Evolutionary Role of Restriction/Modification Systems as Revealed by Comparative Genome Analysis

Eduardo P.C. Rocha,[1,2,5] Antoine Danchin,[2,3] and Alain Viari[1,4]

[1]Atelier de BioInformatique, Université Paris VI, 75005 Paris, France; [2]Unité de Régulation de l'Expression Génétique, Institut Pasteur, 75724 Paris, France; [3]Hong Kong University Pasteur Research Centre, Dexter HC Man Building, Pokfulam, Hong Kong; [4]Action Helix, INRIA-Rhône-Alpes, 38330 Montbonnot-Saint Martin, France

Type II restriction modification systems (RMSs) have been regarded either as defense tools or as molecular parasites of bacteria. We extensively analyzed their evolutionary role from the study of their impact in the complete genomes of 26 bacteria and 35 phages in terms of palindrome avoidance. This analysis reveals that palindrome avoidance is not universally spread among bacterial species and that it does not correlate with taxonomic proximity. Palindrome avoidance is also not universal among bacteriophage, even when their hosts code for RMSs, and depends strongly on the genetic material of the phage. Interestingly, palindrome avoidance is intimately correlated with the infective behavior of the phage. We observe that the degree of palindrome and restriction site avoidance is significantly and consistently less important in phages than in their bacterial hosts. This result brings to the fore a larger selective load for palindrome and restriction site avoidance on the bacterial hosts than on their infecting phages. It is then consistent with a view where type II RMSs are considered as parasites possibly at the verge of mutualism. As a consequence, RMSs constitute a nontrivial third player in the host–parasite relationship between bacteria and phages.

Classic type II restriction modification systems (RMSs) comprise pairs of enzymes with matching DNA sequence specificity. The modification enzyme is a DNA methyltransferase that specifically methylates either adenosyl or cytosyl residues within the recognition sequence (the restriction site; RS), thus making DNA resistant to the restriction activity. The restriction enzyme is an endodeoxyribonuclease that cleaves DNA at a precise location within or around the recognition sequence, when this sequence is not methylated (Redaschi and Bickle 1996). These recognition sequences are symmetrical, comprising 4–8 specific base pairs, and cleavage and methylation occur symmetrically within the sequences. As a consequence, foreign double stranded DNA (dsDNA), unmethylated at the restriction sites recognized by the cell's RMS, is quickly degraded.

Although genes dealing with information processing are rarely horizontally transferred (Jain et al. 1999) and evolve slowly (Woese 1998), RMSs seem to be very frequently exchanged between species (Kita et al. 1999; Kobayashi et al. 1999; Rocha et al. 1999), and to evolve very quickly (Lauster 1989; Jeltsch and Pingoud 1996). This transfer is so frequent that more than 50 different RMSs specificities among natural strains of *Escherichia*

coli have already been found (Roberts and Macelis 2000). Although type II RMSs consist of separate restriction and modification enzymes acting independently of each other, the two genes are usually linked in the chromosome (Wilson 1991). This characteristic suggests tight coregulation and makes horizontal transfer of RMSs much easier.

RMSs were first identified as a cause for the retardation or prevention of phage infection. Consequently, an evolutionary role of "defense tool" was ascribed (Arber 1965; Rambach and Tiollais 1974). The widespread presence of these systems in the bacterial world would be a consequence of the selective advantage of having a defense tool against phage infection. Nevertheless, phages have a non-negligible chance of invading the cell, escaping the action of the RMS, and thereby becoming resistant to RMS by replicating correctly methylated copies of their genetic information. Hence, the intraspecific variety of RMSs would be a natural consequence of selection for variability.

RMSs have several important limitations as defense tools. They act on dsDNA templates and therefore, RNA and many single-strand DNA (ssDNA) phages infect bacteria with relative impunity (Levin 1993). Bacteriophages also use a plethora of strategies to reduce RMSs efficacy, such as inhibition of restriction enzymes, nucleotide base modification, and phage-encoded methylation (Krüger and Bickle 1983; Birge 1994; Campbell 1996). An important argument

[5]Corresponding author.
E-MAIL erocha@abi.snv.jussieu.fr; FAX 33 1 44 27 63 12.

against the purely defensive role of RMSs has been presented by Korona and Levin (Korona et al. 1993; Korona and Levin 1993) as follows: Depending on the cell's RMS, dsDNA phages have a probability of 10^{-6} to 10^{-2} of being correctly methylated before being subject to restriction (Korona et al. 1993). Therefore, in the typically large bacterial and phage colonies, it is certain that at least one phage will successfully invade at least one bacteria. This successful phage produces correctly methylated copies of its chromosome, thereby creating the conditions for the immediate invasion of the clonal population (Korona and Levin 1993). Another argument derives from the existence of rare-cutter RMSs recognizing RS of 8 bp that are typically absent phage genomes because of their small sizes (Naito et al. 1995). Even more striking is the requirement for two recognition sites in the sequence of many of these 8-bp RMSs (Bilcock and Halford 1999). Given these difficulties, it seems troublesome to explain the maintaining of such systems based only on the defense hypothesis.

From these observations, Kobayashi and colleagues (1995) suggested that RMSs might be considered as selfish genetic elements that invade genomes without necessarily providing selective advantages (Naito et al. 1995). In fact, they have shown that once RMSs are acquired they become essential for the survival of the bacteria (Kusano et al. 1995). If the genome loses the RMS, the long half-life of the nuclease (by comparison to the methylase) will eventually lead to the cell's death. Therefore, bacterial chromosomes become dependent on the invading RMS. RMSs have been shown to enhance plasmid segregation stability in *E. coli* and *Bacillus subtilis* as a result of this property (Kulakauskas et al. 1995; Handa et al. 2000).

Early statistical studies revealed that palindromes and RS of size 4 and 6 are avoided in many dsDNA phage genomes (Sharp 1986; Blaisdell et al. 1996). This was taken as evidence in favor of the defense hypothesis, as experimental work indicated that RS avoidance is the best phage strategy to avoid the action of RMSs (Korona et al. 1993). However, further studies revealed that RS avoidance is characteristic of phages and their bacterial hosts (Karlin et al. 1992, 1997). In fact, strong RS avoidance is found in phages, eubacteria, and archaea, and it is well correlated with the host's RMSs (Sharp 1986; Burge et al. 1992; Gelfand and Koonin 1997).

Because RS avoidance is the best way to escape the action of RMSs and as statistical analysis have revealed that RMSs strongly shape the abundance of RS in genomes, one may consider RS avoidance as a measure of the mutational load imposed by RMS on the genome. Bacterial and phage genomes are better off if they avoid RS (therefore, there is a benefit in avoidance), but these mutations are not neutral with regard to other properties (therefore, there is a trade-off). It follows

that if RMSs impose a larger selective load on bacteria than on bacteriophages, then the parasitic hypothesis looks more plausible. On the other hand, if avoidance is more important on phages than on bacteria, one may think that the defense role of RMSs carries a sufficient advantage to be maintained in the population. Note, that for this discussion we consider that phages, bacteria, and RMSs have coevolved for a long time, which seems reasonable (Lauster 1989; Hendrix et al. 1999).

We have sought to characterize the impact of RMSs on the genomes of bacteria and phages in terms of palindrome avoidance. To infer ecological and evolutionary implications on RMSs, we have designed a test to identify which of the two evolutionary theories seems more plausible.

Avoidance is observed in phages and it has been suggested that in some circumstances RMSs may play an important selective role, for example, in the colonization of new habitats (Levin 1993). Therefore, we have tried to shift the current discussion from a purely selfish versus a purely utilitarian theory, to a more contemporary view of the host–parasite relationship. This goes in the direction of current theories about the evolution of mutualism, from parasitism exploring the paths between conflict and cooperation (Herre 1999). The existence of selfish genes and multiple levels of selection units have been the subject of extensive debate (Dawkins 1976; Doolittle and Sapienza 1980; Wilson and Sober 1994; Depew and Weber 1995). Although a general discussion on this theme is far from being the main topic of this paper, our data bring forth some new evidence favoring a change of emphasis on the evolutionary theory (i.e., from the action of selection units strictly at the individual level, to selection units acting at different hierarchical levels, from the gene to the group [Gould and Lloyd 1999]).

RESULTS

Palindrome Avoidance in Bacterial Genomes

We have computed word biases using Markov chains for all words of size 4 and 6 and for all complete genomes of bacteria and phages. The comparison between observed and expected counts (under the Markov model) was done through the use of Z-value statistics (Schbath et al. 1995; see Methods for statistical details). Because other models and statistics have been proposed to calculate these biases, we have also tested the O/E Markov approach and the τ index (Karlin et al. 1997). They provided some differences in the identification of individual biased palindromes, but yielded similar results when comparing the population of palindromes to the other words. It is important to emphasize that our method does not aim at identifying individual biased oligonucleotides, but at

comparing the relative biases of two populations (e.g., differences between palindromes and the other words). Therefore, after computing the oligonucleotides' biases, we compared the average avoidance of palindromes to the remaining words of the same size using the Wilcoxon test (a nonparametric test for comparison of means). We found significant palindrome avoidance in the majority, but not all, of the bacterial genomes (Table 1). The presence of significant avoidance does not have a clear taxonomic pattern. Within the six archaea analyzed, two substantially avoid palindromes of size 4 and 6 (*Archaeoglobus fulgidus* and *Methanococcus jannaschii*), and three do not avoid them (*Aeropyrum pernix*, *Pyrococcus horikoshii*, and *Methanobacterium thermoautotrophicum*). *Pyrococcus abyssi*, although taxonomically very close to *P. horikoshii*, moderately avoids palindromes of size 6. Similarly, among

the Gram-positive bacteria, *B. subtilis* and *Mycoplasma genitalium* avoid palindromes of size 4 and 6, *Mycobacterium tuberculosis* and *Mycobacterium leprae* do not avoid palindromes of any size, *Mycoplasma pneumoniae* avoids only palindromes of size 6 and *Ureaplasma urealyticum* only avoids palindromes of size 4. Finally, within proteobacteria all species (*E. coli*, *Pseudomonas aeruginosa*, *Helicobacter pylori*, *Haemophilus influenzae*, *Neisseria meningitidis*, *Campylobacter jejuni*, and *Rickettsia prowazekii*) avoid palindromes of size 4 and 6, and within Chlamydia, both species avoid palindromes of size 4 but not of size 6. Therefore, palindrome avoidance is widespread among prokaryotes but can occur quite independently of taxonomic proximity. Moreover, avoidance of palindromes of size 4 does not necessarily imply avoidance of palindromes of size 6, and vice versa (Table 1).

There are different degrees of palindrome avoidance among the set of bacterial genomes that significantly avoid palindromes. The ratio of expected-to-observed palindrome avoidance reveals that *H. influenzae* is consistently the most biased of all genomes in terms of palindrome avoidance of both size 4 and 6 words (Table 1).

RS Scarceness Within Palindrome Avoidance

Restriction sites are never significantly less avoided than the remaining palindromes (Table 1). Instead, within the genomes that avoid palindromes, RS are more avoided than the other palindromes in two of nine genomes for words of size 4 and in four of six genomes for words of size 6. The difference between the RS of length 4 and 6 is probably due to the existence of six bacteria that have only one known restriction site of size 4 (only two bacteria for size 6). For these cases the difference between the RS and the remaining palindromes cannot be statistically asserted (at 1%) using the Wilcoxon test. To circumvent this problem we have designed a very simple alternative statistical test: We assign a value of 1 to the genome if the average Z-value is smaller for RS than for the median palindrome, otherwise the value assigned is 0. All nine bacteria with known RS of size 4 have a value of 1 using this evaluation, which is statistically significant at 5‰ (binomial test). Therefore, with this test we can assert that RSs are generally more avoided than the remaining palindromes.

Comparative Avoidance of A + T and G + C Palindromes

To determine whether palindrome avoidance is the result of secondary structure avoidance, we divided the palindromes into G + C palindromes (all nucleotides G or C), A + T palindromes (all nucleotides A or T), and others. We then tested whether the average rank of G + C palindromes was smaller than the average rank

Table 1. Analysis of the Avoidance of Palindromes and RS of Length 4 and 6 in Bacterial Genomes (Wilcoxon Tests)

Bacteria	Length 4			Length 6		
	Palindromes bias	O/E	RMS bias	Palindromes bias	O/E	RMS bias
aepe	0	0.78	NA	0	0.92	NA
aqae	—	0.46	NA	—	0.63	NA
arfu	—	0.48	NA	—	0.53	NA
basu	—	0.41	-	—	0.58	-
bobu	—	0.52	NA	0	0.89	NA
caje	—	0.51	NA	—	0.66	NA
chpn	—	0.43	NA	0	1.06	NA
chtr	—	0.40	NA	0	1.12	NA
esco	—	0.44	4*	—	0.44	—
hain	—	0.19	-	—	0.30	-
hepy	—	0.45	0	—	0.46	NA
meja	—	0.47	0	—	0.78	NA
meth	0	1.05	0	0	0.88	NA
myge	—	0.57	NA	—	0.76	NA
myle	0	1.07	NA	0	0.88	NA
mypn	0	0.78	NA	—	0.72	NA
mytu	0	0.95	NA	0	0.90	NA
neme	—	0.40	2*	—	0.35	0
psae	—	0.46	NA	—	0.63	—
pyab	0	0.91	NA	-	0.83	NA
pyho	0	0.93	NA	0	0.90	NA
ripr	—	0.54	NA	—	0.73	NA
sysp	—	0.62	5*	—	0.41	0
thma	-	0.67	NA	-	0.83	NA
trpa	-	0.72	2*	-	0.86	NA
urur	-	0.70	NA	0	0.91	NA

Palindrome bias is the test that palindromes are more biased than the remaining words. O/E displays the ratio observed/expected of the mean rank of palindromes sorted by decreasing avoidance. RMS avoidance is the result of the Wilcoxon test that RS are more biased than the remained palindromes. Abbreviations: NA, unknown RMS on the species; n*, rank of the sole RS known in the species for 4-palindromes (not enough elements for nonparametric statistics); —, underrepresentation (*P*-value <0.001); -, underrepresentation (*p*-value <0.05); 0, no bias. See Methods for species abbreviations.

for A + T palindromes. Results indicate that this is not so, because G + C palindromes are more avoided in 53% (size 4) and 44% (size 6) of the cases (correspondingly, 47% and 56% for A + T palindromes). These differences (from 50%) are not statistically significant (using a χ^2 test).

Palindrome Avoidance in Phage Genomes

RS avoidance in phage follows the same trend as in the bacterial genomes. Palindromes are frequently avoided and RSs are even more avoided than the other palindromes (Table 2). This avoidance depends on the bacterial host and differs significantly with the nature of the phage. Many phage infecting bacterial species that

avoid palindromes also avoid palindromes (e.g., many phages of *B. subtilis*, *E. coli*, and *H. influenzae*), although this is not always true (e.g., fr, GA, MS2, NL95). Moreover, some phages only avoid palindromes of a given size (e.g., SPβc3, G4, φK). These different tendencies should not be due to small sample effects (i.e., statistical artifacts), because they are observed in some of the largest genomes (e.g., SPβc3 and PBSX). As expected, most phages hosted by species that do not avoid palindromes also do not avoid palindromes. Such is the case of *M. thermoautotrophicum* and *M. tuberculosis* (for size 4). Exceptions to this rule are provided by the two *M. tuberculosis* phages D29 and L5, which strongly avoid palindromes of size 6.

Table 2. Analysis of the Avoidance of Palindromes and RS of Length 4 and 6 in the Genomes of Phages

Phage	Host	Length (bp)	Type	%G + C	Length 4 Palindromes bias	Length 4 O/E	Length 4 RMS bias	Length 6 Palindromes bias	Length 6 O/E	Length 6 RMS bias
PBSX	**Basu**	27614	dsDNA	44.9	—	0.63	0	0	0.91	0
SPβc3	**Basu**	134416	dsDNA	34.6	—	0.63	0	0	0.89	0
PZA	**Basu**	19366	dsDNA	39.7	-	0.67	0	—	0.57	0
α3	**Esco**	6087	ssDNA	45.2	—	0.59	1*	0	0.94	0
f1	**Esco**	6407	ssDNA	40.9	—	0.59	1*	—	0.56	0
fd	**Esco**	6408	ssDNA	40.9	-	0.65	1*	—	0.56	-
fr	**Esco**	3575	ssRNA	51.4	0	1.01	7*	0	1.00	0
G4	**Esco**	5577	ssRNA	45.7	-	0.68	1*	0	0.92	0
GA	**Esco**	3466	ssRNA	47.9	0	0.84	16*	0	0.90	0
I2-2	**Esco**	6744	ssDNA	42.7	—	0.57	4*	—	0.65	0
If1	**Esco**	8454	ssDNA	43.7	—	0.44	2*	—	0.67	-
Ike	**Esco**	6883	ssDNA	39.5	-	0.72	3*	—	0.69	-
λ	**Esco**	48502	dsDNA	49.9	—	0.42	1*	—	0.38	—
MS2	**Esco**	3569	ssRNA	52.1	0	0.83	14*	0	1.04	0
MX1	**Esco**	4215	ssRNA	28.6	0	0.89	6*	0	1.03	0
Mu	**Esco**	36717	dsDNA	52.1	—	0.46	1*	—	0.51	—
N15	**Esco**	46375	dsDNA	51.2	—	0.60	1*	—	0.70	-
NL95	**Esco**	4248	ssRNA	50.8	0	1.01	1*	0	1.06	0
P2	**Esco**	33593	dsDNA	51.2	-	0.70	1*	—	0.66	-
P4	**Esco**	11624	dsDNA	49.5	-	0.70	1*	—	0.77	—
φK	**Esco**	6089	ssDNA	45.0	—	0.54	1*	0	0.85	0
φX174	**Esco**	5386	ssDNA	44.8	—	0.38	1*	—	0.67	0
PRD1	**Esco**	14925	dsDNA	47.1	—	0.42	1*	—	0.46	—
S13	**Esco**	5386	ssDNA	44.3	—	0.41	1*	0	0.64	-
T3	**Esco**	19680	dsDNA	50.6	—	0.51	1*	—	0.44	-
T4	**Esco**	168900	dsDNA	35.3	—	0.55	1*	—	0.78	0
T7	**Esco**	39937	dsDNA	48.4	—	0.33	1*	—	0.53	-
HP1	**Hain**	32355	dsDNA	40.0	—	0.24	-	—	0.63	0
PsiM2	**Meth**	26111	dsDNA	46.3	0	1.08	0	0	0.87	NA
D29	**Mytu**	49136	dsDNA	63.5	0	0.95	NA	—	0.36	NA
l5	**Mytu**	52297	dsDNA	62.2	0	0.92	NA	—	0.46	NA
PF1	**Psae**	7349	ssDNA	61.5	0	1.03	NA	—	0.74	0
Pf3	**Psae**	5833	ssDNA	45.4	+	1.29	NA	—	0.69	0
φCTX	**Psae**	35559	dsDNA	62.6	0	1.08	NA	-	0.84	0
PP7	**Psae**	3588	ssRNA	54.2	0	1.12	NA	0	1.02	0

Palindrome bias is the test that palindromes are more biased than the remaining words. O/E displays the ratio observed/expected of the mean rank of palindromes sorted by decreasing avoidance. RMS avoidance is the result of the Wilcoxon test that RS are more biased than the remained palindromes.
Abbreviations: NA, unknown RMS on the species; n*, rank of the sole RS known in the species for 4-palindromes (not enough elements for nonparametric statistics); —, underrepresentation (*P*-value <0.001); -, underrepresentation (*P*-value <0.05); 0, no bias. See Methods for species abbreviations.

The extent of palindrome avoidance in phages depends on the type of phage. RNA phages do not avoid palindromes (Fig. 1), whereas most DNA phages avoid palindromes whether they are ssDNA (75%) or dsDNA (81%). Classification of ssDNA phages into filamentous and isometric shows that filamentous phages avoid palindromes more often than do isometric phages (Fig. 1). When analyzing the genomes of dsDNA phages in terms of their infective behavior (temperate versus virulent), we find that 92% of the virulent and 83% of the temperate phages avoid palindromes (difference not statistically significant, χ^2 test).

Bacterial Chromosomes Avoid Palindromes to a Greater Extent Than do Their Phages

Using the same methodology as described previously, the comparison of the ranks for palindromes in bacteria and phages reveals that phage genomes avoid palindromes to the same or to a lesser extent than the bacterial chromosome (Table 3). However, as phages have smaller genomes, one may be concerned about possible statistical artifacts due to poorer estimation of Markov transition probabilities. To verify that the differences in genome sizes are not producing these statistical artifacts, we have developed a complementary analysis (see Methods). The first approach is a comparison of the ranks of palindromes in the phage and in its host. The second approach consisted in the comparison of the rank of palindromes in the phage with the rank of 100 sequences of the same size, randomly sampled from the bacterial chromosome. In this second approach both samples have the same size, and poorer determination of Markov probabilities should not bias the results in favor of bacteria. Because both
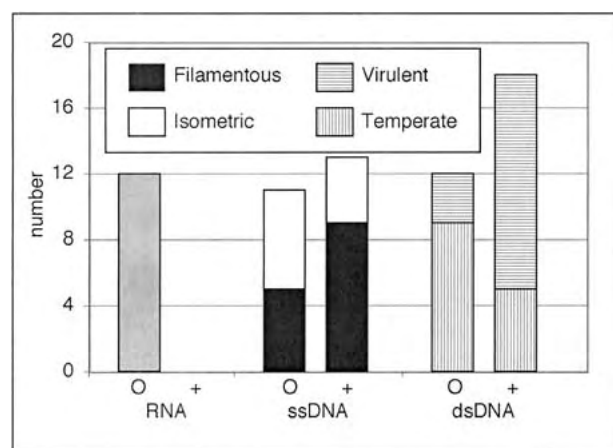
Table 3. Comparison of Palindrome Avoidance in Phages and Respective Bacterial Hosts

Phage	Length 4		Length 6	
	PB	OEp/OEc	PB	OEp/OEc
PBSX	+	1.54	++	1.58
SPBc3	+	1.53	++	1.54
PZA	0	1.63	0	0.98
α3	0	1.35	++	2.15
f1	0	1.34	0	1.28
fd	0	1.47	0	1.28
fr	++	2.30	++	2.27
G4	+	1.54	++	2.08
GA	+	1.90	++	2.06
I2-2	0	1.29	++/+	1.49
If1	0	1.00	++	1.53
Ike	++	1.65	++	1.58
λ	0	0.94	0	0.87
MS2	+	1.88	++	2.36
MX1	+	2.01	++	2.33
Mu	0	1.05	+	1.16
N15	+	1.37	++/+	1.59
NL95	++	2.30	++	2.41
P2	+	1.59	++/+	1.50
P4	+	1.59	++	1.74
ϕK	0	1.23	++	1.93
ϕX174	0	0.86	++/+	1.52
PRD1	0	0.95	0	1.05
S13	0	0.93	+	1.45
T3	0	1.15	+/0	1.00
T4	0	1.25	++	1.77
T7	0	0.75	+/0	1.19
HP1	0	1.24	++	2.11
PsiM2	0	1.03	0	0.99
D29	0	1.00	—	0.40
L5	0	0.96	—	0.51
Pf1	++	2.25	0	1.17
Pf3	++	2.80	0	1.09
ϕCTX	+	2.35	+/0	1.33
PP7	++	2.44	++/+	1.63

PB displays the results of the tests if palindromes are more avoided in phage (— and -) or in bacteria (+ and ++). When the two alternative tests (see Methods) give different results, they are indicated by a slash. Column OEp/OEc is the ratio of observed to expected number of restriction sites on phages, to the same ratio on bacteria. The description of the remaining abbreviations can be found in the Table 1 footnote.

Figure 1 Avoidance of palindromes per type of phage: RNA, ssDNA, and dsDNA. SsDNA phages are divided into filamentous and isometric, and dsDNA phages into virulent and temperate. Each phage counts twice (for the analysis of size 4 and 6 palindromes), (O) no significant avoidance; (+) significant avoidance (5%, Wilcoxon test).

analyses provided similar results (Table 3), and some of the largest phages reveal significantly less avoidance than the chromosome (SPβc3, PBSX, N15, and P2), we may safely conclude that bacterial chromosomes do avoid palindromes to a larger extent than their respective bacteriophages.

The proportion of DNA phages that avoid palindromes significantly less than the host is similar between ssDNA (50%) and dsDNA (50%) phages (Fig. 2). Within ssDNA phages, 36% of filamentous phages avoid palindromes less than the host, whereas this percentage increases to 60% for the isometric phages (Fig. 3). In terms of infective behavior of dsDNA phages, 23% of the virulent phages avoid palindromes signifi-
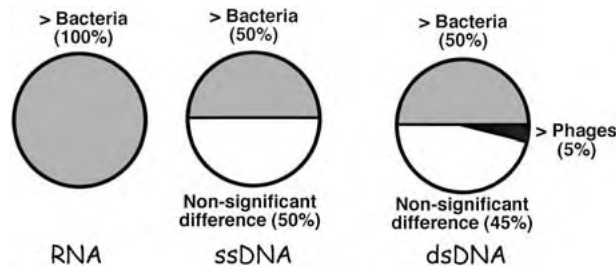
Figure 2 Avoidance of palindromes in phages (per phage type) relative to their bacterial hosts. (> bacteria) larger avoidance in bacteria; (> phages) larger avoidance in phages.

cantly less than their hosts, but for temperate phages this proportion goes up to 68% ($P < 0.01$, χ^2) (Fig. 3).

DISCUSSION

Palindrome Avoidance as a Consequence of RMS

The following observations suggest that palindrome avoidance is caused by RMS. (1) Contrary to prokaryotes and bacteriophages, it has been observed that eukaryotes, eukaryotic DNA viruses, chloroplasts, and mitochondria do not significantly avoid RSs or palindromes (Sharp 1986; Karlin et al. 1992; Gelfand and Koonin 1997). None of these systems contains RMSs. (2) The avoidance of small palindromes is restricted to words of the size typical of RSs (Rocha et al. 1998). (3) Some archaea possess RMSs and they also avoid palindromes (Gelfand and Koonin 1997), although archaea

Figure 3 Avoidance of palindromes of phages (per phage infective type) relative to their bacterial hosts. (NS) nonsignificant difference; (> bacteria) larger avoidance in bacteria; (> phages) larger avoidance in phages.

deal with information in a more "eukaryotic way" (Doolittle and Logdson 1998). (4) One can link DNA phage (but not RNA phage) avoidance of RS with its host genome RMSs (this study; Sharp 1986; Blaisdell et al. 1996).

Through our analysis we can add some more arguments. (5) With one single exception (*M. thermoautotrophicum* for four palindromes), all genomes known for having RS present palindrome avoidance of words of the same length. (6) Known RSs are always at least as much avoided as the remaining palindromes, and frequently they are more avoided. (7) Phages tend to follow the pattern of avoidance of their hosts not only in terms of RS, but also in terms of other palindromes. It is difficult to conceive of some other factor that acts so strongly on the genome to avoid most palindromes frequently found to be RS (see also below). Also, this characteristic is shared by organisms of different phyla, but not among much closer groups. For example, among the Gram-positive bacteria *B. subtilis* and *M. genitalium* avoid palindromes but *M. tuberculosis* and *M. leprae* do not, and *M. pneumoniae* and *U. urealyticum* do it only partially. Interestingly *R. prowazekii* and *Synechocystis* sp. strongly avoid palindromes, whereas mitochondria and chloroplasts do not. Requirements concerning fundamental and well-conserved processes like transcription, replication, or repair should not exhibit such diversity. This also suggests a mobile mechanism for palindrome avoidance along with previous results concerning frequent horizontal transfer of RMSs (Lauster 1989; Jeltsch and Pingoud 1996; Kita et al. 1999).

Other Possible Reasons for Palindrome Avoidance

Four other reasons have been commonly given to explain the avoidance of small palindromes (Karlin et al. 1992, 1997; Hénaut et al. 1996; Gelfand and Koonin 1997): DNA structural requirements, deamination of methyl-cytosine, definition of regulatory regions, and action of Dam-methylase. We shall briefly discuss them and show that they are probably not at the origin of the general bias for palindrome avoidance.

CG compression is a frequent source of sequencing errors due to G:C pairing in single-stranded sequence. One might suppose that G + C-rich palindromes would be avoided to prevent such pairing in vivo. This would imply that part of the palindrome avoidance would have nothing to do with the existence of RMSs. However, closely related bacteria avoid palindromes to a different extent, which would not be expected from such a fundamental physicochemical constraint. Nevertheless, we tested this hypothesis from the point of view of two expected consequences. First, one would expect six palindromes to be more avoided than four palindromes, because structures would be longer, and
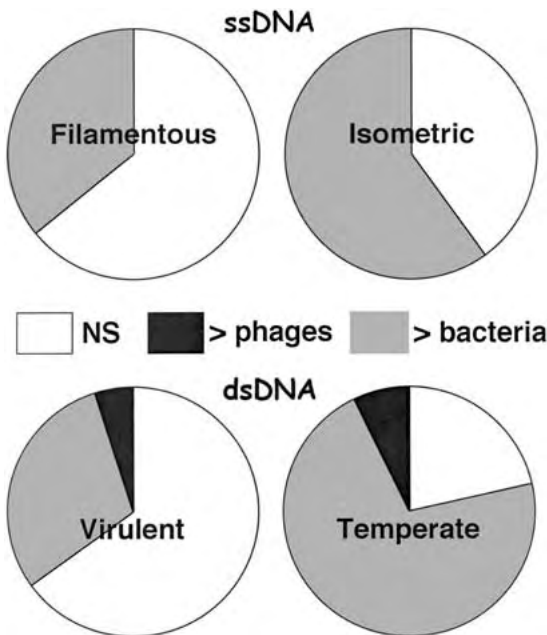
therefore, stronger. This is not the case. Second, one would expect G + C-rich palindromes to be more avoided than A + T-rich palindromes, as they would provide more stable structures. This is not observed either.

Methylases of type II RMSs methylate either cytosine or adenine. Methyl-cytosine is a well-known mutational hotspot because of its hydrolytic deamination that provokes a $C \rightarrow T$ mutation (Lindahl 1993). One might suppose that C-methylated palindromes would tend to disappear due to this mutational load and not as a result of selection pressure on RS avoidance. However, this is contradicted by the observation that A + T palindromes are avoided to the same extent as G + C palindromes.

In *E. coli* GATC is the only known restriction site of size 4, but it is also used by MutH to recognize the newly formed strand in postreplicative mismatch repair (Marians 1992). Because excessive GATC avoidance may render difficult the recognition of the neoformed branch by the mismatch repair system, the results concerning this particular site cannot be interpreted easily (Blaisdell et al. 1996). It has been shown that GATC undermethylation in enterobacteriophages is only deleterious if the MutH endonuclease is active (Deschevanne and Radman 1991). The counterselection of GATC in phages may also be explained by the frequent undermethylation of GATC sites in phages as a result of the low level of Dam methylase in the host, coupled to the rapid phage replication (Hénaut et al. 1996). This may explain why some phages avoid GATC more than *E. coli* (although this difference is never statistically significant). Because the Dam/MutH system seems to be exclusive to enterobacteria and involves just one particular site, it cannot explain widespread palindrome avoidance among bacterial species.

Palindromic sites are frequently involved in protein–DNA interactions at regulatory sites (Mironov et al. 1999). Therefore, one might suppose that bacteria would avoid these sites in the remainder of their genome to optimize promoter recognition, and that phages would avoid these sites to escape coregulation with the host (Blaisdell et al. 1996; Karlin et al. 1997). To test this hypothesis it would be important to study not only the avoidance, but the distribution of palindromes. The study of this distribution falls outside the scope of this article. However, it was shown that palindrome avoidance is present both in genes and in intergenic regions (Rocha et al. 1998) and also that there is a smaller avoidance of palindromes in intergenic regions (Karlin et al. 1992). There is no contradiction between these two observations. Palindrome elimination at regulatory regions that require them is counterselected, and therefore, RS avoidance should be smaller at intergenic regions.

Why Should (Almost) all Palindromes be Avoided?

For phages, RS avoidance seems to be the best strategy to escape RMSs (Korona et al. 1993). A smaller number of RSs makes restriction more unlikely, allows faster full methylation during phage replication, and avoids gene expression problems related to DNA methylation (Reisenauer et al. 1999). Phage may infect many different strains of bacteria carrying different RMSs. The action of these different selective loads leads to the avoidance of all RSs with which a phage was in recent contact. Because almost all palindromes are known to be RS in bacteria (Roberts and Macelis 2000), most palindromes are expected to be avoided. A slightly different reasoning applies to the bacterial avoidance of RS. In this case it is the action of the bacteria's own RMS that produces a selective load toward the avoidance of RS. Because RMSs are very frequently switched and horizontally transferred (Lauster 1989; Jeltsch and Pingoud 1996), bacteria are selected to avoid many different RSs.

Independently of considering RMS as a defense tool or as a molecular parasite, our statistical analysis suggests a frequent degradation of the bacterial chromosome by its own RMS, as bacteria carrying these systems strongly avoid RS. After the acquisition of a new RMS, there is a selective pressure toward RS avoidance and the bias should be established relatively quickly. When the RMS is lost, the bias will fade away mostly by random drift. Because this drift is expected to be nearly neutral and bacterial mutation rates are low ($\sim 10^{-9}$) (Drake et al. 1998), the bias will be slowly lost. As a consequence, vestiges of the presence of very ancient RMSs may still persist.

Predicting New RMSs Through Palindrome Avoidance

If one considers that RMSs are the cause of palindrome avoidance, one may speculate that highly avoided palindromes are likely to be, or have been, RSs of the species. From the analysis of the most underrepresented palindromes we have compiled lists of the five most avoided palindromes that are not known RSs in the species (Table 4). In our opinion this list gathers the most likely candidates for novel RS specificity in strains of each species.

Relative Selective Load of RS Avoidance in Phages and Bacteria

As expected, most phages hosted by species that do not avoid palindromes also fail to exhibit palindrome avoidance. This is the case for *M. thermoautotrophicum* and *M. tuberculosis* (for size 4). The sole exception is provided by the two *M. tuberculosis* phages D29 and L5, which strongly avoid palindromes of size 6. Because *M. tuberculosis* does not avoid palindromes, no RMSs are known in the species, and its complete sequence failed to reveal functional RMSs (Cole et al. 1998); this ob-

Table 4. Known and Potential Restriction Sites (RS) of Size 4 and 6 in Bacterial Genomes

Species	%G + C	Length 4 RS	Length 4 Potential RS	Length 6 RS	Length 6 Potential RS
aepe	56.5	?	NA	?	NA
aqae	43.5	?	TCGA, GGCC, CTAG, CCGG, GCGC	?	CGTACG, AGTACT, CTCGAG, GATATC, GAGCTC
arfu	48.6	?	GATC, CTAG, CGCG, GCGC, GGCC	?	CTCGAG, GAGCTC, GTCGAC, AAGCTT, CAGCTG
basu	43.5	GGCC, CCGG, CGCG, TCGA	AATT, CATG, GATC, TGCA, ACGT	ATCGAT, CTGCAG, GGATCC, TCCGGA	AAATTT, ATATAT, GAGCTC, CAGCTG, AATATT
bobu	28.6	?	TATA, AATT, GTAC, CATG, ACGT	?	NA
caje	30.5	?	GCGC, ACGT, GATC, CATG, GGCC	?	GAATTC, GATATC, GAGCTC, ATGCAT, GTATAC
chpn	40.5	?	CTAG, ATAT, CCGG, GTAC, TGCA	?	NA
chtr	41.3	?	TATA, CTAG, AATT, ATAT, GTAC	?	NA
esco	50.8	GATC	GGCC, CTAG, CGCG, TATA, CATG	AAGCTT, AGCGCT, AGGCCT, AGTACT, ATGCAT, CACGTG, CCGCGG, CGGCCG, GAATTC, GAGCTC, GATATC, GCCGGC, GCGCGC, GGCGCC, GGGCCC, GGTACC, TACGTA	CTGCAG, CTGCAG, TCCGGA, GCATGC, GTCGAC
hain	38.2	CGCG, GCGC, CATG, CCGG	GGCC, TATA, TTAA, AATT, TCGA	AAGCTT	GTTAAC, TAATTA, GAATTC, GATATC, GTGCAC
hepy	38.9	GATC, TCGA, GGCC	GCGC, ACGT, CGCG, AGCT, GTAC	?	GCGCGC, GATATC, GAATTC, GCTAGC, TCTAGA
meja	31.4	CTAG, GATC, GTAC	GGCC, CATG, GCGC, ATAT, TATA	?	GTTAAC, CATATG, GTATAC, GAGCTC, GATATC
meth	49.5	GATC	NA	?	NA
myge	31.7	?	TATA, AGCT, CATG, TGCA, CTAG	?	GAATTC, TAATTA, AAGCTT, GGATCC, GGTACC
myle	57.8	?	NA	?	NA
mypn	40.0	?	NA	?	CTTAAG, ATTAAT, AAGCTT, AAATTT, CTCGAG
mytu	65.6	?	NA	?	NA
neme	51.8	GATC	GGCC, CCGG, CATG, AATT, TGCA	CAGCTG	GGCGCC, CCGCGG, CGCGCG, GCGCGC, CTGCAG
psae	66.6	?	CTAG, CGCG, GTAC, GGCC, AGCT	AGATCT, CAGCTG, CCCGGG, CCGCGG, CTCGAG, CTGCAG, GCATGC, GGATCC	GAGCTC, CGATCG, CGGCCG, GTCGAC, AAGCTT
pyab	44.7	?	NA	?	GTTAAC, GCTAGC, GTCGAC, CCTAGG, AGGCCT
pyho	42.0	?	NA	?	NA
ripr	29.0	?	TATA, AATT, CTAG, TTAA, TGCA	?	TAATTA, GGTACC, GATATC, GGATCC, GAGCTC
sysp	47.7	CCGG	CGCG, GGCC, GCGC, TGCA, TATA	ATCGAT, CCGCGG, TGATCA, TTCGAA	GGCGCC, TGCGCA, GAATTC, CAGCTG, ACCGGT
thma	46.2	CGCG	CTAG, ATAT, TATA, AGCT, GATC	?	CACGTG, GTCGAC, CGATCG, AAATTT, ACGCGT
trpa	52.8	GATC	TATA, GGCC, CTAG, ACGT, ATAT	?	AAATTT, GCGCGC, CTGCAG, GAATTC, ACATGT
urur	25.5	?	TATA, TTAA, CATG, AGCT, TCGA	?	NA

NA, genomes where palindromes are not avoided.
?, unknown RS in the species.
See Methods for species abbreviations.

servation is not explainable by any of the two theories (defense or selfish). One may, therefore, speculate that because these two phages have a very wide range of hosts, infecting both fast-growing and slow-growing Mycobacteria (Fullner and Hatfull 1997; Ford et al. 1998), palindrome avoidance is motivated by the existence of RMS in some other host. Indeed, the existence of RMS has been reported in five different species of closely related Mycobacteria (Roberts and Macelis 2000).

Interestingly, we have found that ssDNA phages avoid palindromes and RS to the same extent as dsDNA. This is not expected if RMSs are only used at the time of cell invasion, as they are usually active only on dsDNA. However, at the replication stage ssDNA phages are made double-stranded using the cell's enzymatic machinery (Campbell 1996), and may be targeted by RMSs at that time.

E. coli ssDNA phages are typically classified as filamentous or isometric (Campbell 1996). We have observed that isometric phages avoid palindromes to a lesser extent than filamentous phages, both in absolute terms and in comparison with the host (Figs. 2, 3). This is consistent with observations pointing to high sensitivity of filamentous phages to the RMSs of the host (Krüger and Bickle 1983). Both types of phages replicate in a similar way (Birge 1994); therefore, the explanation for this observation is probably found in their different infective behavior. Isometric phage infection is fast and terminates by cell lysis, whereas filamentous phages can replicate for a long period of time, exporting the new phage particles without destroying the bacterial cell. This means that filamentous phages coexist with a functional cell that may contain a RMS. One may then suppose that RSs avoidance in filamentous phages is the result of the pressure of optimizing invasion capacities and minimizing restriction of the dsDNA forms at the replication stage.

dsDNA phages are typically divided into temperate and virulent phages (Campbell 1996). Our results indicate a larger avoidance of RSs in virulent phages. Again, in this case, one may speculate that their different infective behavior is the basis of the observed difference. Because temperate phages replicate horizontally (through infection) and vertically (through inheritance), they may rely less on the importance of a successful infection than the virulent phages that replicate exclusively in a horizontal fashion. Further work is necessary on experimental evolution relating phage fitness to RS avoidance to better understand these differences.

RMSs as Parasites at the Verge of Symbiosis

Many ecological interactions that are called either parasitic or mutualistic are complex mixtures of antagonistic and mutualistic aspects (Herre et al. 1999).

Because it seems unlikely that the interests of mutualists will ever be completely concordant, the number of conflicts will depend on the extent to which the survival and reproductive interests of the symbiont/parasite align with those of the host (Doebeli and Knowlton 1998). Many mutualistic interactions are thought to have arisen through previous parasitic interactions, and the line dividing parasitic and mutualistic behavior may be a very thin one (Herre et al. 1999). Moreover, it may be a changing line, depending on the environmental context. For example, although bacteriophages are unanimously regarded as parasites, lysogenic phages confer advantages in certain contexts, such as bacterial resistance to antibiotics (Stewart and Levin 1984). In a similar way, it is conceivable that RMS contribute positively to the fitness of bacteria in special circumstances (see below).

The difference between the two theories for the evolutionary role of RMSs relies on the relative load that they impose on the cell's survival. If the action of RMS against phages and plasmids is very efficient and chromosome restriction is a rare event, then RMSs are mostly beneficial for the bacterial cells and the defense hypothesis may constitute the best explanation for their evolutionary role. However, if action against phages is inefficient and chromosome restriction is frequent, then RMS impose such a negative load on the cell that a parasitic behavior provides the most satisfactory explanation for the widespread distribution of these elements in the bacterial world.

We have put forward a statistical analysis to decide between these two theories. Both bacteria and phages avoid extensively RSs; therefore, the simple analysis of both organisms does not result in the elimination of either hypothesis. Hence, one may primarily conclude that RMSs are a burden to bacteria and to phages. The question is then purely one of balance between the components of the system.

Because bacterial chromosomes avoid palindromes—and restriction sites in particular—more than phages, we are inclined to conclude that the most important load is on the bacterial chromosome. Hence, our data indicate that within the delicate balance between parasitic behavior and mutualistic interaction, RMS should be regarded as parasites, possibly at the verge of mutualism.

Just Selfish?

Along the balance between the parasitic and the mutualistic roles, there are several circumstances where RMSs may "cross the line" and be positively selected in bacteria. For example, an interaction between nonhomologous recombination, homologous recombination, and restriction was identified in a special type of nonhomologous recombination dependent on a type I RMS (Kusano et al. 1995). Previously, it had been pro-

posed that type II RMS might constitute a precious tool for natural genetic engineering (Arber 1991). RMS genes are associated with the artificial endpoints necessary for the alignment of two *H. pylori* genomes in three cases (Alm et al. 1999). This question is intimately related to the evolutionary relevance of bacterial recombination and, although it is difficult to conceive how this side effect alone could justify RMS evolution and maintenance, it may be quite important for the molecular evolution of bacteria (Levin 1993).

Within an ecological framework, it has been suggested that RMS might constitute a bacterial adaptation for invasion of new habitats in which phages are present, rather than to persist in them (Korona and Levin 1993). This advantage would come from its capacity to reduce the minimum colonization density required for an invading bacterial population. We do not know of any further experimental or theoretical work on this hypothesis.

Extensibility and Weaknesses of the Parasitic Theory

Type II RMSs composed of a methylase and a separate nuclease constitute the majority of known restriction systems. However, it is important to understand how other types of RMS might be placed in relation to the evolutionary theories just discussed.

In type I RMSs the methylase and the nuclease are subunits of the same protein. This means that the ratio of methylase to nuclease does not change over time and that a hypothetical parasitic behavior in these systems should not follow the patterns of type II RMSs. In fact, *Eco*KI, a type I RMS, does not exhibit selfish behavior and does not enhance plasmid maintenance (O'Neill et al. 1997). Type I RMSs are much less common than type II, and are frequently subject to antigenic variation either through slipped-mispair among simple sequence repeats or by switch systems similar to those used in antigenic variation (Barcus et al. 1995; Sitaraman and Dybvig 1997). These strategies, which seem typical of type I and type III RMS, are difficult to explain from the point of view of the parasitic theory, even if one supposes that the ancestral system might be a selfish one, for example, derived from a fusion of the two genes of a type II RMS. Also, it is not clear how bacteria such as *M. pulmonis* manage to survive such extensive variation on RMS specificity in bacteria (Sitaraman and Dybvig 1997). The extensive variation of type I and type III RMSs in *Mycoplasma* is even more puzzling, considering that we know very few phages and nearly no other genetic elements capable of entering into these (noncompetent) bacteria.

Function and distribution of type III RMSs are not well known, as very few examples have been identified. As for the type I RMSs, the nuclease and the methylase functions are in a single hetero-oligomeric protein and, in the presence of the required cofactors, both func-

tions compete in type III systems (Redaschi and Bickle 1996). The protein is composed of two modules of which both are required for restriction but only one for methylation. Therefore, it is likely that parasitism is not (or is no longer) the driving force of these systems.

The existence of systems for restricting only methylated DNA has also been reported, but very few examples are known (Redaschi and Bickle 1996). Their evolutionary role is intriguing. One speculates that they may constitute a system of defense against plasmids and bacteriophages encoding RMSs. Hence, they could constitute a defense utility against invasion of RMSs.

Conclusion

We have presented evidence in favor of a parasitic role, eventually at the verge of mutualism, regarding the widespread presence of type II RMSs in bacteria. We have also shown that palindrome avoidance is not an universal feature of bacterial genomes, neither archaeal nor eubacterial, because several genomes display the expected number of palindromes. Even more intriguing is the finding that some bacteria avoid very strongly the palindromes of just one particular length. This suggests that they have only acquired RMSs of a given size.

The problem of the evolutionary role of RMSs is interesting, not only in itself, but also because it is a manageable experimental system by which to study a complex population dynamics. In fact, there are three elements playing interconnected roles. Bacteria strive to prevent phage invasion, but also RS. Phages attack bacteria but are attacked by RMSs within bacteria. RMSs are simultaneously bacterial parasites and protectors against bacterial parasites.

Ultimately, the determination of the burden that RMSs impose on bacteria and phage must be determined by experimental bacterial evolution studies. Particularly relevant would be the determination of the burden that a RMS imposes on bacteria alone and on bacteria in the presence of bacteriophages.

METHODS

Data

We analyzed the following complete genomes (abbreviations in parentheses): *Aeropyrum pernix* (aepe), *Aquifex aeolicus* (aqae), *Archaeoglobus fulgidus* (arfu), *Bacillus subtilis* (basu), *Borrelia burgdorferi* (bobu), *Campylobacter jejuni* (caje), *Chlamydia pneumoniae* CWL029 (chpn), *Chlamydia trachomatis* (chtr), *Escherichia coli* (esco), *Haemophilus influenzae* (hain), *Helicobacter pylori* 26695 (hepy), *Methanococcus jannaschii* (meja), *Methanobacterium thermoautotrophicum* (meth), *Mycoplasma genitalium* (myge), *Mycoplasma pneumoniae* (mypn), *Mycobacterium tuberculosis* (mytu), *Mycobacterium leprae* (myle), *Neisseria meningitidis* MC58 (neme), *Pseudomonas aeruginosa* (psae), *Pyrococcus abyssi* (pyab), *Pyrococcus horikoshii* (pyho), *Rickettsia prowazekii* (ripr), *Synechocystis* spp C125

(sysp), *Treponema pallidum* (trpa), *Thermotoga maritima* (thma), *Ureaplasma urealyticum* (urur). We also analyzed the following phage complete genomes: PBSX, SPβc3, PZA (*B. subtilis*), α3, f1, fd, fr, G4, GA, I2–2, If1, Ike, λ, MS2, Mu, MX1, N15, NL95, P2, P4, φK, φX174, PRD1, S13, T3, T4, T7 (*E. coli*), HP1 (*H. influenzae*), psiM2 (*M. thermoautotrophicum*), D29, L5 (*M. tuberculosis*), Pf1, Pf3, φCTX, PP7 (*P. aeruginosa*). Sequences of all complete genomes were taken from National Center for Biotechnology Information Entrez Genomes (http://www.ncbi.nlm.nih.gov). Information on restriction sites of the different bacterial species was taken from REBASE v.904 (Roberts and Macelis 2000) (http://rebase.neb.com).

Word Counts

The analysis of the relative frequencies of palindromes was performed using statistics based on Markov chains (Karlin and Cardon 1994; Schbath et al. 1995; Rocha et al. 1998). Let us denote by $W = (w_1 w_2 \ldots w_m)$ the word made of the concatenation of the m nucleotides w_i, and $N(W)$ its observed count in a sequence. Under the Markov maximal order model, the expected count $E(W)$ of W is (Karlin et al. 1992):

$$E(W) = \frac{N(w_1 w_2 \ldots w_{m-1}) N(w_2 w_3 \ldots w_m)}{N(w_2 w_3 \ldots w_{m-1})} \qquad (1)$$

Several statistics have been proposed to compare this theoretical expectation with the observed value in a meaningful way (for review, see Leung et al. 1996). In this work we used the Z-value statistics that follows a reduced normal distribution for large sequences (Schbath et al. 1995):

$$z_w = \frac{N(W) - E(W)}{\sqrt{Var(W)}}, \qquad (2)$$

where $Var(W)$ represents the variance of $N(W)-E(W)$.

Z-values provide a statistically meaningful measure of the distance between the actual count of the word and the expected value. Stated differently, Z-values are a measure of the bias of the word, with values close to zero meaning no bias, large negative values meaning under-representation, and large positive values meaning over-representation of the word in the genomic text. This means that an abundant word may turn out to be regarded as under-represented if the two smaller words composing it are themselves very abundant. The most important and relevant difference between regarding frequency and bias is that bias is a direct measure of the mutation/selection load acting on the word, whereas frequency may be just a consequence of the composition of the word. Therefore, bias is to be regarded as the result of evolutionary pressures acting upon the frequency of the word, either mutational or selective.

For large sequences, and large counts, the variance for the Markov maximal case can be well approximated by (Schbath 1997):

$$Var(W) =$$

$$E(W) \frac{[(N(w_2 w_3 \ldots w_{m-1}) - N(w_1 w_2 \ldots w_{m-1})) (N(w_2 w_3 \ldots w_{m-1}) - N(w_2 w_3 \ldots w_m))]}{N(w_2 w_3 \ldots w_{m-1})^2} \qquad (3)$$

Because Z-values are distributed following a reduced normal distribution, the significance of the test is given by this distribution (e.g., $z = 3.29$ for 1‰ or $z = 1.96$ for 5% when doing bilateral tests). The counts and Z-values were computed for all genomes for all words of size 4 and 6.

Other indices have been proposed to compute the bias acting on the word, among which N(W)/E(W), and the Karlin τ (Karlin et al. 1992, 1997). In the first indexes N(W) and E(W) are as above, whereas in the second index, the expected value is computed to access log regression contingency interactions. Because these methods provide some different values of individual biases, we have tested if they granted similar results when testing if the population of all palindromes is more avoided than the remaining words. The use of N/E assessed nearly identical results. The τ formula is only available for words of size 4 and therefore, we could not test it for palindromes of size 6. However, for words of size 4 we have not observed very significant differences to our method in terms of biases of the populations of palindromes (data not shown).

Comparison of Biases

Word biases were compared within genomes and between the genomes of phages and their bacterial hosts. Comparison of biases within genomes aimed at answering the following questions: (1) Are palindromes more avoided within the genome than the remaining words of the same size? (2) Are known RS in the species more avoided in the genome than the remaining palindromes? Biases between genomes were computed to compare each phage with its host to answer the following questions: (1) Are palindromes more avoided in bacteria or in phages? (2) Are RS more avoided in bacteria or in phages?

Within genomes, rank comparisons were performed between the following sets for words of size 4 and 6: (1) palindromes versus nonpalindromes; (2) RSs versus the remaining palindromes. Rank comparisons were done using the Wilcoxon test, which is usually regarded as the most powerful robust nonparametric test for this purpose (Zar 1996). Rank comparisons between phages and bacterial genomes were performed separately for words of size 4 and 6. For each phage, we computed the difference of ranks of each palindrome in the phage with the ones of the host. Then, we performed a signed-rank test to check whether the average was significantly different to zero. We did the same for the known restriction sites in the host.

Whatever the model and statistics used, the significance of the deviation between the observed count of a word and its expected value depends on the number of counts. This happens because the statistical test is able to distinguish with a better accuracy a small significant deviation when the counts are larger (in other terms a statistical test is more powerful for large counts). Therefore, larger genomes will produce larger Z-values for similar biases, and for a given genome, the larger the word size, the less powerful will be the test, because there are fewer occurrences. This may give rise to difficulties when comparing large genomes (bacteria) to smaller ones (phages). This is why we prefer to use rank comparisons in all cases.

Rank comparisons should be quite robust to differences in effective sizes between the two sets. However, when the word counts become too small (very small phages), Markov probabilities are poorly estimated and even rank statistics may perform badly. To verify whether the large differences in effective sizes between bacterial and phage genomes was not biasing our results, we developed an additional approach. For each phage we computed the Z-values for all words of size 4 and 6, and ranked them. Then, we randomly extracted 100 sequences with the length of the phage genome, sampled

from the corresponding bacterial host genome. Subsequently, we compared the ranks of palindromes from the phage to the average ranks obtained from the samples of the bacterial chromosome. This was done using a signed-rank test on the difference of ranks for each palindrome. In this case the problems in estimating reliable Markov probabilities should be similar both for phages and for the 100 samples of bacterial chromosome. Moreover, we added a control to this test. We examined whether the 100 samples from the bacterial chromosome assigned significant palindrome avoidance as the analysis of the complete genome. This was always the case (data not shown). The comparison between both approaches for the determination of whether the bias is stronger in phages or in bacteria, also revealed nearly identical results, thereby indicating that our conclusions do not derive from statistical artifacts.

ACKNOWLEDGMENTS

REFERENCES

Alm, R.A., Ling, L.-S.L., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., De Jonge, B.L., et al. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397:** 176–180.

Arber, W. 1965. Host specificity of DNA produced by *E. coli*. V. The role of methionine in the production of host specificity. *J. Mol. Biol.* **11:** 247–256.

———. 1991. Elements of microbial evolution. *J. Mol. Evol.* **33:** 4–12.

Barcus, V.A., Titheradge, A.J., and Murray, N.E. 1995. The diversity of alleles at the *hsd* locus in natural populations of *Escherichia coli*. *Genetics* **140:** 1187–1197.

Bilcock, D.T. and Halford, S.E. 1999. DNA restriction dependent on two recognition sites: Activities of the SfiI restriction-modification system in *Escherichia coli*. *Mol. Microbiol.* **31:** 1243–1254.

Birge, E.A. 1994. *Bacterial and bacteriophage genetics*, 3rd ed., Springer-Verlag, New York, NY.

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

Burge, C., Campbell, A.M., and Karlin, S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci.* **89:** 1358–1362.

Campbell, A.M. 1996. Bacteriophages. In Escherichia coli *and* Salmonella: *Cellular and molecular biology* (ed. H. Neidhardt et al.), pp. 2325–2338. ASM Press, Washington DC.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537–544.

Dawkins, R. 1976. *The selfish gene*. Oxford University Press, Oxford, UK.

Depew, D.J. and Weber, B.H. 1995. *Darwinism evolving*. MIT Press, Cambridge, MA.

Deschevanne, P. and Radman, M. 1991. Counterselection of GATC sequences in enterobacteriophages by the components of the methyl directed mismatch repair system. *J. Mol. Evol.* **33:** 125–132.

Doebeli, M. and Knowlton, N. 1998. The evolution of interspecific mutualisms. *Proc. Natl. Acad. Sci.* **95:** 8676–8680.

Doolittle, W.F. and Logdson, J.M. 1998. Archaeal genomics: Do archaea have a common heritage? *Curr. Biol.* **8:** R209–R211.

Doolittle, W.F. and Sapienza, C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601–603.

Drake, J.W., Charlesworth, B., Charlesworth, D., and Crow, J.F. 1998. Rates of spontaneous mutation. *Genetics* **148:** 1667–1686.

Ford, M.E., Sarkis, G.J., Belanger, A.E., Hendrix, R.W., and Hatfull, G.F. 1998. Genome structure of mycobacteriophage D2: Implications for phage evolution. *J. Mol. Biol.* **279:** 143–164.

Fullner, K.J. and Hatfull, G.F. 1997. Mycobacteriophage L5 infection of *Mycobacterium bovis* BCG: Implications for phage genetics in the slow growing mycobacteria. *Mol. Microbiol.* **26:** 755–766.

Gelfand, M.S. and Koonin, E. 1997. Avoidance of palindromic words in bacterial and archaeal genomes: A close connection with restriction enzymes. *Nucleic Acids Res.* **25:** 2430–2439.

Gould, S.J. and Lloyd, E.A. 1999. Individuality and adaptation across levels of selection: How shall we name and generalise the unit of Darwinism? *Proc. Natl. Acad. Sci.* **96:** 11904–11909.

Handa, N., Ichige, A., Kusano, K., and Kobayashi, I. 2000. Cellular responses to postsegregational killing by restriction-modification genes. *J. Bacteriol.* **182:** 2218–2229.

Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I., and Danchin, A. 1996. Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. *J. Mol. Biol.* **257:** 574–585.

Hendrix, R.W., Smith, M.C.M., Burns, R.N., Ford, M.E., and Hatfull, G.F. 1999. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci.* **96:** 2192–2197.

Herre, E.A., Knowlton, N., Mueller, U.G., and Rehner, S.A. 1999. The evolution of mutualisms: Exploring the paths between conflict and cooperation. *TREE* **14:** 49–52.

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

Jeltsch, A. and Pingoud, A. 1996. Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction–modification systems. *J. Mol. Evol.* **42:** 91–96.

Karlin, S. and Cardon, L.R. 1994. Computational DNA sequence analysis. *Annu. Rev. Microbiol.* **48:** 619–654.

Karlin, S., Burge, C., and Campbell, A.M. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* **20:** 1363–1370.

Karlin, S., Mrazek, J., and Campbell, A. M. 1997. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179:** 3899–3913.

Kita, K., Tsuda, J., Kato, T., Okamoto, K., Yanese, H., and Tanaka, M. 1999. Evidence of horizontal transfer of the *Eco*O1091 restriction modification gene to *Escherichia coli* chromosomal DNA. *J. Bacteriol.* **181:** 6822–6827.

Kobayashi, I., Nobusato, A., Kobayashi-Takahashi, N., and Uchiyama, I. 1999. Shaping the genome-restriction-modification systems as mobile genetic elements. *Curr. Opin. Genet. Dev.* **9:** 649–656.

Korona, R. and Levin, B.R. 1993. Phage-mediated selection for restriction–modification. *Evolution* **47:** 565–575.

Korona, R., Korona, B., and Levin, B.R. 1993. Sensitivity of naturally occurring coliphages to type I and type II restriction and modification. *J. Gen. Microbiol.* **139:** 1283–1290.

Krüger, D.H. and Bickle, T.A. 1983. Bacteriophage survival. Multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. *Microbiol. Rev.* **47:** 345–360.

Kulakauskas, S., Lubys, A., and Ehrlich, S.D. 1995. DNA restriction–modification systems mediate plasmid maintenance. *J. Bacteriol.* **177:** 3451–3454.

Kusano, K., Naito, T., Handa, N., and Kobayashi, I. 1995. Restriction–modification systems as genomic parasites in competition for specific sequences. *Proc. Natl. Acad. Sci.* **92:** 11095–11099.

Lauster, R. 1989. Evolution of type II DNA methyltransferases: A gene duplication model. *J. Mol. Biol.* **206:** 313–321.

Leung, M.-Y., Marsh, G.M., and Speed, T.P. 1996. Over- and

under-representation of short DNA words in Herpesvirus genomes. *J. Comput. Biol.* **3:** 345–360.

Levin, B.R. 1993. The accessory genetic elements of bacteria: Existence conditions and (co)evolution. *Curr. Opin. Genet. Dev.* **3:** 849–854.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA. *Nature* **362:** 709–715.

Marians, K.J. 1992. Prokaryotic DNA replication. *Annu. Rev. Biochem.* **61:** 673–719.

Mironov, A.A., Koonin, E.V., Roytberg, M.A., and Gelfand, M.S. 1999. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.* **27:** 2981–2989.

Naito, T., Kusano, K., and Kobayashi, I. 1995. Selfish behavior of restriction–modification systems. *Science* **267:** 897–899.

O'Neill, M., Chen, A., and Murray, N.E. 1997. The restriction–modification genes of *Escherichia coli* K-12 may not be selfish: They do not resist loss and are readily replaced by alleles conferring different specificities. *Proc. Natl. Acad. Sci.* **94:** 14596–14601.

Rambach, A. and Tiollais, P. 1974. Bacteriophage λ having *Eco*RI endonucleases sites only in the nonessential sites of the genome. *Proc. Natl. Acad. Sci.* **71:** 3927–3930.

Redaschi, N. and Bickle, T.A. 1996. DNA restriction and modification systems. In Escherichia coli *and* Salmonella: *Cellular and molecular biology* (ed. H. Neidhardt et al.), pp. 773–781. ASM Press, Washington DC.

Reisenauer,., Kahng, L.S., McCollum, S., and Shapiro, L. 1999. Bacterial DNA methylation: A cell cycle regulator? *J. Bacteriol.* **181:** 5135–5139.

Roberts, R.J. and Macelis, D. 2000. REBASE—Restriction enzymes and methylases. *Nucleic Acids Res.* **28:** 306–307.

Rocha, E.P.C., Viari, A., and Danchin, A. 1998. Oligonucleotide bias in *Bacillus subtilis*: General trends and taxonomic comparisons. *Nucleic Acids Res.* **26:** 2971–2980.

Rocha, E.P.C., Danchin, A., and Viari, A. 1999. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. *Mol. Biol. Evol.* **16:** 1219–1230.

Schbath, S. 1997. An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comput. Biol.* **4:** 189–192.

Schbath, S., Prum, B., and de Turckheim, E. 1995. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.* **2:** 417–437.

Sharp, P.M. 1986. Molecular evolution of bacteriophages: Evidence of selection against the recognition sites of host restriction enzymes. *Mol. Biol. Evol.* **3:** 75–83.

Sitaraman, R. and Dybvig, K. 1997. The *hsd* loci of *Mycoplasma pulmonis*: Organization, rearrangements and expression of genes. *Mol. Microbiol.* **26:** 109–120.

Stewart, F.M. and Levin, B.R. 1984. The population biology of bacterial viruses: Why be temperate? *Theor. Pop. Biol.* **26:** 93–117.

Wilson, D.S. and Sober, E. 1994. Re-introducing group selection to the human behavioral sciences. *Behav. Brain Sci.* **17:** 585–654.

Wilson, G.G. 1991. Organization of restriction–modification systems. *Nucleic Acids Res.* **19:** 2539–2566.

Woese, C. 1998. The universal ancestor. *Proc. Natl. Acad. Sci.* **95:** 6854–6859.

Zar, J.H. 1996. *Biostatistical analysis*, 3rd ed. Prentice Hall, Upper Saddle River, NJ.