Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation

Philippe Lopez, Hervé Philippe*

Équipe phylogénie, bio-informatique et génome, UMR 7622, bâtiment B, 6e étage, case 24, 9, quai Saint-Bernard, 75252 Paris cedex 5, France

Received 27 November 2000; accepted 4 December 2000

Communicated by André Adoutte

Abstract – Most prokaryotic genomes display strand compositional asymmetries, but the reasons for these biases remain unclear. When the distribution of gene orientation is biased, as it often is, this may induce a bias in composition, as codon frequencies are not identical. We show here that this effect can be estimated and removed, and that the residual base skews are the highest at third base codon positions and lower at first and second positions. This strongly suggests that compositional asymmetries result from 1) a replication-related mutational bias that is filtered through selective pressure and/or from 2) an uneven distribution of gene orientation. In most cases, the mutational bias alters the codon usage and amino acid frequencies of the leading and the lagging strand. However, these features are not ubiquitous amongst prokaryotes, and the biological reasons for them remain to be found. © 2001 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

strand composition asymmetry / GC skew / prokaryotic genomes / transcription sense / mutational bias

Résumé – Asymétries de composition des génomes procaryotes: biais mutационnel et orientation non aléatoire des gènes. La plupart des génomes procaryotes présentent une asymétrie de composition nucléotidique entre les deux brins, mais ce biais reste largement inexpliqué. Quand l’orientation des gènes est distribuée de façon hétérogène, on peut s’attendre à un biais de composition, puisque les fréquences des codons ne sont pas identiques. Nous montrons ici que cet effet peut être estimé et corrigé, et que le biais résiduel est plus prononcé en troisième base, plus faible en première et deuxième bases. Cela suggère que les asymétries de composition sont le résultat 1) d’un biais mutационnel lié à la réplication filtré par les pressions de sélection et/ou 2) d’une distribution hétérogène de l’orientation des gènes. Dans la plupart des cas, le biais mutационnel modifie l’utilisation des codons et les fréquences des acides aminés des deux brins. Cependant, ces deux phénomènes ne sont pas communs à tous les procaryotes, et leur explication biologique reste à découvrir. © 2001 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

asymétrie de composition / biais de GC / génome procaryote / sens de transcription / biais mutационnel

* Correspondence and reprints.
E-mail address: herve.philippe@snv.jussieu.fr (H. Philippe).
Les génomes procaryotes présentent généralement des asymétries à grande échelle: utilisation des codons ou composition en base et en oligonucléotides. Même si ces biais ont été utilisés avec succès pour détecter in silico les origines de réplication, les processus biologiques qui en sont responsables restent mal connus. On sait que les gènes sont préférentiellement codés sur le brin avancé, ce qui peut générer une asymétrie de composition, puisque tous les codons ne sont pas utilisés à fréquence égale. De plus, l’existence d’un biais mutationnel, pendant la réplication par exemple, reste à étudier. Nous avons donc calculé pour tous les génomes procaryotes disponibles le biais d’orientation des gènes, ainsi que le biais de GC (excès de G par rapport à C) et d’AT en distinguant les trois bases du codons et les régions intergéniques. La majorité des gènes est codée sur le brin avancé et la répartition de ces gènes est parfois remarquablement régulière, ce qui suggère une pression de sélection favorisant cette distribution. Les asymétries observées sont parfois contre-intuitives: biais plus fort en première base, ou biais en deuxième et troisième bases opposés. Nous avons développé une méthode pour estimer le biais de GC et d’AT moyen attendu, connaissant l’orientation des gènes, l’utilisation des codons et la fréquence des acides aminés. Ce biais moyen peut alors être soustrait aux biais observés. Cette analyse est détaillée ici pour deux organismes, Bacillus subtilis et Borrelia burgdorferi. La plupart du temps, le biais résiduel se révèle être de même signe pour les trois bases, mais plus fort en troisième base qu’en première et deuxième bases, ce qui est compatible avec le filtrage d’un biais mutationnel par les pressions de sélection associées à ces positions. Ce résultat suggère fortement l’existence d’un biais mutationnel favorisant la présence de G et de T sur le brin avancé, parfois suffisamment fort pour modifier l’utilisation des codons et la composition en acides aminés entre les deux brins. Chez les procaryotes, la répartition biaisée des gènes et un biais mutationnel expliquent donc totalement les asymétries observées. Il peut cependant suffire d’une de ces deux composantes pour expliquer les asymétries, comme chez les Mycoplasma, qui ne présentent pas de biais mutationnel résiduel. Les asymétries de composition semblent pourtant se maintenir, alors qu’on sait que les génomes procaryotes sont soumis à de vastes réarrangements au cours de l’évolution. L’intensité relative des pressions de sélection mises en évidence et des biais mutationnels observés, ainsi que leur cinétique, pourraient expliquer la variabilité des asymétries observées chez les procaryotes. Le séquençage des génomes complets d’espèces proches reste la meilleure approche pour répondre à ces questions.

1. Introduction

The recent achievements in complete genome sequencing have provided a wealth of material for studying large-scale strand asymmetries in genomes. Among these biases, the focus has been set on demonstrating a wide range of skews: base composition [1, 2], oligonucleotide composition [3, 4], codon usage [5], or amino acid composition [6, 7]. An important reason for this was that these analyses allow in silico prediction of replication origins [4, 7–9], which have then been confirmed in vivo [10, 11]. Yet, the processes responsible for strand asymmetries remain rather unclear, especially because not much is known about the constraints that a genome is under. Neutralist explanations have been advocated, like strand-specific mutational mechanisms during replication and repair processes [1, 12], or fluctuation of the dNTP-pool composition during the cell cycle [13]. On the other hand, selectionist arguments, like selective pressures on gene orientation, amino acid composition or codon usage can be hypothesized [14, 15]. Both kinds of explanations are likely to be true, which even puzzled Frank and Loby in a recent review, stating that “base composition skews could be influenced by differences between coding and non-coding strand because of biased gene orientation” [16]. The recent work of Tillier and Collins has shown by ANOVA approaches that the directions of replication and transcription sometimes played significant and separated roles on the observed asymmetries [17]. We provide here a simpler method that confirms and extends these results and displays graphically the previous effects.

Coding regions generally account for about 90 % of the genome in prokaryotes. This feature has two main consequences for base strand asymmetries. First, there seems to be a selective pressure for having a majority of genes transcribed in the replication sense in most prokaryotic genomes [14, 18], as will be detailed below. Although no biological reason has been demonstrated for this, except perhaps avoidance of head-to-head collisions between replication and transcription units [19], the distribution of gene orientation is skewed between the leading and lagging strand. Second, since the genetic code is degenerated, it is well known that the probability of fixing mutations is the highest when nucleotide mutations occur at third codon positions, lower at first and second codon positions. If there was an overall mutational bias [1, 12], its amplitude should thus be the highest at third base codon positions and lower at first and second ones. However, composition asymmetries are sometimes stronger in first and second codon positions in prokaryotic genomes [14].
Yet, a biased gene orientation is bound to induce strand base asymmetries in coding regions (i.e. in about 90% of the genome) on codon positions 1 and 2. For example, Leucine is the most frequent amino acid in the genome of *Bacillus subtilis*, accounting for 10.5% of the proteome. As all Leucine codons bear a U at the second position, it is likely to produce an excess of T over A at second base codon positions, since the majority of genes are coded on the leading strand. Similarly, McLean et al. [14] have suggested that an unevenly distributed codon usage would have an impact on codon position 3. Taking into account the role of gene orientation, we shall investigate the effect of amino acid frequencies and codon usage in strand compositional asymmetries of prokaryotic genomes. We will then demonstrate that once the effect of gene orientation is removed, the residual skew can be attributed to a mutational bias.

### 2. Materials and methods

For every available prokaryotic genome, we have computed raw GC skews (G - C/G + C), i.e. the relative excess of G over C, in windows of 1/50th of the genome, incremented by values of 1/240th of the genome. These computations were performed for subsets of the genome, namely the intergenic regions (GCi), and the positions corresponding to first, second and third base positions in coding regions (GC1, GC2, GC3). The same analyses were performed on AT skews, leading to AT1, AT2 and AT3. To reflect the organization of gene orientation along the genome, i.e. the colinearity of transcription and replication, we computed transcription skews (TS), as the relative excess of genes that were transcribed in the arbitrary positive sense over those transcribed in the opposite sense (see figure legends for details about the method). For better readability, these skews were all displayed in a cumulative way, meaning the raw value is integrated from the arbitrary start of the genome [2].

To study the impact of gene orientation, average expected base skews were computed by taking into account the local proportion of gene encoded in the arbitrary positive sense \( P_{\text{forward}} \) as well as the amino acid composition and codon usage. It has been shown that some organisms, like the spirochete *Borrelia burgdorferi*, harbor significantly different subsets of codon usage depending on whether the leading or lagging strand is considered [5]. Codon usage was thus globally estimated over the whole genome (global), or split into two subsets by distinguishing between leading and lagging strands (distinct). The same estimations were performed for amino acid frequencies. This led to four possible expected local base compositions, depending on whether codon usage and/or amino acid frequencies are considered global or distinct. The average local proportion of nucleotide base at pos\(^{th}\) codon position was then estimated using the following generic formula, where *forward* (backward) means leading (lagging) if computed on the leading strand, and lagging (leading) if computed on the lagging strand:

\[
P_{\text{base,pos}} = \sum_{i=1}^{61} (F_{i,\text{forward}} \cdot RFSC_{i,\text{forward}} \cdot P_{\text{forward}} \cdot \delta_{\text{base,pos}}(i) + F_{i,\text{backward}} \cdot RFSC_{i,\text{backward}} \cdot (1 - P_{\text{forward}}) \cdot \delta_{\text{base,pos}}(i))
\]

The term \( P_{\text{base,pos}} \) is the frequency of nucleotide base at pos\(^{th}\) codon position and Base stands for the complement of base. \( F_i \) stands for the frequency of amino acid corresponding to codon \( i \), and RFSC\( _i \) is the relative frequency of synonymous codon \( i \). The value of \( \delta_{\text{base,pos}}(i) \) is set to 1 if codon \( i \) has the base nucleotide at pos\(^{th}\) position, or 0 if otherwise. When global codon usage (or amino acid frequencies) is used, then *forward* and *backward* terms are set to the same global value, computed without distinguishing the strands. Corrected skews are then defined as the raw observed skews minus the expected ones, and indicated by the suffix cor followed by two letters (G or D) indicating whether amino acid frequencies and/or codon usage are considered global or distinct. For instance, GC\(_{1,\text{corG}}\) refers to G excess over C at first codon position, corrected by distinct amino acid frequencies and global codon usage.

By removing the average effect of gene content, GG corrections should display any other bias that affects strand composition, such as a mutational bias for instance. However, it is known that both strands have distinct compositional properties in some organisms. In such a case, DG corrections should remove this effect on first and second codon positions (i.e. the positions essentially affected by amino acid frequencies), while GD corrections should do the same on third codon positions (i.e. the ones affected by codon usage). As a consequence, DD corrections should show any residual bias other than gene content and global strand properties, such as a bias correlated to the distance to the origin of replication for example [6].

### 3. Results and discussion

#### 3.1. Transcription sense in prokaryotes

While it has often been observed that genes are preferentially transcribed on the leading strand in prokaryotes, the evenness of this distribution must also be considered. In the genome of *B. subtilis* [20], gene orientation was remarkably biased, as 74% of its genes were transcribed on the leading strand, and the transcription (TS) skew displayed a very steady trend (figure 1). Such an organization was unlikely to have occurred by chance alone, and some selective pressure had to be responsible for it, all the more so as quite a few other organisms display the same kind of organization. For instance, the genome of *Mycobacterium tuberculosis* also displayed a very regular TS skew, although with only 59% of its genes on the leading strand (figure 1). Similarly, the spirochete *Borrelia burg-
dorferi [21] showed only minor irregularities on its TS skew (figure 1). However, such a selective pressure seemed either very weak or absent in many other prokaryotes. For instance, no clear-cut trend could be found in the TS skew of the genomes of Chlamydia trachomatis MoPn [22] or Thermotoga maritima [23], even if they transcribe more than 50% of their genes on the leading strand. In some cases, such as the archaean Pyrococcus furiosus [24], the majority of the genes were transcribed on the lagging strand and TS skew was rather irregular. An interesting feature about closely related genomes, such as the Chlamydia or Pyrococcus species, is that they can display very different TS skews, confirming that massive genomic rearrangements occur in prokaryotes. Gene orientation in prokaryotic genomes has thus to be described with at least two parameters: the percentage of genes transcribed in the replication sense and the regularity of this distribution, as seen on TS skews. As far as could be told from the available genomes, there did not seem to be a correlation between these two factors (figure 1). For instance, the genomes of Helicobacter pylori [99] and Mycobacterium tuberculosis displayed a similar percentage of genes transcribed in the replication sense, although their TS skews were quite different. As for a possible selective pressure for having genes transcribed in the replication sense, this was likely true for most, but not all, prokaryotes (especially Gram-positive ones).

3.2. The case of B. subtilis

The genome of B. subtilis has already been thoroughly analyzed, and we focused on base skews at different codon positions. GC skew was found to be the highest, while GC1 and GC3 skews were less but similarly pronounced (figure 2A). The most striking feature was that second codon positions were oppositely biased, harboring more C than G on the leading strand. AT skews were quite regular as well, but AT1 was the most pronounced, while AT2 and AT3 displayed lower amplitudes. Unexpectedly, AT3 exhibited the opposite behavior, being negative on the leading strand. An overall mutational bias thus could not at first sight explain the raw excess of G over C in the B. subtilis genome, as the order of the amplitudes of the skews was GC1 > GC3 > 0 > GC2 [14].

The corrections we designed (see Materials and methods) allow the removal of the effect of gene orientation, which is generally biased (figure 1). GC1corGG was reduced to a 15-fold lower amplitude, and GC2corGG was reestablished to the same sign as GC1corGG and GC3corGG (figure 2B). Because the degeneracy of the genetic code mainly concerns third base codon positions, GC3 was rather unaffected by the correction. Moreover, corrected GC skews still displayed regular trends, whose order was GC ~ GC3corGG >> GC1corGG ~ GC2corGG. Such a feature is fully congruent with the hypothesis of an overall mutational bias related to replication (favoring G over C on the leading strand) that is filtered in the coding regions through the genetic code and the selective pressure acting on proteins. In the case of B. subtilis, this mutational bias was weaker than the average base skew due to gene orientation.

Corrections on AT skews produced similar results. The very marked AT1 skew, which argued for an excess of A over T on the leading strand, was reduced to mere stochastic noise after correction. AT1corGG and AT2corGG displayed a residual skew, but they were reduced by an approximately two-fold factor. Considering the low amplitudes, no decisive conclusion could be drawn about a mutational bias on A and T nucleotides in the B. subtilis genome. Surprisingly, AT1 and AT3corGG displayed opposite trends, which we shall discuss later on.

The correction can be refined in order to take into account differences between strands for amino acid frequencies and codon usage. Skews on codon positions 1 and 2 were reduced to stochastic noise by DG corrections, i.e. with distinct amino acid frequency subsets (figure 2C). This was not so surprising since amino acid frequencies are to a great extent related to those positions. Similarly, skews on codon position 3 were reduced to an extremely low amplitude by GD corrections (figure 2D), as this subset of position is mostly concerned by codon usage. When both kinds of corrections were applied (figure 2E), no residual trend could be observed for any codon position. Such results strongly suggest that there is indeed a strong mutational bias in the B. subtilis genome, at least for nucleotides G and C, and that this bias modifies the amino acid frequencies and the codon usage of the two strands. Most of all, DD corrections showed that strand composition asymmetries in this genome were fully explained by the combined effects of nonrandom gene orientation and a mutational bias that is filtered through selective pressure.

3.3. The case of Borrelia burgdorferi

The same analyses were performed on the genome of B. burgdorferi [21], for which codon usage is significantly different between the two strands [5]. In this case, the order of raw base composition skews could be interpreted as the result of a mutational bias only (figure 3A). Yet, some incongruities challenged this interpretation: GC1 amplitude similar to that of GC3 and considerably higher than GC2, no skew for AT1 whereas one was displayed by AT2. However, once the average effect of gene orientation was removed (figure 3B), GC1corGG was brought to a much likelier level as well as GC2corGG while AT1corGG was slightly reduced and also smoothed. AT1corGG displayed a pronounced and regular residual trend, where AT1 had none. Moreover, the correction recovered GC2corGG >> GC1corGG ~ GC3corGG and AT3corGG >> AT1corGG ~ AT2corGG, which was again compatible with the hypothesis that was expressed for B. subtilis. In this case the mutational bias seemed to affect both GC and AT skews, nucleotides G and T being the most numerous on the leading strand. As for B. subtilis, refining the correction showed a mutational bias in this genome that altered both codon usage and amino acid frequencies on the two strands (figure 3C-E). Once these effects were removed,
Figure 1. Transcription skew (TS) is computed as the relative excess of genes transcribed in the arbitrary positive sense over those transcribed in the opposite sense.

All skews are measured in a sliding window of 1/50th of the genome for smoothness of the curves, and this window is incremented by 1/240th of the genome for precision. For better readability, skews are displayed in a cumulative way, meaning the values are integrated along the genome. For instance, when the curve goes upwards, it means that locally there are more genes transcribed in the positive sense than in the negative sense. For comparison, the origin of replication is chosen as an arbitrary start for the genome. Here displayed prokaryotic genomes are those whose origin of replication is known. The name of the organism is followed by the percentage of codons transcribed in the replication sense.
no residual trends could be observed. For these two prokaryotic genomes, we have thus shown that the raw strand composition asymmetries could be explained in coding regions by two factors: a nonrandom gene orientation combined with genetic code, amino acid and codon usage constraints, and an overall mutational bias, be it GC and/or AT.

3.4. Asymmetries in intergenic regions

However, intergenic regions, which account for approximately 10 % of the prokaryotic genomes, also displayed strand asymmetries, for which we have little explanation. If intergenic regions were constraint-free, they should be exposed to mutational bias only and GC, and AT, would be expected to display biases similar to (or even greater than) GC_{3corGG} and AT_{3corGG}, respectively. This was roughly observed in the genome of B. burgdorferi, even if AT_{3corGG} had a higher amplitude than AT, In contrast, it was clearly rejected in the genome of B. subtilis, as AT, displayed a pronounced skew opposite to AT_{3corGG}, and GC_{3corGG} displayed a lower amplitude than GC. Such results are in agreement with the fact that intergenic regions are also exposed to heavy functional constraints that yield strand asymmetries, such as Shine-Dalgarno sequences, promoters or terminators. For example, the fact that the Bacillus genome is eight times richer in transcription terminators than that of Borrelia (Claude Thermes, pers. com.) might explain why skews in intergenic regions and third base positions are so different in the genome of Bacillus. In this respect, the skews are much more pronounced in the 25-bp extremities of intergenic regions, where most terminators are supposed to be (data not shown). The extremities of intergenic regions harbor the greater part of the asymmetries, although they represent but 20 % of these regions. The inside of intergenic regions only shows a weak bias, still opposed to that of third codon position, suggesting that other but weaker constraints are at work. Consequently, a better understanding of the constraints affecting these regions is required before a similar correction can be designed [25].

4. Conclusion

We have so far extended our study to all prokaryotic genomes sequenced (the corresponding curves can be...
found at http://sorex.snv.jussieu.fr/gcskew/main.htm), as well as a few eukaryotic complete chromosomes (data not shown). For most of these genomes, the proposed method readily demonstrated an overall residual mutational bias, that behaves congruently with the selective pressures that are imposed on codon positions. However, for Mycoplasma genitalium, M. pneumoniae, and chromosome I of the eukaryote Leishmania major, the remarkable nonrandomness of gene orientation was enough to explain all the base skews, as the residues from GG corrections proved to be very weak and probably random. In this respect, the finding of the origin of replication of these two Mycoplasma genomes through GC skew [8] was due to the very pronounced nonrandomness in gene orientation, and not to a residual mutational bias.

This feature was all the more interesting as the majority of available prokaryotic genomes displayed a mutational bias (21 genomes out of 28, see Web site). Whenever clear-cut results were obtained, residual base skews showed an excess of G over C on the leading strand, and also an excess of T over A. However, this last bias was weaker than GC bias in Eubacteria, or even absent. This proved to be especially true for B. subtilis and Helicobacter pylori. In contrast, AT bias seemed more pronounced than GC bias in Archaea such as Pyrococcus species or Methanobacterium thermoautotrophicum. As the replication machinery is very different between Eubacteria and Archaea [26], it is not unlikely that the mutational bias is different as well.

We have thus demonstrated that strand compositional asymmetries in prokaryotic genomes can be readily explained by 1) an overall mutational bias and/or 2) a biased gene orientation distribution. Even if so far no biological mechanisms have been proven responsible for the mutational bias, there has recently been some controversy as to whether it could be associated with replication or transcription processes [15]. Considering that our method took gene orientation into account and revealed asymmetries that depended only on the considered strand, our results show that the transcriptional hypothesis could probably be ruled out. Moreover, the residual bias was remarkably constant along the strand, suggesting that the mutational bias is replication-associated.

Depending on the organisms, the two previous biases can show different intensities. Yet, prokaryotic genomes are known to undergo rearrangements [27] and this plasticity is likely to upset skews induced by mutational bias. An explanation for the conservation of the asymmetries could be that rearrangements that do not upset the skews (such as translocation without inversion on the same replicore), which we call synonymous rearrangements, are preferentially selected. This is, for example, the case between the Pyrococcus abyssii and P. horikoshii genomes, where only one nonsynonymous rearrangement has been demonstrated. Interestingly, the mutational bias had the AT3 skew restored, while GC3 was still not, meaning that in this archaean the AT bias was greater than the GC bias (unpublished data). However, little if anything is known about the relative kinetics of genomic rearrangements and of the mutational bias. If the latter is faster, then some compositional asymmetries will be observable. In some organisms, such as the four completely sequenced Chlamydia, a high plasticity has been shown [22], but the corresponding genomes still display very pronounced GC skews, meaning that the mutational bias has had enough time to settle. However, it is not unlikely that some genomes have too high a plasticity for asymmetries to be observable, even if a mutational bias exists. For instance, the genomes of Synechocystis sp. and Methanococcus jannaschii did not conserve the major part of their operon structures [28], suggesting a very high rate of genomic rearrangement, and they are also remarkable for the absence of clear trends in strand composition asymmetries. The understanding of the relative kinetics of rearrangements and mutational bias will require the analyses of numerous closely related genomes.

Acknowledgements. The authors would like to thank A. Viari and E. Rocha for helpful discussions and ideas about this paper, and D. Moreira and A. Germot for careful reading of the manuscript.

References


