

Genome Update: AT content in sequenced prokaryotic genomes

Genomes of the month – the good, the bad and the ugly

The sequences of three bacterial genomes have been published in the month since the last Genome Update was written, with each representing bacteria with different and interesting lifestyles. The three genomes include that of a member of the acidophilus group of intestinal lactobacilli (a 'good bacterium'), that of a pathogen that causes a highly contagious respiratory disease in cattle (a 'bad bacterium', certainly from the perspective of the ranchers) and that of a bacterial predator that invades and consumes other bacteria ('ugly' from the point of view of its prey).

The genome of *Lactobacillus johnsonii* strain NCC 533 (Pridmore *et al.*, 2004) is just under 2 Mbp in size, has an AT content of 65% and encodes about 1800 proteins (see Table 1). Surprisingly, this genome is missing key enzymes for the biosynthesis of amino acids, many co-factors and purine nucleotides (Pridmore *et al.*, 2004). However, this appears to be compensated for by extra amino acid permeases, peptidases and small-molecule transporters, which bring in the necessary molecules from the environment. Perhaps this is not surprising for a bacterium which is known for living as an intestinal commensal, living in the rich and relatively constant intestine environment. *L. johnsonii* is a probiotic bacterium and, as such, is claimed to be involved in pathogen inhibition, epithelial cell attachment and immunomodulation. Several large adhesion proteins have been found in this genome, as well as bile salt hydrolases and transporters, which are likely to be involved in persistence in the gastrointestinal tract (Pridmore *et al.*, 2004). Obviously it is better, from a human's perspective, to have one's intestinal tract populated by a healthy population of lactobacilli which can prevent the growth of potential pathogens.

In contrast to the probiotic bacterium, *Mycoplasma mycoides* subsp. *mycoides* SC type strain PG1^T (Westberg *et al.*, 2004) is the causative agent of pleuropneumonia in cows. Its genome is 1.2 Mbp in size, quite AT-rich (76%) and full of DNA repeats and insertion sequences. The genome appears to have a great deal of plasticity, and several potential virulence factors are found in the genome, including genes encoding putative variable surface proteins. As in other *Mollicutes*, the universal stop codon UGA encodes tryptophan, and this codon (UGA) occurs 24 times more often than the Trp UGG codon, most likely because of the selective pressure towards a more AT-rich genome (Westberg *et al.*, 2004).

Bdellovibrio bacteriovorus HD100 is a bacterial predator that feeds on other

Gram-negative bacteria (Rendulic *et al.*, 2004). Its genome is 3.8 Mbp long, with an AT content of 50%, as shown in Table 1. The authors say that this genome is 'surprisingly large', although it is about the same size as those of other δ -*Proteobacteria* (3.8 Mbp for *Geobacter sulfurreducens* and *Desulfovibrio vulgaris*). Perhaps the genome is large compared to the small size of the bacterium (as little as 200 nm wide and 500 nm long). *B. bacteriovorus* can use its flagella to quickly propel itself towards its prey (another Gram-negative bacterium, which is usually larger in cell size), then it attaches to the membrane, makes a hole and squeezes itself inside, and essentially consumes the other bacterium, from the inside out. *Bdellovibrio* prey can also include plants, animals and human pathogens, so understanding the

Table 1. Summary of the published genomes discussed in this Update

Note that the accession number for each chromosome is the same for GenBank, EMBL and the DNA DataBase of Japan (DDBJ).

Genome	Size (bp)	AT content (%)	rRNA operons	tRNAs	CDS	Accession no.
<i>Bdellovibrio bacteriovorus</i> HD100	3 782 950	49.4	2	36	3583	BX842601
<i>Lactobacillus johnsonii</i> NCC 533	1 992 676	65.4	6	79	1821	AE017198
<i>Mycoplasma mycoides</i> PG1 ^T	1 211 703	76.0	2	30	1016	BX293980

Microbiology Comment provides a platform for readers of *Microbiology* to communicate their personal observations and opinions in a more informal way than through the submission of papers.

Most of us feel, from time to time, that other authors have not acknowledged the work of our own or other groups or have omitted to interpret important aspects of their own data. Perhaps we have observations that, although not sufficient to merit a full paper, add a further dimension to one published by others, or we may have a useful piece of methodology that we would like to share.

Guidelines on how to submit a *Microbiology Comment* article can be found in the Instructions for Authors at <http://mic.sgmjournals.org>

It should be noted that the Editors of *Microbiology* do not necessarily agree with the views expressed in *Microbiology Comment*.

Chris Thomas, Editor-in-Chief

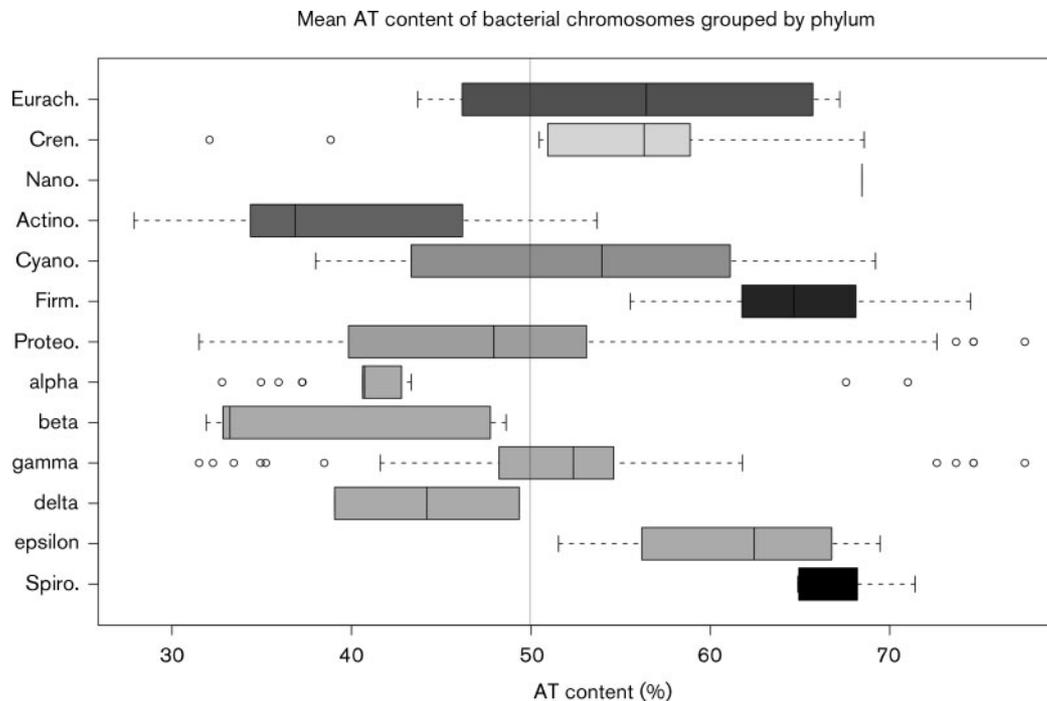


Fig. 1. AT content ranges of sequenced prokaryotic genomes, within different phyla. The data are shown as a 'box and whiskers' plot, where the shaded box represents the middle 50% of the observations (interquartile range) and the vertical line in each box is the median value. The 'whiskers' represent plus or minus 1.5 times the interquartile range. Outlier points are designated with circles. An AT content of 50% (e.g. a random distribution of the four DNA bases) is marked with a vertical line on the graph. Note that some phyla, such as *Actinobacteria*, are mainly more GC-rich, whilst other phyla, such as the *Firmicutes*, are more AT-rich. Eurach., *Euryarchaeota*; Cren., *Crenarchaeota*; Nano., *Nanoarchaeota*; Actino., *Actinobacteria*; Cyano., *Cyanobacteria*; Firm., *Firmicutes*; Proteo., *Proteobacteria* (note that alpha, beta, gamma, delta and epsilon refer to subdivisions within the *Proteobacteria*); Spiro., *Spirochaetes*.

mechanism of how this predator works has broader implications in terms of the development of antimicrobial agents.

Method of the month – AT content comparison in bacterial genomes

The mean AT content of sequenced prokaryotic genomes is shown in Fig. 1. The range is from 27.9% AT in *Streptomyces coelicolor* (Bentley *et al.*, 2002) to 77.5% in the endosymbiont *Wigglesworthia glossinidia* (Akman *et al.*, 2002). As stated above, the *M. mycoides* genome is very AT-rich. In fact, when compared to other sequenced bacteria (at the time of writing, 152 bacterial sequenced genomes – see supplemental web table), only *W. glossinidia* is more AT-rich. Fig. 1 shows that the range of AT contents for several different prokaryotic phyla. For example, all the *Firmicutes* are AT-rich, whilst the *Actinobacteria* are GC-rich. Within the *Proteobacteria*, where more genomes have

been sequenced, the alpha, beta and delta subdivisions tend to be more GC-rich, whilst the gamma and epsilon divisions are more AT-rich.

The AT content of a genome is a mean value, and for many genomes there are regions which differ substantially from this mean. To visualize these differences, we have plotted the AT content along the genome for all 152 sequenced genomes (see supplemental web pages). These differences have been postulated to be representative of horizontal gene transfer, although some of the differences in AT content appear to have biological meaning. As two examples, we will consider local and global differences in AT content, as shown in Fig. 2. Fig. 2(a) plots the difference in AT content for 200 bp upstream of translation start versus 200 bp downstream. A clear and significant difference is seen in nearly all of the 152 genomes examined and is reflective of mechanical properties of the

double helix; i.e. there is a general trend for upstream regions to be more curved, melt more easily and be more rigid (Pedersen *et al.*, 2000). Individual plots for each genome can be seen on the supplemental web pages.

At a more global level, for a few sequenced genomes, we have observed that the region around the replication terminus appears to be more AT-rich than the rest of the genome and that the region around the replication origin is less AT-rich. To test this hypothesis on all 152 sequenced bacterial genomes, for each genome we calculated the mean AT content for a region containing 8% of the genome length flanking either side of the predicted replication terminus and origin (P. Worning, L. J. Jensen, P. F. Hallin, H.-H. Stærfeldt & D. W. Ussery, unpublished data). The results are shown in Fig. 2(b). Again, individual plots for each genome can be found on our

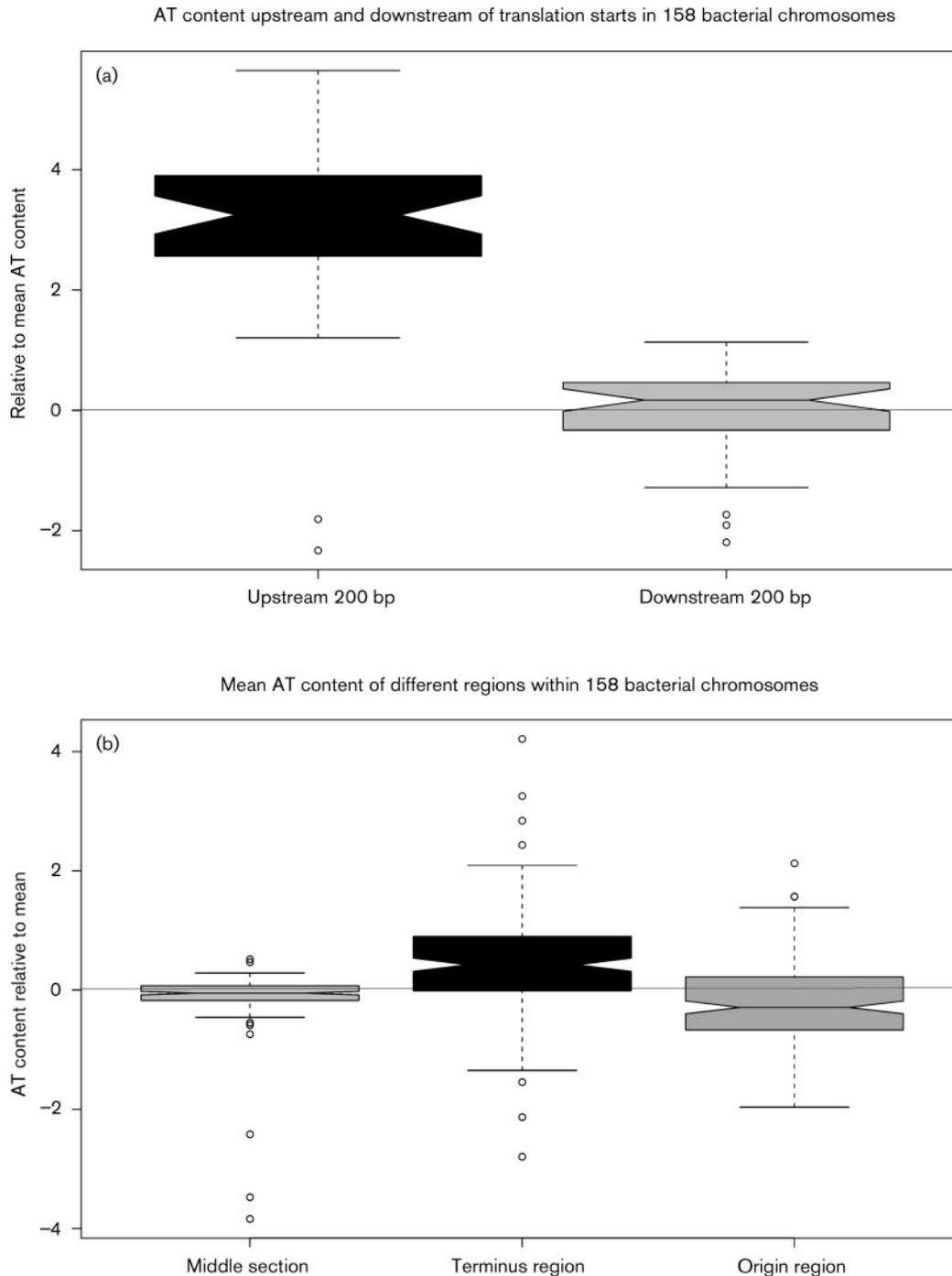


Fig. 2. Comparison of AT content within different parts of prokaryotic chromosomes. (a) For each of the 158 sequenced bacterial chromosomes, the AT content of 200 bp upstream and 200 bp downstream of translation start sites for all genes was calculated. The mean value for each genome was used to generate the 'box and whiskers' plot shown. The notch in the middle of each box represents the 95% confidence interval for the median. Thus, the two medians differ significantly. Note that the values plotted are the relative differences in AT content, to allow for comparison of genomes with different mean AT content. (b) For each chromosome, the mean AT content was calculated for a region on either side of the replication origin and terminus (representing 8% of the length of the chromosome). The difference of the medians for the replication terminus and origin regions is significant. [Note that the 158 chromosomes come from only 152 genomes, since some bacterial genomes contain multiple chromosomes.]

supplemental web pages. It should be stressed that we are talking about large regions here – 8 % of the genome on either side of the replication terminus and origin. Of course, there is a very small region (a few hundred base pairs) around the replication origin that is more AT-rich and must open up for replication initiation, but on a larger scale of hundreds of thousands of bases, the region around the origin is more GC-rich, whilst the region around the replication terminus is more AT-rich. Since AT-rich regions tend to be more curved, this is consistent with our previous observation that a large region around the replication terminus in *Escherichia coli* and *Bacillus subtilis* genomes is more curved than the rest of the genome (Pedersen *et al.*, 2000). We are currently testing whether the increased average curvature seen around the replication terminus is more significant than the differences in AT content.

As a final point of discussion, the chances of finding a repeat in a genome depends on the AT content. Thus, it is not surprising that we find that the *M. mycoides* genome has the second highest level of global Direct repeats [26.9 vs 27.7 % for the *Phytoplasma asteris* genome (Oshima *et al.*, 2003)], as well as the second highest level of global Inverted repeats (15.7 vs 24 % for the *P. asteris* genome). Similarly, at the level of local repeats, the *M. mycoides* genome has the second highest level of local Direct repeats (24.2 %; this time, the highest goes to the most AT-rich genome, *W. glossinidia*, at 33.4 %). Finally, at the level of local Inverted repeats, the *M. mycoides* genome comes in fifth place (17.4 %; again, *W. glossinidia* is the highest, with 27.8 %). However, many genomes have high levels of repeats despite having an AT content of near 50 %. For example, one of the sequenced *E. coli* O157 genomes has a substantial fraction of global Direct repeats (11.8 % of the genome, making it eighth highest on the list of more than 150 genomes), even though it has an AT content of 49.5 % (see the supplemental tables for lists of genomes sorted by AT content and various types of global and local repeats). Thus, AT content is not always the major driving force in determining repeat levels in bacterial genomes.

Next month, the ‘method’ of genome comparison discussed will take into account

rRNA operons in prokaryotic genomes. The number of rRNA operons per genome varies from one to 13, and there is some occasional heterogeneity within the rRNA sequences of an organism (Coenye & Vandamme, 2003).

Supplemental web pages

Several hundred supplemental web pages have been generated to display various aspects of AT content for each sequenced genome discussed in this article. They can be accessed from the following url: <http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/genomeUpdate003/>

Acknowledgements

This work was supported by a grant from the Danish National Research Foundation.

David W. Ussery and Peter F. Hallin

Center for Biological Sequence Analysis, Department of Biotechnology, Building 208, The Technical University of Denmark, Lyngby, DK-2800, Denmark

Correspondence: David W. Ussery (dave@cbs.dtu.dk)

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M. & Aksoy, S. (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat Genet* **32**, 402–407.

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M. & 40 other authors (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147.

Coenye, T. & Vandamme, P. (2003). Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol Lett* **228**, 45–49.

Oshima, K., Kakizawa, S., Nishigawa, H. & 8 other authors (2003). Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat Genet* **36**, 27–29.

Pedersen, A. G., Jensen, L. J., Brunak, S., Staerfeldt, H. H. & Ussery, D. W. (2000). A DNA structural atlas for *Escherichia coli*. *J Mol Biol* **299**, 907–930.

Pridmore, R. D., Berger, B., Desiere, F. & 12 other authors (2004). The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. *Proc Natl Acad Sci U S A* Epub ahead of print, 10.1073/pnas.0307327101

Rendulic, S., Jagtap, P., Rosinus, A. & 10 other authors (2004). A predator unmasked:

life cycle of *Bdellovibrio bacteriovorus* from a genome perspective. *Science* **303**, 689–692.

Westberg, J., Persson, A., Holmberg, A., Goesmann, A., Lundeberg, J., Johansson, K. E., Pettersson, B. & Uhlen, M. (2004). The genome sequence of *Mycoplasma mycoides* subsp. *mycoides* SC type strain PG1^T, the causative agent of contagious bovine pleuropneumonia (CBPP). *Genome Res* **14**, 221–227.

DOI 10.1099/mic.0.27103-0