La transformation en ondelettes continue : un microscope mathématique adapté à l'étude des propriétés d'invariance d'échelle et de corrélations à longue portée des séquences d'ADN

Alain ARNÉODO[1], Cédric VAILLANT[2], Benjamin AUDIT[3], Yves D'AUBENTON-CARAFA[4] et Claude THERMES[4]

[1]Laboratoire de Physique, École Normale Supérieure de Lyon,
46 Allée d'Italie, 69364 LYON Cedex 07, France

[2]FSB, Institut Bernoulli, Bat. MA-Ecublens, EPFL, Lausanne 1015, Suisse

[3]Computational Genomics Group, EMBL-European Bioinformatics Institute,
Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

[4]Centre de Génétique Moléculaire du CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France

alain.arneodo@ens-lyon.fr, cedric.vaillant@epfl.ch, audit@ebi.ac.uk,
daubenton@cgm.cnrs-gif.fr, thermes@cgm.cnrs-gif.fr

Résumé – Depuis le début des années 90, l'intérêt des mathématiciens, physiciens et informaticiens pour l'analyse statistique des séquences d'ADN n'a pas cessé de croître. En effet, les immenses progrès de la biologie moléculaire et les grands projets de séquençage ont révélé l'extraordinaire complexité des génomes. Afin de mieux comprendre l'organisation et l'évolution des génomes, il est apparu nécessaire d'introduire de nouveaux concepts et de nouvelles techniques d'analyse du signal. Ainsi la possibilité que les séquences d'ADN présentent des propriétés d'invariance d'échelle associées à l'existence de corrélations à longue portée (CLP) a été le sujet d'une longue controverse. La raison principale de ce malentendu est le caractère non stationnaire des séquences d'ADN résultant de l'hétérogénéité de composition des génomes. Cette observation nous a conduit à proposer l'utilisation de la transformation en ondelettes continue (TO) comme outil naturel d'analyse des séquences d'ADN : par un choix adéquat de l'ondelette analysatrice, on peut s'affranchir de la "structure mosaïque" de ces séquences et quantifier l'existence de CLP associées à des propriétés d'invariance d'échelle monofractales. L'exploration de séquences d'ADN du génome humain sous l'optique du microscope TO, nous a permis de démontrer l'existence de CLP dans les séquences exoniques (codantes pour les protéines) comme dans les séquences introniques (non codantes), remettant par là en cause les différentes interprétations de ces corrélations à l'aide de modèles de dynamique évolutive (plasticité) des génomes. En profitant de la disponibilité de génomes complets (levure, *E. coli*, ...) et en utilisant différentes tables expérimentales associant des di- ou tri- nucléotides à des grandeurs de nature structurelle (courbure, flexibilité), nous avons montré récemment qu'il est possible d'extraire des séquences d'ADN des informations sur l'organisation spatiale et dynamique de la double hélice dans les cellules via l'interaction avec certaines protéines de structure telles que les histones pour la formation du nucléosome eucaryote. En particulier, l'existence et la nature des CLP jusqu'à des distances de l'ordre de $3\,10^4$ nucléotides dans certains profils de courbure et/ou de flexibilité locales permettent de diagnostiquer la présence de nucléosomes dans les génomes étudiés. L'observation de certaines CLP dans tous les organismes eucaryotes ainsi que dans les organismes des deux autres règnes (eubactéries et archaeabactéries) suggère fortement que ces corrélations pourraient être essentielles aux phénomènes de condensation-décondensation de la chromatine en relation avec les processus de réplication, transcription et division cellulaire.

Abstract – The possible relevance of scale invariance and fractal concepts to the structural complexity of genomic sequences is the subject of considerable increasing interest. During the past ten years, there has been intense discussion about the existence, the nature and the origin of long-range correlations in DNA sequences. Different techniques including mutual information functions, autocorrelation functions, power spectra, "DNA walk" representation, Zipf analysis and entropies, were used for statistical analysis of DNA sequences. For years there has been some permanent debate on rather struggling questions like the fact that the reported long-range correlations might be just an artefact of the compositional heterogeneity of the genome organization. Another controversial issue is whether or not long-range correlation properties are different for protein-coding (exonic) and non-coding (intronic, intergenic) sequences. One of the main obstacles to fractal analysis is the mosaic structure of DNA sequences which are well known to be formed of "patches" ("strand biases") of different underlying compositions. When using the "DNA walk" representation, these patches appear as trends in the DNA walk landscapes that are likely to break scale invariance. Most of the techniques, e.g. the variance method, used so far for characterizing the presence of long-range correlations are not well adapted to study non stationary sequences. In previous works, we have emphasized the wavelet transform (WT) as a well suited technique to overcome this difficulty. By considering analysing wavelets that make the "WT miscroscope" blind to low frequency trends, any bias in the DNA walk can be removed and the existence of power-law correlations with specific scale invariance properties can be revealed accurately. In this paper, we review recent results obtained when exploring the scaling properties of eucaryotic, eubacterial and archaeal genomic sequences using the WT space-scale decomposition. These results suggest that the existence of long-range correlations up to distances $\sim 20-30$kbp is the signature of the nucleosomes and of the 30nm chromatin fiber that participate in DNA packaging in eucaryotic nuclei. We propose some understanding of these correlations as a necessity for chromosome condensation-decondensation processes in relation with DNA repication, gene expression and cell division.

1 Introduction

La relation entre la structure primaire de l'ADN et ses fonctions biologiques est un des grands enjeux de la biologie moderne. Il apparaît de plus en plus que la fonction des séquences d'ADN n'est pas seulement de coder pour les protéines mais également de contrôler la configuration spatiale des chaînes ADN ainsi que l'accessibilité et l'interaction des protéines avec la double hélice [1, 2]. D'un prime abord, les séquences d'ADN semblent désordonnées, sans organisation particulière exceptées certaines structures régulières tels que les motifs répétitifs. Mais globalement, mises à part ces organisations périodiques, la majeure partie de l'ADN génomique est indistinguable d'une séquence aléatoire, markovienne avec des corrélations à courte portée. Pour étudier ces propriétés de corrélation, ainsi que l'existence d'éventuelles périodicités (ou périodicités "cachées"), le recours à la Transformée de Fourier (TF) et à l'étude des fonctions de corrélation semble naturel ; de nombreuses études ont ainsi été menées avec ces techniques que ce soit sur des séquences eucaryotes que sur des séquences procaryotes [3–5]. Différentes oscillations ont effectivement été observées dans l'étude des distributions de di- et trinucléotides, telles que la périodicité de 3 paires de bases (*pb*) dans les séquences codantes [3, 5] qui reflète la structure en "codons" de ces régions et plusieurs autres périodicités autour de $10-11$ *pb* révélées également dans les trois règnes, voisines donc de la périodicité naturelle de 10.55 *pb* de la double hélice (pour l'ADN libre à l'équilibre) [3–5]. D'autres périodicités plus grandes ont été associées aux séquences répétées (Alu, LINE,...).

Parallèlement à l'étude de ces périodicités, les techniques basées sur la TF et le calcul des fonctions de corrélation ont été utilisées pour caractériser les propriétés d'invariance d'échelle des séquences d'ADN [6–11]. La mise en évidence et la caractérisation de ces propriétés d'invariance d'échelle permettent de rendre compte de la possible existence de CLP entre nucléotides ou groupes de nucléotides et par conséquent du caractère non essentiellement markovien des séquences génomiques. De nombreuses études, s'appuyant sur différentes méthodes d'analyse statistique, ont été mises en oeuvre afin de caractériser la nature "fractale" des séquences d'ADN. Nous citerons comme exemples les études basées sur la notion d'information mutuelle [6, 10, 11], les fonctions d'auto-corrélation [10, 11], sur la caractérisation de la "rugosité" des marches ADN [7, 12–15], les analyses de Zipf [16] ou les études basées sur le calcul d'entropies [17, 18]. La question fondamentale qui demeure cependant toujours sans réponse claire est la véritable origine de ces CLP. Un effort important a été consacré pour démontrer que ces observations étaient bien le fait d'un processus stochastique présentant des CLP et non le simple reflet de l'hétérogènéïté de composition relative à l'organisation des génomes [8, 10–14, 19]. Dans la mesure où la plupart des modèles proposés pour expliquer ces CLP sont basés sur la plasticité des génomes [6, 15, 20–22], encore fallait-il résoudre la controverse sur la relation entre le caractère codant ou non codant de la séquence et les propriétés de corrélations de celles-ci [6, 7, 9–11, 13, 15, 16, 23].

Il existe toutefois des raisons objectives quant à l'origine de cette controverse. La plupart des études des CLP dans les séquences d'ADN ont été effectuées avec différentes techniques qui possèdent toutes des avantages et des limitations. Celles-ci consistent toujours à caractériser les propriétés d'invariance d'échelle à travers l'évolution en loi de puissance de certaines quantités comme la variance des marches ADN, les fonctions de corrélation, le spectre de puissance... En pratique, de telles mesures font face à une limitation due aux effets de taille finie des séquences [24–26] qui peuvent affecter dramatiquement les lois d'échelle. Certaines précautions doivent par ailleurs être prises pour ce qui est de la définition de l'échantillon statistique. Une autre limitation de ces techniques est d'origine plus fondamentale. L'exposant de la loi de puissance mesuré ne représente que les propriétés globales d'invariance d'échelle des séquences et ne rend donc pas compte du caractère monofractal ou multifractal de la statistique [27, 28]. Pour une séquence monofractale, les divers exposants mesurés s'expriment tous en fonction d'un seul et unique exposant, l'exposant de Hurst H associé à la marche ADN [15, 28]. $H = 1/2$ correspond à une marche aléatoire (Brownienne) classique décorrélée ; $H > 1/2$ caractérise les marches persistantes qui présentent des CLP ; $H < 1/2$ correspond aux marches anti-persistantes avec pas anti-corrélés. Dans le cas des séquences multifractales, l'exposant de Hurst ne suffit plus à caractériser les propriétés d'invariance d'échelle ; il est alors nécessaire de mesurer un continuum d'exposants (spectres multifractals) afin de rendre compte statistiquement des fluctuations locales des propriétés d'invariance d'échelle [27, 28].

Une des principales difficultés de l'analyse des CLP dans les séquences d'ADN est la présence d'une inhomogénéité dans la composition moyenne des nucléotides le long des séquences (structure "mosaïque" des génomes) [29, 30]. Les lois de tirage associées aux différents nucléotides sont donc souvent "biaisées" ce qui se traduit, par exemple, par la présence de tendances dans les marches ADN. Selon la technique utilisée, ce biais statistique peut conduire à une erreur dans l'estimation des propriétés d'invariance d'échelle [7, 8, 12–15, 19], voire masquer totalement l'invariance d'échelle associée aux fluctuations de composition. Plusieurs méthodes, plus ou moins "ad hoc", ont été ainsi mises en oeuvre pour s'affranchir des ces tendances, comme la méthode du "min-max" [7] et la "detrended fluctuation analysis" [23, 31]. Récemment, la transformée en ondelettes (TO) [32–35] a été introduite avec un certain succès pour l'étude des séquences ADN [27, 28, 36]. Déjà utilisée dans de nombreux domaines des sciences fondamentales et appliquées, la TO apparaît très bien appropriée pour l'étude des propriétés fractales de certains processus et ce, même en présence de composantes polynômiales non invariantes d'échelle [34–36]. En considérant simplement des ondelettes analysatrices capables de "filtrer" les tendances basses fréquences, il est possible de détecter et de quantifier les propriétés d'invariance d'échelle des marches ADN [27, 28, 36]. Dans cette communication, nous passons en revue les résultats les plus significatifs obtenus en appliquant le méthode des maxima du module de la transformée en ondelettes (MMTO) à diverses séquences génomiques appartenant aux trois règnes eucaryotes, procaryotes et archaebactéries.

2 Un formalisme multifractal basé sur la transformation en ondelettes

2.1 TO et analyse des singularités

Un comportement singulier d'un signal $f(x)$ en un point x_0 est généralement caractérisé par un exposant, appelé exposant de Hölder $h(x_0)$, qui est défini comme le plus grand exposant h tel qu'il existe un polynome P_n de degré n et une constante C vérifiant [34–36] :

$$|f(x) - P_n(x-x_0)| \leqslant C|x-x_0|^h \ \ (x \to x_0). \tag{1}$$

L'exposant de Hölder $h(x_0)$ quantifie donc bien la force de la singularité localisée en x_0 : plus $h(x_0)$ est grand, plus la singularité est faible. En fait, si $h(x_o)$ n'est pas un entier, $n < h(x_0) < n+1$, on peut se convaincre facilement que f est n-fois dérivable en x_0, $f \in C^n(x_0)$, mais pas $n+1$-fois dérivable. Le spectre des singularités est défini par la fonction [28, 34, 36] :

$$D(h) = d_F\left(\{x_o \in R / h(x_o) = h\}\right), \tag{2}$$

où d_F symbolise la dimension fractale. Il quantifie statistiquement les contributions relatives des différentes singularités présentes dans le signal.

L'estimation locale de l'exposant de Hölder nécessite l'utilisation d'une technique permettant de s'affranchir des comportements polynomiaux tels que $P_n(x-x_o)$ dans l'equation (1). La TO est tout à fait adaptée en la circonstance. La TO consiste à décomposer un signal sur une base de fonctions génératrices construites à partir d'une seule fonction, l'ondelette analysatrice ψ, par simples translations et dilatations. La TO d'une fonction f est définie par [32–36] :

$$T_\psi[f](x_o,a) = \frac{1}{a}\int_{-\infty}^{+\infty} f(x)\psi\left(\frac{x-x_0}{a}\right)\mathrm{d}x, \tag{3}$$

où x définit le paramètre d'espace et a (>0) le paramètre d'échelle. En choisissant ψ telle que ses N premiers moments soient nuls ($\int x^n\psi(x)dx = 0, \ \forall n, \ 0 \leqslant n < N$), on peut démontrer que pour $N > h(x_0)$, alors $h(x_0)$ peut être estimé à partir du comportement en loi de puissance de la TO [37, 38] :

$$T_\psi[f](x_o,a) \sim a^{h(x_0)}, \ \ a \to 0^+. \tag{4}$$

Dans la pratique, l'exposant de Hölder est mesuré le long des lignes de maxima du module de la TO qui définissent son squelette [34, 35, 39]. Dans le présent travail, nous utiliserons principalement les dérivées successives de la Gaussienne [34, 36] :

$$g^{(N)}(x) = \frac{d^N e^{-x^2/2}}{dx^N} \tag{5}$$

comme ondelettes analysatrices d'ordre N.

2.2 La méthode MMTO

Une façon naturelle de généraliser le formalisme multifractal aux fonctions et plus généralement aux distributions fractales consiste à revoir les algorithmes classiques de comptage de boites en remplaçant les boites par des "boites oscillantes", à savoir les ondelettes. La méthode MMTO [40, 41] consiste à utiliser le squelette de la TO pour positionner ces "boites oscillantes" dans le demi-plan espace-échelle et définir ainsi la fonction de partition :

$$\mathscr{Z}(q,a) = \sum_{x_i \in \mathscr{S}(a)} |T_\psi[f](x_i,a)|^q \sim a^{\tau(q)}, \tag{6}$$

où $q \in R$ et $\mathscr{S}(a)$ est l'ensemble des maxima locaux de $|T_\psi[f]|$ à l'échelle a. Le résultat principal de la méthode MMTO est que le spectre $D(h)$ des singularités (Eq. (2)) peut être déterminé par simple transformation de Legendre du spectre d'exposants $\tau(q)$:

$$D(h) = \min_q(qh - \tau(q)). \tag{7}$$

Il est important de remarquer que les signaux homogènes monofractals mettant en jeu des singularités de même exposant de Hölder $h(x) = H$, sont caractérisés par un spectre $\tau(q)$ linéaire ($H = \partial\tau/\partial q$). Au contraire, un spectre $\tau(q)$ non linéaire est la signature de la nature multifractale du signal analysé avec un exposant de Hölder $h(x)$ qui dépend de x et prend généralement des valeurs $h_{min} \leqslant h \leqslant h_{max}$. La méthode MMTO a été testée sur des signaux synthétiques monofractals (escaliers du diable, signaux Browniens fractionnaires) et multifractals (cascades aléatoires sur des bases ondelettes) [40, 41] et appliquée avec certains succès à divers domaines des sciences fondamentales comme la turbulence pleinement développée, les phénomènes de croissance fractale, les signaux financiers, les signaux médicaux... [34, 36, 39, 40].

3 Analyse en ondelettes de la complexité des séquences d'ADN

3.1 Codages

Pour caractériser quantitativement les propriétés statistiques, les lois de distribution des nucléotides et les possibles corrélations, il faut associer à la séquence (n_i), et donc au texte d'alphabet {A,C,T,G} pour les quatre bases adenine, cytosine, thymine et guanine, un signal digital (u_i). En fait, le choix du codage utilisé est une étape préalable importante et essentielle. Dans cette étude, nous utilisons dans un premier temps des codages binaires relatifs à la distinction entre couples de nucléotides : pour caractériser, par exemple, la distinction relative des purines et pyrimidines, il suffit d'associer la valeur $u = +1$ aux purines (A ou G) et $u = -1$ aux pyrimidines (C ou T). Ces codages sont notamment pratiques pour représenter les séquences d'ADN sous forme de "marche aléatoires" [7] : en utilisant u_i comme variable d'incrément, la "marche ADN" correspond à construire une fonction f correspondant au déplacement du marcheur en fonction du nombre n de nucléotides lus :

$$f(n) = \sum_{i=1}^{n} u_i. \tag{8}$$

Le codage doit donc être adapté à ce que l'on veut caractériser. Si on s'intéresse à la répartition des mononucléotides, un codage naturel est d'attribuer la valeur $u = +1$ à la position du nucléotide considéré et $u = 0$ aux autres positions [9]. Ce codage peut évidemment être étendu à n'importe quel autre motif, di, tri, ...nucléotides [9, 42, 43]. Il est possible également d'utiliser d'autres codages reflétant des propriétés particulières associées à ces motifs nucléotidiques, comme par exemple la flexibilité ou la courbure locale de la chaîne ADN. A ce propos,

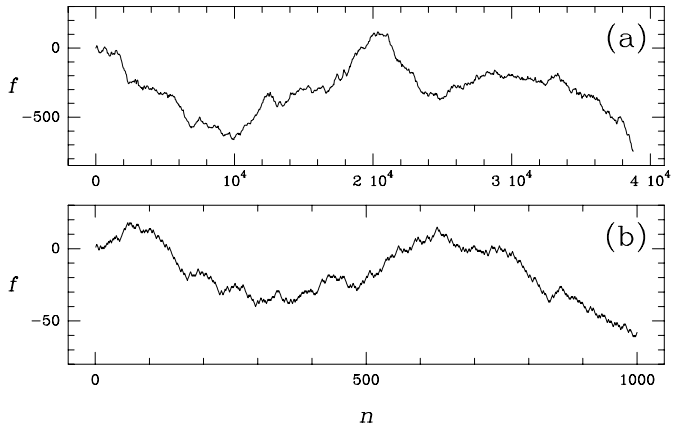

FIG. 1 – Exemples de marches ADN (Eq. (8)) pour la distinction purine-pyrimidine ($A = G = 1$ et $C = T = -1$). (a) Intron du gène de la dystrophine humaine ($L = 38770$). (b) 1000 premiers nucléotides de l'intron représenté en (a).

nous utiliserons dans la section 4, certaines tables trinucléotidiques établies expérimentalement à partir de la structure des nucléosomes eucaryotes (Table Pnuc) [44] et des données de digestion de l'ADN par la DNase I (Table DNase) [45].

Pour illustration, nous avons représenté dans la Fig. 1(a), la marche ADN obtenue par codage purine-pyrimidine d'un grand intron humain. Comme on peut le constater, le signal ainsi obtenu est très chahuté et très irrégulier. En particulier, lorsqu'on compare ce signal à celui obtenu sur la Fig. 1(b) pour les 1000 premiers nucléotides (environ 2.6 % de la longueur totale), on constate une très forte similitude de comportement ce qui suggère que comme les signaux fractals, les marches ADN présentent des propriétés d'invariance par rapport aux dilatations. La suite de ce travail va consister à quantifier ces propriétés d'invariance d'échelle à l'aide des techniques ondelettes et de démontrer l'existence possible de CLP.

3.2 Stationnarisation des séquences d'ADN

Sur la Fig. 2(a) est représenté la marche ADN obtenu par codage purine-pyrimidyne de la séquence du bactériophage λ. Le signal ainsi obtenu est caractéristique de la présence de tendances (auxquelles se superposent des fluctuations) correspondant à une prépondérance de purines ou de pyrimidines et qui traduisent l'hétérogénéïté de composition des génomes, c'est-à-dire la non stationnarité des marches ADN. C'est cette caractéristique fondamentale, quelquefois appelée structure "mosaïque" des génomes, qui est à l'origine de la controverse existant dans la littérature quant à la réelle existence de CLP dans les séquences d'ADN [6–26]. Dans la Fig. 2(b) est représentée la TO du signal de la Fig. 2(a), calculée avec la dérivée première de la Gaussienne $g^{(1)}$ [27, 28]. La structure arborescente de cette représentation espace-échelle, est caractéristique de celles obtenues pour les signaux fractals et multifractals synthétiques comme expérimentaux (Browniens fractionnaires, vitesse turbulente, séries boursières,...) [34, 36]. Cette structure branchée est un fait observée quelle que soit la séquence considérée. Dans les Figs. 2(d) et 2(e) sont représentées deux coupes de $T_{g^{(1)}}[f](x,a)$ aux échelles $a_1 = 12$ et $a_2 = 384$ nucléotides. A petite échelle ($a = a_1$), l'ondelette $g^{(1)}$ est orthogonale aux

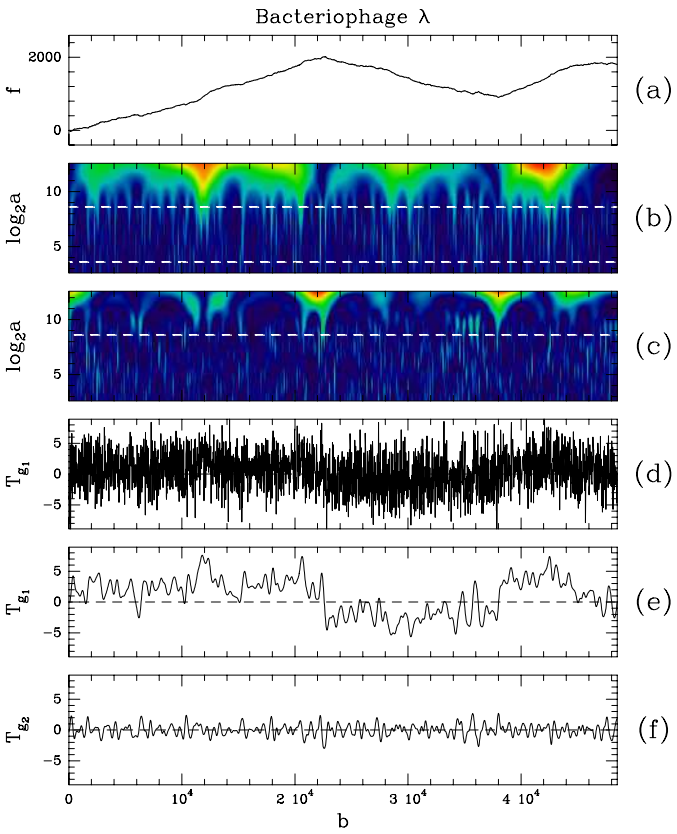

FIG. 2 – Analyse en ondelettes du génome du bactériophage λ ($L = 48502$). (a) Marche ADN $f(n)$ pour la distinction purine-pyrimidine en fonction de la position des nucléotides n. (b) TO de $f(n)$ calculée avec $g^{(1)}$ (Eq. (5)) ; $T_{g^{(1)}}(b,a)$ est codé, indépendemment à chaque échelle a, en utilisant 256 couleurs du noir ($\min_b |T_{g^{(1)}}|$) au rouge ($\max_b |T_{g^{(1)}}|$). (c) Même analyse qu'en (b) mais en utilisant l'ondelette d'ordre 2 $g^{(2)}$. (d) $T_{g^{(1)}}(b, a = a_1)$ vs b pour $a_1 = 12$ (nucléotides). (e) $T_{g^{(1)}}(b, a = a_2)$ vs b pour $a_2 = 384$ (nucléotides). (f) Même analyse qu'en (e) mais pour $g^{(2)}$.

constantes ($N = 1$) et seules les fluctuations locales de $f(x)$ sur une distance caractéristique d'une dizaine de nucléotides sont filtrées. A plus grande échelle ($a = a_2$), on s'aperçoit que la TO commence par osciller autour d'une valeur finie positive jusqu'à $n \sim 22000$, puis oscille autour d'une valeur finie négative jusqu'à $n \sim 38\,000$, pour enfin osciller autour d'une valeur finie positive, qui ne sont rien d'autre que les pentes des 3 tendances (linéaires) visibles sur la Fig. 2(a). Ainsi la présence de ces tendances perturbe le comportement dans les échelles des coefficients en ondelettes et biaise les comportements en lois de puissance espérés (Eqs. (4) et (6)) [27, 28]. Sur la Fig. 2(c) est représentée la TO calculée avec la dérivée seconde de la Gaussienne $g^{(2)}(x)$ ($N = 2$) qui elle est "aveugle" aux comportements constants ainsi qu'aux comportements linéaires. Cette fois, quelle que soit l'échelle, comme cela est illustré dans la Fig. 2(f) pour $a = a_2$, les coefficients en ondelettes oscillent autour de la valeur zero, véritable preuve que le signal peut être désormais considéré comme stationnaire [27, 28]. En utilisant des ondelettes analysatrices d'ordre plus élevé ($g^{(N)}$, $N = 3, 4, \ldots$), on peut ainsi "se débarasser" de tendances non linéaires à priori plus complexes susceptibles

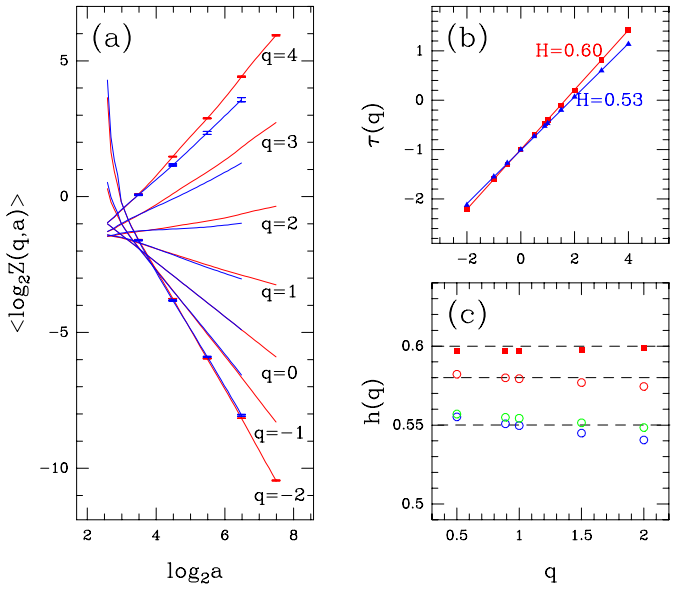

FIG. 3 – Analyse MMTO comparative des marches ADN pour le codage "G" de 2184 introns (rouge) ($L \geq 800$) et 226 exons (bleu) ($L \geq 600$) du génome humain. Les résultats présentés correspondent à la moyenne sur notre échantillon statistique d'introns et d'exons. (a) $< \log_2 \mathscr{Z}(q,a) >$ vs $\log_2 a$ pour diverses valeurs de q. (b) $\tau(q)$ vs q ; les lignes correspondent au spectre théorique $\tau(q) = qH - 1$ pour le mouvement Brownien fractionnaire avec $H = 0.60 \pm 0.02$ (introns) et 0.53 ± 0.02 (exons). (c) $h(q) = (\tau(q)+1)/q$ vs q pour les sous-séquences codantes relatives aux positions 1 (cercles bleus), 2 (cercles vert) et 3 (cercles rouges) des bases dans les codons ; les données pour les introns (carrés rouges) sont présentées pour comparaison ; les lignes pointillées horizontales correspondent au valeurs suivantes de l'exponent de Hurst $H = 0.55$, 0.58 et 0.60.

d'induire des CLP de caractère "trivial". En pratique les résultats obtenus avec $g^{(3)}$ et $g^{(4)}$ étant identiques à ceux obtenus avec $g^{(2)}$, nous nous limiterons dans cette étude à passer en revue les résultats obtenus avec cette dernière ondelette d'ordre 2.

3.3 Démonstration du caractère monofractal des marches ADN

Dans la Fig. 3 sont rapportés les résultats de l'analyse statistique de séquences codantes et non codantes dans le génome humain à l'aide de la méthode MMTO décrite dans la section 2.2 [27, 28, 36]. Le codage utilisé est le codage mononucléotide "G" et l'ensemble statistique est constitué de 2184 introns (non codants) de longueur $L \geqslant 800$ et de 226 exons de longueur $L \geqslant 600$, sélectionnés dans la base de données de l'EMBL. Le critère de longueur utilisé résulte d'un compromis entre la nécessité de maitriser les effets de taille finie (les séquences analysées sont relativement courtes puisque $< L >_{introns} \simeq 800$ et $< L >_{exons} \simeq 150$) et d'assurer la convergence statistique des fonctions de partition [46]. Dans la Fig. 3(a) sont présentés les résultats du calcul des fonctions de partition $\mathscr{Z}(q,a)$ (Eq. (6)) calculées avec l'ondelette $g^{(2)}$, en fonction de l'échelle a en représentation logarithmique. Pour un domaine raisonablement important de valeurs de q : $-2 \leqslant q \leqslant 4$, $\mathscr{Z}(q,a)$ manifeste d'incontestable propriétés d'invariance d'échelle. Les spectres $\tau(q)$ estimés, par régression linéaire sur la gamme d'échelles $20 \lesssim a \lesssim 120$, pour les séquences introniques et exoniques sont représentés dans la Fig. 3(b). Pour ces séquences respectivement non codantes et codantes, l'ensemble des données numériques obtenues pour différentes valeurs de q, se placent remarquablement sur une droite apportant par là, la preuve de la nature monofractale des marches ADN analysées [27, 28, 36]. Toutefois, il existe une différence importante entre ces deux classes de séquences, à savoir que la pente obtenue pour le spectre $\tau(q)$ des introns, $H = \partial\tau(q)/\partial q = 0.60 \pm 0.02$, est significativement plus grande que celle estimée pour les exons $H = 0.53 \pm 0.02$. Il s'agit là d'une indication claire que les marches des séquences introniques présentent des CLP ($H > 1/2$), alors que celles des séquences exoniques ressemblent plus à des marches au hasard non corrélées ($H \simeq 1/2$). A première vue, ces résultats sont en bon accord avec les conclusions de précédents travaux concernant l'existence de CLP uniquement dans les séquences non codantes [7, 15, 16, 23, 47]. Des résultats tout à fait similaires sont obtenus en analysant individuellement les introns et exons de plus grande longueur repertoriés dans la base de données EMBL [27, 28, 36].

Un des résultats les plus marquants de notre analyse MMTO dans la Fig.3 est le fait que les spectres $\tau(q)$ obtenus pour les introns et les exons sont remarquablement reproduits par le spectre théorique $\tau(q) = qH - 1$ prédit pour les marches aléatoires Browniennes fractionnaires [27, 28, 36]. Remarquons que ce résultat est tout à fait robuste et ne dépend pas du codage mononucléotidique utilisé. Nous verrons dans la section 4 que la nature monofractale des marches ADN peut être vérifiée et confirmée par le calcul de la densité de probabilité $\rho_a(T)$ des coefficients en ondelettes à différentes échelles a. En effet, la TO d'un processus auto-similaire d'exposant H étant elle aussi auto-similaire, alors $\rho_a(T)$ vérifie la relation d'auto-similarité [27, 28, 36] :

$$a^H \rho_a(a^H T) = \rho_1(T). \qquad (9)$$

La densité à n'importe quelle échelle a se déduit de celle calculée à l'échelle 1, par une simple dilatation mettant en jeu un exposant H unique.

3.4 Démonstration de l'existence de CLP dans les séquences d'ADN codant pour les protéines

Dans cette section, nous apportons un soin tout particulier dans la mesure des propriétés d'invariance d'échelle des séquences exoniques. A cause de la périodicité 3 induite par la structure en codons des séquences d'ADN codant pour les protéines, il est naturel d'examiner séparément les trois sous-séquences relatives aux premieres, deuxièmes et troisièmes bases de chaque codon [48]. Les résultats de cette analyse sont résumés dans la Fig. 3(c) [28, 36]. Les données obtenues par la méthode MMTO pour le spectre d'exposant $\tau(q)$ sont à nouveau tout a fait compatibles avec un spectre linéaire caractéristique de l'existence d'un seul exposant de Hölder. Pour les deux sous-séquences correspondant aux positions 1 et 2 dans les codons, on obtient la même pente $h(q) = \partial\tau/\partial q = 0.55 \pm 0.02$ qui est quasiment indistinguable de la valeur précédemment obtenue pour

la moyenne des séquences exoniques. De façon tout à fait surprenante, les données correspondant à la position 3 dans les codons présentent une pente $h(q) = 0.58 \pm 0.02$ significativement plus grande que $1/2$ et très proche de la valeur $H = 0.60 \pm 0.02$ obtenue précédemment pour l'ensemble des introns. Cette observation suggère que la troisième sous-séquence codante présente les mêmes propriétés de CLP que les séquences non codantes [28, 36, 48]. Plusieurs interprétation évolutives, c'est-à-dire faisant référence à la plasticité des génomes [6, 15, 20–22], ont été proposées pour l'existence de CLP dans les séquences d'ADN. Parmi les mécanismes invoqués, nos résultats permettent d'exclure les insertions et délétions de fragments de taille variable qui sont rares dans les régions exoniques à cause des contraintes imposées par leurs propriétés codantes. Des arguments de pression de sélection peuvent aussi être invoqués pour expliquer la différence entre les corrélations observées entre les bases en première et deuxième position respectivement. En effet, il est bien connu que la dégénérescence du code génétique est principalement contenue dans la troisième position de chaque codon.

4 Vers une interprétation de nature structurale et dynamique des CLP dans les séquences d'ADN

La disponibilité du premier génome eucaryote complètement séquencé, la levure *Saccharomyces cerevisiae*, permet d'effectuer une analyse comparative des propriétés d'invariance d'échelle des différents chromosomes d'un même organisme [42, 43]. Sur la Fig. 4(a) sont surperposées les courbes correspondant à l'estimation de l'écart type des coefficients en ondelettes $\sigma(a)$ ($\sim a^H$) pour les marches ADN obtenues avec le codage "A" pour les 16 chromosomes de la levure. Elles présentent toutes des comportements identiques avec une échelle caractéristique qui sépare deux régimes d'invariance d'échelle. A petite échelle, $20 \lesssim a \lesssim 200$ (exprimé en nombre de nucléotides), nous observons des CLP caractérisées par $H = 0.59 \pm 0.02$, une valeur significativement plus grande que $1/2$. A grande échelle, $a \gtrsim 200$, apparaissent des CLP plus fortes avec $H = 0.82 \pm 0.01$ jusqu'à une échelle de l'ordre de 10 *kpb* au dessus de laquelle on n'observe plus de corrélations. Lorsque l'on effectue la même analyse en ondelettes sur les profils de courbure des chromosomes de la levure obtenus à partir de la table Pnuc (Fig. 4(a)), on constate une similarité frappante avec les résultats précédents obtenus pour les marches ADN mononucléotidiques, et ce aussi bien à petite qu'à grande échelle. Cette resemblance n'est pas une conséquence triviale du recodage des mêmes séquences. En effet, si l'on code les séquences par la table DNase, on constate un net affaiblissement de l'exposant H à grande échelle ($H \simeq 0.6$). L'existence de ces deux régimes d'invariance d'échelle est confirmée dans la Fig. 5(a) où les densités de probabilité (dps) des coefficients en ondelettes pour les profils de courbure Pnuc de la levure calculées pour différentes échelles, se superposent sur une même courbe comme le prédit l'équation d'auto-similarité (9) [27, 28]. A petite échelle, les dps sont très bien approximées par une Gaussienne alors qu'à grande échelle les dps présentent des queues étirées de type exponentiel. Le fait que

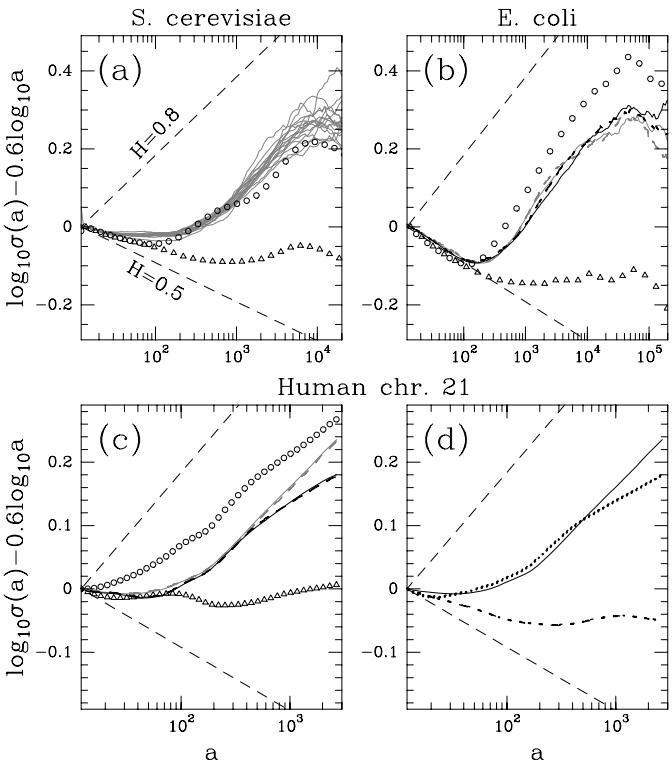

FIG. 4 – Estimation globale de l'écart type des coefficients de TO : $\log_{10}\sigma(a) - 0.6\log_{10}a$ est tracé en fonction de $\log_{10}a$; les lignes pointillées correspondant à l'absence de CLP ($H = 1/2$) et à de fortes CLP ($H = 0.80$) sont dessinées pour guider l'œil. (Une ligne horizontale dans cette représentation correspond à $H = 0.6$). L'ondelette analysatrice est $g^{(2)}$. (a) *S. cerevisiae :* Marches ADN pour le codage "A" des 16 chromosomes de la levure (——) et des profils de courbure Pnuc (○) et DNase (△) moyennés sur les 16 chromosomes. (b) *Escherichia coli :* Marches ADN pour les codages "A" (gris ——), "T" (gris – – –), "G" (noir ——) et "C" (noir – – –) et profils de courbure Pnuc (○) et DNase (△). (c) *Chromosome 21 humain :* Marches ADN "A", "C", "G" et "T" et profils de courbure Pnuc et DNase ; mêmes symboles qu'en (b). (d) *Chromosome 21 humain :* analyse comparative des marches ADN pour les codages "A" (——), "AA" (···) et "A isolé" (– – –) (tiré de la Réf. [42]).

l'équation d'auto-similarité (9) soit vérifiée dans les deux régimes confirme le caractère monofractal des fluctuations de rugosité des profils de courbure de la levure. Ces propriétés d'auto-similarité sont aussi observées sur les marches ADN de cet organisme [42, 43]. Il est important de souligner que les résultats que nous venons de présenter pour les chromosomes de la levure sont représentatifs de l'ensemble des résultats que nous avons obtenus pour un grand nombre de séquences d'ADN issues d'organismes eucaryotes : homme, rongeurs, oiseaux, plantes et insectes (voir la Fig. 4(c) pour le chromosome 21 humain et la Table 1 dans la Réf. [43]). Remarquons que l'échelle caractéristique trouvée pour les eucaryotes supérieurs est légèrement plus petite $a^* \simeq 100 - 140$ *pb* que celle de la levure. Dans une étude en cours, nous étudions de possibles écarts aux statistiques Gaussiennes à petite échelle lorsque la teneur en $(G+C)$ des séquences augmente.

La similitude frappante de l'ensemble des propriétés d'invariance d'échelle observées chez les génomes eucaryotes, nous a

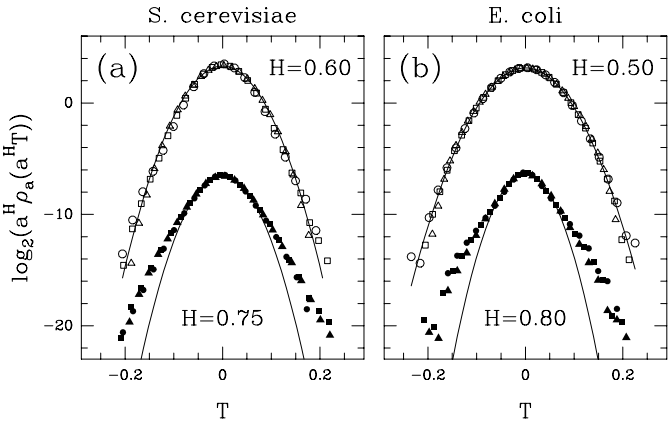

FIG. 5 – Densité de probabilité des coefficients en ondelettes pour les profils de courbure Pnuc. L'ondelettes analysatrice est $g^{(2)}$. (a) *Saccharomyces cerevisiae* : $\log_2(a^H \rho_a(a^H T))$ vs T pour les échelles $a = 12$ (△), 24 (□), 48 (○), 192 (▲), 384 (■), and 768 (●) exprimées en nombre de nucléotides ; $H = 0.60$ ($H = 0.75$) à petite (grande) échelle. (b) *Escherichia coli* : comme dans (a) mais pour $H = 0.50$ ($H = 0.80$) à petite (grande) échelle (tiré de la Réf. [42]).

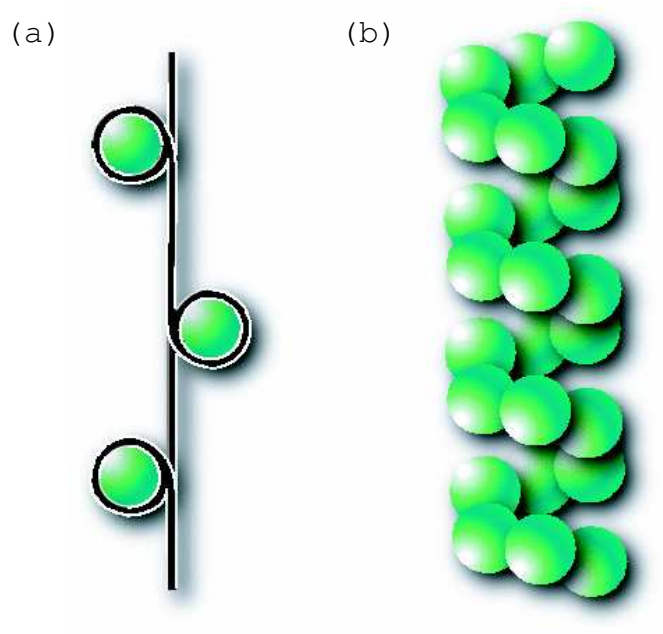

FIG. 6 – Représentation schématique de deux premiers niveaux de compaction de l'ADN dans le noyau eucaryote. (a) Premièrement, 146 *pb* s'enroulent sur presque deux tours autour d'un noyau constitué de 8 protéines histones pour former le nucléosome. Cette structure de 11 *nm* telle des perles sur un fil représente une compression d'un facteur 10 de l'ADN. (b) Deuxièmement, les nucléosomes s'agrègent pour former la fibre de 30 *nm*. Ce niveau polynucléosomal de repliement compacte l'ADN d'un facteur 40 supplémentaire.

incité à vérifier s'il en était de même chez les bactéries [42, 43]. Sur la Fig. 4(b) nous rapportons les résultats obtenus pour *Escherichia coli* qui sont représentatifs de ce que nous avons observé pour d'autres génomes d'eubactéries. Comme pour les eucaryotes, nous observons à nouveau une échelle caractéristique $a^* \simeq 200$ *pb* qui délimite la transition vers un régime de fortes CLP avec $H = 0.80 \pm 0.05$ à grande échelle. Comme pour la levure (Fig. 4(a)), si l'on utilise la table DNase pour coder les séquences de l'homme (Fig. 4(c)) ou de *E. coli* (Fig. 4(b)), on observe un affaiblissement significatif des CLP à grande échelle relativement à celles obtenues avec la table Pnuc. Dans la Fig. 5(b) sont représentées les dps des coefficients en ondelettes pour le profils de courbure Pnuc d'*E. coli*. Comme pour la levure (Fig. 5(a)), ces courbes montrent l'existence d'une échelle de transition entre deux régimes d'invariance d'échelle monofractals caractérisés par $H = 0.50 \pm 0.02$ et $H = 0.80 \pm 0.05$, respectivement. Il existe cependant une différence majeure entre les génomes eucaryotes et eubactériens : aucune CLP n'est observée pour ces derniers dans le régime à petite échelle où un comportement Brownien non corrélé avec $H = 1/2$ est observé (Figs. 4(b) et 5(b)) [42, 43].

Quels mécanismes ou phénomènes pourraient expliquer les CLP observées à petite échelle dans les génomes eucaryotes ? Leur absence totale pour les génomes eubactériens suggèrent qu'elles pourraient être reliées à certains motifs nucléotidiques dans les régions d'ADN de 150 *pb* qui s'enroulent autour d'octamères d'histones pour former les nucléosomes eucaryotes (Fig. 6) [1, 49]. En effet, même si l'ADN génomique eubactérien s'associe aussi à des protéines de structure (e.g. HU), aucun complexe de type nucléosomal n'a été mis à jour chez ces organismes [50]. De même, l'observation de CLP à petite échelle pour les génomes archaebactériens (Fig. 7 pour *Archaeoglobus fulgidus*) est en accord avec l'existence d'une structure analogue au nucléosome eucaryote chez les archaebactéries [51]. Notons que les génomes archaebactériens (comme eubactériens) sont principalement codant ($\simeq 90\%$) ; ce résultat constitue donc une nouvelle illustration que le caractère codant ou non codant des séquences d'ADN n'est pas déterminant pour la présence de CLP.

Nous avons poursuivi cette étude avec les génomes viraux. Des CLP à petite échelle sont clairement detectées pour la plupart des virus eucaryotes à génomes ADN double brin, comme cela est illustré par le virus *Epstein-Barr* dans la Fig. 7. Puisque l'ADN de ces virus forme des nucléosomes dans le noyau cellulaire [52], on confirme ainsi l'hypothèse d'une contrainte nucléosomale à l'origine des CLP. Les seuls virus animaux à génome ADN qui se répliquent dans le cytoplasme de leur cellule hôte, à savoir les poxvirus, codent pour une protéine de structure de type eubactérienne [53] et aucune CLP n'est trouvée dans cette gamme d'échelle comme cela est illustré dans la Fig. 7 pour le virus *Melanoplus sanguinipes* [43]. A nouveau, cette observation est cohérente avec notre hypothèse et suggère que le processus de compaction de l'ADN génomique de ces virus possède des caractéristiques différentes de celui des autres virus animaux. D'autres classes de virus comme les virus à ARN simple ou double brin (à l'exeption des retrovirus) ne forment pas de nucléosomes. Dans tous les cas, nous observons une absence totale de CLP à petite échelle comme cela est illustré dans la Fig. 7(c). Pour les retrovirus, il est connu que le génome viral intégré sous forme ADN dans le génome de l'hôte forme des nucléosomes dans le noyau cellulaire [54] ; nous confirmons de façon claire dans la Fig. 7(c), l'existence de CLP à petite échelle ($H \simeq 0.57 \pm 0.02$). Finalement, les bactériophages ne présentent pas de CLP à petite

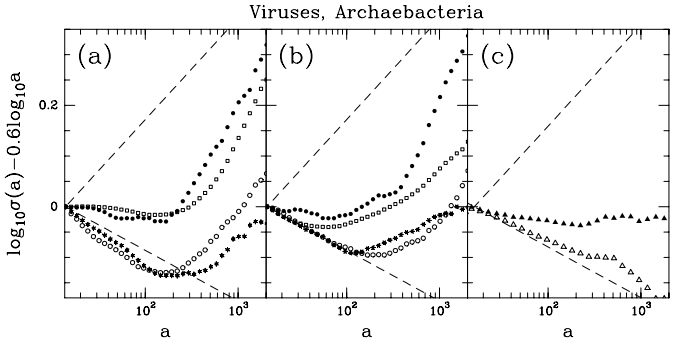

FIG. 7 – Estimation globale de l'écart type des coefficients de TO : (a) Marche ADN pour le codage "A" ; (b), (c) profils de courbure Pnuc. Les différents symboles correspondent aux génomes suivants : *Archaeglobus fulgidus* (carrés), virus *Epstein-Barr* (points), *Melanoplus sanguinipes entomapoxvirus* (cercles), *bacteriophage T4* (étoiles), moyenne sur 21 virus à ARN simple brin (triangles) et 17 retrovirus (triangles noirs). Même représentation que dans la Fig. 4.

échelle (Fig. 7 pour le bactériophage *T4* et Réf. [43]), comme nous l'avons déjà constaté pour leurs hôtes eubactériens. Cette analyse des propriétés d'invariance d'échelle à l'aide de la TO de génomes viraux et cellulaires appartenant aux trois règnes du vivant confirme que les CLP à petite échelle sont la signature de la structure nucléosomale [42, 43].

Afin de tester notre diagnostic nucléosomal basé sur l'existence de CLP à petite échelle, nous avons étudié dans quelle mesure certains dinucléotides particuliers, bien connus pour participer à la formation et au positionnement des nucléosomes [55] (par exemple les dinucléotides AA), présenteraient des CLP specifiquement associées aux génomes eucaryotes. Ceci peut être examiné en analysant différentes marches ADN générées en considérant (i) toutes les adénines, (ii) seulement les adénines qui font partie d'un dinucléotide AA et (iii) les aénines isolées qui ne font pas partie d'un dinucléotide AA. L'analyse du génome humain (Fig. 4(d)) et d'autres génomes eucaryotes (Table 1 dans la Réf. [43]) montre que les marches ADN pour le codage "A isolé" présentent un affaiblissement clair des CLP à petite échelle alors que les marches ADN pour le codage "AA" rendent compte en grande partie des CLP observées pour le codage "A". Ce résultat confirme que les CLP à petite échelle sont bien la signature de la structure nucléosomale. De plus cette observation est une nouvelle illustration que différents codages ne possèdent pas trivialement la même structure de corrélations [43].

5 Discussion

Différentes études ont établi la présence dans les séquences d'ADN génomique de motifs associés à des propriétés de courbure de la double hélice. En particulier l'utilisation de l'analyse de Fourier [3] et le calcul de fonctions de corrélations [4] ont mis en évidence une périodicité de 10.2 *pb* spécifique aux génomes eucaryotes, qui a été interprétée en relation avec la structure nucléosomale. Cependant il existe une différence fondamentale entre le diagnostic nucléosomal basé sur une recherche de périodicité et notre analyse basée sur les propriétés d'invariance d'échelle. Cette dernière suggère fortement que les mécanismes biologiques sous-jacents à la structuration nucléosomale sont des phénomènes multi-échelles qui impliquent l'ensemble de la gamme d'échelles $1 - 200$ *pb*. A cet égard, la périodicité (qui concerne les séquences présentant une affinité pour l'octamère d'histones significativement supérieure à la moyenne [56] et qui représente seulement 5% des séquences) et l'invariance d'échelle (qui concerne l'ensemble de l'ADN génomique dont l'affinité pour l'octamère d'histones est similaire à celle d'une séquence aléatoire et qui représente 95% des séquences [56]), ne doivent pas être considérées comme contradictoires mais plutôt comme complémentaires. Dans la référence [43], nous avons proposé une compréhension dynamique des CLP observées à petite échelle. En contraste avec le lien fort entre histones et ADN créé par une distribution de périodicité adéquate des sites de courbure, les CLP faciliteraient le possible repositionnement de l'octamère d'histones sur tout le génome. Si nous considérons le déplacement du nucléosome comme un processus de diffusion le long de la chaîne ADN et si nous supposons qu'une fois l'octamère d'histones lié à l'ADN, la distribution des sites de courbures a une influence directe sur ce processus de diffusion, alors les CLP sont susceptibles de favoriser la mobilité des nucléosomes le long de l'ADN. La nature persistante de l'organisation spatiale invariante d'échelle serait donc selectionnée pour favoriser la dynamique de formation du chapelet nucléosomal en permettant aux nucléosomes d'explorer, pendant un intervalle de temps donné, de plus grands segments d'ADN et donc un plus grand nombre de configurations possibles. En d'autres termes, le nucléosome dépenserait moins d'énergie pour un même déplacement. La persistance permettrait ainsi de mieux comprendre le fait que, pour la majorité des séquences, l'énergie libre de formation du nucléosome observée soit si modeste. Ainsi le nucléosome serait une structure dynamique qui favoriserait un compromis optimal entre les contraintes de compaction et d'accessibilité aux complexes protéiques mis en jeu dans les processus de transcription et de réplication.

L'interprétation des CLP observées à grande échelle pour les profils de courbure ainsi que pour les marches ADN (Fig. 4) demeure un problème ouvert. Comme cela est suggéré dans la référence [56], les signaux impliqués dans la formation du nucléosome pourraient agir de manière collective sur de grandes distances afin de permettre la compaction du chapelet nucléosomal (filament de 10 *nm*) dans une structure chromatinienne d'ordre supérieur (fibre de 30 *nm*) (Fig. 6) [1, 49, 57]. Etant donné que les sites de courbure sont des éléments clés pour la structure nucléosomale, une étude systématique des CLP à grande échelle (dans la gamme $200 - 5000$ *pb*) observées pour les profils de courbure eucaryotes devrait apporter un nouvel éclairage sur les mécanismes de compaction mis en jeu dans la formation de la structure hiérarchique 3D de la chromatine ainsi que sur sa dynamique. Un élément important de notre étude est que des CLP similaires sont aussi observées pour les profils de courbure des génomes eubactériens (Figs. 4(b) et 5(b)) et archaebactériens (Fig. 7) [43]. En fait, tous les chromosomes sont soumis à des processus de condensation-décondensation (en relation avec la réplication de l'ADN, l'expression des gènes, ...) qui pourraient impliquer l'existence de contraintes structurales et dynamiques communes pour l'ensemble de la vie cellulaire. La véritable compréhension des CLP

à grande échelle et leur interprétation en relation avec de telles contraintes sont le sujet d'un travail en cours d'élaboration.

Ce travail a été soutenu par le GIP GREG (project "Motifs dans les Séquences"), par le Ministère de l'Education Nationale, de l'Enseignement Supérieur, de la Recherche et de l'Insertion Professionnelle ACC-SV (project "Génétique et Environnement") ainsi que par l'Action BioInformatique (CNRS, 2000). BA remercie la Communauté Européenne pour son soutien par une bourse Marie Curie (contrat : HPMF-CT-2001-01321).

Références

[1] van Holde, K. (1989). *Chromatin*. Springer, New York.

[2] Wolffe, A. P. (1995). *Chromatin Structure and Function*. Academic Press, London.

[3] Widom, J. (1996). Short-range order in two eukaryotic genomes : Relation to chromosome structure. *J. Mol. Biol.* **259**, 579–588.

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

[5] Trifonov, E. N. (1998). 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A* **249**, 511–516.

[6] Li, W. (1992). Generating non trivial long-range correlations and $1/f$ spectra by replication and mutation. *Int. J. Bifurc. Chaos* **2**, 137–154.

[7] Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. & Stanley, H. E. (1992). Long-range correlations in nucleotide sequences. *Nature* **356**, 168–170.

[8] Borštnik, B., Pumpernik, D. & Lukman, D. (1993). Analysis of apparent $1/f^{\alpha}$ spectrum in DNA sequences. *Europhys. Lett.* **23**, 389–394.

[9] Voss, R. F. (1994). Long-range fractal correlations in DNA introns and exons. *Fractals* **2**, 1–6.

[10] Azbel', M. Y. (1995). Universality in a DNA statistical structure. *Phys. Rev. Lett.* **75**, 168–171.

[11] Herzel, H. & Grosse, I. (1995). Measuring correlations in symbol sequence. *Physica A* **216**, 518–542.

[12] Nee, S. (1992). Uncorrelated DNA walks. *Nature* **357**, 450–450.

[13] Chatzidimitriou-Dreismann, C. A. & Larhammar, D. (1993). Long-range correlations in DNA. *Nature* **361**, 212–213.

[14] Karlin, S. & Brendel, V. (1993). Patchiness and correlations in DNA sequences. *Science* **259**, 677–679.

[15] Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Ossadnik, S. M., Peng, C.-K. & Simons, M. (1993). Fractal landscapes in biological systems. *Fractals* **1**, 283–301.

[16] Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Simons, M. & Stanley, H. E. (1995). Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev. E* **52**, 2939–2950.

[17] Herzel, H., Ebeling, W. & Schmitt, A. O. (1994). Entropies of biosequences : The role of repeats. *Phys. Rev. E* **50**, 5061–5071.

[18] Li, W. (1997). The measure of compositional heterogeneity in DNA sequences is related to measures of complexity. *Complexity* **3**, 33–37.

[19] Viswanathan, G. M., Buldyrev, S. V., Havlin, S. & Stanley, H. E. (1998). Long-range correlation mesasures for quantifying patchiness : Deviations from uniform power-law scaling in genomic DNA. *Physica A* **249**, 581–586.

[20] Buldyrev, S. V., Goldberger, A. L., Havlin, S., Stanley, H. E., Stanley, M. H. R. & Simons, M. (1993). Fractal landscapes and molecular evolution : Modeling the myosin heavy chain gene family. *Biophys. J.* **65**, 2673–2679.

[21] Li, W., Marr, T. G. & Kaneko, K. (1994). Understanding long-range correlations in DNA sequences. *Physica D* **75**, 392–416.

[22] Herzel, H., Trifonov, E. N., Weiss, O. & Grosse, I. (1998). Interpreting correlations in biosequences. *Physica A* **249**, 449–459.

[23] Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Matsa, M. E., Peng, C.-K., Simons, M. & Stanley, H. E. (1995). Long-range correlation properties of coding and noncoding DNA sequences : GenBank analysis. *Phys. Rev. E* **51**, 5084–5091.

[24] Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Simons, M. & Stanley, H. E. (1993). Finite-size effects on long-range correlations : Implications for analysing DNA sequences. *Phys. Rev. E* **47**, 3730–3733.

[25] Berthelsen, C. L., Glazier, J. A. & Raghavachari, S. (1994). Effective multifractal spectrum of a random walk. *Phys. Rev. E* **49**, 1860–1864.

[26] Li, W. (1997). The study of correlation structure of DNA sequences : A critical review. *Comp. Chem.* **21**, 257–272.

[27] 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.

[28] 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.

[29] Gardiner, K. (1996). Base composition and gene distribution : critical patterns in mammalian genome organization. *Trends Genet.* **12**, 519–524.

[30] Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17.

[31] Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E. & Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49**, 1685–1689.

[32] Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

[33] Meyer, Y. & Roques, S., eds. (1993). *Progress in Wavelets Analysis and Applications*. Editions frontières, Gif-sur-Yvette.

[34] 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.

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

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

[37] Jaffard, S. (1989). Exposants de Hölder en des points donnés et coefficients en ondelettes. *C. R. Acad. Sci. Paris Sér. I* **308**, 79–81.

[38] Holschneider, M. & Tchamitchian, P. (1990). Régularité locale de la fonction non-différentiable de Riemann. *Les Ondelettes en 1989* (P. G. Lemarié, ed.), pp. 102–104. Springer, Berlin.

[39] Mallat, S. & Hwang, W. L. (1992). Singularity detection and processing with wavelets. *IEEE Trans. Info. Theory* **38**, 617–643.

[40] Muzy, J.-F., Bacry, E. & Arneodo, A. (1994). The multifractal formalism revisited with wavelets. *Int. J. Bifurc. Chaos* **4**, 245–302.

[41] Arneodo, A., Bacry, E. & Muzy, J.-F. (1995). The thermodynamics of fractals revisited with wavelets. *Physica A* **213**, 232–275.

[42] 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.

[43] 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.

[44] Goodsell, D. S. & Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. *Nucl. Acids Res.* **22**, 5497–5503.

[45] Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995). Trinucleotide models for DNA bending propensity : comparison of models based on DNase I digestion and nucleosome packaging data. *J. Biomol. Struct. Dynam.* **13**, 309–317.

[46] Audit, B., Bacry, E., Muzy, J.-F. & Arneodo, A. (2002). Wavelet-based estimators of scaling behavior. *IEEE Trans. Info. Theory* **48**, 2938–2954.

[47] Li, W. & Kaneko, K. (1992). Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* **17**, 655–660.

[48] Arneodo, A., d'Aubenton-Carafa, Y., Audit, B., Bacry, E., Muzy, J.-F. & Thermes, C. (1998). Nucleotide composition effects on the long-range correlations in human genes. *Eur. Phys. J. B* **1**, 259–263.

[49] 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.

[50] Murphy, L. D. & Zimmerman, S. B. (1997). Stabilization of compact spermidine nucleoids from *Escherichia coli* under crowded conditions : implications for in vivo nucleoid structure. *J. Struct. Biol.* **119**, 336–346.

[51] Reeve, J. N., Sandman, K. & Daniels, C. J. (1997). Archaeal histones, nucleosomes and transcription initiation. *Cell* **89**, 999–1002.

[52] Challberg, M. D. & Kelly, T. J. (1989). Animal virus DNA replication. *Annu. Rev. Biochem.* **58**, 671–717.

[53] Borca, M. V., Irusta, P. M., Kutish, G. F., Carillo, C., Afonso, C. L., Burrage, A. T., Neilan, J. G. & Rock, D. L. (1996). A structural DNA binding protein of african swine fever virus with similarity to bacterial histone-like proteins. *Arch. Virol.* **141**, 301–313.

[54] Stanfield-Oakley, S. A. & Griffith, J. D. (1996). Nucleosomal arrangement of HIV-1 DNA : maps generated from an integrated genome and an EBV-based episomal model. *J. Mol. Biol.* **256**, 503–516.

[55] Thaström, A., Lowary, P. T., Widlund, H. R., Cao, H., Kubista, M. & Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.* **288**, 213–219.

[56] Lowary, P. T. & Widom, J. (1997). Nucleosome packaging and nucleosome positioning of genomic DNA. *Proc. Natl. Acad. Sci. USA* **94**, 1183–1188.

[57] Polach, K. J. & Widom, J. (1995). Mechanism of protein access to specific DNA sequences in chromatin : A dynamic equilibrium model for gene regulation. *J. Mol. Biol.* **254**, 130–149.