Analysis of Long Repeats in Bacterial Genomes Reveals Alternative Evolutionary Mechanisms in *Bacillus subtilis* and Other Competent Prokaryotes

Eduardo P. C. Rocha,† Antoine Danchin,† and Alain Viari**

*Atelier de BioInformatique, Université Paris VI, Paris, France; and †Unité de Régulation de l'Expression Génétique, Institut Pasteur, Paris, France

Prokaryotic genomes seem to be optimized toward compactness and have therefore been thought to lack long redundant DNA sequences. However, we identified a large number of long strict repeats in eight prokaryotic complete genomes and found that their density is negatively correlated with genome size. A detailed analysis of the long repeats present in the genome of *Bacillus subtilis* revealed a very strict constraint on the spatial distribution of repeats in this genome. We interpret this as the hallmark of selection processes leading to the addition of new genetic information. Such addition is independent of insertion sequences and relies on the nonspecific DNA uptake by the competent cell and its subsequent integration in the chromosome in a circular form through a Campbell-like mechanism. Similar patterns are found in other competent genomes of Gram-negative bacteria and Archaea, suggesting a similar evolutionary mechanism. The correlation of the spatial distribution of repeats and the absence of insertion sequences in a genome may indicate, in the framework of our model, that mechanisms aiming at their avoidance/elimination have been developed.

Introduction

Prokaryotic genomes are compact, with sizes ranging from less than 600 kb in *Mycoplasma* to more than 10 Mb in several cyanobacterial and myxobacterial species. These compact genomes have probably been maintained through selective pressure for rapid DNA replication and cell reproduction (Maniloff 1996). It was therefore expected that repetitive sequences would be kept to a minimum under natural selection for rapid growth.

In recent years, various classes of repetitive DNA have been discovered in many prokaryotes, including genes, intergenic repeats, or insertion sequences (ISs). Repeats in genes typically involve related functions, whether for complete genes, e.g., rDNA genes (Schmidt 1998), or for domains (e.g., protein domains, tRNA genes) (Ohno and Epplen 1983). Interspersed repetitive sequences are a common feature in genomes of enterobacteria and presumably reflect regulatory or structural requirements of the bacterial chromosome, although no clear-cut function has been ascribed to them (Versalovic and Lupski 1998). ISs are frequently regarded as ''selfish'' DNA and are recognized as a major evolutionary driving force (Syvanen 1998). Other repeats have important functions, such as the creation of antigenic variation by recombination of modular genes (Himmelreich et al. 1997) or by phase variation (Dybvig and Voelker 1996), acquisition of novel genetic characteristics (Mazel et al. 1998), and regulation of gene expression by control of mRNA stability (Newbury et al. 1987).

Abbreviations: CDS, coding sequence; IE, inserted element; IS, insertion sequence; Ori, origin of replication; SRS, simple repetitive sequence; Ter, terminus of replication; UFO, unknown function open reading frame.

Key words: prokaryotes, *Bacillus subtilis,* horizontal transfer, DNA repeats, integration, competence.

Address for correspondence and reprints: Eduardo P. C. Rocha, Atelier de BioInformatique, Université Paris VI, 12 Rue Cuvier, 75005 Paris, France. E-mail: erocha@abi.snv.jussieu.fr.

Mol. Biol. Evol. 16(9):1219–1230. 1999

The insertion of heterologous sequences in the genome is usually explained by imperfect deletion of integrated conjugative plasmids or phages through the action of ISs (Syvanen 1994; Lawrence and Ochman 1998). However, *Bacillus subtilis,* the most studied Gram-positive organism, is not naturally conjugative (Dubnau 1993) and has no IS (Kunst et al. 1997). *Bacillus subtilis* possesses bacteriophages, but these usually rely on transposable elements to transfer genetic information, either by transposition on the genome or by imprecise excision (Syvanen 1994). Since these major evolutionary mechanisms are absent, we undertook an analysis of long repeats in prokaryotic chromosomes using *B. subtilis* as the model organism.

We analyzed the long repeats of eight complete bacterial genomes, namely *B. subtilis* (competent Gram-positive bacteria), *Mycoplasma pneumoniae* and *Mycoplasma genitalium* (Gram-positive), *Methanoccocus jannaschii* (Archaea), *Methanobacterium thermoautotrophicum* (competent Archaea), *Escherichia coli* (Gram-negative), *Helicobacter pylori*, and *Haemophilus influenzae* (competent Gram-negative). After a general analysis of the eight genomes, we extensively analyzed the genome of *B. subtilis,* which presents the most spatially constrained and unusual distribution of repeats. Interestingly, this species is the most opposed to *E. coli*, which is the typical model for bacterial evolutionary studies. Our analysis strongly suggests a mechanism relying on the integration of circular DNA, and not requiring ISs, for the acquisition of novel genetic information present along the evolutionary history of *B. subtilis* and other competent organisms.

Materials and Methods
Data

Eight complete genomes were obtained at the following internet addresses: *B. subtilis* (Kunst et al. 1997) (www.pasteur.fr/Bio/SubtiList.html); *H. influenzae* (Fleischmann et al. 1995), *H. pylori* (Tomb et al. 1997), *M. jannaschii* (Bult et al. 1996), and *M. genitalium* (Fraser et al. 1995) (www.tigr.org/tdb/mdb/mdb.html); *E. coli* (Blattner et al. 1997) (www.genetics.wisc.edu); *M. thermoautotrophicum* (Smith et al. 1997) (web.cric.com/genesequences/); and *M. pneumoniae* (Himmelreich et al. 1996) (www.zmbh.univ-heidelberg.de/M_pneumoniae). The sequences of all completely sequenced naturally occurring *B. subtilis* plasmids were downloaded from GenBank (pTA1060 [U3280], pTA1015 [U32379], pTA1040 [U32378]) (Meijer, Venema, and Bron 1995).

Statistical Methods

Statistical significance of long repeats was calculated as in Karlin and Ost (1985), considering a model where the nucleotide bias of the genomes is taken into account. For a large sequence (of length N), we expect the length (L_2) of the largest repeat present at least twice in the genome to be distributed according to a Gaussian distribution with mean and variance given by:

$$E(L_2^{(N)}) = \frac{\log\binom{N}{2}}{-\lambda} - \left[\frac{\log\left(1 - \sum_{j=1}^{4} p_j^2\right) + \lambda}{\lambda} + \frac{0.5772}{\lambda}\right] + 0.5 \qquad (1)$$

$$\mathrm{Var}(L_2^{(N)}) = 1.645\left(\frac{1}{\lambda}\right)^2 \qquad (2)$$

$$\lambda = \log\left(\sum_{j=1}^{4} p_j^2\right) \qquad (3)$$

where p_j is the relative frequency of the nucleotide j in the sequence. We were interested in repeats of very unlikely lengths and have therefore imposed the very conservative threshold of significance of 1‰. With this threshold, we calculate the minimally significant length ($L_{\min}$). Therefore, by chance alone we expect to find at most one repeat of length $L_{\min}$ in 1,000 random genomes (that respect the composition in nucleotides of the true genome). We verified that this model is not significantly altered by the known bias in dinucleotides, trinucleotides, and up to heptanucleotides by simulating random chromosomes for which these compositions were respected.

Repeat Searching

Repeats were searched through an iterative use of the Karp-Miller-Rosenberg (KMR) algorithm (Karp, Miller, and Rosenberg 1972; Soldano, Viari, and Champesme 1995). This algorithm finds the largest subword strictly present at least twice in a sequence (e.g., the genome). If this word (a repeat) is larger than the minimal significant length (i.e., larger than $L_{\min}$), it is kept, its positions in the sequence are excluded from further analysis, and the KMR algorithm is run again. The second run provides the second largest repeat that has no overlap with the first. The process goes on, and it only stops when the largest repeat found by the algorithm is smaller than $L_{\min}$. Through this method, we obtain the largest possible repeats in the genome and avoid the redundancy produced by the fact that two subwords of a repeated word are themselves necessarily repeated. If a word is repeated more than twice (which occurs very rarely), its occurrence is spliced into pairs.

Construction of Trains

If a large repeat presents a few mutations our method will detect two (several) smaller contiguous repeats. We call such contiguous repeats a ‘‘train of repeats,’’ since they constitute an ordered set of sequences of very close repeats. The ‘‘order’’ of a train is the number of contiguous repeats it contains (e.g., a train with two repeats is a 2-train). A train of order 1 corresponds to the case of a single isolated repeat. Trains are built according to the following iterative algorithm: consider the first repeat A (in positions A_1 and A_2 of length l_A); we merge B with A in a single train if the constraint of order $A_1 < B_1$ and $A_2 < B_2$ is respected and if the following constraint of distance (quadratic mean) is also respected:

$$\sqrt{(A_1 + l_A - B_1)^2 + (A_2 + l_A - B_2)^2} < 1{,}000.$$

If both conditions are verified, A and B constitute a train, and one can check whether a third repeat, C, can be included in the train by considering the constraint of order (i.e., $B_1 < C_1$ and $B_2 < C_2$) and the constraint of distance of C to B. The process stops when one repeat fails to obey one of the constraints. Once all trains (including those of order 1) have been built, we call the piece of sequence in between the two occurrences of a train a ‘‘spacer.’’

Search for Similarity

To identify similarities between repeats, we performed sequence comparisons at two different levels. On one hand, the repeats were cross-compared in order to reveal possible families of repeats in terms of sequence similarity. On the other hand, the neighborhoods of the two occurrences of each repeat (or train) were compared to reveal whether similarity extends beyond the strict repeats. In both cases, alignments were performed by using a variant of the classical dynamic programming algorithm for global alignment, whereby one counts 0-weight for gaps at both ends of the largest sequence pattern (Erickson and Sellers 1983). This variant is used to ‘‘fit’’ a sequence into a longer one, and therefore we call it a ‘‘pattern-fit.’’

For the neighborhood analysis, the following procedure was devised. Considering the sequence enclosing the first occurrence of a repeat, a sliding window of size 30 bp was pattern-fitted into the whole sequence enclosing the second occurrence of this repeat. This yielded, for each position of the sliding window, a score associated with the best pattern-fit alignment of that window within the other sequence. The set of the scores for all successive positions resulted in a pattern-fit curve (e.g., fig. 4). This curve is maximal at the exact location

Table 1
Numbers, Sizes, and Positions of Repeats in Various Genomes

	Bs	Ec	Hi	Hp	Mg	Mj	Mp	Mt
Genome length (kb)	4,215	4,639	1,584	1,668	580	1,740	816	1,751
Minimum repeat length (bp)	25	25	24	24	24	26	23	23
Mean repeat length (bp)	62	97	66	100	61	52	58	60
Maximum repeat length (bp)	455	1,811	581	1,890	243	499	470	1,856
Number of repeats	170	397	183	204	139	260	552	280
Number of trains	54	283	75	111	82	187	250	137
Mean train length (bp)	283	156	229	271	254	132	290	256
Median spacer (kb)	15	807	34	90	91	306	232	2
Density (no./Mb)[a]	40	86	116	122	240	149	676	160
Coverage (kb/Mb)[b]	5.00	16.60	15.25	24.46	29.24	15.54	78.47	19.19
Trains density (/Mb)	12.8	61.0	47.3	66.5	141.4	107.5	306.4	91.4
cds/cds (%)	65	26	46	62	12	24	60	51
intr/intr (%)	20	56	34	19	29	69	17	37
intr/cds (%)	5	8	5	10	1	3	6	7
Overlapping (%)	10	10	14	9	58	4	17	5

NOTE.—BS = *Bacillus subtilis*; Ec = *Escherichia coli*; Hp = *Helicobacter pylori*; Mg = *Mycoplasma genitalium*; Mj = *Methanococcus jannaschii*; Mp = *Mycoplasma pneumoniae*; Hi = *Haemophilus influenzae*; Mt = *Methanobacterium thermoautotrophicum*; cds/cds = percentage of repeats with both occurrences completely inside genes; intr/intr = percentage of repeats completely outside genes; intr/cds = percentage of repeats with one occurrence inside and other outside; overlapping = percentage of repeats with occurrences overlapping at least a border of a gene.

[a] Number of repeats (trains) found per Mb of genome.

[b] Coverage is the number of kb of repeated words per Mb of genome.

of the repeats, since the highest similarity of one sequence to the other corresponds to the repeat itself (identity). The visual inspection of the curve, and particularly the way it drops at the repeat boundaries, gives precious clues about the extension of similarity along the neighborhoods of the trains of repeats (e.g., fig. 4*C*).

Global Strategy

For each genome, we computed L_{min} and used it to run the KMR algorithm on the genome. Since we know a priori that there is a large degree of resemblance between tRNA and rRNA genes, we excluded their positions from the analysis by using the annotation tables delivered with the sequences. With the list of significant repeats, we analyzed their contiguity and checked for mutual similarities. For each repeat we also examined whether similarity is exclusive to the repeat or extends on the edges, and we used information entropy approaches to eliminate simple repetitive sequences (SRSs). We used KMR to search for words repeated at least three and four times to complement the results of 2-repeats, checking if words similar (but not identical) to the repeats were present in the genome. For the genome of *B. subtilis*, similar analyses were made with a symmetrized genome, taken as the concatenation of the sequence with its inverse complement, and a leading-strand chromosome, corresponding to the sequence replicating continuously (i.e., the sequence as published up to the *ter* position and the inverse complement of the remaining sequence). This analysis revealed very few further repeats. Finally, all repeats were classified according to their positions in the chromosome and the objects that harbor them.

Results

Comparison of Genomes

Number of Repeats

Although by chance alone (see *Materials and Methods*), we would not expect to find any large repeat in any of the genomes (>23–25 nt), we found 170 such repeats in *B. subtilis*, 397 in *E. coli*, and a maximum of 552 in *M. pneumoniae* (table 1). In order to avoid spurious interpretation of the data, we had previously withdrawn the rDNA sequences—known to be often duplicated—from the analysis. There is a negative correlation between the density of repeats and the genome size that is statistically significant at the 95% level using Spearman's rank association measure. This unexpectedly indicates that smaller genomes have more repeats. This trend applies to purple bacteria but is particularly striking in the set of Gram-positive bacteria because of the contrast between *B. subtilis* and the *Mycoplasma.*

Trains of Repeats

The largest repeats range between 400 and 600 bp in most species but attain much higher values in *E. coli* and *H. pylori* and lower values in *M. genitalium* (table 1). When a large region is duplicated with few point mutations, we observe trains of repeats and not a single large strict repeat (see *Materials and Methods*). Hence, the abundance of repeats in genomes can be expressed by different measures: the number of repeats, the number of bases in repeats (fraction of the genome included in the repeats), and the number of trains (number of events). These different measures present slightly different views of the data, although *B. subtilis* is consistently classified as the genome with less repeats. Trains including more than one repeat are in the majority in *B. subtilis* (78% of the trains) but not in *E. coli* (44%) (table 1). Trains have an average length of ca. 250 bp, although in *E. coli* and *M. jannaschii,* they are significantly smaller (table 1). Trains are never simple runs of letters or small motifs.

Spatial Distribution

Figures 1 and 2 present the positions of both occurrences of each repeat in the chromosomes. These fig-

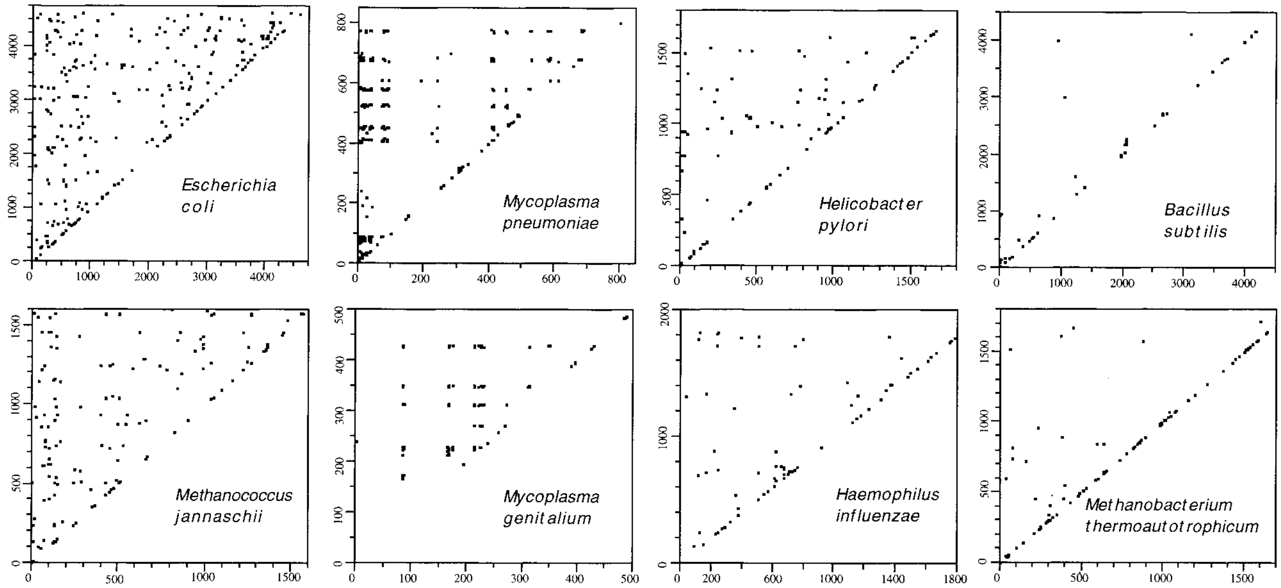

FIG. 1.—Positions (in kb) of both occurrences of repeats along the eight chromosomes. On the x-axis, the position of the first occurrence is presented, and on the y-axis, the position of the second occurrence is presented. The positions are those of the published sequences. Points lying on the diagonal represent repeats with close occurrences (short spacers).

ures reveal near random distribution of the lengths of spacers in *E. coli* and, in contrast, a very strict distribution toward small spacers in *B. subtilis* (fig. 3). The characteristics of the *B. subtilis* chromosome are not completely the opposite of those of the other genomes, since most of them harbor a certain amount of repeats in the main diagonal of the graphic. The *B. subtilis* chromosome is, however, unique in the near complete avoidance of repeats off the diagonal (fig. 1), since more than 50% of the spacers are smaller than 50 kb. Weaker trends are visible in the remaining competent species, with close repeats corresponding to more than a third of the total number of repeats. This similarity is particularly striking in *M. thermoautotrophicum,* a competent but phylogenetically very distant bacterium, in which spacers smaller than 50 kb represent 70% of the total.

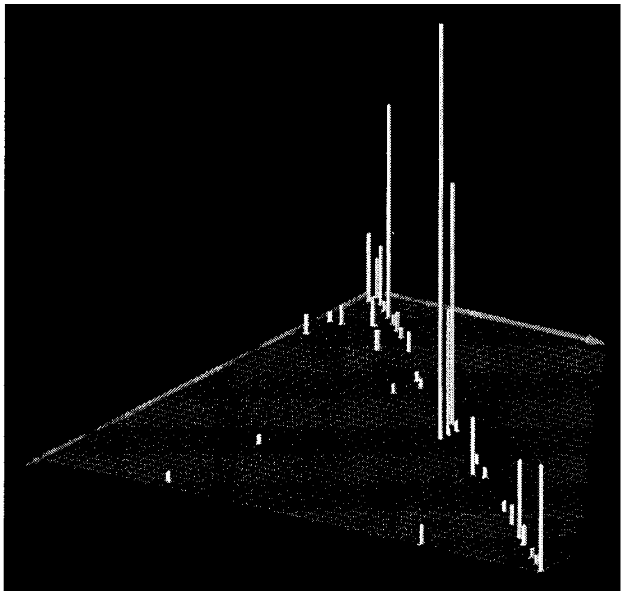

FIG. 2.—Histogram of the numbers of repeats found along the chromosome of *B. subtilis.* On the x-axis, the position of the first occurrence is presented, and on the y-axis, the position of the second occurrence is presented. The large peaks reveal either trains composed of many repeats or large densities of repeats. The bin width is 100 kb.

Antigenic Variation

In both *M. genitalium* and *M. pneumoniae,* we identified domains of repeats, i.e., small regions containing multiple clustered repeats and interrupted by long sequences where repeats are absent. These species are known for possessing a high degree of repetition in genes coding for surface proteins. Recombination between these repeats generates antigen diversity, allowing the bacteria to evade the immune system of the host (Dybvig and Voelker 1996). When one occurrence of a

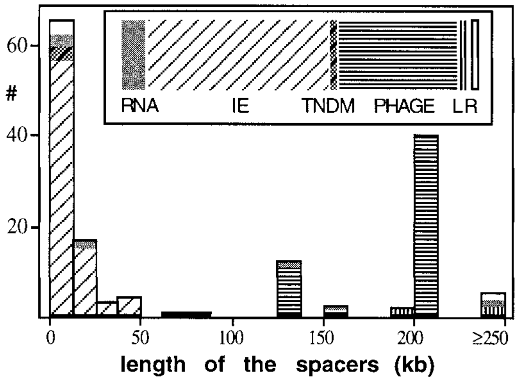

FIG. 3.—Histogram and classification of the repeats found in the sequence of *B. subtilis.* The x-axis groups repeats according to the lengths of spacers. The patterns correspond to the classification of the repeats in rDNA (RNA), intra-prophage-like elements (Phage), IEs (Inserted Elements), tandem trains (Tndm), and large (190 bp) nonstrict repeats (LR, see text), and others (in white). The small insert shows the relative abundance of each class and provides the legend for the patterns.

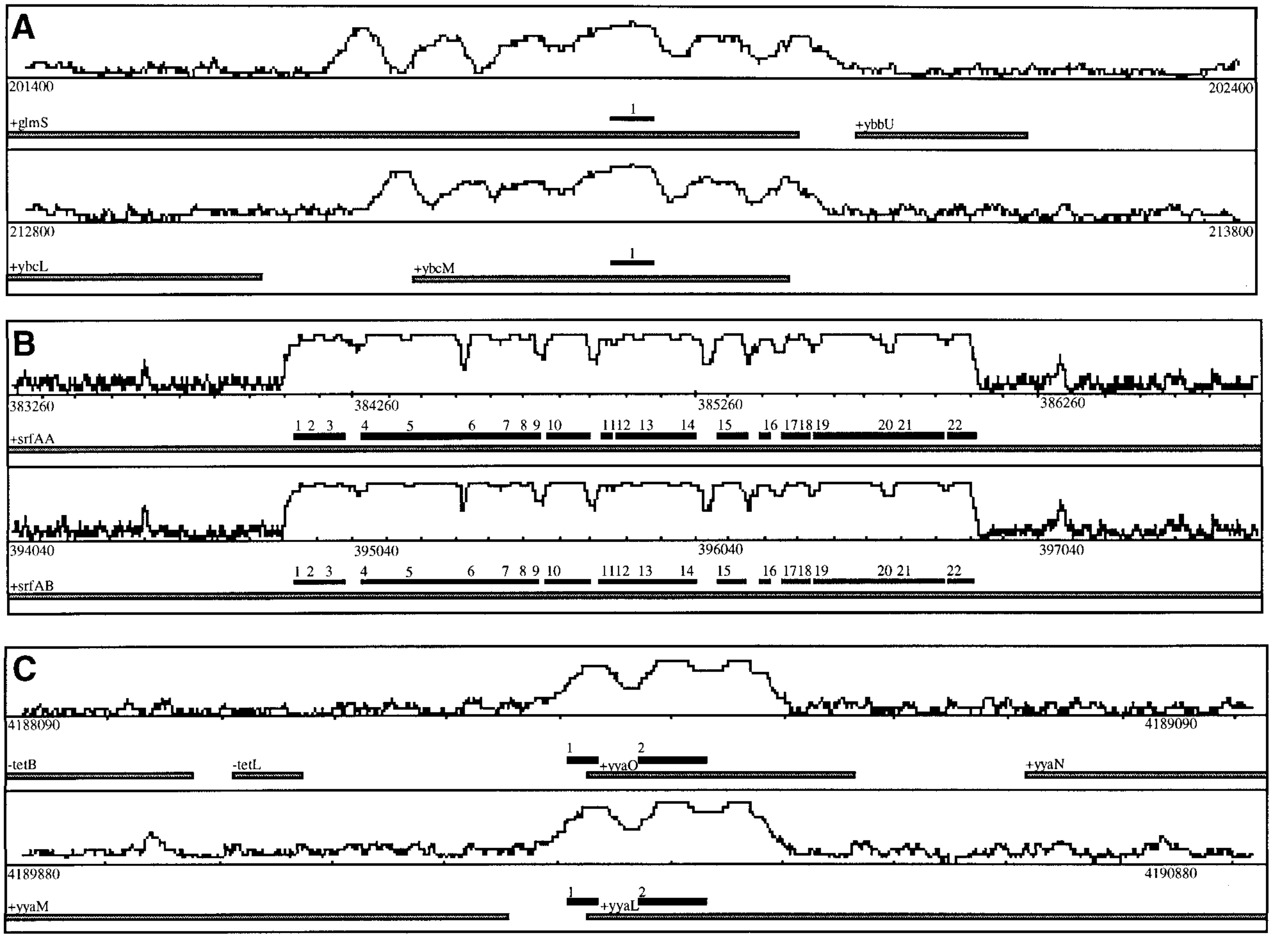

FIG. 4.—Examples of repeats flanking three IEs. Each example represents the neighborhood of the two occurrences of each train of repeats. *A,* A train resulting from an integration in a *prophage-like* element (prophage 1). *B,* A *contiguous gene* train of repeats (repeats in surfactin genes). This train is quite large (around 1.5 kb), and the sharp decline in similarity of the sequences at the edges of the train suggests that it is a recent IE. *C,* An example of the class *others.* Since similarity extends well beyond the repeats and the elements of each occurrence of the train are well apart, this probably represents a repeat that was inserted some time ago. Black boxes represent repeats and are numbered according to their relative positions in trains. Gray boxes represent genes and are labeled with the name and a plus sign if the gene is in the direct strand or a minus sign if it is in the complementary inverted strand. Rules indicate the position in the chromosome. The curves indicate the similarity at each point between the window of 30 bp centered at that position and the other sequence. The similarity is analyzed with the pattern-fit algorithm (see *Materials and Methods*).

repeat is in one of these regions, the other occurrence may be in any of the other regions. Therefore, all of these regions appear to be subject to frequent exchange of genetic material with all of the others, producing a complex net of similarities among regions and therefore dramatically increasing the possibilities of antigenic variation. For example, in *M. pneumoniae,* we found seven regions that share repeats accounting for 71% of the total number of repeats (regions between positions 4 and 44 kb, 71 and 94 kb, 406 and 428 kb, 450 and 457 kb, 523 and 529 kb, 579 and 589 kb, and 681 and 689 kb). The existence of such confined and well-defined regions is possibly a constraint imposed by the existence of operons for these surface proteins, although one can speculate about whether it is not also a strategy to avoid the transmission of recombinational instabilities to the remaining genome (unpublished data).

Environment of Repeats

Nearly two thirds of the *E. coli* repeats are in intergenic sequences, whereas 16% are in coding sequences. In contrast, 31% of *B. subtilis* repeats are in intergenic sequences, and 60% are in coding ones (table 1). This contrast still holds if the comparison is made with the repeats of *B. subtilis* and the subset of *E. coli* repeats with spacers smaller than 50 kb. Since this contrast is not dependent on spacer length, it probably reflects a difference in the type of repeats present in the two species.

General Analysis of the Repeats in the Genome of *B. subtilis*

A more extensive analysis of the repeats in the *B. subtilis* chromosome revealed that 44% of the repeats could be cataloged in well-known classes (figs. 2 and 3). These repeats are described below, and since they reveal known features, they were removed from further analysis.

The edges of rDNA genes (which themselves were withdrawn from the study, see *Materials and Methods*) are strongly homologous and therefore possess several repeats. These repeats represent 8% of the total number and provide very strong multiple alignments (data not shown). Some genes present in different prophages share extensive similarity (Kunst et al. 1997), and this accounts for one third of the total number of repeats.

Ten large (~190 bp) nonstrict repeats located around the origin of replication (Ori) of the chromosome were found in the genome of *B. subtilis* (Amano and Shishido 1995), mostly in the leading strand. Besides these repeats, we further withdrew from the analysis two repeats present within *cotB,* and five small repeats with large spacers whose roles we were not able to determine.

After the purge of these repeats whose presence was already known or expected, we obtained a very homogeneous set of 88 repeats. These 88 repeats form 17 close trains of repeats plus 5 tandem trains. The 17 trains have spacers smaller than 50 kb (and larger than 550 bp) and account for more than 50% of the total number of original repeats. Around 70% of such trains are inside coding sequences (CDSs), although most do not coincide with either of the limits of CDSs (they either exceed or cover just a piece of the CDS). Tandem trains have spacers with null or negative lengths, i.e., their occurrences either overlap or stand side by side. They are present as a 2-train of 410 bp (located at position 4102 kb) and as 1-trains of 182 bp (at 526 kb), 127 bp (at 4095 kb), 52 bp (at 2517 kb), and 25 bp (at 494 kb).

Here, we will focus our attention on the 17 trains that resulted from this purge. These trains include about half of the total number of repeats, have a mean length of 510 bp and a median length of 50 bp, and are not SRSs. For reasons that will become clear, we will call the sequence constituted by one such train and its spacer an inserted element (IE). The average size of the 17 IE is 13.7 kb, and if one excludes the *skin* element (see below) it is reduced to 10.6 kb.

Would an E. coli *Genome Devoid of IS, Transposons, BIMEs, and ERICs Resemble* B. subtilis*?*

One might question whether the differences between the spatial distribution of *B. subtilis* repeats and the remaining chromosomes would disappear if we withdrew ISs, transposons, bacterial interspersed mosaic elements (BIMEs), and enterobacterial repetitive intergenic consensuses (ERICs) from these sequences. Unfortunately, we only have sufficient information to do so in *E. coli,* and therefore the test is limited in range. In figure 5, we observe the spatial pattern of the 167 repeats that remain in the *E. coli* chromosome after removal of all intergenic/intergenic repeats (in order to withdraw intergenic repeats related to RNA secondary structure and remnants of ISs and transposons), and all repeats in genes classified as transposon- or phage-related in the annotation table (Blattner et al. 1997). The median spacer length of this set of repeats is 850 kb, which is significantly different from the value obtained for *B. subtilis* (15 kb). Although we cannot rule out that the repeats on the diagonal are created by a mechanism similar to the one we propose to be present in *B. subtilis,* it is clear that the removal of the IS- and BIME-related repeats does not make the *E. coli* chromosome similar to that of *B. subtilis.*

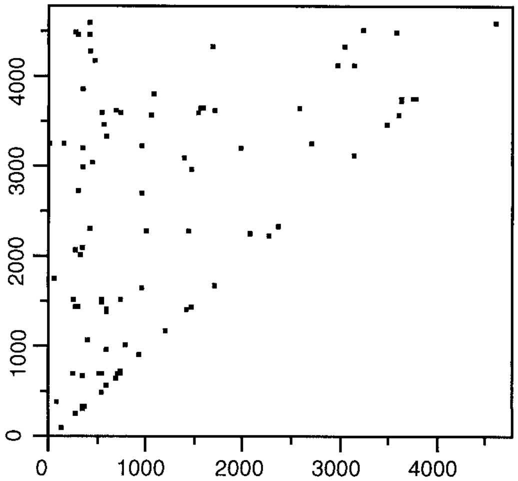

FIG. 5.—Positions (in kb) of both occurrences of repeats along the *E. coli* chromosome after removing all intergenic/intergenic repeats and all repeats in genes classified as transposons or phage-related. On the x-axis, the position of the first occurrence is presented, and on the y-axis, the position of the second occurrence is presented. This set contains 167 repeats, and the median size of the spacers is 850 kb, compared with 15 kb for the 170 *B. sutbilis* repeats.

Analysis of IEs

Classification

A train is classified as contiguous when both occurrences are in contiguous genes or when occurrences are intergenic and there is only one gene in the spacer. There are 5 (out of 17) such contiguous IEs in *B. subtilis* (table 2). Another classification can be made according to the position of the IE inside one prophage-like region of the chromosome. Prophage-like regions were classified in Kunst et al. (1997) based on their atypical codon usage, mostly derived from high A+T content and possibly represent horizontally transferred genes. There are four IE intra-prophage-like elements, all placed in different elements, in which they occupy between one third and one half of the element (table 2). The known recombinogenic *skin* element is itself an IE, with a 2-train occurring at both edges. No close train was found in the integrated bacteriophages PBSX and SPβ, although they represent 40% of the prophage-like regions. The set of IE spacers occupies 218 kb, i.e., 5.5% of the genome (table 2).

Search for Similarity

In order to test whether IEs are present in multiple copies on the chromosome, we analyzed the similarity between the spacers of noncontiguous IEs and the rest of the genome using gapped-BlastN (Altschul et al. 1997). This analysis revealed that each spacer is unique and holds no extensive resemblance to the remaining chromosome. Therefore, IEs are not the result of the duplication of a large region of the chromosome.

If the insertion of the IE is mechanistically led by the repeats independent of the spacer, then repeats could be similar even if spacers are different. In order to test

Table 2
Characterization of the 17 Inserted Elements (IEs) of *Bacillus subtilis*

Position		Length (bp)			IE[b]		Spacer Genes[c]		
First	Second	Repeat	Spacer	Category[a]	cont	Pph	Total	UFO	C3
201908	213307	35	11,399	cds/cds	—	P	11	6	9
384090	394875	1,907[d]	10,785	cds/cds	C	—	1[e]	0	0
555188	566728	27	11,540	cds/cds	—	P	9	4	8
885730	886282	103[d]	552	cds/cds	C	—	0	0	0
1385330	1424124	29	38,794	intr/intr	—	—	38	20	1
1966842	1993051	1,526[d]	26,209	cds/cds	C	—	0	0	0
1977762	1993051	2,572[d]	15,289	cds/cds	—	—	1	0	0
2050181	2060307	34	10,126	cds/intr[f]	—	P	7	3	7
2653979	2701166	171[d]	47,187	cds/intr[f]	—	*skin*	63	55	39
2725068	2735942	30	10,874	cds/intr	—	P	13	5	10
3221825	3223092	42	1,267	cds/cds	C	—	0	0	0
3461665	3469417	49[d]	7,752	cds/intr	—	—	9	5	5
3608500	3634067	35	25,567	cds/cds	—	—	20	4	1
3665044	3672062	433[d]	7,018	cds/intr[f]	—	—	2	0	2
3996569	3997959	31	1,390	intr/intr	C	—	1	1	0
4170397	4175898	1,473[d]	5,501	cds/intr[f]	—	—	5	3	5
4188597	4190389	179[d]	1,792	cds/cds	—	—	2	1	2

[a] Category of the element where the repeat lies (coding sequence [cds] or intergenic region [intr]).

[b] IE-cont is C if the repeats of the IE are contiguous, and IE-Pph is P if the IE is within a prophage-like element or *skin* if the IE is within the *skin* element (there are no IEs within the known prophages SPβ and PBSX).

[c] UFO = unknown function open reading frame; C3 = class 3 genes (horizontally transferred) in the codon usage analysis.

[d] Trains of repeats.

[e] *comS* overlaps with *srfAB*.

[f] Repeats are partially present in CDS and in intergenic regions.

this hypothesis, we performed pairwise alignments of all trains using dynamic programming. These alignments revealed that only two trains show strong resemblance. This similarity, between the trains at 1967/1993 kb and 1977/1993 kb, is inevitable, since the second occurrence of both is at the same location. Since trains are sets of repeats forming a single larger nonstrict repeat, the two occurrences of trains at position 1993 kb share the same location, although the trains at 1967 kb and 1977 kb are not strictly similar. This indicates the existence of a triplet of trains including the genes *ppsD*, *ppsC*, and *ppsA*. The multiple alignment of the trains reveals that the ones at 1977 kb and 1993 kb have diverged less than the one at 1967 kb. The proteins they encode are known to be a mosaic of motifs, but only the portions on which the trains lie are extensively similar at the DNA level.

Although all remaining trains are distinct from one another, we tested the hypothesis that some IEs might be the result of the integration of replicative plasmids. The search for similarity between the genome and the naturally occurring plasmids pTA1060, pTA1015, and pTA1040 revealed no significant similarity in IEs, particularly with the regions involved in plasmid replication.

Environment of Repeats

If trains originate from a recombination process such as the Campbell-like integration we suggest in *Discussion*, then the original repeat has probably been subjected to random drift and presents mismatches and indels. This means that the similarity between the occurrences fades away as time goes by, unless strong gene conversion acts to maintain similarity. To analyze the extent of such erosion, we aligned the regions of both occurrences of each repeat using the pattern-fit algorithm (see *Materials and Methods*). Some trains exhibit extensive similarity at the DNA level on the edges of the train itself, revealing that they were most certainly much larger (e.g., fig. 4*A*). However, some other trains are very strict (e.g., fig. 4*B*), and similarity is restricted to the train itself. The degree of erosion on the edges probably reflects the age of the seminal recombination. Since it is among the smaller trains that similarity extends more on the edges (e.g., fig. 4*C*), it is likely that the average sizes of the original repeats were significantly larger than the values given in table 2.

Locations of the Repeats

IEs can have both repeats inside CDSs (cds/cds), both in intergenic sequences (intr/intr), or one in each (cds/intr). Repeats of IE that are cds/cds are either in genes of similar size (e.g., *srfAA* and *srfAB* in fig. 4*B*) or in genes of very different sizes, where typically one of them is very small (e.g., *ybcM* and *yyaO* in fig. 4*A* and *C*). The former have extensive homology at the amino acid level along all the CDSs, but such extensive homology is much smaller at the DNA level when compared to the train. The latter are similar only near the train. The analysis of the pattern-fit alignment curves of this train suggest that the original repeat had the length of the smaller CDS.

The similarity between some CDSs in which there are two occurrences of a repeat paves the way for the proposition of a simple gene duplication mechanism at short distances. To test this hypothesis, we analyzed the

Table 3
Functional Classification of the IE Genes of *Bacillus subtilis*

Functional Classification	Observed	Expected
Adaptation	3	3
Antibiotic production	6	1
Detoxification	3	2
Regulation	12	8
Restriction/modification and repair	6	1
Cell wall	4	3
Transport/binding	17	13
Mobility and chemotaxis	6	2
Specific pathways	7	8
Others	20	43

distance between all pairs of paralogs in *B. subtilis.* A similar study of *H. influenzae* and *E. coli* revealed random spatial distribution of paralogs (Coissac, Maillier, and Netter 1997). Preliminary results of the analysis of *B. subtilis* also reveals near-random distribution of paralogs on the genome, even when the threshold required for the homology is very high (unpublished data). Therefore, pairs of paralogs are randomly positioned on the genome, and a gene duplication mechanism should require a mechanism for shuffling the duplications. However, *B. subtilis* lacks the most efficient tool to do so (ISs), and this remains an open problem.

Genes in Spacers

According to the codon usage FCA classification (Moszer, Glaser, and Danchin 1995), class 3 genes have atypical codon usage, which may indicate horizontal transfer or phage origin. In fact, half of the IE genes show class 3 codon usage (table 2) (13% on the genome and only 9% withdrawing the SPβ and PBSX phages). This strongly suggests that IEs were horizontally transferred. One should note that the larger IEs either have a large majority of class 3 genes (e.g., IEs at 201 and 2725 kb), or very few (e.g., IEs at 1385 and 3608 kb). This is compatible with the previous analysis, because although class 3 genes typically indicate foreign elements, the inverse is not necessarily true. This is so because since codon usage depends strongly on C+G content (Muto and Osawa 1987), organisms with similar C+G contents are likely to have similar codon usage biases.

We observed that 58% of the IE genes (removing the *skin* element) have unknown function and no significant homology to other genes in the public databases (table 2), which suggests that they are not housekeeping genes. The functional classification of the remaining genes indicates overrepresentation of some frequently horizontally transferred genes (Syvanen 1994), such as genes involved in competence (*srfAA, srfAB*), antibiotic production (*ppsA, ppsB, ppsC, ppsD*), flagellins (*hag*), ABC transporters (*opuBA, opuCD, opuCB, opuCA*), and restriction modification and repair (*adaB, alkA*) (table 3).

The Search for Homologs and Orthologs

The *B. subtilis* biotope is the soil and the surfaces of leaves (Lorenz and Wackernagel 1994), and integrated heterologous DNA should originate from species sharing the same habitat. However, very few of the soil bacteria have been cultivated, much less sequenced, and therefore any attempt to trace the origins of the IE genes will have to wait until more data become available. Nevertheless, we scanned the SwissProt+TrEMBL data bank in search of homologs of IE genes in other *Bacillus* species. The 16S rRNA classification of *Bacillus* (Ash et al. 1991) allows the division of the genus into five smaller groups. We spliced the 1,175 complete genes of 22 different species we found in the data bank into two groups, one including species in group 1 (the *B. subtilis* group, with 814 sequences) and the other including the remaining groups, 2–5 (with 361 sequences). Regarding BlastP hits with P values smaller than 10^{-5}, we found that within *Bacillus* species, the ones closer to *B. subtilis* have significantly larger amounts of IE genes (when compared to the average *B. subtilis* genes). This difference is statistically significant (using a χ^2, $P < 0.05$), but the sequence data on these species are sparse and possibly not very reliable.

Discussion

Possible Mechanisms for the Acquisition of New Information in *B. subtilis*

The results of the analysis of repeats in the genome of *B. subtilis* suggest the existence of recombination events that added new information to the genome. *Bacillus subtilis* is capable of gaining access to new genetic information by becoming competent, and this characteristic is reflected on its nonclonal character (Graham and Istock 1978; Maynard Smith et al. 1993). However, *B. subtilis* does not possess ISs or any kind of transposons; the transformation of monomeric plasmids without chromosomal inserts occurs at very low rates, and conjugation is almost not detectable (Lorenz and Wackernagel 1994). One might then wonder how it evolves, i.e., how it proceeds to incorporate new genetic information into the chromosome.

We suggest that IEs are traces of recent events of horizontal transfer via a Campbell-like integration (fig. 6). The integration is made through recombination between small similar regions in the foreign DNA and the chromosome, which results in an IE flanked by the two occurrences of the similar region. Our study is based on the observation of the remnants of these two regions. Next, we summarize the several steps that may mediate this horizontal transfer and the accordance of the model with the data.

DNA Uptake

In *B. subtilis*, competence is attained in the stationary phase, during which the cell risks death by starvation (Dubnau 1993). In this situation, a strategy of integration of foreign DNA may be advantageous in that it would allow the acquisition of new functions. In a competent *B. subtilis* cell, foreign DNA enters the cell single-stranded, after a nonspecific interaction with the membrane, where it is cut into fragments (Dubnau 1993). The sizes of these fragments have been investi-

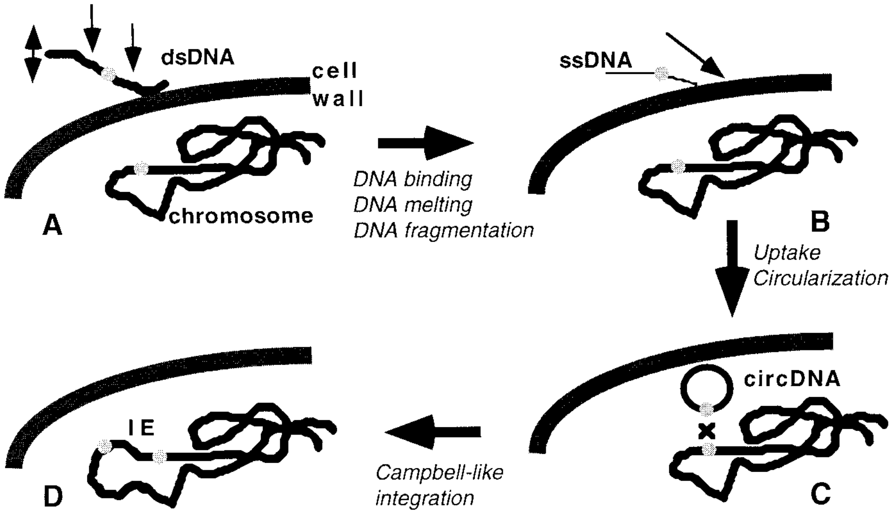

FIG. 6.—Schema of the proposed model for the insertion of exogenous DNA. The gray circle represents the repeat that gives rise to the IE through Campbell-like recombination. Thick lines represent double-stranded DNA, and thin lines represent single-stranded DNA. If the bacterium is in a state of competence, its membrane binds DNA present in the outside environment (*A*), makes it single-stranded, and cuts it into pieces (*B*). Then, the ssDNA enters the cell (*B*), the segment is circularized, and the complementary strand is formed (*C*). Finally, the circle integrates the chromosome through a Campbell-like mechanism, creating the repeat that corresponds to the homologous region (*D*).

gated by physical measurement (Dubnau and Cirigliano 1972) and by electron microscopy (Fornilli and Fox 1977) giving averages of 8.5 and 11 kb, respectively. Once in the cell, 400–500 bp on the extremities of the single strand of DNA are digested (Dubnau 1993). The observed average size of the 16 IEs (excluding the *skin* element) is 10.6 kb and therefore in agreement with the hypothesis of horizontal transfer into a competent *B. subtilis.* It is not clear whether the lengths of these fragments have a distribution such as that of the IEs we observe, but it has been shown that they can attain much higher and lower values (Zawadzki and Cohan 1995).

DNA Circularization

Heterologous single-stranded DNA entering the cell must circularize in order to integrate through a Campbell-like mechanism. DNA circularization has been extensively studied for *B. subtilis* integrative plasmids, since they are also made single-stranded and cut into pieces before entering the cell. Usually, the integration of monomeric plasmids is very difficult in *B. subtilis,* but several mechanisms are known to facilitate it. These mechanisms change the usual second-order dependency on DNA concentration for plasmid integration in *B. subtilis* into a first-order kinetics (Lorenz and Wackernagel 1994). A possible pathway of plasmid rescue is through the mechanism of facilitated plasmid transformation, which requires a region of strong similarity of the donor plasmid molecule with the chromosome (Canosi, Iglesias, and Trautner 1981). DNA synthesis and ligation convert the linear single-stranded DNA into a circular molecule, on which the second strand is synthesized, requiring RecA (Christie et al. 1987). Interestingly, the level of RecA is known to rise 14-fold upon induction of competence in *B. subtilis* (Lovett, Love, and Yasbin 1989). Although these results are for nonreplicative plasmids, one may expect them to hold for foreign chromosomal DNA as long as it possesses sufficient local similarity with the *B. subtilis* chromosome for RecA-mediated recombination.

Insertion

Once circular, the foreign DNA element is like any nonreplicative plasmid, such as the ones used for cloning genes in *B. subtilis* (Mazza and Galizzi 1989; Dubnau 1993). The length of the region of strict similarity necessary for the action of RecA recombination is optimal at values larger than 70 bp (Watt et al. 1985), but can be as small as 24 bp in *B. subtilis* (Roberts and Cohan 1993). As stated above, IEs with trains smaller than 70 bp have extensive similarity at the edges, indicating that the original identical region was larger (fig. 4). Therefore, all IE repeats have lengths compatible with our model. We initially thought that integrative agents such as replicative plasmids might be responsible for IEs, but known plasmid replication genes have no homologs in IEs. Moreover, spacers exhibit low similarities among each other and with the remaining genome, and therefore the plasmid integration hypothesis requires as many different plasmids as repeats.

Selection

The probability of identity between the IE and the genome is higher inside genes, since genes evolve more slowly and frequently present well-conserved motifs. However, if insertion takes place in the middle of a gene, then the outcome for two chimerical genes may render both genes nonfunctional if the genes are very distinct at the amino acid level. Therefore, one may expect a balance between preferential insertion of IEs in genes and preferential selection of IEs inserted at gene borders. In fact, the only three noncontiguous IEs that are well inside genes (at 554, 1977, and 3461 kb) correspond to paralogs, and the high similarity at the amino acid level between the proteins probably renders the chimeras functional. The remaining noncontiguous IEs are in intergenic regions, and genes adjacent to them do not resemble or are at the edges of genes and therefore genes are not disrupted. Contiguous IEs are mostly inside genes that resemble each other strongly, with one exception (at 3996 kb) that is intergenic. As expected, the gene inside this intergenic contiguous IE is not similar to the ones that flank the IE.

Revertants

In line with our model, the large tandem repeats can be regarded as vestiges of imprecise excision of inserted DNA. A large repeat is known to frequently induce deletion of genetic material, and imprecise deletions leave traces of the ancient repeat (Chédin et al. 1994). These tandem repeats are not SRSs, since their entropy is close to that of the genome. Both of their occurrences are always located either completely inside a CDS or in an intergenic region, which is consistent with our hypothesis of imprecise excision.

Further Analysis of the Model and Evolutionary Implications

Similarity and Recombination

Homologous recombination in *B. subtilis* and *E. coli* is restricted to closely related DNA (Cohan 1994), since its frequency decreases exponentially with sequence divergence (Roberts and Cohan 1993; Vulic et al. 1997). Therefore, foreign DNA is inserted through homologous recombination only if it is extensively similar. However, if a small fraction of the invading DNA is very similar, then a mechanism such as the one we propose becomes possible, and invading DNA coming from other organisms may possess regions of very strong similarity to the genome, particularly in slow-evolving genes, very often in horizontally transferred genes, and in protein modules. We observe a strong contrast between the similarity of the two occurrences of a train and the similarities in the remaining region. This is clearly indicated by the drop of the pattern-fit curve at the edges of the train in figure 4*B* and is understandable in light of our model, whereby integration proceeds through a local similarity region. First, the more similar region is the one used as the template for recombination. Second, the integration itself may increase the similarity between the repeats. Third, the regions of higher similarity (i.e., the repeats in the IE) can be kept similar much longer by gene conversion than the more weakly similar neighbor regions.

IS-Free Genomes

The nonspecific competence of *B. subtilis* should facilitate the invasion of IS. Since they are absent and, along the lines of our model, evolutionarily dispensable, one may speculate on the existence of a mechanism aiming at their elimination. This may be through a specific mechanism such as the ones that have been found in several eukaryotes (Goyon, Rossignol, and Faugeron 1996; Hollick, Dorweiler, and Chandler 1997) or through a general mechanism that eliminates repeats at long distances (e.g., due to the fact that they promote the deletion of extensive parts of genetic material).

Stabilization of the Genome

The organization of the genes in the chromosome is constrained by many factors. Since replication and transcription can be coupled, housekeeping genes tend to be in the leading strand (Brewer 1988). Moreover, due to the gene dosage effect, rDNA tends to lie in the third of the chromosome around the replication origin. The disruption of these advantageous strategies by large distant repeats probably involves a decrease in the fitness of the bacteria. Moreover, the existence of multiple repeats in a genome paves the way for frequent deletions of genetic material (Chédin et al. 1994), instabilities due to RecA-independent recombination (Bi and Liu 1996), or even the creation of subgenomes (Itaya and Tanaka 1997). This may explain the absence of large repeats that are far apart: first, repeats are created close to each other; second, there are no mobile elements that can displace close repeats far apart; and third, largely separated repeats are disadvantageous. In fact, *B. subtilis* presents 75% of its genes in the leading strand (comprising all 10 rDNA operons), being in this respect one of the most biased of the published genomes. In *E. coli*, only 55% of the genes are in the leading strand, although the fact that all rDNA operons lie in this strand indicates the advantage of this positioning. The expected larger stability of the genome should also implicate a larger degree of conservation of the genetic map of *B. subtilis.*

Similarities to M. thermoautotrophicum

Surprisingly, *M. thermoautotrophicum* shows repeats as closely spaced as those of *B. subtilis.* Much remains to be known about the functioning of Archaea. Nevertheless, the very strict spatial distribution of repeats exhibited by this species, together with its known natural transformability (Lorenz and Wackernagel 1994) and the apparent absence of ISs in its genome (Smith et al. 1997), suggests the existence of a similar mechanism for the acquisition of genetic information. If this hypothesis holds, then the mechanism we propose is likely to be found in many other naturally transformable bacteria.

Conclusions

It has long been suggested that the near absence of long strict repeats in prokaryotic genomes is due to selective pressures toward compactness (Maniloff 1996). In contrast, our results demonstrate that these repetitions exist in large amounts and are very probably a feature common to all prokaryotes. These long repeats pave the way for genetic recombination, and the spatial distribution of some of them seems to be related to horizontal transfer. The strict avoidance of spatially distant repeats in *B. subtilis* may indicate that large deletions of genetic material are frequent, or that unknown mechanisms act toward avoidance of them.

Several analyses of the genome of *E. coli* have suggested that a large portion (up to 18%) of its genes may be of foreign origin (Médigue et al. 1991; Lawrence and Ochman 1998). Our work indicates that horizontal transfer is also frequent in *B. subtilis* and possibly in *M. thermoautotrophicum,* although the integrative mechanism is quite different. The repeats we observe probably reveal recent recombination events; nevertheless, they indicate that 5% of the genome of *B. subtilis* has been horizontally transferred. Most studies concerning the acquisition of novel information in prokaryotes have been done with enterobacteria, using conjugation in *E. coli* (Syvanen 1994). Within this paradigm, information is acquired by homologous recombination or by the action of ISs (Davidson et al. 1975; Syvanen 1998). However, ISs and conjugation are absent in *B. subtilis,* and it seems that evolution has found alternative pathways.

In accordance with the arguments for counterselection of repeats (Maniloff 1996), one would expect that the density of repeats would be proportional to the size of the genome. However, the highest densities of repeats are found in the genomes of *Mycoplasma,* which are the smallest, and the smallest densities are found in the genomes of *B. subtilis* and *E. coli,* which are the largest. This may be explained by the negative correlation in our set of complete genomes between genome size and pathogenicity, since in this set, the pathogenic organisms have the smallest genomes. Therefore, abundance of repeats may also be an important indicator of pathogenic strategies.

Acknowledgments

We thank F. Dionísio, P. Glaser, F. Kunst, C. Médigue, and I. Moszer for comments and critiques of the manuscript. This work was partially funded by the EU grant Biotech BIO4-CT96-0655. E.P.C.R. acknowledges the support of PRAXIS XXI through grant BD/9394/96.

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHÄFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

AMANO, H., and K. SHISHIDO. 1995. *Bacillus subtilis* strains carry highly homologous direct repeat sequences on their chromosomes. Biosci. Biotechnol. Biochem. **59**:2149–2150.

ASH, C., J. A. E. FARROW, S. WALLBANKS, and M. D. COLLINS. 1991. Phylogenetic heterogeneity of the genus *Bacillus* revealed by comparative analysis of small sub-unit-ribosomal RNA sequences. Lett. Appl. Microbiol. **13**:202–206.

BI, X., and L. F. LIU. 1996. A replication model for DNA recombination between direct repeats. J. Mol. Biol. **256**:849–858.

BLATTNER, F. R., G. PLUNKETT III, C. A. BLOCH et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1461.

BREWER, B. 1988. When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. Cell **53**:679–686.

BULT, C. J., O. WHITE, G. J. OLSEN et al. (40 co-authors). 1996. Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii.* Science **273**:1058–1072.

CANOSI, U., A. IGLESIAS, and T. A. TRAUTNER. 1981. Plasmid transformation in *Bacillus subtilis*: DNA in plasmid pC194. Mol. Gen. Genet. **181**:434–440.

CHÉDIN, F., E. DERVYN, S. D. EHRLICH, and P. NOIROT. 1994. Frequency of deletion formation decreases exponentially with distance between short direct repeats. Mol. Microbiol. **12**:561–569.

CHRISTIE, P. J., R. Z. KORMAN, S. A. ZAHLER, J. C. ADSIT, and G. M. DUNNY. 1987. Two conjugation systems associated with *Streptococcus faecalis* plasmid pCF10: identification of a conjugative transposon that transfers between *S. faecalis* and *Bacillus subtilis.* J. Bacteriol. **169**:2529–2536.

COHAN, F. M. 1994. Genetic exchange and evolutionary divergence in prokaryotes. Trends Ecol. Evol. **9**:175–180.

COISSAC, E., E. MAILLIER, and P. NETTER. 1997. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol. Biol. Evol. **14**:1062–1074.

DAVIDSON, N., R. C. DEONIER, S. HU, and E. OHTSUBO. 1975. Electron microscope heteroduplex structures studies of sequence relations among plasmids of E. coli. Deoxyribonucleic acid sequence organisation of F and F-primes, and the sequences involved in Hfr formation. Pp. 56–65 *in* D. SCHLESSINGER, ed. Microbiology—1974. Ash Press, Washington.

DUBNAU, D. 1993. Genetic exchange and homologous recombination. Pp. 555–584 *in* A. L. SONENSHEIN, J. A. HOCH, and R. LOSICK, eds. *Bacillus subtilis* and other Gram-positive bacteria. Ash Press, Washington.

DUBNAU, D., and C. CIRIGLIANO. 1972. Fate of transforming deoxyrribonucleic acid after uptake by competent *Bacillus subtilis*: size and distribution of the integrated donor sequences. J. Bacteriol. **111**:488–494.

DYBVIG, K., and L. L. VOELKER. 1996. Molecular Biology of mycoplasmas. Annu. Rev. Microbiol. **50**:25–57.

ERICKSON, B. W., and P. H. SELLERS. 1983. Recognition of patterns in genetic sequences. Pp. 55–91 *in* D. SANKOFF and J. B. KRUSKAL, eds. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison Wesley, Reading, Mass.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE et al. (40 coauthors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**:496–512.

FORNILLI, S. L., and M. S. FOX. 1977. Electron microscope visualisation of the products of *Bacillus subtilis* transformation. J. Mol. Biol. **113**:181–191.

FRASER, C. M., J. D. GOCAYNE, O. WHITE et al. (29 co-authors). 1995. The minimal gene complement of *Mycoplasma genitalium.* Science **270**:397–403.

GOYON, C., J.-L. ROSSIGNOL, and G. FAUGERON. 1996. Native DNA repeats and methylation in *Ascolobus.* Nucleic Acids Res. **24**:3348–3356.

GRAHAM, J. B., and C. A. ISTOCK. 1978. Genetic exchange in *Bacillus subtilis* in soil. Mol. Gen. Genet. **166**:287–290.

HIMMELREICH, R., H. HILBERT, H. PLAGENS, E. PIRKI, B.-C. LI, and R. HERRMANN. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res. **24**:4420–4449.

HIMMELREICH, R., H. PLAGENS, H. HILBERT, B. REINER, and R. HERRMANN. 1997. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium.* Nucleic Acids Res. **25**:701–712.

HOLLICK, J. B., J. E. DORWEILER, and V. L. CHANDLER. 1997. Paramutation and related allelic interactions. Trends Genet. **13**:302–308.

ITAYA, M., and T. TANAKA. 1997. Experimental surgery to create sub-genomes of *Bacillus subtilis* 168. Proc. Natl. Acad. Sci. USA **94**:5378–5382.

KARLIN, S., and F. OST. 1985. Maximal segmental match length among random sequences from a finite alphabet. Pp. 225–243 *in* L. M. L. CAM and R. A. OLSHEN, eds. Proceedings of the Berkeley Conference in honour of Jerzy Neyman and Jack Kiefer. Vol. 1. Association for Computing Machinery, New York.

KARP, R. M., R. E. MILLER, and A. L. ROSENBERG. 1972. Rapid identification of repeated patterns in strings, trees and arrays. Pp. 125–136 *in* Proceedings 4th Annual ACM Symposium Theory of Computing. Association for Computing Machinery, New York.

KUNST, F., N. OGASAWARA, I. MOSZER et al. (151 co-authors). 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis.* Nature **390**:249–256.

LAWRENCE, J. G., and H. OCHMAN. 1998. Molecular archaeology of the *E. coli* genome. Proc. Natl. Acad. Sci. USA **95**:9413–9417.

LORENZ, M. G., and W. WACKERNAGEL. 1994. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. **58**:563–602.

LOVETT, C. M., P. E. LOVE, and R. E. YASBIN. 1989. Competence-specific induction of the *B. subtilis* RecA protein analogue: evidence for dual regulation of a recombination protein. J. Bacteriol. **171**:2318–2322.

MANILOFF, J. 1996. The minimal cell genome: ‘‘on being the right size’’. Proc. Natl. Acad. Sci. USA **93**:10004–10006.

MAYNARD SMITH, J., N. H. SMITH, M. O’ROURKE, and B. G. SPRATT. 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90**:4384–4388.

MAZEL, D., B. DYCHINCO, V. A. WEBB, and J. DAVIES. 1998. A distinctive class of integron in the *Vibrio cholerae* genome. Science **280**:605–608.

MAZZA, G., and A. GALIZZI. 1989. Revised genetics of DNA metabolism in *Bacillus subtilis.* Microbiologica **12**:157–179.

MÉDIGUE, C., T. ROUXEL, P. VIGIER, A. HENAUT, and A. DANCHIN. 1991. Evidence for horizontal gene transfer in *E. coli* speciation. J. Mol. Biol. **222**:851–856.

MEIJER, W. J., G. VENEMA, and S. BRON. 1995. Characterisation of single strand origins of cryptic rolling-circle plasmids from *Bacillus subtilis.* Nucleic Acids Res. **23**:612–619.

MOSZER, I., P. GLASER, and A. DANCHIN. 1995. Subtilist: a relational database for the *Bacillus subtilis* genome. Microbiology **141**:261–268.

MUTO, A., and S. OSAWA. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA **84**:166–169.

NEWBURY, S. F., N. H. SMITH, E. C. ROBINSON, I. D. HILES, and C. F. HIGGINS. 1987. Stabilisation of translationally active mRNA by prokaryotic REP sequences. Cell **48**:297–310.

OHNO, S., and J. T. EPPLEN. 1983. The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proc. Natl. Acad. Sci. USA **80**:3391–3395.

ROBERTS, M. S., and F. M. COHAN. 1993. The effect of DNA sequence divergence on sexual isolation in *Bacillus.* Genetics **134**:401–408.

SCHMIDT, T. 1998. Multiplicity of ribosomal RNA operons in prokaryotic genomes. Pp. 221–229 *in* F. J. D. BRUIJN, J. R. LUPSKI, and G. M. WEINSTOCK, eds. Bacterial genomes. Kluwer Academic, Boston.

SMITH, D. R., L. A. DOUCETTE-STAMM, C. DELOUGHERY et al. (37 co-authors). 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. J. Bacteriol. **179**:7135–7155.

SOLDANO, H., A. VIARI, and M. CHAMPESME. 1995. Searching for flexible repeated patterns using a non-transitive relation. Patt. Recogn. Lett. **16**:233–246.

SYVANEN, M. 1994. Horizontal gene transfer: evidence and possible consequences. Annu. Rev. Genet. **28**:237–261.

———. 1998. Insertion sequences and their evolutionary role. Pp. 213–220 *in* F. J. D. BRUIJN, J. R. LUPSKI, and G. M. WEINSTOCK, eds. Bacterial genomes. Kluwer Aademic, Boston.

TOMB, J.-F., O. WHITE, A. R. KLERLAVAGE et al. (42 co-authors). 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature **388**:539–547.

VERSALOVIC, J., and J. R. LUPSKI. 1998. Interspersed repetitive sequences in bacterial genomes. Pp. 38–48 *in* F. J. D. BRUIJN, J. R. LUPSKI, and G. M. WEINSTOCK, eds. Bacterial genomes. Kluwer Academic, Boston.

VULIC, M., F. DIONISIO, F. TADDEI, and M. RADMAN. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. USA **94**:9763–9767.

WATT, V. M., C. J. INGLES, M. S. URDEA, and W. J. RUTTER. 1985. Homology requirements for recombination in *E coli.* Proc. Natl. Acad. Sci. USA **82**:4768–4772.

ZAWADZKI, P., and F. M. COHAN. 1995. The size and continuity of DNA segments integrated in *Bacillus* transformation. Genetics **141**:1231–1243.

MANOLO GOUY, reviewing editor

Accepted May 25, 1999