Published online July 16, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 13 ***3781–3791***
doi:10.1093/nar/gkh699

Where does bacterial replication start? Rules for predicting the *oriC* region

Paweł Mackiewicz, Jolanta Zakrzewska-Czerwińska[1,*], Anna Zawilak[1], Mirosław R. Dudek[2] and Stanisław Cebrat

Department of Genomics, Institute of Genetics and Microbiology, University of Wrocław, Przybyszewskiego 63/77, 51-148 Wrocław, Poland, [1]Ludwik Hirszfeld Institute of Immunology and Experimental Therapy, Weigla 12, 53-114 Wrocław, Poland and [2]Institute of Physics, University of Zielona Góra, Szafrana 4A, 65-516 Zielona Góra, Poland

Received May 5, 2004; Revised June 11, 2004; Accepted June 22, 2004

ABSTRACT

Three methods, based on DNA asymmetry, the distribution of DnaA boxes and *dnaA* gene location, were applied to identify the putative replication origins in 120 chromosomes. The chromosomes were classified according to the agreement of these methods and the applicability of these methods was evaluated. DNA asymmetry is the most universal method of putative *oriC* identification in bacterial chromosomes, but it should be applied together with other methods to achieve better prediction. The three methods identify the same region as a putative origin in all Bacilli and Clostridia, many Actinobacteria and γ Proteobacteria. The organization of clusters of DnaA boxes was analysed in detail. For 76 chromosomes, a DNA fragment containing multiple DnaA boxes was identified as a putative origin region. Most bacterial chromosomes exhibit an overrepresentation of DnaA boxes; many of them contain at least two clusters of DnaA boxes in the vicinity of the *oriC* region. The additional clusters of DnaA boxes are probably involved in controlling replication initiation. Surprisingly, the characteristic features of the initiation of replication, i.e. a cluster of DnaA boxes, a *dnaA* gene and a switch in asymmetry, were not found in some of the analysed chromosomes, particularly those of obligatory intracellular parasites or endosymbionts. This is presumably connected with many mechanisms disturbing DNA asymmetry, translocation or disappearance of the *dnaA* gene and decay of the *Escherichia coli* perfect DnaA box pattern.

INTRODUCTION

As replication is controlled at the initiation stage, the events that occur at the replication origin play a central role in the cell cycle. The initiation of DNA replication is a complex process involving several regulated steps: (i) binding the initiator protein(s) to sites located within the origin region (*ori*); (ii) local unwinding of the *ori* region; and (iii) loading the DNA helicase and other proteins required to form replication forks (1,2). Chromosomal replication starts from one (Bacteria and Archaea, e.g. *Pyroccocus abyssi*) (3), two (Archaea, e.g. *Sulfolobus solfataricus*) (4) or multiple replication origins (Eukaryota) [for review see (5)]. In bacteria, chromosomal replication initiated at the *oriC* region proceeds bi-directionally until the replication forks reach the termination site, *terC* (in the case of circular chromosomes) or chromosomal ends (in the case of linear chromosomes) (6). The initiation of bacterial chromosome replication is mediated by the initiator DnaA protein, which interacts with repetitive non-palindromic nonamer sequences, the DnaA boxes, located within the *oriC* region. Among bacteria, the initiation of replication is best understood in *Escherichia coli* (7). The binding of 10 to 20 DnaA protein molecules to five DnaA boxes promotes unwinding within the AT-rich region of *oriC*.

The sequences of bacterial *oriC* regions are conserved only among closely related organisms. Bacterial replication origins vary in sizes (from ∼200 up to 1000 bp), but (nearly) all contain several DnaA boxes and an AT-rich region. A cluster of four or more DnaA boxes is an indication of a functional chromosomal origin. However, only one and two DnaA boxes, respectively, were identified within the functional *oriC* regions of *Caulobacter crescentus* (8) and *Coxiella burnetti* (9). The *oriC* region is always located within the intergenic region, frequently within the *rnpA–rmpH–dnaA–dnaN–recF–gyrB–rnpA* genes cluster, usually next to the *dnaA* gene.

During the last few years, the genomes of approximately 150 bacteria have been completely sequenced, and there are approximately 500 ongoing bacterial genome projects. Thus, within a decade, the genomes of nearly every significant bacterial species will have been sequenced. Since the events that occur at the replication origin are central to the process regulating DNA replication and the cell cycle, it is important to define the rules required for a proper *oriC* localization within analysed genomes.

Initially, the *oriC* regions were identified by direct measurement of the replication time of various genetic markers, by construction of minichromosomes (plasmids containing *oriC* instead of their own replication origin) or by two-dimensional gel electrophoresis. So far, however, only 11 *oriC* regions have been shown *in vivo* to be functional. Now that long stretches or the complete sequences of bacterial chromosomes have

*To whom correspondence should be addressed. Tel: +48 71 337 11 72; Fax: +48 71 337 13 82; Email: zakrzew@iitd.pan.wroc.pl

become available it is possible to identify a potential *oriC* region *in silico*. Lobry (10,11) was the first to find bias (asymmetry) in nucleotide composition in some bacterial chromosomes. He noticed that the asymmetry changes its polarity at the origin and the terminus of chromosome replication, where DNA strands change their character from leading to lagging and *vice versa*. This asymmetry is usually measured as the normalized difference in the content of complementary nucleotides [e.g. GC skew = (G − C)/(G + C)]. This observation was then confirmed by other researchers for many bacterial chromosomes [e.g. (12–18]. There are many reasons for this asymmetry, but the main one seems to be an asymmetric mutational pressure (different nucleotide substitution patterns) associated with the replication of the leading and lagging strands [for review see (19–22)]. Since DNA strands change their mode of synthesis from lagging to leading in the *oriC* region, it is possible to identify the region *in silico* by analysis of chromosome asymmetry, using, for instance, the computer program Oriloc created by Frank and Lobry (23). This method is commonly used to identify the putative *oriC* region within chromosomes, particularly before experimental analyses (24–26). However, the switch in asymmetry does not always exactly correspond to the position of a functional replication origin; e.g. the *in silico* predicted origin in *Helicobacter pylori* 26695 was located 2500 bp away form the functional one (26). Therefore, in order to locate the replication origin precisely, GC skew analysis should be accomplished by additional analyses to locate the elements involved in the initiation of chromosomal replication: cluster(s) of DnaA boxes and the *dnaA* gene. Searching for a cluster of DnaA boxes requires first of all a proper definition of the DnaA box. The perfect sequence for the DnaA box of a model organism, *E.coli*, is 5′-TTATC-CACA-3′. However, the preferred DnaA box sequence for other bacteria differs slightly from that of *E.coli*; it depends on the G + C content of the analysed chromosome and/or binding specificity of a given DnaA protein (for details see Results and Discussion).

The aim of this study is to evaluate the universality of three different methods of *oriC* identification (DNA asymmetry, DnaA box distribution and *dnaA* gene location) in their application to the analysis of completely sequenced bacterial chromosomes. We have tried to group the analysed genomes according to the concordance of these methods and to analyse the problem from the phylogenetic point of view.

MATERIALS AND METHODS

The analyses were performed on 120 complete sequences of bacterial chromosomes representing different taxonomic groups (see Supplementary Table 1). The chromosomal sequences were downloaded from the Genbank (ftp://www.ncbi.nlm.nih.gov). The assumed location of the *dnaA* gene on the chromosome was that of the database annotations. DNA asymmetry was analysed by DNA walks describing the asymmetry between differently replicating DNA strands [(16,17); see also: http://smorfland.microb.uni.wroc.pl]. This method differs from that originally proposed (10,11,13), as it cumulates local deviations of a parameter of asymmetry ([G–C] or [A–T]) from the average value specific to the whole chromosome. This method eliminates the global compositional trend of the whole chromosome and smoothes random fluctuations. By this method the single origin and terminus of replication (if they are present) are easily recognized as the two main extrema of the DNA walk, i.e. where asymmetry changes its sign (Figure 1; Supplementary Figure 6).

In order to identify the potential cluster(s) of DnaA boxes we looked for all possible 28 nonamers which differed by no more than one position from the perfect DnaA box: TTATC-CACA. The density of the box distribution on the chromosome was presented on charts by $b = 1/d$, calculated for each box, where d is a sum of the distances of the given box to its two adjacent boxes. The b value corresponds to the density of boxes in the given region of the chromosome, with a high value of b indicating a cluster of three boxes.

To visualize the correlation of the three methods of *oriC* identification (*dnaA* gene location, DNA asymmetry and DnaA box distribution), in the charts presenting these analyses (Figures 1–3; Supplementary Figure 6), the x-coordinate begins and ends in the presumed (according to DNA asymmetry) terminus of replication. Then the presumed *oriC* region is usually located near the middle of the chart. This was done for all circular chromosomes showing two distinct global extrema of DNA asymmetry. The charts for linear chromosomes and chromosomes without distinct global extrema of DNA asymmetry have the original sequence coordinates on the x-axis. In the charts, the DNA walk is shown only for the one parameter of asymmetry ([G–C] or [A–T]) which exhibited the strongest and most distinct DNA asymmetry.

Supplementary material

Table 1 (general characteristics of the analysed chromosomes); Table 2 (locations of boxes and their patterns in the presumed *oriC* region); and Figure 6 (DNA asymmetry, DnaA box distribution and *dnaA* gene location on 120 bacterial chromosomes).

RESULTS AND DISCUSSION

General remarks

In order to elucidate the rules for predicting the origin of bacterial chromosome replication, we performed a comprehensive *in silico* analysis of 112 complete genome sequences of bacteria representing different taxonomic groups (Supplementary Table 1). We analysed the DNA asymmetry and the presence of elements involved in the initiation of replication (DnaA boxes and *dnaA* gene). Most (101) of the analysed bacteria possess a single circular chromosome, while three have a linear chromosome. Eight organisms contain two chromosomes (either with both circular or with one circular and the other linear).

A distinct DNA asymmetry between the leading and lagging strands is present in 93% of the analysed chromosomes (Supplementary Table 1). Such asymmetry is not present in the chromosomes of *Aquifex aeolicus*, *Bifidobacterium longum*, *Deinococcus radiodurans* (chromosome II), *Gloeobacter violaceus*, *Nostoc* sp., *Phytoplasma asteris*, *Streptomyces avermitilis*, *Synechocystis* sp. and *Wolbachia pipientis*. In most of the chromosomes stronger asymmetry was observed for the parameter [G–C] than for [A–T].

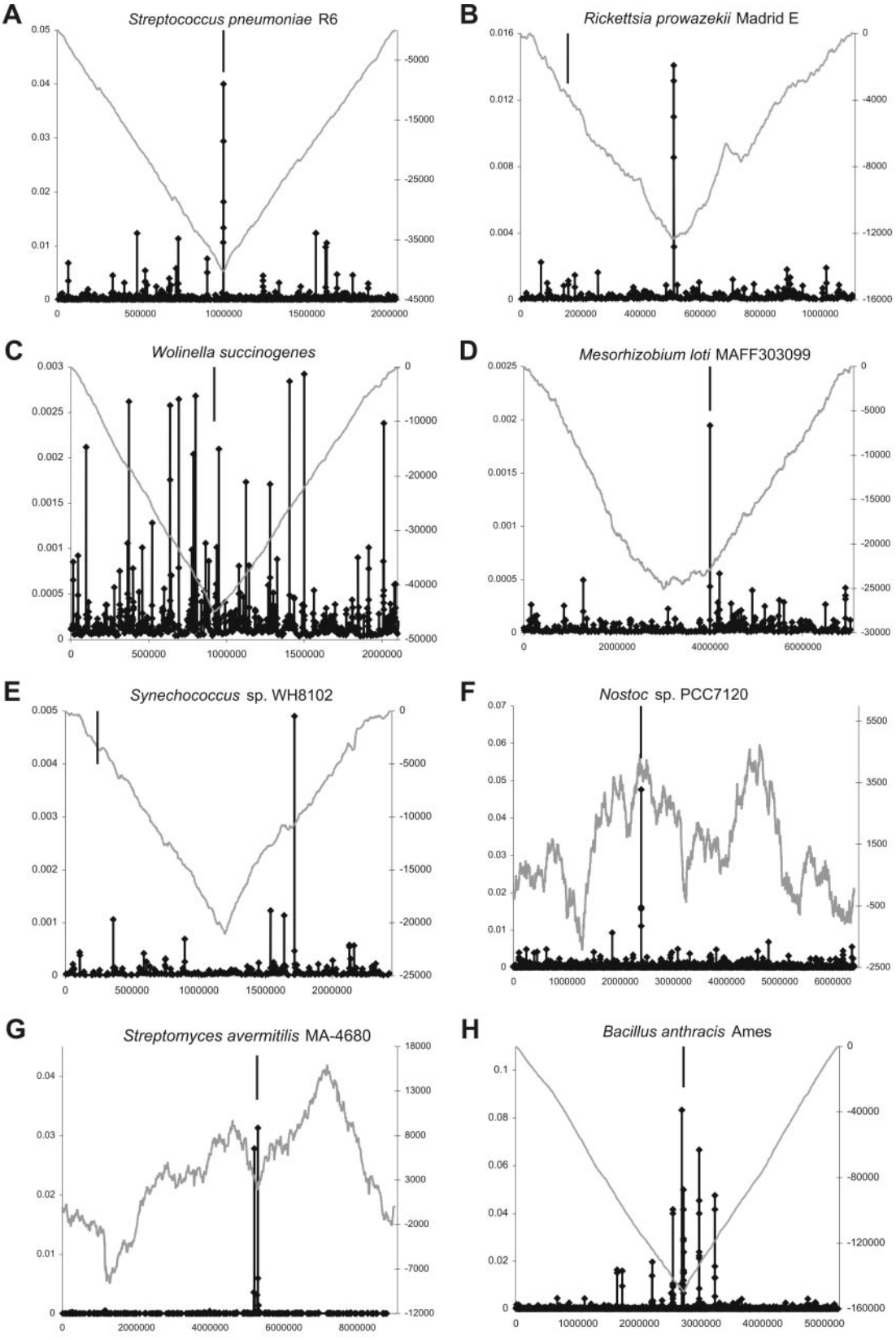

Figure 1. DNA asymmetry, DnaA box distribution and *dnaA* gene location on bacterial chromosomes. *x*-axis—chromosome coordinates in base pairs, in (**A**–**E** and **H**) beginning and ending in the presumed terminus of replication (according to the DNA asymmetry). The chromosomal coordinates in (**F** and **G**) are as in the databases. The grey line and the right *y*-axis show the DNA walk describing DNA asymmetry (asymmetry for [A–T] in (F); asymmetry for [G–C] for the other chromosomes). Black peaks with diamonds and the left *y*-axis show the DnaA box distribution expressed by the *b* value. The short vertical line indicates the *dnaA* gene location.

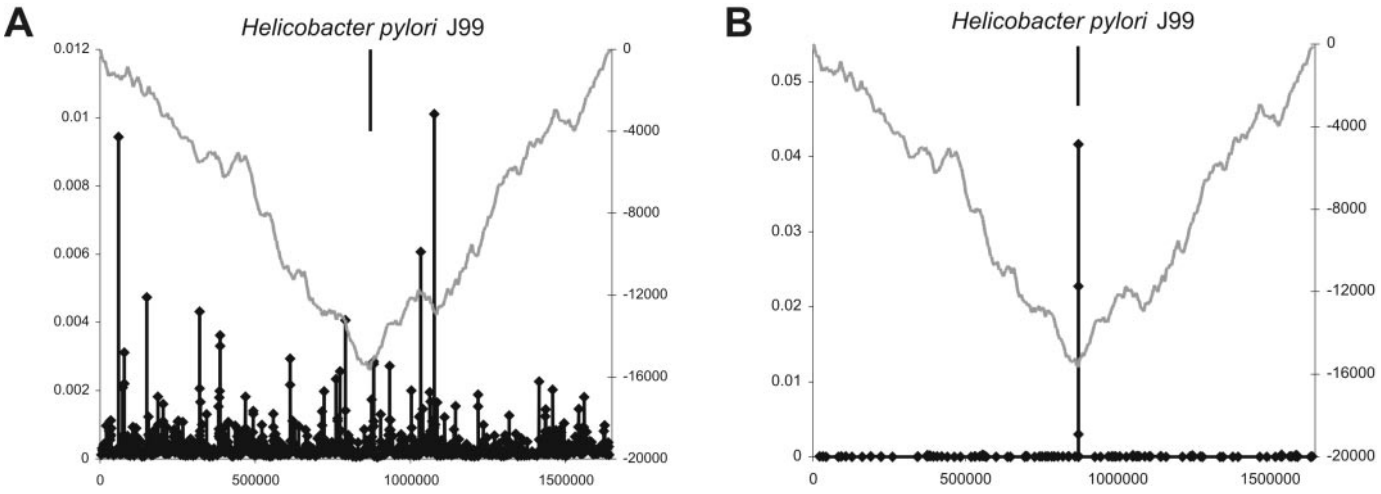

Figure 2. DNA asymmetry in [G–C], DnaA box distribution, and *dnaA* gene location on the *H.pylori* J99 chromosome. (**A**) Distribution of DnaA boxes which differ in more than one position from the perfect DnaA box TTATCCACA; (**B**) distribution of DnaA boxes experimentally proved to be functional in this genome according to Zawilak *et al*. (26). The *x*-axis describing the coordinate position on the chromosome in base pairs begins and ends in the presumed (according to DNA asymmetry) terminus of replication. See Figure 1 for further explanation.

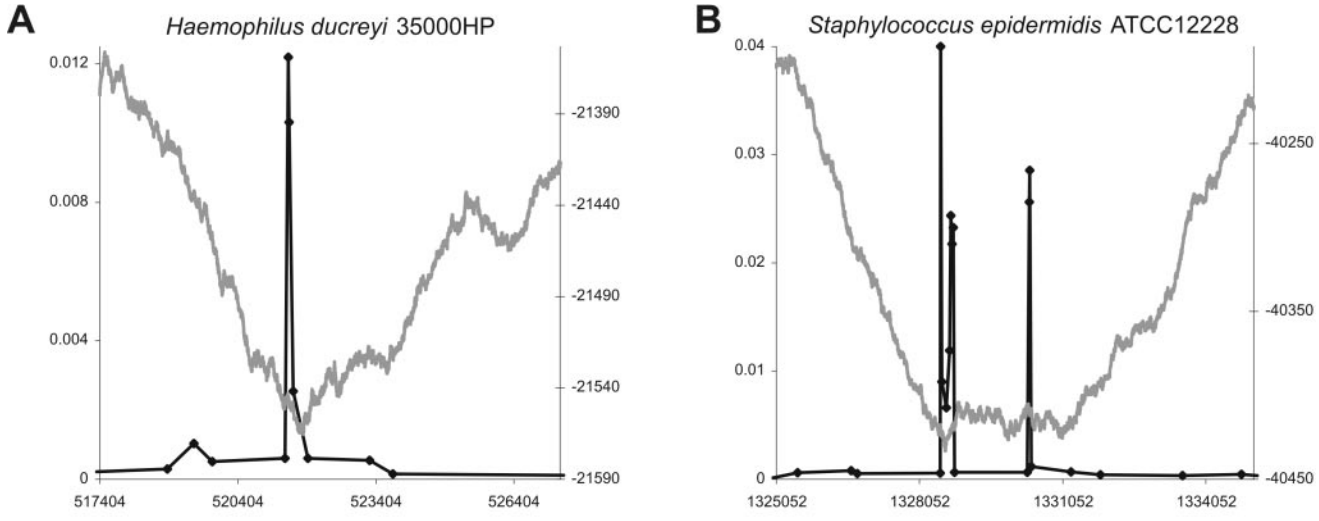

Figure 3. DNA asymmetry in [G–C], DnaA box' distribution and *dnaA* gene location in the *oriC* region. Example of single cluster of DnaA boxes (**A**) and a few separated clusters in the region (**B**). See Figure 1 for further explanation.

dnaA genes have been found in all bacterial genomes studied so far, with the exception of the *Wigglesworthia glossinidia* and *Blochmannia floridanus* genomes (which are reduced in size). In addition, the chromosomes of both endosymbionts do not contain any clusters of DnaA boxes. The lack of typical bacterial elements involved in the initiation of replication may reflect the extreme dependency of *W.glossinidia* and *B.floridanus* on the host genome functions. These bacteria are freely located in the cytosol of bacteriocytes. Bacteria in the cytosol might be a potential danger for the host cell and more difficult to control. The development of a stable symbiosis with cytosolic bacteria probably requires more direct control of DNA replication of the symbionts by the host, involving the loss of the *dnaA* gene (27).

If a genome consists of two chromosomes, one of them lacks the *dnaA* gene (Supplementary Table 1). Two *dnaA* genes are present in chromosomes of the following organisms: *Chlamydia muridarum*, *Chlamydia trachomatis*, *Chlamydia pneumoniae*, *Chlamydia caviae* and *Pirellula* sp. However, it is not yet known whether only one of the *dnaA* genes or both are transcriptionally active. In *Chlamydia* it cannot be excluded that two *dnaA* genes are differentially transcribed during the organism's unusual intracellular developmental life cycle. In *Brucella*, the additional *dnaA* gene present on the larger chromosome was probably transferred from the smaller chromosome (via interchromosomal exchange).

As expected, DnaA boxes (differing in at the most one position from the perfect *E.coli* DnaA box TTATCCACA) were found in all chromosomes. Interestingly, the DnaA boxes were overrepresented (compared with their random occurrence expected from the base composition of a given chromosome) in 70% of the analysed chromosomes. In this

group of chromosomes, the overrepresentation was statistically significant (with $p < 0.05$ in 72% cases). We assume that a potential cluster of DnaA boxes consists of a set of three boxes with an average distance between them of <100 bp (the distance 100 bp corresponds to a *b* value of 0.005 on the left *y*-axis in Figures 1–3). Such clusters (located anywhere on a chromosome) were present in 83% of chromosomes (85% of species).

Evaluation of the three methods in *oriC* identification

We assumed that the methods will identify the same region as the putative origin of replication if the distances between the location of the cluster of DnaA boxes and/or the *dnaA* gene and/or the position of the extreme of the DNA walk are <1% of the chromosome's length. Some doubtful cases (e.g. when the distance was >1% and <7% of chromosome length) were assigned by using parentheses and described by a suitable notation (see Supplementary Table 1). The assumed 1% limit corresponds to the distance between the *dnaA* gene and the experimentally identified *oriC* region in *E.coli*. In most cases, the discrepancy between the three methods results from fluctuation of DNA asymmetry in the extreme in the local scale (see below for potential causes of the fluctuations).

Taking these assumptions into account, we have classified the chromosomes into five distinct groups, depending on the co-localization of the putative *oriC* indicated by the three methods (the abbreviations mean: ***a***—method of DNA asymmetry; ***b***—DnaA box distribution; ***d***—*dnaA* gene location):

(i) ***abd***: all three methods identify the same region as *oriC*—67 chromosomes (Figure 1A);
(ii) ***ab***: the extreme of the DNA walk (asymmetry) is adjacent to the cluster of DnaA boxes—15 chromosomes (Figure 1B);
(iii) ***ad***: the extreme of the DNA walk (asymmetry) corresponds to the location of the *dnaA* gene—18 chromosomes (Figure 1C);
(iv) ***bd***: the *dnaA* gene is in the vicinity of the cluster of DnaA boxes—3 chromosomes (Figure 1D);
(v) ***O***: each method identifies a different region—17 chromosomes (Figure 1E).

The close vicinity of these three features suggests that the analysed region contains a functional origin. So far, only 11 of the putative *oriC* regions have been shown *in vivo* to be functional. Eight of them belong to the first group, ***abd***; their *in silico* predicted *oriC* regions co-localized the functional origins.

The three methods applied to the identification of *oriC* agree rather well. In 55% of cases all methods indicate the same region. The most universal method seems to be DNA asymmetry, which was in agreement with at least one of the other methods in 83% of cases. There are only three cases (***bd***) where the extreme of DNA asymmetry is distant from both the cluster of DnaA boxes and the *dnaA* gene. In some cases [groups ***bd***, (***ab***), (***a***)***bd***] there is imperfect agreement between the location of the cluster of DnaA boxes and/or the *dnaA* gene and DNA asymmetry, which performs some fluctuations near the global extreme. The asymmetry is probably disturbed by other phenomena (such as transcription, chromosome rearrangements, oligomers and gene distributions [for review see (19–22)]. In some cases, the observed shift in the extreme of asymmetry from the cluster of DnaA boxes and the *dnaA* gene may reflect a very recent translocation of the origin region as well. In this case, replication-associated mutational pressure probably had not enough time to erase the old compositional bias and introduce the new one. The smooth global extremes observed in the chromosomes of *Mesorhizobium loti* (Figure 1D) and *Nitrosomonas europaea* may suggest such a possibility.

In the chromosomes grouped in ***ab*** and ***ab***(***d***), the *dnaA* gene was probably transferred to the other region of the chromosome. This may indicate that a close proximity of the *dnaA* gene to the origin region is not necessary for its proper functioning (its product can act in *trans*). Furthermore, it has been reported that mutants of *Synechocystis* sp. with inactivated *dnaA* gene were viable (28). It cannot be excluded that replication of this chromosome is initiated by another protein. However, in most cases the *dnaA* gene is in the vicinity of the *oriC* region. The reason for the close proximity of both elements is not clear. It has been noted that genomic rearrangements rarely occur in the close vicinity of the origins of replication (5). It is possible that the juxtaposition ensures that the DnaA protein is able to associate with the origin as soon as it is synthesized. Alternatively, the proximity to the origin minimizes the possibility of disruption of the linkage between the well-cooperating *dnaA* gene–*oriC* region system.

In some cases [***ad***], a cluster of DnaA boxes could not be identified in the entire chromosome. However, it must be noted that the preferred DnaA box sequences of certain bacteria may differ slightly from the *E.coli* perfect DnaA box (TTATCCACA), e.g. in high-G + C organisms such as *Micrococcus*, *Mycobacterium* or *Streptomyces* (GC > 70%), the third position is substituted by G or C. In addition, our own recent experiments [(26,29) and data not shown] have demonstrated that particular DnaA proteins may exhibit different specificity towards DNA than that from *E.coli*, e.g. the affinity of the DnaA protein from *H.pylori* to the TCATTCACA sequence (two mismatches) is higher than to the *E.coli* perfect box. Therefore, the possibility of a different consensus sequence for the DnaA box (and/or binding specificity of DnaA protein) should be considered when the search of a putative DnaA protein binding site is performed. For example, our first search for DnaA box motifs whose sequences differ by no more than one base from the *E.coli* perfect box failed to find a single cluster of DnaA boxes in the entire *H.pylori* chromosome (Figure 2A). However, when we looked for DnaA box motifs with two to three mismatches, we found a cluster of DnaA boxes upstream of the *dnaA* gene (Figure 2B). This region, designated as the *H.pylori* putative origin of replication, is the only place on the chromosome that contains a cluster of DnaA boxes. The *H.pylori* DnaA exhibits a high affinity to this region (29). There are also many other examples of origin regions that contain 'species-specific' DnaA boxes, e.g. *Borrelia burgdorferi*, TTTTAAACA, TTATCAAAA, TTTTAAAAA (24); *Thermotoga maritima*, AAACCTACCACC (30); *Caulobacter crescentus*, TGATCCACA (8); and *Prochlorococcus marinus*, [AT]TTCCACA (31).

The chromosomes that show asymmetry with more than two distinct extrema are supposed to be replicated from more than one *oriC*. However, we found in some of them a correlation

between the main local extreme of asymmetry and the location of the cluster of DnaA boxes and/or the *dnaA* gene: *Aquifex aeolicus*, *Bifidobacterium longum*, *Gloeobacter violaceus*, *Nostoc* sp. (Figure 1F) and *Synechocystis* sp. The region may be suspected to be the main (the strongest or earlier activated) origin of replication. However, there are many phenomena that influence the course of DNA asymmetry which may disturb the asymmetry introduced by replication-associated mutational pressure [for review see (19–22)]. Only experimental studies can definitely solve this problem. Interestingly, in the linear chromosome of *S.avermitilis* there are two distinct minima suspected of being the origin region (Figure 1G). One of these strongly correlates with the location of distinct clusters of DnaA boxes and the *dnaA* gene. A few DnaA boxes were also found to be close to the second minimum, but they were rather scattered. The arrangement of the DnaA boxes in close vicinity of the *dnaA* gene is the same as within the functional *oriC* region of *Streptomyces coelicolor*. Interestingly, despite selective pressures or the advantage of central positioning of the origin (to finish replication synchronously), the putative *oriC* of *S.avermitilis* is shifted ∼800 000 bp (∼10% of the chromosome's length) from the centre. It has been proposed that linear *Streptomyces* chromosomes resulted from recombination between a linear plasmid and an ancestral circular chromosome (32). Probably, in the case of *S.avermitilis*, linearization events resulted in the formation of chromosomal arms with different lengths (the *oriC* is asymmetrically positioned) and therefore the second minimum is observed (see Figure 1G).

Lack of any agreement between the three methods was usually observed in the second chromosomes (the smaller ones). In chromosome II of *Brucella suis* the extreme of DNA asymmetry correlates with the location of *repC*, a typical plasmid replication gene. It has been suggested that the chromosome of *B.suis* is derived from a megaplasmid that was captured by ancestral *Brucella* (33). Recent experimental studies on the initiation of replication of two *Vibrio cholerae* chromosomes revealed distinct replication requirements for the two chromosomes. The functional origins, $oriCI_{VC}$ and $oriCII_{VC}$, of *V.cholerae* chromosomes are unrelated; the structure of $oriCI_{VC}$ resembles that of *E.coli*, whereas $oriCII_{VC}$ shares some features with certain plasmid replicons (34). The distinct replication requirements of the two chromosomes may minimize competition and thereby help ensure the maintenance of the divided genome. The situation is similar to the observed incompatibility of two plasmids which require identical replication initiation factors.

In organisms which contain a single chromosome and belong to group ***O***, the initiation of chromosome replication is probably also mediated by another protein(s).

Distribution, organization and composition of DnaA boxes

We have found in numerous analysed chromosomes (Supplementary Table 1) that the clusters of DnaA boxes are not distributed randomly in a chromosome, but are densely packed within a relatively short distance around the putative *oriC* region (Figure 1H). The distance between the distal cluster of DnaA boxes and the *oriC* varies from 27 to 800 kb (this being up to 23% of chromosome length). The DnaA boxes, most probably involved in initiation of replication, were identified in 76 chromosomes. Detailed data (locations of these boxes and their sequence patterns) are included in the Supplementary Material (Supplementary Table 2), whereas in Table 1 we present the general organization of the clusters (number of boxes in a given cluster, mean distance between boxes in the cluster and distances between clusters). 41 chromosomes contain at least two separated clusters of DnaA boxes near the putative *oriC* region. In most cases, the distance between the clusters of DnaA boxes is only a few kilobase pairs (up to 7 kb). Longer distances are in *Neisseria* (12–13 kb) and Actinobacteria (mycobacteria, 15–150 kb; *S.coelicolor* and *Corynebacterium efficiens*, 30 kb; *Corynebacterium diphtheriae*, 25 kb).

There are a few possible explanations for the occurrence of many clusters of DnaA boxes. One explanation assumes that more than one cluster participates in the initiation of replication. In several chromosomes which have at least two clusters of DnaA boxes in the putative origin region, we observed two local extrema and a plateau between them in the DNA walk describing asymmetry, e.g. *Staphylococcus epidermidis* (Figure 3, see for comparison the V-shaped asymmetry of the *Haemophilus ducreyi* chromosome). This may suggest that replication starts from one or both sites. In this case, the region corresponding to the plateau would be replicated sometimes as the leading strand and sometimes as the lagging strand. However, it has been experimentally demonstrated that the single *Bacillus subtilis oriC* region is composed of two DnaA clusters (separated by the *dnaA* gene, ∼1500 bp), which are both required for a functional origin. The unwinding of DNA occurs at the 27 bp AT-rich cluster in the vicinity of *dnaN* gene. In *Mycobacterium* (*Mycobacterium avium* and *Mycobacterium tuberculosis*), the *dnaA* gene is also flanked by two clusters of DnaA boxes. However, only one of them acts as *oriC*. Interestingly, in *M.tuberculosis* the 3′ region of the *dnaA* gene exhibits *oriC* activity, while in *M.avium* the 5′ region functions as *oriC* (35).

The second explanation assumes that the clusters of DnaA boxes (which are not involved in the initiation of replication) may contribute to the negative control of replication by reducing the level of free DnaA protein. Since DnaA proteins interact cooperatively with closely spaced DnaA boxes, the clusters of boxes can bind more molecules than single boxes dispersed randomly along the chromosome. In *E.coli* there are five regions of the chromosome which bind DnaA protein (36). DnaA protein exhibits particularly high affinity towards one of them, *datA*, which is located close to the *oriC*. The *datA* region has been suggested as playing a pivotal role in controlling replication initiation; it can bind eight times more DnaA protein molecules than the *oriC* region (37,38). Thus, shortly after initiation of replication, the *datA* region is duplicated and therefore is able to reduce significantly free DnaA below the level that provokes over-initiation. The presence of many clusters of DnaA boxes near the *oriC* region in chromosomes other than that of *E.coli* suggests that such replication control may be a common mechanism for bacteria.

Table 2 presents the general (mean) composition of DnaA boxes, counted only for boxes found in the most probable *oriC* regions in 76 chromosomes. The most variable are the third and the fifth positions in the consensus. This is in agreement with the recently established crystal structure of the *E.coli*

Table 1. Organization of the clusters of DnaA boxes in the putative *oriC* regions

Chromosome	Cluster(s) of DnaA boxes and number of boxes in the cluster(s)[a]	Mean distance between boxes in the cluster and distances between clusters[b]
Agrobacterium tumefaciens C58 Cereon circular chr.	(3)	(**70**)
Aquifex aeolicus VF5	(9)	(**35**)
Bacillus anthracis Ames	(4)(9)(4)	(**47**)*2049*(**23**)*1505*(**18**)
Bacillus cereus ATCC14579	(3)(9)(4)	(**8**)*2195*(**23**)*1503*(**18**)
Bacillus halodurans	(4)-(9)(2)	(**21**)-*6855*-(**29**)*1543*(**34**)
Bacillus subtilis 168	(4)(8)(3)	(**20**)*2515*(**27**)*1519*(**29**)
Bdellovibrio bacteriovorus HD100	(3)	(**89**)
Bifidobacterium longum NCC2705	(8)(4)(7)	(**23**)*1687*(**87**)*3490*(**23**)
Bordetella bronchiseptica RB50	(3)	(**91**)
Bordetella parapertussis 12822	(3)	(**91**)
Bordetella pertussis Tohama I	(3)	(**91**)
Brucella melitensis 16 M chr. I	(4)	(**64**)
Brucella suis 1330 chr. I	(3)	(**36**)
Chlorobium tepidum TLS	(3)(3)	(**53**)*1737*(**2**)
Chromobacterium violaceum ATCC12472	(10)	(**70**)
Clostridium acetobutylicum ATCC824	(8)(5)	(**29**)*1496*(**22**)
Clostridium perfringens 13	(3)(3)(4)-(13)(6)	(**24**)*1938*(**156**)*1261*(**17**)-*2463*-(**51**)*1443*(**37**)
Clostridium tetani E88	(3)(5)	(**24**)*1631*(**17**)
Corynebacterium diphtheriae NCTC13129	(7)(4)-(9)	(**33**)*1918*(**51**)-*24746*-(**33**)
Corynebacterium efficiens YS-314	(4)(7)-(11)	(**39**)*2004*(**54**)-*29466*-(**36**)
Corynebacterium glutamicum ATCC13032	(6)(8)	(**36**)*1782*(**33**)
Deinococcus radiodurans R1 chr. I	(3)(13)	(**49**)*4298*(**27**)
Enterococcus faecalis V583	(9)(5)	(**34**)*1388*(**12**)
Escherichia coli K12	(4)	(**51**)
Escherichia coli O157:H7 RIMD0509952	(4)	(**51**)
Geobacter sulfurreducens PCA	(4)	(**29**)
Haemophilus ducreyi 35000HP	(4)	(**51**)
Lactobacillus johnsonii NCC533	(8)(2)	(**36**)*1399*(**9**)
Lactobacillus plantarum WCFS1	(6)(5)	(**43**)*1423*(**12**)
Lactococcus lactis IL1403	(6)(4)	(**33**)*1393*(**101**)
Listeria innocua Clip11262	(11)(5)	(**28**)*1415*(**10**)
Listeria monocytogenes EGD	(12)(5)	(**24**)*1414*(**10**)
Mycobacterium avium paratuberculosis k10	(3)-(3)-(3)	(**3**)-*61937*-(**29**)-*117564*-(**36**)
Mycobacterium bovis AF2122/97	(3)-(3)	(**36**)-*152476*-(**29**)
Mycobacterium leprae TN	(3)-(3)	(**63**)-*15208*-(**3**)
Mycobacterium tuberculosis H37Rv	(3)-(3)	(**36**)-*154798*-(**29**)
Mycoplasma mycoides mycoides SC	(3)	(**61**)
Mycoplasma pulmonis UAB CTIP	(3)	(**33**)
Neisseria meningitidis MC58 B	(4)-(4)(7)	(**27**)-*12400*-(**167**)*736*(**113**)
Neisseria meningitidis Z2491 A	(3)-(4)(7)-(4)	(**46**)-*12399*-(**167**)*736*(**113**)-*13222*-(**116**)
Nitrosomonas europaea ATCC19718	(3)	(**15**)
Nostoc sp. PCC7120	(6)	(**27**)
Oceanobacillus iheyensis HTE831	(4)-(9)(7)	(**10**)-*2834*-(**32**)*1695*(**98**)
Pasteurella multocida PM70	(3)	(**36**)
Photorhabdus luminescens laumondii TTO1	(5)	(**81**)
Pseudomonas aeruginosa PAO1	(5)-(5)	(**36**)-*4952*-(**57**)
Pseudomonas putida KT2440	(7)-(8)	(**35**)-*4433*-(**31**)
Pseudomonas syringae pv. tomato DC3000	(6)-(4)	(**47**)-*4364*-(**79**)
Ralstonia solanacearum GMI1000	(3)	(**81**)
Rickettsia conorii Malish 7	(6)	(**47**)
Rickettsia prowazekii Madrid E	(7)	(**72**)
Salmonella enterica Typhi	(4)	(**51**)
Salmonella typhimurium LT2	(4)	(**51**)
Shewanella oneidensis MR-1	(5)-(4)	(**78**)-*6170*-(**107**)
Shigella flexneri 2a 301	(3)	(**81**)
Sinorhizobium meliloti 1021	(3)	(**71**)
Staphylococcus aureus Mu50	(6)-(4)	(**25**)-*1498*-(**19**)
Staphylococcus epidermidis ATCC12228	(9)(4)	(**28**)*1522*(**19**)
Streptococcus agalactiae 2603V/R	(4)(5)(3)	(**24**)*2159*(**39**)*1385*(**33**)
Streptococcus mutans UA159	(4)-(3)	(**27**)-*1416*-(**33**)
Streptococcus pneumoniae R6	(6)(3)	(**23**)*1429*(**33**)
Streptococcus pyogenes M1 GAS	(3)-(8)-(3)	(**34**)-*2146*-(**24**)-*1372*-(**33**)
Streptomyces avermitilis MA-4680	(4)	(**54**)
Streptomyces coelicolor A3(2)	(4)-(6)	(**59**)-*30243*-(**62**)
Synechocystis sp. PCC6803	(3)	(**57**)
Thermoanaerobacter tengcongensis MB4T	(12)(3)	(**25**)*1406*(**11**)
Thermosynechococcus elongatus BP-1	(7)	(**25**)
Treponema denticola ATCC35405	(3)	(**70**)

Table 1. *Continued*

Chromosome	Cluster(s) of DnaA boxes and number of boxes in the cluster(s)[a]	Mean distance between boxes in the cluster and distances between clusters[b]
Treponema pallidum Nichols	(3)	(**62**)
Tropheryma whipplei TW08/27	(4)	(**50**)
Vibrio cholerae O1 N16961 chr. I	(5)-(3)	(**51**)-*6882*-(**67**)
Vibrio parahaemolyticus RIMD2210633 chr. I	(5)	(**51**)
Vibrio vulnificus CMCP6 chr. I	(4)-(6)	(**41**)-*6926*-(**62**)
Xanthomonas axonopodis pv. citri 306	(2)(3)	(**26**)*1993*(**90**)
Xylella fastidiosa 9a5c	(3)	(**53**)
Yersinia pestis CO92	(6)	(**63**)

[a]Number in parentheses indicates the number of boxes in the given cluster.
[b]Bold number in parentheses indicates the mean distance between boxes in basepairs, the numbers between them indicate distances between clusters; dashes indicate that between two clusters some single DnaA boxes are present.

Table 2. The composition of DnaA boxes derived from the putative origins identified in 76 chromosomes

Base	Percentage of base in the box position[a]								
	1	2	3	4	5	6	7	8	9
A	1.3	1.4	**78.9**	0.2	5.2	7.6	**96.6**	2.3	**95.3**
T	**93.7**	**94.9**	11.9	**98.7**	5.0	1.4	0.6	0.6	1.3
G	1.9	0.9	6.9	0.0	1.7	0.4	0.0	1.1	0.7
C	3.2	2.8	2.3	1.1	**88.1**	**90.6**	2.8	**96.0**	2.7
Consensus	T	T	A	T	C	C	A	C	A

[a]The highest percentage values for a given position are in bold.

DnaA binding domain in complex with the perfect DnaA box, which showed that the nucleotide at the fifth position does not play an essential role in the interaction with the DnaA protein. We observed statistically significant correlation (r = 0.62) between G + C content of the DnaA boxes and G + C content counted for the given genome (Figure 4). Assuming that global genomic G + C reflects, to some extent, the composition generated by mutational pressure (39), one can assume that composition of the DnaA boxes may be moderated by mutational pressure and that some changes may be neutral. Actually, a sequence comparison of DnaA boxes derived from the origins of organisms with high G + C content revealed that G is frequently present instead of A at the third position of the DnaA box. Bases in the second, fourth, seventh, eighth and ninth positions are highly conserved. Extensive binding studies on *E.coli* revealed that these positions are of particular importance (40,7).

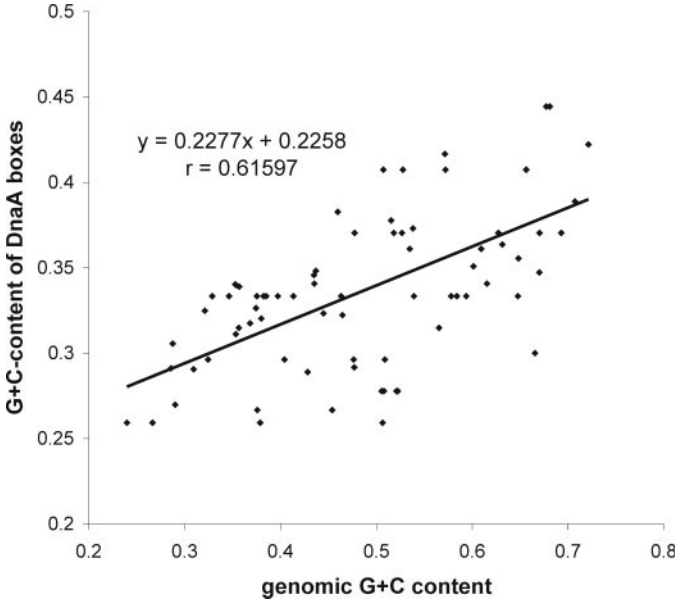

Figure 4. Correlation between G + C content of DnaA boxes (derived from the putative origins identified in 76 chromosomes) and genomic G + C content.

Phylogenetic approach

We compared the phylogenetic (taxonomic) status of the analysed organisms with their classification based on DNA asymmetry, DnaA box distribution, and *dnaA* gene location (see Supplementary Table 1 and Figure 5). Strains belonging to the same species or even the same genus do not exhibit significant differences in DNA asymmetry, distribution of DnaA boxes and *dnaA* gene location on the chromosome. There are a few exceptions: (i) among *Mycoplasma* genomes, only *Mycoplasma pulmonis* shows full agreement of the three methods (it belongs to group ***abd***); (ii) the genome of *Bordetella pertussis* shows more rough asymmetry (with some local extrema) than other *Bordetella* genomes; (iii) *Haemophilus ducreyi* and *Haemophilus influenzae* were assigned to different groups, ***ad*** and ***O***, respectively; (iv) *Streptomyces* genomes differ in DNA asymmetry; and (v) *Treponema denticola* shows some shift of the *dnaA* gene compared with *Treponema pallidum*.

We observed a strong connection between the phylogenetic relations among the analysed bacteria and the classification based on the three parameters: DNA asymmetry, DnaA box distribution and *dnaA* gene location. The phylogenetic relationships of the organisms were evaluated by an analysis of the 16S rRNA gene sequences (see phylogenetic tree, Figure 5). All analysed organisms belonging to two subgroups of the Firmicutes (G + C-low Gram positive), Bacilli and Clostridia,

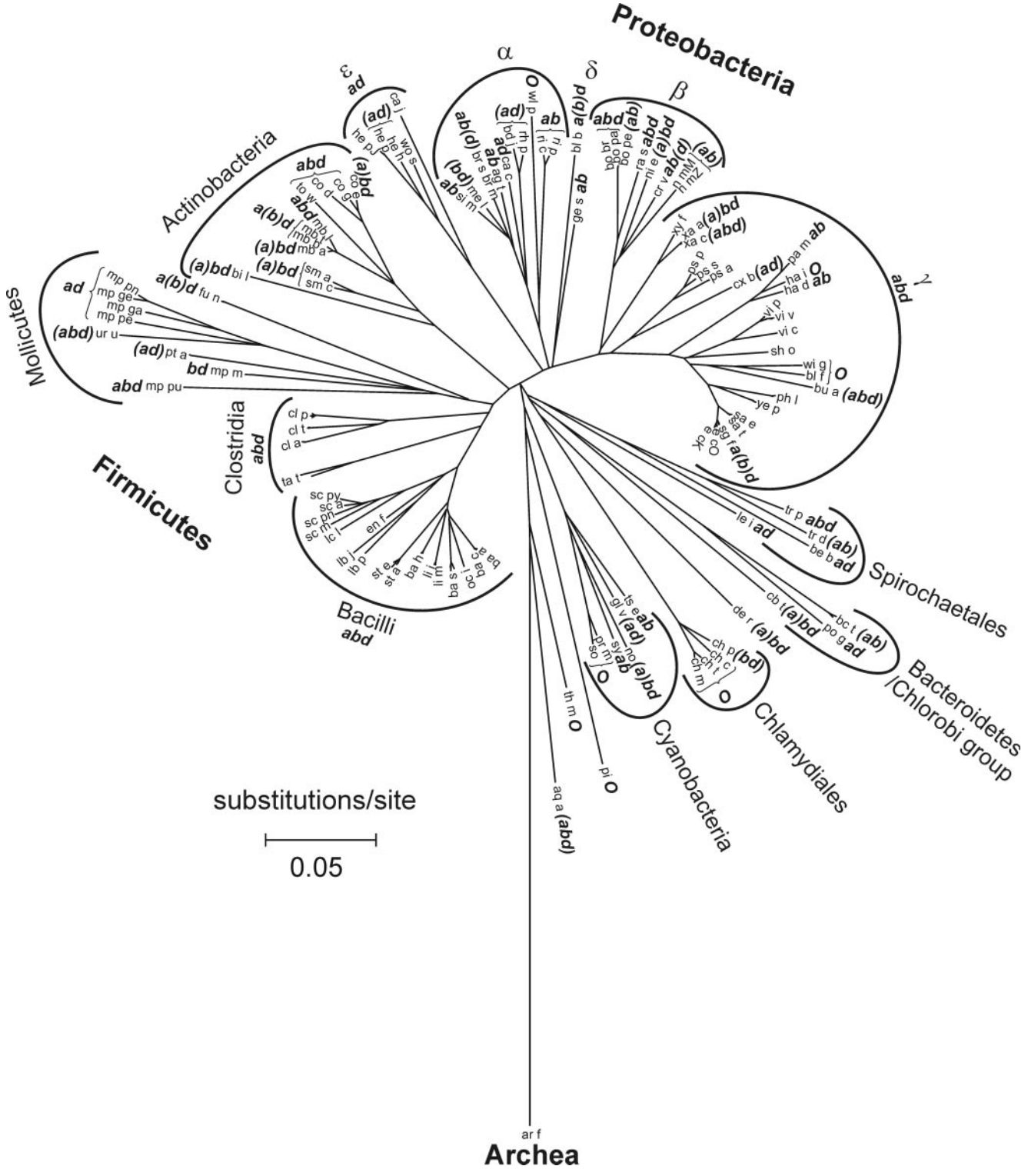

Figure 5. The 16S rRNA phylogenetic tree for 112 analysed bacteria and their classification based on concordance of methods of *oriC* identification: DNA asymmetry, DnaA box distribution and the *dnaA* gene location. The tree was built with the minimum evolution method (ME) by the MEGA 2.1 program (42) assuming the Tamura–Nei model of nucleotide substitutions (43). Abbreviations of the bacterial names are listed in Supplementary Table 1. *Archaeoglobus fulgidus* (ar f) was chosen as representative of the Archaea.

and many organisms of the Actinobacteria (G + C-high Gram positive) belong to the ***abd*** group. Full correlation of these methods is not observed in other Firmicutes, in most Mollicutes whose genomes underwent considerable reduction because of a parasitic style of life. In all the analysed genomes of ε Proteobacteria, the extreme of DNA asymmetry co-localizes only with the *dnaA* gene (the group ***ad***). However, it cannot be excluded that the 'specific' DnaA boxes for this

lineage have evolved and, therefore, in some cases it is difficult to find a cluster of DnaA boxes. The DnaA boxes found within the *H.pylori oriC* region support this hypothesis. In many chromosomes of organisms belonging to α Proteobacteria, a translocation of the *dnaA* gene from the putative origin of replication to another place on the chromosome has been observed. Some of these bacteria are obligatory intracellular parasites or symbionts. The β Proteobacteria genomes exhibit the highest diversity. Nevertheless, some of them show full (or almost full) agreement of the three analysed parameters. Most of the γ Proteobacteria genomes belong to the ***abd*** group. Within this lineage we observed three kinds of exceptions: (i) a distinct cluster of DnaA boxes in the putative origin region is not present in *Coxiella burnetii* (obligatory intracellular parasite); (ii) the *dnaA* gene is translocated from the origin region in *Pasteurella multocida* and *H.ducreyi*; and (iii) lack of correlation of the three parameters is observed in *Haemophilus influenzae* and two endosymbionts, *W.glossinidia* and *B.floridanus* (neither contains the *dnaA* gene).

In Chlamydiales (obligatory intracellular parasites), *Pirellula* and hyperthermophiles (*Aquifex aeolicus* and *Thermotoga maritima*), lack of correlation of the three parameters has also been observed. No universal rules can be determined for Spirochaetales, the Bacteroidetes/Chlorobi group and Cyanobacteria.

The results suggest that the state of ***abd*** (full agreement of the three methods) is probably the ancestral state for Firmicutes, Actinobacteria and Proteobacteria and was then modified in some lineages belonging to these groups, particularly in Proteobacteria and Mollicutes (usually because of some disturbed DNA asymmetry, translocation or disappearance of the *dnaA* gene and decay of the pattern or clusters of the *E.coli* perfect DnaA boxes). In contrast to large genomes, the small genomes of obligatory intracellular parasites or endosymbionts and extremophiles lost many regulatory elements (41), including an extra cluster(s) of DnaA boxes that in other organisms regulates the level of free DnaA protein (titration mechanism). Even the loss of the *dnaA* gene was observed. This aspect may also be partially attributable to a lack of need; living continuously within the host or under extreme conditions (e.g. high temperature) eliminates the environmental fluctuations encountered by free-living bacteria (41).

CONCLUSIONS

The results presented allow us to draw some major conclusions:

(i) The most universal method of putative *oriC* identification in bacterial chromosomes is DNA asymmetry, although in some cases it is necessary to apply all three methods.
(ii) Many bacterial chromosomes show an overrepresentation of DnaA boxes and the presence of more than one cluster of boxes near the *oriC* region. The boxes may be involved in the titration of free DnaA protein and play a role in the regulation of chromosome replication.
(iii) The *E.coli* DnaA-like boxes are present in most bacterial chromosomes. However, they have not been found in some lineages and groups, e.g. ε Proteobacteria. The sequence of DnaA boxes in these organisms was probably modified by mutational pressure.
(iv) Signals indicating the *oriC* region are ambiguous or disappear in bacteria which have become obligatory intracellular parasites or endosymbionts with significantly reduced genomes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The work was supported by grant number 1016/IGM/04 and by the Ministry of Scientific Research and Information Research (The State Committee for Scientific Research, grant 3 P04A 079 22).

REFERENCES

1. Kornberg,A. and Baker,T. (1992) *The DNA Replication, 2nd edn*. Freeman W.H. and Co., New York, NY.
2. Baker,T.A. and Bell,S.P. (1998) Polymerases and the replisome: machines within machines. *Cell*, **92**, 295–305.
3. Matsunaga,F., Forterre,P., Ishino,Y. and Myllykallio,H. (2001) *In vivo* interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. *Proc. Natl Acad. Sci. USA*, **98**, 11152–11157.
4. Robinson,N.P., Dionne,I., Lundgren,M., Marsh,V.L., Bernander,R. and Bell,S.D. (2004) Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell*, **116**, 25–38.
5. Kelman,L.M. and Kelman,Z. (2003) Archaea: an archetype for replication initiation studies? *Mol. Microbiol*., **48**, 605–615.
6. Messer,W. and Weigel,C. (1996) Initiation of chromosome replication. In Neidhardt,F.C. (ed.), *Escherichia coli and Salmonella Cellular and Molecular Biology, 2nd edn*. ASM Press Washington, DC, pp. 1579–1601.
7. Messer,W. (2002) The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication. *FEMS Microbiol. Rev*., **26**, 355–374.
8. Marczyński,G.T. and Shapiro,L. (2002) Control of chromosome replication in *Caulobacter crescentus*. *Annu. Rev. Microbiol*., **56**, 625–656.
9. Suhan,M., Chen,S.Y., Thompson,H.A., Hoover,T.A., Hill,A. and Williams,J.C. (1994) Cloning and characterization of an autonomous replication sequence from *Coxiella burnetii*. *J. Bacteriol*., **176**, 5233–5243.
10. Lobry,J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol*., **13**, 660–665.
11. Lobry,J.R. (1996) A simple vectorial representation of DNA sequence for the detection of replication origins in bacteria. *Biochimie*, **78**, 323–326.
12. Freeman,J.M., Plasterer,T.N., Smith,T.F. and Mohr,S.C. (1998) Patterns of genome organization in bacteria. *Science*, **279**, 1827.
13. Grigoriev,A. (1998) Analysing genomes with cumulative skew diagrams. *Nucleic Acids Res*., **26**, 2286–2290.
14. McLean,M.J., Wolfe,K.H. and Devine,K.M. (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol*., **47**, 691–696.
15. Salzberg,S.L., Salzberg,A.J., Kerlavage,A.R. and Tomb,J-F. (1998) Skewed oligomers and origins of replication. *Gene*, **217**, 57–67.
16. Mackiewicz,P., Gierlik,A., Kowalczuk,M., Dudek,M.R. and Cebrat,S. (1999) How does replication-associated mutational pressure influence amino acid composition of proteins? *Genome Res*., **9**, 409–416.
17. Mackiewicz,P., Gierlik,A., Kowalczuk,M., Dudek,M.R. and Cebrat,S. (1999) Asymmetry of nucleotide composition of prokaryotic chromosomes. *J. Appl. Genet*., **40**, 1–14.
18. Rocha,E.P., Danchin,A. and Viari,A. (1999) Universal replication biases in bacteria. *Mol. Microbiol*., **32**, 11–16.

19. Mrazek,J. and Karlin,S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl Acad. Sci. USA*, **95**, 3720–3725.
20. Frank,A.C. and Lobry,J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, **238**, 65–77.
21. Tillier,E.R. and Collins,R.A. (2000) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.*, **50**, 249–257.
22. Kowalczuk,M., Mackiewicz,P., Mackiewicz,D., Nowicka,A., Dudkiewicz,M., Dudek,M.R. and Cebrat,S. (2001) DNA asymmetry and the replicational mutational pressure. *J. Appl. Genet.*, **42**, 553–577.
23. Frank,A.C. and Lobry,J.R. (2000) Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics*, **16**, 560–561.
24. Picardeau,M., Lobry,J.R. and Hinnebusch,B.J. (1999) Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome. *Mol. Microbiol.*, **32**, 437–445.
25. Picardeau,M., Lobry,J.R. and Hinnebusch,B.J. (2000) Analyzing DNA strand compositional asymmetry to identify candidate replication origins of *Borrelia burgdorferi* linear and circular plasmids. *Genome Res.*, **10**, 1594–1604.
26. Zawilak,A., Cebrat,S., Mackiewicz,P., Król-Hulewicz,A., Jakimowicz,D., Messer,W., Gościniak,G. and Zakrzewska-Czerwinska,J. (2001) Identification of a putative chromosomal replication origin from *Helicobacter pylori* and its interaction with the initiator protein DnaA. *Nucleic Acids Res.*, **29**, 2251–2259.
27. Gil,R., Silva,F.J., Zientz,E., Delmotte,F., Gonzalez-Candelas,F., Latorre,A., Rausell,C., Kamerbeek,J., Gadau,J., Holldobler,B. *et al.* (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc. Natl Acad. Sci. USA*, **100**, 9388–9393.
28. Richter,S., Hagemann,M. and Messer,W. (1998) Transcriptional analysis and mutation of a dnaA-like gene in *Synechocystis* sp. strain PCC 6803. *J. Bacteriol.*, **180**, 4946–4949.
29. Zawilak,A., Durrant,M.C., Jakimowicz,P., Backert,S. and Zakrzewska-Czerwinska,J. (2003) DNA binding specificity of the replication initiator protein, DnaA from *Helicobacter pylori*. *J. Mol. Biol.*, **334**, 933–947.
30. Lopez,P., Forterre,P., le Guyader,H. and Philippe,H. (2000) Origin of replication of *Thermotoga maritima*. *Trends Genet.*, **16**, 59–60.
31. Richter,D., Hess,W.R., Krause,M. and Messer,W. (1998) Unique organization of the dnaA region from *Prochlorococcus marinus* CCMP1375, a marine cyanobacterium. *Mol. Gen. Genet.*, **257**, 534–541.
32. Chen,C.W., Huang,Ch.-H., Lee,H.-H., Tsai,H.-H. and Kirby,R. (2002) Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends in Genet.*, **18**, 522–529.
33. Paulsen,I.T., Seshadri,R., Nelson,K.E., Eisen,J.A., Heidelberg,J.F., Read,T.D., Dodson,R.J., Umayam,L., Brinkac,L.M. and Beanan,M.J. (2002) The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proc. Natl Acad. Sci. USA*, **99**, 13148–13153.
34. Egan,E.S. and Waldor,M.K. (2003) Distinct replication requirements for the two *Vibrio cholerae* chromosomes. *Cell*, **114**, 521–530.
35. Madiraju,M.V., Qin,M.H., Yamamoto,K., Atkinson,M.A. and Rajagopalan,M. (1999) The *dnaA* gene region of *Mycobacterium avium* and the autonomous replication activities of its 5′ and 3′ flanking regions. *Microbiology*, **145**, 2913–2921.
36. Roth,A. and Messer,W. (1998) High-affinity binding sites for the initiator protein DnaA on the chromosome of *Escherichia coli*. *Mol. Microbiol.*, **28**, 395–401.
37. Boye,E., Løbner-Olesen,A. and Skarstad,K. (2000) Limiting DNA replication to once and only once. *EMBO Rep.*, **1**, 479–483.
38. Ogawa,T., Yamada,Y., Kuroda,T., Kishi,T. and Moriya,S. (2002) The *datA* locus predominantly contributes to the initiator titration mechanism in the control of replication initiation in *Escherichia coli*. *Mol. Microbiol.*, **44**, 1367–1375.
39. Muto,A. and Osawa,S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl Acad. Sci. USA*, **84**, 16616–16619.
40. Schaefer,Ch. and Messer,W. (1991) DnaA protein/DNA interaction. Modulation of the recognition sequence. *Mol. Gen. Genet.*, **226**, 34–40.
41. Moran,N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, **108**, 583–586.
42. Kumar,S., Tamura,K. and Nei,M. (1993) *MEGA: Molecular Evolutionary Genetics Analysis*. Pennsylvania State University, University Park, PA.
43. Tamura,K. and Nei,M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**, 512–526.