genomes. The automatic prediction of prophages was not possible until very recently (Bose and Barber, 2006; Fouts, 2006). Based on the set of phage proteins in ACLAME, we developed a completely automated prophage prediction tool, called Prophinder (accessible at http://aclame.ulb.ac.be/Prophinder, Lima-Mendez et al. submitted for publication). Prophinder was run over 404 bacterial and archaeal genomes, generating around 550 prophage predictions distributed over 200 genomes. All predictions can be browsed on the web site, where new bacterial genomes can also be submitted for prophage prediction. The predicted prophage positions are displayed on a map of the host genome, along with prophages predicted by Phage-Finder and those manually annotated by Casjens (2003). Predicted prophages are also represented graphically. Each gene product matching a protein in ACLAME is tagged and linked with the corresponding protein family and function, providing an automated tool for the annotation of prophages. A matrix representation of all the predicted prophage proteins having

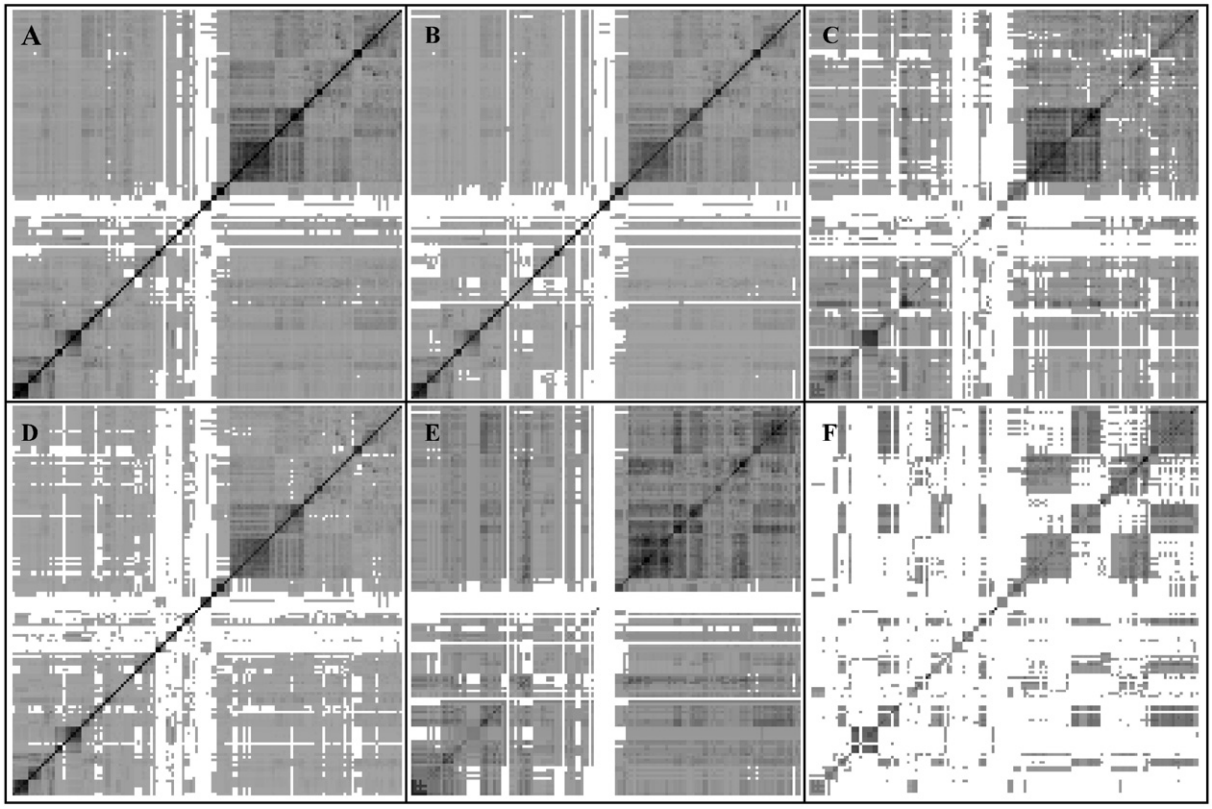

Fig. 5. Modularity of phages at higher resolution. Heat maps were built using one hundred and forty seven temperate phages (list available at http://aclame.ulb.ac.be/Classification/Phages/life_style.html) identified among the 306 phages represented in Fig. 4. To illustrate the contribution of major functional categories (see http://aclame.ulb.ac.be/perl/Aclame/show_cluster.cgi?mode=list&cat=vir) in phages similarities, different sets of protein families were used to calculate the similarity matrices. (A) All protein families were included; (B) structural protein families identified through their ACLAME annotation as *components* were excluded; (C) only families excluded in panel B were used; (D) non-structural protein families identified through their ACLAME annotation as *molecular function* were excluded; (E) only functions excluded in panel D were used; (F) only protein families annotated as terminase or portal protein were used. The pair-wise similarity score is calculated as the number of families in common from the selected set normalized to the total number of families in the shorter of the two phages. For all heat maps, the more intense colour corresponds to the highest similarity score of the corresponding similarity matrix (from score 0 to maximum value in the matrix). Differences between panels C and D and between panels B and E are mainly due to protein families with unknown or unassigned function.

orthologs in ACLAME phages can be obtained from the web interface. Such representation provides a direct visualization of the relationship between the predicted prophage and phage genomes in the database. Prophinder will automatically benefit from any update brought to the ACLAME phage genomes content and list of phage functions.

Conclusions and perspectives

The set of phage genomes readily available for global analysis is small and yet, their integration within a database such as ACLAME provides access to a higher level of biological information thanks to the procedure of classification into protein families and the various features organized in the database. The better conservation of head genes, in particular, terminases, portal and scaffold proteins, and of their organization on the phage genomes was largely documented for comparisons between smaller numbers of related phages (Desiere et al., 1999; Juhala et al., 2000; Pedulla et al., 2003). It could be easily verified here across a set of 140 dsDNA phages and will no doubt hold for a larger set. Many ACLAME protein families are readily seen to contain proteins encoded by phages that infect Gram− and Gram+ bacteria as well as Archaea, pointing towards the phage protein pool spreading across prokaryotic kingdoms. Other protein families appear specific to phages that infect a particular type of bacteria, such as the holin families, also as described earlier (Wang et al., 2000). Many other features become readily recognizable from the ACLAME classification. For instance, in the TerL families cluster:vir:48 and cluster:vir:152, the intein within the TerL protein of mycobacteriophages CJW1 and Omega (Pedulla et al., 2003) is readily visible in the results of the SCOP database searches. Surprisingly, in the intein database (http://www.neb.com/inteins/int_reg.html; Perler, 2002), the CJW1 and Omega inserts are listed as being in a gene coding for a DnaB ortholog. A total of 57 full size or truncated homing endonucleases belonging to 24 dsDNA phages out of 140 analyzed may fuel further discussion on the frequency of introns in phages (e.g. Edgell et al., 2000; Foley et al., 2000). A heat map of the gene-content-based similarity matrix provides a direct view of the modular and combinatorial nature of phage genomes with no loss of information as occurs in a phylogenetic tree, further supporting the notion that different phage types grade into one another.

Some phage properties would be trivial to incorporate into a database such as ACLAME are not necessarily easy to access. Using the links to the NCBI taxonomy, a direct count of *Myo*-, *Sipho*- and *Podoviridae* would be incomplete because many phages are unclassified. The temperate *vs*. virulent nature of a phage is another type of information not always available from the GenBank sequence files or taxonomy. Further large-scale analysis of the gene content of virulent vs. temperate phages may provide a clue into the type of features that could be used to automatically distinguish them.

In conjunction with a standard format for the deposition of phage genome sequences, the phage ontology developed for the functional annotation of the ACLAME protein families will allow for a more robust annotation of phage proteins and greatly facilitate further computer-based analysis. The contribution from the phage community has already been valuable and will continue to be essential to reach a consensus over a uniform nomenclature and a complete set of terms and definitions for phage functions.

Acknowledgments

We thank M. Salas, H. Krish, and E. Haggard-Ljungquist for their help in putting together more ‘user-friendly’ annotations of their favourite genome, S. Casjens and I. Molineux, G. Chaconas, A. Landy and G. Hatfull for their contribution to the assembly of phage terms and definitions and Jeanne Gouello for her contribution to the analysis of ORFan proteins. This work was supported by the Fonds de la Recherche Scientifique Médicale (FRSM) and by ESA (European Space Agency contract ESTEC 16370/02/NL/CK), and the Université Libre de Bruxelles within a collaboration with M. Mergeay and P. de Boever at the Laboratory for Microbiology, SCK-CEN, Mol Belgium. G.L-M is a fellow from the Fonds Xenophilia, ULB.

References

Abuladze, N.K., Gingery, M., Tsai, J., Eiserling, F.A., 1994. Tail length determination in bacteriophage T4. Virology 199 (2), 301–310.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 32, D226–D229 (Database issue).

Banks, J., Poole, S., Nair, S.P., Lewthwaite, J., Tabona, P., McNab, R., Wilson, M., Paul, A., Henderson, B., 2002. *Streptococcus sanguis* secretes CD14-binding proteins that stimulate cytokine synthesis: a clue to the pathogenesis of infective (bacterial) endocarditis? Microb. Pathog. 32 (3), 105–116.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R., 2004. The Pfam protein families database. Nucleic Acids Res. 32, D138–D141 (Database issue).

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2005. GenBank. Nucleic Acids Res. 33, D34–D38 (Database issue).

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O’Donovan, C., Phan, I., Pilbout, S., Schneider, M., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31 (1), 365–370.

Bose, M., Barber, R.D., 2006. Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. In Silico Biol. 6 (3), 0020.

Boyd, E.F., Brussow, H., 2002. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. Trends Microbiol. 10 (11), 521–529.

Brussow, H., Hendrix, R.W., 2002. Phage genomics: small is beautiful. Cell 108 (1), 13–16.

Brussow, H., Canchaya, C., Hardt, W.D., 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol. Mol. Biol. Rev. 68 (3), 560–602.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., Brussow, H., 2003. Prophage genomics. Microbiol. Mol. Biol. Rev. 67 (2), 238–276.

Casjens, S., 2003. Prophages and bacterial genomics: what have we learned so far? Mol. Microbiol. 49 (2), 277–300.

Casjens, S., Hatfull, G., Hendrix, R., 1992. Evolution of the dsDNA-tailed bacteriophage genomes. Semin. Virol. 3, 383–397.

Casjens, S.R., Gilcrease, E.B., Winn-Stapley, D.A., Schicklmaier, P., Schmieger, H., Pedulla, M.L., Ford, M.E., Houtz, J.M., Hatfull, G.F., Hendrix, R.W., 2005. The generalized transducing *Salmonella* bacteriophage ES18:

complete genome sequence and DNA packaging strategy. J. Bacteriol. 187 (3), 1091–1104.
Crutz-Le Coq, A.M., Cesselin, B., Commissaire, J., Anba, J., 2002. Sequence analysis of the lactococcal bacteriophage bIL170: insights into structural proteins and HNH endonucleases in dairy phages. Microbiology 148 (Pt. 4), 985–1001.
Desiere, F., Lucchini, S., Brussow, H., 1999. Comparative sequence analysis of the DNA packaging, head, and tail morphogenesis modules in the temperate cos-site *Streptococcus thermophilus* bacteriophage Sfi21. Virology 260 (2), 244–253.
Edgell, D.R., Belfort, M., Shub, D.A., 2000. Barriers to intron promiscuity in bacteria. J. Bacteriol. 182 (19), 5281–5289.
Edwards, R.A., Rohwer, F., 2005. Viral metagenomics. Nat. Rev., Microbiol. 3 (6), 504–510.
Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30 (7), 1575–1584.
Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., Ysebaert, M., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260 (5551), 500–507.
Foley, S., Bruttin, A., Brussow, H., 2000. Widespread distribution of a group I intron and its three deletion derivatives in the lysin gene of *Streptococcus thermophilus* bacteriophages. J. Virol. 74 (2), 611–618.
Fouts, D.E., 2006. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. 34 (20), 5839–5851.
Golz, S., Kemper, B., 1999. Association of holliday-structure resolving endonuclease VII with gp20 from the packaging machine of phage T4. J. Mol. Biol. 285 (3), 1131–1144.
Grundy, F.J., Howe, M.M., 1984. Involvement of the invertible G segment in bacteriophage mu tail fiber biosynthesis. Virology 134 (2), 296–317.
Haggard-Ljungquist, E., Jacobsen, E., Rishovd, S., Six, E.W., Nilssen, O., Sunshine, M.G., Lindqvist, B.H., Kim, K.J., Barreiro, V., Koonin, E.V., et al., 1995. Bacteriophage P2: genes involved in baseplate assembly. Virology 213 (1), 109–121.
Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258–D261 (Database issue).
Hatfull, G.F., Pedulla, M.L., Jacobs-Sera, D., Cichon, P.M., Foley, A., Ford, M.E., Gonda, R.M., Houtz, J.M., Hryckowian, A.J., Kelchner, V.A., Namburi, S., Pajcini, K.V., Popovich, M.G., Schleicher, D.T., Simanek, B.Z., Smith, A.L., Zdanowicz, G.M., Kumar, V., Peebles, C.L., Jacobs Jr., W.R., Lawrence, J.G., Hendrix, R.W., 2006. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet. 2 (6), e92.
Hendrix, R.W., 2003. Bacteriophage genomics. Curr. Opin. Microbiol. 6 (5), 506–511.
Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., Hatfull, G.F., 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc. Natl. Acad. Sci. U. S. A. 96 (5), 2192–2197.
Hendrix, R.W., Lawrence, J.G., Hatfull, G.F., Casjens, S., 2000. The origins and ongoing evolution of viruses. Trends Microbiol. 8 (11), 504–508.
Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., Hendrix, R.W., 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. 299 (1), 27–51.
Katsura, I., 1990. Mechanism of length determination in bacteriophage lambda tails. Adv. Biophys. 26, 1–18.
Lawrence, J.G., Hatfull, G.F., Hendrix, R.W., 2002. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J. Bacteriol. 184 (17), 4891–4905.
Leplae, R., Hebrant, A., Wodak, S.J., Toussaint, A., 2004. ACLAME: a CLAssification of Mobile genetic Elements. Nucleic Acids Res. 32, D45–D49 (Database issue).
Leplae, R., Lima-Mendez, G., Toussaint, A., 2006. A first global analysis of plasmid encoded proteins in the ACLAME database. FEMS Microbiol. Rev. 30 (6), 980–994.
Levin, B.R., Bull, J.J., 2004. Population and evolutionary dynamics of phage therapy. Nat. Rev., Microbiol. 2 (2), 166–173.
Liu, J., Glazko, G., Mushegian, A., 2006. Protein repertoire of double-stranded DNA bacteriophages. Virus Res. 117 (1), 68–80.
Nelson, D., 2004. Phage taxonomy: we agree to disagree. J. Bacteriol. 186 (21), 7029–7031.
Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., 1999. Use of contiguity on the chromosome to predict functional coupling. In Silico Biol. 1 (2), 93–108.
Pavesi, A., 2006. Origin and evolution of overlapping genes in the family *Microviridae*. J. Gen. Virol. 87 (Pt. 4), 1013–1017.
Pedersen, M., Ostergaard, S., Bresciani, J., Vogensen, F.K., 2000. Mutational analysis of two structural genes of the temperate lactococcal bacteriophage TP901-1 involved in tail length determination and baseplate assembly. Virology 276 (2), 315–328.
Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J.G., Jacobs Jr., W.R., Hendrix, R.W., Hatfull, G.F., 2003. Origins of highly mosaic mycobacteriophage genomes. Cell 113 (2), 171–182.
Perler, F.B., 2002. InBase: the Intein Database. Nucleic Acids Res. 30 (1), 383–384.
Rohwer, F., Edwards, R., 2002. The phage proteomic tree: a genome-based taxonomy for phage. J. Bacteriol. 184 (16), 4529–4535.
Serres, M.H., Riley, M., 2000. MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. Microb. Comp. Genomics 5 (4), 205–222.
Siew, N., Fischer, D., 2003. Twenty thousand ORFan microbial protein families for the biologist? Structure (Cambridge) 11 (1), 7–9.
Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22), 4673–4680.
Toussaint, A., Merlin, C., 2002. Mobile elements as a combination of functional modules. Plasmid 47 (1), 26–35.
Vegge, C.S., Brondsted, L., Neve, H., Mc Grath, S., van Sinderen, D., Vogensen, F.K., 2005. Structural characterization and assembly of the distal tail structure of the temperate lactococcal bacteriophage TP901-1. J. Bacteriol. 187 (12), 4187–4197.
Wang, I.N., Smith, D.L., Young, R., 2000. Holins: the protein clocks of bacteriophage infections. Annu. Rev. Microbiol. 54, 799–825.
Westmoreland, B.C., Szybalski, W., Ris, H., 1969. Mapping of deletions and substitutions in heteroduplex DNA molecules of bacteriophage lambda by electron microscopy. Science 163 (873), 1343–1348.
Wilhelm, S.W., Suttle, C.A., 1999. Viruses and nutrient cycles in the sea. BioScience 49 (10), 781–788.
Zajanckauskaite, A., Malys, N., Nivinskas, R., 1997. A rare type of overlapping genes in bacteriophage T4: gene *30.3′* is completely embedded within gene *30.3* by one position downstream. Gene 194 (2), 157–162.
Zhang, S., 2003. Fabrication of novel biomaterials through molecular self-assembly. Nat. Biotechnol. 21 (10), 1171–1178.