Bacterial DNA strand compositional asymmetry

Samuel Karlin

Until recently, it was generally thought that long sequences of DNA in most genomes were approximately equal in their intrastrand ratios of G to C nucleotides and of A to T nucleotides[1]. However, there is now substantial evidence for a prevalence of G in excess of C in the leading strand relative to the lagging strand in the genomes of *Escherichia coli*, *Bacillus subtilis*, *Mycoplasma genitalium*, *Chlamydia trachomatis* and *Rickettsia prowazekii*, and marginally in *Haemophilus influenzae*, *Mycoplasma pneumoniae* and *Helicobacter pylori*[2–4]. However, such asymmetry is not observed in the genome of the cyanobacterium *Synechocystis*, nor in the genomes of the Archaea *Methanococcus jannaschii*, *Methanococcus thermoautotrophicum*, *Archaeoglobus fulgidus*, *Pyrococcus horikoshii* and *Pyrobaculum aerophilum*. Also, several eukaryotic chromosomes and long genomic regions, including the entire *Saccharomyces cerevisiae* genome, the chromosomes of *Caenorhabditis elegans*, the human T-cell-receptor β locus (670 kb on chromosome 7), a region of the human *BRCA2* gene (780 kb on chromosome 14) and the 340-kb bithorax region of *Drosophila melanogaster*, show no distinctive strand asymmetry.

Sources of asymmetry

What are the possible sources of strand compositional asymmetry? Is this asymmetry common or special, constant or variable over bacterial genomes? Why is the asymmetry seen in G versus C content but not in A versus T content? Lobry[2] primarily attributes strand compositional asymmetry to differences in replication, mutation and repair biases in the leading versus the lagging strand. Francino and Ochman[5] emphasize a mutational bias associated with transcription-coupled repair mechanisms and deamination events (the substitution of C with T in coding sequences). Other potential sources of strand asymmetry include enzyme and architectural asymmetries at the replication fork and in replication processivity[6], intergenic differences in signal or binding sites in the two strands[4], differences in gene density coupled with amino acid and codon differences between the two strands, and dNTP-pool fluctuations during the cell cycle[7]. Additional factors that could contribute to strand asymmetry relate to gene function, gene expression level, operon organization and differences in single-base or context-dependent mutations. It appears likely that there is no single cause of strand compositional asymmetry.

GC skew and *oriC*

Compositional asymmetries are assessed in a given strand by the GC skew in a sliding window of 10, 20 or 50 kb length; the value of the skew is given by $(n_C - n_G)/(n_C + n_G)$, where n_C (n_G) is the number of C (G) nucleotides in the relative window. Rocha and colleagues[8] introduced a new method to assess strand asymmetry using a statistical linear discriminant function (Box 1). They observed compositional asymmetries between genes on the leading strand versus those on the lagging strand at the level of nucleotides, codons and amino acids. The GC skew switches sign at the origin and terminus of replication in those bacteria possessing a single origin of replication (*oriC*) that is subject to bidirectional replication. The GC switch was first used by Lobry[2,9] to locate *oriC* in *M. genitalium*. Using this method, *oriC* has now been mapped in *R. prowazekii*, *Treponema pallidum* and *C. trachomatis*[10–12]. The GC-skew method has also been applied to the linear *Borrelia burgdorferi* chromosome, where compositional asymmetry was also found and the presence of *oriC* determined, then confirmed experimentally[13].

Higher-order oligonucleotides

Are there also compositional asymmetries between the two strands in dinucleotides and in longer oligonucleotides? In this context, the *E. coli* Chi octamer sequence, GCTGGTGG, which promotes recombination in association with the recBCD complex, is statistically overabundant and skewed towards the leading strand. Specifically, >75% of the Chi motifs in the *E. coli* genome are found in the leading strand[14]. By contrast, the relative abundances of all the dinucleotides are approximately equal with respect to the leading and lagging strands in all prokaryotic and eukaryotic genomes[4,15].

The constancy of the dinucleotide relative abundances in the leading and lagging strands is consistent with the constancy of the genome signature[16]. Generally, trinucleotide and higher-order relative abundances are correlated with dinucleotide relative abundances[16]. The mechanisms generating the asymmetry between the two DNA strands seem to operate at the level of individual nucleotides

S. Karlin is in the Dept of Mathematics, Stanford University, Stanford, CA 94305-2125, USA.
tel: +1 650 723 2204,
fax: +1 650 725 2040,
e-mail: fd.zgg@forsythe.stanford.edu

0966-842X/99/$ - see front matter © 1999 Elsevier Science Ltd. All rights reserved. PII: S0966-842X(99)01541-3

Box 1. Statistical linear discriminant analysis[a]

Statistical linear discriminant analysis was carried out by Rocha and colleagues[a] as follows. Two populations of vector points $\underline{x}^{(k)}=(x_1^{(k)}, x_2^{(k)},\ldots, x_n^{(k)})$ (where $x_i^{(k)}$ refers to the i^{th} individual nucleotide, codon or amino-acid frequency of the k^{th} gene), correspond to genes from the leading and lagging strand, respectively. A linear function is constructed:

$$F\left(\underline{x}^{(k)}\right) = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i^{(k)}$$

[where $n=4$ (when considering nucleotides), 61 (when considering codons) or 20 (when considering amino acids)], that seeks to discriminate ‘optimally’ between the two populations in the manner that $F(\underline{x}^{(k)})>0$ for the genes on the leading strand, whereas $F(\underline{x}^{(k)})<0$ for genes on the lagging strand. This is carried out for every potential origin of replication (*oriC*) of the chromosome, with the termination point (*ter*) situated ~180° opposite the assumed *oriC* position. Accordingly, for each *oriC* and *ter* specification, we obtain the distribution of $\{G_1\}=\{y^{(\text{lead})} = F(\underline{x})\}$ and $\{G_2\}=\{y^{(\text{lag})} = F(\underline{x})\}$ for leading and lagging strand genes. Let $y_i^{(*)}$ be the average of $x_i^{(k)}$ over the genes of G_1 and let $y_i^{(**)}$be the average of $x_i^{(k)}$ over the genes of G_2. The $\{\alpha_i\}$ components are determined using the following formula:

$$\max_{\alpha} \frac{\left(\sum_{i=1}^{n} \alpha_i \left(y_i^{(*)} - y_i^{(**)}\right)\right)^2}{\mathrm{Var}(G_1 + G_2)}$$

where $\mathrm{Var}(G_1+G_2)$ is the variance of $\underline{x}^{(k)}$ across all the genes.

This analysis is based on R.A. Fisher’s two-population classification model[b,c] with the vector ensembles of G_1 and G_2 governed by a multivariate Gaussian distribution with a common covariance matrix. These assumptions are probably not applicable to DNA sequences, although the procedure can be used as an empirical index discriminator.

References

a Rocha, E.P.C., Danchin, A. and Viara, A. (1999) *Mol. Microbiol.* 32, 11–16

b Fisher, R.A. (1936) *Ann. Eugen.* 7, 179–188

c Anderson, T.W. (1958) *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York

and are not context dependent. It should be noted, however, that mutation is now generally regarded as being context dependent[6,17]. Strand asymmetry could exist in the distribution of certain signal sequences, probably as a result of biased selection rather than biased mutation events.

Types of asymmetry

Codon bias

Preferences within the coding sequences in almost all prokaryotes and eukaryotes are significantly biased towards G at codon site 1 and A at codon site 2 (Ref. 15). This reflects the high usage of acidic amino acids encoded by GAN (where N is any nucleotide). The predominance of G at codon site 1 is further amplified by relatively high glycine (GGN), alanine (GCN) and valine (GUN) usages[18]. Blattner *et al.*[14] have also noted compositional asymmetry in the *E. coli* genome in both non-coding and coding sequences and at all three codon sites, although somewhat less at sites 1 and 2. The extent to which codon preferences drive amino acid usages, or vice versa, is unknown[15].

Incorporation bias

Nucleotide-incorporation biases at replication are expected to be most vigorous in intergenic regions. Within gene sequences, such a bias is expected mainly at codon site 3 because codon sites 1 and 2 must adhere to amino acid requirements. Generally, transcription would not discriminate between the leading and lagging strand except, possibly, in response to gene density differences between the two strands. The general agreement of orientation between replication and transcription might diminish conflict between DNA- and mRNA-polymerase processivity[19]. This correlation has been verified near the origin of replication in *E. coli*, in the whole genomes of *M. genitalium* and *M. pneumoniae*, and throughout the *B. subtilis* genome. It is not seen, however, in *H. influenzae*, nor in *E. coli* sequences located more than 100 kb from the origin of replication[4].

Orientation bias

There is a pronounced gene-orientation bias throughout the genomes of *M. genitalium* and *M. pneumoniae*, with ~85% of genes encoded by the leading strand. The base composition at codon site 3 is the same for genes from the leading and lagging strands. These observations suggest that biased gene orientation produces strand compositional asymmetry in these genomes. *B. subtilis* also has a pronounced gene-orientation bias.

The phage T4 genome, and the yeast and *C. elegans* chromosomes, do not have any compositional asymmetry. These genetic elements are known to possess multiple origins of replication (transcriptionally primed T4 replicates from multiple start sites[20]). Strand compositional asymmetry is not apparent in the genomes of organisms known to possess multiple origins of replication distributed, on average, every 40–50 kb. It has therefore been surmised that many archaeal genomes that do not show compositional asymmetry possess multiple replication origins[4].

Replication fidelity

Fijalkowska *et al.*[21] investigated the different replication fidelity of the leading versus the lagging strand in *E. coli*. Their findings suggest that lagging-strand replication on the *E. coli* chromosome could be more accurate than leading-strand replication. However, these results contradict Thomas *et al.*[7], who contend that lagging-strand replication is less accurate. Thomas *et al.* emphasize that in the lagging strand, RNA-primase ‘settings’ are required for initiating replication and producing the

Okazaki fragments, and also at completion of replication to remove the primers. These activities differ from the helicase activity of the leading strand, which involves only slight discontinuities in replication processivity. However, there is currently a paucity of data with which to discern biases in the primer sequences. There may also be different kinds of error associated with template–primer alignments and additions. Moreover, architectural asymmetry at the replication forks could give rise to different types of lesions on the two strands. During replication, possible differences in mutation biases and mutation rates throughout the genome could contribute to the creation of strand compositional asymmetry.

dNTP-pool variations

There is considerable variation in dNTP-precursor pools and dNTP levels, creating a variety of dNTP imbalances. An imbalance of dNTPs can induce single point mutations because the probability of correct or incorrect nucleotide incorporation is related to the amounts of the competing substrates. dNTP-pool disturbances can lead to chromosomal aberrations, mutations and repair modifications. Alterations in dNTP pools can also modulate spontaneous and induced genetic changes. It is known that transcription-coupled repair is more efficient in regions containing active genes. In this context, we might expect that, during rapid growth, the transcription template strand of highly expressed genes would maintain fidelity. Although it is known that highly expressed genes entail highly biased codon usage, this apparently does not correspond with any nucleotide bias. Imbalances in dNTP pools can influence replication and repair efficiency in both eukaryotes and prokaryotes. However, it is not known how, or whether, the dNTP-pool concentrations differ during replication in different genomes.

Methods of detecting *oriC*

To detect more-complex compositional strand asymmetries, Rocha and colleagues[8] used a statistical linear discriminant analysis (Box 1). The method, however, does not take gene density into account and therefore offers no prediction for the location of *oriC* in *M. genitalium* or *M. pneumoniae*. Also, this method does not use information on variation in intergenic sequences on the leading and lagging strands. The order of genes around the genome is not considered, whereas gene order is taken into account with GC-skew analysis. The description ‘universal’ in the title of the paper by Rocha and colleagues[8] is somewhat curious as *oriC* was detected in only nine out of 15 genomes. However, Rocha's method has the advantage that the optimal α_i components incorporate refined information on the contrasts between the leading and lagging genes. The GC-skew method has the drawback that the judgement that a skew exists is in the eye of the beholder, whereas the discrimination method is more objective.

Highly expressed leading-strand genes

Genes that deviate strongly in codon usage from the average gene but are relatively similar in codon usage to ribosomal protein (RP) genes are viewed as highly expressed[22,23]. In most prokaryotic genomes, highly expressed genes include those encoding RPs, transcription and translation processing factors, chaperonins, degradation proteases and the principal proteins of energy metabolism. For example, in the genomes of the fast-growing *E. coli*, *H. influenzae* and *B. subtilis*, many glycolysis genes are highly expressed. In *Synechosytis*, several photosystem II genes are highly expressed, and in methanogens such as *M. jannaschii* and *M. thermoautotrophicum*, many highly expressed genes are essential for methanogenesis. A gene is referred to as alien if the codon-usage difference from the average gene exceeds a high threshold, and if the codon-usage difference from RP genes is also high. Alien genes are mainly open reading frames of unknown function, but can also include transposases, prophage genes and restriction/modification enzymes[22,23].

Highly expressed RPs are encoded predominantly from the leading strand. Remarkably, this finding applies to all the currently available complete genomes that have a single bidirectional origin of replication. Consequently, we can modify the method of Rocha and colleagues[8] as follows: consider an *oriC* and its antipodal termination point; count the number of RP genes encoded by the leading strand; choose the origin and termination points that maximize the number of RP genes encoded by the leading strand. This method works very well and has located the origin quite accurately in all cases to date. Thus, the mapping of *oriC* in *H. pylori* to genome position ~110 000 implies that all highly expressed RPs of this genome are encoded from the leading strand. By contrast, all specifications of *oriC* location in the genomes of *M. jannaschii*, *A. fulgidus*, *P. horikoshii*, *P. aerophilum*, *Synechosystis* and *A. aeolicus* indicate that at least two highly expressed RP genes are encoded by the lagging strand. We interpret this to mean that these genomes possess multiple replication origins. In most cases, highly expressed genes are overrepresented on the leading strand and alien genes are encoded mainly by the lagging strand.

References

1 Fickett, J.W., Torney, D.C. and Wolf, D.R. (1992) *Genomics* 13, 1056–1064

2 Lobry, J.R. (1996) *Mol. Biol. Evol.* 13, 660–665

3 Freeman, J.M. *et al.* (1998) *Science* 279, 1827–1828

4 Mrázek, J. and Karlin, S. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 3720–3725

5 Francino, M.P. and Ochman, H. (1997) *Trends Genet.* 13, 240–250

6 Kunkel, T.A. (1992) *BioEssays* 14, 303–308

7 Thomas, D.C. *et al.* (1996) *Mol. Cell. Biol.* 16, 2537–2544

8 Rocha, E.P.C., Danchin, A. and Viari, A. (1999) *Mol. Microbiol.* 32, 11–16

9 Lobry, J.R. (1996) *Science* 272, 745–746

10 Anderson, S.G. *et al.* (1998) *Nature* 396, 133–140

11 Fraser, C.M. *et al.* (1998) *Science* 281, 375–388

12 Stephens, R.S. *et al.* (1998) *Science* 282, 754–759

13 Picardeau, M., Lobry, J.R. and Hinnebusch, B.J. *Mol. Microbiol.* (in press)
14 Blattner, F.R. *et al.* (1997) *Science* 277, 1453–1462
15 Karlin, S. and Mrázek, J. (1996) *J. Mol. Biol.* 262, 459–452
16 Karlin, S. and Burge, C. (1995) *Trends Genet.* 11, 283–290
17 Echols, H. and Goodman, M.F. (1991) *Annu. Rev. Biochem.* 60, 477–511
18 Karlin, S., Blaisdell, B.E. and Bucher, P. (1992) *Protein Eng.* 5, 729–738
19 Brewer, B.J. (1988) *Cell* 53, 679–686
20 Mosig, G. and Colowick, N. (1995) *Methods Enzymol.* 262, 587–604
21 Fijalkowska, I.J. *et al.* (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 10020–10025
22 Karlin, S., Campbell, A.M. and Mrázek, J. (1998) *Annu. Rev. Genet.* 32, 185–225
23 Karlin, S., Mrázek, J. and Campbell, A.M. (1998) *Mol. Microbiol.* 29, 1341–1355

Bacterial DNA strand compositional asymmetry: Response

Eduardo P.C. Rocha, Antoine Danchin and Alan Viari

The main point of this response is to state the primary aim of our paper[1]: our purpose was not to develop yet another method of locating the *oriC* site based on the GC bias (there are already many excellent methods for this) but to study the effect of this bias on genes. The comparison of our approach with the GC-skew plot, or related techniques, is therefore inappropriate – how can a skew plot predict gene orientation from the protein sequence? It is not a question of us being more (or less) 'objective' as Sam Karlin states; he is referring to a different aim. We would like to address some of his comments specifically.

Firstly, he states that our method 'does not use information on variation in intergenic sequences'. Of course it does not, because we were interested in the genes. Nevertheless, it is simple to modify our approach to enable us to work with fragments of chromosomes instead of genes, although an 'in-frame' description would then be impossible. We have performed such experiments to compare the amplitude of the bias in transcribed and non-transcribed regions (E.P.C. Rocha, A. Danchin and A. Viari, unpublished). Interestingly, these experiments have indicated that transcription-coupled repair is not the cause of such bias.

Secondly, he states our method 'does not take gene density into account'. Again, this is not an error but a requirement. In GC-skew analysis, the high density of genes in the replicative strand means that the skew is dependent on the global composition of genes, but we wanted to remove this effect to study only the contrast in gene composition (see Ref. 2 for a discussion of this topic).

Thirdly, his argument about the covariance matrix is actually related to the question of the optimality of the discrimination (when the condition is not met, one cannot ensure that the discrimination is optimal). However, with a 97%–99% test-set accuracy (using *Borrelia burgdorferi*), one can hardly believe that the discrimination is far from optimal. Of course, this objection could be relevant to other species and we do believe the analysis could be improved by using different discrimination techniques and criteria[3].

Lastly, the term 'universal' in the title did not refer to whether or not the bias exists in all bacterial species. It referred to the fact that, when it exists, it always has the same form in terms of nucleotides, codons and amino acids, a point that was unfortunately not even mentioned in Sam Karlin's comment. The bias could be absent as a result of differences in the DNA-repair mechanisms or simply because the genome shuffles too frequently.

In conclusion, we wish to comment on Sam Karlin's statement that 'it appears likely that there is no single cause of the strand compositional asymmetry'. In our opinion, this is still an open question. The scientific challenge is therefore to design *in silico* experiments to identify, and separate, the effects of the various contributing factors, and not to devise new prediction methods.

E.P.C. Rocha, A. Danchin and A. Viari
Atelier de BioInformatique,
Université Paris VI,
12 rue Cuvier, 75005 Paris,
France

References

1 Rocha, E.P.C., Danchin, A. and Viari, A. (1999) *Mol. Microbiol.* 32, 11–16
2 Mackiewicz, P. *et al.* (1999) *Genome Res.* 9, 409–416
3 Perrière, G., Lobry, J.R. and Thioulouse, J. (1996) *Comput. Appl. Biosci.* 12, 519–524

Coming soon in *Trends in Microbiology*

- The pathogenic potential of endogenous retroviruses, by R. Löwer
- Human contact patterns and the spread of airborne infectious diseases, by J. Wallinga, J.W. Edmunds and M. Kretzschmar
- The tip of a molecular syringe, by M. Fivaz and F. van der Goot

Don't miss these and many more articles of interest; subscribe to *Trends in Microbiology* using the form bound in this issue.

0966-842X/99/$ - see front matter © 1999 Elsevier Science Ltd. All rights reserved. PII: S0966-842X(99)01561-9