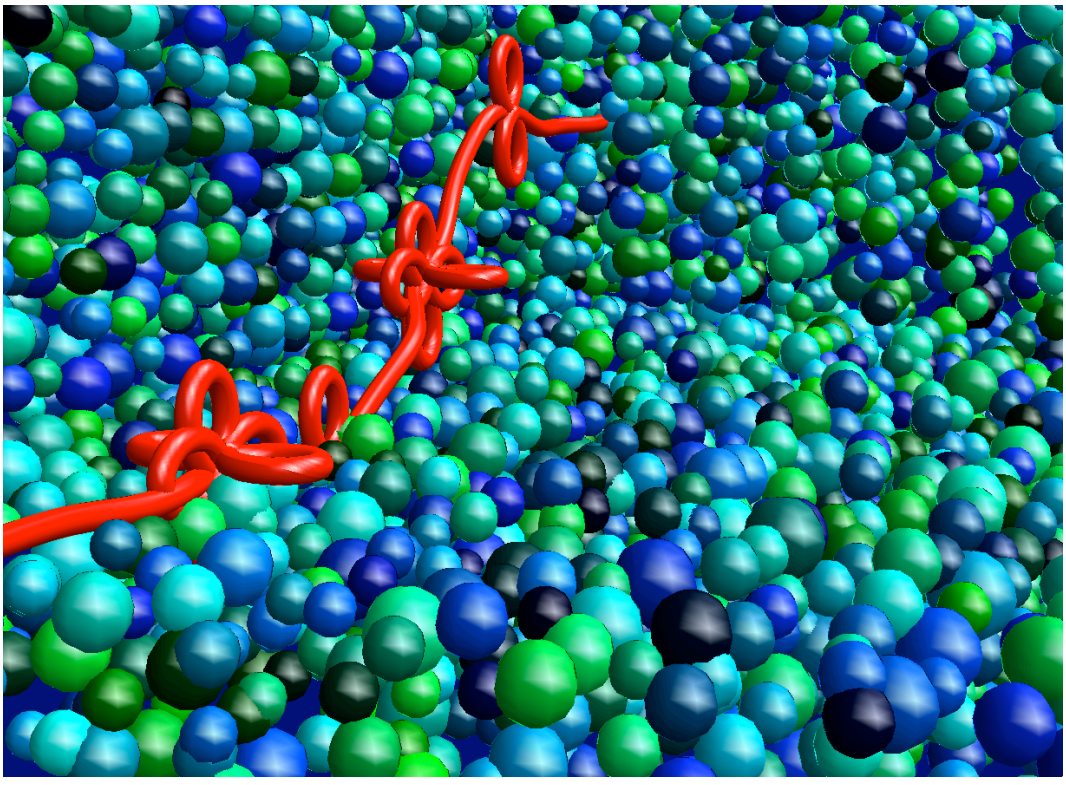
DNA in chromatin: from genome-wide sequence analysis to the modeling of replication in mammals

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 et Laboratoire de Physique, UMR5672, CNRS, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France.
[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.

(January 31st, 2006)



Abstract

Understanding how chromatin is spatially and dynamically organized in the nucleus of eukaryotic cells and how this affects genome functions is one of the main challenges of cell biology. In that context the role of the DNA sequence itself in these condensation-decondensation processes is still debated. In this chapter, we explore large-scale nucleotide compositional fluctuations along the human genome through the optics of the wavelet transform microscope. Analysis of the GC content and of the TA and GC skews reveals the existence of rhythms with characteristic fundamental frequencies that enlighten a remarkable cooperative organization of gene location and orientation. 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 highly relaxational nature of the oscillations observed in the 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 more easily accessible. We show that, in the crowded environment of the cell nucleus, the presence of these intrinsic decondensed structural defects actually predisposes the fiber to spontaneously form multi-looped rosette-like structures that provide an attractive description of genome organization into replication foci that are observed in interphase mammalian nuclei as stable autonomous chromatin domains favoring compartmentalized DNA replication and gene expression. New experimental perspectives are discussed.

Keywords: DNA sequence, Human genome, chromatin fiber loops, compositional bias, skews, wavelet transform, rhythms, replication origins, replication model, transcription.

1 Introduction

The dynamics of folding and unfolding of DNA within living cells plays a major role in regulating many biological processes, such as gene expression, DNA replication, recombination and DNA damage repair[1–4]. As sketched in Figure 1, the genomic DNA of eukaryotic cells is tightly packaged into nucleosomes which constitute the basic units of chromatin[5]. As experimentally detailed by high resolution X-ray analysis[6–8], each nucleosome consists of almost two turns of DNA wrapped around an octamer of core histone proteins. An additional fragment of DNA associated with a linear histone separates successive nucleosomes which are disposed as beads-on-a-string along the DNA. This nucleosomal array is further organized into successive higher order structures[4] including the condensation into the 30nm chromatin fiber, the formation of chromatin loops, up to a full extent of condensation in metaphase chromosomes. Actually, the structure and dynamics of chromatin are under the control of a number of mechanisms involving DNA-protein interactions which may depend upon the nucleotide sequence since DNA is an heteropolymer with locally sequence-dependent physical (mechanical, geometrical, ...) properties. The precise influence of the so-called primary structure (i.e. the sequence) on the organization of chromatin at all scales remains controversial. On a local scale, specific sequence elements have been identified to interact with protein components of chromatin. For instance, some sequence motifs that favor the formation and positioning of nucleosomes were found to be regularly spaced, e.g. the 10bp periodicity[9,10] exhibited by the AA dinucleotide. Alternatively, similar motifs were shown to present long-range correlations along the genome that are a signature of nucleosomes[11–13]. Other DNA regions, the scaffold or matrix attachment regions that constitute the anchor points of chromatin loop domains, are constituted by $\sim$ 1kbp AT-rich sequence patterns[14,15]. On larger scales, the folding of the nucleosomal strings into higher-order structures has been the issue of various models involving, e.g., random packing, coiling into hierarchical helical structures (solenoids)[16–18], or loop-models[14,15,19–21], but the DNA sequence itself was not taken into account. Some recent experimental results suggest that loops might be organized by the active transcription complexes[22–24]. Accordingly gene positions and transcriptional activity would constitute major determinants of the microscopic structure of chromatin that would self-organize in a rather predictable way: the 3D structure would then result to some extent from the DNA primary sequence.

Actually there is much more to be learned about the different stages of compaction of DNA inside the cell nucleus (Fig. 1) from the DNA sequence than commonly thought. The originality of the approach described in this chapter relies on the fact that we are going to extract structural, dynamical and functional informations from the primary DNA sequence using concepts coming from statistical and nonlinear physics and methodologies issued from physics and signal processing[13,26,27]. More precisely, we will mainly use a mathematical microscope, namely the continuous wavelet transform[28,29], to explore the structural complexity of signals generated from adequate codings of the DNA sequences. In a preliminary work[11–13], we have used the space-scale decomposition provided by the wavelet transform to reveal and analyze the scale invariance properties of eukaryotic, eubacterial and archaeal genomic sequences. This study suggests that the existence of long-range correlations, up to distances $\sim$ 20kbp, is the signature of the nucleosomal structure and dynamics of the 30 nm chromatin fiber. Actually these long-range correlations are mainly observed in the DNA bending profiles obtained when using some structural coding of the DNA sequences that accounts for the fluctuations of the local double helix curvature within the nucleosome complex. Because of the approximate

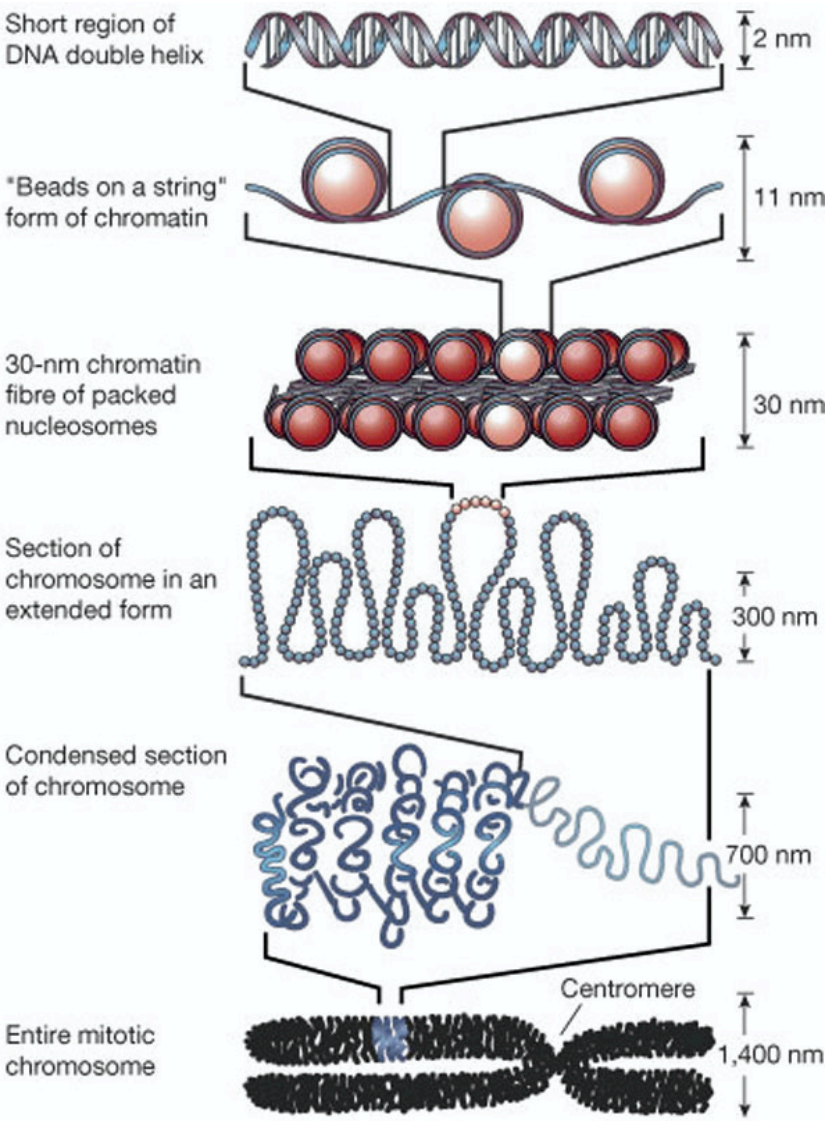


Figure 1: Hierarchical structure of eukaryotic DNA. Each DNA molecule is packed into a mitotic chromosome that is 1/50000 shorter than its extended length (Adapted by permission from Macmillan Publishers Ltd:Nature (Felsenfeld & Groudine[25]), copyright (2003)).

planarity of nucleosomal DNA loops, we have developed some modeling of the thermodynamics of 2D DNA loops in the presence of long-range correlated structural disorder induced by the sequence[30,31]. These long-range correlations clearly favor the autonomous formation of small (i.e. few hundred bp) DNA loops, larger the correlations, smaller the size of the loop, and in turn the propensity of eukaryotic DNA to interact with histones to form nucleosomes. In addition, we have shown that this long-range correlated structural disorder is likely to induce local hyperdiffusion of these loops which provides a very attractive interpretation to the nucleosome repositioning dynamics. Let us emphasize that a recent statistical analysis[32] of nucleosome positioning data obtained recently by Yuan *et al.*[33] for chromosome III of *S. cerevisiae*, has provided a convincing experimental confirmation that long-range correlations in the genomic sequence strongly influence the overall formation and positioning of nucleosomes. In this Chapter, we will keep decreasing the magnification of our mathematical wavelet transform microscope to investigate the complexity of DNA sequences at scales larger than 30 kbp. Our goal is to show that at these large scales, the primary sequence still contains structural and mechanical informations, no longer on DNA, but rather on the 30 nm chromatin fiber and its propensity to form loops and multi-loops structural patterns (Fig. 1) that are likely to stabilize chromatin domains of autonomous DNA replication and gene expression. This study leads us to propose some universal physical mechanism accounting for the self-organization of multi-looped rosettes that would favor the so-called tertiary chromatin structure, prior to the replication and transcription proteic complexes coming into play.

The Chapter is organized as follows. Section 2 is devoted to materials and methods. In Section 3, we show[34,35] that the GC content displays rather regular nonlinear oscillations with two main periods of 110 $\pm$ 20 kbp and 400 $\pm$ 50 kbp, that are well recognized characteristic scales of chromatin loops and loops domains involved in the hierarchical folding of chromatin fibers. These frequencies are also remarkably similar to the size of mammalian replicons and replicon clusters. When further investigating deviations from intrastrand equimolarities between T and A, and between G and C, the so-called TA and GC skews, we corroborate the existence of these rhythms as the footprints of replication and/or transcription mutation bias and we show that the observed relaxational oscillations enlighten a remarkable cooperative gene organization[34,35]. In Section 4, with the specific goal to disentangle the replication and transcription contributions to the TA and GC skews, we analyze 14854 intron-containing genes annotated in the human genome and we show[36,37] that these skews are correlated to each other and display a characteristic step-like profile exhibiting sharp transitions between transcribed (finite bias) and non-transcribed (zero bias) regions. In most sequences, we observe an excess of T over A and of G over C, reflecting transcriptionally coupled mutational processes in germ line cells. In Section 5, we reveal[38,39] the actual existence of replication-associated strand asymmetries by further studying the behavior of the TA and GC skews around the origins of replication experimentally identified. We find that the (TA + GC) skew displays rather sharp upward jumps from negative to positive values at the origin locations. When using the wavelet transform to perform a multiscale analysis of the 22 human autosomal chromosomes skew profiles, we reveal numerous sharp upward jumps that allow us to identify a set of 1012 putative replication initiation zones. Between two neighboring sharp upward jumps, the skew displays a linear decreasing profile leading to a characteristic jagged pattern also observed in mouse and dog genomes[38]. In Section 6, we propose[38,39] a model of replication in mammalian cells with well positioned replication origins and random terminations. A systematic analysis of the gene content around the putative replication origins enlightens a remarkable gene organization with clusters of genes mostly co-oriented with the progression of the replication

fork. This observation suggests that these replication initiation zones are likely to correspond to regions where the chromatin fiber is more open so that DNA be more easily accessible. In Section 7, we show that, in the crowded environment of the cell nucleus, the presence of such intrinsic (sequence dependent) decondensed structural defects actually predisposes the chromatin fiber to spontaneously form rosette-like structures. Prior to any external factors coming into play, these multi-looped rosettes self-organize from the entropy-driven assembling of neighboring defects into clusters by depletive forces. These rosettes provide an attractive description of the compartmentalization of the genome into replication foci that are observed in interphase mammalian nuclei as stable chromatin domains of autonomous DNA replication and gene expression. We conclude, in Section 8, by discussing some new experimental perspectives including *in vivo* visualization of the rosette-like organization of the tertiary chromatin structure via the clustering of replication origins.

2 Materials and Methods

2.1 Data sets

Sequences. Sequence and gene annotation data were retrieved from the Genome Browser of the University of California, Santa Cruz for the human (July 2003 in Sections 3 and 4, May 2004 in Sections 5 and 6), mouse (May 2004) and dog (July 2004) genomes. To delineate the most reliable intergenic regions, transcribed regions were retrieved from "`all_mrna`", one of the largest sets of annotated transcripts. Among transcribed sequences, *sense* (resp. *anti-sense*) genes have the same orientation as the Watson (resp. Crick) strand. To obtain intronic sequences, we used the KnownGene annotation (containing only protein-coding transcripts); when several transcripts presented common exonic regions, only common intronic sequences were retained. For the dog genome, only preliminary gene annotation were available, precluding the analysis of intergenic and intronic sequences.

Sequence repeats. In Section 4, 5, and 6, to exclude repetitive elements that might have been inserted recently and would not reflect long-term evolutionary patterns, we used `REPEATMASKER`[40] leading to a reduction of $\sim 40-50\%$ of the human sequence length.

Human intron sequences. In Section 4, human intron sequences were downloaded from RefGene (April 2003) at UCSC. When several genes presented identical exonic sequences, only the longest one was retained; repeated elements were removed with `REPEATMASKER`. The introns of each gene were taken as a single sequence; introns without repeats were also taken as a single sequence; to avoid the skew associated with splicing signals, 560 bp were removed at both intron extremities. When the resulting intron sequences were shorter than 1120 bp, they were not considered for the analysis, leading to 14854 intron-containing genes.

Human replication origins. Nine replication origins were examined; namely those situated near the genes MCM4[41], HSPA4[42], TOP1[43], MYC[44], SCA-7[45], AR[45], DNMT1[46], Lamin B2 [47] and β-globin[48].

Sequence Alignments. Mouse and dog regions homologous to the six human regions shown in Figure 9 were retrieved from University of California, Santa Cruz (Human Synteny). Mouse intergenic sequences were individually aligned by using `PIPMAKER`[49], leading to a total

of 150 conserved segments longer than 100 bp (> 70% identity) and corresponding to a total of 26 kbp (5.3% of intergenic sequences).

2.2 Coding rules

GC content. GC content fluctuations in the human genome were computed in adjacent (non-overlapping) 1-kbp windows.

Strand asymmetries. The TA and GC skews were calculated in non-overlapping windows of size 1-kbp as:

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

where n_{A}, n_{C}, n_{G} and n_{T} are respectively the numbers of A, C, G and T in the windows. Because of the observed correlation between the TA and GC skews (Section 4), we also considered the total skew

$$S = S_{\mathrm{TA}} + S_{\mathrm{GC}}. \tag{2}$$

From the skews $S_{\mathrm{TA}}(n)$, $S_{\mathrm{GC}}(n)$ and $S(n)$, obtained along the sequences, where n is the position (in kbp units), from the origin, we also computed the cumulative skew profiles:

$$\Sigma_{\mathrm{TA}}(n) = \sum_{j=1}^{n} S_{\mathrm{TA}}(j), \qquad \Sigma_{\mathrm{GC}}(n) = \sum_{j=1}^{n} S_{\mathrm{GC}}(j), \tag{3}$$

and

$$\Sigma(n) = \sum_{j=1}^{n} S(j). \tag{4}$$

2.3 Space-scale analysis based on the continuous wavelet transform

The continuous wavelet transform (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[13,27–29]. When using the successive derivatives of the Gaussian function as analyzing wavelets, namely

$$g^{(N)}(x) = (-1)^N d^N g^{(0)}(x)/dx^N, \tag{5}$$

where

$$g^{(0)}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \tag{6}$$

then the WT of a signal s takes the following simple expression:

$$\begin{aligned} W_{g^{(N)}}[s](x,a) &= \frac{1}{a}\int_{-\infty}^{+\infty} s(y) g^{(N)}\left(\frac{y-x}{a}\right) dy, \\ &= \frac{d^N}{dx^N} W_{g^{(0)}}[s](x,a), \end{aligned} \tag{7}$$

where x and $a(>0)$ are the space and scale parameters respectively. Equation (7) shows that the WT computed with $g^{(N)}$ at scale a is nothing but the Nth derivative of the signal $s(x)$ smoothed by a dilated version $g^{(0)}(x/a)$ of the Gaussian function. This property is at the heart of various applications of the WT microscope as a very efficient multi-scale singularity tracking technique[13,50]. Actually the skeleton of the WT provides a space-scale partitioning that is likely to contain all the information on the singularities of the signal considered. The WT skeleton is defined, at each scale a, by the set of all the points x_i that corresponds to a local maximum of $|W_\psi[s](x,a)|$ and then by connecting these points across scales into the so-called maxima lines[13,27,29]. In Section 6, the ability of identifying sharp jumps in noisy skew profiles from the WT skeleton will be at the heart of the methodology we will propose to detect the origins of replication in mammalian genomes[38,39].

One of the main advantages of the WT is its adaptative ability to perform time-frequency analysis[28,29] when using complex analyzing wavelets like the Morlet's wavelet:

$$\psi_M(x) = \frac{1}{\sqrt{2\pi}} e^{i\omega x} \left(e^{-x^2/2} - \sqrt{2} e^{-\omega^2/4} e^{-x^2} \right), \tag{8}$$

where the second term in the r.h.s. is negligible for large ω values ($\omega \gtrsim 5$). The *scale-spectrum* of a signal s of total length L is defined as:

$$\Lambda(a) = \frac{1}{L} \int_0^L |W_{\psi_M}[s](x,a)|\, dx. \tag{9}$$

3 Low frequency rhythms in Human DNA sequences

3.1 GC content

The recent sequencing of the human genome[51] has opened the door to the statistical analysis of genomic sequence complexity on a chromosomal scale. One of the most striking features of eukaryotic chromosomes is their large-scale variations in base composition. In particular, an extraordinary large heterogeneity of the GC content is observed in mammalian genomes; this has led Bernardi[52,53] to propose a description of these genomes in terms of a mosaic organization of domains of relatively constant GC levels, originally called *isochores*. The isochore model is a topic of controversial discussions[51,54–60]. Nevertheless there is definite evidence that the compositional heterogeneity in a DNA sequence correlates with its GC content[57] which is unanimously recognized as a fundamental property of the chromosomal DNA and is likely to be one of the possible key to the understanding of the organization of eukaryotic genomes[52,53,56,57]. Indeed the GC content has a taxonomy value[61]; it determines the amino acid composition of the encoded proteins and is also related to codon usage in genes [62]. Moreover there is conspicuous evidence[56,57,63] that GC-rich and GC-poor regions respectively match the cytogenic R and G bands and correlate well with early and late replicating domains in the cell cycle. GC-rich regions correspond to regions of very high density of genes including the housekeeping genes and associated CpG islands and also of short inter-dispersed repetitive DNA elements (SINES, Alu)[51]. In contrast, GC-poor regions are definitively poor in genes, predominantly tissue-specific genes containing rather long introns, but are relatively rich in long inter-disperse repetitive DNA elements (LINES)[51] that are significantly more abundant in these regions. The GC content has also some impact on the structure of chromatin. For example, it has been suggested[15] that the proteic chromosomal scaffold that serves

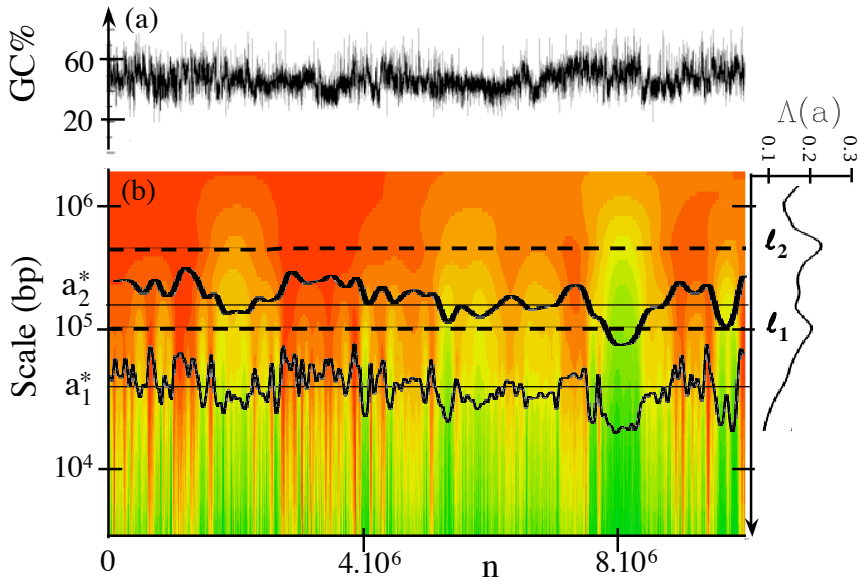


Figure 2: Space-scale representation of the GC content of a 10 Mbp long fragment of human chromosome 22 when using a Gaussian smoothing filter $g^{(0)}(x)$ (Eq. (6)). (a) GC content fluctuations computed in adjacent 1 kbp intervals. (b) Color coding of the convolution product $W_{g^{(0)}}[GC](n,a) = (GC * g^{(0)}(\cdot/a))(n)$ using 256 colors from black (0) to red (max); superimposed are shown the smoothed GC profiles obtained at scales $a_1^* = 40$ kbp and $a_2^* = 160$ kbp. On the right hand side is shown vertically the scale (frequency^{-1}) spectrum $\Lambda(a)$ (Eq. (9)) computed with the complex Morlet wavelet (Eq. (8)) over the entire chromosome 22. The horizontal dashed lines in the color picture correspond to the two main characteristic oscillations length $l_1 = 100$ kbp and $l_2 = 400$ kbp.

as a structural skeleton for the organization of chromatin loops is much less tightly folded (to the benefit of replication and transcription processes) in GC-rich than in GC-poor regions.

In Figure 2 are reported the results[34,35] of a space-scale decomposition of the GC content fluctuations of a 10 Mbp long fragment of human chromosome 22, when using the Gaussian $g^{(0)}(x)$ (Eq. (6)) as smoothing filter. This decomposition reveals that for distances larger than $\sim\ 20 - 30$ kbp, the GC content can no longer be considered as fluctuating homogeneously (at smaller scales the fluctuations cannot be distinguished from a monofractal long-range correlated noise[35]); it instead displays rather regular nonlinear oscillatory behavior. The corresponding scale spectrum $\Lambda(a)$ (Eq. (9)) shown vertically in Figure 2(b), reveals the existence of two main broad peaks corresponding to the scales $l_1 = 100 \pm 20$ kbp and $l_2 = 400 \pm 50$ kbp respectively, that emerge from a continuous background. The former is the characteristic length of the basic oscillations obtained with the low-pass filtering scale $a_1^* = 40$ kbp, although one may observe from time to time oscillations that have a larger length ($\sim 2l_1 = 200$ kbp). If one uses a larger filtering scale $a_2^* = 160$ kbp, in order to smooth

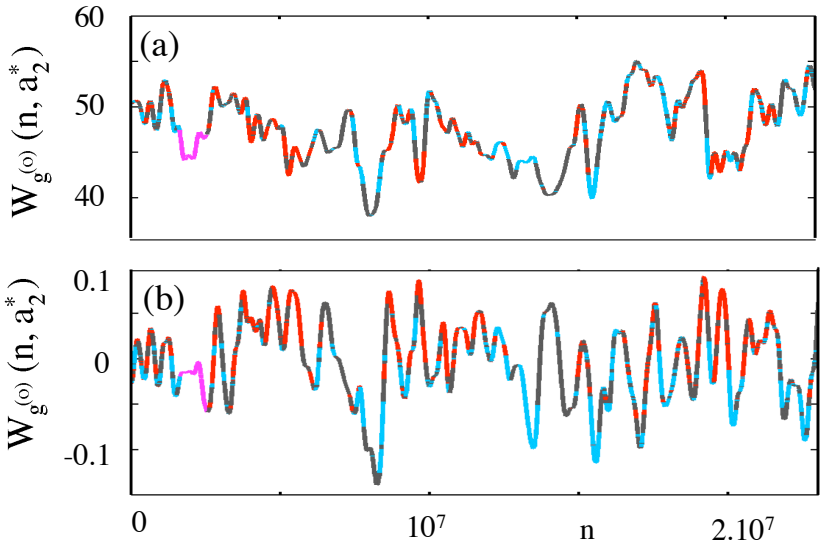


Figure 3: Compositional oscillations observed in the human chromosome 22 fragment (23 Mbp, `NT_011520.8`) after low-pass filtering at scale $a_2^* = 160$ kbp (see Fig. 2). (a) GC content. (b) Total skew $S = S_{TA} + S_{GC}$ (Eq. (2)). The red (blue) portions of the profiles correspond to the location of sense (antisense) genes that have the same (opposite) orientation than the sequence. The location of the immunoglobulin locus is shown in pink.

both the "small scales" (high frequencies) colored noise component and the basic oscillations of scale l_1, one gets some oscillatory profile with a fundamental length $l_2 = 400$ kbp as illustrated in Figure 3(a). Let us point out that the investigation of large GC-rich fragments in various human chromosomes reveals similar periodicities[34], namely $l_1 = 120 \pm 30$ kbp, $l_2 = 410 \pm 60$ kbp for chromosome 11 (24 Mbp, `NT_033899.3`), $l_1 = 130 \pm 30$ kbp, $l_2 = 420 \pm 60$ kbp for chromosome 14 (68 Mbp, `NT_02637.9`), and $l_1 = 110 \pm 20$ kbp, $l_2 = 390 \pm 50$ kbp for chromosome 21 (29 Mbp, `NT_011512.7`).

A possible interpretation of these low-frequency rhythms observed in the GC content is of structural nature and is related to recent experimental and numerical observations of the high-order hierarchical folding of chromatin into fibers and loops of different sizes. 100 kbp is typically the size of DNA loops that are observed by electron microscopy[14,64,65] to be held together by a longitudinal network of scaffolding proteins in histone-depleted chromosomes, favoring a radial loop/scaffold model[15,66,67] of metaphase chromosome structure. 100 kbp is also very close to the chromosome loop size ($\sim$ 80 kbp) estimated from physical measurements of the dynamics of force relaxation in single mitotic chromosomes[68]. Furthermore, 400 kbp is likely to be the size of larger chromatin loops made of a few basic loops that have some coherent dynamical behavior during the cell cycle possibly governed by the mechanisms that underlie replication and transcription processes[69–73]. As an alternative to the loop/scaffold[15,66,67] and multi-loop subcompartment[74] models, 100 kbp and 400 kbp might well be characteristic scales involved into the successive levels of helical coiling of chromatin into fibers or tubes of diameters ranging from 30 to 700 nm observed during interphase using either electron or light microscopy[16,75–80].

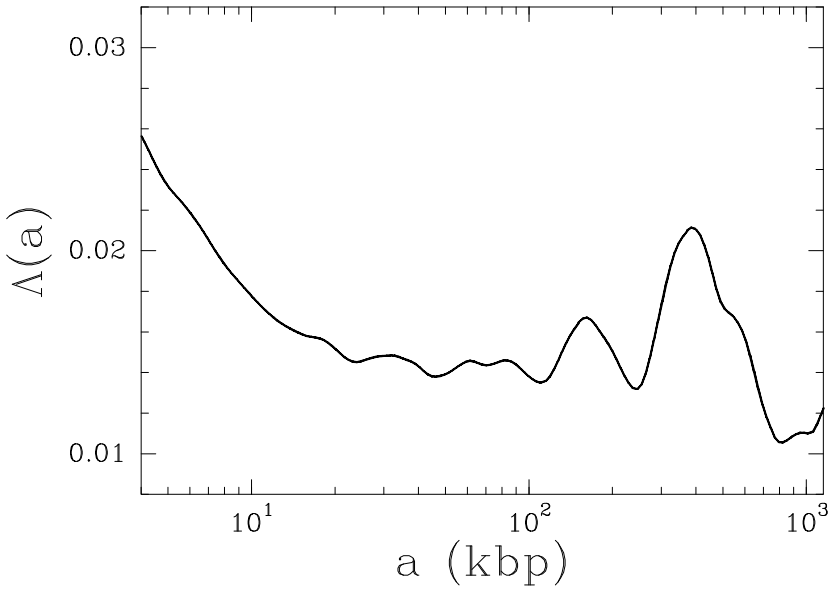


Figure 4: Scale spectrum $\Lambda(a)$ vs (a) of the total skew S (Eq. (2)) computed over the entire chromosome 22 of the human genome.

3.2 Strand asymmetries

An alternative interpretation of the rhythms observed in the GC content is of functional nature and is a direct consequence of the observation that 100 kbp and 400 kbp correlate well with the replicon sizes observed in warm-blooded vertebrate organisms[81–83]. Early experimental investigations[84,85] of replicon size by fiber auto-radiography or fluorography have led to the classical view that mammalian replicons are heterogeneous in size but that most fall into the range of 30-450 kbp with the most frequent sizes in the range 75-150 kbp. Furthermore there are experimental evidences[81,83,84] that replicons are likely to be in groups, the so-called replicon foci, with all the replicons in each group firing at similar time in the S-phase. Newer results obtained with modern replicon-mapping methods clearly show the existence of much larger replicons (the largest ones being as large as a few Mbp) than previously thought, requiring most or all the S-phase to be completed[83,86,87]. In particular, the average size of a mammalian replicon has been reconsidered to be more likely ~ 500 kbp[86,87].

According to the second parity rule[88,89], under no strand-bias conditions, each genomic DNA strand should present equimolarities[90,91] of A and T and of G and C. Deviations from intrastrand equimolarities have been extensively studied in prokaryote, organelle and viral genomes for which they have been used to detect the origins of replication[92–95]. During replication, mutational events can affect the leading and lagging strands (see Section 4.1) differently and one strand can be more efficiently repaired than the other one, leading to strand compositional asymmetry. In eukaryotes, the existence of compositional biases has been debated and most attempts to detect the replication origins from strand compositional asymmetry have been inconclusive. When using our WT microscope to perform a space-scale analysis of both the S_{TA} and S_{GC} skews of human chromosome 22, one gets oscillatory profiles similar to those obtained for the GC content, with still two main characteristic lengths $l_1 = 140 \pm 20$ kbp and $l_2 = 400 \pm 40$ kbp as shown in Figure 4 where the corresponding scale

spectrum of the total skew S (Eq.(2)) displays two main bumps, the latter at 400 kbp being the most pronounced. In Figure 3(b) is shown the oscillatory skew profile obtained for the smoothing scale $a_2^* = 160$ kbp. This profile displays rather regular oscillation trends of basic length $\sim$ 400 kbp. Quite remarkably this oscillatory skew profile provides a guide for the organization of the spatial location and orientation of the (largest) genes[34,35]: *sense* genes with the same orientation as the sequence are located around the positive maxima of the oscillations (among transcribed sequences, this corresponds to $79.6 \pm 1.9\%$ (ch.22), $84.0 \pm 2.6\%$ (ch.11), $89.2 \pm 1.2\%$ (ch.14) and $88.1 \pm 2.4\%$ (ch.21) of 1 kbp fragments that have the same orientation as the Watson strand), while *antisense* genes are quite symmetrically located around the minima (mainly negative). Let us point out that since these skew oscillations are also observed in large intergenic regions (but with smaller amplitude), they may arise from replication mutation biases. Indeed, as we will elaborate about in the next sections, these oscillations of rather marked relaxational character are likely to reflect some correlation between gene organization into clusters with preferential gene orientation and replication.

For more details on the existence of low frequency rhythms in DNA sequences, we refer the reader to the PhD manuscripts of E.B. Brodie of Brodie[96] and S. Nicolay[97].

4 Transcription-coupled strand asymmetries in the human genome

During genome evolution, mutations do not occur at random as illustrated by the diversity of the nucleotide substitution rate values[98–101]. 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[88,89], have been extensively studied during the past decade and the observed skews have been attributed to asymmetries intrinsic to the replication or to the transcription processes. Asymmetries of substitution rates coupled to transcription have been mainly observed in prokaryotes[102–104], with only preliminary results in eukaryotes. In the human genome, excess of T was observed in a set of gene introns[105] and some large-scale asymmetry was observed in human sequences but they were attributed to replication[106]. Only recently, a comparative analysis of mammalian sequences demonstrated a transcription-coupled excess of G+T over A+C in the coding strand[107]. In contrast to the substitution biases observed in bacteria presenting an excess of C→T transitions, these asymmetries are characterized by an excess of purine (A→G) transitions relatively to pyrimidine (T→C) transitions. These might be a by-product of the transcription-coupled repair mechanism acting on uncorrected substitution errors during replication[108]. In this section, we report the results of a genome-scale analysis of human genes that definitely establish the existence of transcription-coupled nucleotide biases[36,37].

4.1 Strand asymmetries in human gene sequences

We have started examining nucleotide compositional strand asymmetries in transcribed regions of human sequences[36]. We have computed the $S_{\rm TA}$ and $S_{\rm GC}$ skews (Eq.(1)) for intron sequences since, in contrast to exonic sequences, they can be considered as weakly selected sequences. For each gene, we have concatenated all the introns in a unique sequence (see Section 2.1). The distributions of the TA and GC skews, computed on the 14854 intro-containing

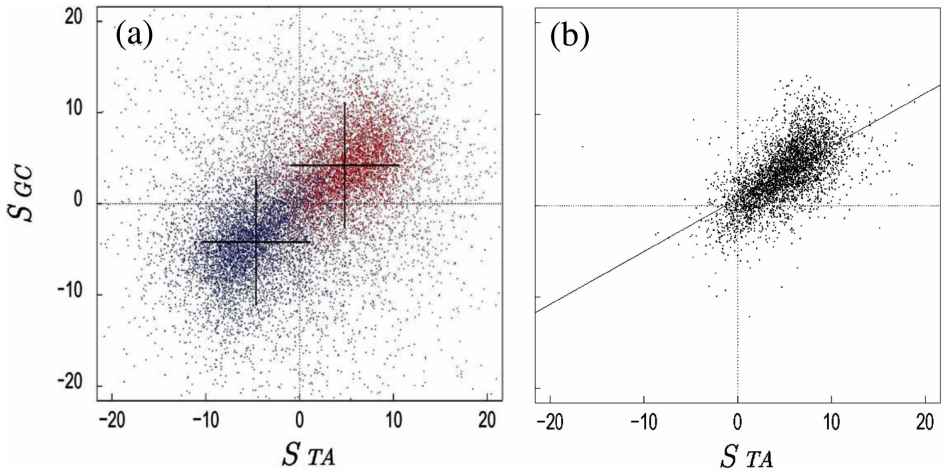


Figure 5: (a) $S_{\rm TA}$ and $S_{\rm GC}$ skews in human introns[36]: each point corresponds to one of the 14854 intron-containing genes; repeated elements are removed from the analysis (Section 2.1); red points correspond to sense genes (7508) with the same orientation as the Watson strand; blue points correspond to anti-sense genes (7346) with opposite orientation; black crosses represent the standard deviations of the distributions. (b) Correlation between $S_{\rm TA}$ and $S_{\rm GC}$ skews determined on the coding strand from intronic regions without repeats: each point corresponds to a gene for which the total length of intronic regions is $l > 25$ kbp (7797 genes); Pearson's correlation coefficient $r = 0.61$ (the slope of the regression line is 0.58).

genes, present positive mean values for sense genes (7508), namely $\bar{S}_{\rm TA} = 4.72 \pm 0.07\%$ and $\bar{S}_{\rm GC} = 2.97 \pm 0.07\%$, and nearly opposed values for anti-sense genes (7346), namely $\bar{S}_{\rm TA} = -4.56 \pm 0.07\%$ and $\bar{S}_{\rm GC} = -3.05 \pm 0.07\%$. When removing the repeated sequences from the analysis (Section 2.1), the TA and GC biases are not strongly altered (Fig. 5(a)). When examined on the coding strand, the mean values for all intron sequences without repeats present significant excess of T over A, namely $\bar{S}_{\rm TA} = 4.49 \pm 0.01\%$, and excess of G over C, namely $\bar{S}_{\rm GC} = 3.29 \pm 0.01\%$ (after appropriately removing intron extremities, see Section 2.1). The corresponding probability density functions (pdf) of $S_{\rm TA}$ and $S_{\rm GC}$ skews when computed from the intronic sequences of the whole set of 14854 intron-containing genes (after removing the non-coding regions closer than 560 bp to an exon) are shown in Figure 6 in a semi-logarithmic representation. For both sense (Figure 6(a)) and anti-sense (Figure 6(b)) genes, one observes the presence, for both skews, of large tails that clearly indicates some departure from a parabolic profile, the signature of Gaussian statistics.

A question of interest is the possible existence of correlations between $S_{\rm TA}$ and $S_{\rm GC}$ skews in intronic sequences without repeats. When all genes are considered, only small correlation is observed (Pearson's correlation coefficient $r = 0.09$). However, the values of the skews from small genes turn out to be highly noisy. When one excludes these small genes, $S_{\rm TA}$ and $S_{\rm GC}$ present larger correlation, e.g. $r = 0.45$ for genes with total intron length $l > 10$ kbp and $r = 0.61$ for genes with $l > 25$ kbp as illustrated in Figure 5(b). Let us point out that $S_{\rm TA}$ and $S_{\rm GC}$ present weak correlation with the intronic GC content as well as with the sequence length and this even if only the large genes are considered[36].

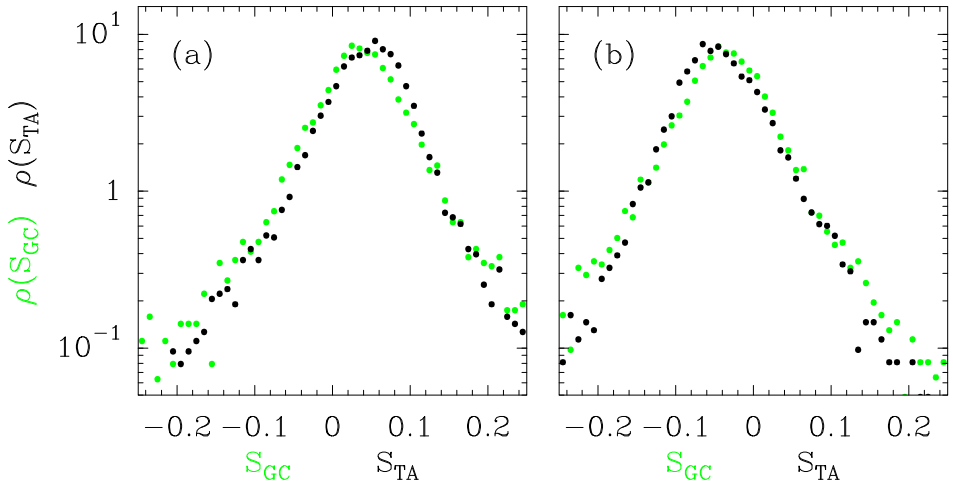


Figure 6: Probability density functions of the skews $S_{\rm TA}$ (•) and $S_{\rm GC}$ (•) values computed from the intronic sequences of the 14854 intron-containing genes after removing repeated sequences. (a) Sense genes; (b) antisense genes.

4.2 Transcription induced step-like skew profiles in the human genome

In order to compare the TA and GC asymmetry values in transcribed regions to those in the neighboring intergenic sequences, we have computed $S_{\rm TA}$ and $S_{\rm GC}$ in adjacent 1 kbp windows along the genome sequence[36,37]. In Figure 7 are reported the mean values of these skews for all genes as a function of the distance to the 5'- or 3'- end. At the 5' gene extremities (Figure 7(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 $\bar{S}_{\rm TA}$ and between 3 and 5% for $\bar{S}_{\rm GC}$. At the gene 3'- extremities (Figure 7(b)), the TA and GC skews also exhibit transitions from significantly large values in transcribed regions to very small values in untranscribed regions. However, 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. In pluricellular organisms, mutations responsible for the observed biases are expected to occur in germ-line cells. It could happen that gene 3'- ends annotated in the databank differ from the poly-A sites effectively used in the germ-line cells. Such differences would then lead to some broadening of the skew profiles.

4.3 A model for transcription-coupled TA and GC skews

As shown in Figure 7, TA and GC biases are specifically observed in transcribed sequences indicating that each of them clearly results from transcription-coupled processes acting in germ-line cells. This observation is reinforced by the observed correlation between $S_{\rm TA}$ and $S_{\rm GC}$ (Figure 5(b)). Indeed, according to this hypothesis, $S_{\rm TA}$ and $S_{\rm GC}$ are likely to increase simultaneously with transcription. How many genes have biased sequences? When comparing [36] the observed biases to those expected for random sequences with same length and same (T+A) composition, 64% of genes are found to present significant TA bias (p-values $< 10^{-2}$).

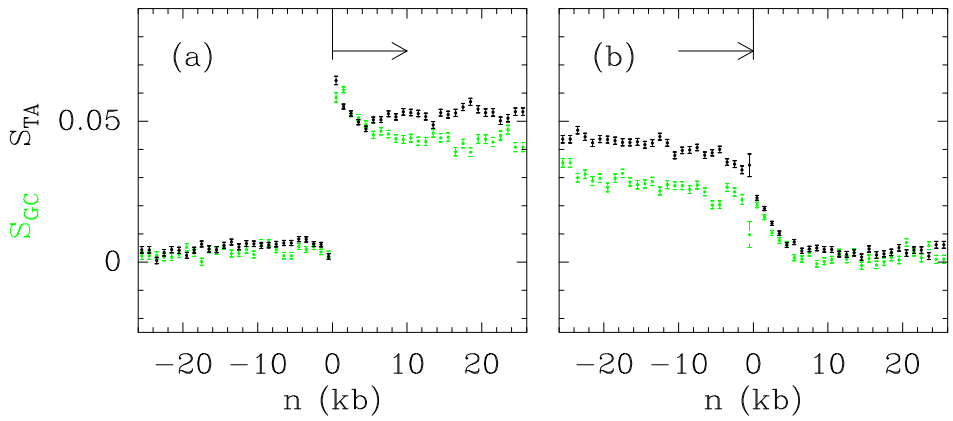


Figure 7: TA (•) and GC (•) skew profiles in the regions surrounding 5' and 3' gene extremities [36]. S_{TA} and S_{GC} were calculated in 1 kbp windows starting from each gene extremities in both directions. In abscissa is reported the distance (n) of each 1 kbp window to the indicated gene extremity; 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 for all 1 kbp windows at the corresponding abscissa. Error bars represent the standard error of the means.

When considering only larger genes, this proportion increases to 82% (total intron length $l > 10$ kbp) and 86% ($l > 25$ kbp) respectively. These results indicate that in germ-line cells, a large majority of genes are expressed.

A recent study [107] showed a transcription-coupled excess of purine transitions and a deficit of pyrimidine transitions in a small set of human genes. To examine if these transition rates might explain the strand asymmetries measured in Figures 5 to 7, for the whole set of genes, we have performed numerical calculations of the composition at equilibrium of a DNA sequence (given the substitution rates) [36]. When supposing that transcription alters transition rates only, we obtained a value of $S_{\mathrm{TA}} = 4.7\%$ similar to our observations, while the value of $S_{\mathrm{GC}} = 7.8\%$ significantly exceeds the value found in our study. This led us to suppose that GC transversions might also produce strand asymmetry in eukaryotes [36]. During evolution, both processes would have been active in germ-line cells.

More recently, we have extended this study of strand asymmetries in intron sequences to evolutionarily distant eukaryotes [37]. When appropriately examined, all genomes present transcription-coupled excess of T over A ($S_{\mathrm{TA}} > 0$) in the coding strand. In contrast, GC skew is found positive in mammals and plants but negative in invertebrates suggesting different mutation repair mechanisms associated to transcription in vertebrates and invertebrates. For more details on the existence of transcription-coupled strand asymmetries in eukaryotic genomes, we refer the reader to the PhD manuscript of M. Touchon [109].

The results reported in Figure 7 suggest that S_{TA} and S_{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 [96,97,109]. 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 replication origins[83]. The following sections will be devoted to the study of replication-associated strand asymmetries in mammalian genomes.

5 Replication-associated strand asymmetries in the human genome

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[110], this process starts with the binding of some "initiator" protein to a specific "replicator" DNA sequence called *origin of replication*. The recruitment of additional factors initiate the bi-directional progression of two divergent replication forks along the chromosome. As illustrated in Figure 8(a), one strand is replicated continuously (leading strand), while the other strand is replicated in discrete steps towards the origin (lagging strand). In eukaryotic cells, this event is initiated at a number of replication origins and propagates until two converging forks collide at a terminus of replication[111]. The initiation of different replication origins is coupled to the cell cycle but there is a definite flexibility in the usage of the replication origins at different developmental stages[112–116]. Also, it can be strongly influenced by the distance and timing of activation of neighboring replication origins, by the transcriptional activity and by the local chromatin structure[113–116]. Actually, sequence requirements for a replication origin vary significantly between different eukaryotic organisms. In the unicellular eukaryote *Saccharomyces cerevisiae*, the replication origins spread over 100-150 bp and present some highly conserved motifs[111]. However, among eukaryotes, *S. cerevisiae* seems to be the exception that remains faithful to the replicon model. In the fission yeast *Schizosaccharomyces pombe*, there is no clear consensus sequence and the replication origins spread over at least 800 to 1000 bp[111]. In multicellular organisms, the nature of initiation sites of DNA replication is even more complex. Metazoan replication origins are rather poorly defined and initiation may occur at multiple sites distributed over a thousand of base pairs[117]. The initiation of replication at random and closely spaced sites was repeatedly observed in *Drosophila* and *Xenopus* early embryo cells, presumably to allow for extremely rapid S phase, suggesting that any DNA sequence can function as a replicator [112,118,119]. A developmental change occurs around midblastula transition that coincides with some remodeling of the chromatin structure, transcription ability and selection of preferential initiation sites[112,119]. Thus, although it is clear that some sites consistently act as replication origins in most eukaryotic cells, the mechanisms that select these sites and the sequences that determine their location remain elusive in many cell types[120,121]. As recently proposed by many authors[122–124], the need to fulfill specific requirements that result from cell diversification may have led high eukaryotes to develop various epigenetic controls over the replication origin selection rather than to conserve specific replication sequence. This might explain that only very few replication origins have been identified so far in multicellular eukaryotes, namely around 20 in metazoa and only about 10 in human[41–48]. Along the line of this epigenetic interpretation, one might wonder what can be learned about eukaryotic DNA replication from DNA sequence analysis.

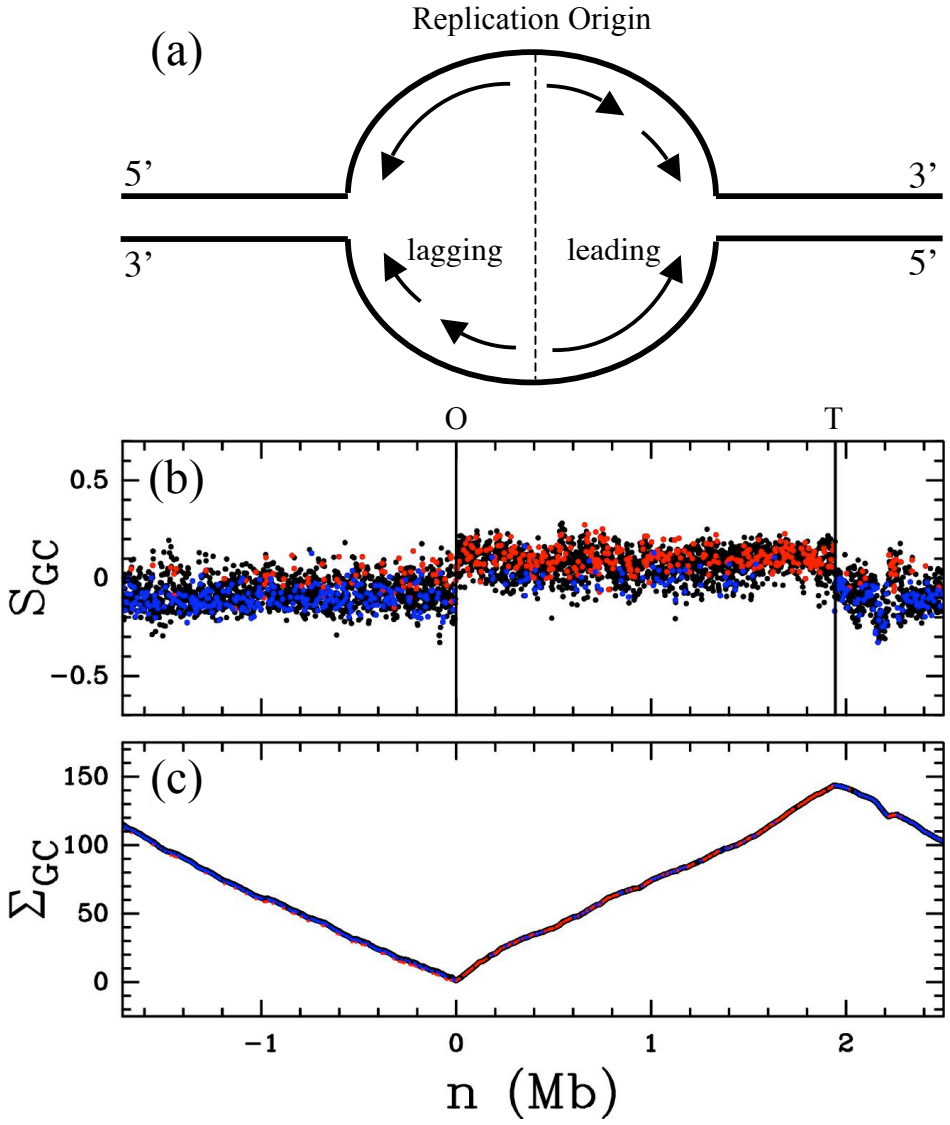


Figure 8: (a) Schematic representation of the divergent bi-directional progression of the two replication forks from the replication origin. (b) S_{GC} calculated in 1 kbp windows along the genomic sequence of *Bacillus subtilis*. (c) Cumulated skew Σ_{GC}. The vertical lines correspond respectively to the replication origin (O) and termination (T) positions. In (b) and (c), red (blue) points correspond to sense (antisense) genes that have the same (opposite) orientation than the sequence.

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

As mentioned in Section 3.2, the existence of replication associated strand asymmetries has been mainly established in bacterial genomes[91–95]. As illustrated in Figure 8, the GC and TA skews abruptly switch sign (over few kbp) from negative to positive values at the replication origin and in the opposite direction from positive to negative values at the replication terminus. This step-like profile is characteristic of the replicon model[110]. In *Bacillus subtilis*, as in most bacteria, the leading (resp. lagging) strand (Fig. 8(a)) is generally richer (resp. poorer) in G than in C (Fig. 8(b)), and to a lesser extent in T than in A (data not shown). This typical pattern is particularly clear when plotting the cumulated skews Σ_{GC} (Fig. 8(c)) and Σ_{TA} (Eq. (3)); both present decreasing (or increasing) profiles in regions situated 5' (or 3') to the origin, displaying a characteristic $\vee$-shape pointing to the replication origin position (similarly a characteristic $\wedge$-shape is observed at the terminus position). The research of $\vee$ patterns in the cumulated skews has been extensively used as a strategy to detect the position of the (unique) replication origin in (generally circular) bacterial genomes[92–95].

As shown in Figures 8(b) and 8(c), when looking at the gene organization around the replication origin of *Bacillus subtilis*, one observes that most of the sense (resp. antisense) genes are preferentially on the right (resp. left) of the replication origin. 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[125–128].

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

In eukaryotes, the existence of compositional biases is unclear and most attempts to detect the replication origins 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* replication origin in primates do not support the existence of mutational bias between the leading and the lagging strands[92,129,130]. 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[131]. With the same methodology as the one developed in Section 4 for gene extremities, we present in this section analyses of strand asymmetries flanking experimentally determined human replication origins[38,39].

As shown in Figure 9, most of the known replication origins in the human genome correspond to rather sharp (over several kbp) transitions from negative to positive S_{TA} and S_{GC} skew values that clearly emerge from the noisy background. This is reminiscent of the behavior observed in Figure 8 for *Bacillus subtilis*, except that the leading strand is relatively enriched in T over A and in G over C. This observation is even more patent when looking at the cumulated skew Σ_{TA} and Σ_{GC} profiles that both display characteristic $\vee$-shapes pointing to the experimentally identified initiation zones. 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 4 that transcription generates positive TA and GC skews on the coding strand[36,37,132], which explains that larger jumps are observed when the sense and/or the antisense genes are

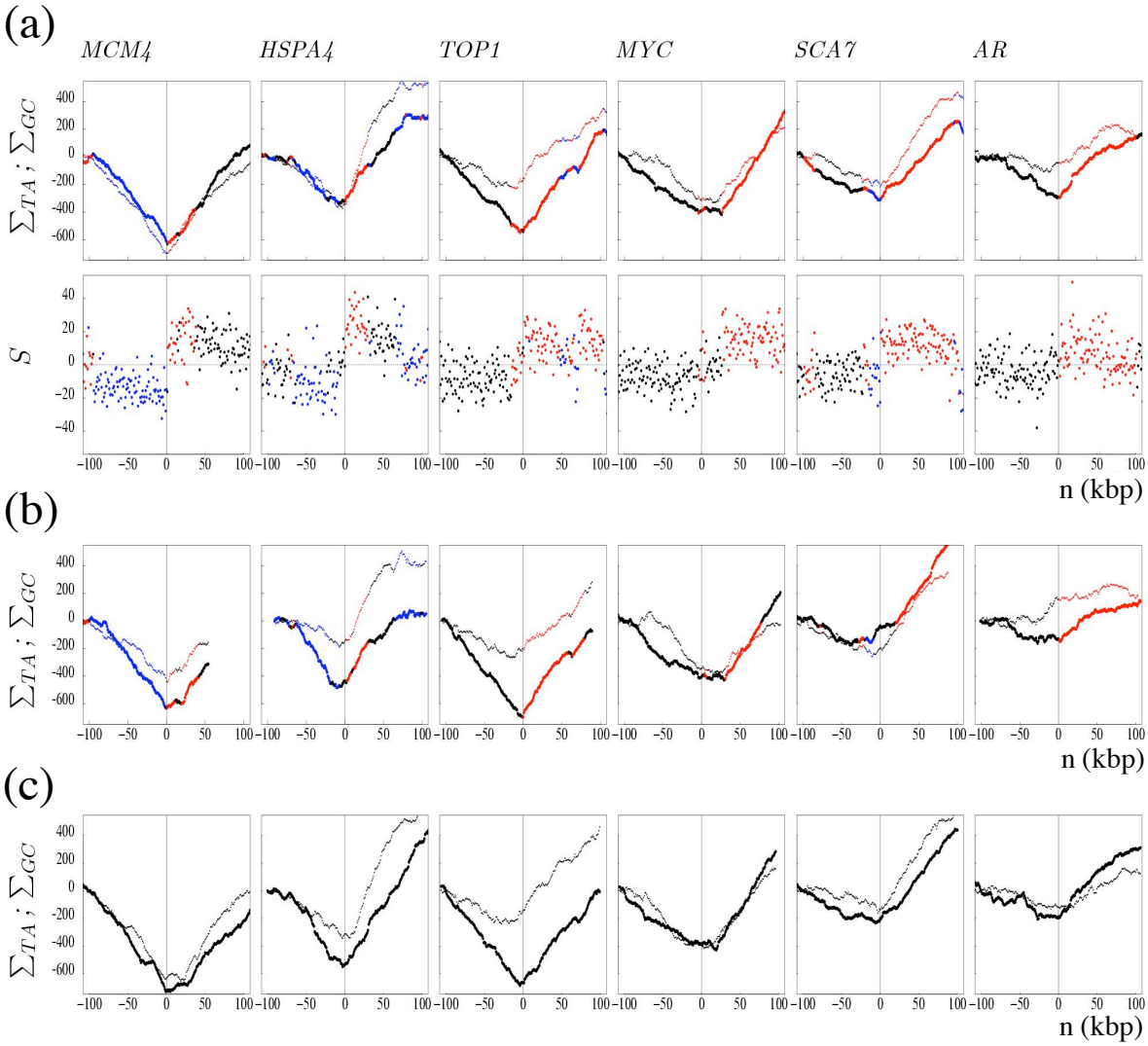


Figure 9: TA and GC skew profiles around experimentally determined human replication origins[38]. (a) The skew profiles were determined in 1 kbp windows in regions surrounding ($\pm$100 kbp without repeats) experimentally determined human replication origins (see Section 2.1). (Upper) TA and GC cumulated skew profiles Σ_{TA} (thick line) and Σ_{GC} (thin line). (Lower) Skew S calculated in the same regions. The ΔS amplitude associated with these origins, calculated as the difference of the skews measured in 20 kbp windows on both sides of the origins, are: MCM4 (31%), HSPA4 (29%), TOP1 (18%), MYC (14%), SCA7 (38%), and AR (14%). (b) Cumulated skew profiles calculated in the six regions of the mouse genome homologous to the human regions analyzed in (a). (c) Cumulated skew profiles in the six regions of the dog genome homologous to human regions analyzed in (a). The abscissa (n) represents the distance (in kbp) of a sequence window to the corresponding origin; the ordinate represents the values of S given in percent. The colors have the following meaning: red, sense genes (coding strand identical to the Watson strand); blue, antisense genes (coding strand opposite to the Watson strand); black, intergenic regions. In (c), genes are not represented.

	S_{TA}	S_{GC}	S	l	G+C, %
Intergenic (*H.s.*) all	3.9 ± 0.4	3.0 ± 0.4	6.9 ± 0.4	487	42
Intergenic (*H.s.*) ncr.	4.0 ± 0.4	3.0 ± 0.5	7.0 ± 0.5	461	42
Intergenic (*M.m.*) ncr.	3.6 ± 0.4	2.2 ± 0.5	5.8 ± 0.5	441	42
S_{lead} (*H.s.* introns)	7.5 ± 0.3	6.8 ± 0.4	14.3 ± 0.4	358	40
S_{lag} (*H.s.* introns)	−1.9 ± 1.0	−0.3 ± 1.4	−2.2 ± 1.3	49	44

Table 1: Strand asymmetries associated with human replication origins[38]. The skews were calculated in the regions flanking the six human replication origins (Fig. 9(a)) and in the corresponding homologous regions of the mouse genome. Intergenic sequences were always considered in the direction of replication fork progression (leading strand); they were considered in totality (all) or after elimination of conserved regions (ncr.) between human (Homo sapiens, H.s.) and mouse (Mus musculus, M.m.) (see Section 2.1). To calculate the mean skew in introns, the sequences were considered on the nontranscribed strand. For S_{lead}, the orientation of transcription was the same as the replication fork progression; for S_{lag}, the situation was the opposite. The mean values of the skews S_{TA}, S_{GC}, and S are given in percent (±SEM). l, total sequence length in kbp.

on the leading strand so that replication and transcription biases add to each other. To measure compositional asymmetries that would result from replication only, we have calculated the skews in intergenic regions on both sides of the origins[38]. 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 replication origin. This result strongly suggests the existence of mutational pressure associated with replication, leading to the mean compositional biases $S_{\rm TA} = 4.0 \pm 0.4\%$ and $S_{\rm GC} = 3.0 \pm 0.5\%$ (Table 1). Let us note that the value of the skew could vary from one origin to another, possibly reflecting different initiation efficiencies. From the calculation of the intron skew values on the leading and lagging strands reported in Table 1, one can estimate the mean skew associated with transcription by subtracting intergenic skews from $S_{\rm lead}$ values giving $S_{\rm TA} = 3.6 \pm 0.7\%$ and $S_{\rm GC} = 3.8 \pm 0.9\%$. These estimations are remarkably consistent with those obtained with our large set of human introns in Section 4, 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 $S = S_{\rm TA} + S_{\rm GC} \sim 7\%$ (Table 1).

In that context, one can wonder to which extent the biases observed in intergenic regions may result from the possible presence of still undetected genes. Two pieces of evidence argue against this eventuality. First, we have been careful enough to retain as transcribed regions one of the largest sets of transcripts available, resulting in a stringent definition of intergenic regions. Second, several studies have demonstrated the existence of hitherto unknown transcripts in regions where no protein coding genes have been previously identified [133–136]. Taking advantage of the set of non-protein-coding RNAs identified in the "H-Inv" database[137], we have checked that none of them are present in the intergenic regions studied here. Finally we have eliminated the possibility that intergenic skews are due to conserved sequences by checking that the removal of homologous segments found in the mouse genome ($\sim$ 5.3% of all intergenic sequences) does not change significantly the skews in intergenic regions[38].

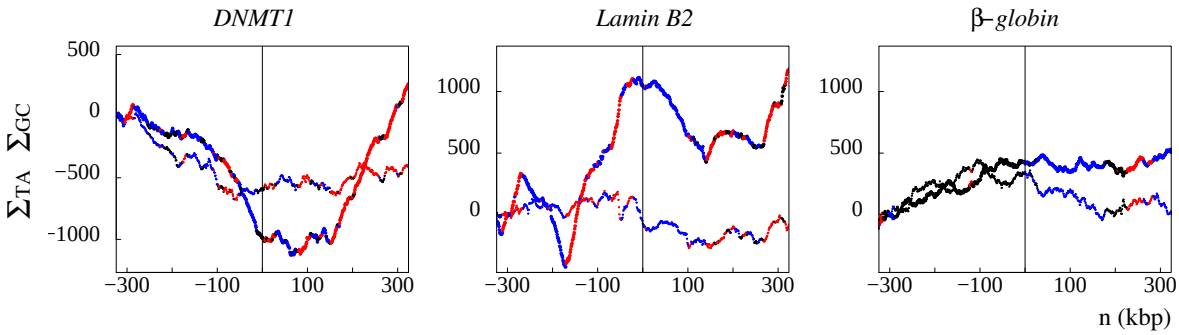


Figure 10: Cumulated skew profiles calculated around the origin of replication DNMT1, Lamin B2, and β-globin in the human genome: Σ_{TA} (thick line) and Σ_{GC} (thin line). The colors have the same meaning as in Figure 9.

5.3 Conservation of replication-associated strand asymmetries in mammalian genomes

As a next step of our study, we have analyzed[38] the $S_{\rm TA}$ an $S_{\rm GC}$ skew profiles in DNA regions of mammalian genomes homologous to the six human origin investigated in Figure 9(a). As shown in Figures 9(b) and 9(c), the human, mouse and dog cumulated skew profiles look strikingly similar to each other, suggesting that in mouse and dog, these regions also correspond to replication initiation zones (note that they are very similar in primate genomes). For each replication origin, one robustly observes a $\vee$-shape characteristic of a sharp upward jump from negative to positive skew values. A detailed examination of the mouse intergenic regions suggests the existence of a compositional bias associated with replication $S = S_{\rm TA} + S_{\rm GC} \sim 5.8\pm0.5\%$ (Table 1). Let us point out that, at these homologous loci, human and mouse intergenic sequences present almost no ($\sim 5.3\%$) conserved elements. Hence, the presence of strand asymmetry in regions that have strongly diverged during evolution further supports the existence of compositional bias associated with replication in both organisms. In the absence of such a process, intergenic sequences would have lost a significant fraction of their strand asymmetry.

Altogether, these results establish the existence of strand asymmetries associated with replication in mammalian germ-line cells[38]. They show that most replication origins experimentally detected in somatic cells coincide with sharp upward transitions of the skew profile. They also imply that for the majority of experimentally determined origins, the position of initiation zones are conserved in mammalian genomes as recently confirmed by the identification of a replication origin in the mouse MYC locus[138]. Let us emphasize that among nine human origins known experimentally, three do not present typical $\vee$-shape cumulated profiles as reported in Figure 10. For DNMT1 (left panel in Fig. 10), the sharp central part of the $\vee$ profile is replaced by a large horizontal plateau (few tens of kbp), possibly reflecting the presence of several origins dispersed over the whole plateau. Note that dispersed origins have been observed, e.g. in the hamster DHFR initiation zone[139]. By contrast, the cumulated skew profiles of the Lamin B2 (central panel of Fig. 10) and β-globin (right panel of Fig. 10) origins present no $\vee$ profile suggesting that they might be inactive in germ-line cells or less active than neighboring origins.

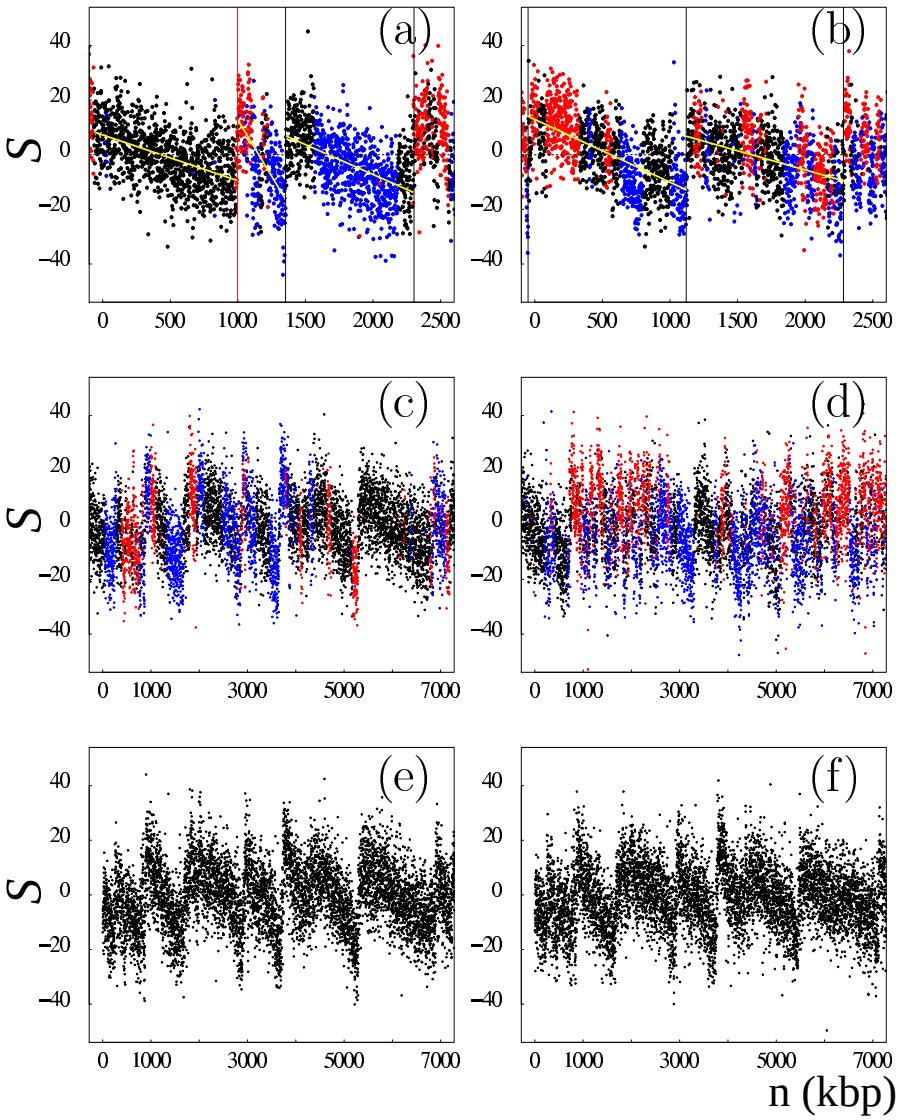


Figure 11: S profiles along mammalian genome fragments[38]. (a) Fragment of chromosome 20 including the TOP1 origin (red vertical line). (b and c) Chromosome 4 and chromosome 9 fragments, respectively, with low GC content (36%). (d) Chromosome 22 fragment with larger GC content (48%). In (a) and (b), vertical lines correspond to selected putative origins (see Section 6.1); yellow lines are linear fits of the S values between successive putative origins. Black, intergenic regions; red, sense genes; blue, antisense genes. Note the fully intergenic regions upstream of TOP1 in (a) and from positions 5290-6850 kbp in (c). (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.

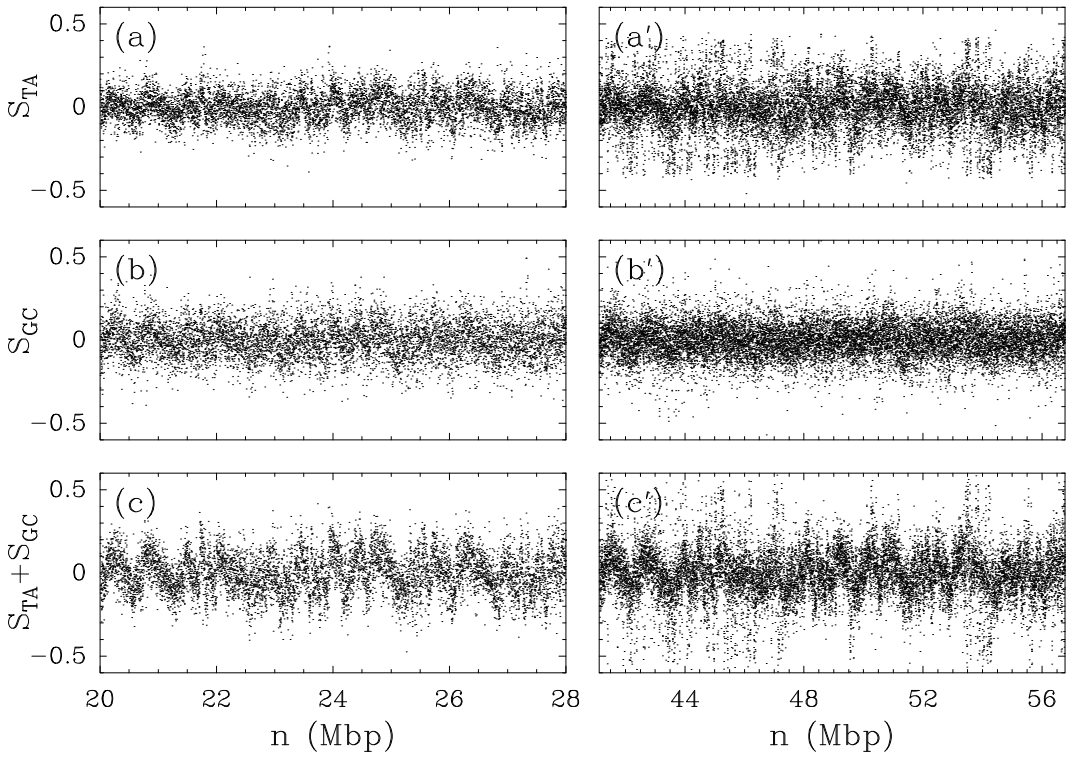


Figure 12: Skew profiles along a large fragment of the human chromosome 12. Repeat-masked sequence (8 Mb): S_{TA} (a), S_{GC} (b), and $S = S_{TA} + S_{GC}$ (c). Native sequence (15.7 Mb): S_{TA} (a'), S_{GC} (b'), and S (c').

5.4 Factory-roof skew profiles in the human genome

As illustrated in Figure 11(a), for TOP1 replication origin, when examining the behavior of the skews at larger distances from the origin, one does not observe a step-like pattern with upward and downward jumps at the origin and termination positions respectively as expected for the bacterial replicon model (Fig. 8(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 11(b–d), sharp upward jumps of amplitude $\Delta S \gtrsim 15\%$, similar to the ones observed for the known replication origins (Fig. 9), 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" [38,39]. For comparison, we show in Figure 12, the $S_{\rm TA}$, $S_{\rm GC}$ and S profiles obtained for a large fragment of the human chromosome 12 after (Figs. 12(a–c)) and before (Figs. 12(a'–c')) removing the repeated sequences (Section 2.1). There is no doubt that repeated sequences increase the level of noise in the skew profiles. Indeed factory roofs are more easily seen on the masked sequences and specially on the total skews $S = S_{\rm TA} + S_{\rm GC}$. As reported in Figure 13, the pdfs of $S_{\rm TA}$, $S_{\rm GC}$ and S are nearly Gaussian for the masked sequences; some large tails are present but for skew amplitudes

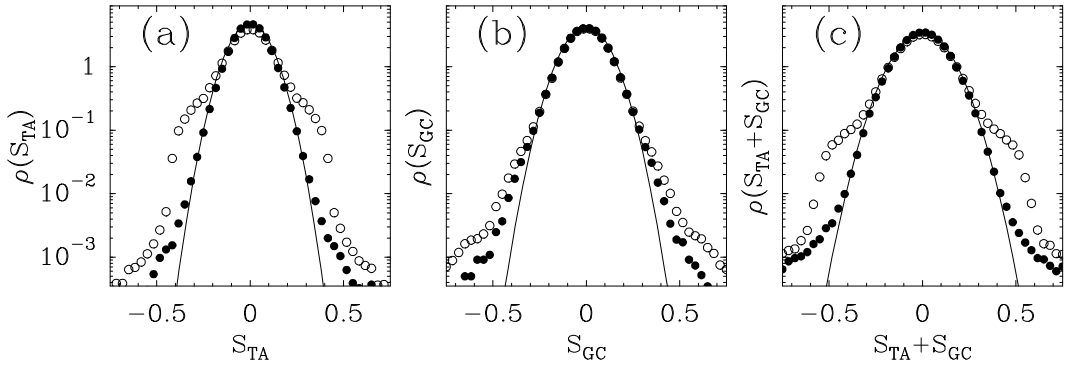


Figure 13: Probability density functions of the skews S_{TA} (a), S_{GC} (b), and $S = S_{TA} + S_{GC}$ (c) values computed in nonoverlapping 1 kbp windows from the DNA sequences of the 22 human autosomal chromosomes. Symbols have the following meaning: (○) native sequences and (●) repeat-masqued sequences.

larger than 40%. The fact that the skew pdfs of the native sequences, and more particularly the $S_{\rm TA}$ pdf, significantly depart from Gaussian distributions justifies, *a posteriori*, the need of removing repeated sequences prior to our statistical analysis. Most of these sequences have been inserted recently in the human genome and do not reflect long-term evolutionary skew patterns.

The jagged S profiles shown in Figures 11(a–d) and 12(a–c) 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 2-3 Mbp ($\sim$ 4-5 Mbp for the native sequences) in agreement with recent experimental studies[83] that have shown that mammalian replicons are heterogeneous in size with an average size ~ 500 kbp, the largest ones being as large as a few Mbp. But what is important to notice is that some of these segments between two successive upward jumps of the skew are entirely intergenic (Figs. 11(a) and 11(c)), clearly illustrating the particular profile of a strand bias resulting solely from replication[38,39]. 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. 11(c)). Importantly, as illustrated in Figures 11(e) and 11(f), the factory-roof pattern is not specific to human sequences but is also observed in numerous regions of the mouse and dog genomes[38].

6 From the detection of putative replication origins to the modeling of replication in the human genome

6.1 A wavelet-based method to detect putative replication origins

We have shown in Section 5 that experimentally determined human replication origins 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 (Fig. 9). To predict replication origins, one thus needs a methodology to detect discontinuities in noisy signals. As introduced in Section

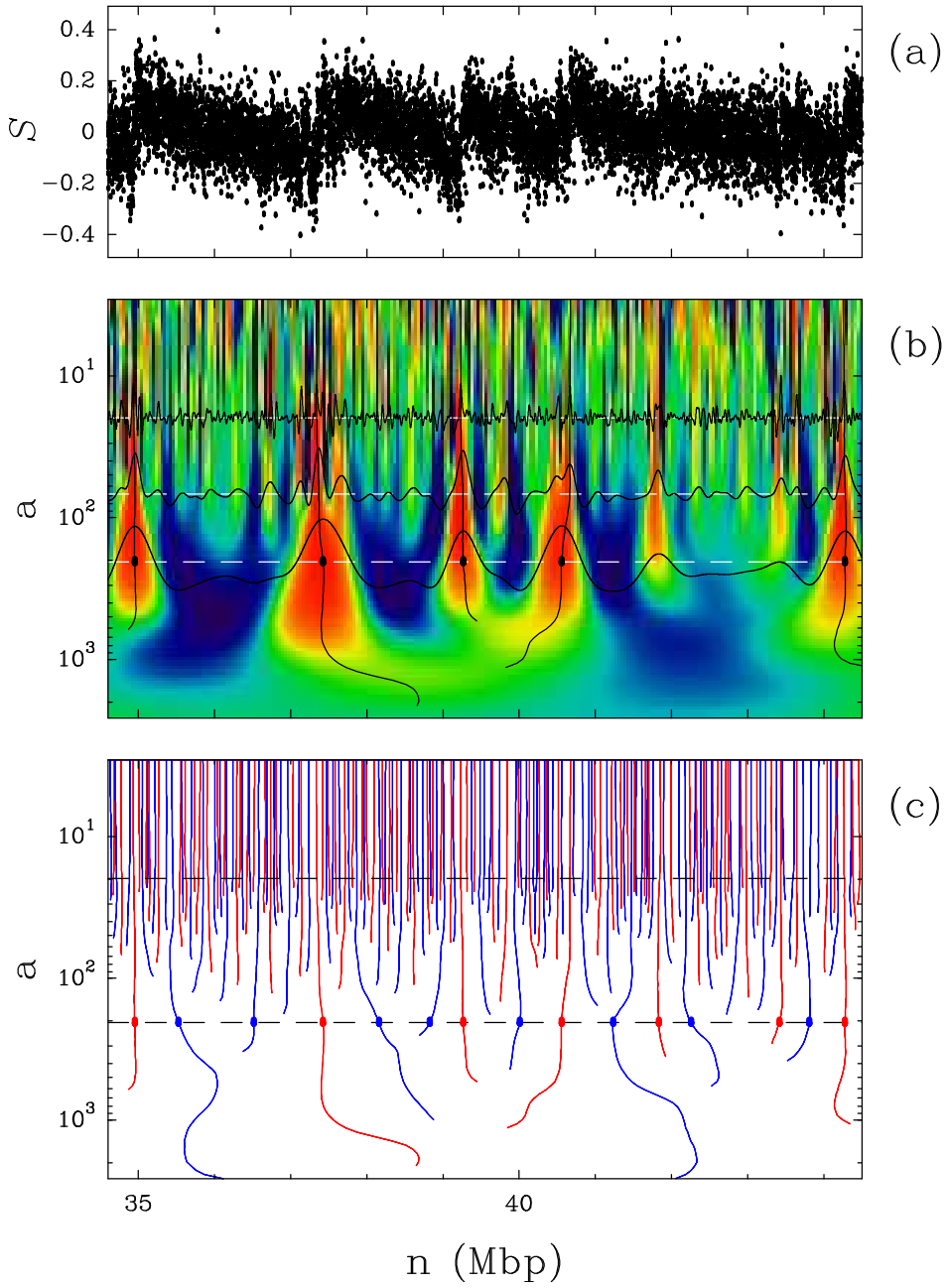


Figure 14: (a) Skew profiles of a fragment of Human chromosome 12. (b) WT of S using $g^{(1)}$; $W_{g^{(1)}}[S](n,a)$ is coded from black (min) to red (max); three cuts of the WT at constant scale $a = a^* = 200$ kbp, 70 kbp and 20 kbp are superimposed together with five maxima lines identified as pointing 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 replication origins; it is clear that the associated maxima lines point to the 5 major upward jumps in the skew profile in the limit $a \to 0^+$.

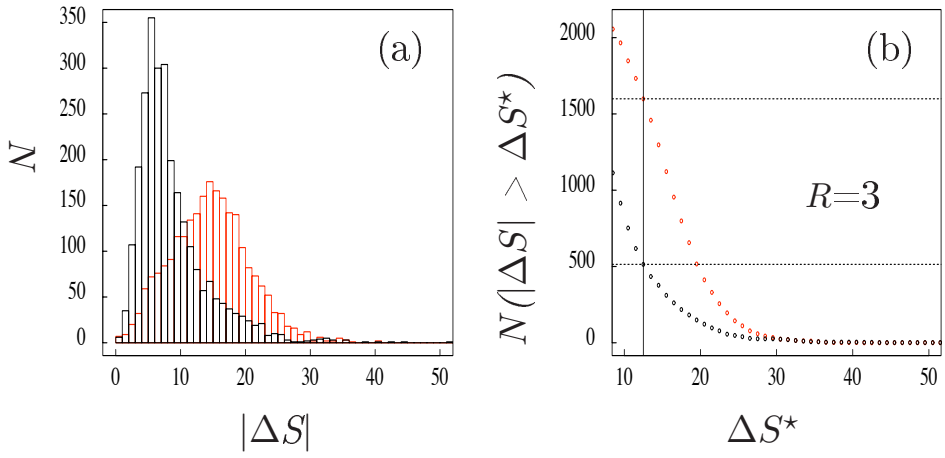


Figure 15: Statistical analysis of the sharp jumps detected in the S profiles of the 22 human autosomal chromosomes by the WT microscope at scale $a^* = 200$ kbp for repeat-masked sequences[38,39]. $|\Delta S| = |\bar{S}(3') - \bar{S}(5')|$, where the averages were computed over the two adjacent 20 kbp windows, respectively, in the 3' and 5' direction from 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 $|\Delta S| \geq 12.5\%$ (see text).

2.3, the continuous wavelet transform is a mathematical microscope that is well adapted for singularity tracking[13,26–29]. The basic principle of the detection of jumps in the skew profiles with the WT is illustrated in Figure 14. From Eq. (7), when using the first-derivative of the Gaussian function 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 14(b). The main issue when dealing with noisy signals like the skew profile in Figure 14(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 14(c), our methodology consists in computing the WT skeleton[13,27,38] 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 only (Fig. 7), to the benefit of maintaining a good sensitivity to replication induced jumps. The maxima lines that exist at that scale a^* are likely to point to jump positions at small scale (Fig. 14(c)). The detected jump locations are estimated as the positions at scale 20 kbp of the so-selected maxima lines. According to Eq. (7), upward (resp. downward) jumps are indentified by the maxima lines corresponding to positive (resp. negative) values of the WT as illustrated in Figure 14(c) by the blue (resp. red) maxima lines. When applying this methodology to the total skew

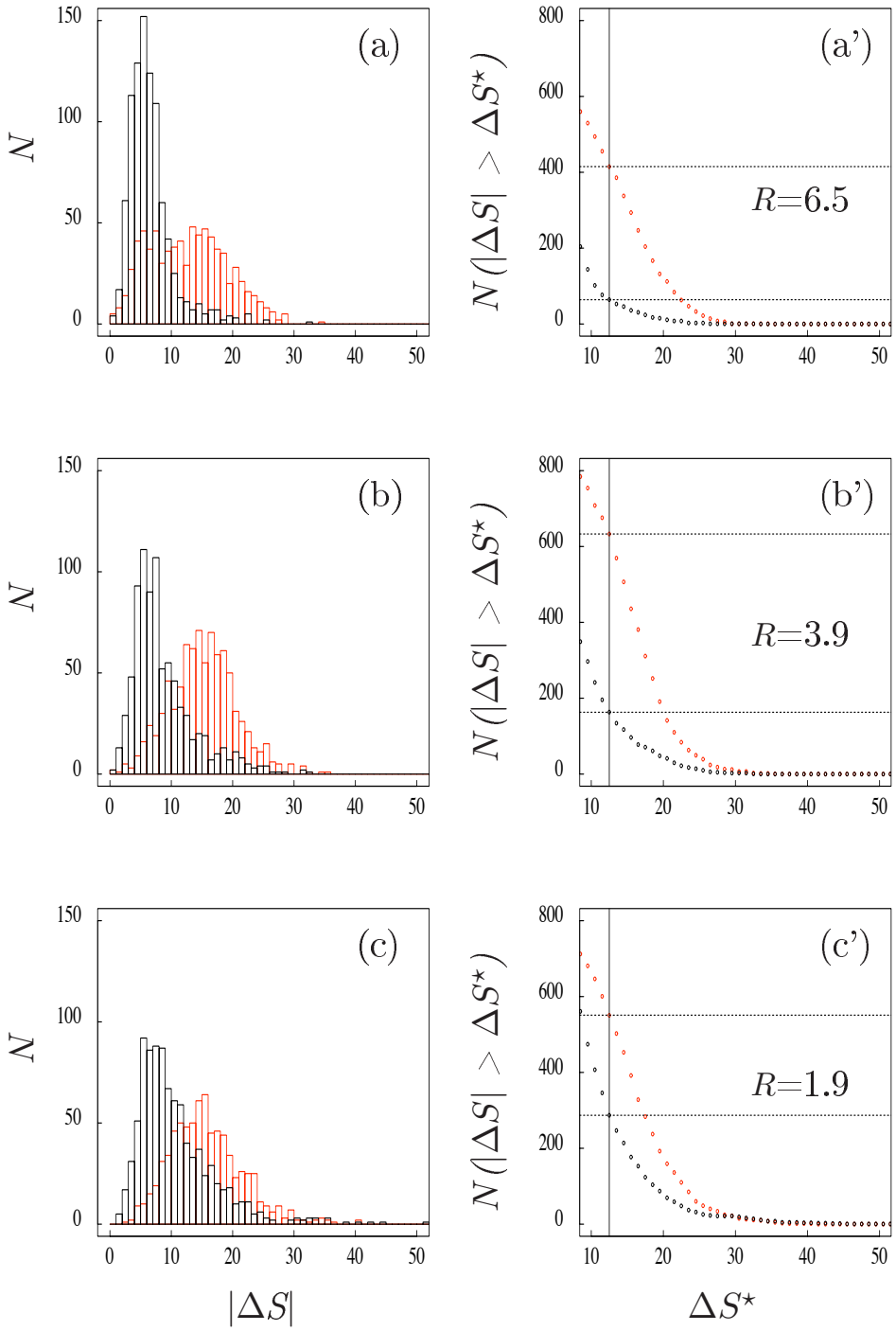


Figure 16: Statistical analysis of the sharp jumps detected in the S profiles of the 22 human autosomal chromosomes by the WT microscope at scale $a^* = 200$ kbp for repeat-masked sequences[109]. The detected jumps have been classified into three categories according to the GC content found in a 100 kbp window centered at the position of the jump. Same as in Figure 15: (a,a') G+C< 37%; (b,b') 37% ≤G+C≤ 42%; (c,c') 42% <G+C.

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 15(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 15(b), one can see that the number of large amplitude upward jumps overexceeds the number of large amplitude downward jumps. These results[38,39] confirm that most of the sharp upward transitions in the S profiles in Figures 11 and 14(a) have no sharp downward transition counterpart. This excess likely results from the fact that, contrasting with the prokaryote replicon model (Fig. 8) where downward jumps result from precisely positioned replication terminations, in mammals termination appears not to occur at specific positions but to be randomly distributed[38,39] (this point will be detailed in Section 6.3). Accordingly the small number of downward jumps with large $|\Delta S|$ is likely to result from transcription (Fig. 7) and not from replication. These jumps are probably due to highly biased genes that also generate a small number of large-amplitude upward jumps, giving rise to false-positive candidate replication origins. In that respect, the number of large downward jumps can be taken as an estimation of the number of false positives. In a first step, we have retained as acceptable a proportion of 33% of false positives. As shown in Figure 15(b), this value results from the selection of upward and downward jumps of amplitude $|\Delta S| \geq 12.5\%$, corresponding to a ratio of upward over downward jumps $R = 3$. Let us notice that the value of this ratio is highly variable along the chromosome (Fig. 16). In G+C poor regions, namely G+C<37%, we observe in Figures 16(a,a') the largest R value, namely $R = 6.5$. In regions with 37%≤G+C≤42%, we obtain $R = 3.9$ (Fig. 16(b,b')) that contrasts with small R values, $R = 1.9$ (Fig. 16(c,c')) found in regions with G+C>42%. In these latter regions (accounting for ∼40% of the genome) with high gene density and small gene length[51], the skew profiles oscillate rapidly with large upward and downward amplitudes (Fig. 11(d)), resulting in a too large estimate of the number of false positives (∼53%).

In a final step, we have decided[38] to retain as putative replication origins upward jumps with $|\Delta S| \geq 12.5\%$ detected in regions with G+C ≤ 42%. This selection leads to a set of 1012 candidates among which our estimate of the proportion of true replication origins is 79% ($R = 4.76$). Some of these putative replication origins are illustrated in Figure 11.

6.2 Gene organization around the 1012 putative replication origins in the human genome

The mean amplitude of the upward jumps associated with the 1012 putative origins is 18%, consistent with the range of values observed for the six experimentally known origins in Figure 9. Let us remark that all six origins have been identified by our detection methodology. When investigating the gene content around these putative origins[109], 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 origin 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 origins mostly results from superimposition of biases (i) associated with replication and (ii) with transcription of the genes proximal to the origin. Determining whether transcription is co-oriented with replication at larger distances is the subject of current study.

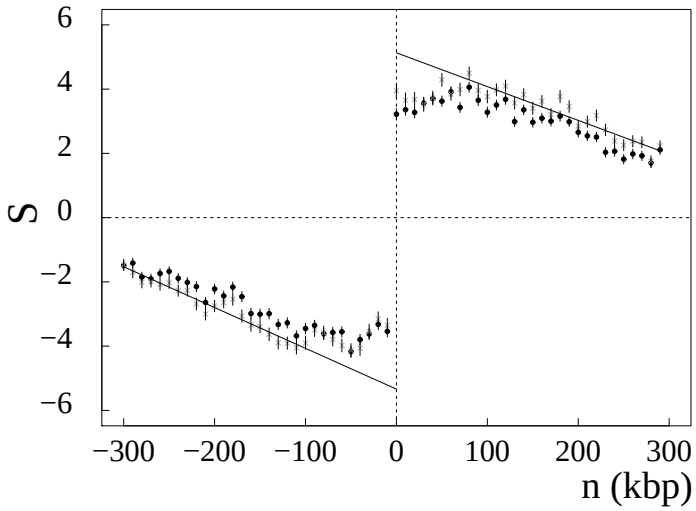


Figure 17: Mean skew profile of intergenic regions around putative replication origins[38]. The skew S was calculated in 1 kbp windows (Watson strand) around the position ($\pm$300 kbp without repeats) of the 1012 detected upward jumps; 5' and 3' transcript extremities were extended by 0.5 and 2 kbp, respectively ($\bullet$), or by 10 kbp at both ends ($*$). The abscissa represents the distance (in kbp) to the corresponding origin; the ordinate represents the skews calculated for the windows situated in intergenic regions (mean values for all discontinuities and for 10 consecutive 1 kbp window positions). The skews are given in percent (vertical bars, SEM). The lines correspond to linear fits of the values of the skew ($*$) for $n < -100$ kbp and $n > 100$ kbp.

In Figure 17 is shown the mean skew profile calculated in intergenic windows on both sides of the 1012 putative replication origins[38]. This mean skew profile presents a rather sharp transition from negative to positive values when crossing the origin position. To avoid any bias in the skew values that could result from incompletely annotated gene extremities (e.g. 5' and 3' UTRs), we have removed 10-kbp sequences at both ends of all annotated transcripts. As shown in Figure 17, the removal of these intergenic sequences does not significantly modifies the mean skew profile, indicating that the observed values do not result from transcription. 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. Note that, due to (i) the potential presence of signals implicated in replication initiation and (ii) the possible existence of dispersed origins[139], one might question the meaningfulness of this flattening that leads to a significant underestimate of the jump amplitude. Furthermore, according to our detection methodology, the numerical uncertainty on the putative origin position estimate may also contribute to this flattening. As illustrated in Figure 17, 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 six experimentally known replication origins namely $7.0 \pm 0.5\%$ (Table 1). Overall, the detection of sharp upward jumps in the skew profiles with characteristics similar to those of experimentally determined replication origins and with no downward counterpart further supports the existence, in human chromosomes, of replication-associated strand asymmetries, leading to the identification of numerous putative replication

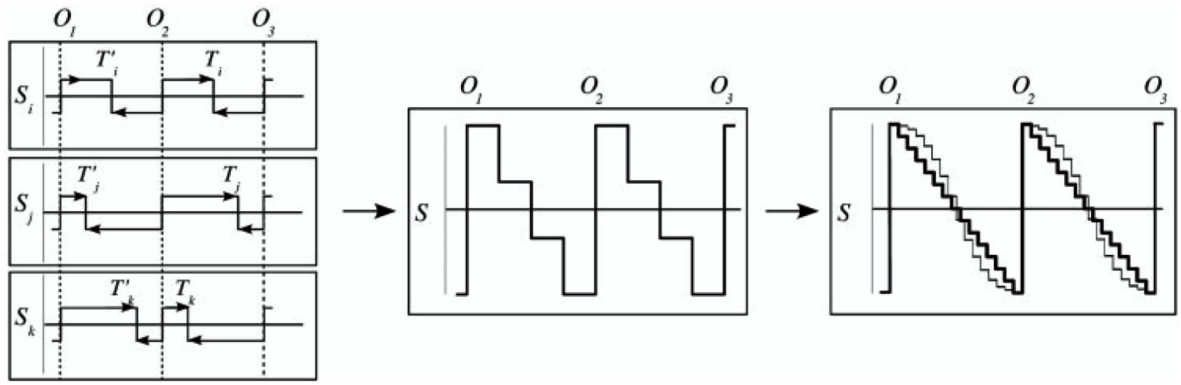


Figure 18: Model of replication termination[38,39]. Schematic representation of the skew profiles associated with three replication origins O_1, O_2, and O_3; we suppose that these replication origins are adjacent, bidirectional origins with similar replication efficiency. The abscissa represents the sequence position; the ordinate represents the S value (arbitrary units). Upward (or downward) steps correspond to origin (or termination) positions. For convenience, the termination sites are symmetric relative to O_2. (Left) Three different termination 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. In the simple model, termination occurs with equal probability on both sides of the origins, leading to the linear profile (thick line). In the alternative model, replication termination is more likely to occur at lower rates close to the origins, leading to a flattening of the profile (gray line).

origins active in germ-line cells.

6.3 A model of replication in mammalian genomes

Following the observation of jagged skew profiles similar to factory roofs in Section 5.4, and the quantitative confirmation of the existence of such (piecewise linear) profiles in the neighborhood of 1012 putative origins in Figure 17, we have proposed, in Touchon *et al.*[38] and Brodie of Brodie *et al.*[39], a rather crude model for replication in the human genome that relies on the hypothesis that the replication origins are quite well positioned while the terminations are randomly distributed. Although some replication origins have been found at specific sites in *S. cerevisiae* and to some extent in *Schizosaccharomyces pombe*[140], they occur randomly between active origins in *Xenopus* egg extracts[141,142]. Our results indicate that this property can be extended to replication in human germ-line cells. As illustrated in Figure 18, replication termination is likely to rely on the existence of numerous termination sites distributed along the sequence. For each termination 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 origins (as in bacteria, Fig. 8(b)). Various termination positions will thus correspond to classical replicon-like skew profiles (Fig. 18, left panel). Addition of those profiles will generate the intermediate profile (Fig. 18, central panel). In a simple picture, we can reasonably suppose that termination 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 origins, leading to a homogeneous distribution of termination sites during successive cell cycles. The final skew profile is then a linear segment decreasing between successive origins (Fig. 18, right panel). Let us point out that firing of replication origins during time interval of the S phase[143] might result in some flattening of the skew profile at the origins as sketched in Figure 18 (right panel, gray curve). In the present state, our results[38,39] support the hypothesis of random replication termination in human, and more generally in mammalian cells (Figs. 9 and 11), but further analyses will be necessary to determine what scenario is precisely at work.

In conclusion, we have revealed a factory roof skew profile as an alternative in mammalian genomes to the replicon step-like profile observed in bacteria (Fig. 8). 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 origins 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[83–85] ($\sim$ 100 kbp to 500 kbp on the native sequences), they are not incompatible with estimations showing that replicon size can reach up to 1 Mbp[83,86], and that replicating units in meiotic chromosomes are much longer than those engaged in sommatic cells[144]. Finally, it is not unlikely that in G+C-rich (gene-rich) regions (Fig. 11(d)) replication origins would be closer to each other than in other regions, further explaining the greater difficulty in detecting origins in these regions. Indeed, the wavelet-based methodology described in Section 6.1 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.

For more details on the existence and modeling of replication associated strand asymmetries in mammalian genomes, we refer the reader to the PhD thesis manuscripts of E.B. Brodie of Brodie[96], S. Nicolay[97] and M. Touchon[109].

7 From sequence analysis to the modeling of the chromatin tertiary structure

Some fifty years ago, Asakura and Oosawa[145] pointed out that two large rigid spheres immersed in a solution of smaller spheres are subject to an attractive force due to the depletion induced by increasing the space available to small spheres as the large ones come close to one another. Snir and Kamien[146] have shown recently that short molecular chains, modelled as stiff (but not rigid) impenetrable tubes, are driven to a helix configuration using the same depletion argument. However this holds only for uniform and relatively short tubes of length of the order of a few persistence lengths l_p of the rod. On longer chains, the picture rapidly grows in complexity with a plethora of optimal configurations (e.g. hairpin, beta-sheet, superhelix, torus) leading to an overwhelmingly rich phase diagram. In this section, our goal is to show that the presence of chromatin fiber rosettes can be explained using a depletion argument for long tubes with "frozen", heterogeneously distributed elastic and/or geometric properties. By frozen we mean that these fluctuations are imprinted on the 30 nm chromatin fiber by the sequence itself. Indeed, the fiber is known to be dependent upon the proper-

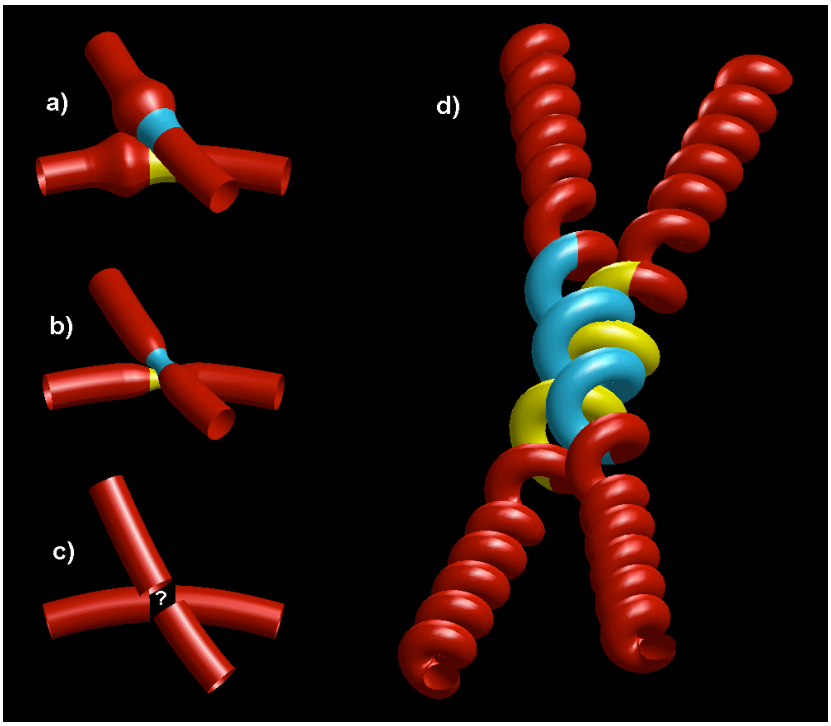


Figure 19: Examples of possible local defects along the fiber. (a) Local swelling or attachment of an external agent (e.g. RNA polymerases in the model of Cook[22,23]); (b) local shrinking; (c) any form of fiber denaturation inducing a depletive potential well, according to the position along the fiber and the entry-exit angle; (d) as an example of (c), the fiber seen as a compact helix (condensed nucleosomal array) with local partial decondensation illustrating a situation where the excluded volume gain is quite important and the entry-exit angle is fixed.

ties of the nucleosomal string-of-beads[147–150], which in turn is influenced by the double helix intrinsic structural disorder induced by the sequence. In essence, it comes down to ask the following question: is there a topological configuration in which the fiber is most likely to self-organize reproducibly?

For a semi-flexible tube in a dilute environment, local repulsive potentials among parts of the fiber induce a self-avoiding random walk configuration (swollen coil[151]). In a crowded environment, the depletive action may dominate and the fiber will tend to collapse on itself, forming a globular phase. We know from standard statistical physics of polymers that this latter phase does not admit a universal description in terms of macroscopic parameters (such as total length, Kuhn length and virial coefficients) but rather depends on a detailed understanding of the interaction potential. However, an important feature of the depletive potential lies in its simplistic geometrical nature. We thus consider a system constituted of a dense fluid of hard spheres bathing a semi-flexible tube. The tube is assumed to be non-uniform, with localized geometrical defects (e.g. local thickening or thinning of the cross-section, see Figure 19). The elastic nature of the tube prevents the appearance of too high curvature 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 en-

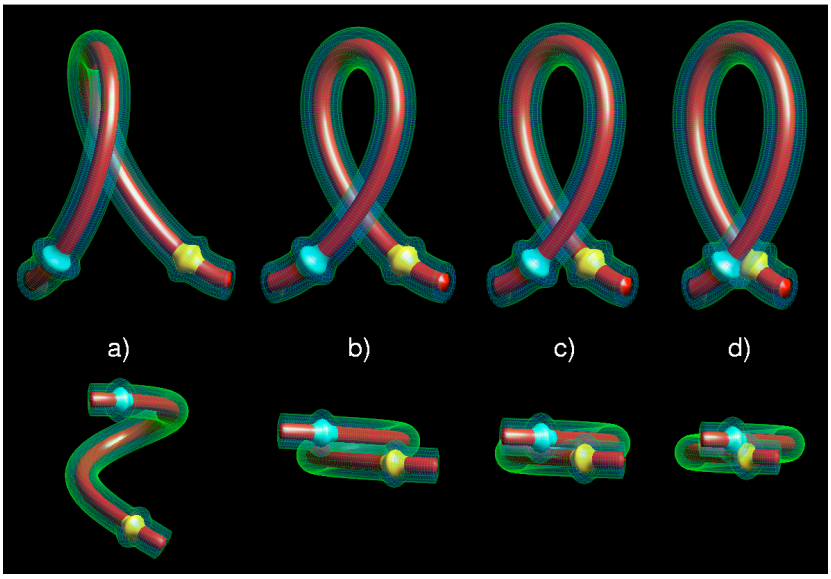


Figure 20: Steps involved in loop formation: (a) free evolution of the tube in depletive environment; (b) formation of an unstable loop at around 3.4 l_p; (c) gliding of the loop governed by the positions of the two contact points along the fiber and the entry-exit angle; (d) trapping of the loop by local defects. The translucent green surface represents the excluded volume for the fluid of hard spheres; in (b,c,d) one sees that some of the excluded volume is reduced from the overlap resulting from formation of the loop.

tropic 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 $3.4l_p$ in the worm-like-chain (WLC) model[152,153].

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 (Fig. 20). 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; 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.

In that context we propose[154] to characterize the distribution of the number of leaves per rosette from minimal parametrization of the system. We consider a dense fluid composed of a large number N_s of identical spheres, bathing a tube which, for simplicity, contains N equidistant defects, separated by a distance l along the tube. We assume that rosettes are formed while respecting sequential order of defects along the tube. Let n denote the number of rosettes along the tube; solitary defects are also considered as trivial "rosettes" with zero leaves. Obviously the case where $n = N$ represents the absence of clustering since all defects are then solitary. On the other hand the case where $n = 1$ corresponds to a single large

rosette assembling all defects.

We separate in a natural fashion the system in two parts, namely the hard sphere fluid and the tube itself. Let F, F_t and F_s denote the free energies of the system, the tube and the spheres respectively (all of which depend on n) such that we can write $F = F_t + F_s$. The most probable value for n is obtained by taking the derivative of F with respect to n and equating to zero. This derivative is simply the *chemical potential* μ_r of a rosette:

$$\mu_r \equiv \frac{\partial F}{\partial n} = \frac{\partial F_t}{\partial n} + \frac{\partial F_s}{\partial n} = 0. \tag{10}$$

The equation of state of a fluid of hard spheres has been extensively studied in the past[155]. We follow the method used by Dinsmore *et al.*[156] and make use of the Carnahan-Starling approximation[157]:

$$\frac{P_s(n)V_s(n)}{N_s k_B T} = \frac{1+\varphi+\varphi^2-\varphi^3}{(1-\varphi)^3}, \tag{11}$$

where $\varphi = N_s v_s / V_s(n)$ is the density of the spheres, $P_s(n)$ the fluid osmotic pressure, v_s the volume of each sphere and $V_s(n)$ is the volume available to the spheres. The F_t term can be expressed as

$$F_t = E_0 + (N-n)\cdot\Delta F_l - k_B T \cdot \log\binom{n}{N}, \tag{12}$$

where $\Delta F_l > 0$ is the free energy cost for the formation of a single loop, and E_0 is an energy term assumed to be independent of n. The last term on r.h.s. of Eq. (12) corresponds to the number of arrangements of n rosettes from N defects and contributes to the entropy of the tube. From Eqs. (11) and (12) and from the thermodynamical identity $\frac{\partial F_s}{\partial V_s} = -P_s$, we get:

$$\begin{aligned}\mu_r &= -k_B T \cdot \log\left(\frac{N-n}{n}\right) - \Delta F_l \\ &+ k_B T \cdot \left(\frac{v_{\text{ovl}}}{v_s}\right)\cdot\left(\frac{\varphi+\varphi^2+\varphi^3-\varphi^4}{(1-\varphi)^3}\right),\end{aligned} \tag{13}$$

where $v_{\text{ovl}} = -\frac{\partial V_s}{\partial n}$ represents the overlapping excluded volume of two interleaved defects, *i.e.* the volume gain for the spheres due to the interaction between two defects. From Eq. (13), we see that this simple model depends on three parameters: the free energy cost of a loop ΔF_l, the normalized overlap volume per loop v_{ovl}/v_s and the sphere density φ. The free energy cost of a loop can be approximated by

$$\Delta F_l = \frac{1}{2} k_b T \cdot l \cdot l_p \kappa^2 \lesssim 6 k_B T, \tag{14}$$

where $l \simeq 3.4 l_p$ is the length of a typical loop and $\kappa \simeq 2\pi/l$ its average curvature. Physiological values for the hard-sphere fluid density $\Phi = \varphi(n=N)$ vary between[158] $\simeq 0.2 \sim 0.3$. For a 30nm fiber we can expect $v_{\text{ovl}} \sim (10\text{nm})^3$; the typical size of proteins is $v_s \sim (5\text{nm})^3$, leading to values of v_{ovl}/v_s around 10.

As illustrated in Figure 21, increasing the sphere density Φ or the normalized overlap volume results in an increase of the average number of leaves per rosette ($N/n - 1$), thus in a more compact structure. On the other hand, increasing the free energy cost of loop

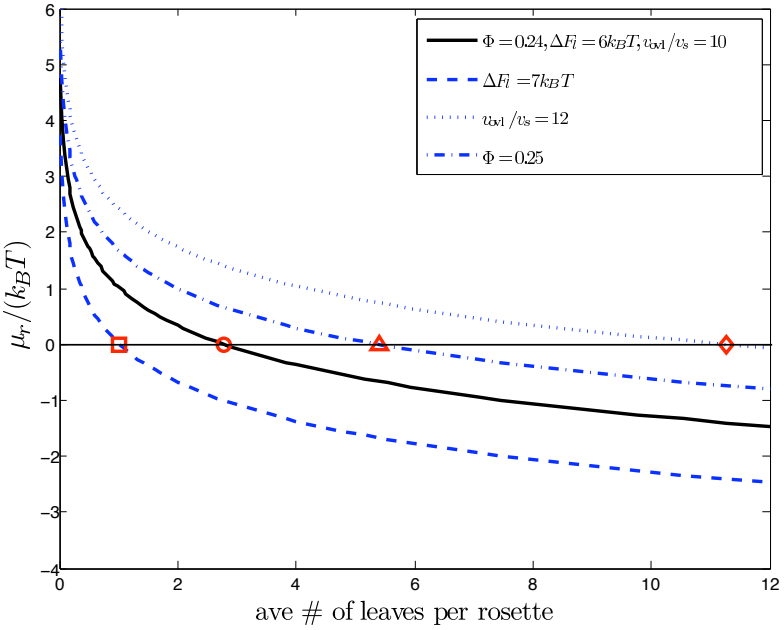


Figure 21: The chemical potential is shown as a function of the number of leaves per rosette $(N/n - 1)$. The solid curve serves as a reference and corresponds to the following parameter values: $\Phi = 0.24$, $\Delta F_l = 6k_BT$ and $v_{\rm ovl}/v_s = 10$. For these values, an average value of 2.77 leaves ($\circ$) per rosette is expected. Increasing Φ to 0.25 increases the average to 5.39 ($\triangle$); similarly, increasing $v_{\rm ovl}/v_s$ to 12 results in an average value of 11.27 ($\Diamond$). Increasing ΔF_l to $7k_BT$ reduces the average to 1.02 ($\square$).

formation (stiffening of the fiber) decreases the average number of leaves. Interestingly, we find that the average number of leaves per rosette can be regulated by fine tuning the values of these parameters within physiological range (Fig. 21). For instance, for $v_{\rm ovl}/v_s = 10$ and $\Delta F_l = 6k_BT$ and varying Φ between 0.24 and 0.25 results in the average number of leaves per rosette running from 2.77, *i.e.* low clustering, to 6.39. This provides attractive scenarios for the spontaneous emergence of chromatin rosettes in the nucleus milieu prior to their possible further stabilization by external factors (*e.g.* specific DNA binding proteins)[23,70,83,123].

Various models of interphase chromatin based on a multi-looped structure of the 30nm fiber have been proposed in the literature[19–21], but they all involve interaction with some nucleoproteic complexes to organize the structure, e.g. the scaffolding proteins that interact specifically at certain DNA regions (Scaffold Associated Regions) to fold the fiber[14,15] or the transcription complexes strung along the genome that clusterize and consequently fold the chromatin fiber[22,23]. The main message of the present work is the possibility that the chromatin fiber self-organizes into rosette-like patterns in the crowded environment of the nucleus thanks to its heterogeneous structure. Recent modeling[148,149] has revealed an extreme sensitivity of the internal fiber conformation to the local structural and mechanical properties of the nucleosomal string, e.g. the linker length, the entry-exit angle between the linkers or the twist angle along a linker. The fiber local structure is known to be controlled by epigenetic modifications of these architectural nucleosomal parameters[147,149] (DNA methylation, his-

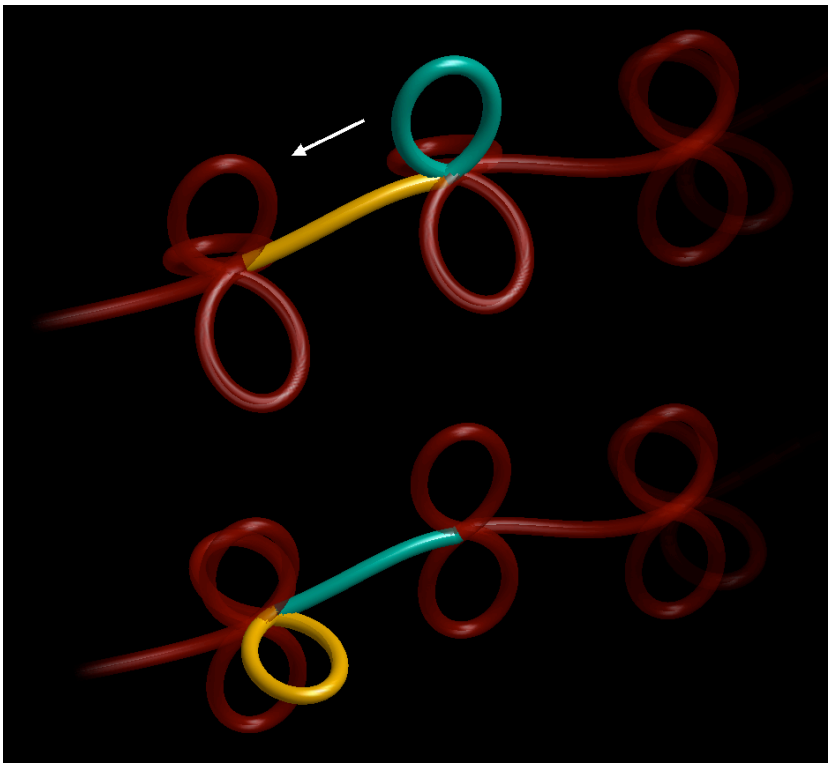


Figure 22: Illustration of the fiber defects clustering dynamics. The number of leaves per rosette fluctuates from one rosette to the next; this is due both to statistical fluctuations and variations in the local environment. From one cell cycle (up) to the next (down), leaves can be exchanged between neighboring rosettes.

tone modifications,...). Yet as suggested by recent modeling of the thermodynamics of DNA loops[30,31], the local properties of the nucleosomal string are also conditioned by the primary DNA sequence which codes for the structural disorder intrinsic to the DNA double helix. Therefore the structural defects of the fiber can be encoded in the sequence. The entropy-driven fiber folding mechanism described above[154] leads to the aggregation of neighboring defects into clusters that ensures high local concentration of distant DNA target sites. This clustering is likely to favor the recruiting of protein complexes involved in the activation of replication and transcription. In this context, the set of 1012 putative replication initiation zones identified in the human genome[38,39] (Section 6) provides priviledged locations for some intrinsic decondensated fiber defects. The spontaneous emergence of rosette patterns (likely stabilized by the Origin Replication Complexes) provides a very attractive description of the so-called replication foci[81,83,84,123] 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 replication origins[38,39], 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 replication origins, the DNA sequence might therefore code, to some extent, for the tertiary chromatin structure. Even though one expects to observe, from one cell cycle to the

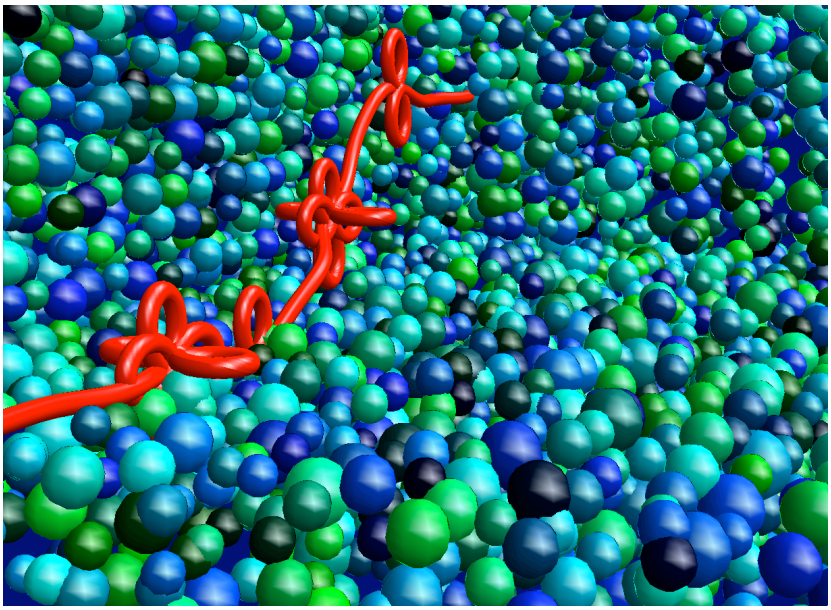


Figure 23: Illustration of the spontaneous emergence of rosette-like folding of the chromatin fiber in the crowded environment of the cell nucleus.

next, fluctuations in the number of loops contained in each rosette as illustrated in Figure 22, the perennity of defects is likely to ensure the inheritance of the interphase chromatin rosette organization. As an illustration, we present in Figure 23 the picture of a rosette-like pattern of the chromatin fiber in a crowded, heterogenous environment mimicking the cell nucleus.

8 Perspectives

In a recent past, the DNA double helix was simply considered as a biological macromolecule (a polymer) that contains our genic heritage (genotype). The regulation and control of DNA replication and expression was supposed to be fully delegated to proteins. Nowadays, DNA is more and more recognized as a complex heteropolymer whose structural and mechanical properties play a relevant part in the management of the gene information it carries. The results reported in this Chapter concerning the analysis at the genome scale of mammalian DNA sequences[36–39], together with the results obtained in a previous study[11–13,30–32] of the long-range correlations exhibited by eukaryotic DNA sequences up to distances of a few tens of kbp, demonstrate that there is a lot of information encoded in the DNA sequences concerning the different stages of compaction of DNA inside the nucleus of mammalian cells (Fig. 1). Surprisingly, if the sequence codes for the local structural and mechanical properties of the double helix and in turn influences the formation and dynamics of the nucleosomal string, our results show that it also conditions to some extent the next levels of compaction, via the self-organized formation of chromatin fiber rosette patterns that are likely to define structural domains of autonomous DNA replication and gene expression. Since introns and intergenic regions constitute more than 95% of the human genome, our study therefore contribute to give a role to the noncoding regions in eukaryotic genomes. These regions actually play a driving role in the condensation and decondensation processes of the chromatin architecture

as well as in many related regulative functions.

The results reported in this Chapter open new persepectives in DNA sequence analysis, modeling as well as in experiment. For a methodological point of view, there is some hope to improve the efficiency of our wavelet-based method to detect the replication origins (Section 6.1). So far our strategy was based on the use of the first-derivative of the Gaussian function as analyzing wavelet to detect sharp upward jumps in the noisy skew profiles. Along the line of the model of replication we propose in Section 6.3 to account for the observed factory roof skew profiles in mammalian genomes, we consider the use of an analyzing wavelet that has exactly the jagged shape predicted by this model as illustrated in Figure 18(c). By adapting the optics of our mathematical WT microscope, we should be in better position to face the observed variability in size of the replication domains. The implementation of a replication pattern matching algorithm in the space-scale representation provided by the WT is in current progress[97]. From a bioinformatics and modeling point of view, we plan to study the lexical and structural characteristics of our set of putative origins. In particular we will search for conserved sequence motifs in these replication initiation zones. Using a sequence-dependent model of DNA-histones interactions, we will develop physical study of nucleosome formation and diffusion along the DNA fiber around the putative replication origins. From an experimental point of view, our study raises new opportunities for future experiments. The first one concerns the experimental validation of the predicted replication origins (e.g. by molecular combing of DNA molecules[159]), which will allow us to determine precisely the existence of replication origins in given genome regions. Large scale study of all candidate origins is in current progress in the laboratory of O. Hyrien (ENS, Ulm). The second experimental project consists in using Atomic Force Microscopy (AFM)[160] and Surface Plasmon Resonance Microscopy (SPRM)[161] to visualize and study the structural and mechanical properties of the DNA double helix, the nucleosomal string and the 30nm chromatin fiber around the predicted replication origins. This work is in current progress in the experimental group of F. Argoul and C. Moskalenko at the Laboratoire Joliot-Curie (ENS, Lyon). Finally the third experimental perspective concerns *in situ* studies of replication origins. Using fluorescence techniques (FISH chromosome painting[80]), we plan to study the distributions and dynamics of origins in the cell nucleus, as well as chromosome domains potentially associated with territories and their possible relation to nuclear matrix attachment sites. This study is likely to provide evidence of chromatin rosette patterns as suggested in Section 7. This study is under progress in the molecular biology experimental group of F. Mongelard at the Laboratoire Joliot-Curie.

Acknowledgement

We thank F. Argoul, M. Castelnovo, P. Cook, O. Hyrien, F. Mongelard and C. Moskalenko for interesting discussions. 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. van Holde, K. E. (1988). *Chromatin*. Springer-Verlag, New York.

2. Wolffe, A. P. (1998). *Chromatin Structure and Function, 3rd ed.* Academic Press, London.

3. Calladine, C. R. & Drew, H. R. (1999). *Understanding DNA*. Academic Press, San Diego.

4. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1994). *Molecular Biology of the Cell, 3rd ed.* Garland Publishing, New-York.

5. Widom, J. (1998). Structure, dynamics and function of chromatin in vitro. *Annu. Rev. Biophys. Biomol. Struct.* 27, 285–327.

6. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature* 389, 251–260.

7. Chantalat, L., Nicholson, J. M., Lambert, S. J., Reid, A. J., Donovan, M. J., Reynolds, C. D., Wood, C. M. & Baldwin, J. P. (2003). Structure of the histone-core octamer in KCl/phosphate crystals at 2.15 A resolution. *Acta Crystallogr. D Biol. Crystallogr.* 59, 1395–1407.

8. Richmond, T. J. & Davey, C. A. (2003). The structure of DNA in the nucleosome core. *Nature* 423, 145–150.

9. Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.* 262, 129–139.

10. Herzel, H., Weiss, O. & Trifonov, E. N. (1999). 10-11bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics* 15, 187–193.

11. Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J.-F. & Arneodo, A. (2001). Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys. Rev. Lett.* 86, 2471–2474.

12. Audit, B., Vaillant, C., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. (2002). Long-range correlations between DNA bending sites: Relation to the structure and dynamics of nucleosomes. *J. Mol. Biol.* 316, 903–918.

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

14. Laemmli, U. K., Käs, E., Poljak, L. & Adachi, Y. (1992). Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. *Curr. Opin. Genet. Dev.* 2, 275–285.

15. Saitoh, Y. & Laemmli, U. K. (1993). From the chromosomal loops and the scaffold to the classic bands of metaphase chromosomes. *Cold Spring Harb. Symp. Quant. Biol.* 58, 755–765.

16. Belmont, A. S. & Bruce, K. (1994). Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure. *J. Mol. Biol.* 127, 287–302.

17. Belmont, A. S., Dietzel, S., Nye, A. C., Strukov, Y. G. & Tumbar, T. (1999). Large-scale chromatin structure and function. *Curr. Opin. Cell Biol.* 11, 307–311.

18. Horn, P. J. & Peterson, C. L. (2002). Molecular biology. Chromatin higher order folding–wrapping up transcription. *Science* 297, 1824–1827.

19. Sachs, R. K., van den Engh, G., Trask, B., Yokota, H. & Hearst, J. E. (1995). A random-walk/giant-loop model for interphase chromosomes. *Proc. Natl. Acad. Sci. USA* 92, 2710–2714.

20. Ostashevsky, J. (1998). A polymer model for the structural organization of chromatin loops and minibands in interphase chromosomes. *Mol. Biol. Cell* 9, 3031–3040.

21. Münkel, C., Eils, R., Dietzel, S., Zink, D., Mehring, C., Wedemann, G., Cremer, T. & Langowski, J. (1999). Compartmentalization of interphase chromosomes observed in simulation and experiment. *J. Mol. Biol.* 285, 1053–1065.

22. Cook, P. R. (1995). A chromomeric model for nuclear and chromosome structure. *J. Cell. Sci.* 108, 2927–2935.

23. Cook, P. R. (2002). Predicting three-dimensional genome structure from transcriptional activity. *Nat. Genet.* 32, 347–352.

24. Mahy, N. L., Perry, P. E. & Bickmore, W. A. (2002). Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *J. Cell Biol.* 159, 753–763.

25. Felsenfeld, G. & Groudine, M. (2003). Controlling the double helix. *Nature* 421, 448–453.

26. Arneodo, A., Bacry, E., Graves, P. V. & Muzy, J.-F. (1995). Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.* 74, 3293–3296.

27. Arneodo, A., d'Aubenton-Carafa, Y., Bacry, E., Graves, P. V., Muzy, J.-F. & Thermes, C. (1996). Wavelet based fractal analysis of DNA sequences. *Physica D* 96, 291–320.

28. Meyer, Y., ed. (1992). *Wavelets and their Applications.* Springer, Berlin.

29. Mallat, S. (1998). *A Wavelet Tour of Signal Processing.* Academic Press, New York.

30. Vaillant, C., Audit, B. & Arneodo, A. (2005). Thermodynamics of DNA loops with long-range correlated structural disorder. *Phys. Rev. Lett.* 95, 068101.

31. Vaillant, C., Audit, B., Thermes, C. & Arneodo, A. (2006). Formation and positioning of nucleosomes: effect of sequence dependent long-range correlated structural disorder. *Eur. Phys. J. E, in press* .

32. Vaillant, C. (2006). First experimental evidence of nucleosome positioning by genomic long-range correlations. Preprint.

33. Yuan, G.-C., Liu, Y.-J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J. & Rando, O. J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* 309, 626–630.

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

35. Nicolay, S., Brodie of Brodie, E.-B., Touchon, M., d'Aubenton-Carafa, Y., Thermes, C. & Arneodo, A. (2004). From scale invariance to deterministic chaos in DNA sequences: towards a deterministic description of gene organization in the human genome. *Physica A* 342, 270–280.

36. Touchon, M., Nicolay, S., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. (2003). Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Letters* 555, 579–582.

37. Touchon, M., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. (2004). Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucl. Acids Res.* 32, 4969–4978.

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

39. Brodie of Brodie, E.-B., Nicolay, S., Touchon, M., Audit, B., d'Aubenton-Carafa, Y., Thermes, C. & Arneodo, A. (2005). From DNA sequence analysis to modelling replication in the human genome. *Phys. Rev. Lett.* 94, 248103.

40. Smit, A. F. A., Hubley, R. & Green, P. (1996-2004). RepeatMasker Open 3.0, http://www.repeatmasker.org.

41. Ladenburger, E.-M., Keller, C. & Knippers, R. (2002). Identification of a binding region for human origin recognition complex proteins 1 and 2 that coincides with an origin of DNA replication. *Mol. Cell. Biol.* 22, 1036–1048.

42. Taira, T., Iguchi-Ariga, S. M. & Ariga, H. (1994). A novel DNA replication origin identified in the human heat shock protein 70 gene promoter. *Mol. Cell. Biol.* 14, 6386–6397.

43. Keller, C., Ladenburger, E.-M., Kremer, M. & Knippers, R. (2002). The origin recognition complex marks a replication origin in the human TOP1 gene promoter. *J. Biol. Chem.* 277, 31430–31440.

44. Vassilev, L. & Johnson, E. M. (1990). An initiation zone of chromosomal DNA replication located upstream of the c-myc gene in proliferating HeLa cells. *Mol. Cell. Biol.* 10, 4899–4904.

45. Nenguke, T., Aladjem, M. I., Gusella, J. F., Wexler, N. S. & Arnheim, N. (2003). Candidate DNA replication initiation regions at human trinucleotide repeat disease loci. *Hum. Mol. Genet.* 12, 1021–1028.

46. Araujo, F. D., Knox, J. D., Ramchandani, S., Pelletier, R., Bigey, P., Price, G., Szyf, M. & Zannis-Hadjopoulos, M. (1999). Identification of initiation sites for DNA replication in the human dnmt1 (DNA-methyltransferase) locus. *J. Biol. Chem.* 274, 9335–9341.

47. Giacca, M., Zentilin, L., Norio, P., Diviacco, S., Dimitrova, D., Contreas, G., Biamonti, G., Perini, G., Weighardt, F. *et al.* (1994). Fine mapping of a replication origin of human DNA. *Proc. Natl. Acad. Sci. USA* 91, 7119–7123.

48. Kitsberg, D., Selig, S., Keshet, I. & Cedar, H. (1993). Replication structure of the human beta-globin gene domain. *Nature* 366, 588–590.

49. Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. & Miller, W. (2000). PipMaker –a web server for aligning two genomic DNA sequences. *Genome Res.* 10, 577–586.

50. Arneodo, A., Argoul, F., Bacry, E., Elezgaray, J. & Muzy, J.-F. (1995). *Ondelettes Multifractales et Turbulences : de l'ADN aux croissances cristallines.* Diderot Editeur, Arts et Sciences, Paris.

51. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

52. Bernardi, G. (1989). The isochore organization of the human genome. *Annu. Rev. Genet.* 23, 637–661.

53. Bernardi, G. (1995). The human genome: organization and evolutionary history. *Annu. Rev. Genet.* 29, 445–476.

54. Nekrutenko, A. & Li, W. H. (2000). Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* 10, 1986–1995.

55. Häring, D. & Kypr, J. (2001). No isochores in the human chromosomes 21 and 22? *Biochem. Biophys. Res. Commun.* 280, 567–573.

56. Bernardi, G. (2001). Misunderstandings about isochores. Part 1. *Gene* 276, 3–13.

57. Eyre-Walker, A. & Hurst, L. D. (2001). The evolution of isochores. *Nat. Rev. Genet.* 2, 549–555.

58. Li, W. (2002). Are isochore sequences homogeneous? *Gene* 300, 129–139.

59. Cohen, N., Dagan, T., Stone, L. & Graur, D. (2005). GC composition of the human genome: in search of isochores. *Mol. Biol. Evol.* 22, 1260–1272.

60. Pavlícek, A., Paces, J., Clay, O. & Bernardi, G. (2002). A compact view of isochores in the draft human genome sequence. *FEBS Lett.* 511, 165–169.

61. Hori, H. & Osawa, S. (1986). Evolutionary change in 5S rRNA secondary structure and a phylogenic tree of 352 5S rRNA species. *Biosystems* 19, 163–172.

62. D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C. & Bernardi, G. (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* 32, 504–510.

63. Graur, D. & Li, W. H. (1999). *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA.

64. Paulson, J. R. & Laemmli, U. K. (1977). The structure of histone-depleted metaphase chromosomes. *Cell* 12, 817–828.

65. Gasser, S. M. & Laemmli, U. K. (1987). A glimpse at chromosomal order. *Trends Genet.* 3, 16–22.

66. Rattner, J. B. & Lin, C. C. (1985). Radial loops and helical coils coexist in metaphase chromosomes. *Cell* 42, 291–296.

67. Boy de la Tour, E. & Laemmli, U. K. (1988). The metaphase scaffold is helically folded: sister chromatids have predominantly opposite helical handedness. *Cell* 55, 937–944.

68. Poirier, M. G., Nemani, A., Gupta, P., Eroglu, S. & Marko, J. F. (2001). Probing chromosome structure with dynamic force relaxation. *Phys. Rev. Lett.* 86, 360–363.

69. Bridger, J. M. & Bickmore, W. A. (1998). Putting the genome on the map. *Trends Genet.* 14, 403–409.

70. Cremer, T. & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* 2, 292–301.

71. Belmont, A. S. (2001). Visualizing chromosome dynamics with GFP. *Trends Cell Biol.* 11, 250–257.

72. Gasser, S. M. (2002). Visualizing chromatin dynamics in interphase nuclei. *Science* 296, 1412–1416.

73. Zink, D., Cremer, T., Saffrich, R., Fischer, R., Trendelenburg, M. F., Ansorge, W. & Stelzer, E. H. (1998). Structure and dynamics of human interphase chromosome territories in vivo. *Hum. Genet.* 102, 241–251.

74. Münkel, C. & Langowski, J. (1998). Chromosome structure predicted by a polymer model. *Phys. Rev. E* 57, 5888–5896.

75. Manuelidis, L. (1990). A view of interphase chromosomes. *Science* 250, 1533–1540.

76. Manuelidis, L. & Chen, T. L. (1990). A unified model of eukaryotic chromosomes. *Cytometry* 11, 8–25.

77. Woodcock, C. L., Woodcock, H. & Horowitz, R. A. (1991). Ultrastructure of chromatin. I. Negative staining of isolated fibers. *J. Cell Sci.* 99, 99–106.

78. Woodcock, C. L. (1994). Chromatin fibers observed in situ in frozen hydrated sections. Native fiber diameter is not correlated with nucleosome repeat length. *J. Cell Biol.* 125, 11–19.

79. Belmont, A. S. (1997). Large-scale chromatin organization. In *Genome Structure and Function*, p. 261. Kluwer Academic Publishers, Dordrecht.

80. Müller, W. G., Rieder, D., Kreth, G., Cremer, C., Trajanoski, Z. & McNally, J. G. (2004). Generic features of tertiary chromatin structure as detected in natural chromosomes. *Mol. Cell. Biol.* 24, 9359–9370.

81. Jackson, D. A. & Pombo, A. (1998). 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.* 140, 1285–1295.

82. Ma, H., Samarabandu, J., Devdhar, R. S., Acharya, R., Cheng, P. C., Meng, C. & Berezney, R. (1998). Spatial and temporal dynamics of DNA replication sites in mammalian cells. *J. Cell Biol.* 143, 1415–1425.

83. Berezney, R., Dubey, D. D. & Huberman, J. A. (2000). Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma* 108, 471–484.

84. Edenberg, H. J. & Huberman, J. A. (1975). Eukaryotic chromosome replication. *Annu. Rev. Genet.* 9, 245–284.

85. Hand, R. (1978). Eucaryotic DNA: organization of the genome for replication. *Cell* 15, 317–325.

86. Yurov, Y. B. & Liapunova, N. A. (1977). The units of DNA replication in the mammalian chromosomes: evidence for a large size of replication units. *Chromosoma* 60, 253–267.

87. Liapunova, N. A. (1994). Organization of replication units and DNA replication in mammalian cells as studied by DNA fiber radioautography. *Int. Rev. Cytol.* 154, 261–308.

88. Chargaff, E. (1951). Structure and function of nucleic acids as cell constituents. *Fed. Proc.* 10, 654–659.

89. Rudner, R., Karkas, J. D. & Chargaff, E. (1968). Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proc. Natl. Acad. Sci. USA* 60, 921–922.

90. Fickett, J. W., Torney, D. C. & Wolf, D. R. (1992). Base compositional structure of genomes. *Genomics* 13, 1056–1064.

91. Lobry, J. R. (1995). Properties of a general model of DNA evolution under no-strand-bias conditions. *J. Mol. Evol.* 40, 326–330.

92. Mrázek, J. & Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA* 95, 3720–3725.

93. Frank, A. C. & Lobry, J. R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238, 65–77.

94. Rocha, E. P., Danchin, A. & Viari, A. (1999). Universal replication biases in bacteria. *Mol. Microbiol.* 32, 11–16.

95. Tillier, E. R. M. & Collins, R. A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* 50, 249–257.

96. Brodie of Brodie, E.-B. (2005). *De l'analyse des séquences d'ADN à la modélisation de la réplication chez les mammifères.* Ph.D. thesis, ENS de Lyon, France.

97. Nicolay, S. (2006). *Analyse des séquences d'ADN par la transformée en ondelettes : extraction d'informations structurelles, dynamiques et fonctionnelles.* Ph.D. thesis, University of Liège, Belgium.

98. Gojobori, T., Li, W. H. & Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* 18, 360–369.

99. Li, W. H., Wu, C. I. & Luo, C. C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* 21, 58–71.

100. Petrov, D. A. & Hartl, D. L. (1999). Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc. Natl. Acad. Sci. USA* 96, 1475–1479.

101. Zhang, Z. & Gerstein, M. (2003). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* 31, 5338–5348.

102. Freeman, J. M., Plasterer, T. N., Smith, T. F. & Mohr, S. C. (1998). Patterns of genome organization in bacteria. *Science* 279, 1827.

103. Beletskii, A., Grigoriev, A., Joyce, S. & Bhagwat, A. S. (2000). Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *J. Mol. Biol.* 300, 1057–1065.

104. Francino, M. P. & Ochman, H. (2001). Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences. *Mol. Biol. Evol.* 18, 1147–1150.

105. Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* 12, 640–649.

106. Shioiri, C. & Takahata, N. (2001). Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *J. Mol. Evol.* 53, 364–376.

107. Green, P., Ewing, B., Miller, W., Thomas, P. J. & Green, E. D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* 33, 514–517.

108. Svejstrup, J. Q. (2002). Mechanisms of transcription-coupled DNA repair. *Nat. Rev. Mol. Cell Biol.* 3, 21–29.

109. Touchon, M. (2005). *Biais de composition chez les mammifères : rôle de la transcription et de la réplication.* Ph.D. thesis, University Denis Diderot, Paris VII, France.

110. Jacob, F., Brenner, S. & Cuzin, F. (1963). On the regulation of DNA replication in bacteria. *Cold Spring Harb. Symp. Quant. Biol.* 28, 329–342.

111. Bell, S. P. & Dutta, A. (2002). DNA replication in eukaryotic cells. *Annu. Rev. Biochem.* 71, 333–374.

112. Hyrien, O. & Méchali, M. (1993). Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos. *EMBO J.* 12, 4511–4520.

113. Gerbi, S. A. & Bielinsky, A. K. (2002). DNA replication and chromatin. *Curr. Opin. Genet. Dev.* 12, 243–248.

114. Schübeler, D., Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J. & Groudine, M. (2002). Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. *Nat. Genet.* 32, 438–442.

115. Fisher, D. & Méchali, M. (2003). Vertebrate HoxB gene expression requires DNA replication. *EMBO J.* 22, 3737–3748.

116. Anglana, M., Apiou, F., Bensimon, A. & Debatisse, M. (2003). Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell* 114, 385–394.

117. Gilbert, D. M. (2001). Making sense of eukaryotic DNA replication origins. *Science* 294, 96–100.

118. Coverley, D. & Laskey, R. A. (1994). Regulation of eukaryotic DNA replication. *Annu. Rev. Biochem.* 63, 745–776.

119. Sasaki, T., Sawado, T., Yamaguchi, M. & Shinomiya, T. (1999). Specification of regions of DNA replication initiation during embryogenesis in the 65-kilobase DNApolalpha-dE2F locus of Drosophila melanogaster. *Mol. Cell. Biol.* 19, 547–555.

120. Bogan, J. A., Natale, D. A. & Depamphilis, M. L. (2000). Initiation of eukaryotic DNA replication: conservative or liberal? *J. Cell. Physiol.* 184, 139–150.

121. Gilbert, D. M. (2004). In search of the holy replicator. *Nat. Rev. Mol. Cell Biol.* 5, 848–855.

122. Méchali, M. (2001). DNA replication origins: from sequence specificity to epigenetics. *Nat. Rev. Genet.* 2, 640–645.

123. Demeret, C., Vassetzky, Y. & Méchali, M. (2001). Chromatin remodelling and DNA replication: from nucleosomes to loop domains. *Oncogene* 20, 3086–3093.

124. McNairn, A. J. & Gilbert, D. M. (2003). Epigenomic replication: linking epigenetics to DNA replication. *Bioessays* 25, 647–656.

125. Brewer, B. J. (1988). When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. *Cell* 53, 679–686.

126. Rocha, E. P. C., Guerdoux-Jamet, P., Moszer, I., Viari, A. & Danchin, A. (2000). Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *J. Biotech.* 78, 209–219.

127. Lopez, P. & Philippe, H. (2001). Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. *C. R. Acad. Sci. III* 324, 201–208.

128. Rocha, E. P. C. (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes. *Trends Microbiol.* 10, 393–395.

129. Bulmer, M. (1991). Strand symmetry of mutation rates in the beta-globin region. *J. Mol. Evol.* 33, 305–310.

130. Francino, M. P. & Ochman, H. (2000). Strand symmetry around the beta-globin origin of replication in primates. *Mol. Biol. Evol.* 17, 416–422.

131. Gierlik, A., Kowalczuk, M., Mackiewicz, P., Dudek, M. R. & Cebrat, S. (2000). Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome? *J. Theor. Biol.* 202, 305–314.

132. Louie, E., Ott, J. & Majewski, J. (2003). Nucleotide frequency variation across human genes. *Genome Res.* 13, 2594–2601.

133. Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. A. & Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919.

134. Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J. D. & Wang, S. M. (2002). Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl. Acad. Sci. USA* 99, 12257–12262.

135. Rinn, J. L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N. M., Hartman, S., Harrison, P. M., Nelson, F. K., Miller, P. *et al.* (2003). The transcriptional activity of human Chromosome 22. *Genes Dev.* 17, 529–540.

136. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S. *et al.* (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* 14, 331–342.

137. Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y. *et al.* (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2, e162.

138. Girard-Reydet, C., Grégoire, D., Vassetzky, Y. & Méchali, M. (2004). DNA replication initiates at domains overlapping with nuclear matrix attachment regions in the xenopus and mouse c-myc promoter. *Gene* 332, 129–138.

139. Vassilev, L. T., Burhans, W. C. & DePamphilis, M. L. (1990). Mapping an origin of DNA replication at a single-copy locus in exponentially proliferating mammalian cells. *Mol. Cell. Biol.* 10, 4685–4689.

140. Codlin, S. & Dalgaard, J. Z. (2003). Complex mechanism of site-specific DNA replication termination in fission yeast. *EMBO J.* 22, 3431–3440.

141. Santamaria, D., Viguera, E., Martinez-Robles, M. L., Hyrien, O., Hernandez, P., Krimer, D. B. & Schvartzman, J. B. (2000). Bi-directional replication and random termination. *Nucleic Acids Res.* 28, 2099–2107.

142. Little, R. D., Platt, T. H. & Schildkraut, C. L. (1993). Initiation and termination of DNA replication in human rRNA genes. *Mol. Cell. Biol.* 13, 6600–6613.

143. White, E. J., Emanuelsson, O., Scalzo, D., Royce, T., Kosak, S., Oakeley, E. J., Weissman, S., Gerstein, M., Groudine, M. *et al.* (2004). DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc. Natl. Acad. Sci. USA* 101, 17771–17776.

144. Callan, H. G. (1972). Replication of DNA in the chromosomes of eukaryotes. *Proc. R. Soc. Lond. B Biol. Sci.* 181, 19–41.

145. Asakura, S. & Oosawa, F. (1954). On interaction between two bodies immersed in a solution of macromolecules. *J. Chem. Phys.* 22, 1255–1256.

146. Snir, Y. & Kamien, R. D. (2005). Entropically driven helix formation. *Science* 307, 1067.

147. van Holde, K. & Zlatanova, J. (1996). What determines the folding of the chromatin fiber? *Proc. Natl. Acad. Sci. USA* 93, 10548–10555.

148. Mergell, B., Everaers, R. & Schiessel, H. (2004). Nucleosome interactions in chromatin: fiber stiffening and hairpin formation. *Phys. Rev. E* 70, 011915.

149. Lesne, A. & Victor, J. M. (2005). Chromatin fiber functional organization: some plausible models. *Eur. Phys. J. E, in press* .

150. Woodcock, C. L., Grigoryev, S. A., Horowitz, R. A. & Whitaker, N. (1993). A chromatin folding model that incorporates linker variability generates fibers resembling the native structures. *Proc. Natl. Acad. Sci. USA* 90, 9021–9025.

151. Grossberg, A. Y. & Khoklov, A. R. (1994). In *Statistical Physics of Macromolecules, AIP series in Polymers and Complex Materials* (R. Larson & P. A. Pincus, eds.). AIP Press, Woodbury.

152. Jun, S., Bechhoefer, J. & Ha, B.-Y. (2003). Diffusion-limited loop formation of semiflexible polymers: Kramers theory and the intertwined time scales of chain relaxation and closing. *Europhys. Lett.* 64, 420–426.

153. Yamakawa, H. & Stockmayer, W. H. (1972). Statistical mechanics of wormlike chains. II. Excluded volume effects. *J. Chem. Phys.* 57, 2843–2854.

154. St-Jean, P., Vaillant, C., Audit, B. & Arneodo, A. (2006). Spontaneous emergence of rosette like folding of chromatin: a keystone to replication and transcription regulation. Preprint.

155. Reiss, H., Frisch, H. L. & Lebowitz, J. L. (1959). Statistical mechanics of rigid spheres. *J. Chem. Phys.* 31, 369–380.

156. Dinsmore, A. D., Yodh, A. G. & Pine, D. J. (1995). Phase diagrams of nearly-hard-sphere binary colloids. *Phys. Rev. E* 52, 4045–4057.

157. Carnahan, N. F. & Starling, K. E. (1969). Equation of state for nonattracting rigid spheres. *J. Chem. Phys.* 51, 635–636.

158. Minton, A. P. (2001). The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *J. Biol. Chem.* 276, 10577–10580.

159. Herrick, J., Stanislawski, P., Hyrien, O. & Bensimon, A. (2000). Replication fork density increases during DNA synthesis in X. laevis egg extracts. *J. Mol. Biol.* 300, 1133–1142.

160. Zlatanova, J. & Leuba, S. H. (2003). Chromatin fibers, one-at-a-time. *J. Mol. Biol.* 331, 1–19.

161. Tassius, C., Moskalenko, C., Minard, P., Desmadril, M., Elezgaray, J. & Argoul, F. (2004). Probing the dynamics of a confined enzyme by surface plasmon resonance. *Physica A* 342, 402–409.