2286–2290 *Nucleic Acids Research, 1998, Vol. 26, No. 10*

Analyzing genomes with cumulative skew diagrams

Andrei Grigoriev

Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received February 20, 1998; Revised and Accepted April 1, 1998

ABSTRACT

A novel method of cumulative diagrams shows that the nucleotide composition of a microbial chromosome changes at two points separated by about a half of its length. These points coincide with sites of replication origin and terminus for all bacteria where such sites are known. The leading strand is found to contain more guanine than cytosine residues. This fact is used to predict origin and terminus locations in other bacterial and archaeal genomes. Local changes, visible as diagram distortions, may represent recent genome rearrangements, as demonstrated for two strains of *Escherichia coli*. Analysis of the diagrams of viral and mitochondrial genomes suggests a link between the base composition bias and the time spent by DNA in a single stranded state during replication.

INTRODUCTION

The human mitochondrial genome of <17 kilobase pairs (kb) was sequenced 17 years ago (1). Today, the 50–750 times longer complete genome sequences of the yeast *Saccharomyces cerevisiae* (2), bacteria *Bacillus subtilis* (3), *Borrelia burgdorferi* (4), *Escherichia coli* (5), *Haemophilus influenzae* (6), *Helicobacter pylori* (7), *Mycoplasma genitalium* (8), *Mycoplasma pneumoniae* (9), *Synechocystis sp*. strain PCC6803 (10) and archaea *Archaeoglobus fulgidus* (11), *Methanococcus jannaschii* (12) and *Methanobacterium thermoautotrophicum* (13) are available. *In silico* analysis of such sequences, cataloguing open reading frames (ORFs) and other structural elements, accelerates molecular biology studies by traditional methods.

Studies of genome replication and rearrangements can also benefit from the analysis of complete sequences. I demonstrate this here by exploiting an early hypothesis of differential mutation occurring in leading and lagging strands due to asymmetry of the replication mechanism: combined with natural selection, this factor may be responsible for a skewed base distribution along the genome (14–16). A report on the *E.coli* genome sequence (5) included a GC skew plot, earlier used for smaller regions by Lobry (15), for the whole chromosome. The skew, calculated as (G–C)/(G+C) for a window sliding along the sequence, was shown to switch polarity in the vicinity of the terminus (ter) and origin (ori) of replication, with the leading strand manifesting a positive skew. Such plots may not always be very illustrative due to visible fluctuations for a small window size, while larger windows may hide precise coordinates of polarity switches. Also, as shown below, some of the local polarity switches are important.

Instead, one can use a far more convenient cumulative GC skew: a sum of (G–C)/(G+C) in adjacent windows from an arbitrary start to a given point in a sequence. Similar to integration of the skew function over the DNA length, this value reaches its global maximum at the *E.coli* terminus, while the minimum resides over the replication origin. Figure 1A gives an example of the *M.pneumoniae* sequence (9), where polarity switches are much harder to detect in the skew plot, compared to *E.coli* (5), but the cumulative plot readily reveals them.

For an imaginary genome which underwent no other sequence changes apart from constant rate strand-specific substitutions (with constant skew values for each strand) such a plot will consist of straight lines. Plots of cumulative GC (or AT) skew are termed GC (or AT) diagrams in this paper. For sequences starting at a ter/ori site, diagrams will have characteristic V/inverted V-shapes for bi-directional replication between singular ori and ter. The two global extremities separated by half of the chromosome length will indicate the points where the properties of two strands switch. Recent sequence inversions will be visible as local extremities located at the inversion boundaries.

RESULTS

Microorganisms

Analysis of 14 microbial genomes has shown that GC diagram extremities nearly coincide in every case with known and putative ori/ter sites, the distances between global extremities are very close to 50% of the respective genome lengths and the minimum is always located at the origin (Fig. 1A–D and Table 1). The largest discrepancy is with the *B.subtilis* terminus assigned to a region between a pair of terminator sequences, arresting a replication fork near the *rtp* gene at 2017 kb (3). However, the diagram maximum is located at 1.94 Mb, precisely between *rtp* and another pair that includes a recently identified terminator next to *glnA* (17).

A simple GC skew analysis has been presented for the *B.burgdorferi* sequence (4). While finding a polarity switch for its linear chromosome, it failed to detect any in its multiple plasmids. However, diagrams for some of them (e.g., lp25 and lp17; Fig. 1C) display clear V-shapes, suggesting centrally located origins, as is the case for the chromosome (18).

The diagrams for the archaea *M.jannaschii* and *M.thermoautotrophicum* also feature pairs of global extremities half-chromosomes away from each other (Fig. 1D). Near the minimum of the *M.thermoautotrophicum* diagram there is a putative homolog of the *B.subtilis* gene *soj* (10 kb from the *B.subtilis* origin), which may be involved in replication control and chromosome partitioning (19).

Tel: +49 30 8413 1404; Fax: +49 30 8413 1384; Email: andy@rag3.rz-berlin.mpg.de

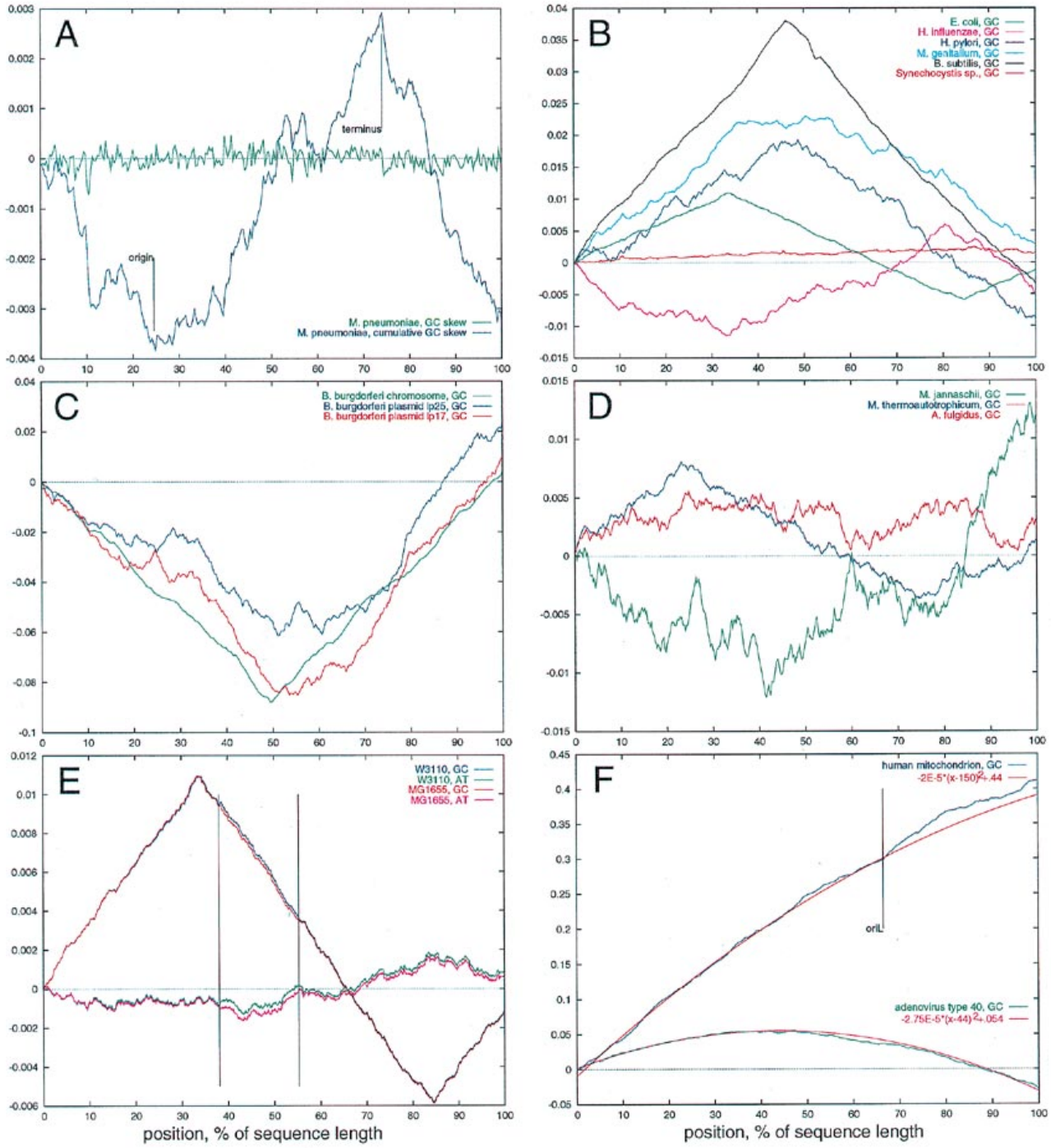

Figure 1. Cumulative diagrams. Organism and *E.coli* strain names are color-coded as shown in each diagram. Y axis (note how the ranges vary between different panels) shows the cumulative skew, calculated as described in the text. To avoid the dependence on the window size *w* and chromosome length *c*, the skew values are multiplied by *w/c*. For all diagrams, *w/c <0.5*%. To allow for plotting graphs for several genomes, the bp coordinate is replaced by a percentage of the genome size on the X axis; zero is selected at base pair 1 of each sequence. The point where a plot intersects the right-hand vertical axis gives a total skew value over the whole genome (for any base pair in a circular sequence, selected as a starting point, the total skew would be the same as well as the distance between global extremities). GC skew versus cumulative skew, putative origin and terminus locations (**A**). GC diagrams for circular bacterial chromosomes (**B**), linear chromosome and plasmids (**C**), circular archaeal chromosomes (**D**). GC and AT diagrams for two *E.coli* strains, vertical lines at 39 and 55% show locations of an inversion and deletion, respectively. (**E**) GC diagrams and second order polynomial trendlines for adenoviral and mitochondrial genomes; for the latter, oriL is marked by the vertical line, oriH is at 0/100% (**F**).

Table 1. Replication ori/ter sites and diagram extremities of microbial genomes

Organism	Replication		GC diagram			AT diagram	
(*bacteria*, ***archaea***)	ori	ter	min	max	min-max	min	max
Diagrams in the opposite phase (GC minimum is AT maximum)							
E.coli (see Fig. 1E)	84.5	34.1	84.5	33.4	48.9	**43.1**	84.5
H.influenzae	32.9	80.1	33.1	80.4	47.3	82.4	32.5
C.trachomatis[a]	n/a	n/a	68.9	18.6	49.7	**25.2**	68.9
M.tuberculosis[a]	0	n/a	0	46.2	46.2	48.4	0
T.pallidum[a]	n/a	n/a	35.5	87.0	48.5	87.0	35.5
B.burgdorferi	50.1	0	49.5	0	49.5	0	49.5
Diagrams in the same phase (GC minimum is AT minimum)							
B.subtilis	0	47.8	0	46.0	46.0	0	46.0
M.pneumoniae	25.1	n/a	24.8	74.1	49.3	25.1	**63.8**
M.genitalium	0	n/a	0	50.4	49.6	0	**61.2**
H.pylori	n/a	n/a	0[b]	48.7	48.7	0	**51.9**
M.jannaschii	n/a	n/a	41.4	98.7	42.5	41.4	98.7
M.thermoautotrophicum	n/a	n/a	74.7	23.2	48.5	**96.9**	22.9

Positions of ori/ter sites (n/a when position is unknown), global minima (min) and maxima (max), and shortest distances between the global extremities (min-max) are given as percentage of genome size. Numbers in bold in the AT diagram column indicate coordinate shifts (of >3%) from corresponding peaks in GC diagrams.
[a]The sequences on the Web sites are not final, so the starts may shift in later releases.
[b]Although ori/ter have not been identified, the diagram minimum coincides with the base pair 1 assignment (7). Another global minimum is at 9% and may be a result of a recent large inversion in this region.

Table 1 includes the data on three genomes awaiting completion and publication: *Treponema pallidum* (sequence available at http://www.tigr.org), *Mycobacterium tuberculosis* (http://www.sanger.ac.uk) and *Chlamydia trachomatis* (http://chlamydia-www.berkeley.edu:4231). Partial contiguous sequences can also be analyzed with this method: a segment of *Mycobacterium leprae* genome (GenBank accession number L39923) features a minimum for oriC site between *dnaA* and *dnaN* on its diagram (data not shown).

AT versus GC skew

The cumulative diagrams provide an easy way to compare the behavior of AT and GC skews over a genome. Unlike the GC skew, an AT diagram maximum does not always correspond to a terminus site but to an origin of replication instead. Also, there are frequent shifts of the positions of AT diagram peaks relative to those in GC diagrams (Table 1, also see Fig. 1E). Such markedly dissimilar behavior of the skews is likely to result from different mechanisms of respective base pair substitution in the above organisms.

Local minima and maxima in AT diagrams correspond to less pronounced local extremities of GC diagrams (Fig. 1E). This may reflect a faster 'smoothing' of sharp distortions that are caused by genome rearrangements, indicating a higher rate of substitutions contributing to the GC skew.

Genome rearrangements

Local extremities may be introduced by sequence inversions or direct translocations to another half of a chromosome (swapping the leading and lagging strands), or integration of foreign DNA into the chromosome. To confirm this, the AT and GC diagrams have been compared for two *E.coli* strains: MG1655 (5) and W3110, available at http://mol.genes.nig.ac.jp/ecoli . The main differences (as clearly seen in Fig. 1E) are located in two regions: a tiny peak in the GC diagram at the 39% position indicates an inversion, while convergence at 55% points to either a smaller-size inversion or presence of a short fragment with a local skew opposite to that of the surrounding region in MG1655, missing in W3110. Zooming in to the plot, one finds no counterpart of that fragment in the W3110 diagram. In other words, this inverted fragment, possibly acquired by horizontal transfer, was deleted at this position in W3110.

Indeed, one will find an inversion between positions 1 756 051–1 775 569 in MG1655 and 1 760 880–1 780 398 in W3110. At 55%, a new cryptic prophage (missing in W3110) was detected between positions 2 556 711 and 2 563 508 in MG1655 (5). As expected, the local GC skew of this prophage is positive, while the chromosome GC skew around the insertion site is negative. This example shows how the long-range (up to the whole genome length) sequence comparisons can be facilitated with the aid of cumulative diagrams.

Other genomes

While there are examples of V-shaped GC diagrams reflecting bi-directional replication in viruses, e.g. SV40 (20) and chloroplasts, e.g., *Euglena gracilis* (21), one can attempt to interpret diagram shapes resulting from other replication mechanisms.

Sequential replication of two strands of the human adenovirus type 40 linear DNA (22) leaves one of them in single-stranded state while another is being duplicated (23). The time the DNA spends single-stranded decreases linearly from one end of the molecule to another, and an integral of such linear function will be an upward pointing parabola. In fact, its GC diagram has a nearly parabolic shape (Fig. 1F).

In vertebrate mitochondria each DNA strand also has its own replication origin: the heavy (H) strand is synthesized first from its oriH origin until the light (L) strand origin oriL is reached; then the remainder of the H strand and the whole L strand are replicated (24). This process in effect divides the DNA into two segments between oriH and oriL, so that in the first segment (2/3 of the genome) the

oriH, and in the second segment the oriL spend the longest time single-stranded. The GC diagram of the H strand of the human mitochondrion (Fig. 1F, trendline shown for the longest segment) reflects this mechanism in that it consists of two distinct parts both displaying parabolic trends (with the GC skew increase towards oriH and oriL in each segment). This holds for all 20 vertebrate mitochondria analyzed (although their total skews vary) and, as one would expect, the closest diagrams are those between closely related species, e.g., human and chimpanzee (data not shown).

Eucaryotic chromosome replication is more complex, it starts from multiple origins. In *S.cerevisiae*, autonomous replicating sequences (ARS) constitute parts of such origins (25), located close to each telomere and at several other sites. GC diagrams for the yeast chromosomes show the most distinct local minima near telomeres but not at other ARS sites (data not shown). Nucleotide substitution mechanisms may differ from those in bacteria and also be chromosome-specific. In fact, diagrams drastically vary between different chromosomes: an AT diagram for chromosome III and GC diagram for chromosome I are both positive (their counterparts negative), while both diagrams for chromosome II are negative and relatively close to each other.

DISCUSSION

With the aid of cumulative skew diagrams, positive (and close to constant for long genome stretches in some of the organisms) leading strand GC skew is observed in 12 out of 14 microbial genomes. This may be related to a presumably average constant time which equal segments of the leading strand spend single-stranded (being more prone to damage) during chromosomal replication.

The diagrams of the differently replicated mitochondrial and adenoviral genomes support this hypothesis. Strand composition bias in relatively short human mitochondrial DNA (26) may result from the prolonged exposure of the H strand during replication with polymerization rate some 200-fold slower than that of *E.coli* (25). As demonstrated in Figure 1F, the GC skew of this strand increases towards oriH and oriL, the longest exposed sites within respective replication segments.

Two origins (one for each strand) are located on the opposite ends of the adenovirus genome. Starting from either end, replication proceeds in the first phase along one strand to produce a full duplex while the other parental strand is gradually displaced as a single strand (23). It means that in the displaced strand the exposure of DNA to damage will decrease linearly from the point where the synthesis starts on the other, template strand. In the second phase of replication the displaced strand becomes a template, and its previously most exposed 5′-end again remains single-stranded until the end of the synthesis. The GC diagram of the adenovirus type 40 genome is strikingly close to a perfect parabola (Fig. 1F), indicating that its derivative (non-cumulative GC skew) also linearly decreases from the left 5′-end of the shown sequence, changing from positive to negative at ~44% of the genome length. Replication may start at either origin, so both strands have a higher GC skew at their 5′-ends.

Spontaneous deamination of C or 5-methylcytosine on the single-stranded template can lead to pairing it with A and a relative abundance of G and T on that strand since deamination rates raise over 100-fold when DNA is single-stranded (27). Mismatch repair peculiarities, such as the fact that a C-C mispair is a very poor substrate for the methyl-directed repair pathway (28) may also contribute to the positive GC skew of the leading strand.

However, it is difficult to propose a single mechanism that can also account for all the differences in the AT diagrams in the above genomes. Clearly, mutational sequence changes arise from multiple independent and organism-specific factors related to relative asymmetries and fidelities of replication and repair processes combined with natural selection. As shown earlier (15,5), the GC skew is most pronounced in intergenic regions and third codon positions where relaxation of selective pressure increases bias, as might be expected with mutational bias.

A recent review (29) strongly advocated DNA strand substitution asymmetry being a result of unequal exposure of the two strands to damage and of differential opportunity for repair during transcription (and admitted that this hypothesis cannot explain the bias in mitochondria). The direction of transcription, rather than replication, would then determine the strand sequence change. While such transcription-coupled repair contribution to the observed bias cannot be ruled out, it does not seem to have a global effect since the numbers of ORFs transcribed in the direction of replication vary between the genomes analyzed in this paper and can be as low as 55% for *E.coli* (5). A notable distortion to the right of the *B.subtilis* diagram maximum (Fig. 1B) corresponds to a large group of genes (from *yonP* to *yomA*) encoded on the lagging strand, while 75% of *B.subtilis* ORFs are collinear with the direction of replication (3). Other divergently transcribed regions do not produce any comparable distortions, so this peak probably indicates a recent inversion or an insertion of foreign DNA. One also has to bear in mind that the direction of transcription may be a consequence of earlier inversions.

Of all analyzed bacterial genome sequences, only the *Synechocystis sp*. diagram (Fig. 1B) displays an atypical behavior. Its skew amplitude is very low, the *dnaA* gene is between the global minimum at zero and maximum at 87%. The replication mechanism of this chromosome is unknown and seems to differ from that of other bacteria in the resulting nucleotide substitution effects.

Little is known about replication in archaea and it is hard to explain the diagram behavior of *A.fulgidus* (Fig. 1D) without speculating about large rearrangements or a possible case of more than one pair of origins and termini: it features two global minima and two global maxima located between them. This resembles the behavior of diagrams for eucaryotic chromosomes. One can observe a certain degree of similarity between the diagrams of *A.fulgidus* and *M.jannaschii*: their global and prominent local extremities are unexpectedly close (e.g. the 60% position) and many of the changes occur in the opposite phase. However, this is likely to be a coincidence for the two genomes that differ in size by 500 kb, with starts arbitrarily located upstream of the repeat-rich regions (11,12). Repeats are found in some of the large local peaks (data not shown) suggesting potential points of rearrangements.

Homology searches with sequences from the regions around global extremities have not revealed any consensus for potential replication origins or termini in archaea. However, the fact that circular chromosomes of *M.jannaschii* and, more clearly, *M.thermoautotrophicum* are divided into two nearly equal parts by their GC diagrams suggests that the similarity with bacterial diagrams may be not coincidental. The regions surrounding the diagram extremities can be further tested experimentally for their ability to replicate. Such studies may also determine if the prominent local (in addition to the global) extremities point to more than one pair of ori/ter sites in *A.fulgidus* and *M.jannaschii*.

Microbial diagrams show potential rearrangement events, with disruptions of linearity of up to 5% of a genome length. In

addition to homologous recombination, one can envisage a mechanism for such inversions related to the observed nucleoid structure of *E.coli*, with 20–50 loops emanating from a dense node of DNA (30). A hypothetical event involving two double-strand breaks at the root of one loop and a subsequent recombination could produce a loop-sized inversion. If the diagram non-linearities indeed reflect inversions undergone by the analyzed genomes, this may represent a way the microorganisms regulate expression or even 'create new genes' randomly combining parts of existing ORFs.

Relative numbers and sizes of diagram distortions may be indicative of the genome stability in different organisms, characterized by the quantities of horizontally transferred DNA in their chromosomes and recombination frequencies. Any distortion manifests a change in the compositional properties of DNA, thus reflecting a local balance of many mutational events (substitutions and rearrangements) and selective constraints. Other estimates of the relative genome stability are provided by the data on some 1500 DNA uptake sites in *H.influenzae* (31), and diverse gene arrangement in different strains of *H.pylori* (32), although the latter fact has been recently contested (33). These findings are in agreement with the more distorted patterns in the diagrams of those organisms. The 'jagged' diagrams of the archaea may reflect the mutation forces specific to their extreme environments or the similarity of their replication machinery to that of eucaryotic genomes (11–13). Independently replicating genome halves seem to accumulate different amounts of rearrangements, this (and unequal lengths of the 'halves') may explain the non-zero total skews at the 100% coordinate (Fig. 1).

As demonstrated for two *E.coli* strains (Fig. 1E), long-range homologies and sequence rearrangements (e.g., an inversion of 20 kb) can be easily found using cumulative diagrams for the whole genome sequence comparisons. Even the diagrams which follow complex patterns and are harder to interpret may help with comparative analysis: e.g., remarkably similar diagrams of chloroplasts of higher plants point to several clear differences between their genomes (data not shown). The simplicity and effective visual representation of cumulative diagrams warrant their usefulness for other studies of global sequence organization.

While this paper was under review, I became aware of a publication by Freeman *et al.* (34) who analyzed a subset (nine microbial genomes) of the data using plots of cumulative purine, keto and coding-strand excess, calculated by 'walking' along a sequence in the four nucleotide space (a similar method was earlier described in ref. 35). Global extremities in their plots are also close to the known ori/ter sites but not in every case. The reason for this is clear, from the important differences between the AT and GC skews discussed above and summarized in Table 1. Purine and keto excesses are aggregate measures which fail to detect such differences in the contributions of the AT and GC skews.

Diagrams for the genomes not shown here are available at http://www.mpimg-berlin-dahlem. mpg.de/~andy/diagrams

ACKNOWLEDGEMENTS

Thanks are due to Igor Ivanov for stimulating discussions and helpful suggestions, Hans Lehrach for continuing support, Leo Schalkwyk, Elmar Maier and John O'Brien for critical reading of the manuscript. Generous data release policies of the organizations governing the Web sites mentioned in the text, are gratefully acknowledged. This work was in part supported by the grant 01KW 9608 of the BMBF Human Genome Project.

REFERENCES

1 Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., *et al.* (1981) *Nature* **290**, 456–465.
2 Goffeau, A., Aert, R., Agostini-Carbone, M. L., Ahmed, A., Aigle, M., Alberghina, L., Albermann, K., Albers, M., Aldea, M., Alexandraki, D., *et al.* (1997) *Nature* **387**, 5–105.
3 Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., *et al.* (1997 *Nature* **390**, 249–256.
4 Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., *et al.* (1997) *Nature* **390**, 580–586.
5 Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., *et al.* (1997) *Science* **277**, 1453–1474.
6 Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A, Merrick, J.M., *et al.* (1995) *Science* **269**, 496–512.
7 Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., *et al.*. (1997) *Nature* **388**, 539–547.
8 Fraser, C.M, Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., *et al.* (1995 *Science* **270**, 397–403.
9 Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., Herrmann, R. (1996) *Nucleic Acids Res*. **24**, 4420–4449.
10 Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., *et al.*. (1996) *DNA Res*. **3**, 109–136.
11 Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn. M., Hickey, E.K., Peterson, J.D., *et al.* (1997) *Nature* **390**, 364–370.
12 Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., Fitzgerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., *et al.* (1996) *Science* **273**, 1058–1073.
13 Smith, D. R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., *et al.*. (1997) *J. Bacteriol.* **179**, 7135–7155.
14 Wu, C.-I. and Maeda, N. (1987) *Nature* **327**, 169–170.
15 Lobry, J.R. (1996) *Mol. Biol. Evol.* **13**, 660–665.
16 Perna, N.T. and Kocher, T.D. (1995) *J. Mol. Evol.* **41**, 353–358.
17 Griffiths, A.A. and Wake, R.G. (1997) *J. Bacteriol.* **179**, 3358–3361.
18 Old, I.G., MacDougall, J., Saint Girons, I. and Davidson, B.E. (1992) *FEMS Microbiol. Lett.* **78**, 245–250.
19 Ireton, K., Gunther, N.W. and Grossman, A.D. (1994) *J. Bacteriol.* **176**, 5320–5329.
20 Li, J.J. and Kelly, T.J. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 6973–6977.
21 Schlunegger, B. and Stutz, E. (1984) *Curr. Genet.* **8**, 629–634.
22 Davison, A.J., Telford, E.A., Watson, M.S., McBride, K. and Mautner, V. (1993) *J. Mol. Biol.* **234**, 1308–1316.
23 Kornberg, A. and Baker, T.A. (1992) *DNA Replication* (2nd ed.) Freeman, and Co., New York.
24 Clayton, D.A. (1982) *Cell* **28**, 693–705.
25 Newlon, C.S. and Theis, J.F. (1993) *Curr. Opin. Genet. Dev.* **3**, 752–758.
26 Tanaka, M. and Ozawa, T. (1994) *Genomics* **22**, 327–335.
27 Frederico, L.A., Kunkel, T.A. and Shaw, B.R. (1990) *Biochemistry* **29**, 2532–2537.
28 Su, S.S., Lahue, R.S., Au, K.G. and Modrich, P. (1988) *J. Biol. Chem.* **263**, 6829–6835.
29 Francino, M.P. and Ochman, H. (1997) *Trends Genet.* **13**, 240–245.
30 Hinnebusch, B.J. and Bendich, A.J. (1997) *J. Bacteriol.* **179**, 2228–2237.
31 Smith, H.O., Tomb, J.F., Dougherty, B.A., Fleischmann, R.D. and Venter, J.C. (1995) *Science* **269**, 538–540.
32 Jiang, Q., Hiratsuka, K. and Taylor, D.E. (1996) *Mol. Microbiol.* **20**, 833–842.
33 Hancock, R., Alm, R., Bina, J. and Trust, T.(1998) *Nature Biotechnol.* **16**, 216–217.
34 Freeman, J.M., Plasterer, T.N., Smith, T.F. and Mohr, S.C. (1998) *Science* **279**, 1827.
35 Lobry, J.R. (1996) *Biochimie* **78**, 323–326.