14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

LARGE-SCALE ANALYSIS OF THE HUMAN GENOME: FROM DNA SEQUENCE ANALYSIS TO THE MODELING OF REPLICATION IN HIGHER EUKARYOTES

A. Arneodo[1], *Y. d'Aubenton-Carafa*[2], *B. Audit*[1], *E.B. Brodie of Brodie*[1], *S. Nicolay*[1], *P. St-Jean*[1], *C. Thermes*[2], *M. Touchon*[2] *and C. Vaillant*[3]

[1] Laboratoire Joliot-Curie and Laboratoire de Physique, UMR5672, CNRS, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France. web: www.ens-lyon.fr/Joliot-Curie
phone: +(33) 4 72 72 87 57, fax: +(33) 4 72 72 80 80, email: Alain.Arneodo@ens-lyon.fr
[2]Centre de Génétique Moléculaire, CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France.
[3] Laboratoire Statistique et Génome, 523 Place des Terrasses de l'Agora, 91000 Evry, France.

ABSTRACT

We explore large-scale nucleotide compositional fluctuations along the human genome through the optics of the wavelet transform microscope. Analysis of the TA and GC skews reveals the existence of strand asymmetries associated to transcription and/or replication. The investigation of 14854 intron-containing genes shows that both skews display a characteristic step-like profile exhibiting sharp transitions between transcribed (finite bias) and non-transcribed (zero bias) regions. As we observe for 7 out of 9 origins of replication experimentally identified so far, the (AT+GC) skew exhibits rather sharp upward jumps, with a linear decreasing profile in between two successive jumps. We describe a multi-scale methodology that allows us to predict 1012 replication origins in the 22 human autosomal chromosomes. We present a model of replication with well-positioned replication origins and random termination sites that accounts for the observed characteristic serrated skew profiles. We emphasize these putative replication initiation zones as regions where the chromatin fiber is likely to be more open so that DNA be easily accessible. In the crowded environment of the cell nucleus, these intrinsic decondensed structural defects actually predisposes the fiber to spontaneously form rosette-like structures that provide an attractive description of genome organization into replication foci that are observed in interphase mammalian nuclei.

1. INTRODUCTION

During genome evolution, mutations do not occur at random as illustrated by the diversity of the nucleotide substitution rate values [1]. This non-randomness is considered as a by-product of the various DNA mutation and repair processes that can affect each of the two DNA strands differently. Deviations from intrastrand equimolarities, the so-called Chargaff's second parity rule [2], have been extensively studied during the past decade and the observed skews have been attributed to asymmetries intrinsic to the replication and/or to the transcription processes. Asymmetries of substitution rates coupled to transcription have been mainly observed in prokaryotes [3, 4], with only preliminary results in eukaryotes [5]. Strand asymmetries (i.e., G≠C and T≠A) associated with the polarity of replication have been found in bacterial, mitochondrial and viral genomes [6–9] where they have been used to detect replications origins. In most cases, the leading replication strand presents an excess of G over C and of T over A. Along one DNA strand, the sign of this bias changes abruptly at the replication origin (*ori*) and terminus (*ter*).

In eukaryotes, the existence of compositional biases is unclear and most attempts to detect the *ori* from strand compositional asymmetry have been inconclusive. Several studies have failed to show compositional biases related to replication, and analysis of nucleotide substitutions in the region of the *β-globin ori* in primates do not support the existence of mutational bias between the leading and lagging strands [7, 10, 11]. Other studies have led to rather opposite results. For instance, strand asymmetries associated with replication have been observed in the subtelomeric regions of *Saccharomyces cerevisiae* chromosomes, supporting the existence of replication-coupled asymmetric mutational pressure in this organism [12]. The aim of the present work is to show that with an adequate multi-scale methodology, one can to some extent disentangle the contributions to the strand asymmetries induced by transcription and replication respectively and challenge the issue of detecting putative *ori* directly from genomic sequences.

2. TRANSCRIPTION-INDUCED STEP-LIKE SKEW PROFILES IN THE HUMAN GENOME

We have started examining nucleotide compositional strand asymmetries in transcribed regions of the human genome [13, 14]. Sequences and gene annotation data were downloaded from the UCSC Genome server, for the human (July 2003 in section 2, May 2004 in sections 3 and 4), mouse (May 2004) and dog (July 2004) genomes. To exclude repetitive elements that might have been inserted recently and would not reflect long-term evolutionary patterns, we used REPEATMASKER leading to a reduction of ∼40-50% of the human sequence length. All analyses were carried out using KNOWNGENE annotations. The TA and GC skews were calculated in non-overlapping windows of size 1-kbp as:

$$S_{\rm TA} = \frac{n_{\rm T} - n_{\rm A}}{n_{\rm T} + n_{\rm A}}, \qquad S_{\rm GC} = \frac{n_{\rm G} - n_{\rm C}}{n_{\rm G} + n_{\rm C}}, \tag{1}$$

where $n_{\rm A}$, $n_{\rm C}$, $n_{\rm G}$ and $n_{\rm T}$ are respectively the numbers of A, C, G and T in the windows. Because of the observed correlation between $S_{\rm TA}$ and $S_{\rm GC}$, we also considered the total skew $S = S_{\rm TA} + S_{\rm GC}$.

In Figure 1 are reported the mean values of these skews for 14854 intron containing genes as a function of the distance to the 5'- or 3'- end. At the 5' gene extremities (Figure 1(a)), a sharp transition of both skews is observed from about zero values in the intergenic regions to finite positive values in transcribed regions ranging between 4 and 6% for $S_{\rm TA}$ and between 3 and 5% for $S_{\rm GC}$. At the gene 3'- extremities (Figure 1(b)), the TA and GC skews also exhibit transitions from significantly large values in transcribed regions to very small values in untranscribed regions. In comparison to the steep transitions observed at 5'- ends, the 3'- end profiles present a slightly smoother transition pattern extending over ∼5 kbp and including regions downstream of the 3'- end likely reflecting the fact that transcription continues to some extent downstream of the polyadenylation site. The results reported in Figure 1 suggest that $S_{\rm TA}$ and $S_{\rm GC}$ are constant along introns. Since introns amount for about 80% of gene sequences, this means that skew profiles induced by transcription processes have a characteristic step-like shape [13, 14]. However, the absence of asymmetries in intergenic regions does not exclude the possibility of additional replication associated biases. Such biases would present opposite signs on leading and lagging strands and would cancel each other in our statistical analysis as a result of the spatial distribution of multiple unknown *ori* [15].

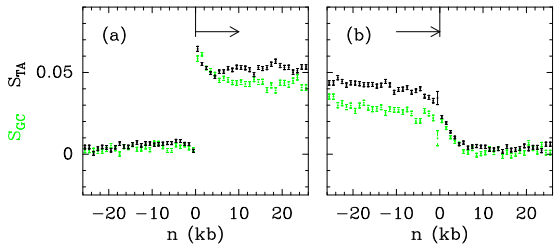

Figure 1: S_{TA} (•) and S_{GC} (•) calculated in 1 kbp windows at the distance (n) on the native sequences to the indicated gene extremity; zero values of abscissa correspond to 5'- (a) or 3'- (b) gene extremities. In ordinate is reported the mean value of the skews over our set of 14854 intron-containing genes. Error bars represent the standard error of the means. In order to avoid the skews associated with exonic and splicing signals, S_{TA} and S_{GC} were calculated only on intronic sequences after removing 560 bp at both intron extremities.

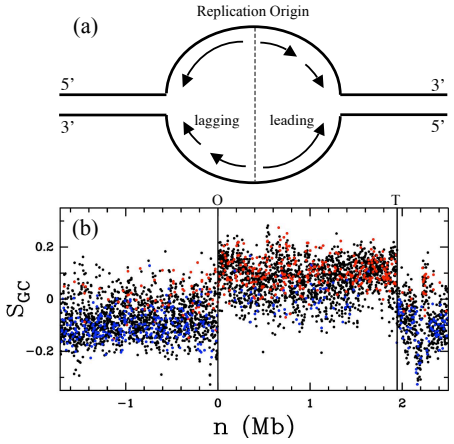

Figure 2: (a) Schematic representation of the divergent bi-directional progression of the two replication forks from the *ori*. (b) S_{GC} calculated in 1 kbp windows along the sequence of *Bacillus subtilis*; the vertical lines correspond to the *ori* (O) and *ter* (T) positions; the red (blue) points correspond to sense (antisense) genes that have the same (opposite) orientation than the sequence.

3. REPLICATION-INDUCED FACTORY-ROOF-LIKE SKEW PROFILES IN MAMMALIAN GENOMES

DNA replication is an essential genomic function responsible for the accurate transmission of genetic information through successive cell generations. According to the so-called "replicon" paradigm derived from prokaryotes [16], this process starts with the binding of some "initiator" protein to a specific "replicator" DNA sequence called *origin of replication* (*ori*). The recruitment of additional factors initiate the bi-directional progression of two divergent replication forks along the chromosome. As illustrated in Figure 2(a), one strand is replicated continuously (leading strand), while the other strand is replicated in discrete steps towards the *ori* (lagging strand). In eukaryotic cells, this event is initiated at a number of *ori* and propagates until two converging forks collide at a *ter* [17]. The initiation of different *ori* is coupled to the cell cycle but there is a definite flexibility in the usage of the *ori* at different developmental stages [18–22]. Also, it can be strongly influenced by the distance and activation timing of neighboring *ori*, by the transcriptional activity and by the local chromatin structure [19–22]. Actually, sequence requirements for an *ori* vary significantly between different eukaryotic organisms. In the unicellular *Saccharomyces cerevisiae*, the *ori* spread over 100-150 bp and present some highly conserved motifs [17]. In the fission yeast *Schizosaccharomyces pombe*, there is no clear consensus sequence and the *ori* spread over at least 800 to 1000 bp [17]. In multicellular organisms, the *ori* are rather poorly defined and initiation may occur at multiple sites distributed over thousands of base pairs [23]. Actually cell diversification may have led higher eukaryotes to develop various epigenetic controls over the *ori* selection rather than to conserve specific replication sequence. This might explain that very few *ori* have been identified so far in multicellular eukaryotes, around 20 in metazoa and only about 10 in human [24, 25]. Along the line of this epigenetic interpretation, one might wonder what can be learned about eukaryotic DNA replication from DNA sequence analysis.

3.1 Replication-associated strand asymmetries in prokaryotic genomes: the replicon model

The existence of replication associated strand asymmetries has been mainly established in bacterial genomes [6–9]. As illustrated in Figure 2, the GC and TA skews abruptly switch sign (over few kbp) from negative to positive values at the *ori* and in the opposite direction from positive to negative values at the *ter*. This step-like profile is characteristic of the replicon model [16]. In *Bacillus subtilis*, as in most bacteria, the leading (resp. lagging) strand (Fig. 2(a)) is generally richer (resp. poorer) in G than in C (Fig. 2(b)), and to a lesser extent in T than in A (data not shown). When looking at the gene organization around *ori*, one observes in Figure 2(b) that most of the sense (resp. antisense) genes that have the same orientation as the Watson (resp. Crick) strand are preferentially located on the right (resp. left) of the *ori*. This suggests that the replication forks progression is co-oriented with transcription, as to minimize the risk of frontal collision between DNA and RNA polymerases [26, 27].

3.2 Analysis of strand asymmetries around experimentally determined replication origins in the human genome

As shown in Figure 3(a) for the TOP1 *ori*, 6 among the 9 *ori* that have been experimentally identified in the human genome, correspond to rather sharp transitions from negative to positive S_{TA} and S_{GC} skew values that clearly emerge from the noisy background [24, 25] (note that among the exceptions, the Lamin B2 and β-globin *ori*, might well be inactive in germline or less frequently used than the adjacent *ori*). This is reminiscent of the behavior observed in Figure 2 for *Bacillus subtilis*, except that the leading strand is relatively enriched in T over A and in G over C. According to the gene environment, the amplitude of the jump observed in the skew profiles can be more or less important and its position more or less localized (from a few kbp to a few tens of kbp). Indeed, we have seen in Section 2 that transcription generates positive TA and GC skews on the coding strand [13, 14, 28], which explains that larger jumps are observed when the sense and/or the antisense genes are on the leading strand so that replication and transcription biases add to each other. The total skew jump amplitude ΔS calculated as the difference of the skews measured in 20 kbp windows on both sides of the 6 *ori*, are: MCM4 (31%), HSPA4 (29%), TOP1 (18%), MYC (14%), SCA7 (38%), and AR (14%). To measure compositional asymmetries that would result from replication only, we have calculated the skews in intergenic regions on both sides of the *ori* [24]. The total skew S definitely shifts from negative ($S = -6.2 \pm 0.4\%$) to positive ($S = 11.1 \pm 1\%$) values when crossing the *ori*. This result strongly suggests the existence of mutational pressure associated with replication, leading to the mean compositional biases $S_{TA} = 4.0 \pm 0.4\%$ and $S_{GC} = 3.0 \pm 0.5\%$. Let us note that the value of the skew could vary from one *ori* to another, possibly reflecting different initiation efficiencies. From the calculation of the intron skew values on the leading ($S_{TA} = 7.5 \pm 0.3\%$, $S_{GC} = 6.8 \pm 0.4\%$) and lagging ($S_{TA} = -1.9 \pm 1.0\%$, $S_{GC} = -0.3 \pm 1.4\%$) strands, one can estimate the mean skew associated with transcription by subtracting intergenic skews from S_{lead} values giving $S_{TA} = 3.6 \pm 0.7\%$ and $S_{GC} = 3.8 \pm 0.9\%$. These estimations are remarkably consistent with those obtained with our large set of human introns in Sec-

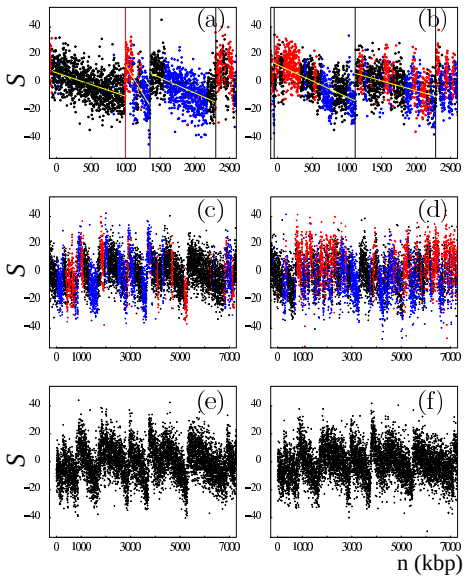

Figure 3: S profiles along (repeat-masked) mammalian genome fragments [24]. (a) Fragment of human chromosome 20 including the TOP1 *ori* (red vertical line). (b and c) Human chromosome 4 and chromosome 9 fragments, respectively, with low GC content (36%). (d) Human chromosome 22 fragment with larger GC content (48%). In (a) and (b), vertical lines correspond to selected putative *ori* (see Section 4.1). Black, intergenic regions; red, sense genes; blue, antisense genes. (e) Fragment of mouse chromosome 4 homologous to the human fragment shown in (c). (f) Fragment of dog chromosome 5 syntenic to the human fragment shown in (c). In (e) and (f), genes are not represented.

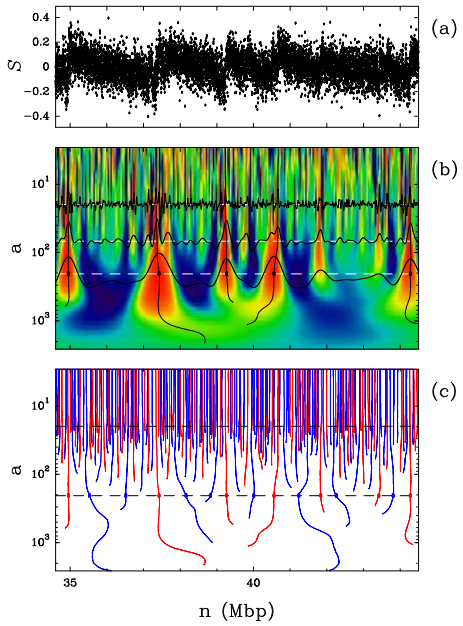

Figure 4: (a) Skew profiles of a fragment of Human chromosome 12. (b) WT of S using $g^{(1)}$; $W_{g^{(1)}}[S](n,a)$ is coded from black (min) to red (max); three cuts of the WT at constant scale $a = a^* = 200$ kbp, 70 kbp and 20 kbp are superimposed together with five maxima lines identified as pointing to upward jumps in the skew profile. (c) WT skeleton defined by the maxima lines in blue (resp. red) when corresponding to positive (resp. negative) values of the WT. At the scale $a^* = 200$ kbp, one thus identify 7 upward (blue dots) and 8 downward (red dots) jumps. The black dots in (b) correspond to the 5 WTMM of largest amplitude that have been identified as putative *ori*; the associated maxima lines point to the 5 major upward jumps in the S profile in the limit $a \to 0^+$.

tion 2, further supporting the existence of replication-coupled strand asymmetries. Overall, these results indicate that the mean replication bias on the leading strand and the mean transcriptional bias on the coding strand are of the same order of magnitude, namely $S = S_{\rm TA} + S_{\rm GC} \sim 7\%$ [24, 25].

3.3 Factory-roof skew profiles in the human genome

As illustrated in Figure 3(a), for TOP1 *ori*, when examining the behavior of the skews at larger distances from the *ori*, one does not observe a step-like pattern with upward and downward jumps at the *ori* and *ter* positions respectively as expected for the bacterial replicon model (Fig. 2(b)). Surprisingly, on both sides of the upward jump, the noisy S profile decreases steadily in the 5' to 3' direction without clear evidence of pronounced downward jumps. As shown in Figures 3(b–d), sharp upward jumps of amplitude $\Delta S \gtrsim 15\%$, similar to the ones observed for the known *ori* (Fig. 3(a)), seem to exist also at many other locations along the human chromosomes. But the most striking feature is the fact that in between two neighboring major upward jumps, not only the noisy S profile does not present any comparable downward sharp transition, but it displays a remarkable decreasing linear behavior. At chromosome scale, one thus gets jagged S profiles that have the aspect of "factory roofs" [24, 25]. These S profiles look somehow disordered because of the extreme variability in the distance between two successive upward jumps, from spacing $\sim 50-100$ kbp ($\sim 100-200$ kbp for the native sequences) up to 1-2 Mbp ($\sim$ 3-4 Mbp for the native sequences) in agreement with recent experimental studies [15] that have shown that mammalian replicons are heterogeneous in size with an average size ~ 500 kbp. But what is important to notice is that some of these segments between two successive upward jumps of the skew are entirely intergenic (Figs. 3(a,c)), clearly illustrating the particular profile of a strand bias resulting solely from replication [24, 25].

In most other cases, one observes the superimposition of this replication profile and of the step-like profiles of sense and antisense genes, appearing as upward and downward blocks standing out from the replication pattern (Fig. 3(c)). Importantly, as illustrated in Figures 3(e) and 3(f), the factory-roof pattern is not specific to human sequences but is also observed in numerous regions of the mouse and dog genomes [24], which strongly suggests that replication-associated strand asymmetry is conserved in mammalian genomes.

4. FROM THE DETECTION OF PUTATIVE *ORI* TO THE MODELING OF REPLICATION IN HUMAN

4.1 A wavelet-based method to detect putative *ori*

We have shown in Section 3 that experimentally determined human *ori* coincide with large-amplitude upward transitions in noisy skew profiles. The corresponding ΔS ranges between 14% and 38%, owing to possible different replication initiation efficiencies and/or different contributions of transcriptional biases. To predict *ori*, one thus needs a methodology to detect discontinuities in noisy signals. As originally introduced in Refs. [29, 30], the continuous wavelet transform (WT) is a mathematical microscope that is well adapted for singularity tracking. The WT is a space-scale analysis which consists in expanding signals in terms of wavelets that are constructed from a single function, the analyzing wavelet ψ, by means of dilations and translations. When using the successive derivatives of the Gaussian function as analyzing wavelets, namely

$$g^{(N)}(x) = (-1)^N d^N g^{(0)}(x)/dx^N, \qquad (2)$$

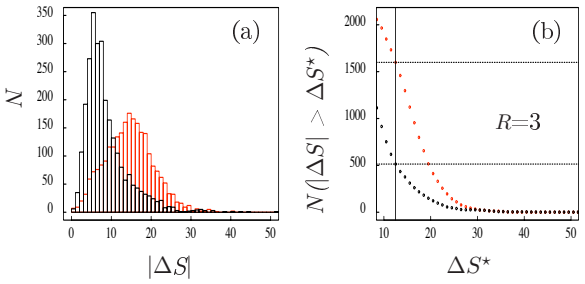

Figure 5: Statistical analysis of the sharp jumps detected in the S profiles of the 22 human autosomal chromosomes by the WT microscope at scale $a^* = 200$ kbp for repeat-masked sequences [24, 25]. $|\Delta S| = |S(3') - S(5')|$, where the averages were computed over the two 20 kbp windows on both sides of the detected jump location. (a) Histograms $N(|\Delta S|)$ of $|\Delta S|$ values. (b) $N(|\Delta S| > \Delta S^*)$ vs ΔS^*. In (a) and (b), the black (resp. red) line corresponds to downward $\Delta S < 0$ (resp. upward $\Delta S > 0$) jumps. $R = 3$ corresponds to the ratio of upward over downward jumps presenting an amplitude $|\Delta S| \geq 12.5\%$. This ratio increases to $R = 4.76$ when considering only the jumps in regions with G+C<42%.

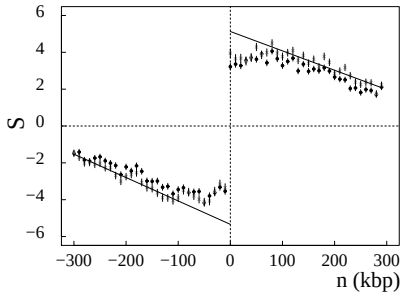

Figure 6: Mean skew profile of intergenic regions around putative *ori* [24, 25]. S was calculated in 1 kbp windows (Watson strand) around the position (without repeats) of the 1012 detected upward jumps; 5' and 3' transcript extremities were extended by 0.5 and 2 kbp, respectively (•), or by 10 kbp at both ends (∗). The abscissa represents the distance to the corresponding *ori*; the ordinate represents S (in percent) calculated for the windows situated in intergenic regions.

where $g^{(0)}(x) = (2\pi)^{-1/2} e^{-x^2/2}$, then the WT of a signal s takes the following simple expression:

$$W_{g^{(N)}}[s](x,a) = \frac{1}{a}\int_{-\infty}^{+\infty} s(y) g^{(N)}\left(\frac{y-x}{a}\right) dy, = \frac{d^N}{dx^N} W_{g^{(0)}}[s](x,a), \quad (3)$$

where x and $a(>0)$ are the space and scale parameters. The basic principle of the detection of jumps in the skew profiles with the WT is illustrated in Figure 4. From Eq. (3), when using $g^{(1)}(x)$ as analyzing wavelet, it is obvious that at a fixed scale a, a large value of the modulus of the WT coefficient corresponds to a strong derivative of the smoothed skew profile. In particular, jumps manifest as local maxima of the WT modulus as illustrated for three different scales in Figure 4(b). The main issue when dealing with noisy signals like the skew profile in Figure 4(a), is to distinguish the local WT modulus maxima (WTMM) associated to the jumps from those induced by the noise. In this respect, the freedom in the choice of the smoothing scale a is fundamental since, whereas the noise amplitude is reduced when increasing the smoothing scale, an isolated jump contributes equally at all scales.

As shown in Figure 4(c), our methodology [24] consists in computing the WT skeleton [29, 30] defined by the set of maxima lines obtained by connecting the WTMM across scales. Then we select a scale $a^* = 200\ kbp$ which is smaller than the typical replicon size and larger than the typical gene size. In this way, we not only reduce the effect of the noise but we also reduce the contribution of the upward (5' extremity) and backward (3' extremity) jumps associated to the step-like skew pattern induced by transcription (Fig. 1). The maxima lines that exist at that scale a^* are likely to point to jump positions at small scale (Fig. 4(c)). The detected jump locations are estimated as the positions at scale 20 kbp of the so-selected maxima lines. According to Eq. (3), upward (resp. downward) jumps are indentified by the maxima lines corresponding to positive (resp. negative) values of the WT as illustrated in Figure 4(c) by the blue (resp. red) maxima lines. When applying this methodology to the total skew S along the repeat-masked DNA sequences of the 22 human autosomal chromosomes, 2415 upward jumps are detected and, as expected, a similar number (namely 2686) of downward jumps. In Figure 5(a) are reported the histograms of the amplitude $|\Delta S|$ of the so-identified upward ($\Delta S > 0$) and downward ($\Delta S < 0$) jumps respectively. These histograms do not superimpose, the former being significantly shifted to larger $|\Delta S|$ values. When plotting $N(|\Delta S| > \Delta S^*)$ versus ΔS^* in Figure 5(b), one can see that the number of large amplitude upward jumps overexceeds the number of large amplitude downward jumps. These results [24, 25] confirm that most of the sharp upward transitions in the S profiles in Figures 3 and 4(a) have no sharp downward transition counterpart. In a final step, we have decided to retain as putative *ori* upward jumps with $|\Delta S| \geq 12.5\%$ detected in region with G+C≤42%. This selection leads to a set of 1012 candidates, some of those putative *ori* are illustrated in Figures 3(a) and 3(b).

4.2 Gene organization around the 1012 putative human *ori*

The mean amplitude of the upward jumps associated with the 1012 putative *ori* is 18%, consistent with the range of values observed for the 6 experimentally known *ori* in Section 3.2, that all have been identified by our detection methodology. When investigating the gene content around these putative *ori* [24, 25], one finds that in a close vincinity (±20 kbp), most DNA sequences (55% of the analyzing windows) are transcribed in the same direction as the progression of the replication fork (namely sense genes on the 3'- side of the *ori* and antisense genes on the 5'- side). By contrast, only 7% of the sequences are transcribed in the opposite direction (38% are intergenic). These results show that the $|\Delta S|$ amplitude at putative *ori* mostly results from superimposition of biases associated with (i) replication and (ii) transcription of the genes proximal to the *ori*.

In Figure 6 is shown the mean skew profile calculated in intergenic windows on both sides of the 1012 putative *ori* [24, 25]. This mean skew profile presents a rather sharp transition from negative to positive values when crossing the *ori* position. On both sides of the jump, we observe a linear decrease of the bias with some flattening of the profile close to the transition point that might be due to (i) the potential presence of signals implicated in replication initiation, (ii) the possible existence of dispersed *ori* [31], (iii) the numerical uncertainty on the putative *ori* position estimate. As illustrated in Figure 6, when extrapolating the linear behavior observed at distances > 100 kbp from the jump, one gets a skew of 5.3%, i.e. a value consistent with the skew measured in intergenic regions around the 6 experimentally known *ori* namely $7.0 \pm 0.5\%$.

4.3 A model of replication in mammalian genomes

Following the observation of jagged skew profiles similar to factory roofs in Section 3.3, and the quantitative confirmation of the existence of such (piecewise linear) profiles in the neighborhood of 1012 putative origins in Figure 6, we have proposed [24, 25], a rather crude model for replication in the human genome that relies on the hypothesis that the *ori* are quite well positioned while the *ter* are randomly distributed. As illustrated in Figure 7, replication

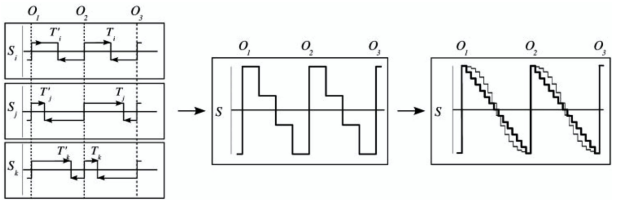

Figure 7: Model of replication termination[24, 25]. Schematic representation of the skew profiles associated with three bidirectional *ori* O_1, O_2, and O_3 with similar replication efficiency. Upward (or downward) steps correspond to *ori* (or *ter*) positions. For convenience, the *ter* sites are symmetric relative to O_2. (Left) Three different *ter* positions T_i, T_j, and T_k, leading to elementary skew profiles S_i, S_j, and S_k. (Center) Superposition of these three profiles. (Right) Superposition of a large number of elementary profiles leading to the final factory-roof pattern.

termination is likely to rely on the existence of numerous *ter* sites distributed along the sequence. For each *ter* site (used in a small proportion of cell cycles), strand asymmetries associated with replication will generate a step-like skew profile with a downward jump at the position of termination and upward jumps at the positions of the adjacent *ori* (as in bacteria, Fig. 2(b)). Addition of those profiles (Fig. 7, left panel) will generate the intermediate profile (Fig. 7, central panel). In a simple picture, we can reasonably suppose that *ter* occurs with constant probability at any position on the sequence. This behavior can, for example, result from the binding of some termination factor at any position between successive *ori*, leading to a homogeneous distribution of *ter* sites during successive cell cycles. The final skew profile is then a linear segment decreasing between successive *ori* (Fig. 7, right panel).

5. CONCLUSIONS AND PERSPECTIVES

In conclusion, we have revealed [24, 25] a factory roof skew profile as an alternative in mammalian genomes to the replicon step-like profile observed in bacteria (Fig. 2). This pattern is displayed by a set of 1012 upward transitions, each flanked on each side by DNA segments of ~ 300 kbp (without repeats), which can be roughly estimated to correspond to 20-30% of the human genome. In these regions, which are characterized by low and medium G+C content (G+C$\leq$ 42%), skew profiles reveal a portrait of germ-line replication consisting of putative *ori* separated by rather long DNA segments ($\sim 1-3$ Mbp on the native sequences). Although such segments are much larger than expected from the classical view [15] ($\sim$ 100 kbp to 500 kbp on the native sequences), they are not incompatible with estimations showing that replicon size can reach up to 1 or 2 Mbp [15, 32], and that replicating units in meiotic chromosomes are much longer than those engaged in sommatic cells [33]. Finally, it is not unlikely that in G+C-rich (gene-rich) regions (Fig. 3(d)) *ori* would be closer to each other than in other regions, further explaining the greater difficulty in detecting *ori* in these regions. Indeed, the wavelet-based methodology described in Section 4.1 remains efficient as long as there exists a clear separation between the characteristic size of a replicon and the characteristic size of a gene; while this separation is unquestionable at low and medium G+C content, this is no longer obvious in high GC regions.

Because most of the 1012 putative *ori* are found to lie close to the promotor regions of sense and/or antisense genes [24, 25], they provide priviledged locations for some intrinsic structural defects where chromatin fiber is likely to be more open so that DNA be more easily accessible. In the crowded environment of the cell nucleus, the presence of such sequence dependent decondensed structural defects actually predisposes the chromatin fiber to spontaneously forms rosette-like structures. Indeed, when considering a semi-flexible tube in a dilute environment of hard spheres, the elastic nature of the tube prevents the appearance of too high curvature

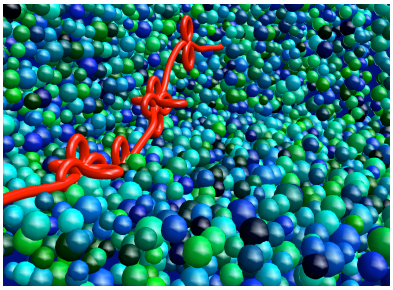

Figure 8: Spontaneous emergence of rosette-like folding of the chromatin fiber in the crowded environment of the cell nucleus.

points; consequently the first step in the condensation of the tube is the formation of loops. Loop formation involves a competition between the bending energy of the tube and the entropic gain of the hard sphere fluid. The free energy cost is dominated by elastic energy for small loops and by entropy for large ones. This results in a preferential length of $3.4 l_p$ in the worm-like-chain (WLC) model [34]. Once a loop is formed, contact will be maintained by depletive forces; hence the loop will preferentially relax through local gliding of the two contact points. This is where local defects come into play: when they meet from this gliding process, they act as local geometrical wells and "stick" together. This defect-induced stabilization is important since it prevents further depletive mechanisms to take place. Indeed by modifying locally the angle of tangent vectors at the contact points, the depletion force could drive them to align in opposite directions, forming the first turn of an helix or toroidal condensate [35]; alternatively it could align them in the same direction, favoring the formation of hairpins. The presence of defects, by favoring a specific contact geometry, *breaks* the symmetries (translational, axial) essential to the formation of these compact structures, drastically modifying the phase diagram. The condensation rather occurs via the aggregation of defects, inducing rosette-like patterns as illustrated in Figure 8. This clustering is likely to favor the recruiting of protein complexes involved in the activation of replication and transcription that will further stabilize the muli-looped patterns [36]. Let us emphasize that the spontaneous emergence of rosette patterns provides a very attractive description of the so-called replication foci [15, 37, 38] that have been observed in interphase mammelian nuclei as stable structural domains of autonomous replication that persist during all cell cycle stages. Furthermore, the remarkable gene organization discovered around the putative *ori* [24, 25, 39], strongly suggests that these rosettes contribute to the compartmentalization of the genome into autonomous domains of gene transcription. Via the self-organizing structural role of the *ori*, the DNA sequence might therefore code, to some extent, for the tertiary chromatin structure [40]. Even though one expects to observe, from one cell cycle to the next, fluctuations in the number of loops contained in each rosette, the perennity of defects is likely to ensure the inheritance of the interphase chromatin rosette organization. Since introns and intergenic regions constitute more than 95% of the human genome, our study therefore contributes to giving a role to the noncoding regions in eukaryotic genomes. These regions are likely to play a driving role in the condensation and decondensation processes of the chromatin architecture as well as in many related regulative functions. In situ studies of the distributions and dynamics of *ori* in the cell nucleus, using fluorescence techniques (FISH chromosome painting [41]), are currently under progress at the Laboratoire Joliot-Curie.

This work was supported by the Action Concertée Incitative Informatique, Mathématiques, Physique en Biologie Moléculaire 2004 under the project "ReplicOr", the Agence Nationale de la Recherche under the project "HUGOREP", the program "Emergence" of the Conseil Régional Rhônes-Alpes and by the Natural Science and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] D. Graur and W. H. Li, *Fundamentals of Molecular Evolution*, Sinauer Associates, Sunderland, MA, 1999.

[2] E. Chargaff, "Structure and function of nucleic acids as cell constituents.", *Fed. Proc.*, vol. 10, pp. 654–659, 1951.

[3] A. Beletskii, A. Grigoriev, S. Joyce and A. S. Bhagwat, "Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome.", *J. Mol. Biol.*, vol. 300, pp. 1057–1065, 2000.

[4] M. P. Francino and H. Ochman, "Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences.", *Mol. Biol. Evol.*, vol. 18, pp. 1147–1150, 2001.

[5] P. Green, B. Ewing, W. Miller, P. J. Thomas and E. D. Green, "Transcription-associated mutational asymmetry in mammalian evolution.", *Nat. Genet.*, vol. 33, pp. 514–517, 2003.

[6] J. R. Lobry, "Properties of a general model of DNA evolution under no-strand-bias conditions.", *J. Mol. Evol.*, vol. 40, pp. 326–330, 1995.

[7] J. Mrázek and S. Karlin, "Strand compositional asymmetry in bacterial and large viral genomes.", *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 3720–3725, 1998.

[8] A. C. Frank and J. R. Lobry, "Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms.", *Gene*, vol. 238, pp. 65–77, 1999.

[9] E. R. M. Tillier and R. A. Collins, "The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes", *J. Mol. Evol.*, vol. 50, pp. 249–257, 2000.

[10] M. Bulmer, "Strand symmetry of mutation rates in the beta-globin region", *J. Mol. Evol.*, vol. 33, pp. 305–310, 1991.

[11] M. P. Francino and H. Ochman, "Strand symmetry around the beta-globin origin of replication in primates", *Mol. Biol. Evol.*, vol. 17, pp. 416–422, 2000.

[12] A. Gierlik, M. Kowalczuk, P. Mackiewicz, M. R. Dudek and S. Cebrat, "Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome?", *J. Theor. Biol.*, vol. 202, pp. 305–314, 2000.

[13] M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa and C. Thermes, "Transcription-coupled TA and GC strand asymmetries in the human genome", *FEBS Letters*, vol. 555, pp. 579–582, 2003.

[14] M. Touchon, A. Arneodo, Y. d'Aubenton-Carafa and C. Thermes, "Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes", *Nucl. Acids Res.*, vol. 32, pp. 4969–4978, 2004.

[15] R. Berezney, D. D. Dubey and J. A. Huberman, "Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci.", *Chromosoma*, vol. 108, pp. 471–484, 2000.

[16] F. Jacob, S. Brenner and F. Cuzin, "On the regulation of DNA replication in bacteria", *Cold Spring Harb. Symp. Quant. Biol.*, vol. 28, pp. 329–342, 1963.

[17] S. P. Bell and A. Dutta, "DNA replication in eukaryotic cells", *Annu. Rev. Biochem.*, vol. 71, pp. 333–374, 2002.

[18] O. Hyrien and M. Méchali, "Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos.", *EMBO J.*, vol. 12, pp. 4511–4520, 1993.

[19] S. A. Gerbi and A. K. Bielinsky, "DNA replication and chromatin.", *Curr. Opin. Genet. Dev.*, vol. 12, pp. 243–248, 2002.

[20] D. Schübeler, D. Scalzo, C. Kooperberg, B. van Steensel, J. Delrow and M. Groudine, "Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing.", *Nat. Genet.*, vol. 32, pp. 438–442, 2002.

[21] D. Fisher and M. Méchali, "Vertebrate HoxB gene expression requires DNA replication.", *EMBO J.*, vol. 22, pp. 3737–3748, 2003.

[22] M. Anglana, F. Apiou, A. Bensimon and M. Debatisse, "Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing.", *Cell*, vol. 114, pp. 385–394, 2003.

[23] D. M. Gilbert, "Making sense of eukaryotic DNA replication origins.", *Science*, vol. 294, pp. 96–100, 2001.

[24] M. Touchon, S. Nicolay, B. Audit, E.-B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo and C. Thermes, "Replication-associated strand asymmetries in mammalian genomes: Towards detection of replication origins", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 9836–9841, 2005.

[25] E.-B. Brodie of Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes and A. Arneodo, "From DNA sequence analysis to modelling replication in the human genome", *Phys. Rev. Lett.*, vol. 94, p. 248103, 2005.

[26] P. Lopez and H. Philippe, "Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation", *C. R. Acad. Sci. III*, vol. 324, pp. 201–208, 2001.

[27] E. P. C. Rocha, "Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes", *Trends Microbiol.*, vol. 10, pp. 393–395, 2002.

[28] E. Louie, J. Ott and J. Majewski, "Nucleotide frequency variation across human genes", *Genome Res.*, vol. 13, pp. 2594–2601, 2003.

[29] A. Arneodo, B. Audit, N. Decoster, J.-F. Muzy and C. Vaillant, *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes*, Springer Verlag, Berlin, chap. Wavelet based multifractal formalism: Application to DNA sequences, satellite images of the cloud structure and stock market data, pp. 26–102, 2002.

[30] A. Arneodo, Y. d'Aubenton-Carafa, E. Bacry, P. V. Graves, J.-F. Muzy and C. Thermes, "Wavelet based fractal analysis of DNA sequences", *Physica D*, vol. 96, pp. 291–320, 1996.

[31] L. T. Vassilev, W. C. Burhans and M. L. DePamphilis, "Mapping an origin of DNA replication at a single-copy locus in exponentially proliferating mammalian cells.", *Mol. Cell. Biol.*, vol. 10, pp. 4685–4689, 1990.

[32] Y. B. Yurov and N. A. Liapunova, "The units of DNA replication in the mammalian chromosomes: evidence for a large size of replication units.", *Chromosoma*, vol. 60, pp. 253–267, 1977.

[33] H. G. Callan, "Replication of DNA in the chromosomes of eukaryotes", *Proc. R. Soc. Lond. B Biol. Sci.*, vol. 181, pp. 19–41, 1972.

[34] S. Jun, J. Bechhoefer and B.-Y. Ha, "Diffusion-limited loop formation of semiflexible polymers: Kramers theory and the intertwined time scales of chain relaxation and closing", *Europhys. Lett.*, vol. 64, pp. 420–426, 2003.

[35] Y. Snir and R. D. Kamien, "Entropically driven helix formation", *Science*, vol. 307, p. 1067, 2005.

[36] P. R. Cook, "Predicting three-dimensional genome structure from transcriptional activity", *Nat. Genet.*, vol. 32, pp. 347–352, 2002.

[37] C. Demeret, Y. Vassetzky and M. Méchali, "Chromatin remodelling and DNA replication: from nucleosomes to loop domains.", *Oncogene*, vol. 20, pp. 3086–3093, 2001.

[38] D. A. Jackson and A. Pombo, "Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells.", *J. Cell Biol.*, vol. 140, pp. 1285–1295, 1998.

[39] S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes and A. Arneodo, "Low frequency rhythms in Human DNA sequences: A key to the organization of gene location and orientation?", *Phys. Rev. Lett.*, vol. 93, p. 108101, 2004.

[40] K. E. van Holde, *Chromatin*, Springer-Verlag, New York, 1988.

[41] W. G. Müller, D. Rieder, G. Kreth, C. Cremer, Z. Trajanoski and J. G. McNally, "Generic features of tertiary chromatin structure as detected in natural chromosomes", *Mol. Cell. Biol.*, vol. 24, pp. 9359–9370, 2004.