Genome Analysis

Host-parasite Relations of Bacteria and Phages can be unveiled by *Oligostickiness*, a Measure of Relaxed Sequence Similarity

Shamim Ahmed, Ayumu Saito, Miho Suzuki, Naoto Nemoto and Koichi Nishigaki

Graduate School of Science and Engineering, Saitama University, Saitama 338-8570, Japan

ABSTRACT

Motivation: The recent metagenome analysis has been producing a large number of host-unassigned viruses. Although assigning viruses to their hosts is basically important not only for virology but also for prevention of epidemic, it has been a laborious and difficult task to date. The only effective method for this purpose has been to find them in a same microscopic view. Now, we tried a computational approach based on genome sequences of bacteria and phages, introducing a physicochemical parameter, SOSS (Set of *Oligostickiness* Similarity Score) derived from *oligostickiness,* a measure of binding affinity of oligonucleotides to template DNA.

Results: We could confirm host-parasite relationships of bacteria and their phages by SOSS analysis: all phages tested (25 species) had a remarkably higher SOSS value with its host than with unrelated bacteria. Interestingly, according to SOSS values, lysogenic phages such as lambda phage (host: *E. coli)* or SPP1 (host: *B. subtilis*) have distinctively higher similarity with its host than its non-lysogenic (excretive or virulent) ones such as fd and T4 (host; *E. coli*) or phages gamma and PZA (host; *B. subtilis*). This finding is very promising for assigning host-unknown viruses to its host. We also investigated the relationship in codon usage frequency or G+C content of genomes to interpret the phenomenon revealed by SOSS analysis, obtaining evidences which support the hypothesis that higher SOSS values resulted from the cohabitation in the same environment which may cause the common biased mutation. Thus, lysogenic phages which stay inside longer resemble the host.

Contact: koichi@fms.saitama-u.ac.jp

1 INTRODUCTION

Currently, the most common way to tentatively assign a virus to its host organism is to observe the putative virus in a host cell (this is mainly depending on electron microscopy, which needs to be confirmed through Koch's postulates). Therefore, a lot of viruses remain unconnected to their hosts (Edwards *et al*., 2005). This fact is becoming more marked as a result of metagenome analysis, a widely applied, powerful technique which has yielded the genome sequences of numerous viruses, many of which have yet to be assigned to their hosts (Venter *et al*., 2004). Undoubtedly, the knowledge of host-parasite relationships is essential for elucidating the life cycle of viruses, investigating the whole ecosystem, and also revealing latent and pathogenic viruses. Therefore, it is important to develop Bioinformatic analyses to reveal host-parasite relationships based on genome sequences. In some cases, a host and its parasite must have common sequences usable for the assignment as has been observed in some of lysogenic phages (Blaisdell *et al*., 1996). However, this approach cannot be universal since some phages such as Qβ and fd are too simple and streamlined to have extra sequences shared with the host, prompting the development of novel methodology for this purpose.

The dynamic nature of genomes has now been well-established: i.e., horizontal transfer and frequent recombination as demonstrated by genome sequence information which has become available in this decade (Dawson *et al*., 2002, Nakamura *et al*., 2004). Most of these findings have been made by analyzing the sequence similarity between genomes or their parts: insertions and/or deletions of particular sequences. Naturally, most of these analyses have dealt with sequence information in a strict one-to-one manner and were able to reveal a large amount of information about genomes. The others contain the dinucleotide relative abundance profiles of DNA: those from the same organism are generally much more similar to each other than those from other organisms (karlin *et al*., 1994). In contrast, another approach, namely *oligostickiness*, exploits hidden genome information. *Oligostickiness* analysis, which is based on the binding (or hybridization) stability of an oligonucleotide to a genome sequence of interest (Nishigaki *et al*., 2002), is an example of a technique where one does not try to find unique sequences but rather relaxed and ambiguous ones. This approach was used in the finding of the phenomenon of *chromosome homogenization* during evolution, a phenomenon not seen by the other approaches (Saito *et al*., 2004). Thus, the chromosomes contained in a cell of an organism have a similar tendency in *oligostickiness*, suggesting frequent recombination between chromosomes in the same nucleus (Saito *et al*., 2004) (Supplementary figure 1). In this study, we adopted this method for the determination of the host-parasite relationships of bacteria and their phages.

There are two categories of phages: namely lytic (or excreting) and lysogenic. Lytic phages reproduce themselves, lyse the host cell and release progeny phages after infection (although the excreting type such as fd does not lyse the host cell but excretes its progeny out of the host cell). Lysogenic phages enter a quiescent state by integrating their genomic DNA into the host chromosome until the lytic cycle begins upon triggering by stimuli. Interestingly, our approach clearly discriminates between these two types. We also tested other relevant methods such as G+C content and codon usage analyses for the current purpose and found that the *oligostickiness* analysis has the strongest power of prediction for this purpose. We therefore discuss the reasons for the utility of *oligostickiness*, a measure of relaxed sequence similarity in this paper.

2 METHODOLOGY

2.1 *Oligostickiness*

Calculation of *oligostickiness* has been described in detail in previous papers (Nishigaki *et al*., 2002, Saito *et al*., 2004). In brief, *oligostickiness*, σ, is a parameter defined as follows:

$$\sigma=\frac{1}{n}\sum_{i=l_0+1}^{l_0+n}\delta(p,T(i)) \quad (0 \le \sigma \le 1) \qquad (1)$$

where $l_0 + 1$ and n are the genome sequence position at which the sampling region begins and the sampling size for *oligostickiness*, respectively, and δ

is a determinant that takes the value of 1 when the probe (*p*) binds stably to the *i*-th local sequence of the genome (*T*(*i*)) (in other words, a fragmental sequence that has a fixed 5'-end at the sequence position *i*) or the value of 0 when not bound (ΔG is larger than a fixed value (Nishigaki *et al*., 2002)). In this formula, the *p*–*T* binding is determined based on the thermodynamic stability of the *p* –*T* complex (Figure 1).

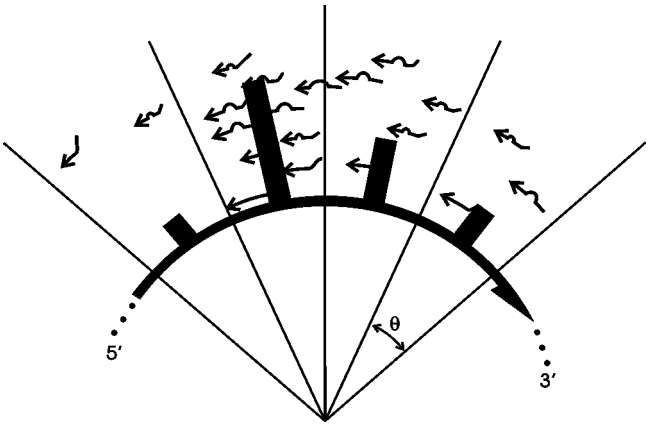


Figure 1. Definition of *oligostickiness*. *Oligostickiness* can be defined as the normalized frequency of binding to the template DNA of a fixed portion (angle θ) in various manners as drawn here. Each binding structure denotes that it makes a sufficiently stable binding to a template DNA at a particular site. The stability of each structure is calculated thermodynamically (see text) and shown with more stable one on a lower layer in this figure. The frequency of probe-binding is accumulated within a sector of the angle θ, normalized by the actual size of the fractional template, and drawn by a pillar, of which height is proportional to the normalized frequency. For the convenience sake, *oligostickiness* is usually defined to the registered genome sequence (or database sequence). This figure was taken from Nishigaki *et al*. (2002).

2.2 Spider-web representation

The representation of chromosome/genome properties used here is called the '*spider-web* presentation of global *oligostickiness*' (i.e., the σ value calculated against the whole entity) (Saito *et al*., 2004). Each global *oligostickiness* value with respect to a chromosome probed by a particular oligonucleotide was plotted on an axis radially extended from a common center, following a logarithmic scale (supplementary figure 1). In this paper 12 axes per round were adopted with 12 different oligonucleotide probes (which were empirically selected), taken from Ref. Saito *et al*., 2004. The nearby plots were connected with a line to define a characteristic pattern for each chromosome. This type of representation appears to be more effective in presenting features of a chromosome in depth.

2.3 Calculation of 'set of *oligostickiness* similarity score (SOSS)' between genomes

SOSS was calculated to detect the similarity of chromosomal texture (Nishigaki *et al*., 2002) as well as to detect the relationship among the genomes. It can be expressed as follows:

$$\mathrm{SOSS} = 1 - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\log S_i^1 - \log S_i^2}{\log S^{max} - \log S^{min}}\right)^2} \quad (0 \le \mathrm{SOSS} \le 1) \qquad (2)$$

where n is the total number of oligonucleotides, S_i^1 and S_i^2 are the *oligostickiness* values against i-th oligo of genome 1 and genome 2, respectively. LogS^{max} and $\log S^{min}$ were set to be 0.1 (maximum) and 0.0001 (minimum) in order to present the chart clearly and informatively based on the data thus far obtained. At the same time, these values were used for normalization by making the ($\log S^{max} - \log S^{min}$) be the unity. The axes were plotted in the logarithmic scale (see Fig. 2)

To examine the effect of codon usage bias or mutational bias during lysogenic state, we made a simulation by replacement of the 3rd positions of codons in coding sequences in genomes with A, T, G or C after which SOSS was calculated. More specifically, s_i^1 and s_i^2 are the *oligostickiness* values of simulated genomes against i-th oligo when the 3rd base of both genome 1 and genome 2 were replaced with any of the single bases.

2.4 Calculation of 'codon usage similarity score (COUSS)' between genomes

Codon usage frequencies of genomes were obtained from Kazusa DNA Res. (http://www.kazusa.or.jp/codon/) and then COUSS was calculated as defined below:

$$\mathrm{COUSS} = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\left|C_h^i - C_p^i\right|}{C_h^i + C_p^i} \quad 0 \le \mathrm{COUSS} \le 1 \qquad (3)$$

where n is the total number of codons (i.e., 64), C_h^i and C_p^i are the codon frequencies of host and phage with regard to the i-th codon, respectively.

2.5 Data sources for whole genome sequences, coding sequences and G+C contents

The whole genomes of phages and hosts were obtained from NCBI and NIG (see supplementary table 2). The assignment of phages to their relevant host and class of lysogenic or non-lysogenic was based on the data collected from literatures. The G+C content of genomes and the discrimination of coding sequences in genomes were obtained from the annotation of genomes contained in NCBI database and the annotation of codon usage database (Kazusa DNA Res. Inst.), respectively.

3 RESULTS AND DISCUSSION

In this study we dealt with a considerable number (25 phages + 7 hosts) of lysogenic, non-lysogenic phages and their corresponding bacteria: *Vibrio cholerae* vs. phage VSK; *Streptococcus thermophilus* vs. phage Sfi21 and O1205; *Clostridium difficile* vs. phage ΦC2 and others as listed in Table 1. These are experimentally well-established regarding the host-parasite relationship besides being available of their genome sequences.

To reveal the host-parasite relationships of phages, it is natural to try to exploit their genome sequences. However, an approach that compares the sequences directly is less likely to provide the desired information since there may be no common sequences between the host and parasite. Here, we employed a method that can unveil the hidden information: namely *oligostickiness* analysis. The *oligostickiness*, a measure of affinity, which had been introduced successfully to characterize genome sequences of various organisms (Nishigaki *et al*., 2002), was applied to the study of host-parasite relationships. The twelve probes (oligonucleotides) were empirically selected with each sequence mutually quasi-orthogonalized i.e., making them different from each other as much as possible. In brief, the oligonucleotides were selected with the following considerations: i) oligonucleotides with different properties (G+C content, sequence complexity, and thermodynamic stability) and ii) representative (mutually pseudo-orthogonal) oligonucleotides from the viewpoint of *oligostickiness* based on data collected from over 20 genomes (Nishigaki *et al*., 2002, Saito *et al*., 2004). As partly shown in Fig. 2, all probes have the same or similar *oligostickiness* values between host bacteria and their phages in the four spider-web charts shown here. As clearly shown, the two categories of phages, i.e., lysogenic and non-lysogenic, display different patterns,

with the former closely overlapping with the host (a and c) while the latter have similar but non-overlapping values to those of their hosts (b and d). This means that the genomes of the host and the lysogenic phages are formed of a similar 'texture' of sequences (Nishigaki *et al*., 2002) and that the homogenization of these chromosomes has notably advanced (Saito *et al*., 2004). This finding was further confirmed by introducing a similarity measuring parameter; SOSS (set of *oligostickiness* similarity score) defined in Methodology (Eq. 2) and applying it to the groups of bacteria and phages, for which genomes had been determined. Fig. 3 shows the SOSS values obtained for five groups of bacteria (*Escherichia coli, Bacillus subtilis, Vibrio cholerae, Streptococcus thermophilus/pyogenes, and Clostridium difficile/botulinum*) and their phages (see Table 1). It is evident that relevant phages, regardless of their lysogenic or non-lysogenic property, have higher score (0.86 ± 0.04 for non-lysogenic and 0.96 ± 0.013 for lysogenic) than that of unrelated phages (0.83 ± 0.07) for all cases tested (Fig. 3). Remarkably, lysogenic phages have a sharply higher SOSS value than for non-lysogenic ones for all bacteria tested here. This notion was further confirmed by investigating about all relationships by way of SOSS (Table 2). Such high scores as greater than 0.95 appear only in the established host-lysogenic phage relationships except two cases (*E. coli* phages (ld and st1) against *V. cholerae* and *S. thermophilus* and *S. pyogenes* against the phages of ss, so, and sp1~6), enabling us to use it as a discriminator of such relationships. In these two exceptions, it is noteworthy that two bacteria for both cases (*E. coli* and *V. cholerae*; *S. thermophilus* and *S. pyogenes*) are known to be genetically close. Table 2 also presents that some non-lysogenic phages such as *E. coli* phages fd and *B. subtilis* phages gamma (symbolized as bγ in Table 2) behave quite singularly with respect to the SOSS value since they resemble more the other bacteria (in case of bγ, *S. thermophilus*, and *S. pyogenes*) than their authentic host (e.g. *B. subtilis* for bγ). This fact may indicate that those phages have a wide host range as has been demonstrated with phages Mu (Harshey, 1988) and lytic bacteriophages of *Sphaerotilus natans* (Jensen, 1998). This is another view point to be explored in future employing this SOSS analysis. These findings demonstrate that analysis of genome sequences can be useful in the determination of host-parasite relationships without direct observation of the interaction. In the same vein, dinucleotide relative abundance profiles of host-parasite supported that lysogenic phages were close to their host, whereas lytic phages were relatively distant (Blaisdell *et al*., 1996). These are especially

Table 1. SOSS and the other characteristic values for bacteria and phages*

Hosts and parasites	Abbreviation	Phage type	G+C content†			COUSS§	SOSS§
			α_c	α_w	Ratio of α_w‡		
E.coli	E.c	(Host)	52	50			
Lambda phage	ld	Lysogenic	51	49	0.98	0.874	0.969
Siga toxin 1 (stx1)	st1	Lysogenic	51	49	0.98	0.853	0.947
Siga toxin 2 (stx2)	st2	Lysogenic	51	49	0.98	0.854	0.947
BacteriophageT4	t4	non-lysogenic	35	35	0.70	0.672	0.796
Bacteriophage T7	t7	non-lysogenic	48	48	0.96	0.784	0.882
Phage fd	fd	non-lysogenic	40	40	0.80	-	0.900
Bacillus subtilis	B.s	(Host)	44	43			
Bacillus phage SPP1	bs	Lysogenic	44	43	1.00	0.874	0.955
Bacillus phage gamma	bγ	non-lysogenic	36	35	0.81	0.753	0.868
Bacillus phage B103	bb	non-lysogenic	38	37	0.86	0.768	0.833
Bacillus phage GA-1	bg	non-lysogenic	35	34	0.79	0.706	0.792
Bacillus phage PZA	bp	non-lysogenic	40	39	0.90	0.802	0.859
Vibrio choleraeO395 chr1	V.c	(Host)	48	46			
Vibrio phage VSK	vv	Lysogenic	43	43	0.93	0.789	0.954
Vibrio phage fs1	vf	non-lysogenic	43	43	0.93	0.784	0.918
Vibrio phage VP4	v4	non-lysogenic	44	42	0.91	0.742	0.849
Streptococcus thermophilus	S.t	(Host)	40	39			
S. phage Sfi21	ss	Lysogenic	38	37	0.94	0.716	0.952
S. phage O1205	so	Lysogenic	38	38	0.97	0.855	0.955
S. phage DT1	sd	non-lysogenic	40	39	1.00	0.849	0.851
Streptococcus pyogenes315	S.p	(Host)	39	38			
S. pyogene phage315.1	sp1	Lysogenic	38	37	0.97	0.860	0.974
S. pyogene phage315.2	sp2	Lysogenic	39	38	1.00	0.892	0.967
S. pyogene phage315.3	sp3	Lysogenic	38	38	1.00	0.874	0.961
S. pyogene phage315.4	sp4	Lysogenic	39	38	1.00	0.825	0.964
S. pyogene phage315.5	sp5	Lysogenic	38	38	1.00	0.862	0.972
S. pyogene phage315.6	sp6	Lysogenic	40	39	1.02	0.882	0.948
Clostridium difficile630	C.d	(Host)	30	29			
C. difficile phage phiC2	cp2	Lysogenic	29	28	0.96	0.887	0.951
Clostridium botulinum S str	C.b	(Host)	29	28			
C. botulinum phage C-st	cc	Lysogenic	27	26	0.92	0.868	0.924

*Genome sequences and codon usage data were arranged after collecting from NCBI (http://www.ncbi.nlm.nih.gov/) and Kazusa DNA Res. Inst. (http://www.kazusa.or.jp/codon/). †G+C contents surveyed over coding sequences only (α_c) or whole genome sequences (α_w). ‡Ratio taken for α_w of the phage against α_w of the host. §COUSS (codon usage similarity score) and SOSS (set of oligostickiness similarity score) defined in the text.

important in metagenomics analysis of microbiomes where a large number of phages need to be assigned without current knowledge of host-parasite relationships (Edwards *et al*., 2005). Furthermore, the ability to discriminate between lysogenic and non-lysogenic phages should also be usable in the general study of phages.

The result obtained here is clear but yet to be rationalized. There are only two explanations for the observations. The first is the frequent recombination between host and parasite genomic DNAs including the horizontal transfer of genes, where genetic recombination occurs during the events of integration and release of the genes and at the same time such genes can also cohabit with the host genome for a considerable duration depending on the type of the genetic materials (vectors).This has already been established to be plausible and to actually occur (Jain *et al*., 1999, Dawson *et al*., 2002). The other explanation, for which there is less evidence, is that there is a biased selection pressure toward genome sequences which work in the host and parasites exposed to the same environment and that this leads to 'similarly-textured' sequences. In order to test the possibility of the latter case, we have compared the codon usage of host and parasite pairs by defining a COUSS (COdon Usage Similarity Score) parameter as defined in Eq. 3. In this analysis, a similar result to that of the SOSS analysis was obtained as shown in Fig. 4 and Table 1 (average COUSS values of lysogenic phages were found to be higher (0.84 ± 0.049) than those of non-lysogenic (0.76 ± 0.056) and phages unrelated (0.68± 0.080) to the hosts (Fig. 4 & supplementary table 1)). Thus, some biased mutations in the third position of codons have resulted in the generation of similar codon usage between the host and its parasite. Since the codon usage of a phage needs to be optimized

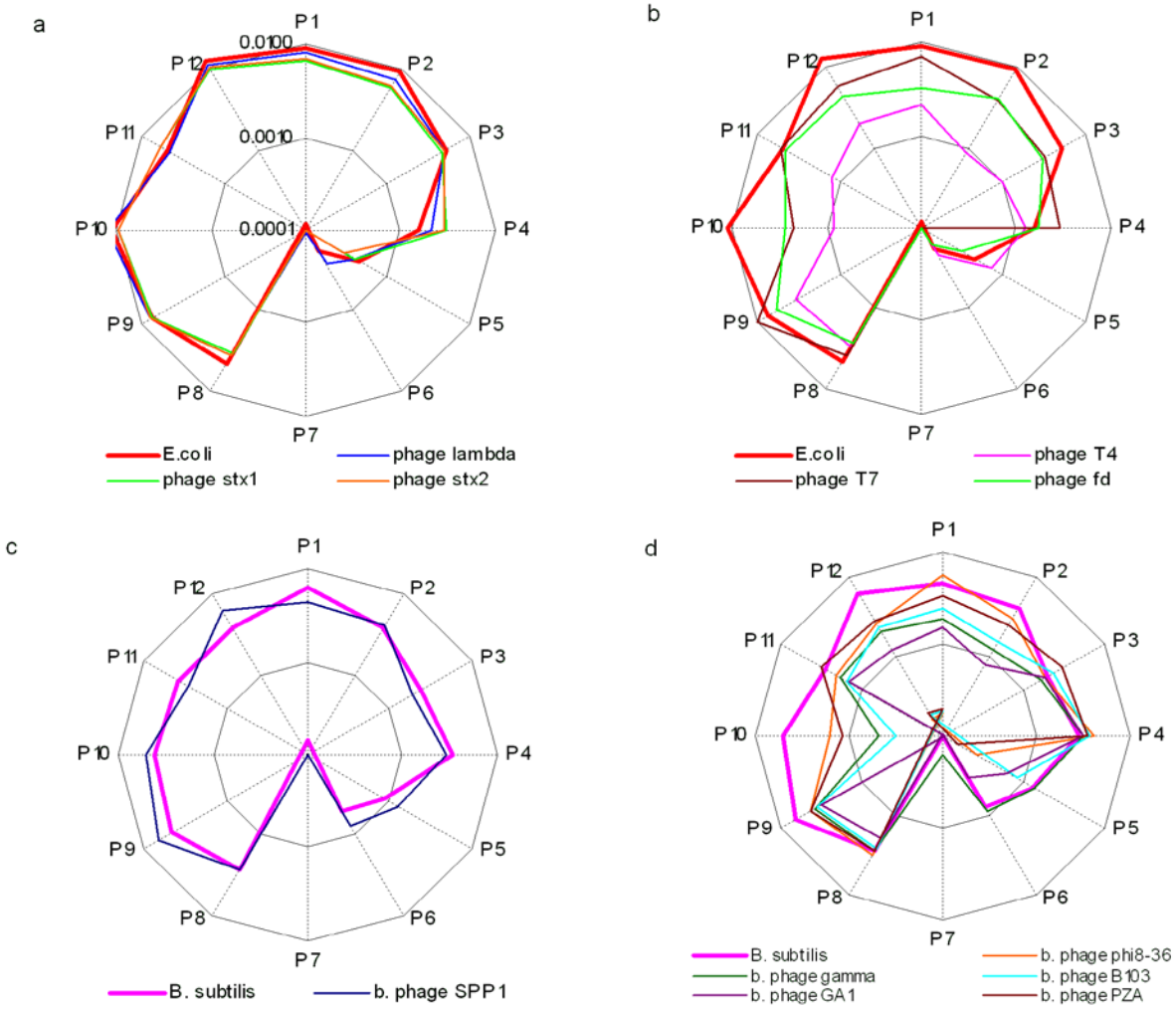


Figure 2. Spider-web chart representation of *oligostickiness*. *Oligostickiness* values obtained for each pair of a phage (as shown) and a host bacterium (a and b *E.coli* or c and d *Bacillus subtilis*) using 12 different probe oligonucleotides were plotted on the radial axes and connected with a line. Circles obtained for different phages (a and c, lysogenic; b and d, non-lysogenic) are superimposed. The logarithmic scale is used for *oligostickiness* (i.e., 10^{-3}, 10^{-2}, 10^{-1}, inner to outer). Note that since the center is not zero but 10^{-3}, some *oligostickiness* values, which are smaller than 10^{-3}, are plotted on the opposite side of the axis. The probe sequences used were: P1, dACGACGACGACG; P2, dGGGGTCGAGGGG; P3, dTGGGTGGGTGGG; P4, dGAGAGAGAGAGA; P5, dGCTAAAAAAAAA; P6, dAAAAAAAAAAAA; P7, dATATATATATAT; P8, dGTGCTGGGATTA; P9, dCCAGGCTGGTCT; P10, dCCGGCCGGCCGG; P11, dGGGGTCGAGGCG; P12, dAGACCGCGCCTG;.

to perform the most efficient proliferation in the host cell, it is apt that it comes to mimic that of the host (Bailly-Bechet *et al*., 2007). The genomes of bacteriophages are well known to be extremely stream-lined so that they leave almost no redundant portion besides necessary gene-coding regions (Kornberg, *et al*., 1992). Considering this fact, the phage genomes are less plausible to have experienced drastic recombination events but rather much possible to have accepted biased point mutation events. This is quite different from the situation of chromosomes in a nucleus in which chromosomes could be rather freely recombined unless it would cause gene disruption or the similar. Thus, the fact that pairs of a host and a parasite have similar codon usage patterns supports the hypothesis that there is a biased substitution-mutation pressure such as the preference of G-to-A changes. Such a possibility is highly plausible since host and parasite replicate in the same molecular environment using the same replication and repair systems. Although the overall tendency is similar to that seen with the SOSS analysis, the discrimination by COUSS is not as clear as SOSS (Figs. 3 and 4). In order to compare SOSS and COUSS discrimination power of lysogenic phages, we also computed sensitivity and specificity of both methods on the same dataset except Clostridium phage C-st (which is pseudo-lysogen to *clostridium botulinum F str*.), where values of ≥ 0.95 and ≥ 0.84 were considered as positive cases (lysogenic) for SOSS and COUSS analyses, respectively. We found, though using a limited number of samples obtained here, that both sensitivity and specificity were higher in SOSS analyses (100% and 92%) than COUSS analyses (71% and 91%), respectively, indicating that the SOSS analysis is more suitable for this purpose.

As both analyses demonstrated that lysogenic phages have genome sequences more closely related to their hosts than non-lysogenic ones to the same host, it is logical to conclude that given their life cycles, the difference comes from the frequency of biased mutation experienced and that this must be nearly proportional to the duration of host-and-parasite cohabitation. Therefore, upon introduction of commonly biased mutations into the third position of codons (mostly synonymous) in the genomes of a host and its non-lysogenic parasite, we would expect the SOSS values for such pairs to get higher. This was indeed the case with all of the tested pairs (for non-lysogenic phages and unrelated ones) while the changes for lysogenic phages were small and non-directional (Fig. 5). As can be seen, this tendency does not depend on the base to which the mutation directed. Most lysogenic pairs seem to be near equilibrium (maximum) state with regard to the genome homogenization phenomenon between host and parasite (Kejnovsky *et al*., 2007) since some mutations did not improve but rather reduced the SOSS value. This fact strongly supports the idea that the biased mutation which is directed to the same mutation product such as A, G, T, or C may be the cause of the higher SOSS values between a host and its lysogenic phages.

We also examined whether G+C content analysis can provide similar predictions to those obtained using SOSS and COUSS. For this purpose G+C content values were collected from the relevant databases (see Methodology) and found that the average value of host-parasite G+C content ratio, α_w (see definition in Table 1) are 0.97 ± 0.02 and 0.86 ± 0.06 for lysogenic and non-lysogenic phages, respectively. While these values are discriminating between lysogenic and non-lysogenic phages to their hosts, some of the non-lysogenic phages of *V. cholerae* and *S. thermophilus* showed similar or higher ratio of α_w than lysogenic phages (Table 1). It is therefore evident that G+C content analysis cannot provide any prediction about host-parasite relationships. This is in agreement with the work of Karlin *et al*. (1998), and they found that prokaryotic genomes tend to be homogeneous in their G+C content and this property was not diagnostic in discriminating among prokaryotes.

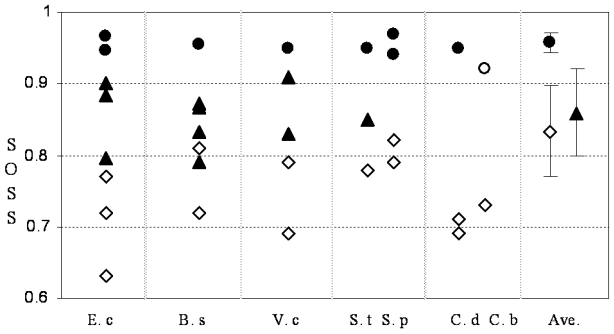


Figure 3. SOSS (Set of *oligostickiness* similarity score) plot of phages against their host. Symbols used are circle (lysogenic), triangle (non-lysogenic), and diamond (non-related). The average SOSS values are plotted rightmost for each categories. The abbreviations E.c., B.s., V.c., S.t., S.p., C.d., and C.b. represent *Escherichia coli*, *Bacillus subtilis*, *Vibrio cholerae O395*, *Streptococcus thermophilus CNRZ1066*, *Streptococcus pyogenes 315*, *Clostridium difficile 630* and *Clostridium botulinum S str*, respectively. The result for the only pseudolysogen is distinctively shown in blank circle.

Table 2. SOSS table between phages and hosts.

Phage	Host bacteria							
	E.c	B.s	V.c1	V.c2	S.t	S.p	C.d	C.b
ld	97	90	94	95	83	82	68	67
st1	95	88	94	95	83	82	68	67
st2	95	87	94	94	83	81	67	66
t4	80	85	85	84	92	91	83	81
t7	88	85	91	91	85	83	70	68
fd	90	90	96	95	89	88	74	73
bs	89	96	94	93	91	90	75	74
bγ	78	87	84	83	95	95	87	85
bb	79	83	85	84	91	89	81	79
bg	73	79	79	78	89	88	85	82
bp	85	86	91	90	89	87	76	74
vv	91	90	95	95	88	87	73	72
vf	88	86	92	92	85	83	71	70
v4	80	82	85	84	87	85	76	74
ss	78	86	84	83	95	96	85	83
so	79	87	85	84	96	96	84	82
sd	77	82	81	80	85	86	82	81
sp1	79	88	85	84	96	97	85	83
sp2	82	91	87	86	95	97	82	81
sp3	78	87	84	83	95	96	86	85
sp4	80	89	85	85	95	96	83	81
sp5	83	92	88	88	96	97	81	80
sp6	80	89	84	84	93	95	84	82
cp2	64	72	69	69	79	80	95	95
cc	63	71	68	68	79	79	94	92

SOSS values are shown taking two digits below the decimal point as 95 for 0.952. Colors represent score range: red; 1~0.95, orange; 0.94~0.90, yellow; 0.89~0.85, blank; below 0.85. Boxed cells are corresponding to a pair of phages sharing a common host. The abbreviations are used as shown in Table 1.

3.1 Lysogenic state stability and high SOSS value of phages

Lysogenic *E.coli* phages, Siga toxin phage 1 (stx 1) and Siga toxin phage 2 (stx 2), were both found to have a SOSS value of 0.947 (Table 1). In contrast, another phage, lambda, has a higher SOSS value of 0.969 and is known to be more stable than the Siga toxin phages on the basis of induction rate (Aurell *et al*., 2002, Livny *et al*., 2004). It is impressive that an excretive phage fd, less virulent than lytic phages has an intermediate value of SOSS between those of lytic and lysogenic phages while the G+C content is not similar. A similar phenomenon was observed with *S. pyogenes* lysogens. The bacteria, *S. pyogenes*, were found to be polylysogenic in sequenced strains with up to 10% of the total host genome being phage DNA. According to an intensive study of the lytic induction of *S. pyogenes MGAS315* prophages, mitomycin C, hydrogen peroxide, and other physiological stimuli were shown to induce prophages with a variable efficiency (Banks *et al*., 2003). Interestingly, the order of SOSS values obtained here (0.94 to 0.97 in Table 3) approximately corresponds to the order of the induction rate of *S. pyogenes*, possibly indicating the degree of phage adaptation to the lysogenic state. This can be explained by the fact that phages which are more stable (non-responsive) against such environmental stimuli (UV, mitomycin C and others) will result in longer coexistence, and thus higher SOSS values as discussed above. Clostridium phage C-st, which has historically been called pseudolysogeny (Sakaguchi *et al*., 2005) and which is an exceptionally unstable lysogenic phage of *C. botulinum F str.*, was found to have a relatively low SOSS value of 0.924 compared to the other relevant lysogenic phages (Table 1). Ultimately, stable, i.e. long-term coexisting, lysogenic phages can increase the frequency of genetic recombination and/or biased mutations, which must lead to a higher SOSS value. Intriguingly, HIV-1 isolated in 2005 (NCBI Accession No. AB287363) has a much higher SOSS value to human chromosome 16 (0.817) than HIV-1 previously isolated in 1976 (NCBI Accession No. U76035) (0.755). Similarly, for HPV16 and Human (chromosome 16) a bit higher SOSS value of 0.827 was found (our unpublished data). Therefore, SOSS may be used to measure the duration of the lysogenic state (or the frequency of interaction with the host) though larger, more detailed amounts of experimental data are required to establish the quantitative relationship between the SOSS and lysogenic stability.

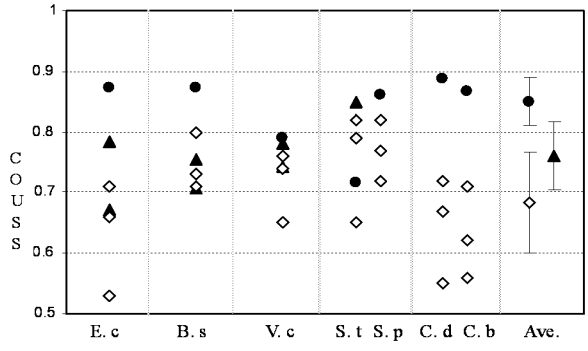


Figure 4. COUSS (Codon usage similarity score) plot of phages against their host. The same symbols and presentations are used as in Fig. 3.

3.2 Measure of the relaxed sequence similarity

We believe that one of the most important concepts presented in this paper is that a relaxed sequence similarity analysis, *oligostickiness*, can extract a large amount of information from genome sequences which is unavailable by conventional strict sequence similarity analyses such as repeat sequence analysis. As has already been well-studied, those sequences (genes) which had the same original sequence (gene) are descended with multiple mutations and are present as homologs and paralogs (Koonin *et al*., 2005). Therefore, the degree of similarity becomes an important concept in the study of such sequences in genomes where there are various sequences of different origins and thus different duration of mutation. Sequences diverged recently should be very close and easily recognized such that they could be analyzed by conventional approaches which deal with sequences strictly based on complete match or similarity (i.e., Hamming distance, $d_H \approx 0$). In contrast, those sequences which diverged a sufficiently long time ago will have changed close to random sequences assuming that there have been no functional constraints on the sequence which would have prevented from altering. Usually, sequences of genomes are intermediately positioned between the two extremes depending on the time from the generation of sequences (genes). *Oligostickiness* analysis calculates the free energies (ΔG) of all of the possible hybridization structures formed between the template and the probe (oligonucleotide) at each position along a genome sequence, which allows the counting up all of the possible structures including a lot of mismatch-containing hybridization ones (which have large d_H value) as long as they have a certain level of stability in terms of ΔG. Therefore, the *oligostickiness* analysis is rather statistical and robust against mutations and is thus endowed with the ability to analyze highly diverged (i.e., relaxed) and veiled sequences. This is why we call *oligostickiness* analysis a measure of relaxed sequence similarity. Since the approach taken is clearly successful as demonstrated here and elsewhere (Nishigaki *et al*., 2002, Saito *et al*., 2004), it is clear that relaxed sequence similarity analyses like *oligostickiness* is another useful genome sequence analysis.

In a preceding study, we selected 12 oligonucleotides as probes for an SOSS-like analysis (Saito *et al*., 2004). In prior to the selection of these probes, we performed the *oligostickiness* analysis with various genomes (virus to human), which were then available, using different probes (more than 20 species). In this analysis each probe generated a specific *oligostickiness* profile for each genome. We think the probes used here are similarly effective and that the set are significant and practical. However, this does not deny the possibility of selecting another set of probes in a more well-defined manner in future.

Table 3: SOSS versus induction of *Streptococcus pyogenes 315* phages.

Prophage	Induction* (PR-THY)			SOSS
	Spontaneous	Mitomycin C	H_2O_2	
Φ315.6	++	+++	+++	0.948
Φ315.3		++	++	0.961
Φ315.4	++	+++	+++	0.964
Φ315.2		±		0.967
Φ315.5		++		0.972

*Relative degree of prophage induction is indicated as follows: ±; variable and weak, ++; intermediate, +++; strong. PR-THY; protein reduced yeast extracts. Induction data was taken and modified from Ref. Banks *et al.*, 2003.

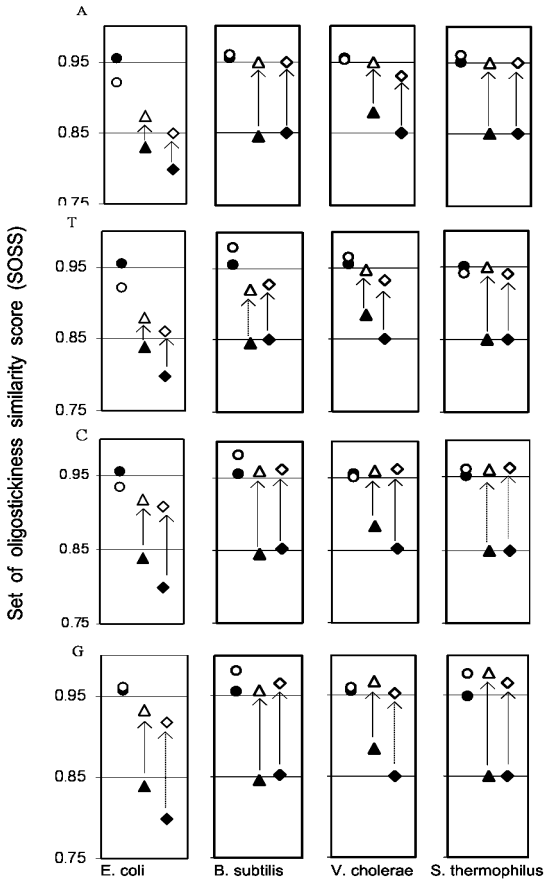


Figure 5. Effect of the biased mutation of the third letter of codons on the SOSS value. The third positions of codons were uniformly changed (e. g., A to G, C, or T) for both hosts and parasites and then SOSS values of the altered gene sequences were calculated. SOSS values are plotted for each host depending on the type of alteration (convergent nucleotide). The symbols used are circle (lysogenic), triangle (non-lysogenic), and diamond (non-related) before (filled) and after (open) mutations. The directions of changes are indicated by an arrow.

CONCLUDING REMARKS

A measure of relaxed sequence similarity, *oligostickiness*, was shown to be effective in finding a cryptic property, host-parasite relationship of bacteria and phages, hidden in genome sequences, which is unavailable by strict sequence similarity analyses. Lysogenic phages were found to be highly similar to its host bacterium in similarity parameters SOSS and COUSS. Especially SOSS, a set of *oligostickiness* similarity score, was excellently predictive of host-parasite relationships. This phenomenon was rationalized by the common suffering of biased mutations for lysogenic phages and bacteria which are long sharing the same physiological environment.

REFERENCES

Aurell, E.S., Brown, S., Johnson, J., Sneppen, K. (2002) Stability puzzles in phage λ. *Phys Rev*, E65, 051914.

Bailly-Bechet, M., Vergassola, M., Rocha, E. (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res*, **17**, 1486-1495.

Banks, D.J., Lei, B., and Musser, J.M. (2003) Prophage induction and expression of prophage-encoded virulence factors in group A *Streptococcus* serotype M3 strain MGAS315. *Infect Immun*, 71, 7079–7086.

Blaisdell, B.E., Campbell, A.M., Karlin, S. (1996) Similarities and dissimilarities of phage genomes. *Proc Natl Acad Sci USA*, 93, 5854-5859.

Dawson, E., *et al.* (2002) A first –generation linkage disequilibrium map of human chromosome 22. *Nature*, 418, 544-548.

Edwards, R.A., Rohwer, F. (2005) Viral metagenomics. *Nature Reviews*, 3, 504-510.

Harshey, R.M. **(**1988) in *The bacteriophages*, ed Callendar R (Plenum Press, New York), pp 193–234.

Jain, R., Rivera, M.C., and Lake, J.A. (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci USA*, 96, 3801-3806.

Jensen, E.C. (1998) Prevalence of Broad-Host-Range Lytic Bacteriophages of *Sphaerotilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*. *Applied and Environmental Microbiology*, 64, 575-580.

Karlin, S. *et al*. (1994) Heterogeneity of genomes: Measures and values. *Proc Natl Acad Sci USA*, 91, 12837-12841.

Karlin, S. *et al.* (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev.Genet* 32, 185–225.

Kejnovsky, E. *et al.* (2007) High intrachromosomal similarity of retrotransposon long terminal repeats: Evidence for homogenization by gene conversion on plant sex chromosome. *Gene*, 390, 92-97.

Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39, 309-338.

Kornberg, A. and Baker, T. A. (1992) DNA Replication (2nd edition), W. H. Freeman & Company, New York.

Little, J.W., Shepley, D.P., Wert, D.W. (1999) Robustness of a gene regulatory circuit. *EMBO J*, 18, 4299–4307.

Livny, J., Friedman, D.I. (2004) Characterizing spontaneous induction of Stx encoding phages using a selectable reporter system. *Mol Microbiol*, 51, 1691–1704.

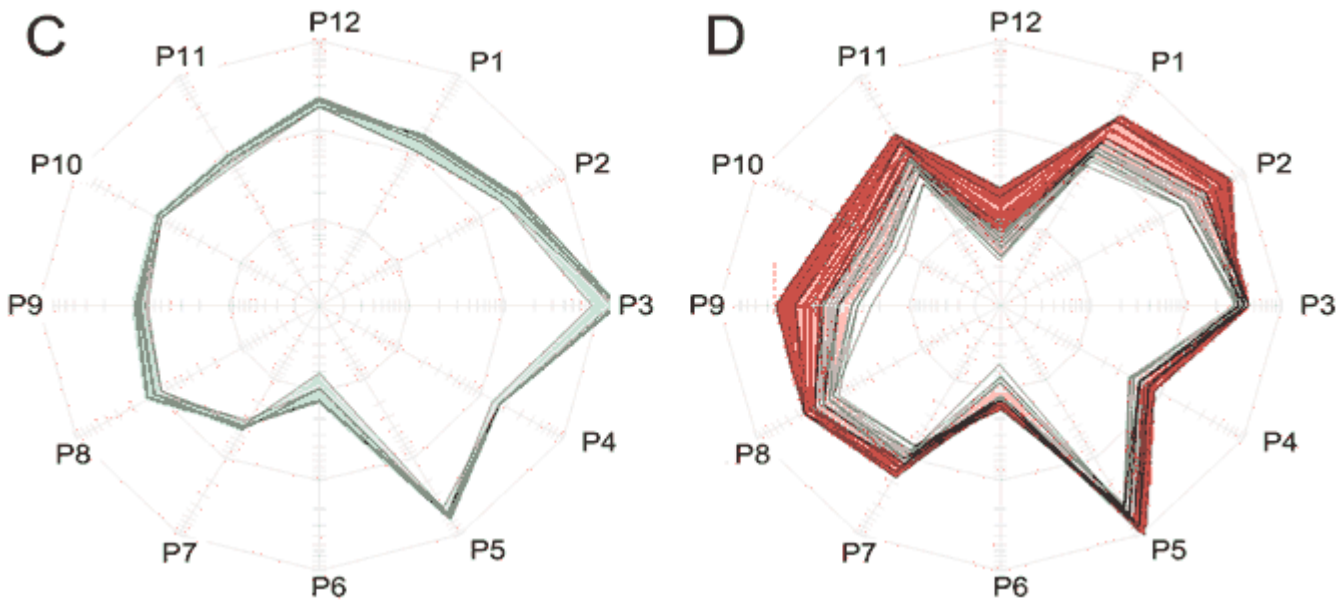
Nakamura, Y., Itoh, T., Matsuda, H., Gojobori, T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics*, 36, 760 – 766.

Nishigaki, K., Saito, A. (2002) Genome structures embossed by oligonucleotide-stickiness. *Bioinformatics*, 18, 1153-1161.

Saito, A., Nishigaki, K. (2004) Homogenization of chromosomes revealed by oligonucleotide-stickiness. *J. Comput. Chem. Jpn*, 3, 145-152.

Sakaguchi, Y., *et al.* (2005) The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. *Proc Natl Acad Sci USA*, 102, 17472–17477.

Venter, J.C., *et al*. (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304, 66 – 74.



Supplementary Figure 1: Chromosome structures of two organisms represented by a *spider-web chart* of *oligostickiness*. *oligostickiness* values plotted on the radial axes are connected with a line to form a circle for each chromosome for multi-chromosomal genomes. These circles are superimposed. *Oligostickiness* is plotted on a logarithmic scale with three polygons crossing at 10^{-3}, 10^{-2}, and 10^{-1} (inner to outer). C, *Caenorhabditis elegans* (6 chromosomes); D, *Homo sapiens* (22 autosomes and X and Y chromosomes). P1, dGGGGTCGAGGGG; P2, dTGGGTGGGTGGG; P3, dGAGAGAGAGAGA; P4, dGCTAAAAAAAAA; P5, dAAAAAAAAAAAA; P6, dATATATATATAT; P7, dGTGCTGGGATTA; P8, dCCAGGCTGGTCT; P9, dCCGGCCGGCCGG; P10, dGGGGTCGAGGCG; P11, dAGACCGCGCCTG; P12, dACGACGACGACG. This figure was taken from ref. Saito *et al* (2004).

Supplementary Table 1: COUSS (Codon usage similarity score) of phages to their hosts

ld	st1	st2	t4	t7	B.s	bs	bγ	bb	bg	bp	V.c1	vv	vf	v4	
.875	.853	.854	.673	.785	.834	.788	.669	.691	.632	.724	.886	.760	.750	.700	E.c
	.914	.915	.679	.786	.842	.829	.692	.700	.646	.746	.822	.793	.745	.689	ld
		.991	.689	.781	.828	.825	.688	.698	.643	.735	.820	.806	.746	.677	st1
			.690	.781	.827	.824	.687	.696	.642	.734	.821	.806	.746	.677	st2
				.732	.738	.734	.848	.813	.804	.796	.709	.777	.765	.777	t4
					.767	.798	.711	.751	.690	.789	.791	.793	.788	.814	t7
						.875	.754	.768	.706	.802	.829	.802	.772	.718	B.s
							.765	.794	.738	.846	.789	.812	.769	.739	bs
								.839	.834	.829	.717	.742	.719	.738	bγ
									.830	.892	.731	.772	.754	.778	bb
										.834	.656	.716	.704	.737	bg
											.746	.793	.768	.782	bp
												.789	.785	.742	V.c1
													.892	.748	vv
														.764	vf

S.t	ss	so	sd	S.p	sp1	sp2	sp3	sp4	sp5	sp6	C.d	cp2	C.b	cc	
.727	.699	.691	.715	.733	.717	.717	.704	.711	.705	.716	.538	.533	.535	.538	E.c
.719	.728	.721	.753	.737	.748	.748	.733	.747	.728	.748	.550	.554	.556	.546	ld
.718	.738	.709	.743	.733	.742	.732	.723	.745	.715	.738	.559	.551	.561	.549	st1
.717	.738	.708	.743	.733	.743	.732	.723	.745	.715	.737	.558	.551	.560	.548	st2
.825	.709	.836	.814	.830	.817	.817	.834	.772	.827	.795	.736	.712	.727	.719	t4
.780	.717	.757	.776	.771	.761	.763	.760	.752	.753	.770	.572	.563	.578	.560	t7
.801	.746	.783	.797	.819	.801	.817	.806	.792	.796	.820	.617	.610	.616	.615	B.s
.778	.781	.795	.837	.787	.830	.831	.822	.841	.823	.833	.612	.622	.612	.605	bs
.801	.735	.844	.830	.825	.826	.856	.871	.803	.848	.833	.776	.768	.757	.748	bγ
.848	.720	.868	.851	.852	.820	.858	.854	.810	.855	.851	.729	.694	.701	.692	bb
.790	.699	.824	.804	.784	.781	.800	.814	.758	.807	.792	.778	.756	.745	.749	bg
.848	.762	.861	.871	.844	.846	.879	.872	.839	.865	.873	.672	.661	.664	.657	bp
.782	.717	.731	.757	.792	.762	.765	.748	.755	.749	.754	.571	.565	.571	.569	V.c1
.783	.754	.781	.814	.795	.815	.805	.797	.803	.791	.798	.619	.601	.613	.597	vv
.803	.714	.770	.795	.795	.800	.780	.770	.767	.775	.779	.599	.594	.592	.576	vf
.807	.693	.789	.778	.782	.770	.765	.768	.745	.766	.770	.625	.612	.613	.609	v4
	.716	.856	.850	.930	.842	.859	.852	.803	.837	.856	.691	.659	.679	.659	S.t
		.734	.766	.730	.782	.756	.739	.823	.730	.746	.608	.613	.603	.611	ss
			.906	.855	.865	.885	.908	.832	.896	.868	.717	.701	.695	.683	so
				.862	.903	.913	.902	.877	.899	.880	.673	.675	.664	.648	sd
					.861	.893	.875	.826	.862	.882	.699	.664	.688	.670	S.p
						.912	.904	.905	.888	.892	.675	.683	.676	.660	sp1
							.934	.897	.937	.926	.693	.687	.685	.669	sp2
								.874	.926	.911	.716	.712	.706	.691	sp3
									.870	.882	.650	.669	.654	.641	sp4
										.904	.705	.698	.693	.680	sp5
											.680	.677	.668	.659	sp6
												.887	.907	.885	C.d
													.858	.876	cp2
														.868	C.b

The abbreviations are used as shown in the main text (table 1).

Supplementary Table 2: Source of genomes

Name of the sample	Size (base)	Source	Accession no.
Streptococcus pyogenes MGAS315	1900521	NCBI	NC_004070
S. pyogenes phage 315.1	39538	NCBI	NC_004584
S. pyogenes phage 315.2	41072	NCBI	NC_004585
S. pyogenes phage 315.3	34419	NCBI	NC_004586
S. pyogenes phage 315.4	41796	NCBI	NC_004587
S. pyogenes phage 315.5	38206	NCBI	NC_004588
S. pyogenes phage 315.6	40014	NCBI	NC_004589
Vibrio cholerae0395 (chromosome 1)	1108250	NCBI	NC_009456
Vibrio cholerae0395 (chromosome 2)	3024069	NCBI	NC_009457
Vibrio phage VSK	6882	NCBI	NC_003327
Vibrio phage fs1	6340	NCBI	NC_004306
Vibrio phage VP4	39503	NCBI	NC_007149
E.coli	4636552	NIG	
Siga toxin converting phage 1 (stx 1)	59866	NCBI	NC_004913
Siga toxin converting phage 2 (stx 2)	62706	NCBI	NC_004914
Enterobacteria phage lambda	48502	NCBI	NC_001416
Enterobacteria phage T4	168903	NCBI	NC_000866
Enterobacteria phage T7	39937	NCBI	NC_001604
Phage fd	6408	EMBL	J02451
Bacillus subtilis	4214814	-	-
Bacillus phage SPP1	44010	NCBI	NC_004166
Bacillus phage gamma	37253	NCBI	NC_007458
Bacillus phage B103	18630	NCBI	NC_004165
Bacillus phage GA-1	21129	NCBI	NC_002649
Bacillus phage PZA (phi29)	19368	NCBI	NC_001423
Clostridium difficil 630	4290252	NCBI	NC_009089
Clostridium phage phiC2	56538	NCBI	NC_009231
Streptococcus thermophilus CNRZ1066	1796226	NCBI	NC_006449
Streptococcus phage sfi21	40739	NCBI	NC_000872
Streptococcus phage O1205	43075	NCBI	NC_004303
Streptococcus phage DT1	34815	NCBI	NC_002072
Clostridium botulinum F str	3995387	NCBI	NC_009699
Clostridium phage c-st	185683	NCBI	NC_007581

NCBI; National Center for Biotechnology Information, NIG; National Institute of Genetics, EMBL; EMBL Nucleotide Sequence Database (EMBL-bank). Genome sequences of *Bacillus subtilis* were taken from the paper: Kunst F, *et al.* (1997), *Nature* 390: 249-256.