ELSEVIER

Gene 217 (1998) 57–67

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

Skewed oligomers and origins of replication

Steven L. Salzberg [a,b,*], Alan J. Salzberg [c], Anthony R. Kerlavage [a], Jean-Francois Tomb [a,1]

[a] *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA*
[b] *Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA*
[c] *KPMG, 2001 M St. NW, Washington, DC 20036, USA*

Received 16 April 1998; received in revised form 9 July 1998; accepted 9 July 1998

Abstract

The putative origin of replication in prokaryotic genomes can be located by a new method that finds short oligomers whose orientation is preferentially skewed around the origin. The skewed oligomer method is shown to work for all bacterial genomes and one of three archaeal genomes sequenced to date, confirming known or predicted origins in most cases and in three cases (*H. pylori*, *M. thermoautotrophicum*, and *Synechocystis* sp.), suggesting origins that were previously unknown. In many cases, the presence of conserved genes and nucleotide motifs confirms the predictions. An algorithm for finding these skewed seven-base and eight-base sequences is described, along with a method for combining evidence from multiple skewed oligomers to accurately locate the replication origin. Possible explanations for the phenomenon of skewed oligomers are discussed. Explanations are presented for why some bacterial genomes contain hundreds of highly skewed oligomers, whereas others contain only a handful.

Keywords: Comparative genomics; Sequence analysis; Computational methods

1. Introduction

Bacteria replicate their genomes beginning at a single origin, and proceeding in opposite directions from that origin (Marczynski and Shapiro, 1993) until termination signals are encountered (Baker, 1995) and the complete genome has been copied. In contrast, eukaryotes have multiple origins of replication, which makes it possible to copy vastly longer genomes. It is not known whether members of the Archaea follow the bacterial or eukaryotic model of replication, but evidence given in this paper suggests that at least some of them may utilize a single origin.

This paper describes a new sequence-based method that finds short oligomers that are highly skewed on opposite strands of the chromosome. The point around which these oligomers are skewed is, in every case where evidence is available, the origin of replication. In recent papers on the complete genomic sequences of *E. coli* (Blattner et al., 1997), *B. burgdorferi* (Fraser et al., 1997), and *T. pallidum* (Fraser et al., 1998), octamers were described for each genome that showed a highly skewed orientation around the origin of replication. (The latter two references used the program described in this paper to find these octamers.) By 'skewed orientation' here, we mean that an octamer occurs much more often on the leading strand in the direction of replication than it does on the lagging strand. In all cases examined thus far, these octamers are significantly over-represented as well, i.e. they occur much more often than would be expected based on the underlying frequencies of the bases comprising them. Thus, there are two phenomena that together seem to typify these oligomers: skewed orientation and over-representation.

Examination of 13 completely sequenced prokaryotic genomes reveals that they fall into three distinct cate-

* Corresponding author. Tel: +1 301 315 2537; Fax: +1 301 838 0208; e-mail: salzberg@tigr.org
[1] Present address: DuPont Central Research and Development, Experimental station, E328-240, Wilmington, DE 19880, USA.

Abbreviations: A, adenine; C, cytosine; Cdc6, cell division control protein 6; dnaA, chromosome replication initiator gene; dnaN, DNA polymerase III beta gene; G, guanine; *gidA*, glucose-inhibited division gene A; *gidB*, glucose-inhibited division gene B; *gyrA*, DNA gyrase subunit A gene; *gyrB*, DNA gyrase subunit B gene; *imp*, inner membrane protein gene; *jag*, SpoIIIJ-associated gene; kb, kilobases; Mb, megabases; R, purine; *recF*, recombination gene F; *rmpH*, ribosomal gene L34; *rnpA*, ribonuclease P gene; *soj*, Spo0J regulator gene; *spo0J*, stage 0 sporulation gene; T, thymine; Y, pyrimidine.

0378-1119/98/$19.00
PII: S0378-1119(98)00374-6

gories: (1) genomes in which numerous skewed oligomers (dozens or hundreds) very clearly indicate the origin of replication, (2) genomes in which very few (less than 10) short oligomers show any substantial skew, and (3) genomes with no detectable skew. Although the underlying mechanisms for the observed oligomer skew cannot be unambiguously determined, it has been observed that both replication and transcription (Francino et al., 1996; Lobry, 1996a) introduce compositional bias in the two strands of DNA.

2. Materials and methods

The data include 13 complete prokaryotic genomes: *H. influenzae* (1.83 Mb), *M. pneumoniae* (816 kb), *M. jannaschii* (1.67 Mb), *M. genitalium* (580 kb), *H. pylori* (1.67 Mb), *A. fulgidis* (2.18 Mb), *E. coli* (4.6 Mb), *B. subtilis* (4.2 Mb), *Synechocystis* sp. (3.57 Mb), *B. burgdorferi* (910 kb), *M. thermoautotrophicum* (1.75 Mb), *T. pallidum* (1.14 Mb), and *M. tuberculosis* (4.30 Mb).

2.1. Algorithm for finding oligomers with skewed distributions

The skew ratio is expressed as the percentage of an oligomer that occurs in one-half the genome, computed for both strands. For any oligomer, its reverse complement will clearly have the identical skew ratio. The algorithm for finding skewed oligomers considers all octamers and all locations in a genome as possible origins. (Although the algorithm described here finds octamers, the program can look for short oligomers of any reasonable length. Octamers were chosen over 9-mers and longer oligomers because these longer oligomers sometimes occur too infrequently for the statistical methods to detect a skewed distribution, even where one exists.) After identifying all skewed oligomers, the algorithm computes the probability of the amount of skew observed in each case.

For a genome of length G, the algorithm first counts, in one pass through the sequence, how many times each octamer occurs in the genome. At the same time, it also counts how often each octamer occurs in positions 1 through $G/2$. It then proceeds from position $G/2$ to G, and at each position i, it increments one counter and decrements another. The counter to be incremented represents the octamer beginning at position i, and the decrement is for the octamer at position $i-G/2$. (Counters for the reverse complementary octamer are adjusted in the opposite direction.) The program simply keeps track of both the maximum and minimum values observed for each octamer during this process. The running time is proportional to G.

Note that for circular genomes, the region of maximal skew for an octamer may 'wrap around' the beginning and end of the genome. However, in that case, the minimal skew will be within the interval from 1 to G. Thus, the algorithm will always find either the minimal or the maximal skewed region.

2.2. Statistical methods

Because of the large number of oligomers being considered by the algorithm (4^8 octamers), some may be highly skewed just by chance around points other than the origin of replication. We used statistical methods to distinguish between oligomers whose skewed distribution is attributable to chance and those where the distribution is so highly skewed that chance can be ruled out.

The basic question that we want to answer is: what are the chances that the observed skew could be found in a sequence in which the oligomers were not skewed around the origin of replication? A naïve estimate of this probability can be obtained by focusing on a single oligomer and determining the probability that the skew found in the distribution of that oligomer could occur by chance. For an oligomer that occurs N times, the probability of observing K oligomers in either direction is given by:

$$P_1 = 2 \times \binom{N}{K} \left(\frac{1}{2}\right)^N.$$

Because we are actually checking all 4^8 octamers, we need to correct P_1 to take this into account.

Treating the skew of each octamer as being independent of the skew of other octamers (although they clearly are not, since many octamers overlap many others) gives a conservative estimate of the adjusted probability, i.e. it gives a higher probability than the true probability. Because each oligomer is treated identically to its reverse complement, we really only have 2^{15} octamers to consider. We therefore adjusted P_1 by:

$$P_2 = 1-(1-P_1)^{2^{15}}.$$

This gives us a conservative estimate of the probability that at least one oligomer out of 2^{15} would be as skewed as that being considered.

Finally, we can correct for the fact that many different possible genomic locations are being considered by the algorithm. Obviously, the skew method has limits on its resolution, i.e. it cannot distinguish locations that are too close together. If we estimate that the method allows one to distinguish locations that are separated by R base pairs, then we can adjust the significance calculation by:

$$P_{\text{corr}} = 1-(1-P_1)^{2^{15}(G/R)}.$$

This gives us a final, conservative estimate of the prob-

ability that a particular oligomer skew would be observed.

2.3. Combination of evidence from multiple oligomers

Using this algorithm, seven genomes are found to contain numerous oligomers with significantly skewed distributions. The exact location of maximal skew varies from one oligomer to another, but by combining the evidence from different oligomers, one can obtain a more precise indication of the location of the origin. We used a likelihood ratio approach to combine evidence from different oligomers and determine the statistically most likely location for the origin of replication. This works to refine the estimates of the origin to within 1–2 kb for genomes with many highly skewed oligomers such as *B. burgdorferi*. We can also use the combining algorithm for genomes in which only a few oligomers are significantly skewed, as is the case in *H. influenzae*. By combining evidence from the 10 most skewed octamers, we found the point of maximum likelihood to be around 615 000, compared to the previously reported origin of replication at 603 000. The individual octamers indicated an origin varying from 502 000 to 726 000, with the average at 567 000. Thus, the combining method was much more accurate than simple averaging. As expected, the method is not as precise when the evidence is weaker, but it still can be useful in narrowing the search for the origin.

Suppose we wish to compare position A and position B on a genome in order to determine which position is the origin. We have the positions of all occurrences of the oligomer O. Assuming that A is the origin of replication, let s be the number of times that O occurs in the direction of replication, $\tilde{s}$ the number of times it occurs in the opposing direction, and $n=s+\tilde{s}$. If we assume that B is the origin, then as many as s oligomers that occurred in the direction of replication relative to A may switch directions. Let s_A be the actual number that switch, $\tilde{s}_A$ the number that switch the opposite way, and $n_A=s_A+\tilde{s}_A$.

Given that position A is the origin, the relative number observed to switch in each direction is random and based only upon the number that are available to switch. The probabilities associated with the number of oligomers that switch in each direction can be modeled by a hypergeometric distribution:

$$P(S=s_A)=\binom{s}{s_A}\binom{\tilde{s}}{\tilde{s}_A}\Big/\binom{n}{n_A}.$$

Call the hypergeometric probability using position A $p(\mathrm{O,A,B})$; that is, the probability of the observed oligomer locations given that A is the true origin and we are comparing it to position B. Considering just one oligomer, then position A should be preferred over B if

$$\frac{p(\mathrm{O,A,B})}{p(\mathrm{O,B,A})}>1.$$

Adding multiple oligomers to this formulation is straightforward; we can simply take the product of their likelihoods. In order to determine the statistically most likely position for the origin, we exhaustively consider all pairs of positions.

In genomes in which many highly skewed oligomers are found, this combining algorithm allows one to pinpoint the origin more precisely (as would be expected) than in genomes with very few skewed oligomers, such as *H. influenzae* and *H. pylori*. By comparing predictions of the skew method for genomes in which the origin is known precisely (from experimental evidence), we provide a validation of the method. In the two cases for which experimental evidence is available, *E. coli* and *B. subtilis*, the skew method agrees with the evidence. For other genomes, the results reported here provide new clues (or further support, in those cases where the origin has been tentatively identified based on characteristic genes or nucleotide motifs) for where the origin should be.

3. Results and discussion

3.1. Oligomer skew

Our analysis revealed that many bacterial genomes contain a large number of highly skewed oligomers, oriented strongly around two points. (Note: the shorthand 'skewed oligomer' is used to mean an oligomer whose occurrences have a skewed distribution.) In the three genomes for which experimental evidence is available [*E. coli*, *B. subtilis*, and *M. tuberculosis* (Marsh and Worcel, 1977; Ogasawara et al., 1984, Salazar et al., 1996)], these skew points correspond quite precisely to the origin and terminus of replication. Among completely sequenced genomes, the most striking examples of skewed oligomers are found in *B. burgdorferi* (Fraser et al., 1997), *T. pallidum* (Fraser et al., 1998), *E. coli* (Blattner et al., 1997), and *B. subtilis* (Kunst et al., 1997). Each of these genomes has hundreds of different octamers (as well as numerous 7-mers, 9-mers, and longer oligomers) all showing a statistically significant skew. In all of these genomes, the multiple skewed oligomers agree on the location of the origin of replication, making it easy to infer a putative origin from these data. One example from each of these genomes is shown in Fig. 1.

As is clear in the figure, the skew occurs with respect to two points on a circular genome, the origin and the terminus. Because the skew algorithm cannot distinguish between these two points, additional evidence such as

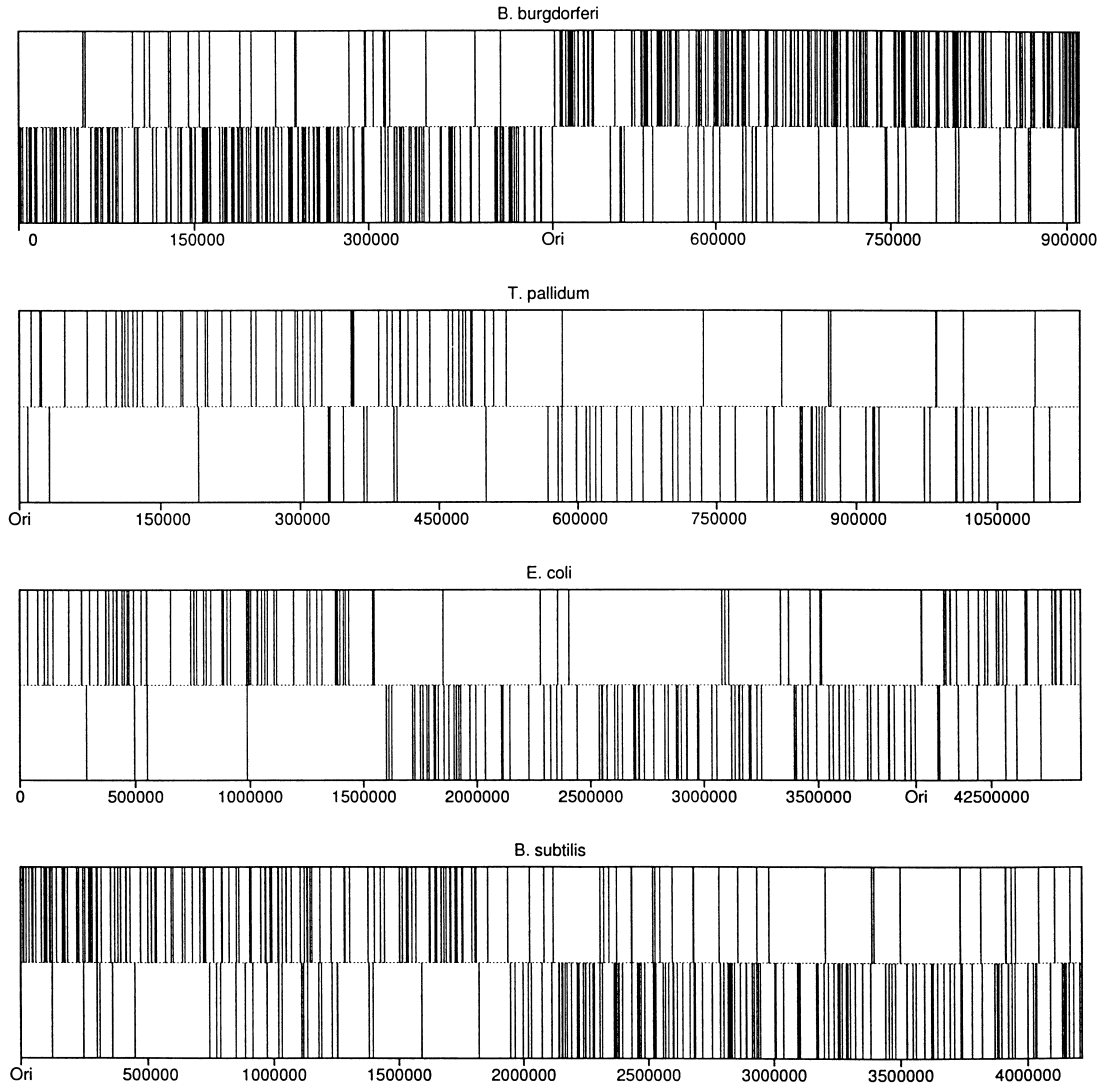

Fig. 1. Examples of skewed oligomers for four genomes: (**A**) tttgtttt in *B. burgdorferi*, (**B**) tgtgtgtg in *T. pallidum*, (**C**) gagcaggg in *E. coli*, and (**D**) tgaagagg in *B. subtilis*. In each graph, vertical lines above the center indicate positions of the oligomer on one DNA strand, and lines below the center indicate positions on the complementary strand. The origins of replication are indicated by a tick mark labeled 'Ori' in each plot. For *B. subtilis* and *E. coli*, the origin's location is known from experimental evidence, whereas for *T. pallidum* and *B. burgdorferi*, the origin has been putatively assigned based on the location of characteristic genes and the oligomer skew described here. Note that the oligomers shown in the figure are illustrative, and by no means represent the only highly skewed oligomers for these genomes.

the presence of the *dnaA* gene or *dnaA*-binding boxes (see Fig. 3) must be used to determine which is the origin. Every completely sequenced bacterial genome to date has at least one skewed oligomer that can be used to locate the origin, but some genomes have very few such oligomers. Table 1 contains examples of skewed oligomers from all bacterial genomes, showing where the origin is in each case and also giving the probability of observing the amount of skew in each case. To the human eye, there may appear to be some clustering of oligomers in the graphs in Fig. 1, suggesting that the distributions of oligomers are not uniform across the leading strand. We ran a test to determine whether or not these distributions differed significantly from the uniform distribution. In particular, we performed Kolmogorov–Smirnov tests (Rohatgi, 1976) on the distributions of the four oligomers plotted in Fig. 1, and obtained p-values of 0.44, 0.37, 0.91, and 0.57. This suggests that there is a close correspondence to the uniform distribution in all four graphs.

It is noteworthy that two of the three completely sequenced archaeal genomes do not appear to contain any oligomers skewed around a single point. [No skewed oligomers were found in an exhaustive search of the genomes of *M. jannaschii* (Bult et al., 1996) and *A. fulgidis* (Klenk et al., 1997).] However, the archaeon

Table 1
Skewed oligomers identifying the origin of replication for 11 prokaryotic genomes

Organism	Number of skewed 8-mers[a]	Example oligomer[b]	Counts[c]		Probability[d]	Origin[e]
			Skewed	Total		
B. burgdorferi	780	tttgtttt	456	520	1.9e−67	457 000
T. pallidum	500	ggagcgtg	120	140	5.1e−13	1
B. subtilis	1160	tgaaaaag	1031	1420	1.7e−58	1
E. coli	150	gctggtgg	763	1009	5.9e−55	3 923 000
M. pneumoniae	140	ctttgatg	92	107	2.9e−09	208 000
M. genitalium	110	ttgatgaa	142	173	5.8e−13	1
M. tuberculosis	60	tgggggag	118	129	1.2e−17	1
H. inflenzae	20	gtttggca	83	115	0.068	603 000
H. pylori	18	rrtagggg	302	368	4.9e−29	1 590 000
M. thermoautotrophicum	17	acttgagg	161	233	0.0030	1 275 000
Synechocystis sp.	1	tcggtcaa	106	157	N.S.	1 310 000

[a]The number of octamers whose distributions are significantly skewed around the predicted or actual origin.
[b]The Example column shows the most significantly skewed oligomer for each genome.
[c]The Counts columns show the number of occurrences of that oligomer on both strands and the number that occur in the direction of replication (Skewed).
[d]The probability score is the corrected probability, P_{corr}, for that oligomer as described in Section 2.2.
[e]The value given in the Origin column corresponds to either (1) the origin known from experimental evidence (for *B. subtilis* and *E. coli*), (2) the first reported prediction of an origin, based on oligomer skew and the other evidence discussed in this study (for *H. pylori*, *M. thermoautotrophicum*, and *Synechocystis* sp.) or (3) the putative origin based on characteristic genes, dnaA boxes, or GC skew (all the remaining genomes). For all genomes whose origin was previously reported, the oligomer skew corresponds to, and confirms, those reports. The *M. tuberculosis* skew results are based on TIGR's unpublished version of the genome [R.D. Fleischmann et al. (TIGR), pers. commun.].

M. thermoautotrophicum (Smith et al., 1997) does have skewed oligomers, implying a single replication origin. The absence of skewed oligomers in the other archaea suggests that either they utilize multiple origins or that their replication machinery does not require skewed oligomers.

Previous reports have observed that a related phenomenon, GC skew, can also be used to locate the origin of replication (Lobry, 1996a,b). GC skew is captured in the formula $(G-C)/(G+C)$, which, when computed in fixed-length windows across the genome, shows a switch in polarity at the origin and terminus. More recently, several similar, but related, measures have been proposed for locating origins, including GT skew and purine skew (Freeman et al., 1998). Several plausible explanations have been offered (Beletskii and Bhagwat, 1996; Francino et al., 1996; Lobry, 1996b; Francino and Ochman, 1997) for GC skew, and these same explanations should be considered here. The processes of both replication and transcription differ in their effects on the two strands of DNA. During replication, the lagging strand undergoes mutations at a different rate from the leading strand due to the differences between continuous and discontinuous replication. During transcription, the non-transcribed strand becomes single-stranded and exposed to DNA damage. In particular, C deaminates to T 100 times faster on single-stranded, versus double-stranded, DNA (Beletskii and Bhagwat, 1996; Francino et al., 1996). Both of these phenomena can explain GC skew, and both may be operative in producing the observed GC skew in prokaryotic genomes. However, as explained in the conclusion, these phenomena are not sufficient to explain oligomer skew.

3.2. Similarities among skewed oligomers

Although there is no single motif common across the genomes, there are clear motifs that can be found within each genome, with the exception of *H. influenzae* and *Synechocystis* sp., which have too few skewed oligomers to produce clear motifs. Fig. 2 shows some of the common motifs from the other nine genomes. Tables containing these motifs, as well as hundreds of oligomers from the various genomes in this study, and statistics on each oligomer will be available online at http://www.tigr.org.

For *B. burgdorferi*, there are over 700 octamers showing a highly skewed orientation around the origin. The most distinctive motif, showing the greatest degree of skew, is the set of octamers tttgtttt, ttttgttt, ttgttttt, and tttttgtt. The 12-mer ttttttgttttt, which contains each of these octamers as a subsequence, also shows a very strong skew. An examination of the locations of all skewed octamers reveals that approximately 89% of them occur in coding regions, which corresponds to the overall percentage of coding DNA. Thus, there is no clear preference of the skewed oligomers for coding or non-coding regions. (A similar pattern was found in other genomes.)

To test this further, we took the 10 octamers with the most skewed distributions in *B. burgdorferi* and retained only those occurrences that fell between coding regions.

B. subtilis		E. coli			M. genitalium	M. pneumoniae		
1	2	1	2	3	1	1	2	3
GAAAAAAG	TTGATGAA	CGAGCAGG	TGAAAGGG	CGCTGGTG	TATTGATG	TTTACTGA	TGAAAAGG	TTTTGTAA
GAAAAAGA	TGATGAAG	GGGGCAGG	GAAGAGGG	TGTCGGTG	TGTTGATG	ACATTGAT	GAAAAGGA	GTTGTTCT
GAAGAAAA	TGATGAAG	AGGGCAGG	GAAAGGGG	GCTGGTGG	TAGTGATG	CTTTGATG	CAAAAGGA	TTGTAAGC
GGAAAAAA	TGATGAAA	CAGCAGGG	GGAAGGGC	TGGTGGCG	AGTGATGA	CATTGATG	CAAGGAAA	TTGTCCTT
GGAAAAAG	GATGAAGA	GAGCAGGG	AAGAGGGC		ATTGATGA	TGGTGATG		TTGTTCTA
GAAAAAGG	ATGAAGTG	GGGCAGGG	AAAGGGCG		TTTGATGA	ATTGATGA		
GAAGAAGA	ATGAAGAA	GGCAGGGG	GAAGGGGA		TTGATGAA	TTAGATGA		
AAAGAAGA	ATGAAAAA	GGCAGGGC	AGAGGGCG		TTGATGGT	TTGATGAA		
AAAAAGAA	TTGAAAAA	AGCAGGGC	AAGGGGAG		GTGATGAA	TGATGAAA		
AAAAGAAG	TGAAAAAG	GCAGGGCA			TGATGAAG	TTGATGGT		
AAAAGAAA	TGAAGAAA	GCAGGGCG			TGATGTTG	TTGAGGTT		
GAAAAAGC	TGAAGAGG				TGATGAAA			
	TGAAGAAG				GATGAAAA			
	TGAAAAAA				GATGAGAA			
Motif	Motif	Motif	Motif	Motif	Motif	Motif	Motif	Motif
RRARRaRR	GaTGAARaRR	gRGCAGGGc	RRRAGGGcR	TGGT	tTGATGaa	tTGAtga	aAAGGa	TTGT

B. burgdorferi	T. pallidum		H. pylori	M. thermoautotrophicum		M. tuberculosis		
1	1	2	1	1	2	1	2	
TTTTTGTT	GCGTGTGT	GGAGCGTG	AATAGGGG	TGCAGGGG	TTGACTTC	GGGTGGGG	GTTGGTGG	
TTTTGTTT	TGTGTGTG	TGCGCGTG	AGTAGGGG	GCAGGGGC	AAACTTGA	GGTGGGGG	GTTGGTGC	
TTTGTTTT	TGCGTGTG	TGTGCGTG	GATAGGGG	AGAGGGCG	TACTTGAA	GCTGGGGG	GGAGGTGG	
TTGTTTTT	GGTGTGTG	GGTGCGTG	GGTAGGGG	GAGGGCGG	ACTTGAGG	GTGGGGGT	TGTGGTGG	
TGTTTTTT	GTGGTGTG	GGCGCGTG	GAGTAGGG	GGGGGAGG	TTGAGATA	GTGGGGGA	GTGGTGGA	
TGTTTTTG	CGCGTGTG	GCGCGTGT	GTAGGGGG	AGGGGTTA		GCGGGGGA	GAGGTGGC	
TTGTTGTT	GCGTGTGC	GTGCGTGC	ATAGGGGG	AGGGCGGT		CGGGGGAG	GTGGTTGC	
TTGTTTTG	GGGTGTGT	GTGCGTGA	ATAGGGGT	AGGGAGGC		TGGGGGAG	CGGTGGTT	
TTTTGTTG	GTGTGTGC	GTGCGGTG	AGTGGGGG	GGGGAGGA		GGGGGAGC	GTGCTGGA	
TTTGTTTG	GTGTGCGC	TTGCGGTG	GAGTGGGG	GGGAGGCA		GGGGGAGA	GAGGCGGT	
TTGTGTTT	TGTGTGCG			GGAGGCAA		GGGGGAGT	GACGAGGA	
ATTTTGTT	GTGTGCGT					GGGGGAGG	TGTGCTGG	
ATTTTTGT						AGGGGACG	GGGTCTGG	
ATTTGTTT						GGGGACGG	AACATCGG	
ATGTTTTT						GGGGAGCG		
						GGGGTGTG		
						GGGGTGGT		
Motif	Motif	Motif	Motif	Motif	Motif	Motif	Motif	
TTTTTGTTTT	GYGTGTGC	GYGCGTg	RRTaGGGG	AGGGagg	ACTTGA	YGGGGGag	tGGtgG	

Fig. 2. Motifs common to the most skewed octamers in nine genomes. This table will also be available at www.tigr.org. *H. influenzae* and *Synechocystis* sp. are omitted because the number of skewed oligomers in those genomes is too small to present a clear motif. A number of genomes (especially *E. coli*, *B. subtilis*, *B. burgdorferi*, and *T. pallidum*) have far too many skewed octamers to show in a single figure; this figure contains only a selection, and other motifs may be present in the skewed oligomers for these genomes.

Of these 312 octamers, 273 occurred in the leading strand (which has a probability of occurring by chance of 7.0×10^{-38}), and they agreed perfectly with the origin indicated by the more complete set of octamers. In order to assess how much of the genome was occupied by oligomers with a skewed distribution, we collected the top 350 skewed octamers and computed how much of the entire genome they represented, being careful to avoid double-counting octamers that overlapped. This analysis revealed that more than 150 000 bases (in a 910-kb genome) are involved in skew about the origin. Thus, the phenomenon is strikingly pervasive, and clearly has a major role in the structure of the genome.

M. genitalium and *M. pneumoniae* are closely related organisms (Fraser et al., 1995; Himmelreich et al., 1996) showing extensive genomic similarity. The skew results confirm this relationship. The most highly skewed octamer in *M. genitalium* is ttgatgaa, which is the fourth most highly skewed in *M. pneumoniae* (where it occurs 126 times, of which 100 are in the direction of replication, with a probability of 5.27×10^{-5}). The most highly skewed oligomer in *M. pneumoniae*, catcaaag, does show a skewed orientation in *M. genitalium*, but not a highly significant orientation. In *M. genitalium*, 14 of the 22 most significantly skewed octamers align to a common motif containing the pentamer tgatg. Not surprisingly, the same pentamer motif appears in eight of the top 25 skewed octamers in *M. pneumoniae*.

E. coli has 150 different octamers with a probability of less than 0.5, and many others that, although not as significantly skewed, are also skewed around the same location (the origin). The octamer known as CHI, tggtggcg (Blattner et al., 1997), contains a short four-base motif common to a number of other skewed oligomers; however, there is another motif that is both distinct from CHI and that appears in many more of the highly skewed octamers. This motif, rrcaggg (where 'r' = purine), appears in 21 of the top 25 most significantly skewed oligomers. Note that neither of the two predominant motifs in *E. coli* seems to be the same as those in the mycoplasma or the spirochaetes.

For *H. pylori*, only a handful of octamers are significantly skewed (after probability correction), but these octamers all agree on the location of the origin. If one examines the most skewed octamers in detail, we find that four of the 10 most skewed oligomers match the motif rrtagggg. Using this pattern to search the genome, we find that 302 out of 368 octamers occur in the direction of replication, which gives us a much more significant result, with a probability of $P = 4.9 \times 10^{-29}$

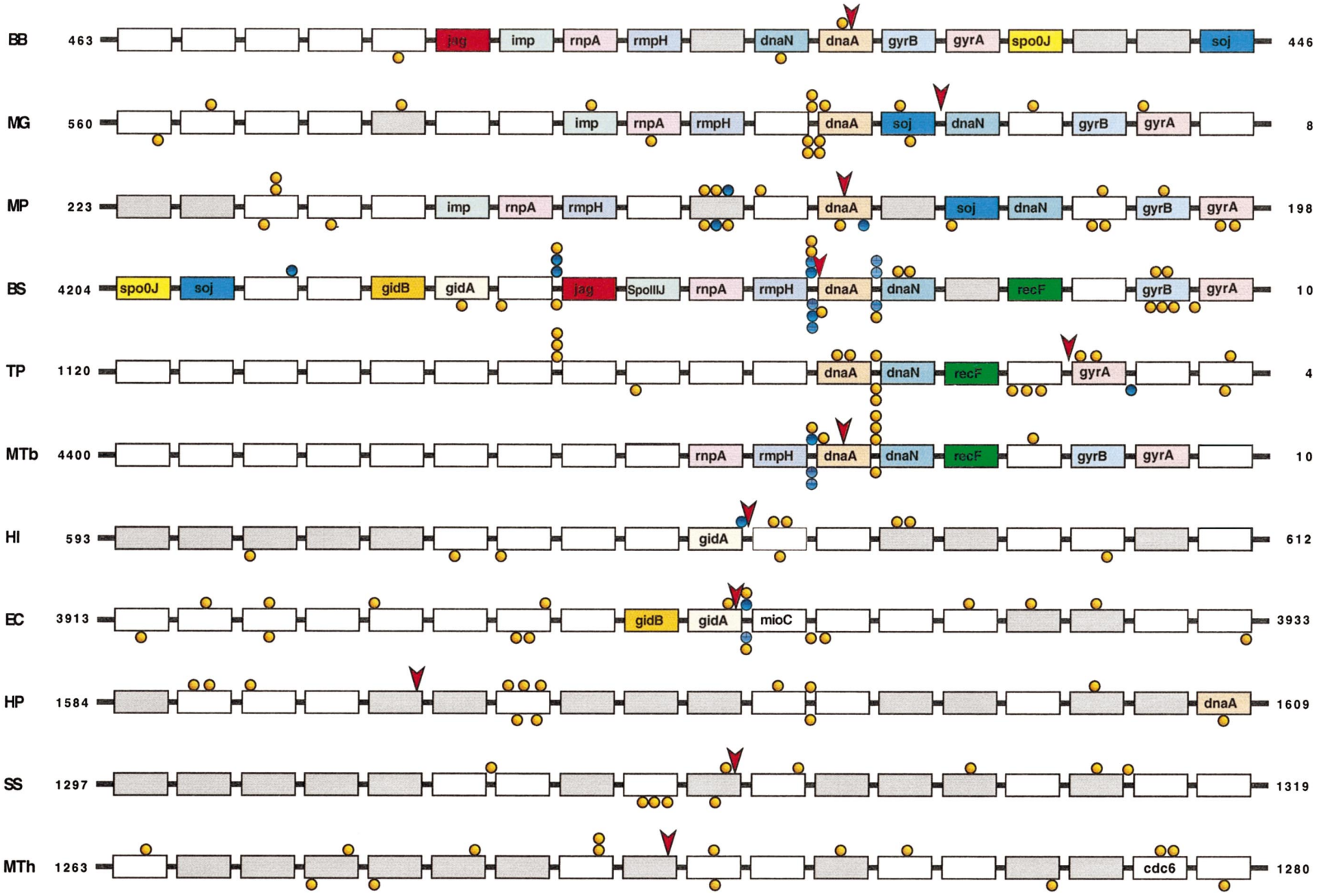

Fig. 3. Origin of replication regions of 11 genomes. The genomes are: *Borrelia burgdorferi* (BB), *Mycoplasma genitalium* (MG), *Mycoplasma pneumoniae* (MP), *Bacillus subtilis* (BS), *Treponema pallidum* (TP), *Mycobacterium tuberculosis* (MTb), *Haemophilus influenzae* (HI), *Escherichia coli* (EC), *Helicobacter pylori* (HP), *Synchocystis* sp. (SS), *Methanobacterium thermoautotrophicum* (MTh). Numbers at the ends of each scale correspond to the position in the genome in thousands. Each box represents a gene (not to scale). Gray boxes represent hypothetical genes with an unknown function. Genes conserved in these regions are indicated by protein and gene name [chromosome partitioning protein (spoOJ), chromosome replication initiator protein (dnaA), DNA gyrase, subunit A (gyrA), DNA gyrase, subunit B (gyrB), DNA polymerase III β (dnaN), glucose inhibited division protein A (gidA), glucose inhibited division protein B (gidB), recombination protein (recF), ribonuclease P (rnpA), ribosomal protein L34 (rmpH), SpoIIIJ-associated protein (jag), inner membrane protein (SpoIIIJ in BS) (imp), Spo0J regulator (soj)]. Predicted origins of replication are indicated by a downward-pointing arrow. Exact matches to a DnaA-box sequence (YTRTCCACA) are indicated (above and below scale by strand) by a blue circle, and single mismatches are indicated by an orange circle.

Table 2
Significantly skewed oligomers from *M. thermoautotrophicum*, a 1.75-Mb genome

Oligomer	Skew ratio[a]	Counts[b]		Predicted origin [c]	Probability (P_1)[d]	P^e_{corr}
		Skewed	Total			
acttgagg	0.691	161	233	1 271 446	5.37e−09	0.0030
tgcagggg	0.699	144	206	1 306 819	1.08e−08	0.0060
agggaggc	0.695	146	210	1 302 986	1.51e−08	0.0084
ggaggcaa	0.829	58	70	1 287 265	2.25e−08	0.012
agagggcg	0.770	77	100	1 311 881	5.51e−08	0.030
aaacttga	0.723	107	148	1 311 026	5.56e−08	0.031
agggcggt	0.795	62	78	1 255 575	1.51e−07	0.081
ttgacttc	0.775	69	89	1 111 293	1.78e−07	0.094
gcaggggc	0.722	96	133	1 336 576	3.19e−07	0.16
aggggtta	0.780	64	82	1 338 286	3.32e−07	0.17
gataatgg	0.711	101	142	1 354 835	5.06e−07	0.25
ttgagata	0.678	135	199	1 298 564	5.34e−07	0.26
gagggcgg	0.721	93	129	1 292 139	5.43e−07	0.26
gggggagg	0.688	106	154	1 238 793	3.39e−06	0.85
gggaggca	0.744	67	90	1 270 053	3.80e−06	0.88
ggggagga	0.679	110	162	1 273 544	6.06e−06	0.97
tacttgaa	0.706	84	119	1 277 272	8.18e−06	0.99

Based on these data, we predict the origin to be near genome position 1 275 000.
[a]For each octamer, the percentage of times that the octamer occurs in the leading strand.
[b]Number of times that the octamer occurs in the leading strand (Skewed) and in the whole genome (Total).
[c]Point around which the distribution is maximally skewed.
[d]Probability of observing a distribution this skewed in a random sequence (see Section 2).
[e]Corrected probability of observing this amount of skew in a genome of this size.

of occurring at random. (Note the similarity between this oligomer and the rrcaggg motif from *E. coli*.) The origin of replication given by this pattern occurs between positions 1 566 000 and 1 640 000.

Surprisingly, one of the three completed archaeal genomes, *M. thermoautotrophicum*, contains 13 highly skewed oligomers, and another four marginally significant oligomers, all indicating an origin and terminus in the vicinity of genome positions 1 275 000 and 413 000, respectively. These oligomers and their associated probabilities are shown in Table 2. The GC-skew plot provides confirmatory evidence for this assignment and also indicates which of the two positions is more likely to be the origin. (For both *E. coli* and *B. subtilis*, the plot of $(G-C)/(G+C)$ makes a transition from negative to positive at the origin. This transition also appears, though less clearly, in *M. thermoautotrophicum*.) This is surprising in light of the fact that our algorithms could find no evidence of oligomers indicating the origin in the other two completed archaeal genomes, *A. fulgidus* and *M. jannaschii*. The skewed oligomers seem to contain three distinct motifs, acttga, ggagg, and aggg. Although it is not known whether or not archaea use a single, rather than multiple, replication origin, the discovery of multiple oligomers clearly skewed around a single pair of points provides evidence for a single origin, at least for *M. thermoautotrophicum*. This contrasts with the prediction, based on the presence of two Cdc6 homologs, that DNA replication initiation is eukaryal for *M. thermoautotrophicum* (Smith et al., 1997).

Finally, *Synechocystis* sp. is the most difficult genome to analyze using octamer skew. This genome does not contain any octamers whose skew is significant after the corrections described below. (*Synechocystis* sp. displays no detectable GC skew.) However, by examining each of the oligomers with a marginally significant skew, we have located one octamer, tcggtcaa, which seems to indicate an origin of replication at genome position 1 310 000. This octamer is weakly skewed, with 106 out of 157 occurrences in the direction of replication (with an uncorrected probability of 1.4×10^{-5}). Based on this evidence, we can assert very tentatively that the origin, whose location has not previously been reported, occurs in this region.

3.3. Genetic characterization of replication origins

The regions surrounding the predicted or experimentally determined origins of replication for the genomes examined are illustrated in Fig. 3. The origin region has been well studied in *E. coli* (Marsh and Worcel, 1977), *B. subtilis* (Ogasawara et al., 1984), and *M. tuberculosis* (Salazar et al., 1996). The origin was experimentally determined for *M. smegmatis* and, because of the significant structural similarity, inferred for *M. tuberculosis* and *M. leprae*. In addition, the origin was predicted for

H. influenzae (Fleischmann et al., 1995), based upon the presence of DnaA-binding boxes and the orientation of rRNA operons. Several distinct classes are evident in the experimentally determined and predicted origin regions.

In six of the genomes, the predicted origin (experimentally determined for *B. subtilis*) is in close proximity to a classical progenitor origin region (Yoshikawa and Ogasawara, 1991) containing the genes *rnpA*, *rmpH*, *dnaA*, *dnaN*, *gyrB*, *gyrA*, and, in some cases, *recF*. In addition, a significant number of DnaA boxes are present close to the center of the predicted origin regions. In addition to the genomes with previously characterized origins, the origins predicted by oligonucleotide skew for *B. burgdorferi*, *M. genitalium*, *M. pneumoniae*, and *T. pallidum* are consistent with this general model for bacterial origins.

H. influenzae and *E. coli* fall into a second class. In these genomes, the origin is associated with the *gidA* gene surrounded by dnaA boxes. In *E. coli*, *gidB* is also present. Both *gidA* and *gidB* are found near the *B. subtilis* origin, and it has been speculated that the *E. coli* replication origin may have evolved from the translocation of the *gidA* region (Ogasawara and Yoshikawa, 1992). *H. pylori*, *Synechocystis* sp., and *M. thermoautotrophicum* fall into a third class. In these genomes, the typical origin genes are scattered throughout the genome. With the exception of *dnaA* in *H. pylori* (found nearly 20 kb from the center of the predicted origin), none of the typical origin genes is found near the predicted origin in this class of organisms. It is interesting that one of the putative eukaryotic-like replication genes, *Cdc6*, is found in the region predicted to be the origin of *M. thermoautotrophicum*. A significant number of hypothetical proteins are present near the predicted origins; it is unclear whether any of these are involved in the replication process.

3.4. Precision of locating origins

If the relative likelihoods of different putative locations are plotted against genome location, an informative picture emerges. The graphs shown in Fig. 4 plot the relative likelihood of the origins for *B. subtilis*, *E. coli*, *B. burgdorferi*, and *M. genitalium*. In each of these genomes, large numbers of different oligomers showed a significant skew. The graphs produced using the combining method described in Section 2.2 reveal that as one moves away from the true origin, the likelihood given by the skew method rapidly decreases. By combining the evidence from multiple oligomers, the indicated origin is much closer to the true origin than some of the oligomers considered separately might indicate. For example, in *E. coli*, the combined likelihood method puts the origin within 1 kb of the actual origin at 3 923 000, although the origins indicated by individual skewed oligomers varied from 3 823 000 to 4 002 000. An interesting side note here is that for *E. coli*, there was a secondary peak observed about 7 kb after the actual origin. Although the true origin had a higher likelihood, in the absence of experimental data, both locations would have to be checked for other evidence, indicating which was correct.

3.5. Conclusion

The skewed octamers reported in the present study cannot be completely explained by differential rates of mutation or repair. These mechanisms can indeed produce a difference in the number of Gs and Cs along a single DNA strand; however, these processes (either mutation or errors in repair) cannot generate numerous exact copies of an eight-base sequence along that strand. Therefore, selective mechanisms must be invoked. One is based on lagging strand synthesis (i.e. priming of Okazaki fragments). This mechanism was suggested as an explanation for the skewed distribution of the CHI sequence in *E. coli* (Blattner et al., 1997); CHI is one of many skewed octamers in *E. coli*, and is perhaps the only such sequence that has been studied in some detail. Even if lagging strand synthesis contributes to the observed skewed distributions, the frequency and the compositional diversity of the skewed oligomers observed in this study would seem to require a further explanation.

A second mechanism that must also be considered is that the oligomer skew is a side-effect of two other processes: transcriptional bias and codon bias. In a number of genomes, most strikingly *M. genitalium*, transcription of genes tends to occur in the same direction as replication (Brewer, 1988). It is well known that different organisms show detectable preferences for synonomous codons (Grantham et al., 1980), and this observation has been used extensively in gene-finding algorithms and other sequence-analysis tasks. Pairs of codons (hexamers) also show distinctive usage patterns and can be useful discriminators between coding and non-coding regions. These two processes together could result in octamers skewed around the replication origin. One potential problem with this mechanism is that transcriptional bias is not nearly as strong as the octamer skew observed in most of the genomes in this study. For example, only 57% of the genes are transcribed from the leading strand *H. pylori*, whereas 82% of the occurrences of the oligomer rrtagggg appear in the leading strand. Thus, whereas a combination of codon usage and transcriptional bias may explain a portion of oligomer skew, it does not seem to be a sufficient explanation. A stronger argument against this explanation is that the skewed octamers do not show any preference for occur-

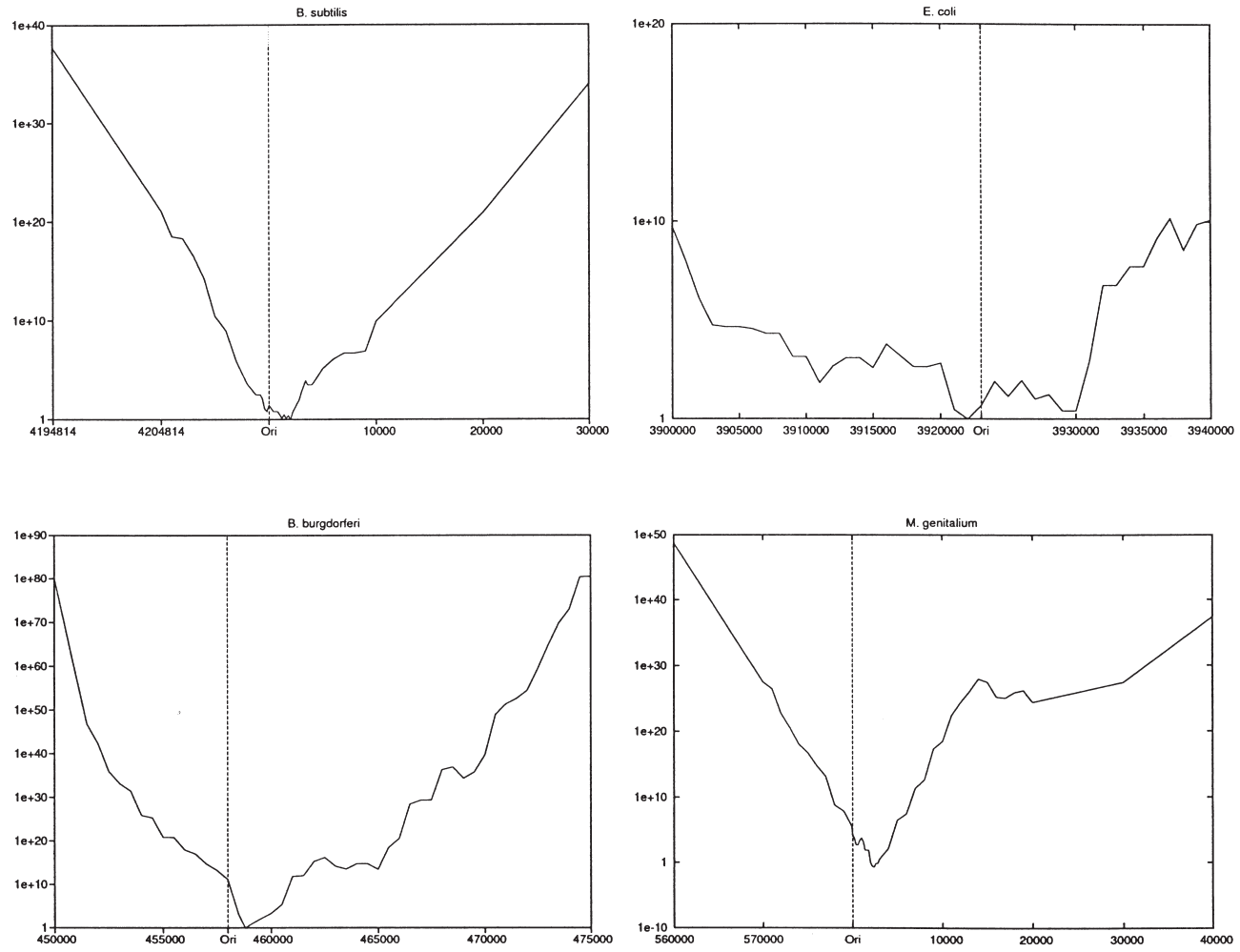

Fig. 4. Graphs of relative likelihood of the origin of replication versus genome position for *B. subtilis*, *E. coli*, *B. burgdorferi*, and *M. genitalium*. The vertical axis is negative log probability; higher values represent less likely locations. The vertical line through each plot indicates the location of the true (for *B. subtilis* and *E. coli*) or putative origin, which is a region covering several hundred bases.

ring within coding regions. As shown in Section 3.2, in at least some genomes, we observe a significant skew even when we consider only those octamers falling in inter-genic regions.

The presence of skew (i.e. a skewed distribution of oligomers) was reported in *E. coli* for two octamers (Blattner et al., 1997), and the current study extends this observation to 11 complete prokaryotic genomes. It also includes an algorithm to find significantly skewed short oligomers and a statistical method to locate the most likely location for the origin of replication based on those oligomers. The skew phenomenon was not observed in two archaeal genomes, but it did appear, surprisingly, in *M. thermoautotrophicum*. The observation of skewed oligomers in one archaeon provides evidence that at least some organisms from this domain of life may replicate using the same mechanism (i.e. a single origin of replication) as bacteria. The coincidence of the predicted origins with those that have been experimentally determined, and with classical origin-region genes and nucleotide motifs in many genomes, increases our confidence that the oligonucleotide skew accurately predicts the origin of replication, even when such other evidence is not present. The algorithm described here should prove useful in locating the origin and terminus in many future prokaryotic genome projects, whether or not they contain the 'classical' origin structure.

Acknowledgement

SLS is supported in part by NIH Grant K01-HG00022-1 and by NSF grant IRI-9530462. The authors thank O. White, C.M. Fraser, R.D. Fleischmann, H.O. Smith, and J.C. Venter for thoughtful comments on the manuscript.

References

Baker, T., 1995. Replication arrest. Cell 80, 521–524.

Beletskii, A., Bhagwat, A., 1996. Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 93, 13919–13924.

Blattner, F. et al., 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1462.

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

Bult, C.J. et al., 1996. Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii*. Science 273, 1058–1073.

Fleischmann, R. et al., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

Francino, M., Chao, L., Riley, M., Ochman, H., 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272, 107–109.

Francino, M., Ochman, H., 1997. Strand asymmetries in DNA evolution. Trends Genet. 13 (6), 240–245.

Fraser, C. et al., 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390 (6660), 580–586.

Fraser, C. et al., 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397–403.

Fraser, C. et al., 1998. Genomic sequence of *Treponema pallidum*, the syphilis spirochete. Science 281, 375–388.

Freeman, J., Plasterer, T., Smith, T., Mohr, S., 1998. Patterns of genome organization in bacteria. Science 279, 1827a

Grantham, R., Gautier, C., Gouy, M., Mercier, M., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, 49–62.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B., Herrmann, R., 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 24 (22), 4420–4449.

Klenk, H. et al., 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing Archaeon *Archaeoglobus fulgidus*. Nature 390, 364–370.

Kunst, F. et al., 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 390, 249–256.

Lobry, J., 1996a. Asymmetric substition patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13 (5), 660–665.

Lobry, J., 1996b. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie 78, 323–326.

Marczynski, G., Shapiro, L., 1993. Bacterial chromosome origins of replication. Curr. Opin. Genet. Dev. 3, 775–782.

Marsh, R., Worcel, A., 1977. A DNA fragment containing the origin of replication of the *Escherichia coli* chromosome. Proc. Natl. Acad. Sci. USA 74, 2720–2724.

Ogasawara, N., Mizumoto, S., Yoshikawa, H., 1984. Replication origin of the *Bacillus subtilis* chromosome determined by hybridization of the first-replicating DNA with cloned fragments from the replication region of the chromosome. Gene 30, 173–182.

Ogasawara, N., Yoshikawa, H., 1992. Genes and their organization in the replication origin region of the bacterial chromosome. Mol. Microbiol. 6, 629–634.

Rohatgi, V.K., 1976. An Introduction to Probability Theory and Mathematical Statistics. Wiley, New York.

Salazar, L., Fsihi, H., deRossi, E., Riccardi, G., Rios, C., Cole, S., Takiff, H., Organization of the origins of replication of the chromosomes of *Mycobacterium smegmatis*, *Mycobacterium lebrae*, and *Mycobacterium tuberculosis* and isolation of a functional origin from *M. smegmatis*. 1996. Mol. Microbiol. 20, 283–293.

Smith, D.R. et al., 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* Δ H: Functional analysis and comparative genomics. J. Bacteriol. 179 (22), 7135–7155.

Yoshikawa, H., Ogasawara, N., 1991. Structure and function of DnaA and the DnaA-box in eubacteria: Evolutionary relationships of bacterial replication origins. Mol. Microbiol. 5, 2589–2597.