Nucleic Acids Research, 1998, Vol. 26, No. 12 **2971–2980**

Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons

Eduardo P. C. Rocha[1,2,*], Alain Viari[1] and Antoine Danchin[2]

[1]Atelier de BioInformatique, Université Paris VI, 12 Rue Cuvier, 75005 Paris, France and [2]Unité de Regulation de l'Expression Génétique, Institut Pasteur, 28 Rue Dr Roux, 75724 Paris, France

Received February 13, 1998; Revised and Accepted April 24, 1998

ABSTRACT

We present a general analysis of oligonucleotide usage in the complete genome of *Bacillus subtilis*. Several datasets were built in order to assign various biological contexts to the biased use of words and to reveal local asymmetries in word usage that may be coupled with replication, the control of gene expression and the restriction/modification system. This analysis was complemented by cross-comparisons with the complete genomes of *Escherichia coli*, *Haemophilus influenzae* and *Methanococcus jannaschii*. We have observed a large number of biased oligonucleotides for words of size up to 8, throughout the datasets and species, indicating that such long strict words play an important role as biological signals. We speculate that some of them are involved in interactions with DNA and/or RNA polymerases. An extensive analysis of palindrome abundances and distributions provides the surprising result that prophage-like elements embedded in the genome exhibit a smaller avoidance of restriction sites. This may reinforce a recently proposed hypothesis of a selfish gene phenomena in the transfer of restriction/modification systems in bacteria.

INTRODUCTION

With the increasing number of full genome sequences available, new and important challenges are emerging in the field of computational biology. The existence of a long single contig representing the full genomic content of a bacterial strain allows the study of a genome as a single whole and not as a mere collection of genes. From this point of view, the question is not to analyse particular features of a protein or a family of proteins, but to consider the average or global properties of the ensemble of the genetic text of a bacteria, which includes protein genes, but also tRNA, rRNA, genetic regulatory elements, restriction sites, recombination hotspots, etc. The first important challenge of a global genome analysis is therefore the description of general rules that allow bacteria to merge together the different types of information present in the genomic text, within its physiological and environmental contexts.

To this end we have analysed several aspects of oligonucleotide usage in the complete genome of *Bacillus subtilis* strain 168 (1). *Bacillus subtilis* is, along with *Escherichia coli*, one of the best biochemically studied eubacteria, being the major model for studies related to the group of Gram-positive bacteria. Due to its ability to express and secrete heterologous proteins it is an organism of great industrial interest and due to its ability to sporulate it seems an interesting organism in which to study basic developmental processes.

The genome of *B.subtilis* is ~4.2 Mbp long, making it the third largest available contig in the databanks, with ~4100 genes, of which about a fourth has unknown or putative function (1). The genome contains 10 prophage-like elements, representing ~10% of the genome (1).

Studies regarding the linguistic properties of nucleotide sequences started long ago, just after the first long DNA sequences became available (2). Within these, methods designed to identify contrast words, i.e. words that are significantly over- or under-represented by comparison with a model, have been extensively developed (3,4). The basic rationale for these methods is that words over-/under-represented in a sequence, in contrast to a model, may indicate a phenomena of positive/negative selection. It must be emphasized that a model is always present behind the statistical procedure for the identification of the biased words. Moreover, if contrast words should be considered as good candidates for biologically relevant signals, one should keep in mind that words present in average amounts may also have important roles.

In this work we have sketched a general method to analyse in general terms word usage in complete genomic sequences. We have focused on the genome of *B.subtilis*, using other genomes only for comparative purposes. Our main intent was to identify biased use of small words, with the goal of assigning them biological interpretations. As usually happens in these studies, this resulted in some answers, but mostly in a bunch of new questions.

MATERIALS AND METHODS

Statistical methods

It has been a standard methodology to use Markov chains to model DNA sequences (5). The basic reasoning behind Markov

*To whom correspondence should be addressed at: Atelier de Bioinformatique, Université Paris VI, 12 Rue Cuvier, 75005 Paris, France. Tel: +33 1 44 27 65 36; Fax: +33 1 44 27 63 12; Email: crocha@abi.snv.jussieu.fr

chains is that one should withdraw from the number of occurrences of a word the effects that are due to the smaller words it contains. For example, it is well known that the C+G content of a genome varies from species to species and it would be misleading to assign a significance to the frequency of a dinucleotide ignoring this information. In fact, for a larger word, e.g. of size 8, one can assign different Markov orders, from 0 to 6, removing the effects of bias in mononucleotides, dinucleotides and so on up to heptanucleotides. In this approach we can immediately see that a particularly important case is that of maximal Markov order, i.e. the model for which we include the information about distribution of words of size 1 nt smaller (heptanucleotides in our previous example). In this maximal model we are maximising the information provided by the counts of all words of smaller sizes.

Let us denote by $W = (w_1w_2\ldots w_m)$ the word made by concatenation of the m nucleotides w_i and $N(W)$ its observed count in a sequence of length n. Under the Markov maximal order model the expected count $E(W)$ of W is

$$E(W) = \frac{N(w_1w_2\ldots w_{m-1})N(w_2w_3\ldots w_m)}{N(w_2w_3\ldots w_{m-1})}$$

Having obtained a theoretical expectation for the count of a word, we need a way to compare it with the real observed count in a statistically meaningful way. Several statistics have been proposed for this purpose (for a review see 7). In this work we use the z value statistic recently proposed by Schbath *et al.* (6)

$$z_w = \frac{N(W)-E(W)}{\sqrt{\mathrm{var}(W)}}$$

where var(W) represents the calculated variance of $N(W) - E(W)$. The main advantage of this z value is that it follows a reduced normal distribution for large n (6).

z values are a measure of the bias of the word, with values close to zero meaning no bias, negative values meaning under-representation and positive values meaning over-representation of the word in the genomic text. The real difficulty of the method lies in calculating the variance term var(W). For sequences large enough (i.e. large counts of each word) and the maximal Markov model the variance can be well approximated by (4)

$$\mathrm{var}(W) = E(W) \times \frac{[(N(w_2w_3\ldots w_{m-1}) - N(w_1w_2\ldots w_{m-1})][(N(w_2w_3\ldots w_{m-1}) - N(w_2w_3\ldots w_m)]}{N(w_2w_3\ldots w_{m-1})^2}$$

For each word W we count the number of occurrences of the word in the sequence and we compute its expected frequency and variance, through the use of counts of smaller words and of the previous formulae. Then we can test if the z value is significant, i.e. if it is compatible with the assumption of the Markov model. Since we know the distribution of z, it suffices to see if it is larger in absolute value than a given threshold, i.e. to a certain statistical significance. In this study, we have chosen a conservative approach and only words that have a chance of at most 1 in 1000 of being erroneously considered over- or under-represented were selected ($|z| > 3.29$). Since the previous formulae are valid only for large counts, we have only analysed words of size up to 8 nt long.

It is important to realize what this model measures and what it does not. Since it evaluates the significance of a word by taking into account the distribution of words of size $m - 1$, it measures the significance of the individual word when all bias due to words of smaller sizes is removed. If a motif is degenerate, i.e. not strictly conserved in comparison with its consensus sequence, this approach can fail to reveal it because we count exact words, whereas counts allowing for errors and deletions would be more appropriate. For very flexible signals, matrix approaches are more convenient (8). However, this kind of signal is usually associated with larger words than the ones we are considering here. Another important point lies in the definition of a 'significantly' biased word. Whatever the statistic (and model) used, the significance of the deviation between $N(W)$ and $E(W)$ varies with the counts $N(W)$, since the statistical test is able to distinguish with better accuracy a small deviation when the counts are larger (in other terms a statistical test is more powerful for large counts). This has two effects on the detection of biased words that are important to keep in mind when comparing different datasets of different sizes or words of different lengths within the same dataset. The first one is related to the total genome (or sequence dataset) size under study: for a given word size, the larger the genome the larger will be the number of detected biased words. Conversely, for a given dataset, the test will be less powerful for larger words than for smaller ones.

Construction of data sets

Having defined a statistical framework for counting exact words, we now have to devise biologically relevant datasets to test and analyse. In order to do that, we first have to identify which biological mechanisms are suspected to bias word usage in the genomic text. This question can be split into four smaller ones, regarding the following biological issues.

Replication. The start of replication is intimately connected to the cell cycle and it has been shown to require the existence of signals near the origin of replication for the attachment and control of *dnaA* activity in *B.subtilis* (9). In the replicating fork mechanism used in *B.subtilis* there are strands of DNA replicated continuously and strands replicated in discrete steps through the use of Okasaki fragments (10).

Coding. The information content of the genome, considered as the set of genes of proteins and RNAs, is usually the most regarded aspect of word usage. Three issues are particularly interesting here: the distribution of genes along the chromosome in terms of function; the peptide signals that guide proteins to their 'working' environment; the bias in usage of the code.

Control. Signals controlling gene expression are a major source of word bias. At the transcription level one is most concerned with signals controlling the synthesis of mRNA, among which are the promoters and the terminators. At the translation level, most concern is directed towards the ribosome binding sites (RBS) and the use of the start and stop codons.

Defence. The distribution of restriction sites has long since been a favourite subject in the field of DNA linguistics due to its importance in the building up of physical maps (11,12). Restriction sites are considered the most important tool for protection of the cell against invasive DNA elements such as phages. Naturally, the study of phages themselves should be included in this category.

According to these biological criteria, one has to define proper datasets allowing isolation of putative signals by using cross-comparisons. Seven different datasets were considered in this study.

Single-strand chromosome. This is simply the chromosome taken as the published single strand.

Symmetrized chromosome. This consists of concatenation of the published sequence and its inverse complement (i.e. of both strands of the chromosome). By construction, the count of a word and of its inverted complement are the same and equal the average of their counts on the single-strand chromosome. Symmetrization is necessary when word usage analysis is performed on sequences whose orientation in the chromosome is unknown. When it comes to analysis of complete genomes the orientation of the sequence is no longer a problem, but there is no *a priori* reason for one strand of DNA to be preferred over the other. One should therefore take the precaution of comparing the counts, on a single chromosome, of a word and its inverse complement or, equivalently, comparing the single strand chromosome with the symmetrized one.

Leading strand and lagging strand. The leading strand set is constituted of two sequences corresponding to the strands replicated continuously in the double-helix of DNA. In the case of *B.subtilis* it is composed of the stretch of chromosome from the origin of replication up to position 172° and of the reverse complement of the remainder. Conversely, the lagging strand is simply the inverted complement of the leading strand and therefore corresponds to the DNA replicated in discrete steps. In *B.subtilis* ~75% of the genes are present in the leading strand, which means that patterns arising from a comparison between the leading and lagging strand will also reveal bias in usage of the code and the presence of regulatory signals.

Genes, non-genes and prophages. These are subsets constituted respectively of protein genes, intergenic regions and prophages.

Finally, all these datasets can also be defined for different bacterial species. The choice of these species was dictated by available data and in order to include two different Gram-negative bacteria (as a model organism *Escherichia coli* and as a competent organism *Haemophilus influenzae*) and an archaebacterium (*Methanococcus jannaschii*).

Data sources

The complete genome sequence of *B.subtilis* was taken from the SubtiList relational database (13) at http://www.pasteur.fr/Bio/SubtiList.html , the complete sequences of *H.influenzae* (14) and *M.jannaschii* (15) were downloaded from the Microbial Database at TIGR (http://www.tigr.org/tdb/mdb/mdb.html) and the complete genome sequence of *E.coli* (16) was downloaded from University of Wisconsin-Madison (http://www.genetics.wisc.edu). Information about restriction sites was taken from the REBASE database at http://www.neb.com/rebase (17) and only restriction sites whose existence in *B.subtilis* strains has been confirmed and published in the literature were selected.

Summary of the strategy

The strategy of analysis is developed at three levels: identification of biased words in each set, selection of words biased differently in different datasets or species and analysis of the distribution of bias along the chromosome.

Word counts and biases. For each of the datasets previously mentioned, all words of size 1–8 were analysed, counts performed and expectations were computed according to maximal order Markov chains. Using this information, we selected words at a significance level below 1‰. A similar analysis was made for the complete genomes of *E.coli*, *H.influenzae* and *M.jannaschii*.

Contrast between datasets. In a tabular form we have displayed all words significantly over-represented in a set and under-represented in the other. This provides a description of words selected for in a set and counter-selected in the other. Additionally, a rank correlation analysis was performed using the *z* values for each word of a given size, through computation of the Kendall-τ association measure:

$$\tau = \frac{2\sum_{i \neq j} \tau_{ij}}{(\sqrt{n(n-1) - 2t_a}\sqrt{n(n-1) - 2t_b})},$$

$$\tau_{ij} = \begin{cases} +1, \text{if } [z_a(i) - z_a(j)][z_b(i) - z_b(j)] > 0 \\ -1, \text{if } [z_a(i) - z_a(j)][z_b(i) - z_b(j)] < 0 \\ 0 \text{ elsewhere} \end{cases}$$

where *i* and *j* are two generic words and t_a and t_b are the number of ties among pairs of elements in each of the two lists of *z* values (*a* and *b*, both of length *n*). The same analysis was performed using other traditional measures of correlation, yielding similar results (data not shown).

Distribution of bias. We have used two different kinds of graphics to represent the distribution of bias along the chromosome. The first is simply the curve, for each word, of either the counts or the *z* value calculated in a sliding window along the chromosome. The size of the window should be adapted to the length of the word being studied, so as to obtain sufficient counts. Naturally, this leads to a very confusing representation when several plots have to be superimposed. A second representation (polarogram) is more adapted to these cases. In this plot, each word is represented as one point in polar co-ordinates. This representative point is computed in the following way: The circular chromosome is first divided into *n* non-overlapping parts (20 parts for words of size 2–6 and 8 parts for words of size 7). In each division the *z* value for the word is computed and is represented by a vector pointing from the origin towards the centre of the division and whose length equals the absolute value of the *z* value. This vector is labelled positive (negative) when the *z* value is positive (negative). To analyse over-representation of the word we compute the vectorial sum of all positive vectors (conversely we sum the negative vectors for under-representation). The representative point for the word in each polarogram corresponds to the extremity of the resulting vector. This gives rise to two polarograms, one for positive and the other for negative bias.

Both analyses are deliberately general and are intended to describe general features, not to focus on particular words. Once such words are identified, other more sophisticated techniques, such as *r* scan (18), can be brought into play.

RESULTS AND DISCUSSION

Word count and bias

Mono, di- and trinucleotides. The G+C content of the genome of *B.subtilis* is 43.5%, with a heterogeneous distribution of nucleotides along the chromosome, whether one counts on the sequence as published, on the leading or lagging strand, or on the genes, which reflects constraints acting at different levels on the chromosome (Table 1).

Table 1a. Basic counts: abundance (%) of nucleotides in the different datasets

Base	Single-strand	Leading	Genes	Intergenic	RNAs	Prophages
A	28.2	29.4	29.9	31.1	25.0	30.8
C	21.8	20.1	20.3	18.9	23.5	19.2
G	21.7	23.5	24.1	18.5	31.4	18.1
T	28.3	27.0	25.7	31.5	20.1	31.9

Table 1b. Basic counts: significance of dinucleotides and relative ranks in five different datasets

	Symmetrized		Single-strand		Leading		Genes		Prophages	
	Rank	*z*	Rank	*z*	Rank	*z*	Rank	*z*	Rank	*z*
AA	1	138	1	139	2	133	3	119	1	37.2
AC	14	–126	15	–128	14	–115	14	–95.0	15	–31.9
AG	12	–44.0	13	–44.3	13	–57.0	13	–66.4	11	–5.1
AT	9	12.2	9	12.2	9	13.2	9	19.6	12	–8.0
CA	4	42.9	5	42.9	6	33.4	6	36.1	5	13.0
CC	10	–14.4	11	–14.5	11	–21.9	12	–46.8	9	7.6
CG	8	15.7	8	15.7	8	18.4	7	25.8	13	–22.1
CT	12	–44.0	12	–43.8	12	–33.1	11	–21.8	10	–1.9
GA	6	32.5	7	31.3	5	37.6	5	36.3	8	7.8
GC	3	121	3	121	3	125	2	125	3	19.9
GG	10	–14.4	10	–14.3	10	–12.9	10	–12.4	7	10.2
GT	14	–126	14	–125	15	–135	15	–138	14	–30.8
TA	16	–204	16	–204	16	–203	16	–197	16	–52.6
TC	6	32.5	6	33.6	7	23.2	8	24.2	6	10.5
TG	4	42.9	4	42.9	4	55.6	4	60.8	4	14.5
TT	1	138	2	136	1	140	1	133	2	32.7

Rank is the position of the word in the list sorted by decreasing *z* value.

Dinucleotide frequencies are the result of a complex combination of factors, among which are conformational stability, mutational hotspots, etc (3). In general, dinucleotide bias follows closely what has been described for the ensemble of prokaryotes (3), namely AA, GC and TT are over-represented and TA is the most under-represented dinucleotide, followed by GT and AC (Table 1). These orders are roughly equivalent in the single-strand and the symmetrized chromosome and, with few exceptions, also in the leading strand and the genes (Table 1).

Bias in trinucleotides is mainly coupled with usage of the code, which is not the main topic of this paper and requires a separate study. We remark, however, that the most over(under)-represented trinucleotides are CGG and GCC (TAG and CTA) for the symmetrized and the single-strand chromosome and TAT and CGG (AAT and TAG) for the leading strand.

General trends in bias of oligonucleotides. Figure 1 presents the total number of biased (either over- or under-represented) words in the single-strand chromosome of each organism as a function of word length. Although the plot is given for the single-strand chromosome, similar results are obtained whatever the dataset chosen (data not shown). The insert in the figure displays the same analysis but in relative terms, i.e. the number of biased words divided by the total number of possible words of that size.

Results indicate that the total number of biased words is much larger than would be expected by chance alone for all organisms and word sizes. It should be noted that the bias is evenly distributed across over- and under-represented words (data not shown) and this is why we pooled the two categories in Figure 1.

Moreover, the figure shows that the total number of detected biased words increases with word length. This is true for words up to 7 nt long. This results from three competing effects: on the one hand, the total number of possible words increases with word length; on the other hand, as mentioned before, the relative number of biased words a statistics can detect tends to decrease with word length. Finally, for a given dataset, longer words usually play a minor role as strict signals, therefore counting exact words also tends to underestimate the importance of larger signals. As a result, we observe in the figure a maximum for the absolute number of biased words for a word size of 7. This is consistently observed in all datasets and in all the analysed organisms. Though we observe more biased words in *E.coli* (when compared with a genome of around the same size, such as *B.subtilis*), the general trend of the curve is maintained for species with very different C+G contents, such as *E.coli* (51%) and *M.jannaschii* (31%).

Recent works regarding DNA and RNA polymerases may provide a precious clue to the biological significance of words of size 7. Doublié *et al*. (19) have found that the processivity domain of T7 bacteriophage DNA polymerase, produced by association of thioredoxin, covers a region of 7 nt in length. Additionally, it has been shown that the upstream part of the RNA polymerase advances along the DNA template in discrete steps of 7 nt long (20). Further extending these results, one might speculate that several of these words are linked with the control of DNA and RNA extension during replication and transcription.

Prophages. Known prophages integrated in the genome of *B.subtilis* are characterized by an A+T content of 63%, as against 55% for the complete chromosome. Due to the relatively low number of phage sequences (10% of the genome), it is difficult to assign statistical significance to words of size larger than 6.

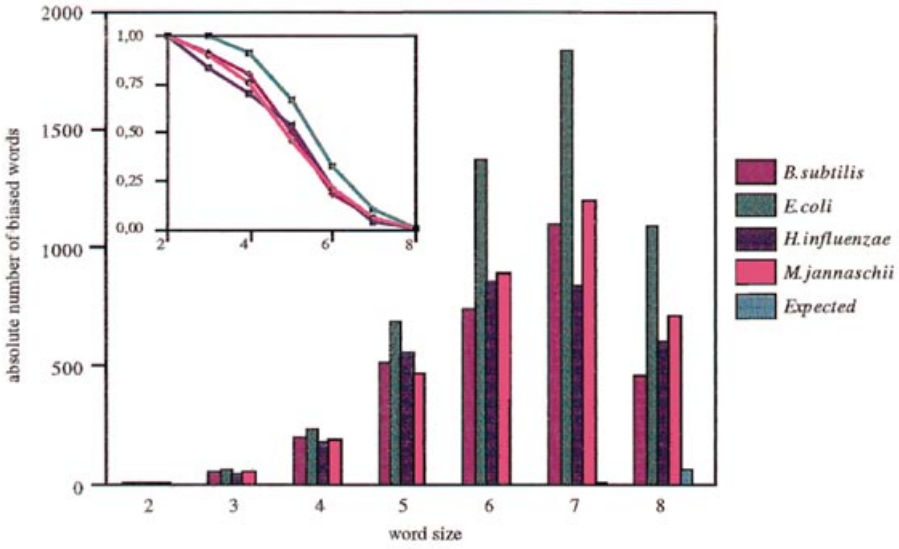

Figure 1. Total number of significantly biased words (either over- or under-represented) found in the single-strand chromosomes of the four bacterial species, for different word lengths. The insert displays the relative number of biased words, i.e. the ratio of the number of biased words to the total number of possible words of that length. Expected stands for the number of significant words expected to be found by chance at the same threshold of significance (1‰).

However, word bias closely follows the genomic pattern, with the relevant exceptions of AT and CG, which are over-represented in the genomic text and under-represented in prophages, and CC, GG and GCAGC, for which the opposite happens. Analysis of the correlation between prophages based on nucleotide frequencies gives values from 0.8 to 0.99 (data not shown), with the exception of the prophages PBSX and *skin*, which seem to differ more significantly in nucleotide distribution. Interestingly, the same analysis using dinucleotide frequencies shows consistently higher correlation coefficients, indicating that dinucleotides are more conserved than mononucleotides within prophage.

A and T runs. Table 2 displays the 10 most over- and under-represented heptanucleotides in the different datasets. A striking feature of this table is that uniform series of T and A are constantly the most under-represented words in all the sets. This is also true for words of size 6 and 8 (data not shown), which precludes series of such letters being strongly selected against in the *B.subtilis* genome. We note that nothing similar is observed for runs of G or C and that this effect is much less prominent in other species (data not shown). We observe that such runs are even under-represented in intergenic regions, which is even more remarkable considering that rho-independent terminators must include a series of T residues.

The behaviour of the runs of a letter also largely explains the pattern of over-represented words, because words such as $T_{n-1}X/XT_{n-1}$ (with $X \neq T$) or $A_{n-1}X/XA_{n-1}$ (with $X \neq A$) are negatively correlated with T_n and A_n respectively. The Markov model identifies bias in words whose frequency is not explained by random construction of its sub-words (See Materials and Methods). Therefore, once we have the frequencies of words of size 6, e.g. T_6, and considering that the biology of the system avoids the construction of words T_7, then the pool of words T_6 will have to be 'spent' elsewhere, i.e. in over-representing T_6X or XT_6 ($X \neq T$). Indeed, examination of the tables of over-represented words of size 6–8 (e.g. size 7 in Table 2) reveals the constant presence of this type of word.

Naturally one might question which is the cause and which is the effect. We believe under-representation of runs of A and T to be the cause because: (i) in absolute values the *z* values for runs of letters are much higher than for their negatively correlated counterparts in the same dataset (of the order of 100% higher; data not shown); (ii) the same runs of letters are always on top in the under-represented words (runs of T and A), whereas the most over-represented anti-correlated counterparts differ with word size, GA_5/T_5C (length 6), T_6A/TA_6 (length 7) and AT_7/A_7T (length 8). More generally, patterns of under-represented heptanucleotides are more constant throughout the datasets than the patterns of over-represented ones. This may indicate that avoidance of heptanucleotides reflects constraints acting on the ensemble of the genome whereas over-representation reveals constraints specific to each dataset.

Palindromes. Palindromes are especially interesting words in the genomic texts due to their special role as signals, particularly concerning restriction/modification systems (RM). RM systems have been considered as the most important biological tool in protecting bacteria from foreign DNA and it has been argued that avoidance of restriction sites in bacterial genomic texts would be caused by accidental deletion of restriction sites. However, this hypothesis fails to explain two important facts: (i) palindromes have also been found to be rare in the genomes of mitochondria and chloroplasts, which do not encode RM systems (11); (ii) most palindromes seem to be avoided, even those that are not recognized by the species' own RM systems. An extensive analysis of palindromes in several complete genomes and a part of the *B.subtilis* genome has recently been published (21). Our results are in line with their conclusions in establishing a relationship between the avoidance of a palindrome and its role in the RM system. The analysis shows that restriction sites appear systematically at the top of the list of most under-represented palindromes (Table 3). In the set of 16 palindromes of length 4, 14 are under-represented and none is over-represented. In the set of 64 palindromes of length 6, 25 are under-represented and only three are over-represented. No palindrome of size 8 is found to be significantly biased.

Table 2. The 10 most significantly under-represented and over-represented heptanucleotides in each dataset

Symmetrized	Single-strand	Genes	Intergenic	Leading
Under-represented				
AAAAAAA/TTTTTTT	TTTTTTT	AAAAAAA	TTTTTTT	AAAAAAA
CTTTTTA/TAAAAAG	AAAAAAA	TAAAAAG	AAAAAAA	TAAAAAG
ATTTTTC/GAAAAAT	CTTTTTA	GAAAAAT	TAAAAAG	TTTTTTT
CAAGCAA/TTGCTTG	TAAAAAG	TTGATGG	CACCTCC	GAAAAAT
CAACCGA/TCGGTTG	ATTTTTC	TCGGTTG	CTTTTTA	TTGCTTG
CGATGAA/TTCATCG	CAAGCAA	TAAATTG	GTTTTTA	TTGATGG
CAATGAA/TTCATTG	GAAAAAT	GGAAAAA	TTTAATC	CTTTTTA
GGAAAAA/TTTTTCC	CAACCGA	TTGCTTG	TACAATC	TCGGTTG
CAACAAA/TTTGTTG	CAACGAA	GTAAAAA	TTCCTTT	TAAGAAG
CTTCTTA/TAAGAAG	CAAGCGA	GCTTTTT	TTAAAAA	TTGATTG
Over-represented				
CTTTTCC/GGAAAAG	TTTTTTA	GGAAAAG	TTTTTTA	GGAAAAG
TAAAAAA/TTTTTTA	GGAAAAG	GTAAAAG	TAAAAAA	TAAAAAA
GTAAAAG/CTTTTAC	CTTTTCC	AAAAAAT	AAAAAAG	GTAAAAG
AAAAAAG/CTTTTTT	GTAAAAG	TAAAAAA	CTTTTTT	AAAAAAT
CAATGAC/GTCATTG	CAATGAC	TAAAAGA	GTTTTTT	GAAATCG
CAAGCTC/GAGCTTG	CTTTTAC	GGAATCG	AAAAAAC	CTTTTTT
CAAGCAC/GTGCTTG	TAAAAAA	ATAAATT	TTTTTTG	GTCATTG
AAATCAA/TTGATTT	GAGCTTG	AATTTGA	CAAAAAA	GTGCTTG
CATTTAC/GTAAATG	CTTTTTT	AAGAGCT	TTCCTTC	AAGAGCT
TAAGAAA/TTTCTTA	CTCCGCC	GCGGCAG	TAAAGAT	AATTTGA

Each list is sorted by decreasing absolute *z* values (i.e. the most biased words appear at the top of each list).

The analysis of the abundance of restriction sites in genes (Table 3) revealed that palindromes are generally less avoided in genes than in the single-strand chromosome (and this difference is not due to the slight difference in dataset sizes). This may be caused by restrictions due to the genetic code. However, within the set of palindromes, restriction sites maintain their relative ranks when sorted in terms of under-representation, which may indicate that selection against restriction sites in genes is of the same magnitude as for the remaining genome. It is merely the selection against palindromes in general that changes. This is consistent with the previous idea of constraints imposed by the code on the evolutionary reduction of palindromes.

Most known restriction sites are not significantly under-represented in prophages, though the small size of the prophages dataset does not allow a straightforward comparison with the single-strand. Below, in the paragraph on palindrome distribution, we will show, by another means, that they are in fact less under-represented in prophages than in the remaining chromosome, contradicting what would seem to be a *sine qua non* condition for the success of transduction.

Several results relating analysis of similarities between C-5 methylases (22) and linkage between the methylase and nuclease genes in RM systems (23) have indicated that these systems are subject to frequent horizontal transfer. In fact, of the eight different restriction sites found in *B.subtilis*, only one is present in strain 168 (YTCGAR) (23), indicating that change in the RM system is much faster than the result of evolutionary tendencies towards the avoidance of its restriction site. Bacteria cannot follow the rapidity of RM systems switch by specifically changing word usage. Therefore, the best evolutionary strategy is to avoid all possible restriction sites, i.e. avoid palindromes of length 4 and 6. It is likely that many more RM systems have occurred in the genome of the ancestors of *B.subtilis*, each of them leaving a trace of under-representation of a word. Through this evolutionary mechanism one might be able to explain the general avoidance of palindromes in most bacterial genomes.

We have scanned REBASE to search for other very under-represented palindromes present within organisms taxonomically related to *B.subtilis*. There is a general tendency in bacteria for G+C-rich restriction sites and no closely related Gram-positive bacteria was found to have AATT as a restriction site. The under-representation of CATG may be explained by the existence in *Bacillus stearothermophilus* of a restriction site RCATGY. GATC is a restriction site in five species of *Bacillus*: *B.stearothermophilus*, *B.cereus*, *B.megaterium*, *B.sphaericus* and *B.thuringiensis*. TGCA does not seem to be a restriction site for any microorganism close to *B.subtilis*. GAGCTC is a restriction site in two genera of Gram-positive (*Nocardia* and *Streptomyces*) and two Gram-negative (*Enterobacter* and *Pseudomonas*) bacteria. TGATCA is a less biased palindrome, but nevertheless significantly so, and is a restriction site in *Bacillus caldolyticus* and *Bacillus coagulans*. We propose that these restriction sites were present at a given time in ancestors of *B.subtilis*. Since they are highly compatible with its genome, they are also very likely to be present in unstudied strains. We will return to this issue in the section dedicated to the spatial distribution of words.

Table 3. Significance and ranks of all the over-represented and of the 10 most under-represented palindromes in the single-strand dataset (first column), while the two other columns (Genes and Phages) give the significance and rank of these words in the Genes and Phages datasets respectively

	Single-strand			Genes			Phages		
	Rank	Rank pal	*z* value	Rank	Rank pal	*z* value	Rank	Rank pal	*z* value
Under-represented									
AATT	1	1	–39.2	6	1	–34.4	4	2	–10.1
GGCC	4	2	–37.7	8	2	–32.7	9	3	–7.8
CATG	12	3	–27.7	13	3	–26.4	2	1	–11.5
CGCG	14	4	–26.0	16	4	–23.9	33	5	–5.2
CCGG	17	5	–24.6	21	5	–21.5	81	10	ns
GATC	32	6	–18.1	26	6	–20.0	25	4	–5.6
TGCA	42	7	–16.1	38	7	–15.7	124	15	ns
TCGA	49	8	–13.9	41	8	–15.0	133	16	ns
ACGT	52	9	–13.6	53	9	–12.6	64	7	–3.4
ATAT	63	10	–11.1	131	15	ns	72	9	ns
GGATCC	1	1	–19.6	3	1	–18.6	3	1	–4.9
CTGCAG	6	2	–11.2	15	3	–10.8	2993	54	ns
AAATTT	7	3	–11.1	9	2	–12.1	193	11	ns
ATATAT	10	4	–10.0	34	4	–8.6	102	7	ns
GAGCTC	13	5	–8.8	150	10	–5.4	1263	37	ns
ATCGAT	20	6	–7.8	61	5	–7.5	528	22	ns
CAGCTG	35	7	–6.5	108	7	–6.1	1925	44	ns
AATATT	36	8	–6.5	101	6	–6.2	630	27	ns
GGCGCC	40	9	–6.3	223	14	–4.6	35	4	ns
AAGCTT	49	10	–6.1	110	8	–6.1	1545	41	ns
Over-represented									
TAGCTA	17	1	7.7	74	1	7.1	786	7	ns
GTATAC	35	2	6.5	101	2	6.4	1416	15	ns
TCATGA	145	3	4.7	238	3	4.6	428	5	ns

Rank refers to the relative position of the word in the list of all words sorted by decreasing *z* values in over-represented and increasing *z* values in under-represented palindromes (i.e. small ranks always correspond to the more biased words). Rank pal refers to the rank in the restricted list of palindromes. ns stands for non-significant word at a *P* value of 1‰. Known restriction sites in *B.subtilis* strains are underlined.

Contrast between datasets

Analysis of correlation between datasets. The analysis of the correlation of *z* scores between the different datasets (see Materials and Methods) shows that the correlation is always positive and decreases with word size for the same dataset, reflecting the larger number of degrees of freedom of the system to accommodate the information. The most striking feature is that the order of the correlation values between the pairs of datasets remains constant whatever the word length. This order is always (symmetrized/single-strand) ≥ (leading/single-strand) ≥ (lagging/single-strand) ≥ (leading/genes) ≥ (lagging/leading) ≥ (genes/single-strand) ≥ (non-genes/single-strand) ≥ (non-genes/lagging) ≥ (non-genes/leading) ≥ (genes/lagging) ≥ (genes/non-genes). Moreover, the analysis of contingency tables (data not shown) reveals that the distribution of word counts on the single-strand and its reverse complement can be considered similar at the 1‰ level of significance for every word size. Hence, the hypothesis of global equivalence between both strands of the *B.subtilis* genome seems to be satisfied. This does not mean that taking a stretch of the genome at random one should expect to find the same distribution of words in each strand, because the distributions differ locally; it is the general character that is similar.

Patterns of replication. In order to reveal bias due to replication, the best choice *a priori* would be to compare the leading with the lagging dataset (in other words, to study in the leading strand the comparative bias of a word and of its reverse complement). However, since in *B.subtilis* most genes are on the leading strand, the analysis has to be complemented with a comparison of the leading strand with the genes.

The result of the analysis of leading versus lagging strand is given in Table 4. Interestingly, at the 1‰ significance level only words of size 4 and 5 were found to be under-represented in the leading strand and over-represented in the lagging strand (the converse situation is obtained by reverse complementing the words in the table). Many fewer words are found in the analysis of the leading strand versus genes (Table 4). It is somehow reassuring that the only 6 nt word found as over-represented in the leading strand and under-represented in genes is AGGAGG, which is the typical RBS consensus signal for *B.subtilis.*

Contrast between species. In Table 5 we display a comparison of *B.subtilis* with all other organisms. This analysis was made by choosing every word significantly over-(under-)represented in *B.subtilis* and significantly under-(over-)represented in all the remaining organisms. These words should be at least Gram-positive specific (though not necessarily species specific). It is interesting

to note that no dinucleotides are found fulfilling this condition, though GA and TC are over-represented in *B.subtilis* and under-represented in the Gram-negative species under consideration. Also, no words of size 7 and 8 are found through this analysis. In fact, words of smaller size strongly reflect codon usage, RM systems and mutational hotspots that may change significantly from species to species in directions that make them inversely biased. Signals of larger size are involved in different phenomena, such as RBS, promoter recognition and χ sites, which, though changing among species, do not have reasons to change in the opposite sense.

In addition, we looked for words simultaneously under-(over-)represented in *B.subtilis* and *H.influenzae* and over-(under-)represented in the other species, with the aim of detecting words that both species could have in common and which could be involved in their competence. No such words were found for oligonucleotides of size 6, 7 or 8. For smaller sizes, the words under-represented in the *subtilis*/*influenzae* group and over-represented in the other are: AAC, GTT (size 3), ACGC, GCGT, GTTA, TAAC (size 4) and AAAGC (size 5). The converse situation yields: AAG, CTT (size 3), ATGG, CCAT, GCAC (size 4), CAGCT, CGCGT and CTCGT (size 5).

Genomic distribution of words

General trends. Like all prokaryote genomes analysed so far, the genome of *B.subtilis* has no isochore-like structure. However, the distribution of nucleotides along the chromosome is far from being constant, as was perceived earlier (24) from a contig representing 5% of the genome. There is an asymmetry in nucleotide frequencies between the first half of the single-strand chromosome and the second half. In fact, G and A are more abundant than C and T in the first half of the chromosome, whereas the converse happens in the second half. This difference is mainly explained by the larger proportion of these bases in genes (Table 1).

Table 4. Contrast words observed in the leading strand versus the lagging strand (top) and versus genes (bottom)

Leading +/lagging –	Leading –/lagging +
CTTG (13.7/–9.5)	CAAG (–9.5/13.7)
GTCG (4.3/–4.1)	CGAC (–4.1/4.3)
GCCC (3.7/–3.7)	CGGG (–3.7/3.7)
GTAG (9.0/–4.5)	CTAC (–4.5/9.0)
GTTC (4.3/–9.2)	GAAC (–9.2/4.3)
ACCC (4.2/–9.5)	GGGT (–9.5/4.2)
GCCGG (4.6/–4.8)	CCGGC (–4.8/4.6)
TATCG (4.2/–6.8)	CGATA (–6.8/4.2)
TAAGC (4.8/–5.3)	GCTTA (–5.3/4.8)
ATTAC (4.6/–4.5)	GTAAT (–4.5/4.6)
ATGAA (3.7/–3.6)	TTCAT (–3.6/3.7)
Leading +/genes –	**Leading –/genes +**
AGG (4.2/–11.4)	
ATA (44.4/–9.6)	
ATAA (4.0/–5.5)	CAGG (–4.0/3.3)
CCTC (4.6/–4.2)	
	ATTTT (–3.7/6.0)
AGGAGG (4.0/–3.3)	

+ indicates over-representation and – indicates under-representation. The *z* values are given in parentheses.

Table 5. Contrast words between *B.subtilis* and all other species

Bacillus subtilis +/other –		*Bacillus subtilis* –/other +	
Single-strand	Genes	Single-strand	Genes
CCT	GTC	CAAC	AAC
AAATG	TCGG	CCAC	CCA
AACCG	AAATG	GTGG	TGG
AATCG	AATAG	GTTG	ATAA
AATTG	ATAAG	AAAAT	CAAC
AGGTG	ATCTG	AAGCC	CCAG
CAATT	CAATT	AATAT	CTTT
CACTC	GGCCA	AGGTT	GCTA
CATTT	TATCC	ATTTT	GGTA
CGATT	TATTT	CAATA	GTTG
GCTAC	TTATT	CAGTA	AAAAT
GTAGC	TTGCG	CCGGG	AATCC
TTGCG	AAGCAA	GCATT	ATGAC
ATTTTA	GTCAAC	GGCTT	CAGTA
TAAAAT		TACTG	GCATT
		TATTG	GGCTT
		TCCTG	TATTG

The left columns present all words significantly over-represented in the *B.subtilis* single-strand and genes dataset respectively and significantly under-represented in all other species (*E.coli*, *H.influenzae* and *M.jannaschii*) in the same sets. The right columns depict the converse situation.

It is possible from a simple graph of the distribution of A+T to devise regions that host potential prophages, since these have a much higher A+T content (1). Conversely, C+G content can be used to find tRNA and rRNA genes, which present high frequencies of these nucleotides (Table 1).

The distribution of dinucleotides presents few large patterns besides those dictated by local nucleotide frequencies (Fig. 2). This seems to indicate that dinucleotide bias is a result of properties that do not have gradients along the genome, i.e. that change solely due to the existence of biological objects, such as genes or prophages. A particularly dramatic decrease/increase is generally observed near the *Ori* and *Ter* sites (Fig. 2). Additional disruptions are also observed for some dinucleotides due to the presence of local features. An example of this is the large peak found for the dinucleotide CT at ~2.15 Mb, which is related to the presence of the SPβ prophage at that location. Similarly, the CG dinucleotide, which is globally over-represented, is actually under-represented within phages (Table 1) and therefore its plot presents a peak near the *Ter* site (where most prophages reside).

Polarograms of bias. Generally, the analysis of polarograms (see Materials and Methods) reveals that, for word lengths up to 6, there is a tendency towards larger outlier moments in the regions 90° and 270° (data not shown). This tendency decreases with increasing word size. However, it is important to note that this tendency is restricted to outliers and the distribution of *z* value for the majority of words is homogeneous across the chromosome.

An interesting phenomenon is visible in the series of T and A that has been raised before. The series of A residues is biased towards negative moments at 90–120° and the series of T towards a negative moment at 270° (see Fig. 3 for words of size 7). For heptanucleotides, all significantly under-represented words in the single-strand chromosome (Table 2) stand out as outliers in the negative polarogram (Fig. 3), but few of the over-represented words (Table 2) appear as outliers in the positive polarogram (data

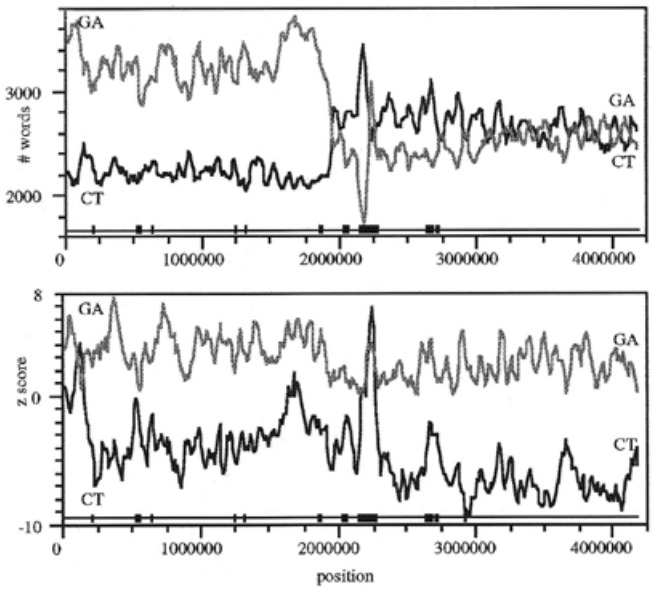

Figure 2. Distribution of the number (top) and significance (bottom) of the dinucleotides GA and CT along the chromosome of *B.subtilis* (the curve was computed using a sliding window of 50 kb and a step of 10 kb). The locations of known prophages of *B.subtilis* are indicated by black boxes above the *x*-axis.

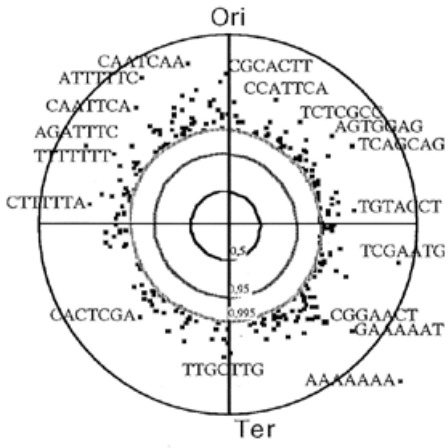

Figure 3. Polarogram of negative bias for heptanucleotides (see Materials and Methods). Inside circles are bivariate normal ellipses of *P* value 0.5, 0.95 and 0.995.

not shown). This reinforces our previous conclusion that avoidance of heptanucleotides is a localized phenomenon, whereas over-abundance tends to be more global.

Distribution of palindromes. We return here to the palindromes, to analyse their spatial distribution. Figure 4 displays the distribution along the chromosome of known restriction sites, together with some of the other more under-represented palindromes in *B.subtilis* for which a noticeable deviation from a flat graph was found. It is apparent that restriction sites are less under-represented near the terminus of replication, which is particularly surprising considering that this is the zone hosting most prophages. Considering that prophages in *B.subtilis* have necessarily achieved transduction, we have the paradox that successful bacteriophages are characteristically careless about the RM system of the host.

This analysis is further reinforced by the analysis of the polarogram of under-representation (Fig. 4). The comparison of the spatial distribution of known restriction sites and other avoided palindromes reveals similar patterns of dispersion and consistently reveals regions of less important avoidance. Particularly, GATC, GAGCTC, TGCA and AATT show polarities similar to the known restriction sites. This reinforces our previous statement that these words may be actual restriction sites, either in ancestors of *B.subtilis* and/or in some other strains of this species.

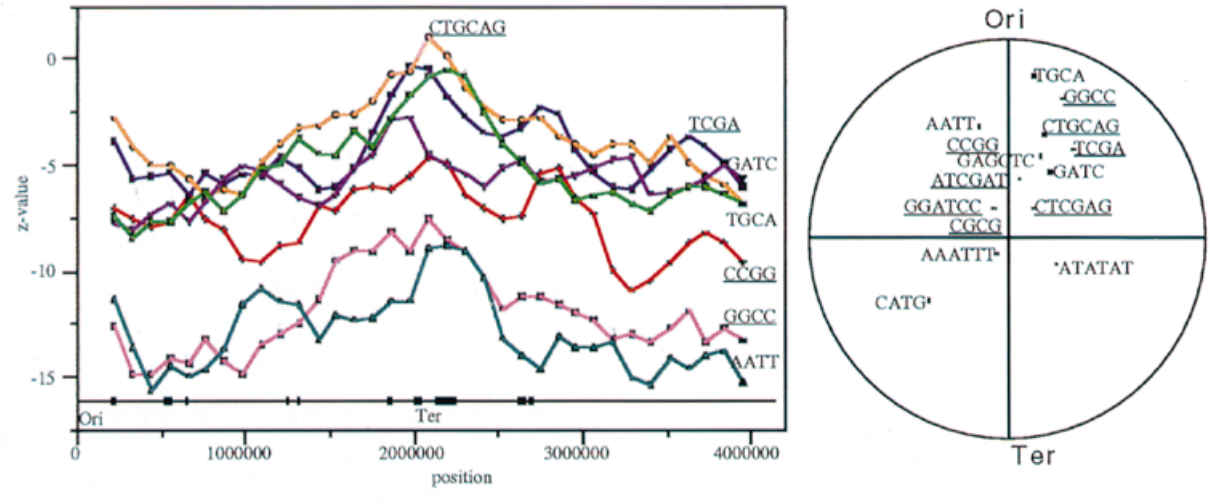

Figure 4. Distribution of some known (underlined) and potential restriction sites along the chromosome. The *y*-axis gives the *z* values computed in sliding windows of 450 kb and steps of 100 kb (left). The locations of known prophages of *B.subtilis* are indicated by black boxes above the *x*-axis. The right part of the figure displays the polarogram of negative bias for all known (underlined) and potential restriction sites.

CONCLUSIONS AND PERSPECTIVES

Many works have been published in the last decade about word usage in different species, mostly using dispersed data provided by different methods, in different species and strains. With the increasing growth of available complete genomes, different analyses are needed, since different questions become pertinent. A genome contains all the genomic information necessary to the life of the cell, therefore an analysis should envisage recognizing patterns related to the major phenomena in cells: replication, functioning, control and defence. Our consideration of different datasets for the delimitation of the phenomena was a first attempt to attain this objective. The cross-comparison of these results with other model organisms for other taxonomic groups may provide substantial insights into the determination of specific signals. Also, analysis of the spatial distribution of words and its bias will be useful in functional analysis of the genome following its complete sequencing. The analysis of word bias through the use of polarograms also provides, in our opinion, an interesting and concise way of looking at this type of data.

The availability of complete genomes allows detection of biased usage of longer words and therefore access to longer biological signals. Here we show that, for the genomes under study, the total number of observed biased words is largest at size 7. We believe that assignment of signals of that size will be found in the study of the interaction of DNA with either DNA or RNA polymerases, though the mechanics and meanings of these types of signal remain to be experimentally observed. Further studies may also explain the heterogeneous distribution of words of this size among different genomic objects, as well as the different character of under- and over-representation of these words.

We have indicated that palindromes are less avoided in prophages, contradicting the standard paradigm of RM systems as defence tools. Moreover, of all the restriction sites of size 4 present in *B.subtilis* strains, it is that present in strain 168 that is least avoided in the strain itself. We advocate that generic palindrome avoidance is the result of the avoidance of restriction sites in the genomes because of the ease with which RM systems are horizontally transferred. This implies that the best strategy is to be protected against all kinds of *possible* restriction sites. Kusano *et al*. (25) have suggested that RM systems are in fact selfish systems, whose role in cell defence is much less important than previously thought. According to this hypothesis one might speculate that avoidance of restriction sites is not only (or not at all) a strategy to prevent accidental failure of methylation, but also a defence against invasion of the bacterium by a new restriction site, hereby seen as a parasite. Moreover, several different works have shown that the influence of the restriction system in transformation of competent bacteria is quite small (26), severely weakening the standard paradigm of RM as the main defence tool of bacteria. This work further indicates that the issue of RM systems is far from being completely understood.

ACKNOWLEDGEMENT

E.R. acknowledges the support of PRAXIS XXI, through grant BD/9394/96.

REFERENCES

1 Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessières,P., Bolotin,A., Borchert,S. *et al.* (1997) *Nature*, **390**, 249–256.
2 Brendel,V., Beckman,J.S. and Trifonov,E.N. (1986) *J. Biomol. Struct. Dyn.*, **4**, 11–21.
3 Burge,C., Campbell,A.M. and Karlin,S. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 1358–1362.
4 Schbath,S. (1997) *J. Comput. Biol.*, **4**, 189–192.
5 Karlin,S. and Cardon,L.R. (1994) *Annu. Rev. Microbiol.*, **48**, 619–654.
6 Schbath,S., Prum,B. and Turckheim,E. (1995) *J. Comput. Biol.*, **2**, 417–437.
7 Leung,M.-Y., Marsh,G.M. and Speed,T.P. (1996) *J. Comput. Biol*, **3**, 345–360.
8 Gribskov,M., Lüthy,R. and Eisenberg,D. (1990) *Methods Enzymol.*, **183**, 146–159.
9 Ogasawara,N. and Yoshikawa,H. (1992) *Mol. Microbiol.*, **6**, 629–634.
10 Yoshikawa,H. and Wake,R.G. (1993) In Sonenshein,A.L., Hoch,J.A. and Losick,R. (eds), *Bacillus subtilis and Other Gram-positive Bacteria*. American Society for Microbiology, Washington, DC, pp. 507–528.
11 Karlin,S., Burge,C. and Campbell,A.M. (1992) *Nucleic Acids Res.*, **20**, 1363–1370.
12 Churchill,G.A., Daniels,D.L. and Waterman,M.S. (1990) *Nucleic Acids Res.*, **18**, 589–597.
13 Moszer,I., Glaser,P. and Danchin,A. (1995) *Microbiology*, **141**, 261–268.
14 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Rerlavage,A.R., Bult,C.J., Tomb,I.-F., Dougherty,R.A., Merrick,I.M. *et al.* (1995) *Science*, **269**, 496–512.
15 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G. *et al.* (1996) *Science*, **273**, 1058–1072.
16 Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V. *et al.* (1997) *Science*, **277**, 1453–1461.
17 Roberts,R.J. and Macelis,D. (1997) *Nucleic Acids Res.*, **25**, 248–262.
18 Karlin,S. and Brendel,V. (1992) *Science*, **257**, 39–49.
19 Doublié,S., Tabor,S., Long,A.M., Richardson,C.C. and Ellenberger,T. (1998) *Nature*, **391**, 251–258.
20 Rice,G.A., Chamberlin,M.J. and Kane,C.M. (1993) *Nucleic Acids Res.*, **21**, 113–118.
21 Gelfand,M.S. and Koonin,E. (1997) *Nucleic Acids Res.*, **25**, 2430–2439.
22 Lauster,R., Trautner,T.A. and Noyer-Weidner,M. (1989) *J. Mol. Biol.*, **206**, 305–312.
23 Trautner,T.A. and Noyer-Weidner,M. (1993) In Sonenshein,A.L., Hoch,J.A. and Losick,R. (eds), *Bacillus subtilis and Other Gram-positive Bacteria.* American Society for Microbiology, Washington, DC, pp. 539–552.
24 Lobry,J.R. (1996) *Mol. Biol. Evol.*, **13**, 660–665.
25 Kusano,K., Naito,T., Handa,N. and Kobayashi,I. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 11095–11099.
26 Lorenz,M.G. and Wackernagel,W. (1994) *Microbiol. Rev.*, **58**, 563–602.