

Genome Arithmetic

Plotting integral purine (A+G versus T+C), keto (G+T versus A+C), and coding-strand excess for nine genomes, James M. Freeman *et al.* (1) found global peaks in many cases near the replication origin (ori) and terminus (ter) sites. They mention earlier findings of similar strand asymmetries with GC skew (2, 3), calculated as (G-C)/(G+C) in a window sliding along a sequence.

In numerical integration with very small windows, purine excess is practically equivalent to a sum of GC and AT skews, and keto excess to their subtraction. This arithmetic is important, as seen from the differences between the cumulative excess (1) and cumulative skew plots (4). DNA strand properties with respect to replication and repair switch at ori/ter, and the leading strand has been shown to contain more G than C in 12 out of 14 microbial genomes (4). This is not the case with the leading strand AT skew: A is less than T in six genomes, for example, *Escherichia coli*, but A is greater than T in others, for example, *Bacillus subtilis*. This variation, and the fact that global switches in AT skew often occur (six cases) far away from ori/ter (4), may negatively affect identification of these sites with the use of the aggregate

values of purine and keto excess (1).

Evolutionary forces seem to affect AT and GC skews differently. Pertinent to transcription and selection, coding-strand excess correlates strongly with AT skew in the first codon position (5), for example, in *Haemophilus influenzae*, where there is only a weak correlation with purine excess (1). GC skew may be linked with replication and repair (2, 3), because it changes linearly with the time the template spends in a single-stranded state during replication of vertebrate mitochondria and viruses (4).

The "rough" plot patterns observed for some species (1) have been explained by uptake of foreign DNA or prophage integration (a common interpretation of A+T-rich islands in A+T content plots). In addition, I would like to suggest another explanation: Some plot distortions correspond to recent inversions, as has been demonstrated (4) for two strains of *E. coli* (6).

Andrei Grigoriev

Genome Pharmaceuticals Corporation,
Lochhamer Str. 29, Martinsried 82152,

Germany

E-mail: andrei.grigoriev@gpc-ag.com

References and Notes

1. J. M. Freeman *et al.*, *Science* **279**, 1827 (summary) (1998). Full text at www.sciencemag.org/cgi/content/full/279/5358/1827a
2. J. R. Lobry, *Mol. Biol. Evol.* **13**, 660 (1996); *Science* **272**, 745 (1996).
3. F. R. Blattner *et al.*, *Science* **277**, 1453 (1997).
4. A. Grigoriev, *Nucleic Acids Res.* **26**, 2286 (1998). (Results available at www.gpc-ag.com/diagrams). On the basis of consistent behavior of GC skew, locations of ori/ter sites have been predicted for several genomes. For *Treponema pallidum*, the predicted ori position coincides with that recently published [C. M. Fraser *et al.*, *Science* **281**, 375 (1998)].
5. A. Grigoriev, unpublished results.
6. Strains MG1655 (3) and W3110 (<http://mol.genesis.nig.ac.jp/ecoli>).

13 April 1998; accepted 20 July 1998

Response: In order to reveal biologically relevant features, complete genome sequence data require suitable methods of graphical analysis and display. Our recent technical comment (1) described the use of cumulative strand asymmetries of purines, keto bases, and coding sequences to reveal the correlations between these functions and replication origins-termini and the directions of gene transcription in nine complete bacterial genomes. We also suggested that the plots indicate positions of DNA segments recently acquired by phage-transposon integration or uptake of transforming DNA.

Grigoriev (2) has used a similar approach by adapting the GC-skew method of Lobry

Table 1. Comparison of bacterial replication-origin predictions by different cumulative strand-asymmetry analyses.*

Genome	Total length (bp)	Cumulative two-base excess*		Cumulative GC/AT skew†			Cumulative one-base excess‡	
		Best function and pt.	Deviation (kb)	Window size (bp)	Best function and pt.	Deviation (kb)	Best function and pt.	Deviation (kb)
<i>E. coli</i>	4,639,222	GT-min	-2.3	38	GC-min	-0.3	G-min	-0.3
<i>B. subtilis</i>	4,214,814	AG-min	-0.05	41	GC-min	-0.04	G-min	-0.01
<i>H. influenzae</i>	1,830,136	GT-min	13.3	60	GC-min	5.43	G-min	5.4
<i>M. pneumoniae</i>	816,395	GT-min	-203.6	34	GC-min	28.9	-	-
<i>M. genitalium</i>	580,074	AG-min	0.2	41	AT-min	0.02	-	-
<i>H. pylori</i>	1,667,868	GT-min	0.001	41	AT-min	9.07	-	-
<i>Synechocystis</i> sp.	3,573,470	AG-max	755	40	AT-max	754	-	-

*In each case the maxima and minima of the relevant functions were computed and compared with published origin positions (1). Deviations are reported as: (maximum/minimum position) - (position of published origin). Where this amounted to more than half the genome length, the length of the genome was added to the published origin position before subtracting so as to compute the shorter distance between predicted and observed origin positions while preserving the sign convention, which gives negative values to predicted positions lying upstream of the observed origin (on the presented

strand). †Calculated by the method described in (2) after optimizing window size to maximize resolution. ‡Analogous to the method described in (1), where the cumulative G excess function is incremented by +1 for a G residue, by -1 for a C, and left unchanged for A or T. The A excess function is computed correspondingly. [These "one-base excess" functions correspond to projections of the sequence trajectory on the the GC and AT planes of the sequence space (1).] Dash (-) indicates not calculated.

Table 2. Comparison of bacterial replication-terminus predictions by different cumulative strand-asymmetry analyses.

Genome	Total length (bp)	Cumulative two-base excess*		Cumulative GC/AT skew			Cumulative one-base excess	
		Best function and pt.	Deviation (kb)	Window size (bp)	Best function and pt.	Deviation (kb)	Best function and pt.	Deviation (kb)
<i>E. coli</i>	4,639,222	AG-max	-38.4	38	GC-max	-38.4	G-max	-38.4
<i>B. subtilis</i>	4,214,814	AG-max	-75.3	41	GC-max	-75.2	G-max	-75.1
<i>H. influenzae</i>	1,830,136	GT-max	-43.2	60	AT-min	-8.9	A-min	8.9

*See notes to Table 1 for explanation of column headings and calculational procedures.

TECHNICAL COMMENTS

(3) to calculate cumulative AT- and GC-skew curves that also show these features (4). He points out that discontinuities in cumulative base-asymmetry curves may correlate with sequence inversions, and correctly observes that amalgamating the A and G (or G and T) bases to create purine- (or keto-) excess curves may adversely affect their utility for locating origins and termini of replication because the strand asymmetry patterns for A track the replication direction less consistently than the patterns for G.

Window size (taken by Grigoriev as an arbitrary parameter) limits the resolution of cumulative base-skew curves, and its minimum in turn depends on the details of the sequence, because any window size that can be filled by an exclusively AT-containing segment of the sequence will produce a (mathematically undefined) singularity in the GC-skew plot when such a sequence is encountered (and vice versa for the AT-skew). Grigoriev does not furnish details of his window size beyond stating that it was always less than 0.5% of the genome length, which, for example, in *E. coli* equals 23 kb. In order to compare the different types of genome plot, we have optimized his approach by determining the minimum window size possible for each of the nine complete genomes described in our earlier comment (1) and then used these to compute cumulative GC- and AT-skew plots at the maximum possible resolution. For the three bacterial genomes for which the positions of DNA replication termini are known, we also computed two "single-base excess" curves (which have the same selectivity as the cumulative GC- and AT-skews without the singularity problem). To calculate the "A excess" for example, we walked along the sequence and counted every A as +1, every T as -1, and G's and C's as 0.

The essential results of this exercise are

summarized in Table 1 (for replication origins) and Table 2 (for termini). For the case of each genome and each method, we have chosen the best match to the reported origin or terminus and calculated the deviation between the predicted and observed chromosomal feature (5). We conclude that (i) with one exception (*H. pylori*), the optimized cumulative skew method comes closer to pinpointing origins than does the cumulative two-base excess method—provided that the correct function is selected (alternate skew functions, which are not shown, do not have minima or maxima closer to the targets than the two-base excess functions listed); (ii) in one case (of three) the cumulative-skew method comes closer to predicting the correct terminus than the two-base excess method; (iii) the "one-base excess" method comes closest of all to the targets—it also works better than the cumulative skew method because it avoids the need to optimize windows and by definition works at single-base resolution; (iv) the best function to use (AG versus GT excess, GC versus AT skew, or A versus G excess) is not a priori clear and, in particular, does not correlate with overall GC content.

Whole-genome strand asymmetry analyses should prove useful for a number of purposes in addition to origin and terminus location. In particular, the striking correlations between coding strand selection and purine excess (1) suggest that this function may be helpful in open reading frame verification, and both the cumulative skew and cumulative excess plots reveal important locations of genome rearrangements. Qualitative and quantitative analysis of the "roughness" of these plots should aid in the understanding of the differing dynamics of genomes in different organisms as well as fundamental differences in the behavior of their

replication, transcription, and repair machinery. This is a broad field and many other variations of these analytic methods can be developed, some of which may be highly revealing of significant genome features. Given the relative ease of creating such "pictures" of chromosomes, it may well be advisable to combine several methods to attack any specific problem.

James M. Freeman

*BioMolecular Engineering Research Center,
Boston University,
Boston, MA 02215, USA*

Thomas N. Plasterer

*Department of Pharmacology,
Boston University*

Temple F. Smith

*BioMolecular Engineering Research Center,
Boston University*

Scott C. Mohr

*Department of Chemistry,
Boston University*

References and Notes

1. J. M. Freeman *et al.*, *Science* **279**, 1827 (summary) (1998). Full text at www.sciencemag.org/cgi/content/full/279/5358/1827a
2. A. Grigoriev, *Nucleic Acids Res.* **26**, 2286 (1998).
3. J. R. Lobry, *Science* **272**, 745 (1996).
4. AT skew for a given DNA segment is defined as $(A-T)/(A+T)$, with a corresponding expression for GC skew. The cumulative plots calculate these quantities for a fixed window and add the values as the window walks along the sequence.
5. Neither the replication origin nor the terminus constitutes a single point in the genome; thus, there is a degree of arbitrariness in these comparisons. The *E. coli* origin of replication (*oriC*) is 245 bp long, and the so-called terminus resides in the middle of a cluster of at least six *Ter* sites extending over 25% of the chromosome (6). For simplicity, we have accepted the positions for origin and termination cited in the primary publications of the relevant complete genome sequences.
6. T. M. Hill, *Annu. Rev. Microbiol.* **46**, 603 (1992).

19 June 1998; accepted 20 July 1998