

Review

From GC skews to wavelets: A gentle guide to the analysis of compositional asymmetries in genomic data

Marie Touchon^{a,b}, Eduardo P.C. Rocha^{a,b,*}^a *Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris 6, Paris, France*^b *Unité GGB, URA CNRS 2171, Institut Pasteur, France*

Received 18 June 2007; accepted 21 September 2007

Available online 29 September 2007

Abstract

Compositional asymmetries are pervasive in DNA sequences. They are the result of the asymmetric interactions between DNA and cellular mechanisms such as replication and transcription. Here, we review many of the methods that have been proposed over the years to analyse compositional asymmetries in DNA sequences. Among these we list GC skews, oligonucleotide skews and wavelets, which among other uses have been extensively employed to delimitate origins and termini of replication in genomes. We also review the use of multivariate methods, such as factorial correspondence analysis, discriminant analysis and analysis of variance, which allow assigning compositional strand asymmetries to the different biological processes shaping sequence composition. Finally, we review methods that have been used to infer substitution matrices and allow understanding the mutational processes underlying strand asymmetry. We focus on replication asymmetries because they have been more thoroughly studied, but the methods may be adapted, and often are, to other problems. Although strand asymmetry has been studied more frequently through compositional skews of nucleotides or oligonucleotides, we recall that, depending on the goal of the analysis, other methods may be more appropriate to answer certain biological questions. We also refer to programs freely available to analyse strand asymmetry.

© 2007 Elsevier Masson SAS. All rights reserved.

Keywords: Comparative genomics; Compositional bias; Bioinformatics; Skews; Molecular evolution**1. Introduction**

The proposal by Watson and Crick [1] that DNA was structured as a double stranded helix strongly relied on an important observation by Chargaff that in double stranded DNA (dsDNA), the number of A equals the number of T and the number of C equals the number of G [2]. In the absence of selective or mutational differences between the two complementary strands of the DNA helix such equality should also hold in single stranded DNA (ssDNA) [3,4]. Indeed, the published strand of a genome often reveals nearly as many A and T and as many C and G. However, under this apparent symmetry,

DNA sequences are highly asymmetric. This is because cellular mechanisms interacting with DNA are themselves intrinsically asymmetric and thus strongly constrain the composition of each strand. Many mechanisms are involved in creating this asymmetry, among which the better studied involve gene expression or chromosome replication. Coding a polypeptide, a stable RNA or a regulatory sequence leads to an asymmetric sequence composition because in general the sequence does not have the same composition as its reverse complement. At a more global level, replication is done differently on the leading and the lagging strands and this leads to different mutational biases (reviewed in Ref. [5]). Finally, DNA is packed with signals that direct other cellular processes such as chromatin condensation [6], homologous recombination [7], chromosome segregation [8], etc. Many of these processes act asymmetrically relative to the two DNA strands and thus leave its imprint in the form of strand asymmetry.

* Corresponding author. Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris 6, Paris, France. Tel.: +33 1 44 27 65 36; fax: +33 1 44 27 63 12.

E-mail address: erocha@pasteur.fr (E.P.C. Rocha).

Historically, several reasons led to a late popularity of the analyses of strand asymmetry. First, Chargaff presented evidence suggesting that the two complementary DNA strands had similar nucleotide compositions [9]. This is most frequently true overall, but as discussed above, hides a tremendous, and most interesting, variability at the local level. Second, the unavailability of DNA sequences containing more than one gene made it difficult to study the problem at a local level. For example, clusters of pyrimidine were found to predominate on transcribed strands very early in the 1960s [10], but less local analyses only started with the availability of larger DNA sequences of eukaryotes [11], and especially with the availability of whole genomes of viruses [12], and organelles [13] in the 1990s. When the first bacterial genomes became available it was immediately clear that most of them were largely asymmetric, with the two replicating strands showing very different sequence compositions [14]. Although the exact reasons of the bias have remained elusive, it involves with very few exceptions an enrichment of the leading strand in G over C, and less frequently of T over A [14–16] (reviewed in Refs. [5,17]). Since bacterial genomes are densely packed with genes, this bias is strongly affected by gene composition. Yet, since many amino acid changes are nearly neutral, the mutational bias also changes the composition of proteins in the two strands [16,18]. Thus, even if most researchers are inclined to consider replication strand asymmetry as the result of a mutational bias that is not under selection, it will result in the change of the composition of genes and proteins and in some extreme cases most severely so. Since genomes rearrange, genes switching from one strand to the other suffer opposite biases and evolve faster [19], mostly through synonymous substitutions [20]. Strand asymmetry is also observed in eukaryotes. This can be related with transcription, in genomes such as in *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* and mammals [21,22], but asymmetry has also been related with replication in the subtelomeric regions of *Saccharomyces cerevisiae* chromosomes [23] and in the human genome [24].

A difficulty with the analysis of compositional bias results from the overlap of codes in DNA sequences [25]. Thus, an asymmetry may reveal the imprint of replication, transcription, translation, recombination, segregation, nucleoid structure, etc. Most frequently a given region of the genome will show

the overlapping imprint of many of these processes. Thus, the most appropriate method to analyse compositional asymmetry will strongly depend on the biological background and on the purpose of the research. In this article we review the methods most commonly used to study strand asymmetry. We focus on the applications associated with replication bias. Yet, most of the methods are generic and can be used to analyse other sources of local strand asymmetry. Our motivation is that in spite of the considerable literature describing strand asymmetry there is no review describing the different methods and explicating their limitations and scope. For example, one often confounds GC skews with replication bias or transcription bias, when in fact it is most often a product of both mechanisms. Also, since there is now enough genomic data to make precise analysis of the changes inducing the strand asymmetries, simply counting nucleotides is no longer a sufficient approach to the problem if one is trying to understand the mechanistic basis of the processes. In fact, we have recently shown that although the enrichment of G over C in the leading strand is a constant in nearly all genomes, the precise mutations implicated in the creation of this bias differ widely [26]. On the other hand, there is no point in using a sophisticated method if a simpler one can easily solve the relevant question, e.g. if one is only interested in identifying the origin of replication in a highly asymmetric genome. It is therefore important to realize the scopes and relevance of each method (Table 1).

2. Skews

GC skews were first used to study mitochondrial strand asymmetry. Given two nucleotides X and Y, with frequencies N_X and N_Y , their skew was defined as [13]:

$$XY_{\text{skew}} = \frac{N_X - N_Y}{N_X + N_Y} \quad (1)$$

Most frequently X is G or T and Y is C or A. Some works have also investigated IUPAC transformations of these data, e.g. for purine versus pyrimidine asymmetry this corresponds to $X = \{G, A\}$ and $Y = \{C, T\}$. Since leading strands are most often richer in G and T versus C and A, a keto/amino skew allows cumulating all replication biases. In transcription, where

Table 1
Uses, relative advantages and disadvantages of different ways of analysing strand asymmetry

	Advantages	Disadvantages	Frequent uses
Skews	<ul style="list-style-type: none"> • Simple and quick • Graphical 	<ul style="list-style-type: none"> • Not very sensitive • Hard to control for other sources of bias • No substitution spectra 	<ul style="list-style-type: none"> • Finding ori/ter • Motif finding
Signal processing	<ul style="list-style-type: none"> • Quick • Graphical 	<ul style="list-style-type: none"> • Not very simple • Hard to control for other sources of bias • No substitution spectra 	<ul style="list-style-type: none"> • Finding ori/ter
Multivariate statistics	<ul style="list-style-type: none"> • Inferential framework • Many available programs • Deals multiple variables 	<ul style="list-style-type: none"> • Requires some statistical knowledge • Usually requires knowing the replichores • No substitution spectra 	<ul style="list-style-type: none"> • Quantifying the contribution of different processes • Effect on codon and amino acid usage
Substitution matrices	<ul style="list-style-type: none"> • Substitution frequencies 	<ul style="list-style-type: none"> • Requires accurate long alignments • Paucity of available software 	<ul style="list-style-type: none"> • Identification of the mutational source of biases

it has often been found that the bias is G and A versus C and T, a purine–pyrimidine skew is more relevant. Some positions in genes reflect more strongly the mutational bias than others, e.g. because of selection on amino acid usage, and therefore the third codon position usually accumulates more signal. Several authors have analysed strand asymmetry at different codon positions and with varied transformations [14–16,27,28]. These analyses may show apparently contradictory results. For example, in *Bacillus subtilis* AT skew is positive in the third codon position, but negative overall [15]. In general, this will occur when replication asymmetry and codon and amino acid composition show conflicting trends. In fact, XY may not even be nucleotides. Gene direction [29], transcription skews [30] and CDS skews [31] have been proposed where X is a gene in the published strand and Y a gene in the complementary strand (Fig. 1b). In *Mycoplasma genitalium* the replication compositional bias is inexistent but since most genes accumulate in the leading strand, i.e. gene strand bias is very important, and these genes show a skew, the origin and termini can still be identified [29].

If one starts with a sequence and wants to investigate local asymmetry, the skew must be computed locally. The most frequent approach is to compute it in overlapping sliding windows. In this case, every window has an associated value of the skew and one plots these values against the position of the window along the chromosome (Fig. 1c). A problem associated with using overlapping sliding windows is that measures are not independent and thus most standard statistical procedures are invalid. As a result, even though there are proposals regarding the standard deviation of the skews [14], these are rarely used. Lobry also introduced an aesthetically pleasing representation

Since replication skews evolve through long periods of time, recent inversions involving switch between replicating strands can be identified by plots of cumulative skews [34,35] (Fig. 2).

All previous formulae put together transcription and replication biases. To circumvent this difficulty, one can compute a difference in skew between two data sets, e.g. genes in the leading versus genes in the lagging strand [14,20]:

$$\Delta XY_{\text{skew}} = XY_{\text{skew,lead}} - XY_{\text{skew,lag}} \\ = \frac{N_{X,\text{lead}} - N_{Y,\text{lead}}}{N_{X,\text{lead}} + N_{Y,\text{lead}}} - \frac{N_{X,\text{lag}} - N_{Y,\text{lag}}}{N_{X,\text{lag}} + N_{Y,\text{lag}}} \quad (3)$$

It is important to note that this formula measures the effect of replication bias independently of gene composition, i.e. it allows controlling for the compositional bias associated with amino acids, codons and transcription. Such control assumes that genes in both strands are identical for all other variables. This may not always be the case. For example, it has been found that essential genes tend to concentrate in the leading strand [36]. Thus, features positively associated with essentiality will also be biased between strands. This is the case of highly expressed genes, which are more likely to be essential and that also have a more biased codon usage. Thus, if one wants to control the effects of translation and transcription, one should remove highly expressed genes before computing ΔXY_{skew} . An alternative to the previous formula that cumulates the absolute values of GC and TA skews has also been proposed [37]. Note that this is not equivalent to introduce purine versus pyrimidine as X and Y in the previous formula:

$$B_I = \sqrt{\left(\frac{N_{G,\text{lead}}}{N_{G,\text{lead}} + N_{C,\text{lead}}} - \frac{N_{G,\text{lag}}}{N_{G,\text{lag}} + N_{C,\text{lag}}}\right)^2 + \left(\frac{N_{T,\text{lead}}}{N_{T,\text{lead}} + N_{A,\text{lead}}} - \frac{N_{T,\text{lag}}}{N_{T,\text{lag}} + N_{A,\text{lag}}}\right)^2} \quad (4)$$

of skews [32] based on an original idea by Mizraji and Ninio [33]. This analysis is made nucleotide by nucleotide and yields a walk in Cartesian coordinates following the rules: if ‘G’ $y = y + 1$, if ‘C’ $y = y - 1$, if ‘T’ $x = x - 1$, if ‘A’ $x = x + 1$ (Fig. 1d). The advantage of this representation is that it allows simultaneously looking at the four nucleotides. The disadvantage is that it is less intuitive to find the origins and terminus of replication because these have to be searched in a 2-D space. This may justify its relatively lower popularity.

Skews only clearly indicate the origins and termini of replication when the biases are very strong. Thus, most researchers resort to cumulative skews [34], which provide a much clearer indication of the extremities of the replichores, because these are the minima and maxima of the graph (Fig. 1e). A cumulative skew XY_{skew}^c at the window i can be computed using the recurrence formula:

$$XY_{\text{skew},i}^c = XY_{\text{skew},i-1}^c + XY_{\text{skew},i} \quad (2)$$

These two last methods assume that the leading and lagging strands are known. Since these are usually detected using strand asymmetries there is a risk of circularity in the method. Their advantage is that they measure the intensity of replication bias, whereas simple cumulative skews represent the accumulation of transcription and replication biases and they depend on genome size. Other closely related methods have been proposed to remove transcription biases from GC skews in sliding windows [16,30].

3. Oligonucleotide biases

If there is a compositional strand asymmetry it follows necessarily that oligonucleotides will also be asymmetrically distributed between the two strands. For example, if a given strand is rich in G over C and T over A, then words composed of Gs and Ts will tend to be over-represented in that strand. Note that the inverse is not necessarily true. There may be no nucleotide skew but significant bias for some oligonucleotides.

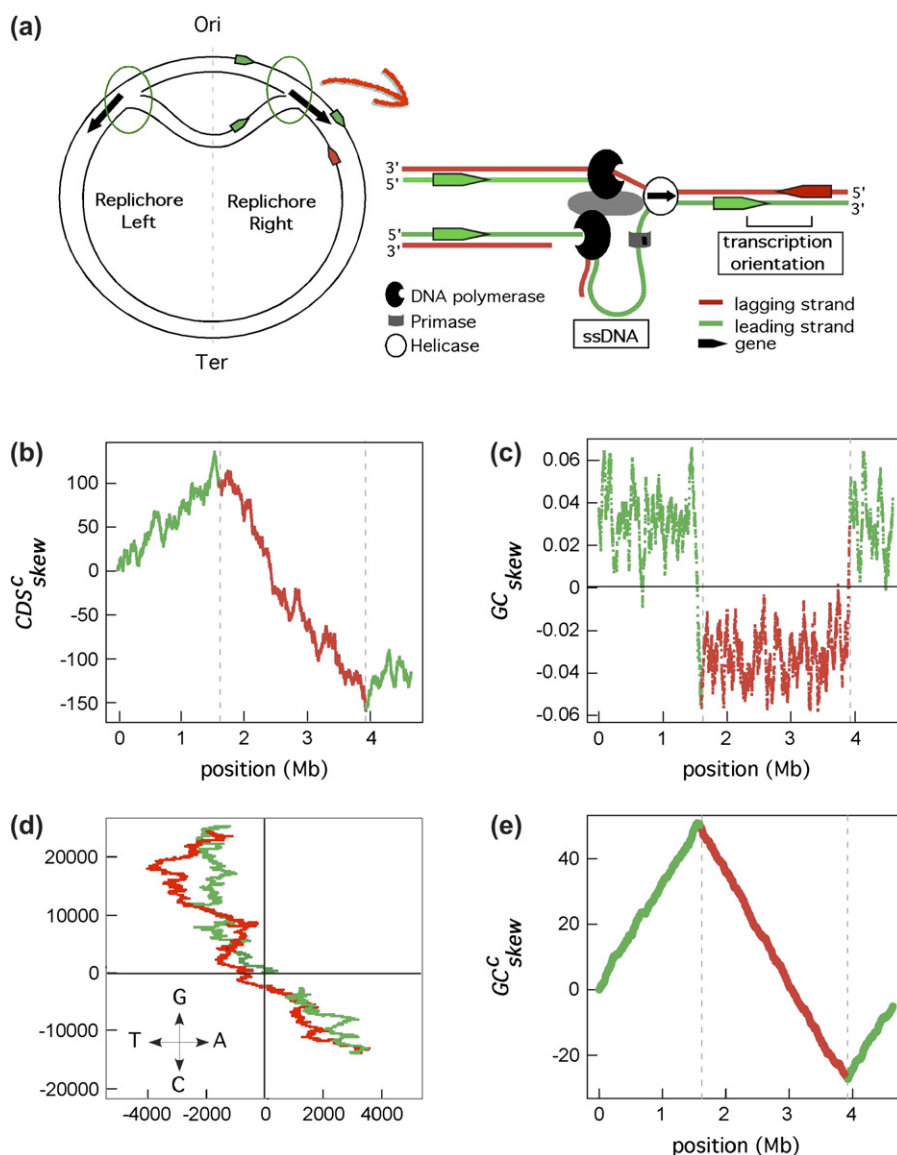


Fig. 1. Examples of skew profiles in the *E. coli* K12-MG1655 chromosome. (a) The bacterial model of replication. Replication follows bi-directionally from the origin to the terminus and at the replication fork there are several asymmetries that can account for strand asymmetry among which are: (i) ssDNA exposure while synthesising the lagging strand; (ii) dedicated DNA polymerases; (iii) transcription orientation relative to replication; and (iv) RNA primer of the Okazaki fragments. (b) Cumulated CDS skew. This curve represents the coding sequence orientation bias. The walker is gliding along the published strand of the chromosome and moves one unit up when the gene is encoded on the examined strand, and moves one unit down when the gene is encoded on the complementary strand. (c) GC skew (window size = 50 kb; step size = 1 kb). (d) 2-D DNA walk is performed by reading the sequence nucleotide by nucleotide and walking into the plane according to the four directions defined by the four bases as indicated on the bottom left of the figure. (e) Cumulative GC skew (window size = 1 kb). The origin is predicted at the maximum skew value while the terminus is predicted at the minimum in this graph. In (a–e), the dashed vertical lines correspond to the experimentally determined limits of replichores; red, lagging strand and green, leading strand.

This will often reflect selective pressures on these oligonucleotides. Typically selected oligonucleotides include optimally translated codons and regulatory signals, although codon usage is usually analysed with the multivariate techniques described in the next section. Here, we concentrate on the methods used to analyse asymmetry between generic oligonucleotides. Blattner and colleagues found several skewed octamers in *Escherichia coli*, some of which correspond to known biological motifs [38]. For example, the chi sequence of *E. coli*, which is involved in the RecBCD homologous recombination pathway, is over-represented in the leading strand because it is G rich and also

for biological reasons [7]. Thus, analysing the distribution of chi often allows determining the leading and the lagging strands (Fig. 3). Salzberg and colleagues analysed the distribution of octamers in several genomes. They made cumulated oligonucleotide skews and systematically found the origins and terminus of replication given by GC skews [39]. The length of the oligonucleotide to be used in this analysis is somewhat arbitrary and octamers are often chosen for bacterial genomes as a good compromise between specificity, the larger the oligonucleotide the better, and statistical power, shorter oligonucleotides are more abundant. Smaller sequences require the use of smaller

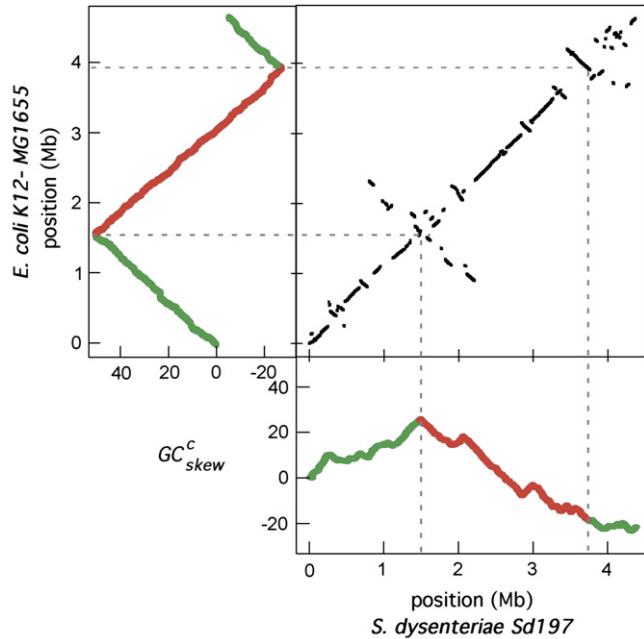


Fig. 2. Example of the effect on cumulative skew plots of the recent chromosomal inversions in the genome of *Shigella dysenteriae* Sd197 when compared with the genome sequence of *Escherichia coli* K12-MG1655. The central panel is a dot plot comparison at protein level, where each point represents the position of an orthologous pair in the respective genomes. Orthologous genes were identified as unique pairwise reciprocal best hits, with at least 80% similarity in protein sequence and <20% difference in length. External panels represent the cumulative skew plots in 1 kb sliding windows along the two chromosomes. The dashed vertical lines correspond to the experimentally determined replicore's extremities; red, lagging strand and green, leading strand.

oligonucleotides. In fact, other authors have used dinucleotides [40], tetranucleotides [41], or every oligonucleotide with size ranging from 2 to 8 [42]. For each oligonucleotide one can draw a statistical test of significance by simply accounting for the deviation from equal partition between the two strands, e.g. using a binomial test [43]. If the tests are done for every word then a correction for multiple tests must be done, e.g. a sequential Bonferroni correction. Note that these statistical analyses assume that words are independent, which they are not. Thus, if a word XYZ is very frequent YZW is more likely to be over-abundant because it overlaps with it. Detailed identification of the asymmetrically distributed oligonucleotides requires the clustering of the most skewed oligonucleotides into degenerate motifs that partially overlap [44,45].

A problem with these approaches is that they do not provide a way of putting together consistent motifs, i.e. oligonucleotides that may not overlap in sequence but that point consistently to the same bias. To circumvent this problem, Worning and colleagues proposed a weighted double Kullback–Leibler distance for each oligonucleotide [46], D_X :

$$D_X = (N_{X,\text{lead}} - N_{X,\text{lag}}) \log_2 \left(\frac{N_{X,\text{lead}} + r}{N_{X,\text{lag}} + r} \right) \quad (5)$$

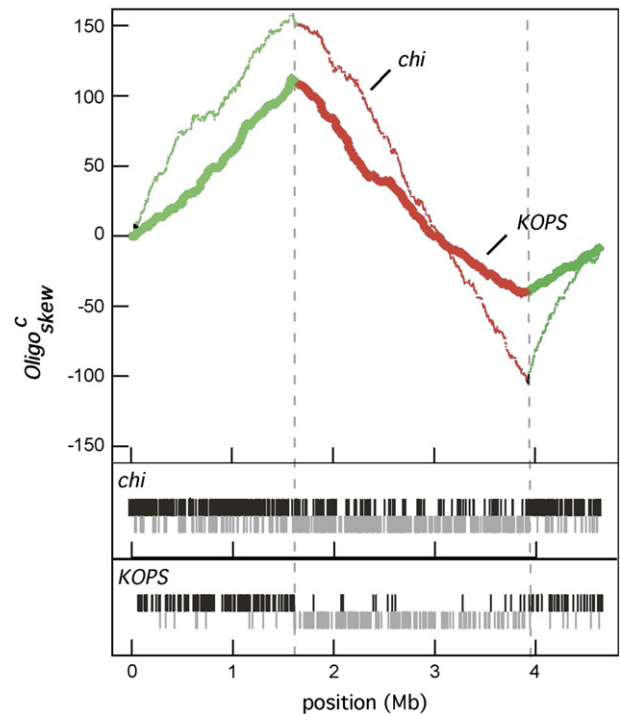


Fig. 3. DNA walks for chi sites (GCTGGTGG) and KOPS motifs (GGGNAGGG) along the *E. coli* K12-MG1655 chromosome. The walker glides along the published strand of the chromosome and moves one unit up when encountering the motif under analysis, and moves one unit down when the complement of the motif is encountered. For instance, in the *E. coli* chromosome, there are about 200 more chi sites in the direct orientation than in the reverse orientation, and this bias is inverted after the terminus. Lower panels show the distribution of the chi and the KOPS motifs. The positions of the replication origin, oriC, and dif are indicated. Each vertical bar represents a motif in the direct orientation (black upper bar) or in the reverse orientation (grey lower bar).

where r is a factor to control for low counts of the oligonucleotide X ($r = 5$ in the original proposal). To cumulate different oligonucleotides one sums to obtain the overall bias:

$$D = \sum_X D_X \quad (6)$$

To identify a point of maximal skew, e.g. the origin of replication for replication asymmetries, it suffices to compute the previous value for sliding windows along the sequence.

It is important to emphasize that the asymmetric distribution of an oligonucleotide between strands does not necessarily indicate selection for that asymmetry: it may simply reflect the matching between the composition of the oligonucleotide and the strand compositional asymmetry. While the previous methods identify the oligonucleotides strongly discriminating between strands they do not allow identifying the oligonucleotides whose asymmetry is under selection. For example, dinucleotides are differently distributed among replicating strands, but the asymmetry disappears when controlling nucleotide bias [40,47]. To assess the biological significance of the over-representation of an oligonucleotide one has to normalise by its composition, which is often done using Markov chains.

Robin and colleagues detail the different statistics that can be used when controlling for oligonucleotide content to identify differences between two sequences [48].

The search for the motifs involved in the translocation of the *E. coli* chromosome by FtsK provides an interesting recent example of a set of different methodologies coming up with the right answer. These elements were expected to show a strong replication strand skew because they were proposed to direct the sense of translocation of the bacterial chromosome towards placing the *dif* site at the septum [49]. Bigot and colleagues analysed octamer frequency taking into account the composition of the genome in heptamers and searching for the most skewed oligonucleotides [8]. Levy et al. [45] searched for highly frequent octamers showing strong replication skew around the *dif* site. Hendrickson and Lawrence [50] searched genomes for octamers that were highly frequent, highly skewed (controlling for the frequency of shorter oligonucleotides), and overabundant near the *dif* site. In the end, parallel experimental works validated the same motif [8,45].

4. Signal processing techniques

The main issue when dealing with noisy signals like the skew profile is to distinguish the local maxima and minima associated to the origins and termini of replication from those induced by the noise (e.g. associated with transcription, translation, gene orientation or stochastic effects). Signal processing techniques such as Fourier transformation (FT) and wavelet transformation (WT) analysis have been used to detect singularities in noisy skew profiles.

Recently, a noise-reduction approach using fast Fourier transform was applied to the GC skew profile [51]. This technique allows removing the noise from a signal when it is possible to estimate the frequency components that are likely to be noise rather than the signal of interest, as is the case for replication compositional asymmetries. The application of Fourier transformation is based on the existence of two regions of opposite polarity (i.e. 1 Hz in the case of the two bacterial replichores), which allows considering higher-frequency components as noise that must be removed. The advantage of this method is that it increases the prediction accuracy of the replication termini and origins for weakly biased genomes. However, this method suffers from some drawbacks: one is that FT performs poorly when the signal is not stationary (i.e. a signal that does not repeat over time); the second one is that it assumes the existence of only two inflexions of the skew (one origin and one terminus of replication), which is not the case in several archaeal and eukaryotic genomes. Finally, this method and most of the approaches we have described so far are dependent on the choice of a window size. The shape of the curves and the accuracy of the predicted sites will strongly depend on this choice: the larger the window, the less accurate the positioning of the sites, the smaller the window, the noisier the signal. Thus, for genomic analysis, computing indexes in overlapping windows of arbitrary fixed lengths may lead to the loss of important information.

All these disadvantages can be overcome by the continuous wavelet transform (WT) approach, which is a very efficient multi-scale singularity tracking technique [52]. The WT is especially useful in detecting singularities in noisy signal by examining and following the modulus maxima of the WT (WTMM) across scales (Fig. 4). The idea underlying the analysis is that these maxima indicate positions of high curvature in a smoothed version of the signal. Therefore, they indicate the presence of breakpoints. At large scales, noise amplitude is reduced and WTMM are easy to identify, although their locations are not precise. At small scales, the smoothing is weaker, and the locations are more precise. On the other hand, at smaller scales the signal-to-noise ratio becomes more important, so the maxima are harder to identify. As a result, following the lines of maxima from large to small scales allows retaining the advantages of both large- and fine-scale

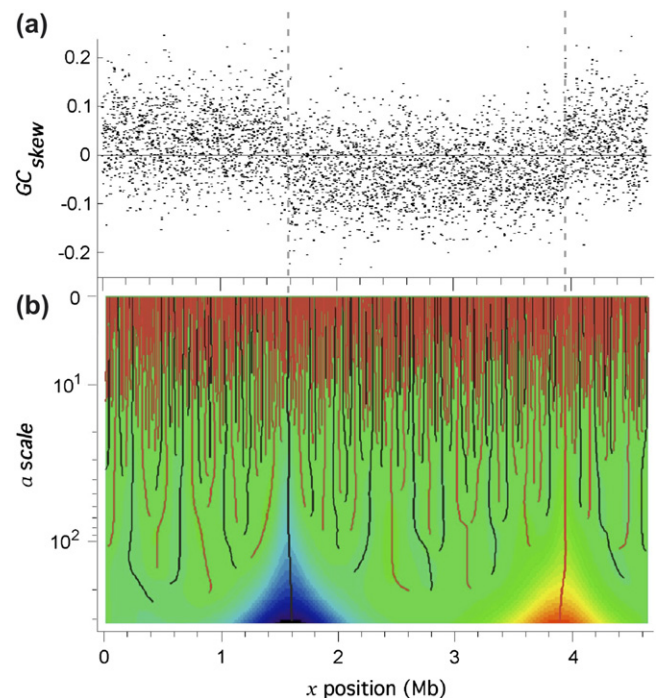


Fig. 4. (a) GC skew profile (S) along the *E. coli* K12-MG1655 genome in adjacent windows ($w = 1$ kb). (b) Space-scale representation of the GC skew profile using the analysing wavelet $g^{(1)}$ (the first derivative of the Gaussian function). Wavelet transform (WT) computed with $g^{(1)}$ is the derivative of the skew profile smoothed by a dilated version $g_a^{(0)}(x) = g^{(0)}(x/a)$ of the Gaussian function. Thus, the wavelet coefficient $T_{g^{(1)}}[S](x, a)$ quantifies to which extent, around position x over a distance a , the skew profile has a similar shape as the analysing wavelet. WT coefficient is coded using 256 colors from blue (min) to red (max); the WT skeleton defined by the set of maxima lines obtained by connecting the WTMM (WT modulus maxima) across scales is shown by thin solid lines. By looking for the maxima of the $T_{g^{(1)}}[S](x, a)$ over the space-scale half plane, the WT can be used as a multi-scale jump detector. Thus, upward (respectively, downward) jumps are identified by the maxima lines corresponding to positive (respectively, negative) values of the WT as illustrated here by the black (respectively, red) lines. The amplitude of the WTMM measures the importance of the jumps to the overall signal. It is clear that the two maxima lines of largest amplitude point to the two major jumps in the skew profile, which, respectively, correspond to the *ter* and the *ori* site.

analyses (Fig. 4). Song et al. [53] used this technique to locate the origins and terminus of replication by documenting GC skew, AT skew, keto excess and purine excess on published bacterial chromosomes and performed statistical significance tests (i.e. Monte-Carlo simulation) for the predicted loci. Other authors used this methodology on skew profiles of the human genome to disentangle the part of the strand asymmetry coming from replication from that induced by transcription. This allowed the identification of a large number of putative replication initiation zones, and to propose a model of replication with well-positioned replication origins and random terminations that accounts for the observed characteristic serrated skew profiles [24,54].

5. Statistical approaches

Multivariate statistical methods such as discriminant analysis and analysis of variance have been used to analyse compositional biases, and especially to quantify the impact of each mechanism and of each variable in each mechanism. Correspondence analysis has also been used to understand the different origins of the skews operating on biological sequences. It is out of the scope of this review to detail these methods and the reader is invited to consult standard references for discriminant analysis [55], analysis of variance [56], and correspondence analysis [57].

Correspondence analysis is mostly a descriptive/graphical method that has been used to understand the patterns of intra-genomic variance in both codon and amino acid usage [58–60]. The method is designed to analyse large contingency tables such as the ones displaying counts where rows correspond to genes and columns to codons or amino acids. For codons, many researchers have found it useful to compare correspondence analysis based on counts and relative synonymous codon usage, although that has been criticized [61]. The method finds the successive most important axes that best capture the trends in the data. If the skew is important it will show among the first axes. The most spectacular description of strand asymmetry using correspondence analysis was the one of *Borrelia burgdorferi* by McInerney [62] (Fig. 5). Because the method allows overlapping columns and rows in the same graph it allows to visually correlate the two and one can then associate codon usage and the structure of the replicore. This association, however, is not very easy to place in an inferential framework.

Some authors have used linear discriminant analysis [18] or correspondence discriminant analysis [63] to rank the discriminant power of different variables. The use of linear discriminant analysis is particularly simple in the context of strand asymmetry because the dependent variable assumes only two states, e.g. the sequence is either in the leading or in the lagging strand. In this case, the linear function is constructed as:

$$F(x) = \alpha_0 + \sum_{i=1}^n \alpha_i x_i \quad (7)$$

where α_i are the parameters to be estimated and x_i is the relative composition of gene i in terms of a given variable.

Variables can be nucleotides ($n = 4$), codons (61), amino acids (20), but also other measures such as nucleotide composition at each position of the codon (12), purines versus pyrimidines (2), etc. The appealing of this function is that the α_i associated with each variable gives an indication of the intensity of its discriminant role. Furthermore, by training the function on a subset of the data and applying it on the rest one can rigorously assess the discriminating power of the set of variables. For example, in *B. burgdorferi* this method shows a maxima of discrimination in excellent agreement with the origin of replication and shows that by simply knowing the amino acid composition of the protein one can predict with more than 95% of accuracy the position of the gene in relation to the replication strands (Fig. 5) [18]. Naturally, sets with more variables discriminate better than subsets of them (e.g. codons discriminate better than nucleotides). In line with compositional asymmetry being the result of a mutational bias, the discriminating power of nucleotides is higher than that of amino acids (Fig. 5). This method also allows identifying the origins and terminus of replication if it is used in sliding windows but since it only uses the information on replication composition bias it is less efficient than methods also including gene orientation bias. On the other hand, discriminant analysis pinpoints the most relevant variables of the bias, which is better if one aims at understanding its biological grounds.

There is often a misunderstanding in the literature between the impact of gene orientation and the replication compositional bias in GC skews. Tillier and Collins [29] have used analysis of variance to understand their relative importance in shaping strand asymmetries. They used a general linear model of the form:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad (8)$$

where y_{ijk} is the (GC or AT) skew for the k th gene at an observed level i (leading/lagging strand) and j (coding/non-coding strand), μ is the mean skew, α_i is the effect of gene direction and β_j is the effect of replication direction. The term γ_{ij} is the term of interaction between replication and transcription, independent of the additive effect. It is an often-overlooked conclusion of this work that this term was found to be not significantly different from zero in most cases [29]. Stated otherwise, replication and gene orientation biases are approximately independent and additive and can then be quantified and analysed separately. In the *Borrelia* example that has followed us through this section (Fig. 5), the effect of gene orientation is less important at third codon positions while the replication effect is most important at these positions. This is because the gene effect reflects largely amino acid composition biases whereas the replication effect reflects a mutational bias that will affect more severely the positions generating more neutral substitutions, i.e. third codon positions.

6. Matrix inference

Given the very large number of papers describing GC skew analyses it is surprising that so few works have aimed at

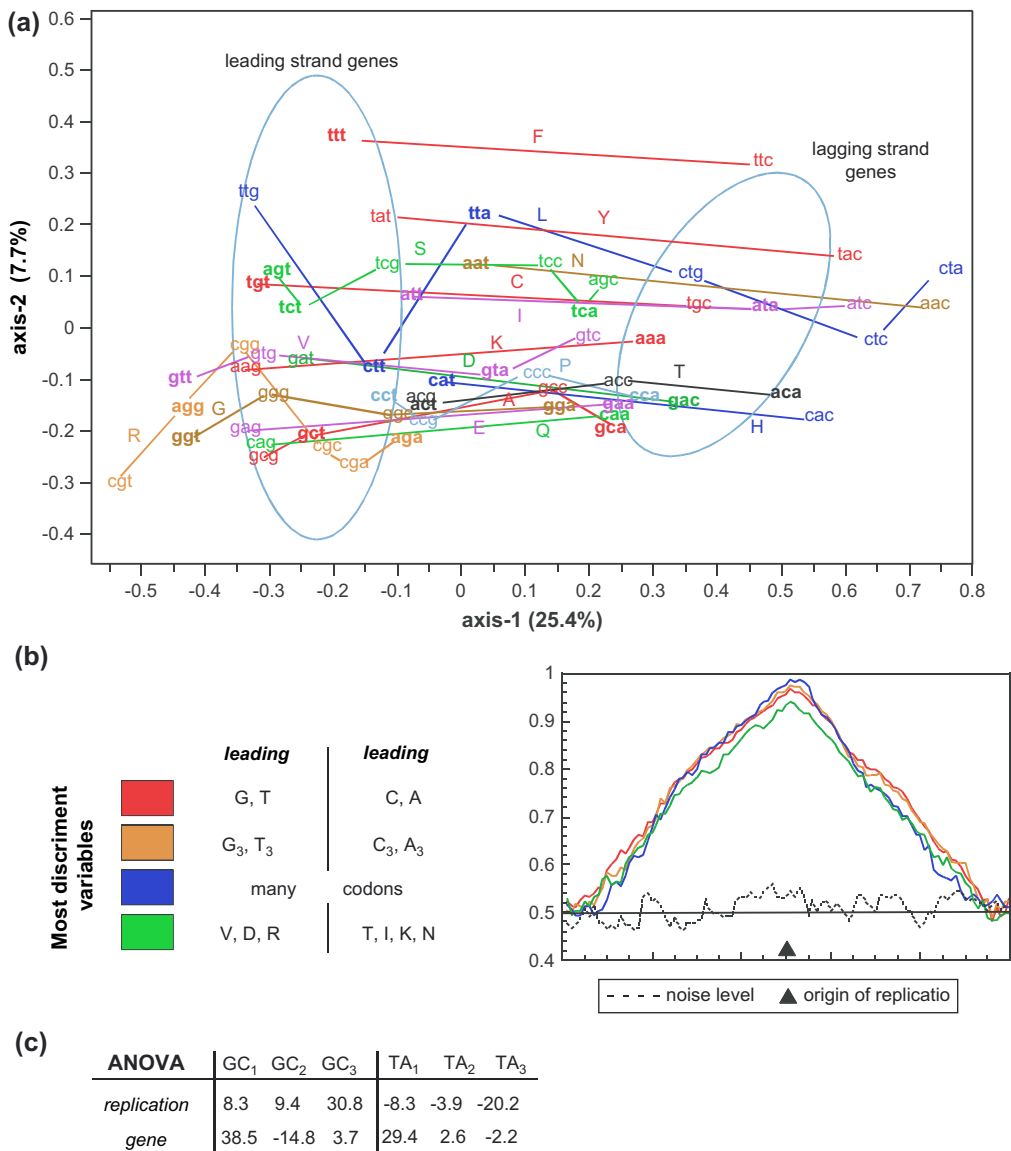


Fig. 5. Statistical analysis of the replication strand asymmetry of *Borrelia burgdorferi*. (a) Correspondence analysis inspired from Ref. [62], but made on absolute counts of codons. The first axis explains 25% of the variance and separates genes in the two leading strands. The points representing the genes were removed for clarity and were encircled by the two ellipses. Codon types are projected in the two first axes and show the association of the different codons with each of the replicating strands. (b) Results of linear discriminant analysis in sliding windows, with indication of the most discriminant variables when using representations such as nucleotides, nucleotides at each codon position, codons and amino acids [18]. (c) Coefficients of the regression analysis of Tillier and Collins [29] to determine the relative contribution of replication and transcription to GC and TA skews.

identifying the relevant mutational mechanisms. Although mutation is commonly thought of as a random process, evolutionary studies show that the different types of “mutation” occur with widely varying rates that presumably reflect biases intrinsic to replication and repair mechanisms. Strand asymmetry associated with DNA replication has been characterized in several bacterial clades [20,26,64] and mitochondria [65,66]. Strand asymmetry associated with transcription and attributable to higher rates of cytosine deamination on the coding strand has been observed in enterobacteria [67–69]. Transcription-associated strand asymmetry has also been observed in mammalian genomes. In this case, the strongest asymmetry occurs for A → G transitions, which may be a by-product of transcription-coupled repair in germline cells [21,22].

Inferring substitution matrices poses two major problems. First, if one wants to study a mutational bias then the analysis should only include neutral changes. In many genomes there are no *a priori* ways of identifying neutral changes, since intergenic regions are packed with regulatory targets and codon usage is under selection. In such compact genomes, given the small size of intergenic regions, and the difficulty in aligning them accurately, most works have analysed the fourfold degenerate positions, i.e. the third codon positions for amino acids’ fourfold degenerate [26,66]. These are under selection in highly expressed genes of fast growing bacteria and therefore such genes should be removed prior to the analysis [26]. To infer substitution matrices one has to unambiguously identify orthologous positions for which the sequence must

have diverged recently. In such weakly divergent sequences the number of polymorphisms is low and only the availability of large sequences, e.g. complete genomes, allows obtaining a sufficiently strong signal. The second difficulty with this type of analysis is that it involves inferring substitution matrices that are not time reversible, which is a very tricky thing to do by maximum likelihood methods. Thus, most works have used parsimony, which in highly biased sequences require the use of data sets with low densities of polymorphisms to avoid artifactual results [70].

The observed rate of substitution of a nucleotide X to another Y (e.g. $A \rightarrow T$) can be converted to relative substitution frequencies $f_{X \rightarrow Y}$ by dividing the number of observed substitutions by the number of nucleotides X in all the positions analysed (i.e. A in the precedent example) [71]. If one analyses fourfold degenerate codon positions, the normalization should be made according to the frequencies of nucleotides at these positions. One must emphasize that this approach is only valid as long as the probability of multiple substitutions is low. Substitution matrices can then be calculated for genes belonging to two strands, e.g. transcribed versus non-transcribed or leading versus lagging, and one can then compare directly the frequencies between different types of substitutions. This involves a statistical test for which two parametric and two non-parametric tests have been proposed.

Francino and Ochman [72] defined an odds-ratio of complementary substitutions (e.g. $A \rightarrow T$ versus $T \rightarrow A$) at either side ($5'$ and $3'$) of the origin of replication. For every pair of complementary substitutions they computed (ω):

$$\omega = \frac{f_{X \rightarrow Y(5')}/f_{\overline{X} \rightarrow \overline{Y}(5')}}{f_{X \rightarrow Y(3')}/f_{\overline{X} \rightarrow \overline{Y}(3')}} \quad (9)$$

where $\overline{X} \rightarrow \overline{Y}$ represents the complementary substitution of $X \rightarrow Y$. The ratio is close to 1 in the absence of bias, which can be easily tested assuming that the logarithm of ω is normally distributed. We have proposed a non-parametric robust, but computationally more demanding, alternative test using non-parametric bootstrap [26]. In this case, one computes $f_{X \rightarrow Y}$ in one strand and in the other (in non-overlapping segments), e.g. genes in the leading and in the lagging strands, and then compute the difference between the two to test the hypothesis:

$$H_0 : f_{X \rightarrow Y, \text{lead}} - f_{X \rightarrow Y, \text{lag}} = 0 \quad (10)$$

The frequency at which one of the rates is higher than the other among all random experiences gives the risk of rejecting the hypothesis. Thus, for example if $f_{C \rightarrow G, \text{leading}} > f_{C \rightarrow G, \text{lagging}}$ more than 99% of the times, one can refuse the H_0 hypothesis with a risk of 1%. Since one is commonly interested in computing the differences among complementary changes, one can cumulate the biases between the two strands. This involves testing the hypothesis:

$$H_0 : [f_{X \rightarrow Y, \text{lead}} + f_{\overline{X} \rightarrow \overline{Y}, \text{lag}}] - [f_{X \rightarrow Y, \text{lag}} + f_{\overline{X} \rightarrow \overline{Y}, \text{lead}}] = 0 \quad (11)$$

This hypothesis, albeit not the method, is equivalent to the one tested by Francino and Ochman [72]. Because one pools the data of a substitution and its complement the resulting statistics is more powerful, although less specific.

In a different approach one may compare the compositional asymmetry between the populations of genes coded in the two strands. In this case, statistics are based on the population of genes not in two large concatenate of leading and lagging strand gene sequences. The disadvantage of this approach is that it is less powerful from a statistical point of view because the whole data are not pooled together. The advantage is that it allows analysing outliers, e.g. genes that have atypical compositions. Rocha and Danchin [20] analysed the substitution frequencies per gene and divided the genes into two classes, leading and lagging strand genes. The differences in substitution frequencies could then be assessed by a non-parametric test of means, e.g. the Wilcoxon test [43]. Alternatively, Klasson and Andersson [64] computed substitution rates and then modelled the process using a multinomial distribution. To test if the two multinomial distributions, corresponding to substitutions in the leading and lagging strands, were significantly different they used the following formula:

$$\sum_{j=1}^2 \sum_{i=1}^k \frac{[x_{ij} - n_j \{(x_{i1} - n_{i2}) / (n_1 + n_2)\}]^2}{n_j \{(x_{i1} - n_{i2}) / (n_1 + n_2)\}} \sim \chi_{k-1}^2 \quad (12)$$

where n_j is the total number of substitutions from distribution j . Multinomial distributions are appropriate to test analyses based on counts. However, to analyse substitution bias one has to normalise the counts of substitutions by the nucleotide frequencies (as mentioned above, $f_{A \rightarrow C} = N_{A \rightarrow C} / N_A$). Since substitution rates are not counts they cannot be accounted for by a multinomial process. The authors work around this problem by multiplying substitution rates by the total number of substitutions, thus transforming them into some sort of counts [64]. This is the term x_{ij} included in the above equation.

To the best of our knowledge, only Bielawski and Gold [66] applied maximum likelihood methods to the understanding of the mutational biases shaping compositional asymmetry. They analysed the asymmetry between the L and the D strands of mitochondria by using a strand-symmetric model as defined by Sueoka [3]:

$$Q = \begin{matrix} & \begin{matrix} T & C & A & G \end{matrix} \\ \begin{matrix} T \\ C \\ A \\ G \end{matrix} & \begin{bmatrix} - & \beta & \chi & \varepsilon \\ \alpha & - & \delta & \phi \\ \chi & \varepsilon & - & \beta \\ \delta & \phi & \alpha & - \end{bmatrix} \end{matrix} \quad (13)$$

This null model has five free parameters and can be compared with models involving a symmetry break. For example, if one replaces α by two different parameters α_1 and α_2 , one allows $C \rightarrow T$ (α_1) to be different from $G \rightarrow A$ (α_2). The model with these two parameters is a generalization of the symmetrical model, which can be recovered by making $\alpha_1 = \alpha_2$. The gain obtained by including one more parameter to the model can be evaluated statistically using the classical likelihood ratio test [73]. If there is a gain, this means that the process is

Table 2
Some freely available programs to analyse strand asymmetry

Free software/package	Availability	Source code	Web sites	Functionalities
SeqinR (Oriloc) [79]	Download	R package	http://pbil.univ-lyon1.fr/Rweb/	<ul style="list-style-type: none"> • Nucleotide skews accounting for codon positions • CDS skew
GenSkew	Online and download	Java	http://mips.gsf.de/services/analysis/genskew/	<ul style="list-style-type: none"> • Simple and cumulated skew of two selectable nucleotides
GraphDNA [80]	Download	Java	http://athena.bioc.uvic.ca/workbench.php?tool=graphdna&db=	<ul style="list-style-type: none"> • Skew of two selectable nucleotides or IUPAC symbols • DNA walker
GeneR	Download	R package	http://bioconductor.org/packages/2.0/bioc/html/GeneR.html	<ul style="list-style-type: none"> • TA and GC skews
Z-curve [81,82]	Online and download	Java	http://tubic.tju.edu.cn/zcurve/	<ul style="list-style-type: none"> • Three-dimensional curve representations
ADE4 [83]	Download	R package	http://pbil.univ-lyon1.fr/Rweb/	<ul style="list-style-type: none"> • Multivariate data analysis
LastWave	Download	C	http://www.cmap.polytechnique.fr/~bacry/LastWave/	<ul style="list-style-type: none"> • Wavelet and Fourier transforms • Extrema representations of wavelet transforms
PAML [84]	Download	C	http://abacus.gene.ucl.ac.uk/software/paml.html	<ul style="list-style-type: none"> • Phylogenetic analysis by maximum likelihood

best described by two different parameters, so that there are significant substitution asymmetries.

7. Available software

The analysis of strand asymmetry can be relatively simple while requiring almost necessarily a careful expert analysis. There are some freely available programs to make strand asymmetry analysis (Table 2). SeqinR is one of the most complete programs to perform skew analysis, which can then be used to identify replichores. It runs in R, as many of the other software we indicate. R is a free software environment for statistical computing that is becoming the norm for serious statistical studies (<http://www.r-project.org/>). These packages have the further advantage of profiting from the excellent statistical and graphical capabilities of R.

One subject that is often put forward when identifying replichores with nucleotide or oligonucleotide skew methods is the accuracy. Unfortunately, there is neither an easy answer to the question of which is the best method nor to which are the best parameters for a given method. This is simply because we know very few experimentally and precisely determined origins of replication. The problem is even more difficult for the assessment of how accurately one can determine the terminus of replication. This is because there are still doubts if replication finishes at one well-defined site or at a range of different sites [74,75]. This obviously prevents a thorough assessment on the best way of delimitating replichores. Still, several works have departed from skew analysis to try to identify experimentally the origins of replication in bacteria and archaea and these showed that GC skews pointed to the correct origins of replication [41,76,77]. There have also been attempts at improving the identification of origins of replication using other external data [78].

There are no specific programs to deal with signal processing or multivariate statistical analysis of strand asymmetry. Yet, many programs available to compute the composition

and the skews of the sequences can be used to produce the input of general-purpose programs that have the possibility of analysing these types of data (Table 2). This is also true for the analysis of substitution matrices. In this case, analysis of maximum likelihood can be done with PAML, while parsimony analysis can be easily scripted. All these analyses can hardly be automated and require a minimum knowledge of general statistics and of the method being used.

8. Conclusion

Many methods have been proposed in the literature to analyse strand asymmetry and it is important to know their possibilities and their limitations. As the field advances it is likely, and desirable, that more sophisticated methods aiming at unravelling the biological origins of the composition asymmetries become routine in genome analysis. This does not mean that simple methods, such as GC skews, will be less used. To identify origins and termini of replication, GC and TA skews work remarkably well as long as replichores are long and the signal is strong. The use of more sophisticated signal processing techniques to delimitate replichores will probably be limited to the cases when the signal is not very clear, i.e. weak asymmetries and small replichores. Yet, this may be the case in many eukaryotic genomes, where the organisation of the genome around replication is largely ignored. The systematic study of strand asymmetry in the last decade has allowed identifying replichores, transcription units and mutational biases in many genomes. If several mechanisms create overlapping biases and one is interested in understanding the biology of genome organisation, the results provided by skew and signal processing methods are hard to interpret.

Understanding the mutational or selective origins of compositional asymmetries will require the identification of the types of substitutions underlying strand asymmetry. Simply counting nucleotides not only does not provide enough

information but may even be misleading as similar overall skews may result from very diverse mutational biases [26].

Acknowledgements

MT is funded by the Conseil Régional de l'Ile de France. We thank Alain Viari for discussions and for parts of Fig. 5.

References

- [1] J.D. Watson, F.C. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid, *Nature* 171 (1953) 737–738.
- [2] E. Chargaff, Chemical specificity of nucleic acids and mechanism of their enzymatic degradation, *Experientia* 6 (1950) 201–240.
- [3] N. Sueoka, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, *J. Mol. Evol.* 40 (1995) 318–325.
- [4] J.R. Lobry, Properties of a general model of DNA evolution under no-strand bias conditions, *J. Mol. Evol.* 40 (1995) 326–330.
- [5] E.P.C. Rocha, The replication-related organisation of the bacterial chromosome, *Microbiology* 150 (2004) 1609–1627.
- [6] E. Bouffartigues, M. Buckle, C. Badaut, A. Travers, S. Rimsky, H-NS co-operative binding to high-affinity sites in a regulatory element results in transcriptional silencing, *Nat. Struct. Mol. Biol.* 14 (2007) 441–448.
- [7] M. El Karoui, V. Biauudet, S. Schbath, A. Gruss, Characteristics of Chi distribution on different bacterial genomes, *Res. Microbiol.* 150 (1999) 579–587.
- [8] S. Bigot, O.A. Saleh, C. Lesterlin, C. Pages, M. El Karoui, C. Dennis, M. Grigoriev, J.F. Allemand, F.X. Barre, F. Cornet, KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase, *EMBO J.* 24 (2005) 3770–3780.
- [9] H.J. Lin, E. Chargaff, On the denaturation of deoxyribonucleic acid II. Effects of concentration, *Biochem. Biophys. Acta* 145 (1967) 398–409.
- [10] W. Szybalski, H. Kubinski, P. Sheldrick, Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis, *Cold Spring Harbor Symp. Quant. Biol.* 31 (1966) 123–127.
- [11] O. Smithies, W.R. Engels, J.R. Devereux, J.L. Slightom, S. Shen, Base substitutions, length differences and DNA strand asymmetries in the human G gamma and A gamma fetal globin gene region, *Cell* 26 (1981) 345–353.
- [12] J. Filipinski, Evolution of DNA sequences. Contributions of mutational bias and selection to the origin of chromosomal compartments, in: G. Obe (Ed.), *Advances in Mutagenesis Research*, vol. 2, Springer-Verlag, Berlin, 1990, pp. 1–54.
- [13] N.T. Perna, T.D. Kocher, Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes, *J. Mol. Evol.* 41 (1995) 353–358.
- [14] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.* 13 (1996) 660–665.
- [15] M.J. McLean, K.H. Wolfe, K.M. Devine, Base composition skews, replication orientation and gene orientation in 12 prokaryote genomes, *J. Mol. Evol.* 47 (1998) 691–696.
- [16] P. Mackiewicz, A. Gierlik, M. Kowalczyk, M.R. Dudek, S. Cebrat, How does replication-associated mutational pressure influence amino acid composition of proteins? *Genome Res.* 9 (1999) 409–416.
- [17] A.C. Frank, J.R. Lobry, Asymmetric patterns: a review of possible underlying mutational or selective mechanisms, *Gene* 238 (1999) 65–77.
- [18] E.P.C. Rocha, A. Danchin, A. Viari, Universal replication bias in bacteria, *Mol. Microbiol.* 32 (1999) 11–16.
- [19] E.R. Tillier, R.A. Collins, Replication orientation affects the rate and direction of bacterial gene evolution, *J. Mol. Evol.* 51 (2000) 459–463.
- [20] E.P.C. Rocha, A. Danchin, Ongoing evolution of strand composition in bacterial genomes, *Mol. Biol. Evol.* 18 (2001) 1789–1799.
- [21] P. Green, B. Ewing, W. Miller, P.J. Thomas, E.D. Green, Transcription-associated mutational asymmetry in mammalian evolution, *Nat. Genet.* 33 (2003) 514–517.
- [22] M. Touchon, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes, *Nucleic Acids Res.* 32 (2004) 4969–4978.
- [23] A. Gierlik, M. Kowalczyk, P. Mackiewicz, M.R. Dudek, S. Cebrat, Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? *J. Theor. Biol.* 202 (2000) 305–314.
- [24] M. Touchon, S. Nicolay, B. Audit, E.B. Brodie, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins, *Proc. Natl. Acad. Sci. U.S.A.* 102 (2005) 9836–9841.
- [25] E.N. Trifonov, The multiple codes of nucleotides sequences, *Bull. Math. Biol.* 51 (1989) 417–432.
- [26] E.P. Rocha, M. Touchon, E.J. Feil, Similar compositional biases are caused by very different mutational effects, *Genome Res.* 16 (2006) 1537–1547.
- [27] S.J. Bell, D.R. Forsdyke, Accounting units in DNA, *J. Theor. Biol.* 197 (1999) 51–61.
- [28] J.M. Freeman, T.N. Plasterer, T.F. Smith, S.C. Mohr, Patterns of genome organization in bacteria, *Science* 279 (1998) 1827a.
- [29] E.R. Tillier, R.A. Collins, The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes, *J. Mol. Evol.* 50 (2000) 249–257.
- [30] P. Lopez, H. Philippe, Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation, *C. R. Acad. Sci. Ser. III* 324 (2001) 201–208.
- [31] C. Nikolaou, Y. Almirantis, A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species, *Nucleic Acids Res.* 33 (2005) 6816–6822.
- [32] J.R. Lobry, A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria, *Biochimie* 78 (1996) 323–326.
- [33] E. Mizraji, J. Ninio, Graphical coding of nucleic acid sequences, *Biochimie* 67 (1985) 445–448.
- [34] A. Grigoriev, Analyzing genomes with cumulative skew diagrams, *Nucleic Acids Res.* 26 (1998) 2286–2290.
- [35] A. Grigoriev, Graphical genome comparison: rearrangements and replication origin of *Helicobacter pylori*, *Trends Genet.* 16 (2000) 376–378.
- [36] E.P.C. Rocha, A. Danchin, Essentiality, not expressiveness, drives gene strand bias in bacteria, *Nat. Genet.* 34 (2003) 377–378.
- [37] J. Lobry, N. Sueoka, Asymmetric directional mutation pressures in bacteria, *Genome Biol.* 3 (2002) 0058.1–0058.14.
- [38] F.R. Blattner, G.P. Iii, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, Y. Shao, The complete genome sequence of *Escherichia coli* K-12, *Science* 277 (1997) 1453–1461.
- [39] S.L. Salzberg, A.J. Salzberg, A.R. Kerlavage, J.F. Tomb, Skewed oligomers and origins of replication, *Gene* 217 (1998) 57–67.
- [40] J. Mrázek, S. Karlin, Strand compositional asymmetry in bacterial and large viral genomes, *Proc. Natl. Acad. Sci. U.S.A.* 95 (1998) 3720–3725.
- [41] H. Myllykallio, P. Lopez, P. Lopez-Garcia, R. Heilig, W. Saurin, Y. Zivanovic, H. Philippe, P. Forterre, Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon, *Science* 288 (2000) 2212–2215.
- [42] P. Lopez, H. Philippe, H. Myllykallio, P. Forterre, Identification of putative chromosomal origins of replication in Archaea, *Mol. Microbiol.* 32 (1999) 883–886.
- [43] J.H. Zar, *Biostatistical Analysis*, Prentice Hall, New Jersey, 1996.
- [44] L. Marsan, M.F. Sagot, Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification, *J. Comput. Biol.* 7 (2000) 345–362.
- [45] O. Levy, J.L. Ptacin, P.J. Pease, J. Gore, M.B. Eisen, C. Bustamante, N.R. Cozzarelli, Identification of oligonucleotide sequences that direct the movement of the *Escherichia coli* FtsK translocase, *Proc. Natl. Acad. Sci. U.S.A.* 102 (2005) 17618–17623.
- [46] P. Worning, L.J. Jensen, P.F. Hallin, H.H. Staerfeldt, D.W. Ussery, Origin of replication in circular prokaryotic chromosomes, *Environ. Microbiol.* 8 (2006) 353–361.

- [47] E.P.C. Rocha, A. Viari, A. Danchin, Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons, *Nucleic Acids Res.* 26 (1998) 2971–2980.
- [48] S. Robin, S. Schbath, V. Vandewalle, Statistical tests to compare motif count exceptionalities, *BMC Bioinformatics* 8 (2007) 84.
- [49] H. Capiiaux, F. Cornet, J. Corre, M.I. Guijo, K. Peral, J.E. Rebollo, J.M. Louarn, Polarization of the *Escherichia coli* chromosome. A view from the terminus, *Biochimie* 83 (2001) 161–170.
- [50] H. Hendrickson, J.G. Lawrence, Selection for chromosome architecture in bacteria, *J. Mol. Evol.* 62 (2006) 615–629.
- [51] K. Arakawa, R. Saito, M. Tomita, Noise-reduction filtering for accurate detection of replication termini in bacterial genomes, *FEBS Lett.* 581 (2007) 253–258.
- [52] S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Low frequency rhythms in human DNA sequences: a key to the organization of gene location and orientation? *Phys. Rev. Lett.* 93 (2004) 108101.
- [53] J. Song, A. Ware, S.L. Liu, Wavelet to predict bacterial ori and ter: a tendency towards a physical balance, *BMC Genomics* 4 (2003) 17.
- [54] E.B. Brodie, S. Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, From DNA sequence analysis to modeling replication in the human genome, *Phys. Rev. Lett.* 94 (2005) 248103.
- [55] C.J. Huberty, *Applied Discriminant Analysis*, Wiley-Interscience, New Jersey, 1994.
- [56] H.R. Lindman, *Analysis of Variance in Complex Experimental Designs*, W.H. Freeman & Co., San Francisco, 1974.
- [57] M.J. Greenacre, *Correspondence Analysis in Practice*, Academic Press, London, 1993.
- [58] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pavé, Codon catalog usage and the genome hypothesis, *Nucleic Acids Res.* 8 (1980) r49–r62.
- [59] J.R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, *Nucleic Acids Res.* 22 (1994) 3174–3180.
- [60] G. Pascal, C. Medigue, A. Danchin, Persistent biases in the amino acid composition of prokaryotic proteins, *Bioessays* 28 (2006) 726–738.
- [61] G. Perrière, J. Thioulouse, Use and misuse of correspondence analysis in codon usage studies, *Nucleic Acids Res.* 30 (2002) 4548–4555.
- [62] J.O. McInerney, Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*, *Proc. Natl. Acad. Sci. U.S.A.* 95 (1998) 10698–10703.
- [63] G. Perrière, J.R. Lobry, J. Thioulouse, Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences, *CABIOS* 12 (1996) 519–524.
- [64] L. Klasson, S.G. Andersson, Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways, *Mol. Biol. Evol.* 25 (2006) 1031–1039.
- [65] M. Tanaka, T. Ozawa, Strand asymmetry in human mitochondrial DNA mutations, *Genomics* 22 (1994) 327–335.
- [66] J.P. Bielawski, J.R. Gold, Mutation patterns of mitochondrial H- and L-strand DNA in closely related Cyprinid fishes, *Genetics* 161 (2002) 1589–1597.
- [67] M.P. Francino, L. Chao, M.A. Riley, H. Ochman, Asymmetries generated by transcription-coupled repair in enterobacterial genes, *Science* 272 (1996) 107–109.
- [68] A. Beletskii, A.S. Bhagwat, Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*, *Proc. Natl. Acad. Sci. U.S.A.* 93 (1996) 13919–13924.
- [69] M.P. Francino, H. Ochman, Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences, *Mol. Biol. Evol.* 18 (2001) 1147–1150.
- [70] A. Eyre-Walker, Problems with parsimony in sequences of biased base composition, *J. Mol. Evol.* 47 (1998) 686–690.
- [71] T. Gojobori, W.H. Li, D. Graur, Patterns of nucleotide substitution in pseudogenes and functional genes, *J. Mol. Evol.* 18 (1982) 360–369.
- [72] M.P. Francino, H. Ochman, Strand symmetry around the beta-globin origin of replication in primates, *Mol. Biol. Evol.* 17 (2000) 416–422.
- [73] Z. Yang, Estimating the pattern of nucleotide substitution, *J. Mol. Evol.* 39 (1994) 105–111.
- [74] H. Hendrickson, J.G. Lawrence, Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites, *Mol. Microbiol.* 64 (2007) 42–56.
- [75] J.M. Louarn, J.P. Bouche, F. Legendre, J. Louarn, J. Patte, Characterization and properties of very large inversions of the *E. coli* chromosome along the origin-to-terminus axis, *Mol. Gen. Genet.* 201 (1985) 467–476.
- [76] M. Lundgren, A. Andersson, L. Chen, P. Nilsson, R. Bernander, Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination, *Proc. Natl. Acad. Sci. U.S.A.* 101 (2004) 7046–7051.
- [77] M. Picardeau, J.R. Lobry, B.J. Hinnenbusch, Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome, *Mol. Microbiol.* 32 (1999) 437–445.
- [78] P. Mackiewicz, J. Zakrzewska-Czerwinska, A. Zawilak, M.R. Dudek, S. Cebrat, Where does bacterial replication start? Rules for predicting the oriC region, *Nucleic Acids Res.* 32 (2004) 3781–3791.
- [79] A.C. Frank, J.R. Lobry, Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes, *Bioinformatics* 16 (2000) 560–561.
- [80] J.M. Thomas, D. Horspool, G. Brown, V. Tcherepanov, C. Upton, GraphDNA: a Java program for graphical display of DNA composition analyses, *BMC Bioinformatics* 8 (2007) 21.
- [81] C.T. Zhang, R. Zhang, H.Y. Ou, The Z curve database: a graphic representation of genome sequences, *Bioinformatics* 19 (2003) 593–599.
- [82] R. Zhang, C.T. Zhang, Z curves, an intuitive tool for visualizing and analyzing the DNA sequences, *J. Biomol. Struct. Dyn.* 11 (1994) 767–782.
- [83] J. Thioulouse, D. Chessel, S. Dolédec, J.M. Olivier, ADE-4: a multivariate analysis and graphical display software, *Stat. Comput.* 7 (1996) 75–83.
- [84] Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood, *CABIOS* 13 (1997) 555–556.