Base composition bias might result from competition for metabolic resources

Eduardo P.C. Rocha and Antoine Danchin

The GC content of bacterial genomes varies from 25 to 75%, but the reason for this variation is unclear. Here, we show that genomes of bacteria that rely on their host for survival (obligatory pathogens or symbionts) tend to be AT rich. Furthermore, we have analysed bacterial phages, plasmids and insertion sequences, which might also be regarded as 'intracellular pathogens', and show that they too are significantly richer in AT than their hosts. We suggest that the higher energy cost and limited availability of G and C over A and T/U could be a basis for the understanding of these differences.

There is a wide variation of genomic guanine (G) and cytosine (C) content in bacterial genomes (25–75%; Ref. [1]), which affects the codon usage and the amino acid composition of genes and proteins [2]. Genetic elements that reproduce inside the cell (chromosomes, plasmids, phages and insertion sequences [IS]) using the cell's replication machinery might be expected to have the same GC content as the host. Exceptions to this rule are thought to indicate recent horizontal transfer [3], allowing unusual GC composition to be used in the identification of pathogenicity islands in bacterial genomes [4]. Such horizontally transferred elements are then progressively altered towards the average nucleotide composition of the host genome [5].

We have explored such variations among bacteria, and among their phages, plasmids and IS, using complete genome sequences (full lists available at wwwabi.snv.jussieu.fr/~erocha/scarceGC). This includes the genome sequences of 52 bacterial species, 59 of their phages (37 dsDNA, 15 ssDNA and seven ssRNA phages), and 54 of their natural plasmids (>5 kb). We also extracted 368 IS from the genomes, using the annotation files. Except where stated, all tests are nonparametric; that is, Wilcoxon tests for pairwise comparisons and Spearman's rho for correlations.

Eduardo P.C. Rocha*
Atelier de BioInformatique, Université Pierre et Marie Curie, 12, rue Cuvier, 75005 Paris, France.
*e-mail: erocha@abi.snv.jussieu.fr

Antoine Danchin
HKU-Pasteur Research Centre, Dexter HC Man Building, 8 Sassoon Road, Pokfulam, Hong Kong.

AT content of bacterial genomes

Free-living bacteria have on average higher GC content than bacteria that are not free living, such as obligatory pathogens and symbionts ($P < 0.001$) (Fig. 1). Analyses excluding archaea (not known to be pathogenic) or considering only proteobacteria provide similar results ($P < 0.005$). Also, among low GC content firmicutes, the mycoplasmas (extracellular obligatory pathogens) have the lowest GC content, and the free-living bacilli have the highest. Among high GC content firmicutes, *Mycobacterium leprae* has a smaller GC content than *Mycobacterium tuberculosis*. Therefore, the adenine (A) and thymine (T) richness in the genomes of non-free-living bacteria seems to occur in each of the different taxonomic branches.

'Here, we show that genomes of bacteria that rely on their host for survival (obligatory pathogens or symbionts) tend to be AT rich.'

AT content of phage genomes

Phages are on average 4% richer in AT than their hosts ($P < 0.001$). This holds true for all types of phage (Fig. 2), but to different extents. Among these phages, there are elements with different replication mechanisms and different infective strategies (Box 1). When we compared phages with similar infective behaviour (RNA, virulent dsDNA and isometric ssDNA phages), we found that the increase in AT content in the phage compared with the host genome (referred to here as the 'AT deviation') is not significantly different between the phage types: +4.4% for RNA phages, +4.2% for dsDNA phages and +5.0% for ssDNA phage. This prompted us to test systematically for differences in AT deviation among phages, taking into consideration their infective behaviour. Both dsDNA phages are richer in AT than their host, but the AT deviation is smaller in the temperate (+1.4%) than in the virulent phages (+4.2%; $P < 0.03$). Among ssDNA phages, filamentous phages show an average AT deviation larger (+10.4%) than that of isometric phages (+5.0%; $P < 0.05$). Thus, the analysis of phage indicates the GC content is generally lower in phages than in their hosts, and that the amount of this decrease is similar for phages with the same infective behaviour, independent of their type. Temperate phages have smaller biases towards AT, and filamentous phages have larger biases than the average phages.

AT bias in plasmids and IS

If parasitic lifestyle underlies this nucleotide bias, one might also expect to find comparatively high AT content in other non-essential self-replicating genetic elements, such as plasmids and IS, that are often regarded as parasitic [6]. Indeed, plasmids are also richer in AT than their hosts (+2.7%; $P < 0.001$) (Fig. 2). We compared 245 plasmid genes with homologues in the host and found that they are on average 2% richer in AT ($P < 0.001$), irrespective of the host (Fig. 2). Therefore, higher AT content is not associated with a bias in the types of gene present in plasmids. IS are also significantly richer in AT than

0168-9525/02/$ – see front matter PII: S0168-9525(02)02690-2

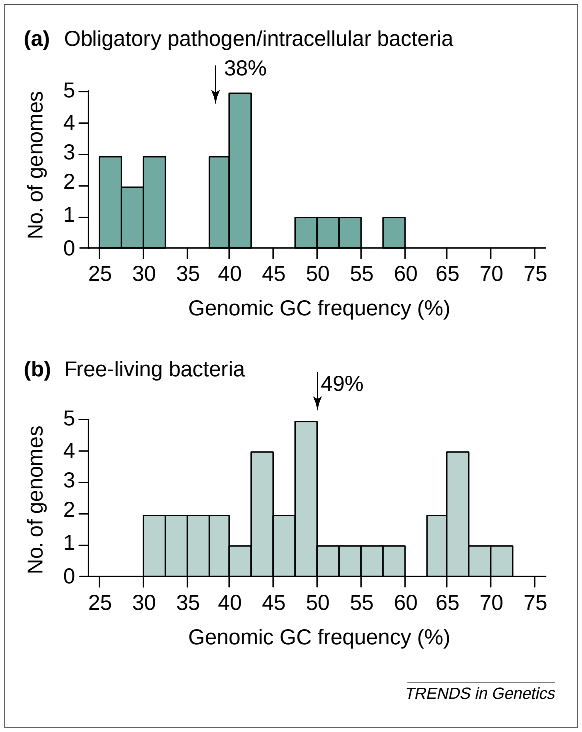

Fig. 1. Distribution of GC content in the sets of obligatory pathogen/symbiont (a) and other bacterial genomes (b) (only one strain for each species was used). Arrows indicate the mean value. It is difficult to define 'obligate parasite' unambiguously, because some bacteria (e.g. some mycoplasma) are able to grow in host-free but very elaborate media. We have labelled those bacteria that are almost exclusively associated with a host, and that are not found to reproduce outside the host, as obligate parasites. Obligate parasites that have been completely sequenced include: *Chlamydia*, *Spirochaetes*, *Mycoplasma*, *Mycobacterium leprae*, ε-proteobacteria and some other proteobacteria (*Haemophilus*, *Buchnera*, *Rickettsia* and *Pasteurella multocida*).

the rest of the genome, although the difference is smaller than for plasmids or phages (+1%; $P < 0.001$).

Explaining high AT content by means of metabolism

We propose that the comparatively high AT content of plasmids, phages, IS and bacteria that are not free-living results from the differential cost and availability of relevant metabolites in the cell (Fig. 3). First, GTP and CTP nucleotides are energetically more 'expensive' than ATP and UTP. Among pyrimidines, cytidine nucleotides follow from transamination of UTP by CTP synthetase, using one ATP. Among purines, GMP requires one more NAD$^+$ than AMP. Second, the major role of ATP in the cell's energetics makes it the most abundant of all nucleotides. As a result, very different pools of purine triphosphates (3.5 mM of ATP and 1.9 mM of GTP) as well as of pyrimidines triphosphates (2.0 mM of UTP and 1.2 mM of CTP) can be found in *E. coli* cells under exponential growth [7]. Third, unlike for the other nucleosides, *de novo* synthesis of the cytidine nucleotides is via dephosphorylation of CTP to CDP (Fig. 3). CDP can only be produced by the turnover of molecules consuming CTP in their biosynthesis, or by the catalysis mediated by nucleoside diphosphate kinase (NDK). However, the ADP:ATP ratio shifts the equilibrium of this reaction towards triphosphate synthesis, rendering dCTP synthesis very complicated [8,9]. As an example, *ndk* inactivation increases the relative levels of dCTP, which is mutagenic, resulting in a mutator phenotype [10]. However, dTTP can be synthesized from dCTP or UDP, avoiding this bottleneck.

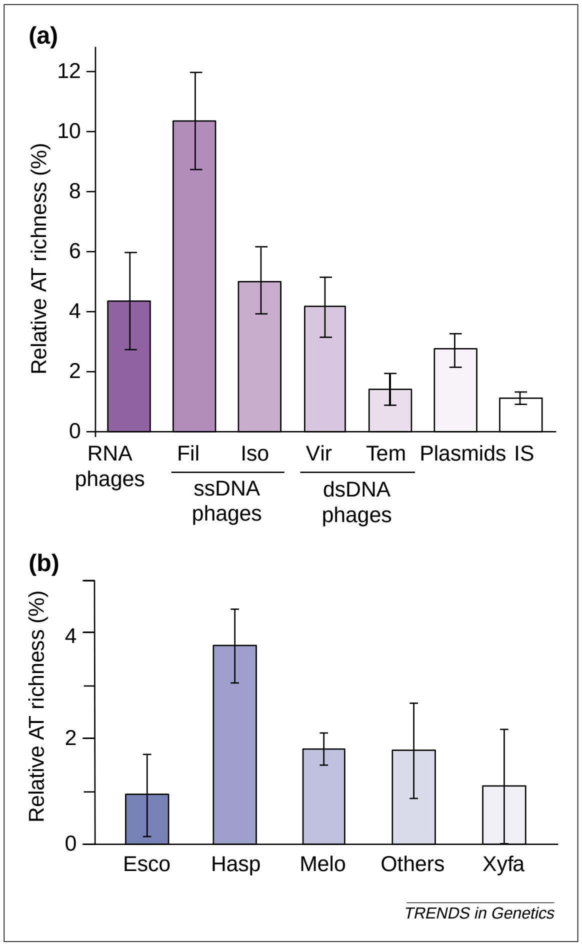

Fig. 2. AT deviations of phages, plasmids and IS relative to their bacterial hosts. (a) Average relative richness in AT content of phages (i.e. $AT_{phage}–AT_{bacterium}$), plasmids, and insertion sequences (IS) in comparison with their bacterial hosts. Fil, filamentous phages; Iso, isometric phages; Tem, temperate phages; Vir, virulent phages. (b) Higher AT content of genes in plasmids relative to their chromosomal homologues. Homologues were defined as reciprocal best hits, with more than 50% similarity in amino acid sequence and less than 20% difference in length. Alignments were done by global alignment, counting a 0-weight for gaps at both ends of the largest sequence. Species with less than ten homologues were merged in the group 'Others'. Esco, *Escherichia coli*; Melo, *Mesorhizobium loti*; Hasp, *Halobacterium* sp.; Xyfa, *Xylella fastidiosa*.

A selective framework

Could the high AT content in bacteria that are not free living be the result of selection by competition for scarce resources? We propose that indeed it might. One generally assumes that the GC content of a bacterial genome is the result of a set of mutational biases, eventually constrained by functional and ecological requirements. We propose

Box 1. Infective behaviour of phages

The infection by RNA, dsDNA virulent and ssDNA isometric phages start by bacterial cell invasion, followed by cell division inhibition and phage replication up to depletion of bacterial resources [a]. These phages only replicate horizontally and replication is typically followed by bacterial lysis. Temperate and filamentous phages are different in this respect. After invasion, temperate phages can replicate and cause cell lysis, or integrate in the chromosome and replicate synchronously with it. Such an integrated phage is called a prophage. Upon induction of the lytic cycle, the prophage is excised from the bacterial chromosome, blocks cell division and replicates. This is followed by bacterial lysis. The ssDNA filamentous phages stay in the bacterial cell in the reproductive state for large periods of time without lysis or cell division inhibition. The newly produced phage particles are continuously exported through the bacterial membrane. Hence, both temperate and filamentous phages reproduce horizontally (by infection) as well as vertically (with the bacterial cell). However, temperate phages either replicate horizontally or vertically. Instead, filamentous phages replicate in a vertical and horizontal way simultaneously and independently of the bacterial chromosome.

Reference

[a] Campbell, A.M. (1996) *Bacteriophages in* Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, H. *et al.*, eds), pp. 2325–2338, ASM Press

that bacteria that evolve to become obligatory pathogens or symbionts tend to shift from such equilibrium to become richer in AT. Thus, in a similar phylogenetic group the bacteria that are not free living tend to be richer in AT, as we observed. A similar reasoning could apply for the other elements analysed in this study: given the equilibrium GC composition of a bacteriuma, imposed by its replication machinery, a bias towards AT enrichment in phages, plasmids or IS could be selected because it would allow them to exploit the cell resources better.

> **‘We propose that the comparatively high AT content of plasmids, phages...and bacteria that are not free-living results from the differential cost and availability of relevant metabolites in the cell.’**

In bacterial genomes, it is unlikely that the individual increase in fitness of each C/G→A/T mutation could carry a sufficient advantage to allow frequent fixation. However, a mutational bias can occur if it has a selective advantage or if it hitchhikes with a selective mutation [11]. In a context of limited resources and limited metabolic capacities, the selection of a mutation bias favouring higher AT content would act on the mutation creating the bias, not on each individual C/G→A/T substitution [11]. This is valid for bacteria, but also for phages coding for elements of their replication machinery. Smaller phages, as well as plasmids and IS have a more limited control on eventual mutational biases affecting their replication. However, their smaller size renders selection at each polymorphic site more effective.

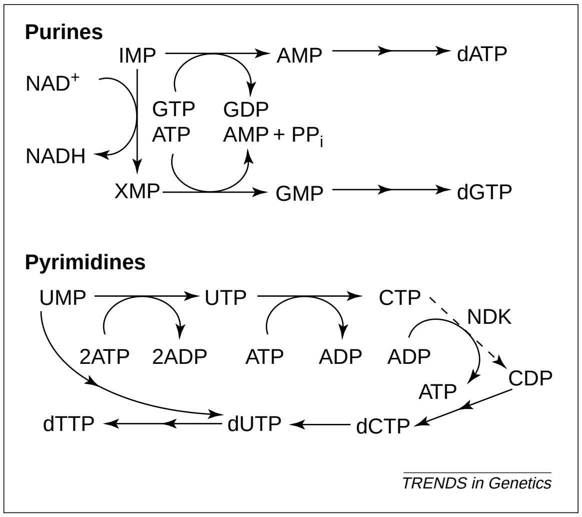

Fig. 3. Key steps of nucleotide biosynthetic pathways. Double arrows represent simplifications in the pathway. Unlike the other nucleosides, *de novo* synthesis of cytidine nucleotides follows directly from transamination of UTP by CTP synthetase, without CDP or CMP intermediates [9]. These cytidine nucleotides can only be produced by the turnover of molecules consuming CTP in their biosynthesis, such as the degradation of mRNA or the decomposition of CDP diglycerides. In many bacteria, the synthesis of the cell envelope can also participate in CMP synthesis. One should add to this list the catalysis mediated by nucleoside diphosphate kinase (NDK; dashed line in the figure). However, the equilibrium between CDP and CTP is driven by the ADP/ATP ratio in the direction of the triphosphate synthesis.

Our proposition can also shed some light on the correlation between the infection mechanism of the phage and the AT content, and in particular on the distinctive behaviour of temperate and filamentous phages. Prophages replicate vertically with the bacterial chromosome and are thus subject to ‘amelioration’ towards the host GC content. This should result in smaller AT deviations in temperate phages, as observed. Filamentous phages, however, co-exist with a living cell for a long period of time, continuously exporting phage particles and being vertically transmitted by bacterial replication [12]. Because they are not integrated as prophages, they are not subject to amelioration, but because they replicate in a replicating bacterial cell, the evolution of higher AT content will be doubly rewarded. By improving the host exploitation, filamentous phages simultaneously augment their chances of horizontal and vertical transmission. For the other phages, only horizontal transmission is improved by selection for higher AT content.

Under our hypothesis, one expects a relationship between the parasites’ progeny number (weighted by the resources each replication consumes) and the selective advantage of higher AT content. Although we lack precise data on phage burst sizes and plasmid copy numbers, one can suppose that the resources required to replicate invading phages are on average larger than for plasmids, and these in turn are larger than for IS. Because all these elements are confined in the cell and require its limited resources to grow, the relative order of GC avoidance should follow the same trend. Indeed, the observed order of ATdeviation is: phages > plasmids > IS.

A neutral framework

Neutral mutational biases have been called on to explain GC variation among bacterial genomes. In particular, the higher AT content of pathogenic bacteria could result from mutational biases operating on their small genomes, which have lost many of the DNA repair systems existing in larger bacteria [13,14]. Indeed, rank correlation between AT content and genome size is significant (0.63; $P < 0.001$). This had been indicated previously in a study of intracellular bacteria [13], and it seems also to apply for extracellular obligatory pathogens, such as mycoplasma. However, it is not clear why such bias should systematically cause higher AT content. Given the metabolic constraints we have described, the differential availability of nucleotides in the host might induce a mutational bias towards higher AT content (a similar model has been proposed for eukaryotic isochores) [15]. As resources get depleted, the relative availability of A and T increases, and mis-incorporation of these nucleotides might produce a bias towards higher AT content. Such neutral mutational bias could explain the avoidance of GC among bacteria that are not free living. In this case only, availability, rather than energetic cost, would be determinant for the direction of the bias.

A neutral bias could also explain the higher AT content of phages, because depletion of bacterial resources could lead to a systematic insertion of the more abundant A and T nucleotides. In this case, one would expect to find more biased genomes in later stages of the infection. Most types of phage deplete bacterial resources and kill the host, whereas filamentous phages do not. Therefore, under the neutral model, filamentous phages should avoid GC to a lesser extent. However, this is the exact opposite of our observations. Also, neutral mutational biases fail to account for the systematically higher AT content in IS and plasmids, which typically do not deplete the host nucleotides. Thus, although neutral mutational biases are certainly at the origin of many biases in the GC composition of genomes, they do not seem to explain our results in a fully satisfactory way.

> **'...our results could be of importance for the understanding of the evolution of AT-rich genomes in important human pathogens, such as *Plasmodium* and HIV.'**

Final remarks

Our hypothesis cannot be a basis for explaining the entire variety of GC content variation in bacteria, otherwise there would be no GC-rich bacterial genomes. Other important factors constrain the nucleotide composition of genomes and further genome research will certainly be illuminating. This hypothesis is intended to provide a common ground for a set of disparate observations, all including the tendency of certain elements to exhibit higher AT content. Its consequences on current topics of research can be significant. For example, the typically higher AT content of horizontally transferred elements [3] could be explained by their passage through AT-enriching vectors, such as phages or plasmids. Also, it can be important for the understanding of evolution of specialization of bacterial genomes suffering reductive evolution, either symbionts or pathogens [14]. For the moment, lack of reliable data and the heterogeneity of eukaryotic genomes, preclude their comparative analysis, even though the basic fundamental biochemical and ecological grounds of our hypothesis suggest extensibility to the eukaryotic world. If so, then our results could be of importance for the understanding of the evolution of AT-rich genomes in important human pathogens, such as *Plasmodium* and HIV.

Acknowledgements
Both authors are in CNRS URA 2171, Unité GGB of Institut Pasteur. We thank Isabelle Gonçalves, Hector Musto and anonymous referees for comments and criticisms.

References

1 Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. U. S. A.* 48, 582–591
2 Muto, A. and Osawa, S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution *Proc. Natl. Acad. Sci. U. S. A.* 84, 166–169
3 Moszer, I. *et al.* (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol.* 2, 524–528
4 Karlin, S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.* 9, 335–343
5 Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the *E. coli* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9413–9417
6 Levin, B.R. and Lenski, R.E. (1983) Coevolution in bacteria and their viruses and plasmids. In *Coevolution* (Futuyama, D.J. and Slatkin, M., eds), pp. 99–127, Sinauer
7 Danchin, A. *et al.* (1984) Metabolic alterations mediated by 2-ketobutyrate in *Escherichia coli* K12. *Mol. Gen. Genet.* 193, 473–478
8 Zak, V.L. and Kelln, R.A. (1978) 5-Fluoroorotate-resistant mutants of *Salmonella typhimurium*. *Can. J. Microbiol.* 24, 1339–1345
9 Danchin, A. (1997) Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP. *DNA Res.* 4, 9–18
10 Lu, Q. *et al.* (1995) The gene for nucleoside diphosphate kinase functions as a mutator gene in *Escherichia coli*. *J. Mol. Biol.* 254, 337–341
11 Sueoka, N. (1993) Directional mutation pressure, mutator mutations and dynamics of molecular evolution *J. Mol. Evol.* 37, 137–153
12 Birge, E.A. (1994) *Bacterial and Bacteriophage Genetics*, Springer-Verlag
13 Heddi, A. *et al.* (1998) Molecular characterization of the principal symbiotic bacteria of the Weevil *Sitophilus oryzae*: a peculiar G+C content of an endocytobiotic DNA. *J. Mol. Evol.* 47, 52–61
14 Ochman, H. and Moran, N.A. (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292, 1096–1099
15 Wolfe, K.H. *et al.* (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337, 283–285