Reticulate Representation of Evolutionary and Functional Relationships between Phage Genomes

Gipsi Lima-Mendez, Jacques Van Helden, Ariane Toussaint, and Raphaël Leplae

Service de Bioinformatique des Génomes et des Réseaux (BiGRe), Université Libre de Bruxelles, Bruxelles, Belgium

Bacteriophage genomes show pervasive mosaicism, indicating the importance of horizontal gene exchange in their evolution. Phage genomes represent unique combinations of modules, each of them with a different phylogenetic history. The traditional classification, based on a variety of criteria such as nucleic acid type (single/double-stranded DNA/RNA), morphology, and host range, appeared inconsistent with sequence analyses. With the genomic era, an ever increasing number of sequenced phages cannot be classified, in part due to a lack of morphological information and in part to the intrinsic incapability of tree-based methods to efficiently deal with mosaicism. This problem led some virologists to call for a moratorium on the creation of additional taxa in the order *Caudovirales*, in order to let virologists discuss classification schemes that might better suit phage evolution. In this context, we propose a framework for a reticulate classification of phages based on gene content. Starting from gene families, we built a weighted graph, where nodes represent phages and edges represent phage–phage similarities in terms of shared genes. We then apply various measures of graph topology to analyze the resulting graph. Most double-stranded DNA phages are found in a single component. The values of the clustering coefficient and closeness distinguish temperate from virulent phages, whereas chimeric phages are characterized by a high betweenness coefficient. We apply a 2-step clustering method to this graph to generate a reticulate classification of phages: Each phage is associated with a membership vector, which quantitatively characterizes its membership to the set of clusters. Furthermore, we cluster genes based on their "phylogenetic profiles" to define "evolutionary cohesive modules." In virulent phages, evolutionary modules span several functional categories, whereas in temperate phages they correspond better to functional modules. Moreover, despite the fact that modules only cover a fraction of all phage genes, phage groups can be distinguished by their different combination of modules, serving the bases for a higher level reticulate classification. These 2 classification schemes provide an automatic and dynamic way of representing the relationships within the phage population and can be extended to include newly sequenced phage genomes, as well as other types of genetic elements.

Introduction

Bacteriophages (phages) are the most abundant biological entities in the biosphere (Chibani-Chennoufi, Bruttin, et al. 2004) and probably the largest source of gene diversity (Breitbart et al. 2002). Virulent phages infect the host bacterium and multiply and exit the cell via lysis or extrusion/budding, whereas temperate phages have the alternative of remaining within the host in a latent state. The current phage taxonomy, devised by the International Committee on Taxonomy of Viruses (ICTV), is based on genome type and phage tail and head/capsid morphologies. Tailed double-stranded DNA (dsDNA) phages form the largest group and are classified in the order *Caudovirales* and in 3 families: *Myoviridae* for phages with contractile tails, *Siphoviridae* for phages with long, noncontractile tails, and *Podoviridae* for phages with short tails (Fauquet et al. 2005).

Early pairwise genome comparisons by DNA–DNA heteroduplex analysis revealed that some dsDNA phages are genetic mosaics, where regions with sequence similarity alternate with unrelated regions in a patchwise fashion (Westmoreland et al. 1969). This observation was further substantiated once phage genome nucleotide sequences became available for comparative analysis. These showed that similar morphology does not imply genetic similarity or vice versa, raising serious debate on the validity of the ICTV taxonomy (Hendrix et al. 2000; Brussow and Hendrix 2002; Lawrence et al. 2002; Nelson 2004). Hendrix et al. (1999) proposed that all dsDNA phages access a common gene pool, although horizontal genetic exchange is more intense among phages with a similar host range. As the sequencing of phage genomes progressed, mosaicism appeared to be so pervasive that it became impossible to reach consensus within the Prokaryote Virus Subcommittee of the ICTV as to whether the traditional hierarchical classification should be replaced with a new system, and if so, what form that system may take (Mayo and Ball 2006).

Several proposals for virus classification have been made based on sequence information. Because phages lack a common locus, Rohwer and Edwards (2002) discarded classical phylogenetic analysis and built a tree based on pairwise dissimilarities between the complete bacteriophage proteomes. A second system classified phages based on the structural head module (Proux et al. 2002). Pride et al. (2006) proposed a phylogeny based on tetranucleotide usage patterns. However, being hierarchical, none of those approaches account for the combinatorial structure of the phage population, a hallmark of phage genome structure.

Lawrence et al. (2002) acknowledged that reticulate relationships cannot be modeled by classical taxonomies. They proposed a classification where, at a given taxonomic level, phages could belong to several groups defined as "modi." Phages within a modus would share a particular module or phenotypic character (e.g., the tail). One phage could belong to several modi (e.g., 1 for the head genes, 1 for the tail genes, and 1 for replication genes, etc.), reflecting their mosaic nature. This system of reticulate classification was illustrated with examples, but was not, and has not yet been implemented.

The new methodological framework presented here aims at filling this gap. We propose to represent relationships across the phage population as a weighted graph where nodes represent phages and edges represent

Key words: bacteriophage classification, phage evolution, network, graph, functional module, aclame.

E-mail: gipsi@scmbb.ulb.ac.be.

Mol. Biol. Evol. 25(4):762–777. 2008
doi:10.1093/molbev/msn023
Advance Access publication January 29, 2008

phage–phage similarities in terms of gene (protein) content. We studied the structure of the phage population using graph theory tools and defined overlapping groups of phages by clustering the graph. Genes with similar phylogenetic profiles grouped into putative functional modules, some of which correlate with the phage clusters. This approach provides an accurate description of the global phage population and is useful for exploring particular regions of the phage sequence landscape.

Methods

Genome Sequences

In this study, we used the set of $G = 306$ bacteriophage genomes downloaded from National Center for Biotechnology Information (NCBI) Web site in February 2006 (http://www.ncbi.nlm.nih.gov/genomes/static/phg.html).

This set consists of 250 dsDNA, 36 single-stranded (ss) DNA, 12 dsRNA, and 8 ssRNA phages. We had previously manually classified the dsDNA phages of this data set according to their temperate or virulent lifestyle (list available at http://aclame.ulb.ac.be/Classification/Phages/life_style.html) (Lima-Mendez et al. 2007). From 250 dsDNA phages, 72 could be classified as virulent, 156 as temperate, and 22 could not be classified.

Composition of Phages in Protein Families

We used BlastP (Altschul et al. 1997) to detect pairwise similarities between all the phage proteins. Protein families were derived from the whole set of pairwise similarities using Markov clustering (MCL) algorithm with the same parameters as in our previous work (Leplae et al. 2004). This procedure allowed us to classify 19,537 phage proteins into $n = 8{,}576$ families.

The composition of the phages in protein families was represented as a $306 \times 8{,}576$ matrix (M), where rows represent bacteriophage genomes and columns protein families, and the element $M_{i,j}$ equals 1 if phage i encodes at least 1 protein belonging to family j and 0 otherwise.

Network of Similarities between Phages

We built a graph representing the similarity relationships between phages derived from the number of shared protein families. Given 2 phages A and B, containing a and b protein families, respectively, we estimated the probability to observe at least c protein families in common using the hypergeometric formula:

$$\text{Pval} = P(X \geq c) = \sum_{i=c}^{\min(a,b)} \frac{C_a^i C_{n-a}^{b-i}}{C_n^b}, \qquad (1)$$

where n is the number of columns of matrix M, that is, the total number of protein families identified in all the phage genomes. The P value (Pval) is the probability of false positive for one comparison between 2 given phages, that is, the probability to consider the number of common families as significant whereas it results from chance.

In total, we performed all $T = (306 \times 305)/2 = 46{,}665$ pairwise comparisons between the 306 genomes. To estimate the expected number of false positives, we calculated the E value by multiplying the P value by the total number of comparisons. This E value is in turn converted into a significance index significance (sig) value by taking the minus logarithm in base 10:

$$\text{sig} = -\log(\text{Evalue}) = -\log(\text{Pvalue} \cdot T). \qquad (2)$$

Pairs of genomes having a value of sig above 1, that is, an E value <0.1, were considered as similar and joined by an edge with weight equal to the sig value.

We represented the phage population as an undirected weighted graph where nodes represent phages and edges the similarities between them. We refer to this graph as S277.

Permutation Test

As a negative control, we produced 100 random matrices by permuting the elements within the rows of M and calculated the similarity between the shuffled genomes.

Network Topology

A "component" is a maximal connected subgraph (Diestel 1997). Two nodes are in the same connected component if, and only if, a path exists between them. In a graphical view of a graph, different components are seen separated by empty space. For each node in the network, we calculated the "degree," the "clustering coefficient," the "closeness," and the "betweenness" as described below.

The degree of a node is defined as the number of links the node has (see [Barabasi and Oltvai 2004] for summary of graph theory measures).

The clustering coefficient CC_v of node v is the number of links between its neighbors divided by the number of links that could possibly exist between them. In a nondirected graph without self-loops, the number of possible links between N nodes is $N(N-1)/2$. Thus, for a node v having N neighbors (Watts and Strogatz 1998):

$$\text{CC}_v = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_{i,j}}{N(N-1)/2}, \qquad (3)$$

where $e_{i,j}$ is 1 if nodes i and j are connected and 0 otherwise.

The betweenness of node v measures the frequency at which it is found in the shortest paths connecting pairs of other nodes. It was defined by Freeman (1977) as:

$$Bv = \sum_{\substack{i=1 \\ i \neq v}}^{G-1} \sum_{\substack{j=i+1 \\ j \neq v}}^{G} \frac{ci,j(v)}{ci,j}, \qquad (4)$$

where G is the number of nodes in the graph, $c_{i,j}$ is the number of shortest paths from node i to node j, and $c_{i,j}(v)$ is the number of shortest paths from i to j running through node v.

Nodes with high betweenness hold different parts of the network together. We calculated the betweenness for

all nodes in the phage network (S277) to identify phages acting as bridges between unrelated phages.

The closeness of a node v is calculated as the inverse of the average of the distance to all other nodes (Sabidussi 1966):

$$\mathrm{Cl}_v = \frac{G-1}{\sum_{\substack{i=1 \\ i \neq v}}^{G} d_{v,i}}, \quad (5)$$

where $d_{v,i}$ represents the shortest path length between nodes v and i.

Extracting Clusters of Phages from the Network

We applied the MCL algorithm (van Dongen 2000) to partition the network into nonoverlapping clusters. The main parameter of MCL algorithm is called "inflation," which influences clusters granularity (number and size). We tested all inflation values between the 2 allowed extremes 1.2 and 5 by steps of 0.2 and analyzed each clustering result as described in the next section.

Evaluating Clustering Results

The clustering coefficient measures the "cliquishness" around a node; hence, its average over the nodes of a cluster can be used as a measure of the cluster homogeneity. We computed an "intracluster clustering coefficient" (ICCC) considering only the edges within clusters. In principle, a clustering result that maximizes the ICCC produces more homogeneous clusters. For each clustering result, we calculated the average ICCC over all nodes of the graph. We excluded clusters with less than 3 members and set to 0, the clustering coefficient for nodes with a single edge to avoid the trivial solution where the optimal cluster is the one with most nodes assigned to singleton or duet clusters. The optimal clustering of the graph was defined as the one with the highest ICCC. For comparison, we generated 1,000 random graphs according to the Erdos–Renyi (ER) model (Erdos and Renyi 1959), randomly assigned weights to their edges (using the S277 weight distribution), and used the same procedure as for S277.

Assigning Phages to Multiple Clusters

The "membership" of a node g to cluster c was calculated as the proportion of edge weights (the sum of the weight of all its edges) linking it to nodes of cluster c:

$$B_{g,c} = \frac{\sum_{i \in c} wg,i}{\sum_{p \in C} \sum_{j \in p} wg,j}, \quad (6)$$

where w is the weight of an edge, c is a given cluster, C is the set of clusters, and i and j are indexes for the nodes of clusters c and p, respectively.

The matrix $B = \{B_{g,c}\}$ contains the membership values for every node–cluster pair. We refer to the matrix B as the "membership matrix."

Comparing with the ICTV Classification

The phage taxonomy can be extracted from 2 resources, the ICTV database (ICTVdb) at http://www.ncbi.nlm.nih.gov/ICTVdb/, the portal of the official classification, and the NCBI taxonomy database at http://www.ncbi.nlm.nih.gov/Taxonomy/, which contains information for more phages. The NCBI taxonomy covers essentially all phages that have been sequenced at the cost of relying on the information provided by the authors submitting the sequences. On the other hand, experts are responsible for the information stored in the ICTVdb, which, as a result, has a limited coverage of sequenced phages (Fauquet and Fargette 2005).

For the reasons outlined above, we chose to retrieve the taxonomy from both databases. Out of the 277 phages in S277, 85 were classified in the ICTVdb. From the NCBI taxonomy database, the phage family could be retrieved for 265 and the genus for 144 phages. For the purpose of comparison and following the procedure described above, we built 3 new graphs, S85, S265, and S144, keeping only the nodes representing the phages classified in ICTVdb and NCBI taxonomy at family and genus level, respectively. Each of the 3 graphs was processed as described for S277; thus, we could directly map the current taxonomical classes (taken from ICTVdb or NCBI taxonomy database) into the clusters obtained in this study.

For the comparison between 2 classifications, we calculated the correspondence matrix $Q = \{Q_{c,k}\}$ between our clusters (indexed by c) and the ICTV clusters (indexed by k) as the matrix product between the transposed membership matrix $B^T = \{B_{c,g}\}$ and the g-by-k matrix describing the ICTV cluster membership for phages g into the k clusters $K = \{B_{g,k}\}$:

$$Q = \{Q_{c,k}\} = B^T \cdot K = \{B_{c,g}\} \cdot \{K_{g,k}\}. \quad (7)$$

We adapted the statistics defined by Brohee and van Helden (2006) to account for nonbinary membership and calculated the "recall" $R_{c,k}$ (known also as sensitivity or coverage) as the proportion of members of ICTV class k found in cluster c, relative to the total number of members of class k assigned to all clusters:

$$R_{c,k} = \frac{Q_{c,k}}{\sum_{i \in C} Q_{i,k}}. \quad (8)$$

The cluster-wise recall R_k is the maximal proportion of members from ICTV class k found in the same cluster:

$$R_k = \max_{i=1}^{c} R_{c,k}. \quad (9)$$

The clustering-wise recall is the weighted average of the cluster-wise recall over all clusters:

$$R = \frac{\sum_{i=1}^{K} G_k R_k}{\sum_{i=1}^{K} G_k}. \quad (10)$$

G_k is the number of phages classified in ICTV class k.

Analogously, we defined "precision" $P_{c,k}$ (known also as positive predictive value) as the proportion of members of cluster c belonging to the ICTV class k, relative to the total number of members of the cluster c assigned to any ICTV class:

$$P_c = \frac{Q_{c,k}}{\sum_{j \dot{I} K} Q_{c,j}}. \quad (11)$$

The "cluster-wise precision" P_c is the maximal proportion of members of cluster c found in the same ICTV class:

$$P_c = \max_{i=1}^{k} P_{c,i}. \quad (12)$$

The clustering-wise precision is the weighted average of the cluster-wise precision over all clusters:

$$P = \frac{\sum_{i=1}^{C} P_i \sum_{j=1}^{G} B_{j,i}}{\sum_{i=1}^{C} \sum_{j=1}^{G} B_{j,i}}, \quad (13)$$

where $B_{j,i}$ is the membership of phage j to cluster i calculated according to equation (6) and its sum over all phages corresponded to the size of the cluster i, which is not necessarily an integer. The term in the denominator equals the number of phages classified in all clusters.

Permutation Tests

As negative control, we permuted 1,000 times the row labels of the membership matrix, while keeping intact the membership matrix. We then followed the procedure described above (eqs. 7–13) to compare the permuted matrices with the ICTV classification.

Shared Gene Content

As supportive information in the comparison between the classifications, we calculated the overlap between the gene content of a pair of phages and related this information to the ICTV classes where the corresponding phages are assigned. The overlap was defined as the size of the intersection divided by the size of the union of the corresponding protein family sets, a measure known as Jaccard similarity (Gordon 1999):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \quad (14)$$

where A and B are the sets of protein families present in the phages A and B, respectively.

The median of the distribution of shared gene content between phages belonging to different classes provides a general measure of the relationships between distantly related phages.

Defining Evolutionarily Cohesive Modules

In order to identify evolutionarily related genes, we analyzed the phylogenetic profiles of all the gene families found in phages (Pellegrini et al. 1999). For each of the 1,486 gene families represented in at least 3 bacteriophage genomes, we constructed a binary vector of presence/absence among the bacteriophage genomes and compared all pairs of such profiles using the hypergeometric formula (eq. 1). The exclusion of families with 2 members was because, statistically, proteins present in only 2 genomes are unlikely to be part of a module. The P values were transformed into significance index (eq. 2). Similar profiles were defined as those having a sig value above a defined threshold. The highest the sig, the more similar the profiles are; we used 3 threshold values, 10, 5, and 1. Genes found in similar profiles were joined into a network with the edge weight set to the sig value and clustered using the MCL algorithm with the inflation value set at 5, which is the highest of the range generally used (http://micans.org/mcl/man/mcl.html#options).

The association of the modules to the clusters of phages obtained with the MCL algorithm was measured using the hypergeometric formula (eq. 1), where the overlap c between a cluster and a module is the number of phages found in a cluster that harbor a given module, a is the number of phages in the cluster, and b is the number of phages having the module. A phage with 70% of module occurrence (measured as protein families represented) was considered to have the corresponding module.

Computation

Graph topology and statistical analyses were performed using Regulatory Sequence Analysis Tools (van Helden 2003) (http://rsat.ulb.ac.be/rsat/), R statistical package (http://www.r-project.org/), and in-house perl scripts. Graph visualization was done using the Cytoscape software (http://www.cytoscape.org/) and its GenePro plugin (Vlasblom et al. 2006).

Results

Snapshot of the Bacteriophage Sequence Landscape

To build the network, we downloaded 306 completely sequenced phage genomes available in February 2006 at NCBI (http://www.ncbi.nlm.nih.gov/genomes/static/phg.html). We classified the phage proteins into families (supplementary table 1S, Supplementary Material online) as described for the A CLAssification of Mobile genetic Elements database (Leplae et al. 2004) (http://aclame.ulb.ac.be/). The number of protein families in common between phage genomes was used as the criterion for estimating pairwise similarity between each pair of phages. Using the hypergeometric formula, we measured the statistical significance of the similarity between all pairs of phages and kept the pairs passing the threshold of sig ≥ 1, (E value ≤ 0.1). As a negative control, we performed the same comparison between pairs of rows from 100 randomized matrices (see Methods). This produced only 1

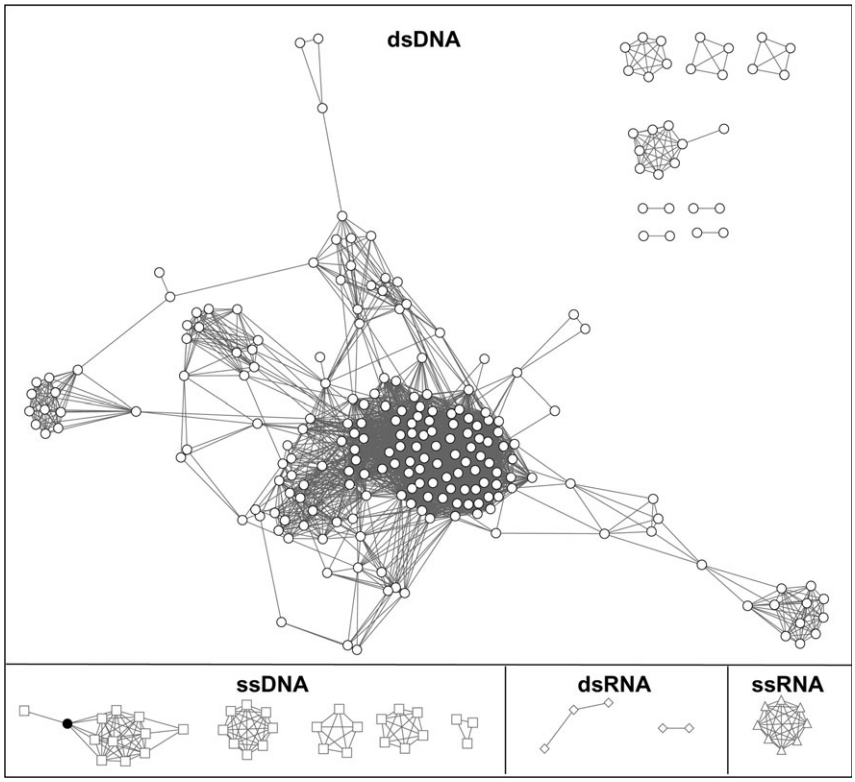


FIG. 1.—The phage population network (S277). Nodes represent phages and edges represent their statistically significant pairwise similarities as estimated with the hypergeometric formula. Phages cluster according to genome type. No reliable information on the genome type was obtained for one of the phages (in black).

pair with a sig value above this threshold, confirming the stringency of our significance threshold.

The network (fig. 1) consists of 277 nodes (phages) and 3,277 edges representing the significant relationships between phages. Twenty-nine phages were not similar to any other and were not further considered. Figure 1 shows the network (called S277 after the number of nodes). A force-directed layout algorithm (Fruchterman and Reingold 1991) places the nodes based on their connections, grouping the most highly connected regions of the graph. Inspection of the network reveals immediately the clustering structure of the global phage population according to the type of genetic material. Out of 17 network components (see Methods for definition), 9 consist of dsDNA phages, 5 of ssDNA phages, 2 of dsRNA phages, and 1 corresponds to the ssRNA phages.

The overlap between the gene pools of phages with different genome type is limited and hence not statistically significant. The distribution of dsDNA phages among the components is far from balanced. There are 8 small components containing from 2 to 9 phages. One hundred and 99 dsDNA phages are found in the largest component. Yet, within the largest component, some groups appear separated from the bulk. This mass of very interconnected phages observed in the largest component is mainly composed of temperate phages. Indeed, another feature in this network is that phylogenetically related phages are found in quasi-cliques; hence, no strict hubs are observed.

Graph Topological Measures Capture Relevant Properties of the Phage Population Structure

The clustering coefficient measures the extent to which the neighbors of a given node are interlinked. We used this coefficient as an indicator of cohesiveness around a phage's neighborhood. Virulent phages have higher clustering coefficients than temperate phages (median temperate = 0.72; median virulent = 1; P value $< 2.683 \times 10^{-07}$ Mann-Whitney U-test). To avoid a possible bias due to the overrepresentation of temperate phages in the network, we built 1,000 subnetworks composed of the 70 virulent phages and a random selection of 70 temperate phages. The clustering coefficient of virulent phages remains higher in those subnetworks (median P value over 1,000 comparisons = 4.364×10^{-05} Mann U-test). The differences between temperate and virulent clustering coefficient distributions may indicate that mosaicism is less crucial in the diversity of virulent phages than it is for temperate ones, as suggested from analysis of T4-like (Chibani-Chennoufi, Canchaya, et al. 2004; Filee et al. 2006) and dairy phages (Chopin et al. 2001).

The closeness of a node is the inverse of its average distance to all other nodes in the graph (see Methods). The higher the closeness, the more central is the node. We compared the distribution of closeness for temperate and virulent phages in the largest component of the graph and found that temperate phages are more central than virulent ones (median temperate = 0.41; median

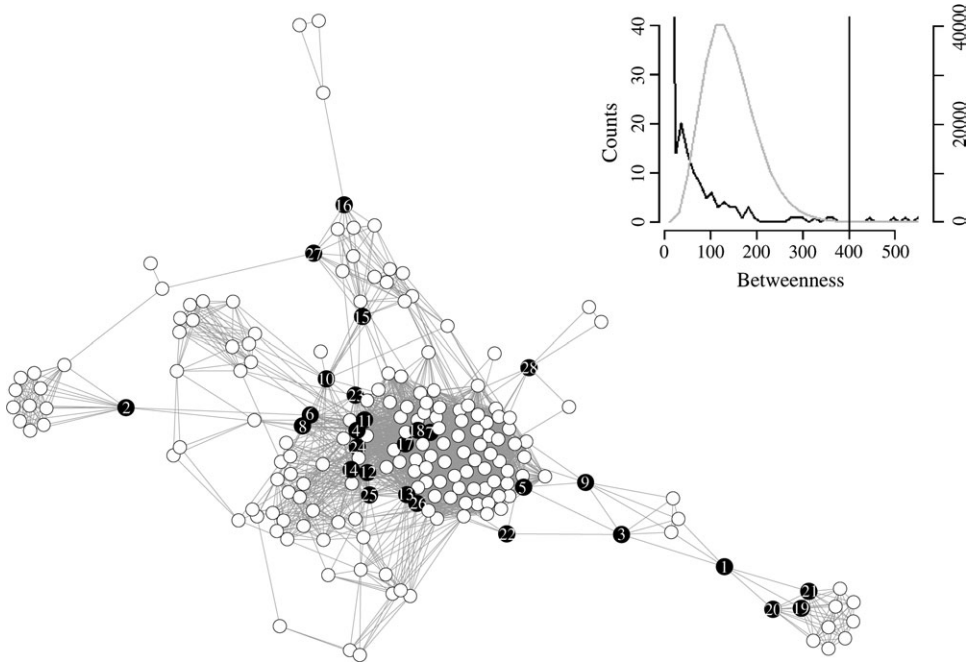


FIG. 2.—Detecting chimeric phages with the betweenness centrality. Main component of the phage network with the 28 phages having betweenness greater than 400 is in black along with their ranks. Values are listed in table 1. The inset shows the betweenness distribution of nodes in this component in black. The gray curve represents the betweenness distribution of 1,000 random graphs built with the same number of nodes and edges according to the ER model. The vertical line indicates the threshold value of 400, associated with a P value of 6×10^{-4}.

virulent = 0.27; P value $< 1.524 \times 10^{-12}$ Mann U-test). In this component, there are 53 virulent and 135 temperate phages. We also computed the closeness of the nodes of the main component of 1,000 balanced subnetworks built from 53 randomly chosen temperate and the 53 virulent phages of S277 main component, and the results confirmed that the observed difference between temperate and virulent phages is not due only to the overrepresentation of temperate phages (median P value over 1,000 comparisons = 1.66×10^{-7} Mann U-test). Temperate phages form essentially a single group in the center of the network; no path is observed between all virulent phages bypassing temperate ones. Thus, the closeness reflects the tendency of temperate phages to act as module donors/recipients between the virulent phages and the rest of the population (Lawrence et al. 2002).

The betweenness measures the extent to which a node is located in the shortest paths between all the other nodes. Nodes with high betweenness values are bridges between otherwise disconnected regions of the network. It is a centrality measure that gives an idea of which nodes are holding the network together. To identify phages combining genes from different phage groups, we calculated the betweenness for all nodes in the phage population graph and in 1,000 random graphs generated according to the ER model (Erdos and Renyi 1959). In the ER graph, every pair of nodes are joined with equal probability; thus, the network formed is homogeneous in nature and suitable as a negative control for estimating the significance of phages with high betweenness values, a feature linked to the network community structure (Girvan and Newman 2002). The P value was estimated as the frequency of node betweenness above a given threshold in the ER graphs. This P value represents the probability of having nodes with betweenness greater than the threshold by chance alone. The distribution of betweenness of the phage network is represented in figure 2 (inset), together with the distribution over the 1,000 ER graphs. The vertical line marks the betweenness value of 400, for which we obtained a P value of 6×10^{-4}. The 28 phages with betweenness above this threshold are ranked in decreasing order of betweenness values in table 1 and appear mapped onto the network depicted in the figure. Some of the phages with high betweenness values have been reported as founding members of novel groups. This is the case for T5—rank 1—(Wang et al. 2005), phiKMV—rank 2—(Ceyssens et al. 2006), and LP65—rank 9—(Chibani-Chennoufi, Dilmann, et al. 2004). Other phages have been described as genetic mosaics; ST64B—rank 4—(Mmolawa et al. 2003), XP10—rank 6—(Yuzenkova et al. 2003), P27—rank 11—(Recktenwald and Schmidt 2002), HK022 and HK97—rank 12 and 14, respectively—(Juhala et al. 2000), SfV—rank 24—(Allison et al. 2002), etc. The common theme among these phages is that they are shortcuts between otherwise unrelated phages. ST64B, SfV, and P27 bridge Mu-like phages with lambda-like phages. T5 bridges T4-like phages with lambda-like phages. PhiKMV and XP10 bridge T7-like phages with the rest of the *Siphoviridae*.

The Network Breakdown

Having seen that the network topology disclosed the relationships underlying the phage population, we

Table 1
Top of the List of Phages Sorted by Decreasing Value of Betweenness

Rank	Phage Accession	Phage Names	Betweenness
1	NC_005859	Bacteriophage T5	2057.00
2	NC_005045	Bacteriophage phiKMV	1969.72
3	NC_007610	Listeria bacteriophage P100	1892.66
4	NC_004313	*Salmonella typhimurium* phage ST64B	1655.41
5	NC_004820	Bacteriophage phBC6A51	1383.14
6	NC_004902	*Xanthomonas oryzae* bacteriophage Xp10	1347.87
7	NC_004112	*Lactobacillus casei* bacteriophage A2	1232.65
8	NC_007709	*Xanthomonas oryzae* phage OP1	1101.88
9	NC_006565	*Lactobacillus plantarum* bacteriophage LP65	1050.51
10	NC_004827	Bacteriophage Aaphi23	1024.02
11	NC_003356	Bacteriophage P27	702.98
12	NC_002166	Enterobacteria phage HK022	682.31
13	NC_004616	*Staphylococcus aureus* phage phi 12	679.63
14	NC_002167	Enterobacteria phage HK97	659.13
15	NC_004681	Mycobacteriophage CJW1	620.86
16	NC_002656	Mycobacteriophage Bxb1	586.72
17	NC_003524	Bacteriophage phi3626	578.48
18	NC_005893	Bacteriophage phi AT3	575.34
19	NC_005066	Enterobacteria phage RB49	501.33
20	NC_005083	Bacteriophage KVP40	501.33
21	NC_007023	Enterobacteria phage RB43	501.33
22	NC_001884	Bacteriophage SPBc2	492.76
23	NC_005069	Bacteriophage PY54	472.21
24	NC_003444	*Shigella flexneri* bacteriophage V	460.20
25	NC_006949	*Salmonella typhimurium* bacteriophage ES18	458.86
26	NC_007054	Bacteriophage 47	449.48
27	NC_001335	Mycobacterium phage L5	426.17
28	NC_002669	Bacteriophage bIL310	400.17

exploited the network representation to retrieve reticulate phage groups.

We first applied the MCL algorithm (van Dongen 2000) to the graph. The main parameter of this algorithm is the inflation factor that modulates cluster granularity. Figure 3*A* shows the number of clusters as a function of the inflation value for the real network (black) and for 1,000 random networks (gray) generated according to the ER model (Erdos and Renyi 1959) and randomly weighted with the weight distribution of the S277 network. To choose the optimal inflation factor, we explored values ranging from 1.2 to 5 by steps of 0.2 and estimated the clusters homogeneity by computing the average ICCC (see Methods). Figure 3*B* shows the evolution of the ICCC as the inflation value increases. The peak at inflation value of 2 suggests that this clustering solution is the best tradeoff between homogeneity and cluster size. Hence, we selected 2 as the inflation value, which produced 48 clusters.

Once the graph was partitioned, we reassessed the membership of the phages to the different clusters. From the S277 network, we calculated the total weight of the connections of a given node to the nodes within a particular cluster. The membership of the node to that cluster is equal to this value divided by the sum of the weight of all the connections of the node. Each phage was associated with a vector describing its membership to each cluster (see Methods). The membership matrix is available as supplementary table 2S (Supplementary Material online). In the network in figure 4, each node is a pie chart where each wedge represents the membership of the node to one cluster.

Comparison of the Clusters with the ICTV Taxonomy

To compare the result of our approach with the traditional phage classification, we measured the overlap between our phage clusters and the phage classes described in the ICTV taxonomy (Fauquet et al. 2005). From the ICTVdb (http://www.ncbi.nlm.nih.gov/ICTVdb/), we could retrieve the families and genera for 85 out of the 306 phages in our data set. From the NCBI taxonomy database, we obtained the families for 265 phages and the genera for 144. With these phages for which the classification was available, we built 3 new graphs S85, S265, and S144 (named after the number of nodes) for the comparison with the annotated classes (see Methods).

Classically, the comparison between 2 classifications is based on a contingency table, where each column represents a class of the first classification (e.g., ICTV taxonomy), each row a class of the second classification (e.g., our clustering result), and a cell indicates the number of elements at the intersection between the corresponding classes. In our reticulate classification though elements do not necessarily belong to a single class; instead, the relationship of each element to each class is estimated with a membership coefficient ranging from 0 to 1. Hence, we calculate the overlap between 1 ICTV class and 1 cluster as the sum of memberships of the phages of the cluster classified in that ICTV class. This overlap can thus not be strictly interpreted in terms of number of phages classified into the same cluster and ICTV class but as membership units.

Three statistics were used to evaluate the overlap between the classifications, recall, precision, and "accuracy (Acc)." The "class-wise recall" (R) represents the coverage

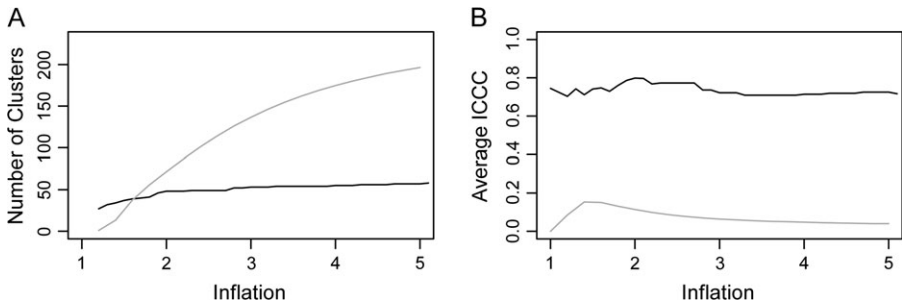


FIG. 3.—MCL of the phage graph. (*A*) The number of clusters is represented as a function of the inflation values. Results for the S277 population network are on the black curve. The gray curve represents the mean of the number of clusters obtained from the 1,000 ER random graphs. (*B*) Evolution of the ICCC as a function of the inflation value. The average clustering coefficients of the unclustered graphs are the first points of the curves. The black curve represents the ICCC values for the S277 population network and the gray curve for mean values obtained from the 1,000 random graphs. The peak at $I = 2$ indicates the inflation value that produces clusters with the highest homogeneity. To avoid trivial clustering solutions where all nodes are assigned to singletons or duets, we assigned a clustering coefficient equal to 0 to the 2- and single-member clusters. Singletons or duets are obtained in increasing amounts for higher inflation values provoking the observed decrease in clustering coefficient at higher inflation values.

of an ICTV class by its best-matching cluster; a value of 1 corresponds to the case where all phages of a given ICTV class are found in the same cluster. Reciprocally, the cluster-wise precision (*P*) measures how well a given cluster corresponds to its best-matching ICTV class; a value of 1 indicates that all phages in the cluster belong to the same ICTV class. From the class- and cluster-wise statistics, we computed the clustering-wise statistics as the weighted means over all classes/clusters of the class/cluster-wise values. The 2 measures are combined in the Acc, the arithmetic mean of *R* and *P* (see Methods for equations a detailed description of the matching statistics).

To estimate the random expectation, we performed 1,000 permutation tests by shuffling the labels of the rows of the membership matrix, that is, phage identifiers. This permutation test preserves the number of clusters and their respective sizes but regroups phages in a random way. These permuted clusters were then compared with the ICTV classes, and the same statistics (*R*, *P*, Acc) were computed as described above.

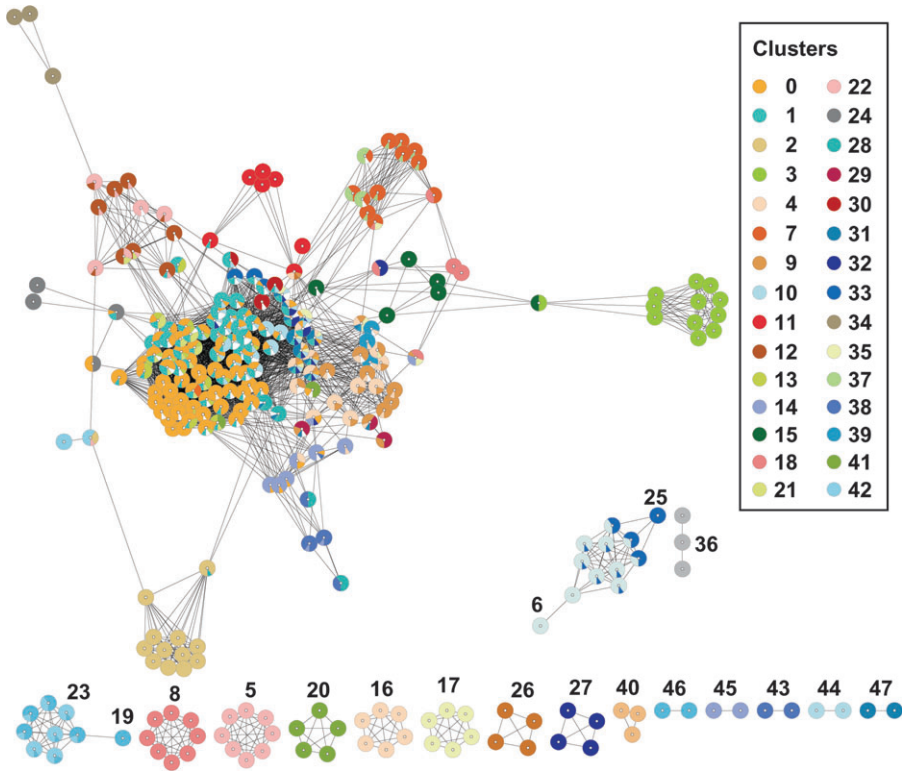


FIG. 4.—Mapping of the phage clusters onto the phage population network. Each node is depicted as a pie chart with the wedges representing the fraction of its edges belonging to the different clusters. A color code is used to differentiate each cluster as defined in the legend box. In the main component, cluster 3 contains T4-like phages, cluster 2 contains T7-like phages, cluster 7 contains P2-like phages, and clusters 34, 22, and 12 contain mycobacteriophages. Supplementary table 2S (Supplementary Material online) provides the membership matrix describing the occurrence of phages across the clusters.

Table 2
Comparison of Our Classification System with ICTV Phage Taxonomy

Classification	ICTV S85		ICTV S85		NCBI S265		NCBI S144	
Taxon Level	Family		Genus		Family		Genus	
Clusters	Real	Permuted	Real	Permuted	Real	Permuted	Real	Permuted
Clustering-wise recall	0.43	0.20 (0.02)	0.78	0.34 (0.02)	0.38	0.18 (0.01)	0.65	0.26 (0.01)
Clustering-wise precision	0.96	0.52 (0.02)	0.93	0.44 (0.02)	0.93	0.52 (0.01)	0.91	0.43 (0.01)
Acc	0.64	0.33 (0.02)	0.85	0.39 (0.02)	0.59	0.31 (0.01)	0.77	0.34 (0.01)

NOTE.—For permuted clusters, the values indicate the mean of 1,000 permutation tests, and the standard deviation is shown between parentheses.

The results are summarized in table 2. The clustering-wise recall was approximately 0.4 for families and 0.7 for genera, both values being twice as high as those obtained for the negative control. Such high values were never encountered in any of the 1,000 permutation tests. The *P* value can thus be roughly estimated as smaller than 10^{-3}, which indicates that the overlap between our reticulate classification and the official taxonomy is highly significant.

Clearly, our classification shows a better correspondence to genera than to families. This is not surprising because we mapped 13 families and 30 genera from the ICTV classification onto 48 clusters, the ICTV classes were inevitably split in our classification so that achieving a recall of 100% would have been impossible.

Table 2 also shows that the precision was approximately 0.92 for the genera and 0.95 for the families. This value indicates that on average 8% of significant evolutionary links join phages that belong to distinct ICTV genera and 5% to distinct ICTV families. The precision is about 2-fold higher for the data than for the negative controls (S144mean = 0.43 and $S85mean_{genera} = 0.46$), indicating that the high precision value is not an artifact of the cluster size distribution.

The main conflicts between the morphology-based and the gene content–based classifications concern the tailed phages. For each pair of tailed phages, we can measure the shared gene fraction as the shared protein families divided by the families found in either of the 2 phages (see eq. 14, Methods). Despite the shared gene content is higher for phages belonging in the same ICTV class (median intrafamily = 0.026; median interfamily = 0.009; median intragenus = 0.106; and median intergenera = 0.008); there are phages within an ICTV class sharing less genes than pairs of phages from different classes, whether families or genera (data not shown; http://aclame.ulb.ac.be/Classification/Phages/).

Note that all the phages of the data set are not assigned to ICTV classes. We can speculate that a good fraction of the phages that remain unclassified are those for which a reticulate system is more needed; thus, the real benefit of the reticulate classification will be higher than what we can estimate by analyzing the subset of phages found in the ICTV classification.

The Reticulate Evolutionary Relationships Are Apparent in the Proposed Classification

Further inspection of the membership to different clusters brings more insight into the added value of our approach. The subset of phages distributed mainly between clusters 0 and 1 defines a continuous gradient between these 2 clusters (fig. 5*A*). At one extreme, 27 *Staphylococcus aureus* phages belong mainly to cluster 0. At the other extreme, phages Sfi19, Sfi21, and DT1 from *Streptococcus thermophilus* belong mainly to cluster 1. Other phages infecting this host (7201, Sfi11, and O1205) and the *Lactobacillus* and *Streptococcus* phages scatter between the 2 clusters.

The well-documented mosaicism of lambda-related phages is easily visible. They are found in clusters 4 and 9 with memberships plotted in figure 5*B*. The shiga toxin–encoding *Siphoviridae* Stx1, Stx2 I and II, 933W, and VT2-Sa have $membership_{cluster9} > 0.8$ and $membership_{cluster4} < 0.2$, forming a subgroup. Cluster 4 appears rich in *Podoviridae* phages such as *Acyrthosiphon pisum* endosymbiont bacteriophage APSE-1, Sf6, ST104T, ST64T, P22, and HK620 with $membership_{cluster4} \sim 0.8$ and $membership_{cluster9} < 0.2$. Interestingly, phage lambda belongs almost equally to both clusters ($membership_{cluster9} = 0.38$ and $membership_{cluster4} = 0.46$). The chimera phages P27, SfV, and ST64B are classified with *Siphoviridae* in clusters 0 and 1 and with the *Myoviridae* phage Mu in cluster 32 (fig. 6*A*). Noteworthy, all 3 have a low membership to clusters 4 and 9, despite their reported lambda genetic structure (Allison et al. 2002; Recktenwald and Schmidt 2002; Mmolawa et al. 2003).

Phage N15, which bears features from lambda-related phages and from linear plasmids (Rybchin and Svarchevsky 1999), belongs to clusters 4, 9, and 35. Cluster 35 also harbors phiKO2 from *Klebsiella oxytoca* and PY54 from *Yersinia enterocolitica*, both known to lysogenize as linear plasmids (Hertwig et al. 2003; Casjens et al. 2004). All 3 have the protelomerase and plasmid partitioning proteins required for this type of lysogeny. Other phages with membership to this cluster are *Vibrio harveyi* bacteriophage *Vibrio harveyi* myovirus-like (VHML) and *Burkholderia cepacia* phage BcepNazgul. Phage BcepNazgul encodes a plasmid partitioning system, but no information is available on its lifestyle. VHML allegedly integrates in the host chromosome via transposition (Oakey et al. 2002); however, no transposase has been found encoded in the genome. VHML does encode a protelomerase and a plasmid-partitioning protein; thus, it is likely that it establishes as a linear plasmid prophage. VHML is another example of a phage challenging the traditional taxonomy that can be classified in our system without apparent contradiction (fig. 6*B*).

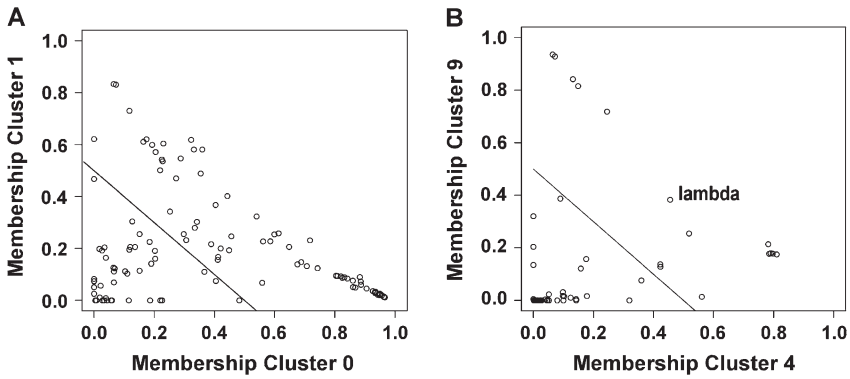
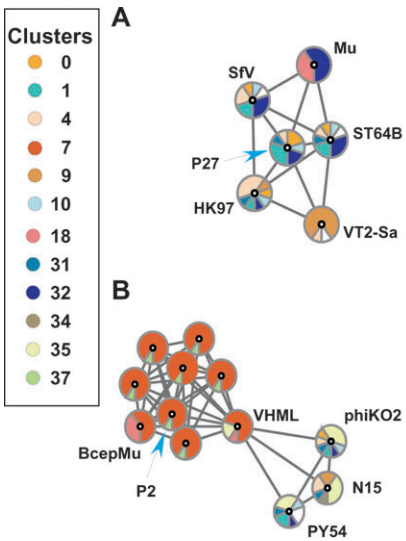


FIG. 5.—Intertwined relationships between clusters 0 and 1 (*A*) and clusters 4 and 9 (*B*). The membership of the phages to clusters 0 and 1 ranges from 0 to 1 and vice versa, almost in a linear fashion at least for phages located above the line, which marks the limits between the phages that mainly belong to these 2 clusters (above) and the phages that belong to other clusters in greater proportion (below). A similar situation holds for phages belonging to clusters 4 and 9, most of which are known as "lambdoid" phages. Ironically, lambda with a membership of 0.46 to cluster 4 and 0.38 to cluster 9 appears as a mosaic between the phages that define those clusters and not as a phage defining a family.

Evolutionary Cohesive Modules

The modular theory of bacteriophage evolution states that the product of phage evolution is a family of interchangeable genetic elements called modules carrying a biological function (Susskind and Botstein 1978; Botstein 1980). Thus, phage genomes are often dissected—not automatically—into modules corresponding to the functional categories: replication, lysis/lysogeny, DNA packaging, and head and tail morphogenesis, etc (Brussow and Desiere 2001). Phylogenetic profiles have been used to predict functional links between genes on the assumption that genes involved in the same function and thus required together co-occur in genomes (Pellegrini et al. 1999). We derived the phylogenetic profiles (see Methods) for each phage gene and identified modules as clusters of proteins with similar phylogenetic profiles (see Methods). We observed that the evolutionary modularity of phage is rather limited. Using strict parameters (sig $\geq$ 10), we discovered 26 modules covering 144 protein families containing 2–27 proteins. Decreasing the sig value threshold, we detected more modules; sig $\geq$5 produced 62 modules spanning 385 protein families and sig $\geq$1 allowed identifying 93 modules covering 754 protein families. Throughout this work, we use the term "module" to refer to those defined with sig threshold of 10. Modules defined at lower sig



Membership matrix

	Clusters														
Phages	0	1	4	7	9	10	18	28	29	31	32	33	34	35	37
Mu	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.00	0.00	0.00	0.58	0.00	0.00	0.00	0.00
SfV	0.12	0.20	0.18	0.00	0.02	0.08	0.03	0.00	0.00	0.05	0.29	0.01	0.00	0.03	0.00
P27	0.15	0.26	0.05	0.02	0.02	0.12	0.03	0.00	0.00	0.06	0.24	0.02	0.00	0.03	0.00
ST64B	0.12	0.21	0.10	0.00	0.03	0.10	0.04	0.01	0.00	0.06	0.27	0.01	0.00	0.04	0.01
HK97	0.07	0.07	0.42	0.00	0.14	0.06	0.00	0.01	0.00	0.09	0.07	0.00	0.03	0.04	0.00
VT2-Sa	0.00	0.00	0.15	0.00	0.82	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.00

Membership matrix

	Clusters														
Phages	0	1	4	7	9	10	14	18	22	28	31	32	34	35	37
VHML	0.00	0.00	0.00	0.78	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.17	0.00
phiKO2	0.07	0.12	0.10	0.02	0.02	0.07	0.00	0.00	0.00	0.02	0.13	0.08	0.01	0.37	0.00
PY54	0.03	0.20	0.05	0.03	0.00	0.09	0.00	0.00	0.00	0.03	0.12	0.09	0.00	0.34	0.00
N15	0.00	0.00	0.18	0.05	0.16	0.00	0.01	0.01	0.00	0.00	0.09	0.00	0.08	0.43	0.00
BcepMu	0.00	0.00	0.00	0.62	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P2	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.08

FIG. 6.—Reticulate relationships become apparent in the proposed classification. The left part of each panel illustrates the clusters mapped into the nodes (as in fig. 4). Only clusters to which the phage has membership greater than 0.1 are indicated. White wedges represent the overall membership to all clusters that individually do not reach that threshold. The right part of each panel is a section of the membership matrix relevant for the represented phages. Shaded boxes indicate the clusters where the phages are assigned by the MCL algorithm (partition step). (*A*) Phages SfV, ST64B, and P27 are similar to lambdoid phages (*Siphoviridae* family) and to Mu (*Myoviridae* family). In ICTVdb (http://www.ncbi.nlm.nih.gov/ICTVdb/), SfV and P27 are classified as *Myoviridae*. ST64B is classified only in the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/) as *Podoviridae*. Notice that phages HK97, VT2-Sa, and Mu are assigned in the multiple assignation step to cluster 32, where SfV, P27, and ST64B are assigned in the partition step. (*B*) Phage VHML is classified as a *Myoviridae* P2-like phage in the NCBI taxonomy database. This phage has also relevant links with phages PY54, N15, and phiKO2 through proteins involved in lysogeny. In this study, these relationships are transparent because phage VHML is classified in cluster 7 with P2-like phages and in cluster 35 with PY54, N15, and phiKO2.

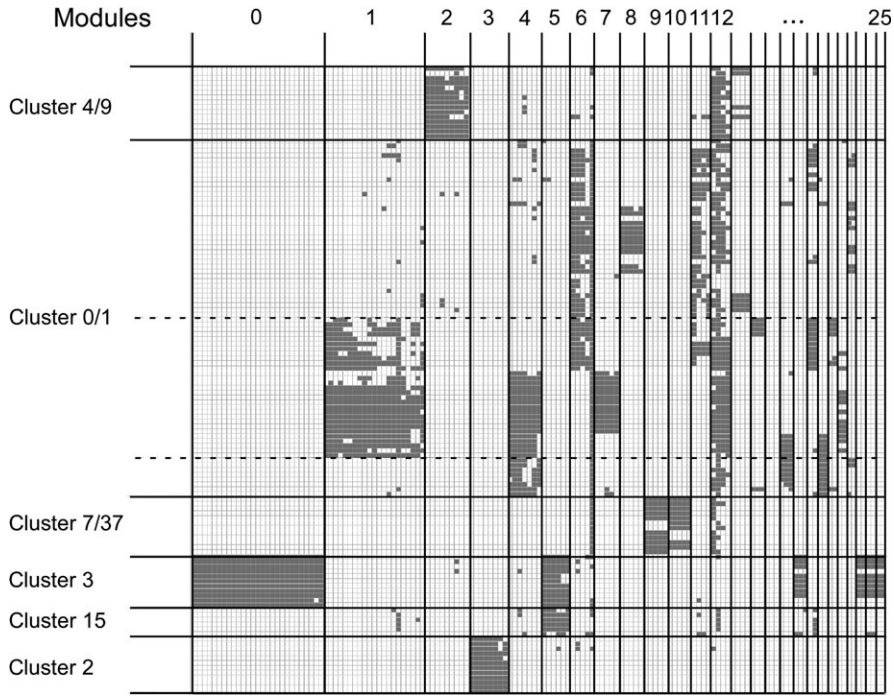


FIG. 7.—Module profile across phages. This is the section of the complete matrix (available as supplementary fig. 1S, Supplementary Material online). Each row represents a phage and each column represents a protein family that is part of a module. Phages are grouped by the clusters they are members of. The 2 horizontal dashed lines delimit the subgroup of phages that have proteins grouped in module 1, all these phages infect *Staphylococcus aureus*. Module 0 is the hallmark of T4 phages because is present in all members of cluster 3 and not elsewhere. Module 5, which probably participates in DNA replication, recombination, and repair, is also found in all phages of the T4 group, but it is shared with phages of cluster 15. Modules 4 and 6 both have large and small terminase subunits. Modules 7 and 11 contain phage morphogenesis proteins. Module 6 combines with module 11 and 4 with 7. Module 13, containing tail proteins and a prohead protein, links some lambdoid phages with phages infecting Gram(+) bacteria. Note that the group of lambdoid phages has conserved only modules involved in transcription regulation and lysis/lysogeny decision: module 12 with integrase, CI and Cro repressors, and module 2 with CII, CIII, N transcription antitermination, and Nin proteins.

thresholds are appended with the corresponding sig values. All modules are available as supplementary table 3S (Supplementary Material online). Replication and recombination, head and tail structural and assembly functions are prominent components of the modules beside numerous protein families with unknown function.

Figure 7 displays the profile of the modules across the main phage clusters (the complete matrix of occurrences of modules in phages is represented in supplementary figure 1S (Supplementary Material online). Important differences exist between temperate and virulent phages. Virulent phages tend to have evolutionary modules that span over different functional categories. For example, module 0 contains head, tail, and DNA replication proteins of T4-like phages. Likewise, the T7 group of phage features module 3, which is composed of tail proteins, head protein, terminase, portal and RNA polymerase (see supplementary table 3S [Supplementary Material online] for the functional annotation).

Modules in temperate phages better corresponded to functional modules. Most temperate phages exhibit module 12, consisting in the integrase, transcriptional repressors (CI and Cro are grouped in the same protein family), a DNA-binding protein and the replication initiation protein. Allegedly lambdoid phages (clusters 4 and 9, fig. 7) have only one additional module: module 2. Module 2 is composed of the transcription antitermination protein N, proteins CII and CIII, which control lysogenization and several Nin and the Rz proteins, the function of which has not been elucidated yet.

In contrast to the lambdoid phages, the group of temperate phages infecting Gram(+) bacteria displays several modules in a combinatory fashion (clusters 0 and 1 in fig. 7). Interestingly, only phages from *S. aureus* have module 1. We do not have clues into the function of this module, but it displays a fragmentary profile across the phages, suggesting that it comprises more than 1 functional module. Because many pathogenicity determinants are phage encoded (Brussow et al. 2004; Ogura et al. 2007), it is important to further investigate the function of these proteins. This group (cluster 0/1) has 2 modules with DNA-packaging functions, modules 4 and 6, and 2 modules with head morphogenesis proteins, modules 7 and 11. Only the combinations module 4/module 7 and module 6/module 11 are observed. Noteworthy, within the subgroup of phages that have module 1 both combinations are found.

Modules 9 and 10 are present in phages of the *Myoviridae* family and correspond to proteins participating in head and tail morphogenesis, respectively (clusters 7 and 37, fig. 7). Module 9, composed of head proteins, is present in all P2-like phages except VHML. Module 10, composed of tail proteins, is strongly linked to but not exclusive of P2-like phages. That module is found in the Mu-like BcepMu but is absent from the P2-like phages HP1, HP2, and K139.

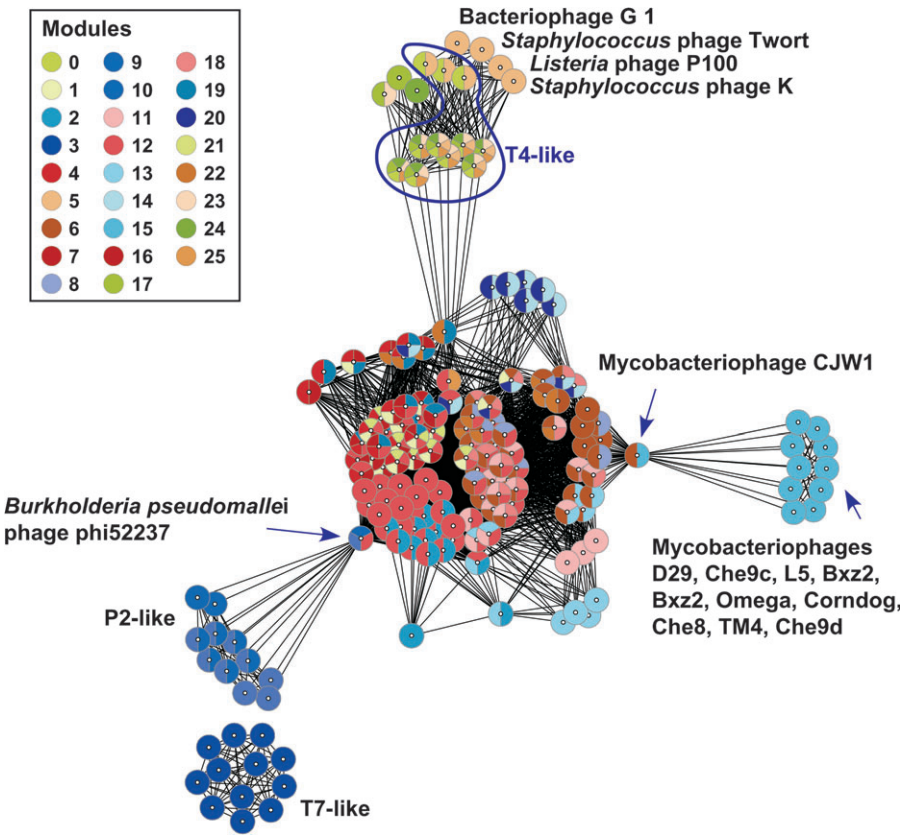


FIG. 8.—Module-based phage network. The nodes represent phages and the edges are drawn when the corresponding phages share at least 1 module. The node wedges represent the different modules present in the corresponding phage. The phage groups that separate from the main bulk are indicated.

BcepMu and VHML are clearly chimeric phages, featuring DNA packaging and capsid proteins related to Mu and lambda, respectively.

A Module-Based Network?

One interesting alternative in defining the phage network is to establish the links between phages based on their modules (fig. 8). We computed the "intersection network" from the edges in common between the 2 networks. The intersection network has 76% of the edges of the protein-based network (S277). At first glance, there are 2 important differences with the S277 network. First, T7-like phages are not connected to the main component in the module-based network. Second, in S277 T4 phages are connected to the bulk of the network through a path containing T5, *Staphylococcus* phage Twort, *Listeria* phage P100, *Staphylococcus* phage K, phage G1, and *Lactobacillus plantarum* phage LP65. In the module-based network, T4-like phages are in the path between those phages and the network bulk, thus inverting the local topology. The explanation for this inversion is straightforward. The small module of 2 genes (module 25) that links T4 phages to the bulk in the module-based network is not statistically significant when building the S277 network. Reciprocally, the genes shared by phages P100 and LP65 with phages in the network bulk do not form a module but their quantities are significant when building the S277 network.

Discussion

Pairwise comparisons between phage genomes have provided a high-resolution picture of phage relationships and revealed that mosaicism is the hallmark of phage genome structure (Hendrix 2003). However, these analyses had not been scaled up at the level of the whole phage population and the information was incomplete. Previous evolutionary representations neglected the mosaicism by making a choice of hierarchical trees based either on whole genomes (Rohwer and Edwards 2002) or on capsid proteins (Proux et al. 2002). Unlike previous approaches developed to classify phages, we count all statistically significant relationships to build a network that places all phages in the right scenario given the current genomic data. Moreover, our methodology does not rely on any knowledge-based training (e.g., supervised classification); hence, the results are not biased by predefined criteria.

Consistent with the consensus (Lawrence et al. 2002; Fauquet et al. 2005) that phages with different genome types—ssDNA, dsDNA, ssRNA, and dsRNA—are essentially different, they are found in separate components in the network. Graph theoretical measures capture relevant properties of the phage population and are suitable to analyze phages in their genetic neighborhood. The near-one clustering coefficient of virulent phages is consistent with the conservation of a core genome among virulent phages of the same group (Scholl et al. 2004; Filee et al. 2006). Virulent phages appear at the periphery of the phage sequence space

centered on temperate phages. This reflects that module exchange occurs within the host, where temperate phages may reside for longer periods and function as "banks" for modules/genes exchange across the whole phage population (Lawrence et al. 2002). Chimeric variants bridging temperate and virulent phages (Yuzenkova et al. 2003; Wang et al. 2005; Ceyssens et al. 2006) have high betweenness. Such values possibly arise as an artifact of the sparse sampling of the sequence space. Yet, the betweenness could prove useful to fill in those gaps with dedicated sequencing projects.

The automatic classification system proposed here involves the definition of clusters with similar phages followed by the reassignment to multiple clusters in order to generate a reticulate classification of the phages. In order to identify clusters of highly connected phages (quasi-cliques), we selected parameters that maximized the ICCC. Thus, phages within the same MCL cluster are likely descendant from a unique module combination. The weight of the intracluster connections represents 79% of the total weight of the connections of the network. This number can be taken as a rough estimate of the contribution of vertical evolution in this network. However, phages from different MCL clusters may be also related through vertical evolution but they might have diverged so much that sequence similarities are no longer recognizable or only some modules may have been vertically inherited, whereas others have been replaced through horizontal gene transfer.

The number of MCL clusters is larger than the number of ICTV genera (Fauquet et al. 2005). From the operational point of view, producing many clusters allowed more combinations to classify the mosaic genomes in the multiple assignment step. Given the importance of recombination in phage evolution, it is reasonable to observe an inflation of the number of phage groups, when considering that separate clusters can correspond to distinct combinations of gene modules. The overlap between the clusters of our reticulate classification thus enables to reflect the traces of the recombination events.

The distribution of the phages among the clusters is described with the membership matrix, where row vectors correspond to the membership of the individual phages. It can be argued that wrong relationships may be inferred from the membership matrix. For example, phage HK97 and phage Mu have no direct relationship although, in the multiple assignments step, they are assigned to cluster 32, due to their links to phages ST64B, P27, and SfV (fig. 6*A*). However, this is not a real problem because keeping track of the original MCL assignment ensures that a false direct link between Mu and HK97 would be dismissed. On the other hand, the inclusion of both phages, Mu and HK97, within the same cluster reflects that ancestors of these phages—and probably these phages too—shared the same gene pool as previously suggested (Hendrix and Casjens 2005).

We derived a second phage network, based on modules shared between phages. Contrary to the S277 network in which the identity of the shared protein families is not encoded, this module-based network allows for a detailed exploration of the nature of the functions shared between the connected phages. This higher functionality of the module-based network runs at the expense of links corresponding to genes that are not part of modules, which do contribute to the S277 network. Noteworthy, if the protein families part of modules are filtered out of the S277 network (significant links due only to the genes part of modules are lost), the main component of the network still harbors temperate phages from Gram(+) and Gram(−) bacteria (data not shown), indicating that those are not linked because of the integrase and repressors alone.

Table 3
Comparison of Reticulate Taxonomic Proposal (Lawrence et al. 2002) with the Evolutionary Modules (this study)

Phage	Modi (Lawrence et al. 2002)	Modules
N15	Lambda-like head genes	Module 73_sig1
	Lambda-like tail	Module 13
	Linear episome–mediated temperate phage	Module 56_sig1/ module 87_sig1
Lambda	Lambda-like head genes	Module 73_sig1
	Lambda-like tail genes	Module 13
	Integrase-mediated temperate phage	Module 12
HK97	HK97-like head genes	Module 6/module11
	Lambda-like tail genes	Module 13
	Integrase-mediated temperate phage	Module 12
SfV	HK97-like head genes	Module 6/module11
	Mu-like tail genes	Module 22_sig1
	Integrase-mediated temperate phage	Module 12
Mu	Mu-like head genes	Module 41_sig1
	Mu-like tail genes	Module 22_sig1
	Transposase-mediated temperate phage	Not found
M13	M13-like structural genes	Module 40_sig1
	M13-like replication genes	Not found
	M13-like maturation genes	Module 40_sig1
I2-2	M13-like structural genes	Module 40_sig1
	I2-2-like replication genes	Not found
	M13-like maturation genes	Module 40_sig1
PhiX174	External scaffolding protein	Module 9_sig5
	Lysis via MraY protein	Module 9_sig5

NOTE.—The term "module" refers to the modules defined with sig = 10 as threshold unless otherwise specified. A suffix (_sig) indicates the threshold when the matching module was obtained either with sig = 5 or with sig = 1 as threshold values.

The taxonomy outlined by Lawrence et al. (2002) provided a few selected examples of phages decomposed in reticulate groups—called modi. The correspondence of those modi with the modules detected here is depicted in table 3, where phages are described by a combination of modules according to each system. The HK97 head modus is split between the packaging module (module 6) and the head structural proteins (module 11), in agreement with reports about cellular organisms, where evolutionary modules recover only fragments of functional modules (Glazko and Mushegian 2004; Snel and Huynen 2004). Nonorthologous gene displacement may occur (Koonin et al. 1996), disrupting the modularity. This is the case for the 2 protease families from modules 6 and 11.

The modus named as "linear episome–mediated temperate phage," which describes the prophage status of bacteriophage N15, corresponds to modules 56_sig1 and 87_sig1 that, respectively, contain the protelomerase and the plasmid-related replication and partitioning proteins, both detected using sig threshold of 1.

The 2 modi present in phage PhiX174 are each composed of a single protein, the external scaffolding and the inhibitor of the MraY protein, respectively. The method of phylogenetic profiles (Pellegrini et al. 1999) cannot detect functional modules composed of a single protein. We detected these 2 proteins in a single module defined with sig $\geq$5 (module 9_sig5) and harboring 8 out of the 11 proteins of the phage. Likely, this situation reflects a low level of recombination and divergence among PhiX174 and its relatives so that co-occurring genes correspond to more than single functional module. The same conclusion holds for modules 0 and 3, present in T4- and T7-like phages, respectively, and possibly to module 1, present in a set of 21 *Staphylococcus* phages.

Despite the limited evolutionary modularity, some modules behave as signature of phage groups. Moreover, the module sharing also allows defining a network that is easier to interpret and may hold the most relevant biological information. The existence of genes that remain tightly linked across several genomes in spite of the pervasive recombination suggests constraints against module disruption. The constraints are not equal for all phages. Lambdoid phages have retained a module comprising proteins involved in transcription regulation, whereas they display poor modularity in phage morphogenesis and DNA packaging. On the other hand, virulent phages have conserved many genes as a group, implying constraints to evolutionary successful new gene combinations. These results argue against the validity of a single module to classify all phages because module conservation is strongly dependent on the phage biology and lifestyle.

The acquisition of more genomic data would—no doubt—reshape the clusters and modules as new evolutionary links will be unveiled. Questions regarding the structure of the phage population, in particular whether there is a defined boundary between temperate and virulent phages, may be answered in the future with this kind of systematic analysis. The global view depicted in this work suggests that the phages acting as bridges between less related phage groups may be the sole representatives of phage families located in the path between such groups. When more members of the family are known, the picture could converge toward the kind of relationship observed between clusters 0 and 1 and clusters 4 and 9 (fig. 5).

Conclusions

Our work addresses the long-standing claim for a classification capable of assigning phages to multiple groups featured by a series of marker modules. We proposed here 2 strategies toward classification. Both achieve the reticulate classification by describing the phages as vectors; cluster membership in one case and module occurrence in the second. The 2 approaches can be combined to further explore the evolutionary links trying to discriminate the contribution of vertical and horizontal evolution.

Graph theory measures captured the genetic differences between phages due to their lifestyle. These measures could be further exploited as predictors of phage lifestyle. The automatic detection of chimeric phages may prove useful in the ongoing assessment of phage diversity.

These methodologies could be extended to the classification of other mosaic genetic entities existing in prokaryotes such as plasmids, conjugative transposons, and other genomic islands. It may also provide a way to reinvestigate the extent to which eukaryotic dsDNA viruses undergo recombination (Iyer et al. 2006), a matter of crucial importance to assess the safety of defective viruses as vectors for gene therapy or the development of live vaccines.

Supplementary Material

Supplementary figure 1S and tables 1S–3S are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to I. Molineux, M. Chandler, and A. Garcia Pino for critical reading of the manuscript. Our work is supported by European Space Agency–PROgramme de Développement d'EXperiences scientifiques (contract C90254), the Fonds de la Recherche Fondamentale Collective, the Belgian Government concerning priority actions for basic research, and the Université Libre de Bruxelles (ULB). G.L.-M. was supported by the Fonds Xenophilia, ULB. The BiGRe laboratory is a partner of the BioSapiens Network of excellence funded under the sixth Framework program of the European Communities (LSHG-CT-2003-503265). Publication costs were sponsored by the Région Wallonne de Belgique (TransMaze project 415925).

Literature Cited

Allison GE, Angeles D, Tran-Dinh N, Verma NK. 2002. Complete genomic sequence of SfV, a serotype-converting temperate bacteriophage of *Shigella flexneri*. J Bacteriol. 184:1974–1987.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Barabasi AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nat Rev Genet. 5:101–113.

Botstein D. 1980. A theory of modular evolution for bacteriophages. Ann N Y Acad Sci. 354:484–490.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA. 99:14250–14255.

Brohee S, van Helden J. 2006. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics. 7:488.

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

Brussow H, Desiere F. 2001. Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. Mol Microbiol. 39:213–222.

Brussow H, Hendrix RW. 2002. Phage genomics: small is beautiful. Cell. 108:13–16.

Casjens SR, Gilcrease EB, Huang WM, Bunny KL, Pedulla ML, Ford ME, Houtz JM, Hatfull GF, Hendrix RW. 2004. The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. J Bacteriol. 186:1818–1832.

Ceyssens PJ, Lavigne R, Mattheus W, Chibeu A, Hertveldt K, Mast J, Robben J, Volckaert G. 2006. Genomic analysis of *Pseudomonas aeruginosa* phages LKD16 and LKA1: establishment of the phiKMV subgroup within the T7 supergroup. J Bacteriol. 188:6924–6931.

Chibani-Chennoufi S, Bruttin A, Dillmann ML, Brussow H. 2004. Phage-host interaction: an ecological perspective. J Bacteriol. 186:3677–3686.

Chibani-Chennoufi S, Canchaya C, Bruttin A, Brussow H. 2004. Comparative genomics of the T4-Like *Escherichia coli* phage JS98: implications for the evolution of T4 phages. J Bacteriol. 186:8276–8286.

Chibani-Chennoufi S, Dillmann ML, Marvin-Guy L, Rami-Shojaei S, Brussow H. 2004. *Lactobacillus plantarum* bacteriophage LP65: a new member of the SPO1-like genus of the family Myoviridae. J Bacteriol. 186:7069–7083.

Chopin A, Bolotin A, Sorokin A, Ehrlich SD, Chopin M. 2001. Analysis of six prophages in *Lactococcus lactis* IL1403: different genetic structure of temperate and virulent phage populations. Nucleic Acids Res. 29:644–651.

Diestel R. 1997. Graph theory. New York: Springer.

Erdos P, Renyi A. 1959. On random graphs i. Publ Math. 6:290–297.

Fauquet CM, Fargette D. 2005. International Committee on Taxonomy of Viruses and the 3,142 unassigned species. Virol J. 2:64.

Fauquet CM, Mayo MA, Maniloff J, Deseelberger U, Ball LA. 2005. Virus taxonomy. Classification and nomenclature of viruses. San Diego (CA): Elsevier.

Filee 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.

Freeman LC. 1977. A set of measures of centrality based on betweenness. Sociometry. 40:35–41.

Fruchterman TMJ, Reingold EM. 1991. Graph drawing by force-directed placement. Softw Pract Exp. 21:1129–1164.

Girvan M, Newman ME. 2002. Community structure in social and biological networks. Proc Natl Acad Sci USA. 99:7821–7826.

Glazko GV, Mushegian AR. 2004. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. Genome Biol. 5:R32.

Gordon AD. 1999. Classification. Washington (DC): Chapman and Hall/CRC.

Hendrix RW. 2003. Bacteriophage genomics. Curr Opin Microbiol. 6:506–511.

Hendrix RW, Casjens S. 2005. Bacteriophage l and its genetic neighborhood. In: Calendar R, editor. The bacteriophages. Oxford: Oxford University Press. p. 409–446.

Hendrix RW, Lawrence JG, Hatfull GF, Casjens S. 2000. The origins and ongoing evolution of viruses. Trends Microbiol. 8:504–508.

Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci USA. 96:2192–2197.

Hertwig S, Klein I, Schmidt V, Beck S, Hammerl JA, Appel B. 2003. Sequence analysis of the genome of the temperate *Yersinia enterocolitica* phage PY54. J Mol Biol. 331:605–622.

Iyer LM, Balaji S, Koonin EV, Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 117:156–184.

Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol. 299:27–51.

Koonin EV, Mushegian AR, Bork P. 1996. Non-orthologous gene displacement. Trends Genet. 12:334–336.

Lawrence JG, Hatfull GF, Hendrix RW. 2002. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol. 184:4891–4905.

Leplae R, Hebrant A, Wodak SJ, Toussaint A. 2004. ACLAME: a classification of mobile genetic elements. Nucleic Acids Res. 32:D45–D49.

Lima-Mendez G, Toussaint A, Leplae R. 2007. Analysis of the phage sequence space: the benefit of structured information. Virology. 365:241–249.

Mayo MA, Ball LA. 2006. ICTV in San Francisco: a report from the plenary session. Arch Virol. 151:413–422.

Mmolawa PT, Schmieger H, Heuzenroeder MW. 2003. Bacteriophage ST64B, a genetic mosaic of genes from diverse sources isolated from *Salmonella enterica* serovar typhimurium DT 64. J Bacteriol. 185:6481–6485.

Nelson D. 2004. Phage taxonomy: we agree to disagree. J Bacteriol. 186:7029–7031.

Oakey HJ, Cullen BR, Owens L. 2002. The complete nucleotide sequence of the *Vibrio harveyi* bacteriophage VHML. J Appl Microbiol. 93:1089–1098.

Ogura Y, Ooka T, Asadulghani , et al. (13 co-authors). 2007. Extensive genomic diversity and selective conservation of virulence-determinants in enterohemorrhagic *Escherichia coli* strains of O157 and non-O157 serotypes. Genome Biol. 8:R138.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA. 96:4285–4288.

Pride DT, Wassenaar TM, Ghose C, Blaser MJ. 2006. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics. 7:8.

Proux C, van Sinderen D, Suarez J, Garcia P, Ladero V, Fitzgerald GF, Desiere F, Brussow H. 2002. The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. J Bacteriol. 184:6026–6036.

Recktenwald J, Schmidt H. 2002. The nucleotide sequence of Shiga toxin (Stx) 2e-encoding phage phiP27 is not related to other Stx phage genomes, but the modular genetic structure is conserved. Infect Immun. 70:1896–1908.

Rohwer F, Edwards R. 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol. 184:4529–4535.

Rybchin VN, Svarchevsky AN. 1999. The plasmid prophage N15: a linear DNA with covalently closed ends. Mol Microbiol. 33:895–903.

Sabidussi G. 1966. The centrality of a graph. Psychometrika. 31:581–603.

Scholl D, Kieleczawa J, Kemp P, Rush J, Richardson CC, Merril C, Adhya S, Molineux IJ. 2004. Genomic analysis of bacteriophages SP6 and K1-5, an estranged subgroup of the T7 supergroup. J Mol Biol. 335:1151–1171.

Snel B, Huynen MA. 2004. Quantifying modularity in the evolution of biomolecular systems. Genome Res. 14:391–397.

Susskind MM, Botstein D. 1978. Molecular genetics of bacteriophage P22. Microbiol Rev. 42:385–413.

van Dongen S. 2000. Graph clustering by flow simulation. Amsterdam (The Netherlands): Centre for Mathematics and Computer science. p. 173.

van Dongen S. 2000. Graph clustering by flow simulation. [PhD thesis]. [Amsterdam (The Netherlands)]: Centre for Mathematics and Computer science. 173p.

van Helden J. 2003. Regulatory sequence analysis tools. Nucleic Acids Res. 31:3593–3596.

Vlasblom J, Wu S, Pu S, Superina M, Liu G, Orsi C, Wodak SJ. 2006. GenePro: a cytoscape plug-in for advanced visualization and analysis of interaction networks. Bioinformatics. 22:2178–2179.

Wang J, Jiang Y, Vincent M, Sun Y, Yu H, Wang J, Bao Q, Kong H, Hu S. 2005. Complete genome sequence of bacteriophage T5. Virology. 332:45–65.

Watts DJ, Strogatz SH. 1998. Collective dynamics of 'small-world' networks. Nature. 393:440–442.

Westmoreland BC, Szybalski W, Ris H. 1969. Mapping of deletions and substitutions in heteroduplex DNA molecules of bacteriophage lambda by electron microscopy. Science. 163:1343–1348.

Yuzenkova J, Nechaev S, Berlin J, Rogulja D, Kuznedelov K, Inman R, Mushegian A, Severinov K. 2003. Genome of *Xanthomonas oryzae* bacteriophage Xp10: an odd T-odd phage. J Mol Biol. 330:735–748.

William Martin, Associate Editor

Accepted January 20, 2008