Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 5854–5859, June 1996
Evolution

Similarities and dissimilarities of phage genomes

B. EDWIN BLAISDELL†, ALLAN M. CAMPBELL‡, AND SAMUEL KARLIN†§

Departments of †Mathematics and ‡Biological Sciences, Stanford University, Stanford, CA 94305-2125

Contributed by Allan M. Campbell, February 22, 1996

ABSTRACT **Genomic similarities and contrasts are investigated in a collection of 23 bacteriophages, including phages with temperate, lytic, and parasitic life histories, with varied sequence organizations and with different hosts and with different morphologies. Comparisons use relative abundances of di-, tri-, and tetranucleotides from entire genomes. We highlight several specific findings. (*i*) As previously shown for cellular genomes, each viral genome has a distinctive signature of short oligonucleotide abundances that pervade the entire genome and distinguish it from other genomes. (*ii*) The enteric temperate double-stranded (ds) phages, like enterobacteria, exhibit significantly high relative abundances of GpC = GC and significantly low values of TA, but no such extremes exist in ds lytic phages. (*iii*) The tetranucleotide CTAG is of statistically low relative abundance in most phages. (*iv*) The DAM methylase site GATC is of statistically low relative abundance in most phages, but not in P1. This difference may relate to controls on replication (e.g., actions of the host SeqA gene product) and to MutH cleavage potential of the *Escherichia coli* DAM mismatch repair system. (*v*) The enteric temperate dsDNA phages form a coherent group: they are relatively close to each other and to their bacterial hosts in average differences of dinucleotide relative abundance values. By contrast, the lytic dsDNA phages do not form a coherent group. This difference may come about because the temperate phages acquire more sequence characteristics of the host because they use the host replication and repair machinery, whereas the analyzed lytic phages are replicated by their own machinery. (*vi*) The nonenteric temperate phages with mycoplasmal and mycobacterial hosts are relatively close to their respective hosts and relatively distant from any of the enteric hosts and from the other phages. (*vii*) The single-stranded RNA phages have dinucleotide relative abundance values closest to those for random sequences, presumably attributable to the mutation rates of RNA phages being much greater than those of DNA phages.**

In previous publications (1–4), data and analyses were presented supporting the hypothesis that the 16-dinucleotide relative abundance values of genomes provide a robust unique genomic signature for prokaryotic and eukaryotic sequences. Genomic sequences are compared with respect to: (*i*) di-, tri-, and tetranucleotide relative abundance extremes; (*ii*) distances between dinucleotide relative abundance values (profiles) for each pair of genomes; and (*iii*) consistent similarity of individual sequences to various consensus sequences. In many cases, distances calculated from dinucleotide relative abundances correlate well with traditional phylogenies (1, 2). Dinucleotide relative abundance values are essentially equivalent to the "general designs" derived from nearest-neighbor frequency analysis evaluated extensively during the 1960s and 1970s in samples of genomic DNA from many organisms (5, 6). Our recent studies have demonstrated that the dinucleotide relative abundance profiles of different DNA sequence samples from the same organism are generally much more similar to each other than they are to profiles from other organisms and that closely related organisms generally have more similar profiles than do distantly related organisms (1–4, 7). These highly stable DNA-doublet patterns suggest that there may be genome-wide factors, such as functions of the replication and repair machinery that impose limits on the compositional and structural patterns of a genomic sequence, and that the set of dinucleotide relative abundance values constitutes a genomic signature that reflects the influence of these factors. In cellular organisms the existence of genomic signatures and their correlation with phylogeny may reflect either the evolutionary development of the causative factors or a very slow rate of change in response to any alterations in them.

We have studied the similarities and differences between the DNA sequences of 23 phages for which complete genomes are available or for which many segments of sufficient aggregate length have been sequenced. These phages possess diverse characteristics including: double-stranded (ds) or single-stranded (ss); DNA or RNA; lytic, temperate, or parasitic life cycles; icosahedral or filamentous capsids; a considerable range of G+C content; a large variety of morphological classifications; and some variety in bacterial hosts although most are *Escherichia coli* (Table 1).

The signatures of bacteriophage genomes have a special interest because there is no accepted molecular taxonomy of phages. Models for phage evolution frequently postulate a chimeric origin for phage genomes (8). At least some lateral transfer between distantly related phages or from host to infecting phage is known (9). Despite these facts, the results in this paper show that the dinucleotide relative abundance profile provides a genomic signature that pervades each phage genome. The following questions are also of interest. To what extent are morphology, life cycle, virion organization, and genomic replication properties reflected in the genomic sequence similarities and contrasts? To what extent are the various bacteriophage genomes similar to or different from their bacterial host genomes or to other possible hosts? Which phage genomes have the most (or least) bias in dinucleotide relative abundances? Finally, to what extent do temperate, lytic, and filamentous phages define coherent groups?

DATA

The data include every phage having a complete genomic sequence and several nonredundant sets of sequences of sufficient aggregate length that were available in GenBank as of July 1994. Very similar replicates were removed. For example, among complete genomes, PX1 (= ϕX174) and S13 are ≈98% identical, and only PX1 is retained. Other similar selections are shown in the legend of Table 1. The data also include dinucleotide relative abundance profiles for collections of appropriate bacterial sequences as listed in the legend of Table 1.

METHODS

The statistical methods of this paper were elaborated in previous publications (1, 3) and are applied in this paper to compare diverse bacteriophage and bacterial host sequences.

Oligonucleotide Relative Abundances. A standard (single strand) measure of dinucleotide bias is the odds ratio $\rho_{XY} = f_{XY}/f_X f_Y$, where f_X denotes the frequency of the nucleotide X and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ds, double-stranded; ss, single-stranded.

§To whom reprint requests should be addressed.

Table 1. Phage sequences studied

Genome name	G + C, % × 10	Length, bp	Host
tmp* enteric dsDNA			
LAM†	498	48,502‡	Eco
MU	494	11,997	Many
P1	428	20,825	Many
P4	495	11,624‡	Eco
P22	474	30,002	Sty
tmp nonenteric dsDNA			
L5	632	52,297‡	Msm
L2	320	11,965‡	Ala
lyt* dsDNA			
T2	378	8,883	Eco
T4	357	168,903‡	Eco
T3	498	26,457	Eco
T7	484	39,936‡	Eco
PZA	397	19,366‡	Bsu
PRD	481	14,925‡	Many
prod* ssDNA			
f1	409	6407‡	Eco
I22	427	6744‡	Eco
IKE	406	6883‡	Eco
PF3	454	5833‡	Pae
Lytic ssDNA			
PX1†	448	5386‡	Eco
G4	457	5577‡	Eco
CP1	365	4877‡	Cps
RNA			
ϕ6 (ds)	559	13,286	Psy
GA (ss)	479	3466‡	Eco
MS2 (ss)	521	3569‡	Eco

Eco, *E. coli*; Bsu, *Bacillus subtilis*; Pae, *Pseudomonas aeruginosa*; Sty, *Salmonella typhimurium*; Cps, *Chlamydia psittaci*; Psy, *Pseudomonas syringae*; Ala, *Acholeplasma laidlawii*; and, Msm, *Mycobacterium smegmatis*. Of the very similar PX1 and S13, only PX1 is kept; of f1, M13 and fd, only f1 is kept; and of PZA and φ29, only PZA is kept.
*tmp, temperate; lyt, lytic; and prod, parasitic or filamentous (fil).
†LAM, λ; and PX1, GenBank name for ϕX174.
‡Signifies a complete genome sequence.

f_{XY} is the frequency of the dinucleotide XY. However, because DNA properties are influenced by oligonucleotide compositions of both strands, the formula for ρ_{XY} is modified by combining the given sequence and its inverted complement sequence. In this way, the frequency f_A is symmetrized to $f^*_A = f^*_T = (f_A + f_T)/2$ and $f^*_C = f^*_G = (f_C + f_G)/2$. Similarly, $f^*_{GT} = f^*_{AC} = (f_{GT} + f_{AC})/2$, etc. A symmetrized dinucleotide relative abundance measure is $\rho^*_{GT} = \rho^*_{AC} = f^*_{GT}/f^*_G f^*_T = 2(f_{GT} + f_{AC})/(f_G + f_C)(f_T + f_A)$, etc. The deviation of ρ^*_{GT} from 1 can be construed as a measure of the dinucleotide bias of GT/AC. Corresponding trinucleotide and tetranucleotide measures are γ and τ, respectively (1). Table 2 shows the dinucleotide relative abundance ρ^* vectors (also referred to as the $\{\rho^*_{XY}\}$ profiles) of all the organisms of Table 1, of three consensus sequences, and of eight host bacteria.

Consensus ρ^* Values. For natural groups of phage sequences (e.g., enteric temperate dsDNA {LAM, Mu, P1, P4, P22}, filamentous {F1, IKE, I22, PF3}, and lytic dsDNA {T2, T4, T3, T7, PZA}), we calculate a consensus dinucleotide relative abundance ρ^* vector by averaging the f^*_{ij} and f^*_i values of the relevant sequences. For example, let tmp denote the enteric temperate consensus phage. We calculate

$$f^*_{ij}(tmp) = \frac{1}{5}[f^*_{ij}(\mathrm{LAM}) + f^*_{ij}(\mathrm{MU}) + f^*_{ij}(\mathrm{P1}) + f^*_{ij}(\mathrm{P4}) + f^*_{ij}(\mathrm{P22})] \quad [1]$$

Table 2. Symmetrized dinucleotide relative abundances (ρ^*)

	CG	GC	TA	AT	CC GG	AA TT	AC GT	AG CT	CA TG	GA TC
tmp enteric dsDNA										
LAM	103	120	71	109	94	115	88	87	116	98
MU	108	124	69	102	96	125	91	81	115	90
P1	98	124	79	100	100	117	86	91	110	96
P4	104	122	74	106	101	124	85	85	110	93
P22	98	126	76	106	92	112	83	96	113	100
tmp nonenteric dsDNA										
L5	108	91	51	96	86	90	105	104	104	132
L2	59	107	92	99	108	103	98	98	118	95
lyt dsDNA										
T2	104	114	89	98	109	113	92	90	102	94
T4	94	117	82	99	102	112	85	95	106	104
T3	90	100	84	86	91	101	104	110	110	106
T7	88	95	88	85	91	97	107	112	108	107
PZA	92	88	86	93	98	105	108	95	112	102
PRD	125	154	91	112	116	146	71	66	96	64
fil ssDNA										
f1	88	119	85	93	112	121	85	98	98	95
I22	101	133	89	97	97	115	88	97	105	90
IKE	99	127	83	96	100	116	89	93	108	93
PF3	80	117	68	92	106	123	88	94	118	93
lyt ssDNA										
PX1	99	129	76	93	90	118	88	98	111	97
G4	99	108	82	95	93	109	95	100	107	104
CP1	81	110	86	93	97	108	80	118	94	116
RNA										
ϕ6	113	101	67	94	83	104	103	99	106	117
GA	104	102	89	99	103	111	91	98	95	105
MS2	110	96	95	94	98	106	100	100	92	107
Consensus										
tmp	103	123	74	105	97	119	86	88	112	95
fil	92	124	82	95	104	119	88	95	107	93
lyt	94	103	87	94	98	108	98	99	106	101
Host										
Eco	116	127	76	110	90	121	89	83	111	93
Sty	123	129	82	114	91	122	85	82	103	93
Bsu	104	126	65	101	96	123	76	92	108	107
Pae	109	117	58	111	87	114	87	99	109	108
Cps	81	120	79	83	99	119	83	115	98	105
Msm	124	100	45	116	81	95	111	83	108	124
Ala	76	108	89	88	109	112	100	101	109	89
Psy	105	121	64	108	88	115	88	93	115	101

All values multiplied by 100. See legends to Tables 1 and 5. tmp, consensus of five temperate dsDNA phages; fil, consensus of four filamentous ssDNA phages; and lyt, consensus of five lytic dsDNA phages excluding PRD, which is an extreme outlier (see Table 3).

and similarly for f^*_i(tmp) and define ρ^*_{ij}(tmp) = f^*_{ij}(tmp)/f^*_i(tmp)f^*_j(tmp). A random sequence would have each of its dinucleotide relative abundances about 1.00, so we designate the 16-component vector with all unit values as equ.

Distances Between ρ^* Vectors. We calculate the pairwise absolute distances between the phage g and h ρ^* vectors, (Table 3).

$$\delta^*(g,h) = \frac{1}{16}\sum_{j=1}^{16}|\rho^*_j(g) - \rho^*_j(h)| \quad [2]$$

RESULTS AND DISCUSSION

A Genomic Signature of Dinucleotide Relative Abundance Values. Our methods of genomic comparisons based on short oligonucleotide (mainly dinucleotide) relative abundance values can be applied to sets of genomic sequences for which common functional gene sequences are not available—e.g., to a broad range of eukaryotic viruses—to diverse sets of protist

Table 3. Absolute distances (δ^*) of dinucleotide relative abundance (ρ^*) vectors of 23 phages

MU	P1	P4	P22	L5	L2	T2	T4	T3	T7	PZA	PRD	
46	43	43	43	185	121	74	67	131	155	109	228	LAM
*	63	40	79	205	140	88	99	160	184	135	200	MU
	*	36	41	190	107	57	37	120	145	95	214	P1
		*	62	210	132	69	68	153	178	126	189	P4
			*	175	116	89	46	108	132	94	239	P22
				*	193	184	165	112	108	132	377	L5
					*	88	99	106	122	88	283	L2
						*	62	130	143	101	199	T2
							*	91	110	73	240	T4
								*	24	58	329	T3
									*	65	342	T7
										*	300	PZA

f1	I22	IKE	PF3	PX1	G4	CP1	ϕ6	GA	MS2	
90	69	55	71	51	78	152	129	97	132	LAM
103	77	64	72	75	107	184	149	122	151	MU
56	45	20	62	36	68	129	134	79	120	P1
68	69	47	62	65	101	159	154	103	146	P4
83	60	55	85	36	60	123	119	87	122	P22
204	185	188	213	167	125	169	73	153	124	L5
113	112	104	86	119	90	148	140	100	115	L2
48	54	47	79	84	75	134	133	54	97	T2
63	56	44	75	60	42	94	109	49	90	T4
144	126	117	141	100	58	108	77	99	84	T3
160	138	138	163	124	77	111	92	111	87	T7
118	101	92	115	97	60	128	89	84	70	PZA
185	203	212	239	243	273	286	321	233	272	PRD
*	59	56	69	70	91	114	152	62	104	f1
	*	30	85	47	70	125	132	71	103	I22
		*	63	40	64	131	132	74	112	IKE
			*	71	104	146	154	111	153	PF3
				*	57	130	108	88	117	PX1
					*	97	72	52	62	G4
						*	130	91	98	CP1
							*	102	80	ϕ6
								*	48	GA

All values multiplied by 1000.

or bacterial sequences (1, 3) and to bacteriophages as we have done here. The traditional means of assessing similarities between organisms is to examine aligned sequences of homologous genes by various tree-building algorithms. In the case of the 23 phages at hand, it is neither possible to find highly similar genes common to all of them, or even to most pairs, nor to make reasonable alignments of them, even pairwise. Conservation of the order of genes coding for proteins of similar function has been used as evidence for the similarity of blocks of contiguous genes in phages. However, different sets of such blocks are conserved in different sets of phages so that no overall similarity of whole genomes is realized (10).

Previous work has shown that dinucleotide relative abundances provide a unique signature for each genome, which is observed in all segments ≥20 kb of that genome (1, 2, 4). In fact, the δ^* distances (see *Methods*) between dinucleotide relative abundance values of long samples from an organism are almost always less than the corresponding distances between similarly long samples from different organisms. This relation has been found to be generally true for a wide variety of organisms (vertebrate, invertebrate, fungus, protist, and prokaryote) and is here confirmed for a variety of phages (see Table 6). We highlight below some of the main results emanating from our dinucleotide relative abundance comparisons and some interpretations of them.

Extreme Relative Abundances of Dinucleotides (Table 4). The dinucleotide TA is significantly low in all the enteric temperate dsDNA phages and in the nonenteric temperate phage L5 but normal in the lytic dsDNA phages. The three parasitic ssDNA phages of *E. coli* generally carry TA in the low normal range, while TA is significantly low in PF3, a phage of *P. aeruginosa*, and in the lytic ssDNA phage PX1. The temperate and parasitic phages but not lytic dsDNA coliphages generally exhibit significantly high relative abundances of the dinucleotide GC (not CG, which is, among the phages, low only in L2 and among the hosts, significantly low only in its mycoplasmal host). Notably, sequences of γ-proteobacteria infecting mammalian hosts (e.g., *E. coli, S. typhimurium, Haemophilus influenzae,* and *Klebsiella pneumoniae*) but not of soil-dwelling γ-proteobacteria (e.g., *P. aeruginosa* and *Azotobacter vinelandii*) are significantly high in GC relative abundances (4).

Enteric Temperate Versus Lytic dsDNA Phages. The set of enteric temperate dsDNA phages and the set of lytic dsDNA phages both show generally significantly low relative abundances for CTAG and GATC. On the other hand, CGCG is low for all the lytic but none of the enteric temperate; CCCC/GGGG is low or low-normal for all the lytic but high-normal for all the enteric temperate. For δ^* distances, the five enteric temperate phages are all close to the host *E. coli*, whereas lytic phages are relatively distant. Nonenteric temperate phages are much closer to their respective hosts than to the enteric hosts and are closer than the dsDNA lytic phages are to their hosts (Table 5).

Low Relative Abundance of the DAM Site GATC in Coliphages. GATC is low in all enteric dsDNA phages except P1 and tends to be lower in lytic coliphages than in temperate coliphages (Table 4). McClelland (11) has noticed that GATC is low in some phages but not in prokaryotic chromosomes (e.g., normal in *E. coli*). This has been attributed to selection against it in phages that would otherwise occasionally be cut at unmethylated GATC sites by the MutH activity of the host (12). This cutting is prevented in P1 by its own EcoP1 protein that methylates GATC (13). The life history of P1 as a single-copy plasmid requires elaborate machinery to stop P1 replication at one additional copy (sequestration) and to ensure that one copy goes to each daughter cell at host cell division (partition). Important to this control is the state of methylation of four GATC sites embedded in the heptad AGATCC(C/A) repeats near *oriR*, where the host sequestration protein SeqA can bind at hemimethylated GATC (14). It is reasonable to guess that P1 developed its own GATC methylase to assure methylation of this vital cluster of GATC.

Relative abundances of GATC in dsDNA phages are, except in PZA, consistently lower in lytic dsDNA phages than in dsDNA temperate phages. GATC is much lower in phages T3 and T7 than in phages T2 and T4. Also, except in PZA, GAAC/GTTC and GACC/GGTC, both differing from GATC by a single substitution, are higher in lytic phages than in temperate phages and are higher in T3 and T7 than in T2 and T4. While integrated into the host genome, the GATCs of temperate dsDNA phages are methylated by the host DAM methylase, whereas lytically replicating dsDNA phage DNA may not be. In this context, protection against MutH cutting is attained by severely reducing the number of GATC occurrences in lytic phages and to a lesser extent in the temperate phages. The T2 and T4 phages contain their own DAM methylases that are different from those of the host *E. coli* and all cytosine bases are converted to hydroxymethylcytosine (frequently glycosylated). These modifications may also inhibit host MutH cutting of GATC. On the other hand, the T-odd phages have no means of specific protection of GATC and presumably compensate by maintaining very low levels of occurrence. The host for PZA is *B. subtilis*, which does not possess a DAM methylase and may not possess an analog of the MutH endonuclease that recognizes GATC. Thus, PZA may not need to reduce its GATC level to that found in the other lytic dsDNA coliphages.

Table 4. Extreme symmetrized relative abundances

Phage	TA	AA*	GC	CCC*	CTA*	CCA*	CGC*	CCCC*	CGCG	CCGG	CTAG	GATC	TATA	TCGA	CCGC*	CTAC*	GAAC*	GACC*	GCCC*
Enteric tmp dsDNA																			
LAM	71		(120)		68	125					52	72							
MU	69	125	123		64	128						63	58	78					
P1	(79)		124								39								
P4	73	123	(122)			(120)						47					(121)		
P22	76		126		(78)	(121)					33	54					(122)		
Nonenteric tmp dsDNA																			
L2									55			76			123		126		
L5	51				(79)						47		53						
lyt dsDNA																			
T2				69		135		70	74		(81)	26						(122)	123
T4						125			71		(81)	48		77			123	(120)	
T3				(80)		(120)			70		52	12	64			123	125	132	
T7				78		124		71	75		57	5					127	128	
PZA								67	11	54		78			142				
PRD		146	154								60	8		59		134	167	147	
Parasitic ssDNA (filamentous)																			
f1		(120)					125	63			26	29	75	00				132	
I22			133			125				58	52		73		123				
IKE			127	71	(80)	125	124	68					73		129				136
PF3	68	123					(121)	70	65		63					125			123
lyt ssDNA																			
PX1	76		129	69	76	132	(120)			65	37	00	67	74				127	
G4										66	29	17	71	76		124		126	
CP1				37		127	135	00		37			77		166			(121)	134
RNA																			
ϕ6	67				66						43					(122)			
PGA					(80)					50									
MS2													73						

All values multiplied by 100.

*Symmetrized value for oligonucleotide and its inverted complement. Significance criteria were low <78 and high >122. Values are shown for di if there was more than one occurrence, tri more than 2, and tetra more than 3. Some values (in parentheses) nearly satisfying the criteria are shown when they occur in accepted blocks.

Low Relative Abundance of CTAG. CTAG is broadly low in Gram-negative proteobacterial sequences (for example, see ref. 15). Interpretations center on structural defects (kinking) or special functional roles associated with this tetranucleotide (15). CTAG is generally low in all classes of our phages. However, CTAG is not low in three dsDNA phages: P4, MU, and PZA ($\tau^* = 0.93, 0.97$, and 1.10, respectively). We speculate that CTAG has an important but as yet undetermined regulatory function in these phages. The CTAG sites tend to occur in small clusters in each of these phages, perhaps as binding sites for regulatory proteins.

Restriction Avoidance. The low values for palindromic tetranucleotides in Table 4 may reflect to some extent restriction avoidance by the various phages (16, 17). Supporting this hypothesis is the observation that all the extreme high relative abundance tetranucleotides can be obtained from one of the extreme low palindromic tetranucleotides via a single base substitution. For example, high relative abundances for GAAC/GTTC and GACC/GGTC are paired with low relative abundances of GATC in the same phages. Similarly, the high values of CTAC/GTAG are paired with low values of CTAG. The extreme occurrences of CGCG and of GACC/GGTC are almost exclusively associated with lytic phages. CGCG is also low and CCCG/CGGG is high in L2, which, although eventually lysogenic, is initially noncytocidally productive (18). TATA is low and TAAA/TTTA is high in L5, that is phenotypically temperate but has its own DNA polymerase, a characteristic generally of lytic phages (19). The tetranucleotide restriction sites in *B. subtilis*, CCGG, CGCG, and GGCC ($\tau^* = 0.08$), are much lower in its phage PZA than in any other lytic dsDNA phage and are accompanied by many high tetranucleotides differing from them by one substitution (data not shown). These facts are consistent with the principle of restriction avoidance.

Dinucleotide Relative Abundance Distances Between Phages. Pairwise distances (see *Methods*) between phages are shown in Table 3 and phage distances to some consensus phages and bacteria are shown in Table 5. The enteric temperate dsDNA phages are among the closest to *E. coli* and are generally closest to each other (in the range of distances of human to mammals, see (2)). The smallest δ^* distances to *E. coli* are attained by MU ($\delta^* = 0.043$), P4 ($\delta^* = 0.041$), and LAM ($\delta^* = 0.045$), and each of these is closer to *E. coli* than to any other bacterium in our sample. Also, moderately close to *E. coli* are P1 and P22. Similarly, temperate L2 is closer to its host than to any other bacterium and to any other phage, and temperate L5 is closer to its host than to any other bacterium and to any other phage except ϕ6. Also, CP1 is closer to its host than to any other bacterium and to any other phage (data not shown). CP1 is assumed to be lytic because of its similarity to PX1 and G4 (20), but our result suggests that it is phenotypically temperate. None of the lytic phages is as close to its host as is any of the enteric temperate phages to its host. Temperate coexistence apparently produces δ^* distances within the set and to the host that are smaller than those of the lytic phages.

In the distance data the five lytic dsDNA phages form three distinguishable subsets, T-even (T2 and T4), T-odd (T3 and T7), and PZA. They are also distinguished in morphology and G+C composition (Table 1). In extreme relative abundances, PZA is often different from the other four, possibly because it infects a different host with its different restriction systems.

Coherence of Temperate Phages. The five enteric temperate dsDNA phages all have in common some distinguishing extreme relative abundance properties not possessed by any of our lytic phage genomes—i.e., low TA and high GC (also true

Table 5. Absolute distances (δ*) of dinucleotide relative abundance vectors of 23 phages relative to four bacterial genomes and three consensus sequences

	Eco	Bsu	Sty	Msm	Ala	tmp*	lyt*	equ*
LAM	45	67	70	164	122	27	93	118
MU	43	72	65	184	129	42	119	144
P1	57	56	70	193	94	25	74	96
P4	41	59	56	189	115	23	107	127
P22	63	62	81	182	111	44	76	104
L5	199	178	211	96	186	194	134	131
L2	152	161	170	212	57	121	74	83
T2	88	102	87	195	67	66	70	81
T4	94	65	97	180	85	60	49	80
T3	147	140	164	153	92	136	61	75
T7	172	157	183	152	106	160	75	85
PZA	138	129	156	154	85	108	40	66
PRD	195	216	168	334	255	210	268	269
f1	98	91	89	235	89	68	87	99
I22	75	83	75	212	82	56	68	91
IKE	59	68	71	198	80	41	67	93
PF3	92	89	114	228	79	65	102	137
PX1	52	69	76	192	99	45	73	107
G4	102	88	113	155	76	84	27	61
CP1	179	129	170	220	129	145	96	121
φ6	137	124	154	89	133	138	77	95
GA	123	101	122	185	87	96	49	55
MS2	150	130	149	168	103	131	51	46

All values multiplied by 1000.
*tmp, consensus of {LAM, MU, P1, P4, P22}; lyt, consensus of {T2, T4, T3, T7, PZA}; and equ, a vector of 16 ones.

for *E. coli* and *S. typhimurium*). In δ* distances, they constitute the five single sequences closest to their consensus sequence. The major host *E. coli* is generally close to each of them and to their consensus. No other group of three or more phage sequences is as cohesive in its behavior. However, the enteric temperate group is diverse in a number of other properties. For example, the set contains exemplars of three capsid forms: short-tailed (P22), long-tailed (LAM), and contractile-tailed (MU, P1, and P4). Also, one occurs as an independent plasmid (P1), one can act as a transposon (MU), one infects *S. typhimurium* (P22), and two infect many hosts (P1 and MU). However, these differences are submerged beneath a similarity of relative abundances, probably because all are replicated and repaired by a common host machinery. The average δ* distance of *E. coli* to *S. typhimurium*, based on many 50-kb samples from each, is ≈0.040 (7). None of the temperate phages is this close to either host. The *Salmonella* phage P22 is farthest, but not very far, from both hosts, presumably for reasons other than its host preference.

Variation among the filamentous phages. The four filamentous parasitic ssDNA phages do not constitute a coherent group. They show substantial differences in relative abundance extremes and in δ* distances. These differences can be associated with morphology, life history, and host differences. In no instance do all four phages possess an extreme relative abundance of the same oligonucleotide. In extreme relative abundance PF3 often stands alone: low in TA, low in CGCG and high in CTAC/GTAG, and not high in GC and not low in TATA. PF3 infects *P. aeruginosa* and, in this host, the TA level is very low ($\rho^* = 0.60$) and the GC level is normal. In δ* distance, PF3 is farthest from the filamentous consensus. I22 and IKE, both of which attach to I pili, are very close (δ* = 0.030), and the next closest pairs are twice as distant (≥0.063). Phage f1, which attaches to F pili, is alone very low in GATC, very low in TCGA (in fact, absent), and high in GACC/GGTC and in CCGA ($\rho^* = 1.29$). These four phages also require variable numbers of host proteins for the three stages of RF form replication and for transcription stage 2.

Similarities and Differences of PX1 (= φX174) and G4. The two lytic ssDNA phages PX1 and G4 both have small icosahedral capsids with no tails, the same host range, approximately the same genome size and G+C content, and the same gene organization. The aligned amino acid sequences are, on average, 66% identical (21) but are substantially different in codon usage. They have six extreme palindromic tetranucleotides in common, more than any other pair of phages, including GATC, which is absent from PX1 and occurs only twice in G4. However, the two also show substantial differences: PX1 has five extreme di- and trinucleotide relative abundances, while G4 has none. Their mutual distance is moderate (δ* = 0.057). Several temperate and filamentous phages and their consensuses, but no lytic phages, are closer to PX1 than is G4. No temperate or filamentous phages is closer to G4 than PX1 is.

There are precedents for substantial amino acid identities associated with only a moderate level of genome similarities. For example, human herpes simplex virus I and varicella zoster virus show significant amino acid identities and highly concordant gene organization in the unique long region. However, herpes simplex virus I entails a genome G+C content of 68%, compared with varicella zoster virus with 46%. Among all vertebrate alpha herpes viruses, herpes simplex virus I and varicella zoster virus are relatively distant with genomic δ* distance = 0.081 (3). Human and mouse are often highly concordant in protein sequence comparisons but have genome δ* distance = 0.057, the same level as PX1 with G4. *E. coli* and *Shigella flexneri* are highly similar on the protein level but with moderate δ* distance (0.062; ref. 7).

Lytic ssRNA Phages GA and MS2 Are Closest to Random. Distances between computer-generated random sequences of length 10 kb fall in the range 0.000–0.018 (2). No phage sample in our collection is very close to random. The closest individual phage sequences to random in distances are the ssRNA MS2 (0.046) and GA (0.055). This observation is consistent with the fact that these two phages show much fewer extreme dinucleotide relative abundances than do any other phages. These results are in accord with the indication that mutation rates of ssRNA are 10,000 times those of dsDNA (22) and that sequences tend to approach randomness with increasing numbers of substitutions. Among the most nonrandom are all the enteric temperate phages (Table 5). This is consistent with the finding that the enteric temperate dsDNA phages are among those having the most extreme oligonucleotides (Table 4).

Phages That Are Far from Most Other Phages. The set of phages {CP1, φ6, L2, L5, MS2, T3, T7, PRD} are generally farthest from any other phage not in this set. It is interesting that CP1, φ6, L2, and L5 are phages of distinct hosts different from *E. coli*. MS2 is the phage closest to random and, therefore, different from the other phages, which generally have their own characteristic profiles. T3 and T7 infect *E. coli*, and PRD infects many bacteria including *E. coli*, but they are lytic and use their own, presumably characteristic, DNA polymerases. PRD is generally farthest from any of the other phages, which is consistent with its having four extreme oligonucleotides possessed by no other phage. PRD is unique among these 23 phages in having lipid inside the capsid.

Distances Between 20-kb Samples of Four Phage and Three Bacterial Sequences. Karlin *et al.* (1, 2) found that distances between ρ* vectors based on 50-kb sample sequences from prokaryotic or eukaryotic species were generally considerably larger than distances between 50-kb samples taken from single

Table 6. Average, minimum, and maximum of absolute distances (δ^*) of dinucleotide relative abundance (ρ^*) vectors of ≈20-kb samples

Organism	Eco	Bsu	Sty	Msm	LAM	P22	L5	T4	T7
Eco	39	87	51	161	61	75	196	98	172
(n = 20)	9	48	10	132	25	48	128	37	104
	90	135	114	194	90	94	226	133	211
Bsu		39	96	183	76	66	178	73	160
(n = 20)		13	53	160	39	50	151	35	127
		87	153	202	106	86	198	115	189
Sty			36	174	82	89	214	105	187
(n = 20)			13	134	30	48	167	62	140
			75	214	120	115	252	133	227
Msm				39	162	182	96	182	155
(n = 2)				39	146	156	68	162	151
				39	177	207	120	204	161
LAM					31	46	179	68	150
(n = 3)					24	32	168	46	131
					36	66	191	104	169
P22						42	175	52	132
(n = 2)						42	152	32	106
						42	197	69	158
L5							33	165	109
(n = 3)							26	147	91
							39	188	121
T4								29	113
(n = 8)								13	95
								42	141
T7									37
(n = 2)									37
									37

All values multiplied by 1000.

species. Because of our phage sequence size limitations, we have taken samples of length ≈20 kb from the larger phage sequence sets: genomes T7 (2 samples), LAM (3 samples), L5 (3 samples), T4 (8 samples), and the sequence collection P22 (2 samples). We also randomly selected twenty 20-kb samples from each of the collections of *B. subtilis*, *E. coli*, and *S. typhimurium*, and two from *M. smegmatis*.

The ranges and averages within and between organism sample distances are shown in Table 6. Examination reveals that the average distance between organism samples always exceeds the corresponding within organism sample average distances.

Conclusion. The dinucleotide relative abundance measure of distance between DNA sequences appears to provide meaningful measures of similarities. We suggest that the short oligonucleotide (di-, tri-, and tetranucleotide) relative abundance values relate to DNA structures (4). Several factors that can impact on DNA structures have been identified—e.g., dinucleotide stacking energies, curvature, superhelicity, methylation and other short oligonucleotide modifications, context-dependent mutation biases, and DNA replication and repair mechanisms (4, 7).

The distances between relative abundance vectors of samples from the same genome are generally smaller than distances between samples from different genomes (Table 6). We reiterate the hypothesis (4) that the replication and repair machinery that processes the whole genome contributes most to maintaining the constancy and uniqueness of the relative abundance vectors of the genome of an organism and that evolutionary differences in this machinery can create evolutionarily meaningful genomic differences between different organisms. Among phages, the relative abundance values should therefore be strongly influenced by the extent to which host machinery is used and by the nature of the host. A comprehensive correlation of distance values with degree of host dependence requires more information than we have about the natural life styles and host preferences of the phages available for analysis. However, our results support a picture where those phages dependent on host replication machinery (temperate phages and, to a lesser extent, parasitic ssDNA phages) converge toward the DNA signatures of the host, whereas autologously replicating phages (T4 and T7) each diverge in their own direction.

Can similarity of temperate phages to their hosts arise through extensive incorporation of host genes? Whereas such incorporation certainly can and does occur, we consider this explanation unlikely. The genome signature depends on the totality of the DNA and cannot be significantly affected by the presence of a few recently acquired DNA segments. For example, in the case of λ and *E. coli* ($\delta^* = 0.046$), a search for long approximately matching common blocks (unpublished results) found three of total length 878 nt, or 1.8% of the length of λ. It is plausible that the signature has been imposed upon the viral sequence during its long residence in *E. coli*, where the replication and repair systems and context dependent mutational biases of the host act on it. We would expect that any foreign sequence replicated in *E. coli* (or any other host) for a long period of time should likewise approach the host signature. The phylogenetic significance of the δ^* distances may be low within the temperate phages, moderate to appreciable for the ssDNA phages, and substantial for the autologously replicating phages as it is for cellular organisms.

This study was supported in part by National Institutes of Health Grants 2R01GM10452-31, 5R01HG00335-07, and 9R01 GM51117-27, and National Science Foundation Grant DMS 9403553.

1. Karlin, S. & Cardon, L. R. (1994) *Annu. Rev. Microbiol.* **48,** 619–654.
2. Karlin, S. & Ladunga, I. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12832–12836.
3. Karlin, S., Mocarski, E. S. & Schachtel, G. A. (1994) *J. Virol.* **68,** 1886–1902.
4. Karlin, S. & Burge, C. (1995) *Trends Genet.* **11,** 283–290.
5. Josse, J., Kaiser, A. D. & Kornberg, A. (1961) *J. Biol. Chem.* **263,** 864–875.
6. Russel, G. J., Walker, P. M. B., Elton, R. A. & Subak-Sharpe, J. H. (1976) *J. Mol. Biol.* **108,** 1–28.
7. Karlin, S. (1996) *Genomic Signature and Bacterial Phylogeny in Bacterial Genomes: Physical Structures and Analysis*, eds. de Bruijn, F. J., Lupski, J. R. & Weinstock, G. (Chapman and Hall, New York), in press.
8. Campbell, A. M. (1994) *Annu. Rev. Microbiol.* **48,** 193–222.
9. Haggard-Ljungquist, E., Holling, C. & Callender, R. (1992) *J. Bacteriol.* **174,** 1462–1477.
10. Casjens, S., Hatfull, G. & Hendrix, R. (1992) *Semin. Virol.* **3,** 383–397.
11. McClelland, M. (1985) *J. Mol. Evol.* **27,** 317–322.
12. Deschavanne, P. & Radman, M. (1991) *J. Mol. Evol.* **33,** 125–132.
13. Yarmolinsky, M. B. & Sternberg, N. (1988) in *The Bacteriophages*, ed. Calendar, R. (Plenum, New York), Vol. 1.
14. Brendler, T. G., Abeles, A. L. & Austin, S. J. (1995) *J. Cell Biol.*, Suppl. 19A, A2–115 (abstr.).
15. Burge, C., Campbell, A. M. & Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 1358–1362.
16. Kessler, C. & V. Manta. (1990) *Gene* **92,** 1–248.
17. Sharp, P. M. (1986) *Mol. Biol. Evol.* **3,** 75–83.
18. Maniloff, J., Campo, G. J. & Dascher, C. C. (1994) *Gene* **141,** 1–8.
19. Hatfull, G. F. & Sarkis, G. J. (1993) *Mol. Microbiol.* **7,** 395–405.
20. Storey, C. G., Lusher, M. & Richmond, S. J. (1989) *J. Gen. Virol.* **70,** 3381–3390.
21. Godson, G. N., Fiddes, J. C., Barrell, B. G. & Sanger, F. (1978) in *The Single-Stranded DNA Phages*, eds. Denhardt, D. T., Dressler, D. & Ray, D. S. (Cold Spring Harbor Lab. Press, Plainview, NY).
22. Holland, J. (1993) in *Emerging Viruses*, ed. S. S. Morse (Oxford, New York).