

Codon Bias is a Major Factor Explaining Phage Evolution in Translationally Biased Hosts

Alessandra Carbone

Received: 27 July 2007 / Accepted: 7 December 2007
© Springer Science+Business Media, LLC 2008

Abstract The size and diversity of bacteriophage populations require methodologies to quantitatively study the landscape of phage differences. Statistical approaches are confronted with small genome sizes forbidding significant single-phage analysis, and comparative methods analyzing full phage genomes represent an alternative but they are of difficult interpretation due to lateral gene transfer, which creates a mosaic spectrum of related phage species. Based on a large-scale codon bias analysis of 116 DNA phages hosted by 11 translationally biased bacteria belonging to different phylogenetic families, we observe that phage genomes are almost always under codon selective pressure imposed by translationally biased hosts, and we propose a classification of phages with translationally biased hosts which is based on adaptation patterns. We introduce a computational method for comparing phages sharing homologous proteins, possibly accepted by different hosts. We observe that throughout phages, independently from the host, capsid genes appear to be the most affected by host translational bias. For coliphages, genes involved in virion morphogenesis, host interaction and ssDNA binding are also affected by adaptive pressure. Adaptation affects long and small phages in a significant way. We analyze in more detail the Microviridae phage space to illustrate the

potentiality of the approach. The small number of directions in adaptation observed in phages grouped around ϕ X174 is discussed in the light of functional bias. The adaptation analysis of the set of Microviridae phages defined around ϕ MH2K shows that phage classification based on adaptation does not reflect bacterial phylogeny.

Keywords Bacteria · Phages · Codon bias · Codon Adaptation Index · Self-Consistent Codon Index · Microbial evolution · Microbial lifestyle

Introduction

Translational selection refers to the benefit of an increased translational output for a fixed investment in the translational machinery (ribosomes, tRNA, elongation factors, etc.) if only a subset of codons (and their corresponding tRNAs) is used preferentially. Since the benefit of using a particular codon depends on how often it is translated, the strength of translational selection, and hence the degree of codon bias, is expected to vary with the expression level of a gene within an organism (Grantham et al. 1980; Sharp and Li 1987). In bacteria and archaea, translational bias provides information on living environment (Carbone et al. 2005; Willenbrock et al. 2006) and on genes involved in essential metabolic functions and stress response, which are crucial for the bacteria wildlife (Carbone and Madden 2005; Carbone 2006). (For translational bias in eukaryotes see Sharp et al. 1988; Stenico et al. 1994; Carbone et al. 2003; Akashi 2001; Kliman and Hey 1994.)

We test the hypothesis that the strong functional signal inherent in codon bias of translationally biased bacteria also provides information on the evolution of phages infecting them. Working on this confined set of hosts does

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9068-6) contains supplementary material, which is available to authorized users.

A. Carbone (✉)
Génomique Analytique, Université Pierre et Marie Curie-Paris 6,
UMR S511, 91 Bd de l'Hôpital, 75013 Paris, France
e-mail: Alessandra.Carbone@lip6.fr

A. Carbone
INSERM, U511, Paris 75013, France

not restrict the analysis of viruses for a combination of reasons: (1) the majority of viruses found in the environment are phages (Weinbauer 2004), especially double-stranded DNA (dsDNA)-tailed bacteriophages (Hendrix et al. 1999; Coetzee 1987); (2) translationally biased bacteria belong to a large variety of phylogenetic classes (Carbone et al. 2005); and (3) most phages whose genomic sequence is available are hosted by translationally biased bacteria.

Phages with translationally biased hosts contain at most a subset of tRNAs in their genomes, and the majority of them contain no tRNA genes. Therefore the translation of their genes sharply rely on the tRNA pool made available by the host (Sharp et al. 1985; Sharp 1986). We thus expect phages to display the codon bias of their hosts, as preferred codons and iso-acceptor tRNA content exhibit a strong positive correlation (Ikemura 1985; Bennetzen and Hall 1982) and tRNA isoacceptor pools affect the rate of polypeptide chain elongation (Varenne et al. 1984).

Phage adaptation to host translational efficiency was observed several times in previous studies run on the few available phages hosted by *Escherichia coli* and *Staphylococcus aureus*, two translationally biased bacteria (Sharp and Li 1987; Carbone et al. 2003). Adjustment of codon usage patterns in coliphage genes to gene expression level has been observed by Gouy (1987) and Kunisawa et al. (1998). In phage T4, adaptation of patterns of codon usage to expression in *E. coli* has been related to the time of gene expression (Cowe and Sharp 1991). For several other T4-like phages and KVP40, much less correlation was found (Nolan et al. 2006), reflecting that the functional role of tRNA genes in coliphages remains unclear. Codon usage variation is influenced by translational selection also in *S. aureus* phages (Sau et al. 2005), and phages 44AHJD, P68, and K were argued to be extremely virulent in nature, as most of their genes have a high translation efficiency.

In a large-scale analysis of 116 genomes hosted by 11 translationally biased bacteria belonging to five different phylogenetic families, we demonstrate that DNA phage genomes are almost always under codon selective pressure imposed by the host and that, throughout phages and hosts, capsid genes consistently appear to be the most biased. Also, we provide a complete functional classification of biased genes in 28 coliphages and observe that genes involved in tail formation, lysis, and host interaction are also undergoing adaptation.

The codon bias measure used here is the Self-Consistent Codon Index (SCCI) (Carbone et al. 2003; Carbone 2006). It measures the dominant bias of a genome from its protein coding sequences without prior knowledge about gene expression or even functional annotation. Yet it is equivalent to the traditional Codon Adaptation Index (CAI)

(Sharp and Li 1987) for translationally biased genomes. Phage genes are indexed with the SCCI of their host. Namely, SCCI values reflect codon composition of phage genes *relative* to host codon composition and provide a numerical index of the advantage taken by phage genes once translated in the host environment. This advantage is expected to be higher when phage gene codon composition is biased toward host codon composition.

The SCCI proved to be useful to capture lifestyle features and essential genes in bacteria (Carbone and Madden 2005; Carbone 2006). Here we use it to reconstruct relationships among phages of the same species. We consider Microviridae phages and examine their pattern of evolution from codon bias analysis. Surprisingly, we can reconstruct the phylogenetic tree of this large phage pool (Rokyta et al. 2006) using exclusively codon bias information. This result highlights that adaptation patterns in phages might be profitably used to unravel the intricate mosaic of phage speciation.

The phage classification method based on codon bias that we propose fits in the spirit of the Phage Proteomic Tree (Rohwer and Edwards 2002) and Phage Orthologous Groups (Liu et al. 2006). Instead of amino acids, we use a refined measure of codon adaptation that allows us to zoom more precisely into phage evolutionary patterns for a detailed understanding of selection in phages with translationally biased hosts. It can profitably be used within phage species belonging to the same host but also between phages with different hosts. The only requirement is that phage species share homologous genes.

The numerical finding provided by this and future studies of phage-host coevolution will hopefully be useful in clarifying the role of phages as therapeutic agents against bacteria (Summers 2001) and in organizing metagenomic data.

Materials and Methods

Genomes and Annotation

Phage and bacterial genomes flatfiles were retrieved from GenBank. Hosts are listed in Table 1, and phages in Supplementary Table 1. Among them, there are six mobile elements found by sequencing the *Streptococcus* MGAS315 genome (Beres et al. 2002). The 116 phages are all DNA phages (either dsDNA or single-stranded DNA [ssDNA]); from a preliminary screening we found no bias in *E. coli* RNA phages examined in agreement with a previous analysis realized on a few RNA coliphages (Gouy 1987). We also considered 42 Microviridae phages reported by Rokyta et al. (2006) and the Microviridae phages G4, α 3, ϕ K, S13, and ϕ X174, all known to infect *E. coli*.

Table 1 Host genomes, SCCI, and Pearson correlation coefficients among different phage properties

Host organism		Host SCCI		Correlations ^a and phage properties					
Name	Class	Mean	SD	No. genes & hi-bias	No. genes & μ_{SCCI}	Length & hi-bias	Length & μ_{SCCI}	Max	No. phages
<i>Escherichia coli</i> K12	γ -Proteobacteria	0.31	0.1	0.82	0.26	0.8	0.24	292	28
<i>Salmonella thyphimurium</i> LT2	γ -Proteobacteria	0.35	0.09	0.18	-0.16	0.52	0.33	72	4
<i>Vibrio cholerae</i> O1 biovar eltor str. N16961	γ -Proteobacteria	0.28	0.08	0.99	0.54	0.98	0.6	381	10
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	Firmicutes <i>bacillales</i>	0.36	0.07	0.89	-0.14	0.95	0.0027	185	6
<i>Listeria monocytogenes</i> EGD-e	Firmicutes <i>bacillales</i>	0.47	0.09	0.58	-0.06	0.52	-0.95	174	3
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> MW2	Firmicutes <i>bacillales</i>	0.43	0.1	0.8	0.011	0.87	0.16	214	18
<i>Lactococcus lactis</i> subsp. <i>lactis</i> II1403	Firmicutes <i>lactobacillales</i>	0.32	0.1	0.12	-0.18	0.18	0.27	63	14
<i>Streptococcus pyogenes</i> MGAS315	Firmicutes <i>lactobacillales</i>	0.3	0.09	-0.08	-0.28	-0.44	-0.64	64	6
<i>Chlamidophyla caviae</i> GPIC	Chlamydiales	0.49	0.05	0.99	0.94	0.98	0.99	12	3
<i>Mycobacterium smegmatis</i> str. MC2 155	Actinobacteria	0.51	0.11	0.4	-0.06	0.69	0.22	237	21
<i>Mycobacterium tuberculosis</i> H37Rv	Actinobacteria	0.5	0.07	0.38	-0.48	-0.39	-0.96	98	3
All phages confounded				0.63	0.18	0.66	0.21	116	

^a *rPs* for genome length (length), number of genes (no. genes), number of highly biased genes (hi-bias), and mean SCCI (μ_{SCCI}) in phages

^b Number of phages (no. phages) and maximum number of genes (max) for the set of phages considered in this paper

Functional annotation of phage genes was extracted from GenBank files for all phages with the exception of mycobacteriophage Che12, for which we referred to Gomathi et al. (2007). Proteins considered to be involved in lysis are holin, lysin, excisionase, integrase, and repressor proteins (immunity maintenance repressor CI, establishment of lysogeny CII and CIII). Virion morphogenesis includes proteins involved in head decoration, scaffolding, maturation, head completion, and capsid assembly. The ensemble of capsid/head/coat proteins includes precursors also. The number of tRNAs in phages (in Supplementary Table 1) was evaluated after annotation in GenBank files.

Calculation of the Self-Consistent Codon Index

Sharp (Sharp and Li 1987) formulated the hypothesis that for translationally biased genomes *G*, there is a *reference set S* of coding sequences, constituting roughly 1% of the genes in *G*, which are representative of codon adaptation in *G*. This bias can be described by listing a set of codon weights calculated on genes in *S*: given an amino acid *j*, its synonymous codons might have different frequencies in *S*; if $x_{i,j}$ is the number of times that the codon *i* for the amino acid *j* occurs in *S*, then one associates with *i* a weight $w_{i,j} = x_{i,j}/y_j$ relative to its sibling of maximal frequency *y_j*

in *S*. Such weights are used to compute the CAI (Sharp and Li 1987) for all genes, $CAI(g) = (\prod_{k=1}^L w_k)^{1/L}$, where *g* is a gene, w_k is the weight of the *k*th codon in *g*, *L* is the number of codons in *g*, and the reference set *S* is manually defined as the set of genes coding for proteins known to be highly expressed, as ribosomal and glycolytic proteins are for fast growers. For fast-growing bacteria, genes with a high CAI value turn out to be the ones which are highly expressed (Sharp and Li 1987).

In Carbone et al. (2003), we extended Sharp's hypothesis, saying that for a genome *G*, there is a reference set *S* of coding sequences, which is representative of *dominating codon bias* in *G*. We consider the *SCCI* to be defined as $SCCI(g) = (\prod_{k=1}^L w_k)^{1/L}$, where the reference set *S* is the most biased set of genes in the organism with respect to this formula; that is, *S* is the (self-consistent) set of genes that take maximum value in the formula when *S* is chosen as a reference set. *SCCI* correlates with the dominating bias in a genome, such as GC content, GC3 content, a leading strand richer in *G+T* than a lagging strand, and translational bias. For translationally biased organisms, *SCCI* computes codon adaptation and it coincides with CAI. The computation of the reference set *S* is based on a pure statistical analysis of all genes in a genome and it does not rely on biological knowledge of the organism. This allows us to compute weights for organisms of unknown lifestyle.

The name SCCI was employed for the first time by Carbone (2006), while in Carbone et al. (2003, 2005) the notion is called CAI, even though it does not exclusively refer to codon adaptation. Notice that CAI is usually employed with a manual and explicit choice of S , while the formula SCCI (i.e., CAI parameterized with S) turns out to be a *universal measure* to study codon bias. Codon weights, reference set S , and SCCI values are calculated with the program CAIJava (Carbone et al. 2003), available at www.ihes.fr/~carbone/data.htm.

Detection of Weak and Strong Forms of Translational Bias for Bacteria

In Carbone et al. (2005), two numerical criteria were introduced to detect translational bias in bacteria. The “ribosomal criterion” defines the z -score $(SCCI(r) - \mu)/\sigma$, for each gene of a ribosomal protein r , where mean μ and standard deviation σ are calculated for the SCCI distribution over all CDS; this allows the definition of the average \bar{z}_{Rib} of z -scores for ribosomal proteins and to say that an organism characterized by translational bias is expected to have high \bar{z}_{Rib} , i.e., >1 . The “strength criterion” computes codon weights based on all genes in the genome G ($w_k(G)$) and on the genes in the reference set S (w_k) and expects the difference between $w_k(G)$ and w_k to be large for translationally biased genomes (Carbone et al. 2005). The combination of the two criteria differentiates those genomes that are strongly translationally biased, that is, those satisfying both criteria, from those that are weakly so, that is, those that only satisfy the ribosomal criterion. Another numerical approach, on the spirit of the ribosomal criteria, has been introduced by Sharp et al. (2005).

Translationally Biased Hosts and Bias Strength of Bacteria

The 11 hosts listed in Table 1 have been shown to display clear signals of translational bias. Among them, *E. coli*, *S. typhimurium*, *V. cholerae*, *B. subtilis*, *L. lactis*, and *S. pyogenes* show a strong form of translational bias (Carbone et al. 2005; see also Sharp and Li 1987; Shields and Sharp 1987; Gupta et al. 2004), while all the remaining bacteria display a weaker form (Carbone et al. 2005). A weak form of translational bias governing *Chlamydomonas reinhardtii* GPIC and *Chlamydomonas reinhardtii* AR39 and a strong form of translational bias governing *Bdellovibrio bacteriovorus* have been shown using tools and numerical criteria described by Carbone et al. (2005; see also Lu et al. 2005). Hosts with a weak form of translational bias show coexisting evolutionary trends, and overlapping of trends

weakens the statistical signal detected within genomes. Even if weak, this signal can be successfully used to predict essential genes (Carbone 2006) and essential metabolic networks in bacteria (Carbone and Madden 2005). Thus, it appears appropriate to search for phage adaptation in hosts with weak translational bias.

Detection of Highly Biased Genes in Phages

We run CAIJava on phage genomes using the codon weights of the host (these weights are computed on the set of most highly biased host genes), and we ranked phage genes by their SCCI value. Highly biased phage genes g are those with $SCCI(g) \geq \mu + \sigma$, with μ and σ being the mean and the standard deviation of the SCCI distribution for the host. This threshold has been used by Carbone and Madden (2005) and Carbone (2006) to identify bacterial genes whose codon bias deviates greatly from the average and that are susceptible to being needed greatly by the cell.

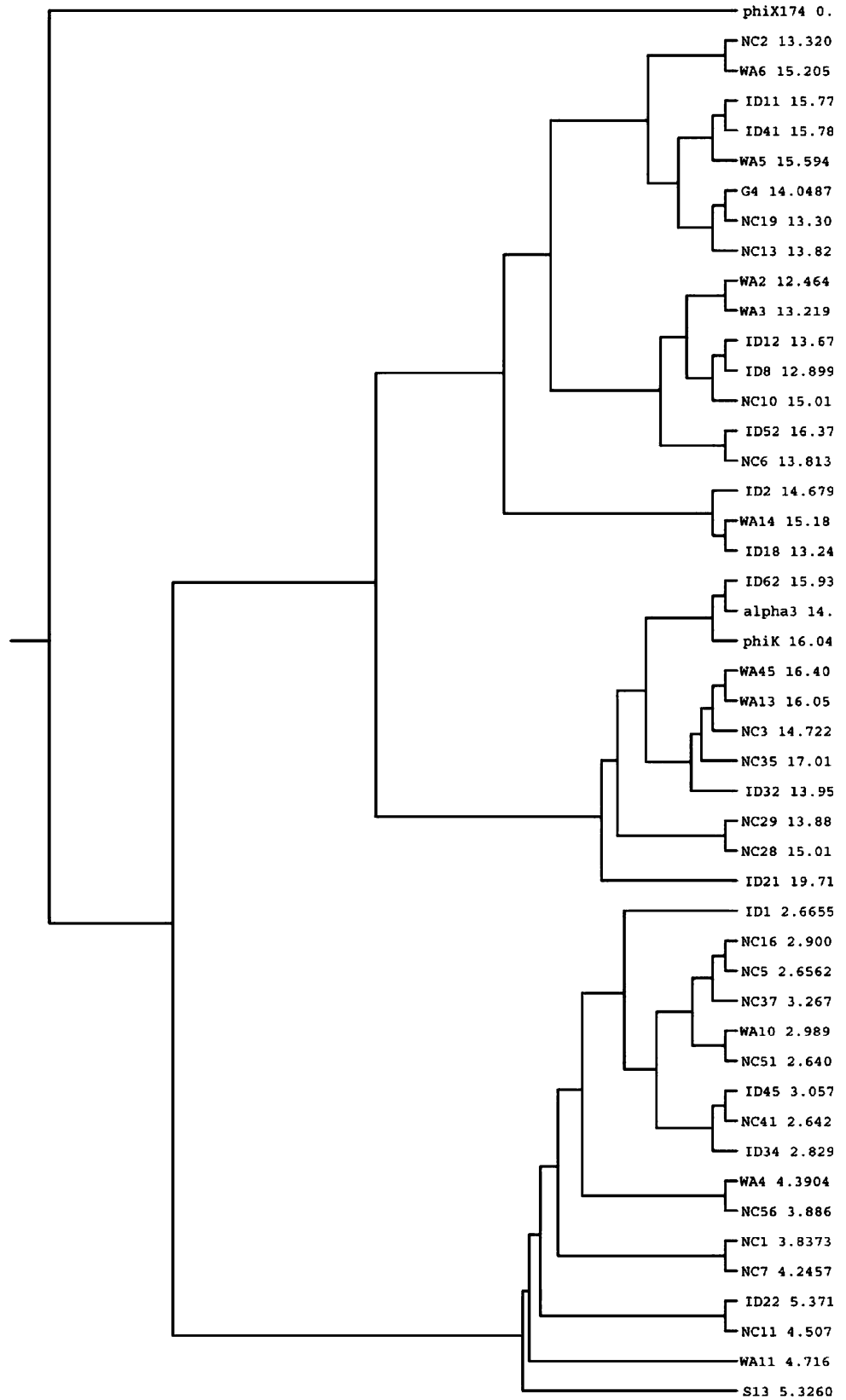
Fitness of a Phage Versus Its Host

SCCI values provide a numerical index of the advantage taken by phage genes once translated in the host cell, and numerically define the *relative* bias of a phage genome versus the host genome. Along the same line, we define the *fitness* of the phage bias as the percentage of highly biased genes in the phage genome. The higher the percentage of highly biased genes, the stronger the fitness. Fitness values are reported in Supplementary Table 1.

Definition of a Space of Phages with a Single Host

The 42 Microviridae phages in Rokyta et al. (2006) are approximately 6K bases long and contain 10 genes, A , B , C , D , E , F , G , H , J , and K , with one more gene, A^* , which is missing in some of the 42 phages. Gene J is very small (78 bp) and is involved in DNA binding; gene F codes for a major capsid protein, and gene G for a major spike protein. We represent a phage as a 10-dimensional normalized vector with entries corresponding to SCCI values of the 10 shared genes. Given two vectors w^1 , w^2 of weights w_i^1 , w_i^2 , corresponding to two genomes G_1 , G_2 , the ℓ_1 -distance between G_1 and G_2 is defined as $\sum_{i=1}^{64} |w_i^1 - w_i^2|$. (Motivations for working with an ℓ_1 -distance instead of an ℓ_2 -distance, that is, the sum of the squared differences, between the two vectors w^1 , w^2 are given by Carbone et al. [2005].) For the 42 phages in the Rokyta set and the Microviridae phages G4, $\alpha 3$, ϕK , S13, and $\phi X174$, we compute the ℓ_1 -distance of each pair of points in the

Fig. 1 Distance tree of 47 Microviridae phages reconstructed in codon space using the ℓ_j -distance



10-dimensional space and we construct the associated tree with UPGMA (see Fig. 1). The construction of a consensus tree based on multiple runnings of the UPGMA

algorithm (bioweb.pasteur.fr/seqanal/interfaces/neighbor-simple.html) confirms the tree structure and the separation of the phages into three distinct groups.

Principal component analysis (PCA) applied to the 47 points sitting in the 10-dimensional space and linear discriminant analysis (LDA) providing discriminant coefficients confirm the three distinct phage groups (see Supplementary Table 2 and Supplementary Fig. 1).

Definition of a Space of Phages with Multiple Hosts

Phages ϕ MH2K, chp1, chp2, CPAR39, and guinea pig Chl form a class of Microviridae phages infecting nonenterobacterial Proteobacteria and Chlamidiae. We do not consider Spiroplasma phage S4 because of the unavailability of the host genome. These phages are compared by looking at the adaptation of their five shared proteins, *F*, *A*, *H*, *C*, and *B*; each phage is represented as a five-dimensional vector whose entries correspond to the SCCI value of the five genes and where the SCCI value of a gene is computed with respect to the SCCI weights of its host. The main difference with a space of phages sharing the same host is that for groups of phages having possibly different hosts, we have to normalize the vector representing each phage by replacing the entry x with $(x - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the host SCCI distribution. Roughly speaking, a vector represents the adaptation of phage genes within the host and measures the proximity of phage composition to host bias.

Formal Definitions of Pattern of Adaptation and of Correlation Distance

We assume organisms to share some set of proteins and to be represented by vectors of SCCI values of genes coding for these proteins, where SCCI values are computed on host weights. In the case of multiple hosts, vectors are normalized as above. We say that two phages have the same *pattern of adaptation* if the Pearson correlation coefficient (rP) of their relative vectors is high. By numerical evaluation realized on the 47 Microviridae phages, we estimated that high is >0.9 . This value might be adaptable, and in general, it should be estimated from the distribution of rPs calculated over all pairs of vectors under study. The classification tree, based on rPs calculated on the 47 Microviridae phages (and computed from the associated distance matrix with UPGMA), is given in Fig. 2. The distance between two organisms is defined as 1 minus their rP and it is computed with the package R (www.r-project.org). We refer to it as the “correlation distance.” Intuitively, the correlation distance allows the identification of pairs of phages that have chosen to optimize the same proteins, and the correlation distance tree

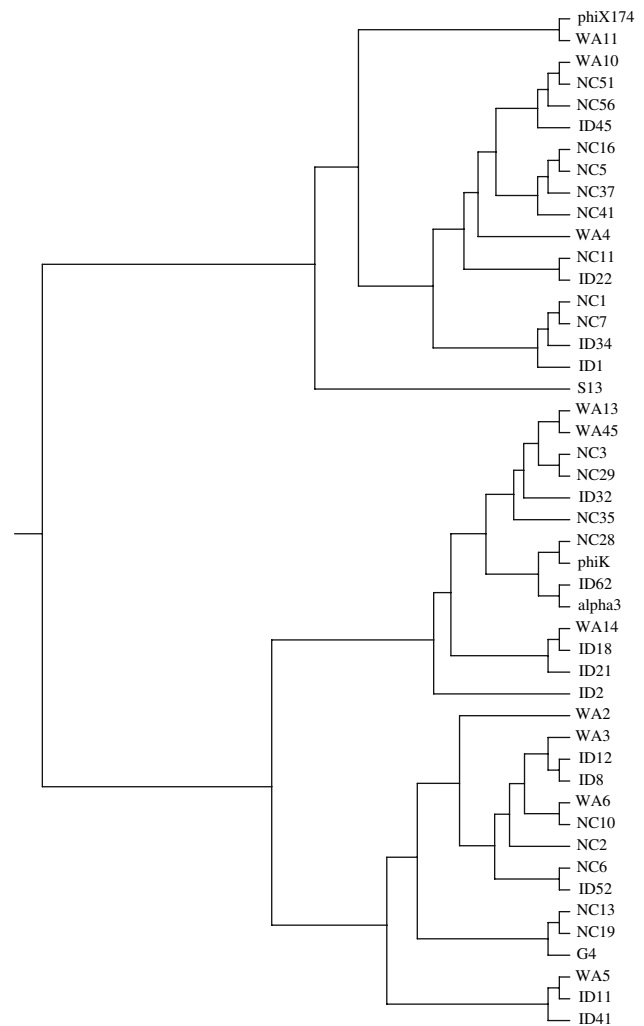


Fig. 2 Distance tree of 47 Microviridae phages reconstructed in codon space using correlations between vectors representing phages

allows grouping of phages undergoing similar functional pressure.

Microviridae Phages Tree Based on Amino Acid Sequences

We concatenated the 10 amino acid sequences corresponding to the 10 genes shared by all 42 phages in the Rokyta set (Rokyta et al. 2006) and by Microviridae phages G4, α 3, ϕ K, S13, and ϕ X174. After aligning the 47 amino acid sequences with Multalin (<http://www.bioinfo.genopole-toulouse.prd.fr/multalin/multalin.html>) (Corpet 1988), we computed the distances with *Protdist* (<http://www.evolution.genetics.washington.edu/phylip.html>) and reconstructed the distance tree with UPGMA. See Supplementary Fig. 2. To verify the robustness of the results we constructed the tree based on ClustalW and PHYML using the JTT model

of amino acid substitution also and obtained the same three group partition (see Results).

Results

Highly Biased Genes in Phages

We considered 116 phages hosted by 11 translationally biased bacteria belonging to different phylogenetic families (listed in Table 1). We studied codon bias tendencies in each phage compared to bias in the host genome and found that 101 of these phages display at least one gene with high bias (see Supplementary Table 1). The number of highly biased genes varies from phage to phage from a few percent up to the 80% of the total number of genes in the phage genome (43 phages have >10% highly biased genes and 32 have between 5 and 10%). Among the 116 phages, 64 have been annotated so far and we report a functional classification of their highly biased genes in Table 2.

A comparison between the SCCI distributions of host genomes (see Table 1, left) and the SCCI distributions of phage genomes (see Supplementary Table 1) shows that phage genes are usually not biased and that only a restricted pool of selected viral genes is. In fact, three different scenarios of phage-host evolutionary relationship arise: either most phage genes are coded with host preferred codons ($\mu_{\text{SCCI}(\text{phage})} > \mu_{\text{SCCI}(\text{host})} + \sigma_{\text{SCCI}(\text{host})}$, where $\text{SCCI}(\text{phage})$ and $\text{SCCI}(\text{host})$ denote the SCCI distributions on phage genes and host genes, respectively), or the phage genome is biased toward the host genome ($\mu_{\text{SCCI}(\text{phage})} > \mu_{\text{SCCI}(\text{host})}$), or a few phage genes are coded with host preferred codons ($\mu_{\text{SCCI}(\text{phage})} < \mu_{\text{SCCI}(\text{host})}$). Namely, 4, 28, and 84 of 116 phages fall into the first, second, and third category, respectively (note that 9 of the 10 phage genomes of *V. cholerae* belong to the second group, and 2 to the first group). The 84 phages in the third group have $\mu_{\text{SCCI}(\text{phage})}$ “close” to the one of the host, that is, $\mu_{\text{SCCI}(\text{host})} - \sigma_{\text{SCCI}(\text{host})} < \mu_{\text{SCCI}(\text{phage})} < \mu_{\text{SCCI}(\text{host})}$. From this partition, it follows that most phage genes employ codons which are not preferential for the host even though codon bias within the phage genome is present and it affects genes belonging to few preferential functional classes. The latter fact establishes the evolutionary pressure on phage genomes imposed by host codon composition.

Evidence of the Adaptation of Capsid Genes to Host Bias

Capsid genes consistently appear to be the most biased, throughout all phages and all hosts. Of the 64 annotated genomes in Table 2, 53 of them have highly biased capsid

proteins (12 phages display just 1 highly biased gene, and capsid proteins are highly biased for 9 of them). Sixteen phages display proteins involved in lysis, and 13 in tail. We can safely conclude that there is a tendency in evolving capsid proteins, and in minor extent tail and lysis proteins, toward host bias. Too many nonannotated proteins are present in the dataset to be able to conclude more than this. A better understanding of genes coding for proteins of unknown function will help to settle the evidence reported here.

Adaptation in Coliphages

Phage genes with a SCCI value greater than $\mu_{\text{SCCI}(\text{host})}$ provide further information on the tendency of phage adaptation versus host codon composition. We call these genes “biased” (in contrast to “highly biased”) and we look at their distribution in coliphage genomes (Supplementary Table 4). On average, the 30% of genes in a coliphage genome are biased. By looking at the distribution of biased genes in all functional classes involving at least a biased gene in some coliphage, we observe that a few specific functional classes tend to be represented by biased genes for most coliphages. Essentially all of the 28 coliphages contain biased genes involved in virion morphogenesis, tail formation, and lysis, besides capsid formation. We also observe that genes coding for host interaction, inhibition of host functions, ssDNA binding, and transcription regulation are well represented within biased genes, although in a minor manner. If phage genomes were to contain a pool of *essential* genes, these functional classes could suggest appropriate candidate genes. (A predictive study of essential genes in bacteria was done by Carbone [2006].) In this direction and in agreement with Kunisawa et al. (1998), we found that virulent phages T7 and T4 follow the pattern of translational bias typically found in essential genes of the host *E. coli*. The same holds true for RB43, RB49, and RB69. It was also claimed that patterns from temperate phages P4, P2, N15, and λ were similar to the pattern of *E. coli* foreign-type genes from prophages and transposons. In contrast, we found that P2, N15, and λ actually contain some highly biased genes and that capsid genes are among those for each phage species (see also Table 2).

Codon Adaptation Affects Small and Large Phages

Large dsDNA phages are generally thought to have evolved by capturing multiple genes from their hosts (Filée et al. 2003), and above, we showed that dsDNA phages having translationally biased hosts are influenced in a

Table 2 Functional classification of highly biased genes in annotated biased phages

Phage ^a	No. genes	Capsid	Morpho	Tail	Lysis	Hyp	Other	No. hi-bias genes
Enterobacteria phage RB43	292	5		4		35	7	51
Enterobacteria phage RB49	274	3	2	2		26	6	39
Enterobacteria phage RB69	275	4	2	1	1	8	4	20
Enterobacteria phage T4	288	2	2			3	3	10
Enterobacteria phage T7	60	2	1			2	3	8
Enterobacteria phage K1E	62	1			1	3		5
Enterobacteria phage K1F	43	1	1			1	1	4
Enterobacteria phage μ	55	1				2		3
Enterobacteria phage P1	110	1		1			1	3
Enterobacteria phage N15	60	2				1		3
Enterobacteria phage α 3	10	1		1				2
Enterobacteria phage λ	71	1				1		2
Enterobacteria phage P2	43	1	1					2
Enterobacteria phage 186	42	1						1
Enterobacteria phage G4	10	1						1
Enterobacteria phage HK022	57	1						1
Enterobacteria phage I2-2	9	1						1
Enterobacteria phage Ike	10	1						1
Enterobacteria phage ϕ K	10	1						1
Enterobacteria phage ϕ X174	11						1	1
Enterobacteria phage 933W	80						1	1
Enterobacteria phage HK97	61	1						1
Enterobacteria phage ϵ 15	40					6	1	7
<i>Salmonella</i> phage SETP3	53	1				2		3
Enterobacteria phage P22	72	1				1		2
Enterobacteria phage S13	12						1	1
<i>Bacillus</i> phage SPBc2	185					50	9	59
<i>Bacillus</i> phage GA-1 virus	35	1	1		2	8	3	15
<i>Bacillus</i> phage PZA	27	1	1		1	4		7
<i>Bacillus</i> phage B103	17	1		2	1	1	1	6
<i>Mycobacterium</i> phage D29	79	1		5	1	38	3	48
<i>Mycobacterium</i> phage Che12	98	2	1	4	3	33	9	52
<i>Lactococcus</i> phage P335	49	2			2	1		5
<i>Lactococcus</i> phage bIL170	63	2				3		5
<i>Lactococcus</i> phage r1t	50	1				4		5
<i>Lactococcus</i> phage BK5-T	63	2			1		1	4
<i>Lactococcus</i> phage jj50	49	2			1	2		4
<i>Lactococcus</i> phage ϕ LC3	51	1	1			2		4
<i>L. lactis</i> Phage ul36	58	1	1		1			3
<i>Lactococcus</i> phage P008	58	2				1		3
<i>Lactococcus</i> phage 712	55	2						2
<i>Lactococcus</i> phage c2	39	1				1		2
<i>Lactococcus</i> phage Tuc2009	56	1				1		2
<i>Listeria</i> phage A118	72	1			1	4		6
<i>Listeria</i> phage P100	174					5	1	6
<i>Streptococcus</i> phage 315-5	55				1	6		7
<i>Streptococcus</i> phage 315-2	60					3	1	4

Table 2 continued

Phage ^a	No. genes	Capsid	Morpho	Tail	Lysis	Hyp	Other	No. hi-bias genes
<i>Streptococcus</i> phage 315-1	56	1		1		1		3
<i>Staphylococcus</i> phage K	115	1		1	2	18	3	25
<i>S. aureus</i> phage ϕ P68	22	3				3	1	7
<i>Staphylococcus</i> phage 44AHJD	21	3				3	1	7
<i>S. aureus</i> phage ϕ NM3 provirus	65					4	1	5
<i>S. aureus</i> phage PVL provirus	62					3		3
<i>S. aureus</i> phage ϕ 11 provirus	53	2					1	3
<i>Staphylococcus</i> phage ϕ ETA	66	1			1	1		3
<i>Staphylococcus</i> phage ϕ NM	64	1			1	1		3
<i>Vibrio</i> phage KVP40	381	2	1	4		178	57	242
<i>Vibrio</i> phage VP4	31	4	1	2		4	15	25
<i>Vibrio</i> phage VP2	47					12	1	13
<i>Vibrio</i> phage K139	44	1	1	1				3
<i>Vibrio</i> phage VGJ ϕ	13	1				1	1	3
<i>Chlamydia</i> phage chp1	12	4				3		9
<i>Chlamydia</i> phage chp2	8	1						1
<i>Chlamydia</i> phage ϕ CPG1	9	1						1

Note. No. genes denotes the number of genes in the phage genome. The number of highly biased genes (no. hi-bias genes) is split up into genes involved in capsid formation (capsid), morphogenesis (morpho), tail formation (tail), lysis (lysis), hypothetical proteins (hyp), and other proteins (other)

^a Phages are grouped by host and listed by decreasing number of highly biased genes

major manner by codon adaptation to the host. We analyze here how the number of genes relates to our SCCI measure and check whether “small” genomes (with fewer than 27 genes and 20,000 bases) are meaningfully affected by translational adaptation as well as “large” ones (with more than 150 genes and 100,000 bases). (Recall that our method is inherently independent on gene number and on genome length since SCCI is computed on host weight values.) As expected, the longer the genome, the higher the number of genes ($rP = 0.96$ between length and gene number) and the higher the number of highly biased genes ($rP = 0.69$ between number of genes and number of highly biased genes), even though in 5 phage families of 11, the rP between gene number and number of highly biased genes is positive but not very high (<0.38). We observe a significant number of small phages with a high percentage of highly biased genes, and this proves that codon adaptation affects small as well as large phages at a variable strength: of the 27 phages displaying $>15\%$ highly biased genes, 6 are large (they constitute half of the large phages within the 116), 8 are small (they make up a third of the small phages), and 13 have between 27 and 150 genes (shown in

Supplementary Table 1). Notice that single-host analysis could lead to wrong conclusions due to rather diversified rP values between gene number and number of highly biased genes: $rP = 0.8$ for *S. aureus* phages (in agreement with Sau et al. 2005), $rP = 0.99$ for *V. cholerae* phages, and $rP = 0.18$ for *L. lactis* phages. rPs between phage gene number and mean SCCI value of phage genes are also very diversified among hosts (see Table 1).

Hydrophilic Bias of Phage Proteins Issued by the Large-Scale Analysis

Even though our focus is on codon adaptation of phage genes, which is highly mediated by host translational bias, the large-scale analysis of phage genomes that we ran led to the observation that phage genes code for proteins displaying a highly hydrophilic character. They imply an amino acid usage that is sharply different from the one observed in the host hydrophobic proteins (such as membrane and permease proteins). This can be shown in codon space where host and phage genes are considered together:

phage genes occupy rather different positions than host genes, and in particular, the host cluster of membrane and permease genes (characterized, in all bacteria, by a biased usage of hydrophobic amino acids [Médigue et al. 1994; Carbone et al. 2003]) is essentially free of phage genes. Specificity and sensitivity of LDA separating the host membrane and permease genes from all phage genes considered for that host are given in Supplementary Table 3. All separating values rank high (with a sensitivity >88% and a specificity >76%) for all hosts.

Phage Organization of Microviridae Phage Groups: A Case Study

The two ssDNA Microviridae phages ϕ X174 (Sanger et al. 1977; Fiers and Sinsheimer 1962; Wichman et al. 1999) and ϕ MH2K (Brentlinger et al. 2002) exhibit extremely close relationships to the Microviridae of *E. coli* and *Chlamydia* respectively, in both genome organization and encoded proteins. In this way they typify two groups splitting the Microviridae world. We study phage organization based on codon adaptation of the two phage groups and suggest that codon adaptation might be a key element to understanding the intricate phage speciation.

Rokyta Sequences

Codon bias of 42 Microviridae phages, isolated in a single strain of *E. coli* (Rokyta et al. 2006), has been compared to codon bias of the related ϕ X174, S13, α 3, ϕ K, and G4 phages, and the analysis (based on ℓ_1 -distance tree reconstruction) identified three groups of phages located in well-separated parts of codon space. These groups were previously identified by Rokyta et al. (2006) by a reconstruction of the maximum a posteriori probability phylogeny based on full-genome comparison. (We found the same grouping by constructing a distance tree based on 10 concatenated proteins shared by all 47 phages, shown in Supplementary Fig. 2.)

The 42 sequences are organized around S13 and ϕ X174, α 3 and ϕ K, and G4 phages, respectively (see Figs. 1 and 2 and Supplementary Fig. 1). It might appear surprising that the sequence-based phylogenetic tree of Rokyta et al. (2006) is found here based on codon bias information and on a very simple distance-based algorithm. Since our tree (based on ℓ_1 -distances) is not aimed at phylogenetic reconstruction, but rather at identifying evolutionary directions of adaptation, the finding strongly suggests that the three phage groups correspond to evolutionary directions based on adaptation, where two of the groups are characterized by coding amelioration (corresponding to

improved SCCI values) for major coat gene *F* (group G4) and major spike gene *G* (group α 3 and ϕ K). The group characterized by ϕ X174 and S13 does not exhibit a significant codon bias variation, maintaining the coding pattern of ϕ X174.

A finer analysis of codon patterns, realized in codon space with correlation distances, highlights a small number of directions in pattern evolution which are guided by specific gene functions. The tree based on correlation distances groups Rokyta sequences roughly in the same way as the ℓ_1 -distance tree with a few exceptions. Phages WA14, ID18, and ID2 grouped with the G4-like phages in the ℓ_1 -distance tree, and display a stronger correlation with α 3-like phages in the correlation distance tree (see Figs. 1 and 2). This might indicate a new separate pattern of adaptation which cannot be clearly identified with current data due to phage underrepresentation. Also, phages WA6 and NC10 cluster together in the ℓ_1 -distance tree but fall into two separate clades in the correlation distance tree. This suggests that at the protein level, adaptation favors different proteins in the two phages and, in particular, different functions.

The ϕ MH2K World

Our quantitative method (determining the relative position of phages in codon space with correlation distances) provides a way to analyze patterns of adaptation of phages with different hosts. Codon adaptation of the five homologous proteins shared by all Microviridae phages known to be close to ϕ MH2K demonstrates that in phage space, *Chlamidophyla caviae* guinea pig Chl phage and *Chlamidophyla pneumoniae* AR39 virus CPAR39 display the same pattern of adaptation ($rP = 0.91$), and that *Bdellovibrio bacteriovorus* phage ϕ MH2K clusters close to *Chlamidophyla caviae* phages chp2 and guinea pig Chl phage and *Chlamidophyla pneumoniae* AR39 virus CPAR39, displaying a pattern of adaptation closer to *Chlamidophyla caviae* phage chp2 ($rP = 0.80$) than chp2 does to *Chlamidophyla caviae* phage chp1 ($rP = -0.05$ between chp2 and chp1; see Table 3). A qualitative analysis of B and C gene similarities in the two latter species allowed Brentlinger et al. (2002) to derive the same observation and argue that, if phages were to reflect their host's phylogeny, this result would be surprising. Even though we could expect phages with the same host to have similar codon usage indeed, there is no reason to think that adaptive evolutionary directions should be the same for phages infecting the same host. In fact, what we observe in codon space with correlation distance is that the evolutionary directions followed by the two Chlamydophilae phages differ and that one of them is functionally closer to

Table 3 Correlations between (vectors representing) ϕ MH2K, chp1, chp2, guinea pig Chl phage, and CPAR39

	ϕ MH2K	chp1	chp2	CPAR39
chp1	-0.45			
chp2	0.80	-0.05		
CPAR39	-0.54	-0.28	-0.73	
Guinea pig Chl phage	-0.4	-0.17	-0.7	0.91

Note. The correlation distance matrix is easily computed as $1 -$ the above matrix. Higher rP values correspond to closer distances

the *Bdellovibrio* phage. In this sense, our methodology is finer and it allows us to derive more subtle relationships between phages than previously done.

Discussion

Phages and Translational Bias Strength of the Host

Almost all of the 116 phages considered in this study are shown to contain some highly biased genes, and capsid genes are within highly biased genes for all annotated phages. This fact is shown, despite the strength, weak or strong, of the translational bias affecting the host.

Promiscuous Phages

Codon usage bias might vary a great deal among genes within a phage genome, and one could wonder whether phages with promiscuous host preferences could manage to deal with different host tRNA abundances. Our finding suggests that phages can be promiscuous, but not in some arbitrary way. In fact, phages need to guarantee that their highly biased genes fit a few specific functional classes and that the codon composition of such genes should reflect the dominant bias of the host. Bacteria with a dominant codon composition that fits these genes could be potential “favorable hosts” for the phage. This hypothesis should be ultimately tested on a large scale in phages with a broad host range, and this will be done elsewhere.

Investigation of promiscuous phages demands an understanding of the role of encoded tRNAs in phage genomes. Among the 116 phages, only 20 of them encode tRNAs, and the number of tRNA genes varies considerably from phage to phage, ranging from 0 to 30 as reported in Supplementary Table 1. Number of genes and number of tRNAs have $rP = 0.56$, implying that longer phage genomes do not “typically” encode more tRNAs, even though a tendency is observed. Phages encoding more than 10

tRNAs have a fitness $<30\%$, and those displaying a fitness $>30\%$ encode either no tRNA at all (usually so) or a very small number (<5 , in rare cases). The *Vibrio* phage KVP40 represents a main exception to the second group, with 29 tRNAs and 63% highly biased genes. Its large collection of tRNAs does not especially supplement *V. cholerae* isoacceptor tRNA species that are present in minor amounts or recognize codons that occur more frequently in the phage genes. In fact, this phage displays a trend to greater use of *V. cholerae* codons or no apparent codon preference in its genes, and the large number of encoded tRNA species seems to reflect mostly its adaptation to a broad host range (Miller et al. 2003). There are two more noteworthy closely related (Hatfull et al. 2006) *Mycobacterium* phages, Bx1 and Catera virus, which display almost 30% highly biased genes and about 30 tRNAs. Bx1 is known to have multiple fast-growing hosts (Lee et al. 2004) and it might be affected by broad host-range adaptation as KVP40. (For an analysis of tRNA content in phages see Bailly-Bechet et al. [2007]. See also Nolan et al. [2006] on the functional role of tRNA genes in coliphages claimed to remain unclear at present.)

Fitness of Codon Composition and Bias Strength in Phages

Bias strength in bacteria is defined with respect to a reference set of very biased proteins, like ribosomal proteins, and numerical methods evaluating the strength are based on this reference set (Carbone et al. 2005; Sharp et al. 2005). For phages, no such reference set of genes has been identified and these methods cannot be applied. Also, a meaningful statistical analysis of single genomes is unfeasible when the genome is small (e.g., displaying ~ 10 genes) and no solid conclusion on gene composition could be drawn for small phages.

In this paper, we study the bias fitness of a phage genome versus the host genome, and we compute the tendency of a phage to use preferred host codons. SCCI values reflect codon composition in phage genes relative to host codon composition, and provide a numerical index of the advantage taken by the phage gene once translated in the host cell. The translational advantage offered by the bacterial environment is higher when the phage gene codon composition is biased toward the host dominant codon bias.

Our analysis points out that a set of genes belonging to specific functional classes, such as capsid and tail proteins, can be identified across phage species as being (almost) always (highly) biased. We propose these genes to play the same role for phages that reference sets containing ribosomal proteins play for bacteria. Their characterization

would provide a way to define “bias strength” for phages. For instance, a high (low) average of z -scores computed for those genes identified to be especially biased in phages (where a z -score of a gene g is $(\text{SCCI}(g) - \mu_{\text{SCCI}(\text{phage})}) / \sigma_{\text{SCCI}(\text{phage})}$) might provide a numerical criterion to determine a strong (weak) phage bias strength, as previously done for bacteria (see “ribosomal criterion” under Materials and Methods [Detection of Weak and Strong Forms of Translational Bias for Bacteria] and Carbone et al. [2005]). Notice that by using this measure of bias strength, most phage genomes would become strongly biased since their $\mu_{\text{SCCI}(\text{phage})}$ is rather far away from the SCCI of their capsid proteins (Supplementary Table 1). A larger number of available annotated phages is necessary to properly define reference sets and evaluate the proposed measures.

Viral-Host Gene Exchange and Codon Bias

The vast majority of phage and viral genes are unique to families of viruses and not of their hosts (Villareal 2001), and about 90% of dsDNA phage genes are phage-specific (either they have no matches in microbial genomes or they are significantly more likely to have been drawn from phage genomes than from cellular genomes) (Liu et al. 2006). This led us to conclude that despite the known propensity of phages to capture and transduce fragments of host genomes, these processes have a relatively small impact on the phage gene repertoire and that the majority of phage genes follow their own evolutionary trajectories. Our result states that in the case of phages with translationally biased hosts, such evolutionary trajectories are highly controlled by codon adaptation. Also, if codon adaptation plays a filtering role during gene acquisition, then gene acquisition from bacteria or other bacteriophages should be expected to be slow compared to the dissemination rate within the bacteriophage population once acquired. This conclusion is in agreement with the analysis done by Pedulla et al. (2003) for mycobacteriophages.

Complexity of Phage Organization and Environmental Pressures

At the genetic level, studies on the mosaicism of phages brought to light that conventional taxonomy fails to correctly represent the reticulate relationships of tail phages due to LGT, and indeed that it misrepresents most of them. Evidence for exchange of large blocks of genes, up to half of the genome, led to the observation that the new functional phages produced are substantially different in genome organization and biological properties from the

inferred parents (Lawrence et al. 2002). In the context of dairy phages the same problem has been observed (Proux et al. 2002). Pedulla et al. (2003) showed that mycobacteriophages cannot be phylogenetically ordered into any single hierarchical relationship. All these studies argue that the biology is not captured by the taxonomy. See Hendrix (2003) for a review.

Also, for dsDNA phages, and therefore for phages and viruses in general, there is no common molecule, such as the 16S rRNA gene for cellular microorganisms, to classify species (Hendrix et al. 1999). However, it might be possible to use conserved or core genes as genetic markers for different viral groups. Sequence differences of such genes would correspond to genotypes and they could be used as an estimation of viral species differences (Weinbauer and Rassoulzadegan 2004). Rohwer and Edwards (2002) proposed a hierarchical phage taxonomy based on the number of genes shared between pairs of phages, and Liu et al. (2006) proposed phage grouping based on shared gene repertoire. Even though our study does not concern the phylogenetic reconstruction of phages, our methodology embraces the latter view, and to study phage groups we consider the set of proteins shared by phages in the group, and not a single gene sequence.

Our distance trees are not phylogenies but rather phage space representations, where distances between phages represent evolutionary adaptation mediated by the host. We do not want to reconstruct ancestors here but, rather, to group phages with respect to their adaptive pressures, which are likely to be environmental pressures. In fact, it has been shown that codon bias for bacteria is highly related to environmental pressure (Carbone et al. 2005; Willenbrock et al. 2006) (and not to phylogeny), and it is reasonable to think that this character will be a fortiori forced on phages that need bacteria to live. In phage space, we represent a virus as the vector of SCCI values associated with a pool of shared genes. Since this information comes from codon analysis and nothing more, it is surprising to see that such limited information provides the same taxonomic organization derived from sequence-based phylogeny. This suggests that adaptation signals are stronger for phages than previously argued (Weinbauer and Rassoulzadegan 2004) and that bacterial environmental organization might influence phage organization at the phylogenetic level.

A Methodology to Analyze a Combinatorics of Phage Functional Preferences

A large number of striking similarities have been recorded between viruses that infect organisms belonging to

the different domains of life. They span morphological similarity, similar organization of the genome and replication processes, and similar capsid structures (Filée et al. 2003). On the other hand, when looking for mechanisms creating evolutionary diversity, we find that parallel adaptation, experimentally observed for ssDNA phages (Wichman et al. 1999, 2005; Rokyta et al. 2006), creates strong diversification within the same species, and that LGT among dsDNA phages (Lawrence et al. 2002) and among ssDNA phages (Rokyta et al. 2006) creates a mosaic spectrum of related phage species (Weinbauer and Rassoulzadegan 2004) possibly hosted by different bacteria. Methodologies to quantitatively study differences in phage spaces are being sought. The methods for comparing phages that we propose allow zooming in on species sharing homologous proteins, possibly accepted by different hosts, and analyze their patterns of codon bias evolution. The number of shared proteins need not be very high and this makes our numerical approach interesting for applications. We introduce two distance measures of codon space, the ℓ_1 -measure and the correlation measure, the first characterizing differences in codon bias between phages and reflecting environmental evolutionary pressures, and the second identifying phage functional patterns of evolution. The ϕ MH2K phage world and the Rokyta sequences discussed above are insightful examples illustrating a niche of the complex mechanisms regulating phage evolution and they are useful to test the mathematical approach. The analysis of the 47 Microviridae phages suggests that, in viral competition, viruses whose capsid proteins or tail proteins are ameliorated in their DNA coding by mutating in favor of codons which are most biased for the host genome are those that prevail. The number of evolutionary directions depends on combinations of genes that need to ameliorate together and the difficulty of predicting such directions for phages is based on the fact that the hypothesis of functional improvement is at the base of gene combinations. In small phages, such as the Microviridae phages, this phenomenon is simpler to observe because of the limited number of genes in the genome and the limited number of pressures for survival (e.g., it is reasonable to think that capsid proteins should be produced rapidly)

If the limited directionality in phage parallel evolution (Wichman et al. 1999, 2005; for a review see also Wood et al. 2005) that we observe is due to a limited variety of phage functional needs, then phage classification based on codon bias might suggest a way to meaningfully organize metagenomic data.

Acknowledgments The authors thank Julie Baussand for remarks and help with R and Hervé Isambert for discussions.

References

- Akashi H (2001) Gene expression and molecular evolution. *Curr Opin Genet Dev* 11:660–666
- Bailly-Bechet M, Vergassola M, Rocha E (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res* 17(10):1486–1495
- Bennetzen JL, Hall BD (1982) Codon selection in yeast. *J Biol Chem* 257:3026–3031
- Beres SB, Sylva GL, Barbian KD et al (2002) Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci USA* 99(15):10078–10083
- Brentlinger KL, Hafenstein S, Novak CR, Fane BA, Borgon R, McKenna R, Agbandje-McKenna M (2002) Microviridae, a family divided: isolation, characterization, and genome sequence of ϕ MH2K, a bacteriophage of the obligate intracellular parasitic bacterium *Bdellovibrio bacteriovorus*. *J Bacteriol* 184(4):1089–1094
- Carbone A (2006) Computational prediction of genomic functional cores specific to different microbes. *J Mol Evol* 63(6):733–746
- Carbone A, Madden R (2005) Insights on the evolution of metabolic networks of unicellular translationally biased organisms from transcriptomic data and sequence analysis. *J Mol Evol* 61:456–469
- Carbone A, Zinovyev A, Képès F (2003) Codon Adaptation Index as a measure of dominating codon bias. *Bioinformatics* 19:2005–2015
- Carbone A, Képès F, Zinovyev A (2005) Codon bias signatures, organization of microorganisms in codon space and lifestyle. *Mol Biol Evol* 22:547–561
- Coetzee JN (1987) Bacteriophage taxonomy. In: Goyal SM, Gerba CP, Bitton G (eds) *Phage ecology*. Wiley and Sons Interscience, New York, pp 45–86
- Corpet F (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* 16(22):10881–10890
- Cowe E, Sharp PM (1991) Molecular evolution of bacteriophages: discrete patterns of codon usage in T4 genes are related to the time of gene expression. *J Mol Evol* 33(1):13–22
- Fiers W, Sinsheimer RL (1962) The structure of the DNA of bacteriophage ϕ X174. III. Ultracentrifuge evidence for a ring structure. *J Mol Biol*, 5:424–434
- Filée J, Forterre P, Laurent J (2003) The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Res Microbiol* 154:237–243
- Gomathi NS, Sameer H, Kumar V et al (2007) In silico analysis of mycobacteriophage Che 12 genome: characterisation of genes required to lysogenise *Mycobacterium tuberculosis*. *Comput Biol Chem* 31(2):82–91
- Gouy M (1987) Codon contexts in enterobacterial and coliphage genes. *Mol Biol Evol* 4:426–444
- Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* 8:r49–r62
- Gupta SK, Bhattacharyya TK, Ghosh TC (2004) Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. *J Biomol Struct Dynam* 21:527–536
- Hatfull GF, Pedulla ML, Jacobs-Sera D et al (2006) Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genet* 2(6):e92, (online, doi: 10.1371/journal.pgen.0020092)
- Hendrix R (1999) The long evolutionary reach of viruses. *Curr Biol* 9:9914–9917
- Hendrix RW (2003) Bacteriophage genomics. *Curr Opin Microbiol* 6:506–511

- Hendrix R, Smith M, Burns R, Ford M, Hatfull G (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci USA* 96:2192–2197
- Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2:13–34
- Kliman RM, Hey J (1994) The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* 137:1049–1056
- Kunisawa T, Kanaya S, Kutter E (1998) Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. *DNA Res* 5:319–326
- Lawrence JG, Hatfull GF, Hendrix RW (2002) Imbrolios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J Bacteriol* 184:4891–4905
- Lee S, Kriakov J, Vilcheze C, Dai Z, Hatfull GF, Jacobs WR Jr (2004) Bxz1: a new generalized transducing phage for mycobacteria. *FEMS Microbiol Lett* 241:271–276
- Liu J, Glazko G, Mushegian A (2006) Protein repertoire of double-stranded DNA bacteriophages. *Virus Res* 117:68–80
- Lü H, Zhao W-M, Zheng Y, Wang H, Qi M, Yu X-P (2005) Analysis of synonymous codon usage bias in *Chlamydia*. *Acta Biochim Biophys Sinica* 37:1–10
- Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* 222:851–856
- Miller ES, Heidelberg JF, Eisen JA, Nelson WC, Durkin AS, Ciecko A, Feldblyum TV, White O, Paulsen IT, Nierman WC, Lee J, Szczypinski B, Fraser CM (2003) Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J Bacteriol* 185(17):5220–5233
- Nolan JM, Petrov V, Bertrand C, Krisch HM, Karam JD (2006) Genetic diversity among five T4-like bacteriophages. *Virology* 330(3) (doi: [101186/1743-422X-3-30](https://doi.org/10.1016/j.virus.2005.11.001))
- Pedulla ML, Ford ME, Houtz JM et al (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell* 113:171–182
- 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
- Rohwer F, Edwards R (2002) The phage proteomic tree: a genome-based taxonomy for phage. *J Bacteriol* 184(16):4529–4535
- Rokyta DR, Burch CL, Caudle SB, Wichman HA (2006) Horizontal gene transfer and the evolution of microvirid coliphage genomes. *J Bacteriol* 188(3):1134–1142
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocumbe PM, Smith M (1977) Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature* 265:687–695
- Sau K, Gupta SK, Sau S, Ghosh TC (2005) Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: implication in phage therapy. *Virus Res* 113(2):123–31
- Sharp PM (1986) Molecular evolution of bacteriophage: evidence of selection against recognition sites of host restriction enzymes. *Mol Biol Evol* 3(1):75–83
- Sharp PM, Li W-H (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acid Res* 15:1281–1295
- Sharp PM, Rogers MS, McConnell DJ (1985) Selection pressures on codon usage in the complete genome of bacteriophage T7. *J Mol Evol* 21(2):150–160
- Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F (1988) Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res* 16:8207–8211
- Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res* 33:1141–1153
- Shields DC, Sharp PM (1987) Synonymous codon usage in *Bacillus subtilis* reflects both traditional selection and mutational biases. *Nucleic Acids Res* 15:8023–8040
- Stenico M, Loyd AT, Sharp PM (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res* 22:2437–2446
- Summers WC (2001) Bacteriophage therapy. *Annu Rev Microbiol* 55:437–451
- Varenne S, Buc J, Llobès R, Lazdunski C (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol* 180:549–576
- Villarreal L (2001) Persisting viruses could play role in driving host evolution. *Am Soc Microbiol News* 67:501–507
- Weinbauer GM (2004) Ecology of procaryotic viruses. *FEMS Microbiol Rev* 28:127–181
- Weinbauer MG, Rassoulzadegan F (2004) Are viruses driving microbial diversification and diversity? *Environ Microbiol* 6:1–11
- Wichman HA, Badgett MR, Scott LA, Boulianne CM, Bull JJ (1999) Different trajectories of parallel evolution during viral adaptation. *Science* 285(5426):422–424
- Wichman HA, Millstein J, Bull JJ (2005) Adaptive molecular evolution for 13,000 phage generations: a possible arms race. *Genetics* 170(1):19–31
- Willenbrock H, Friis C, Friis AS, Ussery DW (2006) An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biol* 7:R114
- Wood TE, Burke JM, Rieseberg LH (2005) Parallel genotypic adaptation: when evolution repeats itself. *Genetica* 123:157–170