Briefings in Bioinformatics Advance Access published July 26, 2008

doi:10.1093/bib/bbn031

Identification of replication origins in prokaryotic genomes

Natalia V. Sernova and Mikhail S. Gelfand

Submitted: 8th April 2008; Received (in revised form): 2nd July 2008

Abstract

The availability of hundreds of complete bacterial genomes has created new challenges and simultaneously opportunities for bioinformatics. In the area of statistical analysis of genomic sequences, the studies of nucleotide compositional bias and gene bias between strands and replichores paved way to the development of tools for prediction of bacterial replication origins. Only a few (about 20) origin regions for eubacteria and archaea have been proven experimentally. One reason for that may be that this is now considered as an essentially bioinformatics problem, where predictions are sufficiently reliable not to run labor-intensive experiments, unless specifically needed. Here we describe the main existing approaches to the identification of replication origin (*oriC*) and termination (*terC*) loci in prokaryotic chromosomes and characterize a number of computational tools based on various skew types and other types of evidence. We also classify the eubacterial and archaeal chromosomes by predictability of their replication origins using skew plots. Finally, we discuss possible combined approaches to the identification of the *oriC* sites that may be used to improve the prediction tools, in particular, the analysis of DnaA binding sites using the comparative genomic methods.

Keywords: *origin of replication; comparative genomics; DnaA-box; genomic composition asymmetry; GC-skew*

OVERVIEW

The replication of genomic DNA is arguably the most important task performed by a cell. Prokaryotic genomes contain one or several chromosomes [1], most of which are circular [2]. The chromosomes consist of two anti-parallel DNA strands, and are supposed to have a single origin of replication (eubacteria) [3] or may have single or multiple origins (archaea) [4]. In eubacteria, the origin locus, *oriC*, is relatively short, from 100 to 1000 bp [5]. The replication starts there and then proceeds bi-directionally, carried out by two replication forks that copy the chromosome in two arcs called replichores and (in circular chromosomes) meet at the replication terminus (*terC*) [6, 7]. The DNA replication in archaea is described in [8].

Semi-conservative DNA synthesis implies that polymerization of a new strand in a replichore involves using each of the old strands as a template. Since elongation is possible only in the $5' \rightarrow 3'$ direction, the new strands are synthesized using different mechanisms: the leading strand is replicated continuously, while the lagging strand is replicated in a fragmented manner, through the assembly of the Okazaki fragments [3, 9].

The first Chargaff parity rule, experimentally established about 60 years ago, says that, in a two-stranded DNA molecule, %A = %T and %C = %G [10]. Later, the same equalities were shown to hold approximately for each separate DNA strand of a bacterial chromosome (*Bacillus subtilis*), yielding the second Chargaff parity rule [11]. The fact that the

Corresponding author. M.S. Gelfand, Institute for Information Transmission Problems (the Kharkevich Institute), Russian Academy of Sciences, Bolshoi Karetny pereulok, 19, Moscow, 127994, Russia. Tel: +7-495-650-4225; Fax: +7-495-650-0579; E-mail: gelfand@iitp.ru

Natalia Sernova obtained PhD in Biophysics at Moscow Institute of Physics and Technology, 2004, and MSc in Proteomics and Bioinformatics, Unige, Switzerland, 2006. The Research interests include Comparative Genomics.

Mikhail Gelfand is the Vice-director for Science, IITP and Professor of Bioinformatics, M.V. Lomonosov Moscow State University, A. A. Baev (2007) and Best Scientist of Russian Academy of Sciences (2004) awards. The Research interests include Comparative Genomics, Evolution of Regulatory Systems, Alternative Splicing.

same approximate equalities, %A≈%T and %C≈%G should be observed within each strand at the equilibrium state, was later derived theoretically [12, 13].

Statistical analyses of the genomic sequences demonstrated local deviations from the second Chargaff rule. One type of such deviations is the differences in the nucleotide composition of the leading and lagging strands [14, 15]. Such differences were observed in almost all bacterial genomes and were called asymmetry (skew).

The hypotheses that could explain the compositional strand bias were listed and discussed in [14, 16]. The diverse mutational biases leading to compositional skews were reviewed in detail in [17]. The list includes cytosine deamination, DNA polymerase processivity, strand-specific protein-binding sites, genome rearrangements, the length of the Okazaki fragments; strand preferences of protein-coding genes, uneven codon usage and transcription-related mutation bias. The current consensus explanation seems to be single-strand cytosine deamination, which causes C → T mutations, and hence the deficit of C in the leading strand that spends more time in the single-strand state [15, 18, 19]. However, most authors agree that the bias is multifactorial.

In many genomes, more genes are transcribed from the leading strand [15, 20], this bias being especially marked for essential or highly expressed genes [21, 22], the transcription biases are combined with replication ones in nonobvious ways. In particular, in *B. subtilis*, A and T are preferred in third codon positions, but not overall on the leading strand [20]. A technique of decoupling of replication and transcription or translation effects using the artificial genome rearrangement approach [23] was developed in [24]. The method allows one to discriminate visually between the contributions of replication-related and coding sequence-related mechanisms to base composition asymmetry.

The accumulated bias is diluted by genome rearrangements, such as insertions, inversions (especially gene switches between strands), deletions and gene losses, etc. Such rearrangements were studied in many bacterial genomes, in particular, the *Neisseria* spp. (translocations, inversions and insertion/deletions) [25], *Salmonella paratyphi* C (insertions and deletions) [26], the *Thermotogales* (large-scale inversions) [27], *Rickettsia typhi* (insertions and inversions) [28], chromosome 2 of *Brucella abortus* (inversions and deletions) [29], the *Leptospira* spp. (gene inversions, duplications and losses) [30], *Yersinia pestis* (insertions, accompanied by recombinations, gene loss and inversions of gene blocks) [31], etc. The genomic rearrangements rarely happen in the vicinity of *oriC*, and usually they are symmetric relative to the axis between the replication origin and terminus [32, 33]. On the other hand, rearrangements in the neighborhood of the replication origin may mask its position, e.g. as in *Helicobacter pylori* [34].

One consequence of the nucleotide skew is the general preference of G-rich oligonucleotides on the leading strand [35]. However, in some cases one may observe uneven distribution of oligonucleotides on the leading and lagging strand that cannot be reduced to the nucleotide skew [36–39]. For example, in *Escherichia coli*, 75% of the ChiEc sites GCTGGTGG occur in the leading strand [40]. Similarly, in *Lactococcus lactis*, 78% of ChiLl sites GCGCGTG occur in the leading strand [41]. The information on Chi sites for several bacteria is summarized in [42]. However, the prevalence of Chi sites on the leading strand is not universal: e.g. it is much less pronounced in the genome of *Haemophilus influenzae* where the Chi-like sites are GNTGGTGG and GSTGGAGG [43, 44]. Other oligonucleotides with clear biological effect and uneven distribution on the leading and lagging strands (around the *dif* site) are KOPS motifs GGGNAGGG directing the movement of the *E. coli* translocase FtsK [39, 45, 46].

Whatever the cause, this skew can be used to distinguish between the leading and lagging strands and hence to identify candidate replication origins and termini. This analysis was pioneered in 1996 by J. Lobry who studied the replication origins of *E. coli*, *H. influenzae*, *B. subtilis*, *Mycoplasma genitalium* and the replication terminus of *H. influenzae* and observed that the relative *CG*-skew $= (C-G)/(C+G)$ changes the sign crossing the *oriC* and *terC* regions [47, 48]. These observations were confirmed for the complete genomes of *B. subtilis* [49] and *E. coli* [50]. The developed computational tools identify the replication origins and termini using the single nucleotide (*CG* and *AT*) skews [47, 48], as well as keto-amino skew, $[(A+C)-(G+T)]$, or purine–pyrimidine skew, $[(A+G)-(C+T)]$ [51]. We shall use the term *nucleotide skew* for all measures of this type.

The nucleotide skew is maintained in evolution, since genes that move to a different strand adapt their nucleotide composition to the new (receiving) strand. Indeed, orthologous genes of *Borrelia burgdorferi* and *Treponema pallidum* that changed strand after the divergence of these two Spirochaetes have the

amino acid and codon usage corresponding to their current replication strand [52]. Similarly, switched genes acquired the skew of their new replication strand in the *Chlamydia* species [53]. The highest divergence rate for genes in closely related genomes was observed for the genes that had changed strand, since they had acquired more substitutions than sequences staying on the same DNA strand [54].

Some software tools use also oligonucleotide-based approaches and either search for particular oligonucleotides with the most pronounced difference in the leading and lagging strand frequencies, as identified by preliminary analysis [55], or consider all oligonucleotides of some fixed length [36, 56].

The biology of replication initiation in prokaryotes was reviewed in many papers that addressed, e.g. regulation of replication initiation in *E. coli* [57]; recent studies with the emphasis on organisms other than *E. coli* [58]; structural features of DNA replication in bacteria and especially traits common to all three domains of life [59]. Details of the archaeal DNA-replication machinery with the emphasis on structural features of the major proteins involved were described in a recent review [8].

Initiation of replication is a tightly regulated process. *E. coli* is conventionally used as a model bacterium for replication studies [60]. The *dnaA* gene, almost ubiquitous in bacteria, encodes the major component of the DNA-replication machinery, the replication initiation factor DnaA. DnaA is a member of the ATPase family (AAA+), it may bind both ATP and ADP, but initiates replication only as ATP–DnaA. Complexed with ATP, DnaA binds to DnaA-boxes, which are repeated several times in the *oriC* region. In *E. coli*, the interactions between DnaA and DnaA-boxes start the replication process and determine all subsequent events at *oriC*, in particular, DNA duplex unwinding and formation of replication forks [61]. The processes in other bacteria are assumed to be the same. Importantly, replication should happen only once per cell cycle [62], and this is guaranteed by multiple mechanisms, such as sequestration of the chromosomal replication origin, *oriC*, autogenous repression, and DnaA-excess titration [58, 63].

The DnaA-boxes are 9-nt nonpalindromic sites [58]. Experimentally proven DnaA-boxes are limited to *E. coli* [64] and few other bacteria, in particular, *Thermus thermophilus* [65], *H. pylori* [66], *Mycobacterium tuberculosis* [67]. On the genome scale, the ChIP–chip technique, based on the chromatin affinity precipitation assay was applied to find DnaA-binding sites in *B. subtilis* [68]. The crystal structure of the complex of DnaA with a DNA fragment was determined, and it elucidated the interactions of DnaA protein with the DnaA-box [69]. This DNA fragment was derived from the DnaA-binding site R1 of the *E. coli oriC* region and contained the consensus DnaA-box of *E. coli*.

The skew characteristics are often only rough indicators of the replication origin position in a chromosome [70], and when more accurate localization of *oriC* is needed recognition of clusters of DnaA-boxes may be attempted. Indeed, the DnaA-box motif in most considered cases is an over-represented word within the *oriC* locus [70]. On the other hand, in several cases where the replication origin was identified experimentally, it did not contain any candidate boxes similar to the *E. coli* DnaA-box motif, e.g. in *B. burgdorferi* [71] and endosymbionts [72]. In some bacteria, like *H. pylori* [66], the DnaA-binding motif is known to be different from the *E. coli* one. Nevertheless, to identify DnaA-sites computationally, genome-wide search for sites similar (up to one mismatch) to the *E. coli* DnaA-box consensus have been applied [70; 73, http://tubic.tju.edu.cn/doric/].

All experimentally confirmed origins are in intergenic regions [70, 74]. The global sequence similarity in these loci is limited to close relatives, as is the case for all intergenic regions. In many cases, the dnaA gene is positioned near the *oriC* site, but this is not a universal rule [74, 75]. For instance, *oriC* in *Coxiella burnetii* is between the *gidA* and *rpmH* genes [76], and in *Caulobacter crescentus* it is in the *hemE/RP001* region [77]. Similar to the eubacteria, the origins in the archaea are commonly located adjacent to genes for initiator proteins Orc1/Cdc6. Almost all archaeal genomes sequenced to date contain at least one gene with homology to both Orc1 and Cdc6 [8].

Much less is known about *terC*. The strand asymmetry switches polarity at the *terC* region [47, 48]. In the past, the main role in the termination of replication has been ascribed to *Ter*-sites [78], but recently this view has been challenged, and currently, the replication termination is believed to occur in the vicinity (within a kilobase) of the *dif* site [6]. The alignment of *dif*-sites from the γ-proteobacteria, Firmicutes and Actinobacteria, produced a consensus sequence DBBBCSBATAAT RTAYATTATGTHAANT [6].

Until recently, it was generally accepted, that archaea follow the bacterial mode of replication, with a single origin and terminus. Indeed, a single origin was found in *Pyrococcus abyssi* [37]. However, subsequent studies demonstrated the existence of multiple replication origins in *Sulfolobus solfataricus* [79, 80] and *Sulfolobus acidocaldarius* [80].

METHODS

The skew calculation is performed as follows. Relative *CG*-skew = $(C-G)/(C+G)$ [47], is measured in percentage. It is different from the absolute skew *CG*-skew = $C-G$, *GC*-skew = $G-C$, *AT*-skew = $A-T$, *TA*-skew = $T-A$; purine–pyrimidine [(A + G)−(C + T)]-skew; keto–amino [(A + C) −(G + T)]-skew; weak hydrogen (H) bond − strong H bond bases [(A + T) − (G + C)]-skew. Cumulative GC-skew is the sum of (G − C) in adjacent windows from an arbitrary start to a given point in a sequence [81], and similarly for other kinds of skew. Most of the considered programs calculate the absolute cumulative skew, measured in kb (Table 1).

Sequences for 486 complete genomes were taken from *GenBank* [87, ftp://ftp.ncbi.nih.gov/genomes/Bacteria].

Sequence alignments were constructed using *Muscle* [88, http://www.drive5.com/muscle]. To display multiple sequence alignments, program *GeneDoc Editor* version 2.6.002 was used (Nicholas, Karl B and Nicholas, Hugh B. Jr. 1997, GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the authors). Positional weight matrices were constructed using *SignalX* [89]. Sequence logos were drawn using *WebLogo* [90]. Clusters of motifs were constructed using *ClusterTree-RS* [91]. Mycobacterial trees were drawn with the help of NJplot program [92]. The skew-curves for eubacteria were downloaded from the *Comparative Genometrics* website (http://www2.unil.ch/comparativegenometrics/index.html) on 01 July 2007, and for the archaebacteria on 20 March 2008.

PROGRAMS

In this section, we review six freely available computational tools for the analysis of bacterial compositional skew and the prediction of the putative *oriC* and *terC* position [56, 73, 74, 82–86] (Table 1).

The programs can process the genomes in the FASTA, GenBank or EMBL formats. The prediction

Table 1: Programs for prediction of origin and terminus of replication in eubacterial chromosomes

Program	*Oriloc*	*CG-software*	*GraphDNA*	*Z-curve*	*Oligonucleotide skew method*	*Ori-Finder, DoriC database*
References	[82]	[83]	[84]	[85, 86]	[56]	[73, 74]
URL	http://pbil.univ-lyon1.fr/software/Oriloc/oriloc.html	http://www2.unil.ch/comparativegenometrics/index.html	http://athena.bioc.uvic.ca/workbench.php?tool=graphdna&db=	http://202.113.12.55/zcurve/	http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/origin/	http://202.113.12.55/Ori-Finder/; http://tubic.tju.edu.cn/doric/
Programming language	R	R	Java 1.4 (JRE)	Fortran77	C	C++, Perl
Source code availability	Yes	Yes, upon request	Yes	Yes	Yes	Yes (*Ori-Finder*)
Input format	GenBank or FASTA	GenBank	GenBank, EMBL, FASTA	FASTA	FASTA	GenBank or FASTA
Type(s) of asymmetry	CG-skew, TA-skew, CDS-skew	GC-skew, TA-skew	GC-skew, AT-skew, AC-skew, GA-skew, GT-skew, TC-skew, ((A + G) − (C + T))-skew, ((A + C) − (G + T))-skew	((A + G) − (C + T))-skew, ((A + C) − (G + T))-skew, ((A + T) − (G + C))-skew, AT-skew, GC-skew	oligonucleotide skew	((A + G) − (C + T))-skew, ((A + C) − (G + T))-skew, ((A + T) − (G + C))-skew AT-skew GC-skew
Output	Plots of CG-, TA-, CDS-skew, combined skew	Plots of GC-skew and TA-skew	Plots of different types of skew (see above)	3D skew plot	Curve-plot and predicted origin position	*Z*-curve, DnaA-boxes, *dif* sequences, location of indicator genes, phylogenetic data
Preprocessed chromosomes	612	658	None	412	325	59 (*Ori-Finder*) 632 (*DoriC*)

algorithms are based on computation of various skew types (Table 1) and present output data as plots and/or tables of plot coordinates. The websites differ in the type of information they provide. Most of them contain collections of plots for pre-processed chromosomes, and allow the user to download the source code of corresponding programs, so that a plot may be built for a sequence missing in the database.

The initial step of the analysis for most of the programs is construction of a DNA-walk, a plot based on the nucleotide distribution along a chromosome [83, http://www2.unil.ch/comparative genometrics/DNA_walk.html; 93]. The genome sequence is broken into windows, and DNA-walk detects local changes of nucleotide composition for each window. The DNA-walk forms the base for calculation of various cumulative skews to build skew-plots, similar to the ones suggested for the cumulative GC-skew [81]. (See Methods section and Table 1 for differences in the skew calculations between the programs.) As a result, all these programs build different cumulative skew plots.

Windows in *Oriloc* are genes taken from a GenBank file (for annotated genomes) or predicted with Glimmer2 or Glimmer3 (for nonannotated genomes). *Oriloc* examines only third codon positions in each gene. Intergenic sequences are not considered. Then *Oriloc* constructs the 'combined' *CG*(*TA*) plot [82, http://pbil.univ-lyon1.fr/software/Oriloc/howto.html], which is used as the main plot for the *oriC*/*terC* prediction. This plot, together with *CG*-skew, *TA*-skew and CDS-skew-plots is presented at the *Oriloc* website.

In the *CG-software (Comparative Genometrics software)*, the sequence is also broken into windows, defined as overlapping sequence fragments (1000 nt), but, unlike *Oriloc*, disregarding the gene annotation. This program calculates the *GC*-skew and *TA*-skew and draws the corresponding plots. It also draws the DNA-walk plot and provides all plot coordinates, which is convenient from the user's point of view.

A similar DNA-walk procedure is implemented in the *GraphDNA* program. This program allows the user to select between many skew types (Table 1).

The *Z-curve* analysis is based on a 3D DNA-walk, where the three dimensions correspond to the $[(A+G)-(C+T)]$ skew, $[(A+C)-(G+T)]$-skew, and $[(A+T)-(G+C)]$-skew [94]. The corresponding website contains also an accessory program *Z-plotter*, which is a web server that accepts sequences supplied by a user and returns the set of plot coordinates (http://202.113.12.55/zcurve/).

The approach to the *oriC* prediction implemented in four programs discussed above is determination of turning points in DNA-walk plots, which correspond to global extrema in the cumulative-skew plots.

At the terminus region, the skew changes sign in the opposite direction. Thus, in general, the same programs may also be used to identify the replication termination site.

Dependent on the details of the computational procedure and the type of the cumulative skew, the origin may correspond to the global minimum or the global maximum of the plot; the reverse holds for the terminus, respectively.

For the 'combined' skew plots in *Oriloc*, the putative origin corresponds to the global maximum (http://pbil.univ-lyon1.fr/software/SeqinR/SEQINR_CRAN/DOC/html/Oriloc.html); for the *GC*-skew plots in *CG-software*, to the global minimum [83]. It relates to the use of CG-skew in Oriloc, and GC-skew in CG-software (Table 1). For *GraphDNA*, it depends on the skew type selected by the user (for example, for the purine-skew, to the global minimum) [95]. The identification criteria of the origins are not well defined for *Z-curve*. The origin is expected to be an extremum, but it may be either the minimum or the maximum [94].

As mentioned above, in some species the direction of the skew for chromosomal DNA (all nucleotides) is opposite to that for the third codon positions (in genes). In particular, this has been observed in *B. subtilis* for the *AT*-skew and in *M. genitalium* and *Mycoplasma pneumoniae* for the *AT*- and GC-skew [20]. We have extended the list of such bacteria. A total of 357 chromosomes (*Oriloc* curves currently available) were analyzed in detail for this 'inverted skew' relation between the third codon position skew and the overall nucleotide skew. To accomplish that, plots, corresponding to particular skew-types in *Oriloc* and *CG-software* were compared. The 'inverted skew' relation was found in 14 cases (13 if strains are merged) for the GC-skew and 37 (respectively, 27) cases for the TA-skew, of which 6 cases had both GC-skew and TA-skew inverted. The complete list of these bacteria is presented in Supplementary List 3. While we could not find a general rule that explained this set of genomes, some observations could be made. Both skews are inverted for many *Mycoplasma* spp., and the

TA-skew is inverted for many other Firmicutes (*Bacillus* spp., *Lactobacillus* spp. and *Streptococcus* spp.). In any case, simultaneous use of two programs, *Oriloc* and *CG-software* allows one to check for this phenomenon in any genome.

Differences in computational procedures influence not only the direction of the skew, but also the range of plot amplitude. In particular, the amplitudes for the CG-skew plot in *Oriloc* are approximately 3-fold smaller, than the amplitudes of the corresponding GC-skew plots in other programs (as an example Figure 1A–C) and also links to the colored plots for *B. subtilis* in the Oriloc (http://pbil.univ-lyon1.fr/software/Oriloc/NC_000964.png), *Z*-curve (http://202.113.12.55/zcurve/img/bacteria/AL009126_xpy.png) and *DoriC* (http://tubic.tju.edu.cn/doric/img/NC_000964.o1.png) collections).

In the *Oligonucleotide skew* method [56], the DNA-walk is not used. Instead, the program selects the optimal *oriC* positions among a set of candidate positions spaced by 1000 bp. For each position, a large window (50–100% of the genome) centered at this position is considered and all oligonucleotides

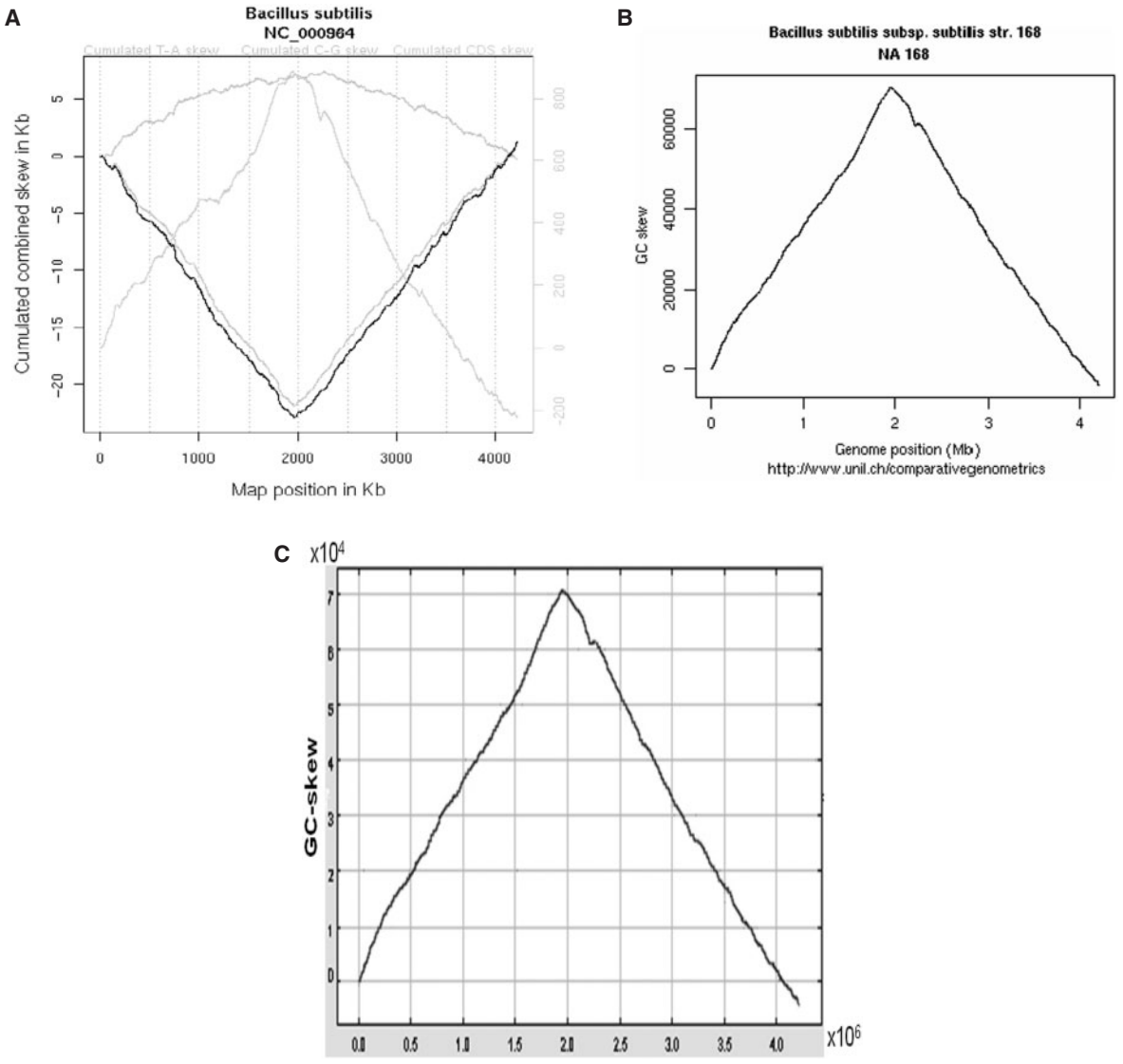

Figure 1: The range of plot amplitudes for *B. subtilis*. (**A**) CG(TA)-skew plot (black curve) for *B. subtilis* from the *Oriloc* website. (The plot was taken from the *Oriloc* website, with permission). (**B**) GC-skew plot for *B. subtilis* from *CG-software* website. (The plot was taken from the *Comparative Genometrics* website, with permission). (**C**) GC-skew plot for *B. subtilis*, *GraphDNA* (the software was downloaded from the corresponding website).

with lengths up to 8 are counted separately on the leading and lagging strand. Then the weighted double Kullback–Leibler distance between the two frequency distributions is calculated as the sum over all such oligonucleotides. It measures the difference between oligonucleotide frequencies in the lagging and leading strands for a given candidate position. Finally, the program output coordinates of candidate origins and the corresponding Kullback–Leibler distances. For oligonucleotide skew both origin and terminus corresponds to maximuma.

From the user's point of view, it is somewhat inconvenient when a program does not provide the extremum coordinates, because the scale unit of the horizontal axis, reflecting the genome position, is 0.2 Mb through 1 Mb. The *GraphDNA* program provides a special tool that allows the user to determine the coordinate of any point of the plot, and it is useful when the plot contains sharp peaks. The *Oligonucleotide skew* site presents the origin coordinates with the precision of 1 kb, but only for 325 analyzed bacterial chromosomes. The users of the remaining programs, *Oriloc*, *CG-software* and *Z-curve*, have to resort to the analysis of the attached coordinate tables.

The computational tools considered above were based only on the global statistical measures of genomic DNA. However, other types of data may also be used to supplement this analysis. The distribution of DnaA-boxes and the location of the *dnaA* gene were used to improve the precision of mapping of the *oriC* sites [70]. These features were used in the *DoriC* database [73] and the *Ori-Finder* web server [74], created by the same group. The *DoriC* database (http://tubic.tju.edu.cn/doric/) currently contains 578 eubacterial genomes (632 chromosomes). The predictions were made based on compositional strand asymmetry (estimated using the *Z-curve* program), the distribution of DnaA-boxes (either of the *E. coli* type or species-specific), location of the indicator genes (*dnaA*, *hemE*, *gidA*, *dnaN*, *hemB*, *maf*, *repC*, *etc.*), and the *dif* sequences.

The *Ori-Finder* site (http://202.113.12.55/Ori-Finder/) contains a separate database of newly sequenced bacterial genomes. The program uses *Z-curve*, DnaA-boxes, indicator genes, takes into account phylogenetic relationships, but does not consider *dif* sites.

Most programs considered above were also used to predict putative replication origin and terminus sites in the archaea. Archaeal predictions were made by *Z-curve* [94]; *CG-software*, which produced numerous archaeal skew plots [83 and *Comparative genometrics* website]; *Oligonucleotide skew* method that also was used to predict whether an archaeal genome contains single or multiple replication origins [56]; and *Oriloc* (http://pbil.univ-lyon1.fr/software/Oriloc/index.html). The origin prediction is complicated by the fact that some archaeal chromosomes may have multiple replication origins [80]. However, several origins predicted by computer programs were proven experimentally [24, 94]. A recent list of experimentally confirmed origins in archaea includes *Archaeoglobus fulgidus* [96], *Halobacterium* NRC-1 [97], *Haloferax volcanii* [98], *P. abyssi* [37], *S. acidocaldarius* [80] and *S. solfataricus* [79, 80].

CLASSIFICATION OF GC-SKEW PREDICTION CURVES OF PROKARYOTES

We classified 486 eubacterial chromosomes from the *Comparative Genometrics* website collection of asymmetry curves [83] into several classes according to the overall shape of the plots (Supplementary List 1 and Table 2).

The ideal curve is Λ-shaped when the chromosomal sequence starts at the *oriC* site (Figure 2A). If the sequence starts at an arbitrary point of the chromosome, the corresponding curve has one minimum and one maximum (Figure 2B).

The ordinates of the points at the beginning and the end of the curve in the majority of cases are about the same. If this is not the case, it would mean either that the statistical properties of the replichores

Table 2: Distribution of prokaryotic chromosomes in the GC-curve classes

Class	The number of chromosomes
Eubacterial chromosomes	
1A	284
1B	56
2A	71
2B	39
3A	19
3B	17
Total	486
Archaeal chromosomes	
2B	22
3A	3
3B	20
Total	45

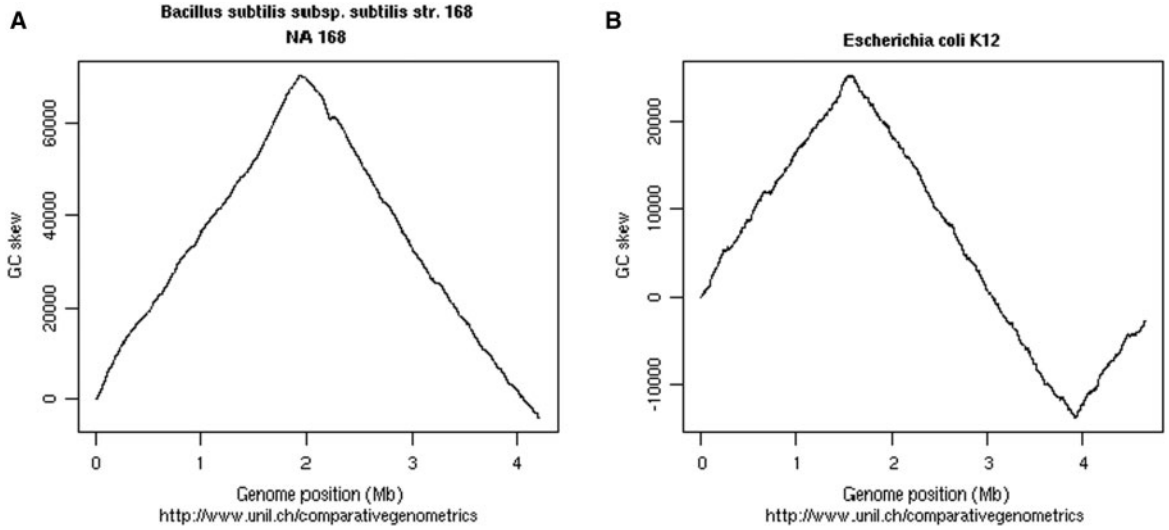

Figure 2: *GC*-skew curves for Class I genomes. (**A**) Class IA, Λ-shaped curves: *B. subtilis* subsp. *subtilis* str. 168. (**B**) Class IB, curves with a single minimum and single maximum: *E. coli* K12. The plots were taken from the *Comparative Genometrics* website, with permission.

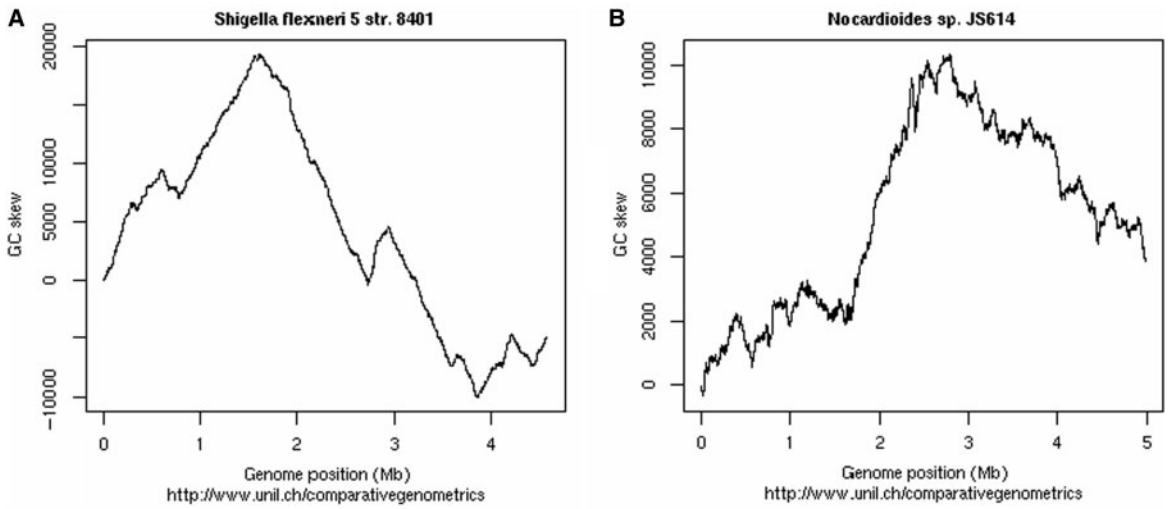

Figure 3: *GC*-skew curves for Class 2 genomes. (**A**) Class 2A, plots with a global minimum and several local extrema: *Shigella flexneri* 5 str. 8401. (**B**) Class 2B, plots with a global minimum and multiple local extrema: *Nocardioides* sp. JS614. The plots were taken from the *Comparative Genometrics* website, with permission.

are different (which is unlikely due to the inherent symmetry of the genome) or, more likely, that the replichores in the analyzed chromosome have different lengths. Visually, this manifests in *terC* being not directly opposite to *oriC* in the circular representation of the chromosome.

Genomic rearrangements, in particular, inversions, create a mosaic of ancestral leading and lagging strands and such genomes yield plots with multiple minima and maxima corresponding to the recombination sites.

The main criterion used to assign a genome to Class 2 is the presence of the global minimum, accompanied by additional local extrema, and, dependent on the intensity of recombinations, we subdivide Class 2 further into Class 2A with few rearrangements, expressed as several local extrema (Figure 3A) and Class 2B with multiple rearrangements, seen as multiple local extrema (Figure 3B).

Class 3 contains genomes without a clear global minimum in the GC-skew plots, and at the same time these plots have several positions with

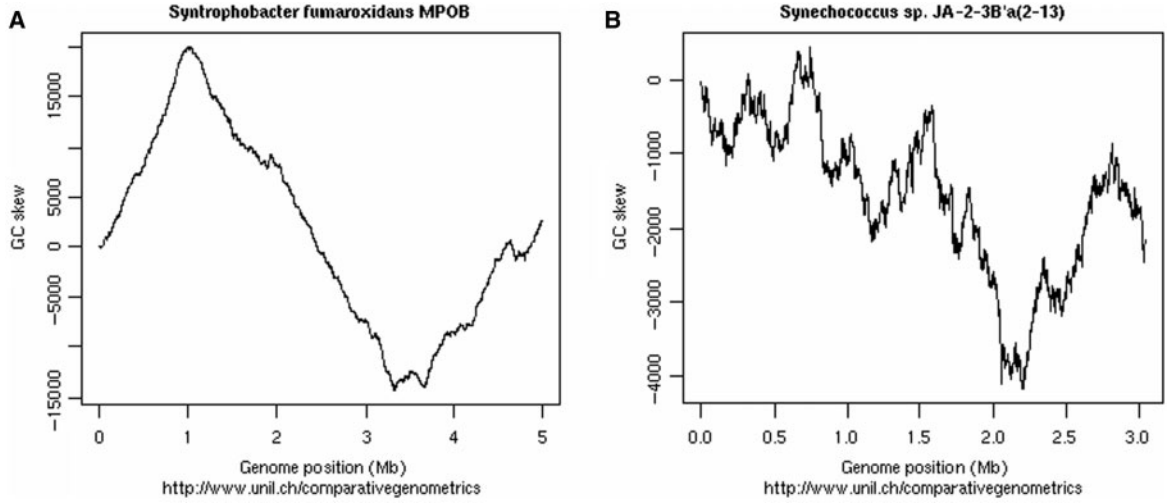

Figure 4: *GC*-skew curves for Class 3 genomes. (**A**) Class 3A, plots without a well-defined global minimum, but with several positions of approximately equal minimal values and several local extrema: *Syntrophobacter fumaroxidans* MPOB. (**B**) Class 3B, plots without a well-defined global minimum, but with several positions of approximately equal minimal values and multiple local extrema; *Synechococcus* sp. JA-2-3B'a(2–13). The plots were taken from the *Comparative Genometrics* website, with permission.

Table 3: The mean range of amplitude for GC-skew plots (plots of 486 genomes, presented by CG-software website, were used)

Class	1A	1B	2A	2B	3A	3B
Mean range of amplitude (kb)	50	37	24	12	22	8

approximately equal minimal values. At that time, the Class 3A *GC*-skew plots have several local extrema (Figure 4A), whereas in Class 3B, the *GC*-skew plots have multiple local extrema (Figure 4B).

Naturally, the plot amplitudes for the GC-skew depend on the class: the more rearrangements have happened the smaller are the amplitudes. The average amplitudes for different classes are presented in Table 3.

Notably, according to our classification, chromosomes of numerous close relatives (species or strain level) may be assigned either to the same class, or to different classes. As an example, bacterial chromosomes of the *Bacillus* or *Staphylococcus* species are assigned to Class 1A. At the same time, bacterial chromosomes from the *Burkholderia* genus are distributed among different classes. Most *Burkholderia* representatives belong to Class 1A, among them *Burkholderia mallei* ATCC 23344 chromosome 2 and *B. mallei* NCTC 10247 chromosome II. At the same time, *B. mallei* ATCC 23344 chromosome 1 and *B. mallei* NCTC 10247 chromosome I belong to Class 2A. Class 3A also contains representatives of this genus, *B. mallei* SAVP1 chromosome I and *B. mallei* NCTC 10229 chromosome I. The genus *Synechococcus* also is represented in all classes. *Synechococcus* sp. WH8102 belongs to Class 1A, *Synechococcus* sp. CC9311 and *Synechococcus* sp. CC9605 to class 2A, *Synechococcus* sp. JA-3-3Ab to Class 2B, *S. elongatus* PCC 6301 to Class 3A, and *Synechococcus* sp. JA-2-3B'a(2–13) to Class 3B.

Using the established eubacterial classification, we also classified the archaeal chromosomes. We assigned 45 archeal plots to different classes of our classification (Supplement 2, Table 2). All considered archaeal chromosomes belonged to Classes 2B, 3A and 3B.

The skew plots of eubacterial and archaeal chromosomes are rather similar. Considering plots assigned to class 2B, it is possible to suggest the presence of a single origin in these chromosomes, and indeed this has been proven experimentally for *P. abyssi* [37]. Plots of Classes 3A and 3B yield the presence of several origins in corresponding chromosomes, also proven experimentally [79]. The impossibility to distinguish between the trace of recent recombinations and the impact of multiple origins in the absence of closely related genomes has lead to the suggestion of existence of 'anomalous' eubacteria with multiple replication origins [23]. No such bacteria have been yet observed in experiment.

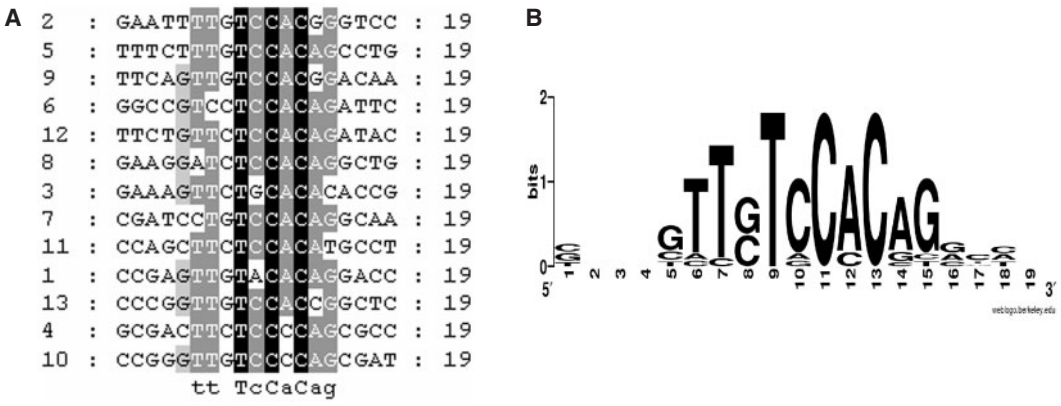

Figure 5: Candidate DnaA-boxes of *S. coelicolor*. (**A**) Alignment of candidate DnaA boxes identified in the *oriC* region. (**B**) Sequence logo of the DnaA-boxes.

Analysis of the *GC*-skew plots of the *Comparative Genometrics* website demonstrates that information given by the asymmetry curves sometimes is not sufficient for identification of the origin. Indeed, only Class 1 and probably, Class 2A plots yield reliable predictions. Class 2B chromosomes with their multiple chromosomal rearrangements are problematic, and Class 3 curves do not allow one to make a prediction.

Currenly, to predict the replication origins in archaea, procedures developed for eubacteria are used. There are too few experimentally defined archaeal origins to develop special computational procedures.

IDENTIFICATION OF DNAA BOXES

Since the replication initiation regions usually contain clusters of DnaA-boxes [70], this property can be used to identify the *oriC* sites. However, the DnaA-binding motifs may be taxon-specific, as observed in *M. tuberculosis* [67, 99] and *H. pylori* [66] and expected for *B. burgdorpheri* [71]. Thus the use of the standard *E. coli* consensus TTATNCACA may produce inconclusive results. To account for this possibility, the following combined approach may be used. For a taxon under analysis, the regions of the replication initiation may be defined for genomes from Class 1 and 2A using the *GC*-skew plots and location of the *dnaA* gene. Over-represented motifs identified in this region are likely DnaA-boxes. The positional weight matrix (PWM) may be derived from these boxes and used to find clusters of candidate DnaA-boxes in the remaining genomes from this taxon.

Two cautionary notes should be made at this point. First, in chromosomes of likely plasmid origin, such as chromosome II of *Vibrio cholerae*, the structure of the replication origin may be different from that of the *bona fide* bacterial chromosomes [100]. Second, clusters of the DnaA-boxes may occur not only at the replication origins, but also in other loci. For example, the *datA* locus of *E. coli* contains a cluster of the DnaA-boxes that are involved in the regulation of replication initiation, more exactly, control of over-initiation [101, 102].

As a proof of principle, we demonstrate the general applicability of this approach on a set of the *Streptomyces* species. This taxon contains a genome with an experimentally determined *oriC* site (*Streptomyces coelicolor*) [103] and a genome belonging to Class 3A of our classification (*S. avermilitis*). The replication origin of *S. coelicolor* contains more than 10 sites with the consensus TTSTCCACA (S = G or C) (Figure 5A). Nineteen putative DnaA-boxes for *S. lividans*, whose *oriC* sequence is identical to the *oriC* sequence of *S. coelicolor* [104], were suggested in [105]. We selected 13 sites with the length of 9 nt; all sites had at most two mismatches to the *E. coli* consensus TTATNCACA (the logo is presented in Figure 5B).

We constructed a PWM [89] using these sites and then used this PWM to scan the genome of *S. avermilitis*. A cluster of 14 candidate DnaA-boxes was identified (Figure 6A); the logo of these sites is presented in Figure 6B. There are two *dnaA* genes in *S. avermilitis*, and one is localized upstream of the identified cluster; this may be considered as independent collaboration, since we did not use the *dnaA* location at previous steps. The *GC*-skew plot is presented on Figure 6C, and the predicted *oriC*

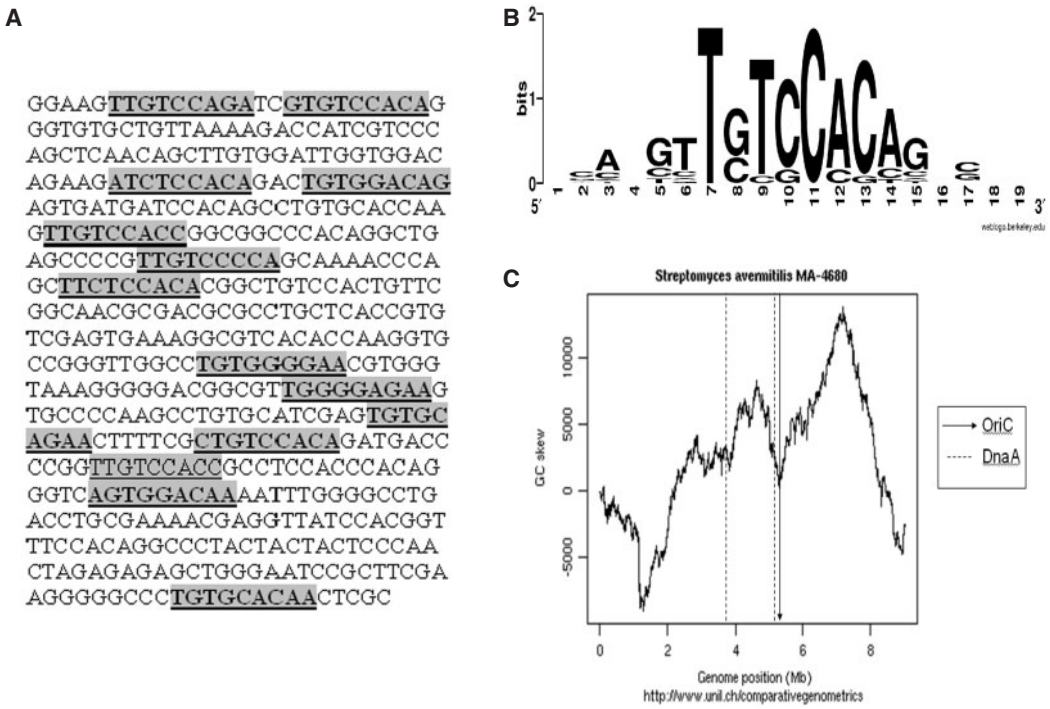

Figure 6: Candidate *oriC* region of *Streptomyces avermitilis*. (**A**) Candidate DnaA boxes identified in the *oriC* region. (**B**) Sequence logo of the DnaA-boxes. (**C**) *GC*-skew plot (Comparative Genometrics website, with permission). The arrow shows the locations of the predicted *oriC* region and the dotted lines—the location of two *dnaA* genes.

region coincides with the prediction in the *DoriC* database [73].

We further considered the *Synechococcus* genus as an example of a taxon with no experimentally studied replication origins. This taxon includes genomes from different classes in our classification. The candidate replication *oriC* regions could be identified by visual analysis of the *GC*-skew curves for *Synechococcus* sp. CC9311 (Class 2A) and *Synechococcus* sp. CC9605 (Class 2A). Analysis of over-represented oligonucleotides (searching for matrices with the maximal informational content) yielded six candidate DnaA-boxes in *Synechococcus* sp. CC9311 and five candidate DnaA-boxes in *Synechococcus* sp. CC9605. The logos are presented in Figure 7A (a,b).

Thus we identified a conserved, over-represented motif TTTTCCMCA (M = A or C) common to both genomes [Figure 7A (a,b)]. The constructed PWM was then used to identify clusters of candidate DnaA-boxes in *Synechococcus* sp. JA-2-3B'a(2–13) (http://www.genome.jp/dbget-bin/www_bget?refseq+NC_007775) belonging to Class 3A and *Synechococcus* sp. JA-3-3Ab (http://www.genome.jp/dbget-bin/www_bget?refseq+NC_007776) from Class 2B [Figure 7B (a,b) and C (a,b), respectively].

Again, the predicted origin locations coincided with the predictions of *DoriC*. In both cases, the search for clusters of potential DnaA-boxes was done in the whole GenBank sequence of corresponding chromosomes, not just in intergenic sequences (to account for possible genome annotation inaccuracies).

Note that the obtained motifs show some minor species-specific differences, but this has not influenced the result.

Finally, we considered the replication origins of the *Mycobacterium* species. The *oriC* locus, situated between the *dnaA* and *dnaN* genes, was determined experimentally for *M. tuberculosis*, as well as for *M. smegmatis* and *M. leprae* [99]. In an experimental paper [67], 13 previously proposed boxes [106] as well as 2 newly predicted ones, were tested on their ability to bind the DnaA protein by dimethylsulfate (DMS) footprinting and surface plasmon resonance (SPR). Binding of the DnaA protein was confirmed for 10 of them.

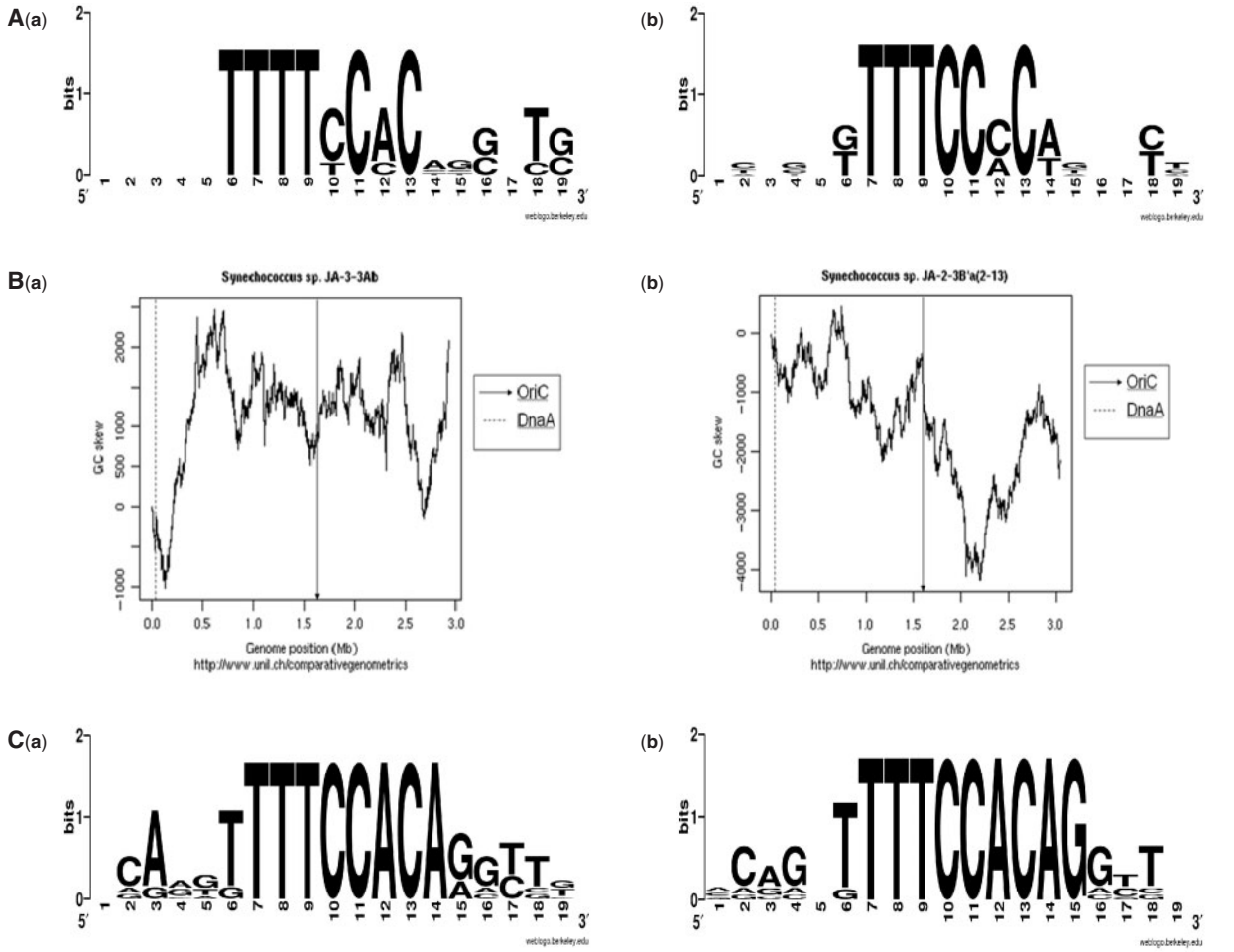

Figure 7: Candidate DnaA-boxes and *oriC* regions of Synechococcus spp. (**A**) Logo of candidate DnaA-boxes of *Synechococcus* sp. CC9311 (a), *Synechococcus* sp. CC9605 (b). (**B**) *GC*-skew plots of *Synechococcus* sp. JA-2-3B'a(2–13) (a), *Synechococcus* sp. JA-3-3Ab (b). (Comparative Genometrics website, with permission). The arrow shows the locations of the predicted *oriC* region and the dotted line—the location of *dnaA* gene. (**C**) Logo of candidate DnaA-boxes of *Synechococcus* sp. JA-2-3B'a(2–13) (a), *Synechococcus* sp. JA-3-3Ab (b).

In an attempt to identify candidate DnaA-sites for other mycobacteria, we aligned this region to the corresponding regions (defined by synteny) of other *Mycobacterium* species. The visual analysis of this alignment identified a cluster of conserved islands containing CACA motifs similar to the DnaA motifs from other species. It allowed us to select seven motifs of the type N_5-CACA-N_5. Five of seven *M. tuberculosis* boxes, derived from the alignment, coincided with experimentally determined sites [67] (Figure 8A). However, positions near this motif, while conserved between species, showed little conservation between individual islands (Figure 8B).

Thus, the exact span of the mycobacterial DnaA-motif remains unknown. Since the resolution of the footprinting experiments is not very high, further experimental analysis is needed to explain the discrepancies. However, they do not put into doubt the usefulness of such analysis, since both the general motif and the *oriC* location were identified correctly.

CONCLUSIONS

We considered six currently available *in silico* tools for the prediction of eubacterial replication origins and termini, and classified the skew-curves from one of the largest on-line collections (*Comparative Genometrics*). The majority of the curves are perfectly Λ-shaped, while the rest are disturbed by inter-genomic rearrangements of varying intensity. We observed clusters of putative DnaA-boxes,

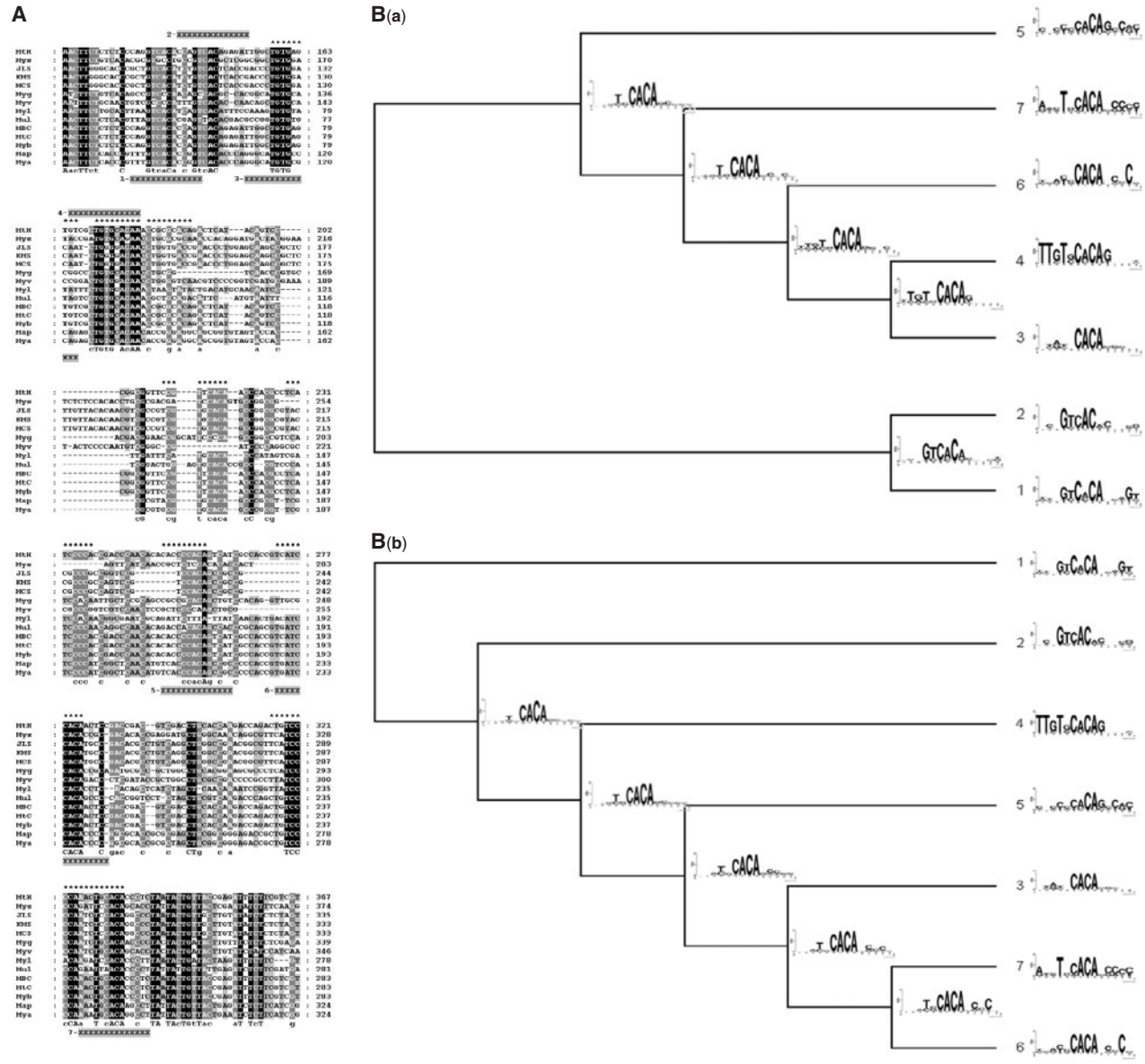

Figure 8: Candidate DnaA-boxes of *Mycobacterium* spp. (**A**) Alignment of the *dnaA–dnaN* intergenic regions with seven conserved motifs N_5-CACA-N_5, experimental (***) DnaA-boxes in the *oriC*-site of *M. tuberculosis* [67] and predicted (XXX) DnaA-boxes. (Only 9/10 experimental boxes are shown, the 10th one is located at a distance exceeding 250 nt from the region in the picture). (Genome abbreviations: Myv, Mycobacterium_vanbaalenii_PYR-1; Mul, Mycobacterium_ulcerans_Agy99; MtH, Mycobacterium_tuberculosis_H37Rv; MtC, Mycobacterium_tuberculosis_CDC1551; Mys, Mycobacterium_smegmatis_MC2_155; MCS, Mycobacterium_MCS; Myl, Mycobacterium_leprae; KMS, Mycobacterium_KMS; JLS, Mycobacterium_JLS; Myg, Mycobacterium_gilvum_PYR-GCK; MBC, Mycobacterium_bovis_BCG_Pasteur_1173P2; Myb, Mycobacterium_bovis; Map, Mycobacterium_avium_paratuberculosis; Mya, Mycobacterium_avium_104). (**B**) Cluster trees of individual motifs constructed by *ClusterTree-RS*. Sequence logos are shown for each node. The two trees correspond to different definitions of the DnaA-motif [N_5CACA (a) and CACAN_5 (b)].

common to the predicted *oriC* sites within separate taxonomic groups, but sometimes varying between groups. As a result we suggested an additional approach which may facilitate identification of the replication *oriC* in cases of ambiguous or 'fuzzy' skew-based predictions in eubacteria, independent of synteny or 'typical neighborhood' gene arrangements.

Key Points

- Existing software tools based on analysis of compositional asymmetry of chromosomes may predict origin of replication, but approximately and not in all cases.
- The use of additional features such as DnaA-boxes and location of indicator genes enhances resolution and allows for reliable predictions in complicated cases.
- Comparative genomics is a powerful approach to description of DnaA-box motifs that may have taxon-specific deviations from the common consensus.

Acknowledgements

We are grateful to the *Comparative Genometrics* team for the source-code of their program and the permission to use plots from their website as figures in this article. We are also grateful to Jean Lobry for the permission to use plots from the *Oriloc* website as figures in this article. We thank Andrey Mironov for his assistance with *SignalX* and Elena Stavrovskaya for the help with the program *ClusterTree-RS*. This study was partially supported by grants from the Howard Hughes Medical Institute (55005610), INTAS (05-8028), and the Russian Academy of Sciences (program 'Molecular and Cellular Biology').

SUPPLEMENTARY MATERIALS

Supplementary Materials are available at *Briefings in Bioinformatics* Online.

References

1. Egan ES, Fogel MA, Waldor MK. Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. *Mol Microbiol* 2005;**56**:1129–38.
2. Campbell AM. Genome organization in prokaryotes. *Curr Opin Genet Dev* 1993;**3**:837–44.
3. Kornberg A, Baker TA. *DNA Replication*. 2nd edn. New York: W.H. Freeman & Company, 1992.
4. Kelman LM, Kelman Z. Multiple origins of replication in archaea. *Trends Microbiol* 2004;**12**:399–401.
5. Boulikas T. Common structural features of replication origins in all life forms. *J Cell Biochem* 1996;**60**:297–316.
6. Hendrickson H, Lawrence JG. Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites. *Mol Microbiol* 2007;**64**:42–56.
7. Higgins NP. Mutational bias suggests that replication termination occurs near the *dif* site, not at *Ter* sites: what's the *Dif*? *Mol Microbiol* 2007;**64**:1–4.
8. Barry ER, Bell SD. DNA replication in the archaea. *Microbiol Mol Biol Rev* 2006;**70**:876–87.
9. Okazaki R, Okazaki T, Sakabe K, *et al*. In vivo mechanism of DNA chain growth. Cold. Spring Harbor Symp. *Quant Biol* 1968;**33**:129–43.
10. Chargaff E. Structure and function of nucleic acids as cell constituents. *Fed Proc* 1951;**10**:654–9.
11. Rudner R, Karkas JD, Chargaff E. Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. *Proc Natl Acad Sci USA* 1968;**60**:921–2.
12. Sueoka N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* 1995;**40**:318–25.
13. Lobry JR. Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol* 1995;40: 326–30; Erratum in: *J Mol Evol* 1995;**41**:680.
14. Frank AC, Lobry JR. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 1999;**238**:65–77.
15. Tillier ER, Collins RA. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* 2000a;**50**:249–57.
16. Rocha EP, Touchon M, Feil EJ. Similar compositional biases are caused by very different mutational effects. *Genome Res* 2006;**16**:1537–47.
17. Touchon M, Rocha EP. From GC skews to wavelets: a gentle guide to the analysis of compositional asymmetries in genomic data. *Biochimie* 2008;**90**:648–59.
18. Rocha EP, Danchin A. Ongoing evolution of strand composition in bacterial genomes. *Mol Biol Evol* 2001;**18**: 1789–99.
19. Rocha EP. The replication-related organization of bacterial genomes. *Microbiol* 2004;**150**:1609–27.
20. McLean MJ, Wolfe KH, Devine KM. Base composition skews, replication orientation and gene orientation in 12 prokaryote genomes. *J Mol Evol* 1998;**47**:691–696.
21. Rocha EPC, Danchin A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res* 2003a;**31**:6570–7.
22. Rocha EPC, Danchin A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* 2003b;**34**: 377–8.
23. Nikolaou C, Almirantis Y. A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. *Nucleic Acids Res* 2005;**33**:6816–22.
24. Necşulea A, Lobry JR. A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol Biol Evol* 2007;**24**:2169–79.
25. Kawai M, Nakao K, Uchiyama I, *et al*. How genomes rearrange: genome comparison within bacteria *Neisseria* suggests roles for mobile elements in formation of complex genome polymorphisms. *Gene* 2006;**383**:52–63.
26. Liu WQ, Liu GR, Li JQ, *et al*. Diverse genome structures of *Salmonella paratyphi* C. *BMC Genomics* 2007;**8**:290.
27. DeBoy RT, Mongodin EF, Emerson JB, *et al*. Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. *J Bacteriol* 2006; **188**:2364–74.
28. McLeod MP, Qin X, Karpathy SE, *et al*. Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. *J Bacteriol* 2004;**186**: 5842–55.
29. Tsoktouridis G, Merz CA, Manning SP, *et al*. Molecular characterization of *Brucella abortus* chromosome II recombination. *J Bacteriol* 2003;**185**:6130–6.
30. Bulach DM, Zuerner RL, Wilson P, *et al*. Genome reduction in *Leptospira borgpetersenii* reflects limited transmission potential. *Proc Natl Acad Sci USA* 2006;**103**: 14560–5.

31. Parkhill J, Wren BW, Thomson NR, *et al*. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* 2001;**413**:523–7.
32. Kelman LM, Kelman Z. Archaea: an archetype for replication initiation studies? *Mol Microbiol* 2003;**48**:605–15.
33. Eisen JA, Heidelberg JF, White O, *et al*. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol* 2000;**1**: research0011.1–0011.9.
34. Grigoriev A. Graphical genome comparison: rearrangements and replication origin of *Helicobacter pylori*. *Trends Genet* 2000;**16**:376–8.
35. Uno R, Nakayama Y, Arakawa K, *et al*. The orientation bias of Chi sequences is a general tendency of G-rich oligomers. *Gene* 2000;**259**:207–15.
36. Salzberg SL, Salzberg AJ, Kerlavage AR, *et al*. Skewed oligomers and origins of replication. *Gene* 1998;**217**:57–67.
37. Myllykallio H, Lopez P, López-García P, *et al*. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* 2000;**288**: 2212–5.
38. Lopez P, Forterre P, le Guyader H, *et al*. Origin of replication of Thermotoga maritima. *Trends Genet* 2000;**16**: 59–60.
39. Hendrickson H, Lawrence JG. Selection for chromosome architecture in bacteria. *J Mol Evol* 2006;**62**:615–29.
40. Arakawa K, Uno R, Nakayama Y, *et al*. Validating the significance of genomic properties of Chi sites from the distribution of all octamers in *Escherichia coli*. *Gene* 2007;**392**: 239–46.
41. El Karoui M, Schaeffer M, Biaudet V, *et al*. Orientation specificity of the *Lactococcus lactis* Chi site. *Genes Cells* 2000; **5**:453–61.
42. El Karoui M, Biaudet V, Schbath S, *et al*. Characteristics of Chi distribution on different bacterial genomes. *Res Microbiol* 1999;**150**:579–87.
43. Sourice S, Biaudet V, El Karoui M, *et al*. Identification of the Chi site of *Haemophilus influenzae* as several sequences related to the *Escherichia coli* Chi site. *Mol Microbiol* 1998;**27**: 1021–9.
44. Bell SJ, Chow YC, Ho JY, *et al*. Correlation of chi orientation with transcription indicates a fundamental relationship between recombination and transcription. *Gene* 1998;**216**:285–92; Erratum in: Gene 1999;231:213.
45. Bigot S, Saleh OA, Lesterlin C, *et al*. KOPS: DNA motifs that control E. coli chromosome segregation by orienting the FtsK translocase. *EMBO J* 2005;**24**:3770–80.
46. Levy O, Ptacin JL, Pease PJ, *et al*. Identification of oligonucleotide sequences that direct the movement of the Escherichia coli FtsK translocase. *Proc Natl Acad Sci USA* 2005;**102**:17618–23.
47. Lobry JR. Origin of replication of *Mycoplasma genitalium*. *Science* 1996a;**272**:745–6.
48. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* 1996b;**13**:660–5.
49. Kunst F, Ogasawara N, Moszer I, *et al*. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 1997;**390**:249–56.
50. Blattner FR, Plunkett G, III, Bloch CA, *et al*. The complete genome sequence of *Escherichia coli* K-12. *Science* 1997;**277**: 1453–74.
51. Freeman JM, Plasterer TN, Smith TF, *et al*. Patterns of genome organization in bacteria. *Science* 1998;**279**:1827a.
52. Lafay B, Lloyd AT, McLean MJ, *et al*. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res* 1999;**27**:1642–9.
53. Tillier ER, Collins RA. Replication orientation affects the rate and direction of bacterial gene evolution. *J Mol Evol* 2000b;**51**:459–63.
54. Mackiewicz P, Mackiewicz D, Kowalczuk M, *et al*. High divergence rate of sequences located on different DNA strands in closely related bacterial genomes. *J Appl Genet* 2003;**44**:561–84.
55. Lopez P, Philippe H, Myllykallio H, *et al*. Identification of putative chromosomal origins of replication in Archaea. *Mol Microbiol* 1999;**32**:883–6.
56. Worning P, Jensen LJ, Hallin PF, *et al*. Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* 2006;**8**:353–61.
57. Leonard AC, Grimwade JE. Building a bacterial orisome: emergence of new regulatory features for replication origin unwinding. *Mol Microbiol* 2005;**55**:978–85.
58. Zakrzewska-Czerwińska J, Jakimowicz D, Zawilak-Pawlik A, *et al*. Regulation of the initiation of chromosomal replication in bacteria. *FEMS Microbiol Rev* 2007;**31**:378–87.
59. Mott ML, Berger JM. DNA replication initiation: mechanisms and regulation in bacteria. *Nat Rev Microbiol* 2007;**5**: 343–54.
60. Messer W. The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication. *FEMS Micro Rev* 2002;**26**:355–74.
61. Kaguni JM. *Escherichia coli* DnaA protein: the replication initiator. *Mol Cells* 1997;**7**:145–57.
62. Christensen BB, Atlung T, Hansen FG. DnaA boxes are important elements in setting the initiation mass of *Escherichia coli*. *J Bacteriol* 1999;**181**:2683–8.
63. Kato J. Regulatory network of the initiation of chromosomal replication in *Escherichia coli*. *Crit Rev Biochem Mol Biol* 2005;**40**:331–42.
64. Weigel C, Schmidt A, Rückert B, *et al*. DnaA protein binding to individual DnaA boxes in the *Escherichia coli* replication origin, oriC. *EMBO J* 1997;**16**:6574–83.
65. Schaper S, Nardmann J, Lüder G, *et al*. Identification of the chromosomal replication origin from *Thermus thermophilus* and its interaction with the replication initiator DnaA. *J Mol Biol* 2000;**299**:655–65.
66. Zawilak A, Durrant MC, Jakimowicz P, *et al*. DNA binding specificity of the replication initiator protein, DnaA from Helicobacter pylori. *J Mol Biol* 2003;**334**: 933–47.
67. Madiraju MV, Moomey M, Neuenschwander PF, *et al*. The intrinsic ATPase activity of *Mycobacterium tuberculosis* DnaA promotes rapid oligomerization of DnaA on oriC. *Mol Microbiol* 2006;**59**:1876–90.
68. Ishikawa S, Ogura Y, Yoshimura M, *et al*. Distribution of stable DnaA-binding sites on the *Bacillus subtilis* genome detected using a modified ChIP-chip method. *DNA Res* 2007;**14**:155–68.
69. Fujikawa N, Kurumizaka H, Nureki O, *et al*. Structural basis of replication origin recognition by the DnaA protein. *Nucleic Acids Res* 2003;**31**:2077–86.

70. Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, *et al*. Where does bacterial replication start? Rules for predicting the *oriC* region. *Nucleic Acids Res* 2004;**32**:3781–91.
71. Picardeau M, Lobry JR, Hinnebusch BJ. Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome. *Mol Microbiol* 1999;**32**: 437–45.
72. Gil R, Silva FJ, Zientz E, *et al*. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc Natl Acad Sci USA* 2003;**100**:9388–93.
73. Gao F, Zhang CT. DoriC: a database of oriC regions in bacterial genomes. *Bioinformatics* 2007;**23**:1866–7.
74. Gao F, Zhang CT. Ori-Finder: a web-based system for finding *oriC*s in unannotated bacterial genomes. *BMC Bioinformatics* 2008;**9**:79.
75. Tobiason DM, Seifert HS. The obligate human pathogen, Neisseria gonorrhoeae, is polyploid. *PLoS Biol* 2006;**4**:e185.
76. Suhan M, Chen S-Y, Thompson HA, *et al*. Cloning and characterization of an autonomous replication sequence from *Coxiella burnetii*. *J Bacteriol* 1994;**176**:5233–43.
77. Marczynski GT, Shapiro L. Cell-cycle control of a cloned chromosomal origin of replication from *Caulobacter crescentus*. *J Mol Biol* 1992;**226**:959–77.
78. Neylon C, Kralicek AV, Hill TM, *et al*. Replication termination in *Escherichia coli*: structure and antihelicase activity of the Tus-Ter complex. *Microbiol Mol Biol Rev* 2005;**69**:501–26.
79. Robinson NP, Dionne I, Lundgren M, *et al*. Identification of two origins of replication in the single chromosome of the archaeon Sulfolobus solfataricus. *Cell* 2004;**116**:25–38.
80. Lundgren M, Andersson A, Chen L, *et al*. Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination. *Proc Natl Acad Sci USA* 2004;**101**:7046–51.
81. Grigoriev A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* 1998;**26**:2286–90.
82. Frank CA, Lobry JR. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* 2000;**16**:560–1.
83. Roten C-AH, Gamba P, Barblan J-L, *et al*. Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. *Nucleic Acids Res* 2002;**30**: 142–4.
84. Thomas JM, Horspool D, Brown G, *et al*. GraphDNA: a Java program for graphical display of DNA composition analyses. *BMC Bioinformatics* 2007;**23**:21.
85. Zhang CT, Zhang R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res* 1991;**19**:6313–7.
86. Zhang R, Zhang CT. Z curves, an intuitive tool for visualizing and analyzing DNA sequences. *J Biomol Struc Dynamics* 1994;**11**:767–82.
87. Benson DA, Karsch-Mizrachi I, Lipman DJ, *et al*. GenBank. *Nucleic Acids Res* 2007;**35**:21–5.
88. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**:1792–97.
89. Gelfand MS. Recognition of regulatory sites by genomic comparison. *Res Microbiol* 1999;**150**:755–71.
90. Crooks GE, Hon G, Chandonia JM, *et al*. WebLogo: a sequence logo generator. *Genome Res* 2004;**14**:1188–90.
91. Stavrovskaia ED, Makeev VIu, Mironov AA. ClusterTree-RS: the binary tree algorithm for identification of co-regulated genes by clustering regulatory signals. *Mol Biol (Mosk)* 2006;**40**:524–32.
92. Perrière G, Gouy M. WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie* 1996;**78**: 364–9.
93. Lobry JR. Genomic landscapes. *Microbiol Today* 1999;**26**: 164–5.
94. Zhang R, Zhang CT. Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea* 2005;**1**:335–46.
95. Song J, Ware A, Liu SL. Wavelet to predict bacterial ori and ter: a tendency towards a physical balance. *BMC Genomics* 2003;**4**:17.
96. Maisnier-Patin S, Malandrin L, Birkeland NK, *et al*. Chromosome replication patterns in the hyperthermophilic euryarchaea Archaeoglobus fulgidus and Methanocaldococcus (Methanococcus) jannaschii. *Mol Microbiol* 2002;**45**: 1443–50.
97. Berquist BR, DasSarma S. An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, Halobacterium sp. strain NRC-1. *J Bacteriol* 2003;**185**:5959–66.
98. Norais C, Hawkins M, Hartman AL, *et al*. Genetic and physical mapping of DNA replication origins in Haloferax volcanii. *PLoS Genet* 2007;**3**:e77.
99. Salazar L, Fsihi H, de Rossi E, *et al*. Organization of the origins of replication of the chromosomes of Mycobacterium smegmatis, *Mycobacterium leprae* and *Mycobacterium tuberculosis* and isolation of a functional origin from *M. smegmatis*. *Mol Microbiol* 1996;**20**:283–93.
100. Egan ES, Waldor MK. Distinct replication requirements for the two Vibrio cholerae chromosomes. *Cell* 2003;**114**: 521–30.
101. Boye E, Løbner-Olesen A, Skarstad K. Limiting DNA replication to once and only once. *EMBO Rep* 2000;**1**: 479–483.
102. Ogawa T, Yamada Y, Kuroda T, *et al*. The datA locus predominantly contributes to the initiator titration mechanism in the control of replication initiation in *Escherichia coli*. *Mol Microbiol* 2002;**44**:1367–75.
103. Calcutt MJ, Schmidt FJ. Conserved gene arrangement in the origin region of the *Streptomyces coelicolor* chromosome. *J Bacteriol* 1992;**174**:3220–6.
104. Zakrzewska-Czerwińska J, Majka J, Schrempf H. Minimal requirements of the Streptomyces lividans 66 oriC region and its transcriptional and translational activities. *J Bacteriol* 1995;**177**:4765–71.
105. Jakimowicz D, Majka J, Messer W, *et al*. Structural elements of the *Streptomyces* oriC region and their interactions with the DnaA protein. *Microbiol* 1998;**144**: 1281–90.
106. Dziadek J, Rajagopalan M, Parish T, *et al*. Mutations in the CCGTTCACA DnaA box of *Mycobacterium tuberculosis* oriC that abolish replication of oriC plasmids are tolerated on the chromosome. *J Bacteriol* 2002;**184**: 3848–55.