Université de Liège
Faculté des sciences

Analyse de séquences ADN par la transformée en ondelettes : extraction d'informations structurelles, dynamiques et fonctionnelles

Samuel Nicolay

Dissertation originale présentée en vue
de l'obtention du grade académique de
docteur en sciences
Avril 2006

« Il faut imaginer Sisyphe heureux »
A. Camus, *Le mythe de Sisyphe*.

Remerciements

Je tiens à remercier Alain Arneodo pour m'avoir encadré durant ces années. Plus que disponible, tu as été omniprésent. Ce manuscrit est aussi une façon de t'exprimer ma gratitude. Ton ombre m'accompagnera encore longtemps.

Sans le soutien inconditionnel de Françoise Bastin, ce travail n'aurait pu aboutir. Je sais que m'appuyer n'a pas toujours été chose facile. J'espère pouvoir te rendre un jour la pareille.

Les résultats obtenus sont le fruit d'une collaboration étroite entre scientifiques de tous bords, et dieu sait qu'il n'est pas facile de collaborer avec Alain ! Merci à Claude Thermes et son équipe d'avoir relevé le défi.

Merci aussi à Stéphane Jaffard pour avoir plus d'une fois éclairé ma lanterne et d'avoir accepté de faire partie des membres de mon jury.

Je tiens à exprimer ma gratitude envers messieurs Jean-Pierre Antoine, Paul Gérard, Alex Grossmann et Jean Schmets pour me faire l'honneur de prendre part à mon jury.

Un grand merci à tous les (ex-)membres de l'équipe au sens large ! Vous êtes trop nombreux à énumérer (et puis j'en oublierais la moitié et en plus je suis pressé par le temps).

Enfin, un grand merci à la famille pour son soutien, en particulier à ma femme, Clara.

Table des matières

Notations

Dans ce mémoire, nous avons tenté de nous en tenir aux conventions les plus usitées. Nous les énumérons ici pour éviter toute confusion. Il ne s'agit en rien de définitions ; celles-ci seront données en temps opportun ou supposées connues.

Ensembles

$\varnothing$	l'ensembe vide.
$\mathbb{R}$	l'ensemble des réels.
$\mathbb{Q}$	l'ensemble des rationnels.
$\mathbb{Z}$	l'ensemble des entiers.
$\mathbb{N}$	l'ensemble des naturels.
$\mathbb{T}$	le tore unité $\mathbb{T} = \mathbb{R}/\mathbb{Z}$.
$\mathbb{R}_*^+$	l'ensemble des éléments de $\mathbb{R}$ strictement positifs.
$\mathbb{R}^n$	l'espace euclidien à n dimensions.
$\mathbb{N}_0$	l'ensemble des naturels non nuls.
$[a,b]$	l'ensemble des réels compris (non-strictement) entre a et b.
$]a,b[$	l'ensemble des réels strictement compris entre a et b.
$E1\backslash E_2$	l'ensemble des points de E_1 n'appartenant pas à E_2.
$\mathcal{K}$	la classe des compacts non-vides.
K^*	un ensemble invariant.
$[\,.\,]$	le support.

dom()	le domaine de définition.
$\overset{\circ}{.}$	l'intérieur.
$\bar{.}$	l'adhérence.
$\partial\,.$	la frontière.
$B_\delta(\,.\,)$	la boule fermée de rayon δ centrée en un point.

Espaces fonctionnels

C^j	l'ensemble des fonctions j fois continûment dérivables.
D	l'ensemble des fonctions de C^∞ à support compact.
L^p	l'ensembles des fonctions de puissance p intégrables.
$\mathcal{S}$	la classe de Schwartz.
$\mathcal{C}^s$	l'espace de Hölder d'exposant s.
$\dot{\mathcal{C}}^s$	l'espace de Hölder homogène d'exposant s.
$\mathcal{C}^{s,s'}(t)$	l'espace de 2-microlocal d'exposants s et s' relatif au point t.
$\mathcal{C}^{s,*}(t)$	$f \in \mathcal{C}^{s,*}(t)$ si $\lvert f(t+l) - P(l)\rvert \leqslant \lvert l\rvert^s$.
$\mathcal{C}^s(t)$	la version locale de $\mathcal{C}^{s,*}(t)$.
$\dot{H}^s$	l'espace de Sobolev homogène d'exposant s.
$B_q^{s,p}$	l'espace de Besov d'indices s, p et q.
$\dot{B}_q^{s,p}$	l'espace de Besov homogène d'indices s, p et q.
l^p	l'espace des suites de Lebesgue.

Mesures, distances et dimensions

$\mathcal{H}^s$	la mesure ou mesure extérieure de Hausdorff à s dimensions.
$\mathcal{L}^n$	la mesure de Lebesgue à n dimensions.
diam	le diamètre.
dist	la distance euclidienne.
$\text{dist}_\mathcal{H}$	la distance de Hausdorff.
dim	la dimension topologique.
$\dim_\mathcal{H}$	la dimension de Hausdorff.
$\dim_M$	la dimension de Minkowski ou de boîte.
$\dim_s$	la dimension de similarité.

Probabilités

P	la probabilité d'un évènement.
E	la moyenne.
var	la variance.
cov	la covariance.
$. \overset{d}{=} .$	l'égalité en distribution.
B	un mouvement brownien.

B_H	un mouvement brownien fractionnaire d'indice H.
Δ_H	le bruit gaussien associé à un mouvement brownien.
b_H	une marche binaire.

Applications

S	une similitude.
ψ	une ondelette.
φ	le père d'ondelette.
$\lfloor x \rfloor$	le plus grand entier inférieur ou égal à x.
$\lceil x \rceil$	le plus grand entier supérieur ou égal à x.
$[x]$	$\lfloor x \rfloor$ si $x - \lfloor x \rfloor \leqslant \lceil x \rceil - x$, $\lceil x \rceil$ sinon.
$\{x_n\}_{n \in \mathbb{N}_0}$	une suite x_n.
$\binom{x_1}{x_2}$	le coefficient binômial $x_1!/(x_2!(x_1 - x_2)!)$.
$\pi(m,n,N)$	le sous-mot du mot m, de taille N et commençant à la n-ième lettre de m.
$\chi_E(x)$	1 si $x \in E$, zéro sinon.
Γ	la fonction gamma.
H_n	le polynôme d'Hermitte de degré n.
erf	la fonction d'erreur.

Opérateurs et symboles

$D_x^n f$	la dérivée n-ième de f par rapport à x.
$\partial_x^n f$	la dérivée n-ième partielle de f par rapport à x.
$\bar{f}$	le complexe conjugué de f.
$\hat{f}$	la transformée de Fourier négative de f.
$. * .$	le produit de convolution.
$W_\psi f$	la transformée en ondelette de f en utilisant l'ondelette ψ.
$\overline{\lim}$.	la limite supérieure.
$\underline{\lim}$.	la limite inférieure.
$. \circ .$	la composition de fonctions.
$f \sim g$	$\lim f/g = 1$.
$f = o(g)$	$\lim f/g = 0$.
$f = \mathcal{O}(g)$	$\lim \lvert f \rvert / g \leqslant C$.
$f = \bar{\mathcal{O}}(g)$	$\underline{\lim} \log \lvert f \rvert / log \lvert g \rvert \geqslant 1$.
$f = \widetilde{\mathcal{O}}(g)$	$\lim \log \lvert f \rvert / log \lvert g \rvert = 1$.
I	l'opérateur identité.
Δ	le laplacien.
Δ_h^j	la différence d'ordre j.

Avant-propos

Les récents programmes de séquençage du génome humain, et plus généralement des génomes des eucaryotes supérieurs, ont révélé que seule une faible proportion de l'ADN code pour la synthèse de protéines. L'origine et le rôle de l'ADN non-codant ont dès lors constitué une problématique majeure de la génomique. Récemment, il a été montré que ces séquences sont porteuses d'information de nature structurelle et qu'elles présentent certaines signatures des mécanismes de régulation du processus de transcription. Le séquençage d'organismes procaryotes a quant à lui montré l'existence d'asymétries de composition nucléotidique, résultant à la fois des mécanismes sous-jacents à la transcription et à la réplication. Les études menant à de telles conclusions chez l'homme sont rares, voire même inexistantes en ce qui concerne le rôle de la réplication dans l'apparition de telles asymétries de composition.

L'application d'outils tels que la transformée en ondelettes et le formalisme multifractal à des signaux issus de codages nucléotidiques des séquences ADN a précédemment permis, entre autres choses, de mettre en évidence le rôle des séquences non-codantes dans la structure nucléosomale de l'ADN eucaryote et plus généralement dans les mécanismes de condensation et décondensation de la chromatine. Nous poursuivrons ici cette démarche en généralisant à grande échelle les études précédentes qui s'étaient principalement focalisées sur la caractérisation des propriétés de corrélations à longue portée[+] existant jusqu'á des distances de quelques dizaines de milliers de paires de base. Ainsi, l'analyse de codages structurels nous conduira à mettre en évidence l'existence de rythmes

+. Comme nous le verrons, la dénomination « corrélations à longue portée » est quelque peu abusive.

de basses fréquences, signature de l'existence de domaines structurels (boucles de fibres de chromatine) jouant potentiellement le rôle de domaines fonctionnels. En effet, nous montrerons que ces rythmes sont aussi présents dans les asymétries de composition observées dans le génome humain. Notre principal objectif est de montrer le rôle prépondérant des mécanismes de transcription et de réplication dans l'apparition d'asymétries de composition nucléotidiques chez les mammifères. Ce manuscrit est organisé en deux parties. La première présente les concepts mathématiques nécessaires à l'étude de signaux issus de codages nucléotidiques. La mise en évidence de rythmes de basses fréquences dans ces signaux, l'analyse des profils de biais de composition relatifs à l'homme, l'interprétation de ces résultats et la mise au point d'une méthodologie de prédiction des origines de réplication sont effectuées dans la seconde partie.

La première partie de ce manuscrit contient trois chapitres. Le premier traite de la dimension d'un ensemble : il passe en revue différentes notions de dimension, à savoir la dimension topologique, la dimension de Hausdorff, la dimension de Minkowski ou de boîte ainsi que la dimension de similitude et explore les différentes relations qui peuvent exister entre celles-ci. Les bases du formalisme multifractal pour les mesures, généralisant la notion même de dimension, sont aussi exposées. Dans ce contexte, le concept d'ensemble fractal est défini naturellement.

Le deuxième chapitre introduit la transformée en ondelettes et présente la manière dont celle-ci peut être utilisée pour effectuer soit une étude de type spectral, soit la détection des singularités d'un signal. Le spectre d'ondelettes, les lignes de maxima et les exposants de Hölder sont d'abord définis. Une méthodologie permettant de caractériser les singularités isolées dans un signal basée sur la transformée en ondelettes est ensuite présentée. Le formalisme multifractal, permettant l'étude statistique des propriétés d'invariance d'échelle de signaux hautement irréguliers, est alors abordé. Différentes méthodes de calcul du spectre des singularités sont proposées, en particulier la méthode des maxima du module de la transformée en ondelettes (MMTO), qui sera utilisée par la suite. Nous porterons aussi notre attention sur les fonctions du type $\sin {}^{1}\!/_{t}$, caractéristiques des fonctions présentant des singularités oscillantes, et sur les diverses mises en oeuvre envisageables pour les étudier.

Le troisième chapitre définit le mouvement brownien fractionnaire comme une généralisation du mouvement brownien traditionnel, en s'intéressant plus particulièrement aux propriétés de corrélations à longue portée du bruit gaussien associé. Un algorithme de synthèse exacte est présenté, ainsi qu'une méthode d'obtention de marches à incréments discrets présentant des corrélations à longue portée. Il est par ailleurs démontré que la suite obtenue en prenant les signes d'un bruit gaussien fractionnaire présente les mêmes

propriétés de corrélations à longue portée que le bruit gaussien original. Il n'en va pas de même pour la suite définie par les modules d'un bruit gaussien fractionnaire, cette suite pouvant même être dé-corrélée.

La dimension de Hausdorff, le formalisme multifractal et l'étude spectrale *via* la transformée en ondelettes ont été introduits afin d'être appliqués à des signaux issus de codages des séquences ADN. Les marches à incréments discrets nous permettront quant à elles de synthétiser des séquences ADN artificielles. L'originalité de cette première partie réside essentiellement dans les travaux de synthèse réalisés. De nouvelles considérations sur la détection de singularités oscillantes y sont présentées. Enfin, les propositions concernant les corrélations à longue portée des signes et des modules des incréments d'un mouvement brownien fractionnaire sont, à notre connaissance, originales.

La seconde partie de ce manuscrit présente l'essentiel des résultats obtenus lors de ce travail de thèse concernant l'étude des séquences ADN par l'intermédiaire de codages nucléotidiques. Les concepts théoriques et les méthodologies développées dans la première partie vont nous permettre de révéler l'existence de biais de composition nucléotidique dûs aux mécanismes de transcription et de réplication. Grâce à ces observations, nous serons à même de proposer un modèle pour la réplication dans les génomes des mammifères. Finalement, à partir de ce modèle, nous mettrons en oeuvre un algorithme de détection des origines de réplication dans le génome humain. Cet algorithme permettra de déterminer un nombre d'origines de réplication putatives environ cent fois supérieur au nombre d'origines expérimentalement connues.

Le premier chapitre de la seconde partie regroupe les notions de biologie moléculaire nécessaires à notre étude. Sont entre autres abordées les processus de transcription, de réplication et d'empaquetage de l'ADN.

Le deuxième chapitre formalise la notion de codage nucléotidique et présente les codages les plus usuels. Chez l'homme, nous montrons que les signaux issus de codages nucléotidiques présentent pour la plupart des rythmes de basses fréquences mettant en jeu des tailles caractéristiques de l'ordre de la centaine voire de quelques centaines de milliers de paires de base. Nous resituons ensuite cette étude par rapport aux résultats précédemment obtenus concernant l'existence de corrélations à longue portée dans ces signaux, jusqu'à des distances de l'ordre de quelques dizaines de milliers de paires de base. Finalement, nous proposons une méthode de synthèse de séquences nucléotidiques artificielles dont les signaux ADN associés possèdent les mêmes propriétés de corrélations à longue portée que celles observées dans les séquences naturelles.

Le troisième chapitre présente l'analyse multifractale du signal biais associé aux vingt-deux chromosomes asexués de l'homme à l'aide de la méthode MMTO. Une propriété caractéristique de ce signal bruité est ainsi mise en évidence, à savoir la présence de nombreux sauts dont le caractère ascendant ou descendant dépend de l'échelle envisagée. À petite échelle, une « composante » saut se superpose à une « composante » bruit gaussien corrélé à longue portée. Pour les échelles plus grandes, les sauts ascendants de grande amplitude dominent l'analyse multifractale et n'ont quasiment pas d'équivalents descendants.

Le quatrième chapitre met en relation les sauts observés dans le signal biais au chapitre précédent avec les mécanismes de transcription et de réplication. Les sauts ascendants et descendants symétriquement distribués selon leur amplitude observés à petite échelle sont induits par la transcription. Les sauts ascendants de grande amplitude observés à grande échelle, sans équivalents descendants, sont dûs à la réplication. Cette dissymétrie entre sauts ascendants et sauts descendants s'explique par la présence de motifs en forme de « toit d'usine » dans les profils de biais caractéristiques des mécanismes de réplication, puisqu'aussi présents dans les régions intergéniques.

Le dernier chapitre propose une modèle pour la réplication des mammifères permettant d'expliquer les profils en forme de toit d'usine omniprésents dans le signal biais. Ce modèle consiste à supposer que les origines de réplication sont fixes alors que les terminaisons sont aléatoirement réparties et uniformément distribuées entre deux origines voisines. Finalement, à partir des prédictions de ce modèle, nous proposons un algorithme de détection des origines de réplication basé sur une méthode de reconnaissance de forme dans la représentation espace-échelle associée à la transformée en ondelettes.

Les données relatives aux séquences nucléotidiques proviennent pour la plupart du site internet de l'UCSC[⁕]. Les logiciels utilisés pour effectuer les transformées en ondelettes sont LASTWAVE[⁎] version 1.7 et un logiciel que nous avons personnellement écrit[⁕]. Les mouvements browniens fractionnaires ont été générés avec une librairie que nous avons écrite*, utilisant elle-même la librairie FFTW*.

La distinction qui est faite entre la première partie du manuscrit, où les notions mathématiques sont présentées, et la deuxième partie où celles-ci sont utilisées en tant qu'outil physique constitue une des particularités de la présente thèse. Elle illustre les différences

⁕. http://genome.ucsc.edu
⁎. http://www.cmap.polytechnique.fr/~bacry/LastWave/
⁕. S.Nicolay@ulg.ac.be
*. http://www.ulg.ac.be/sectmath/fbm.tar.gz
*. http://www.fftw.org/

qui existent entre les sciences mathématiques, où toutes les précautions sont prises dans la formalisation d'une notion, et les sciences physiques pour lesquelles ces concepts doivent pouvoir être appliqués en pratique. Ainsi, les physiciens m'excuseront pour la grande prudence avec laquelle je définis les outils mathématiques dans la première partie, et les mathématiciens me pardonneront pour les amalgames et les raisonnements heuristiques qui sont faits dans la seconde partie.

Première partie

Fondements mathématiques

Chapitre 1

Fractales et notions de dimension

La dimension topologique peut révéler certaines limites lorsque l'objet étudié se révèle trop complexe ; typiquement, un tel objet sera appelé fractale. Il est dès lors utile d'introduire d'autres notions de dimension, pour permettre l'étude de ces fractales. Même si nous serons essentiellement intéressé dans la suite par la dimension de Hausdorff, qui nous permettra d'étudier les fonctions multifractales (*cf.* chapitre 2), nous allons ici présenter diverses approches permettant de définir une dimension relative à un ensemble, en nous attardant sur les relations existants entre elles. Nous terminerons ce chapitre par une présentation du formalisme multifractal pour les mesures, généralisant la notion même de dimension. Nous verrons dans la suite que les points communs entre le formalisme multifractal pour les mesures et le formalisme multifractal pour les fonctions sont nombreux.

Les théories exposées dans ce chapitre sont classiques. Pour ne pas alourdir le texte, nous ne le référencerons qu'en début de sous-section, voire de section.

1.1 Quelques définitions de la dimension

Jusqu'au début du XX^{e} siècle, la notion de dimension permettait de préciser le nombre de paramètres réels nécessaire pour décrire la position d'un point matériel. C'est ainsi d'ailleurs qu'était, en général, définie la dimension d'un système. Les inconsistances de cette approche ont été mises en évidence par deux célèbres découvertes de l'époque. Tout d'abord, la bijection de Cantor [91] entre les points d'une ligne et les points d'un plan qui mettait à mal l'idée qu'un plan est « plus riche » qu'une droite. Ensuite l'application continue de Peano [311] d'un intervalle sur un carré, qui contredit l'idée que la dimension représente le plus petit nombre de paramètres réels continus requis pour décrire la position d'un point de l'espace. La question naturelle résultant de ces découvertes était de savoir s'il est possible de trouver une correspondance entre les espaces euclidiens $\mathbb{R}^n$ et $\mathbb{R}^m$ lorsque n est différent de m, avec les propriétés combinées des constructions de Cantor et Peano ; autrement dit, de savoir s'il existe un homéomorphisme[*] entre ces deux espaces euclidiens. La réponse (par la négative) donnée par Brouwer [80] à cette question laissa la porte ouverte à toute une série d'apports. La dimension pouvait enfin être définie de manière satisfaisante et d'autres notions de dimension allaient permettre de caractériser les « curiosités mathématiques » telles celles introduites par Peano, pour donner lieu à ce que l'on appelle aujourd'hui la géométrie fractale. Ces dimensions complémentaires, dites « fractales », peuvent prendre des valeurs non entières et sont définies à l'aide de recouvrements. En général, une dimension de grande portée théorique peut difficilement être évaluée en pratique. C'est la raison pour laquelle il n'existe pas une dimension fractale unique.

Dimension topologique

La notion de dimension (ou dimension topologique) a été rigoureusement définie au début du siècle dernier avec les travaux de Menger [277] et Urysohn [377, 379], fondés sur ceux de Brouwer [80] et les idées de Poincaré [318]. Il semble difficile de trouver une autre définition s'accordant aussi bien avec l'intuition et donnant une théorie si élégante.

Dans cette section nous ne considérerons que les espaces métrisables séparables. Des hypothèses plus générales peuvent être données, mais les résultats sont dès lors moins abondants. La définition, telle que donnée par Menger [277], de la dimension topologique adoptée ici est la suivante :

[*]. Nous entendons par là une application bijective, continue et d'inverse continu.

Définition 1.1 La *dimension topologique* $\dim(X)$ d'un espace métrisable séparable X est un entier supérieur ou égal à -1 qui se définit par induction :

- l'ensemble vide $\varnothing$ est le seul ensemble à être de dimension topologique égale à -1,
- X est de dimension $\leqslant n$ ($n \geqslant 0$) au point x s'il existe des voisinages arbitrairement petits de x dont les frontières sont de dimension $\leqslant n-1$,
- X est de dimension $\leqslant n$ si X est de dimension $\leqslant n$ en chacun de ses points,
- X est de dimension n au point x si X est de dimension $\leqslant n$ en x et n'est pas de dimension $\leqslant n-1$ en ce même point,
- X est de dimension n si X est de dimension $\leqslant n$ et n'est pas de dimension $\leqslant n-1$,
- X est de dimension infinie si X n'est pas de dimension $\leqslant n$ quel que soit n.

Si $\dim(X) = n$, avec n fini, alors X contient un sous-ensemble de dimension m, pour tout $m \leqslant n$. De fait, comme $\dim(X) > n-1$, il existe un point x de X et un voisinage de ce point pour lequel tout ouvert Ω inclus dans ce voisinage et contenant x est tel que $\dim(\partial\Omega) \geqslant n-1$. De plus, puisque $\dim(X) \leqslant n$, il existe un ensemble ouvert Ω_0 inclus dans le même voisinage, contenant lui aussi x et tel que $\dim(\partial\Omega_0) \leqslant n-1$. Ainsi $\partial\Omega_0$ est un ensemble de X de dimension $n-1$. La conclusion s'ensuit. De la même manière, on peut démontrer la proposition suivante.

Proposition 1.2 *Un sous-espace d'un espace de dimension $\leqslant n$ est de dimension $\leqslant n$.*

On peut aussi obtenir le résultat classique suivant.

Proposition 1.3 *La dimension topologique de l'espace euclidien $\mathbb{R}^n$ est n.*

La définition générale de la dimension porte sur les espaces métrisables séparables. Elle peut être étendue, mais revêt alors des propriétés rendant difficile le développement d'une véritable théorie de la dimension.

Remarque 1.4 Pour les espaces n'ayant pas la propriété d'être métrisable et séparable, la dimension comme définie en 1.1 peut être non nulle pour des espaces dénombrables. En effet, il existe un exemple, dû à Urysohn [378], d'espace de Hausdorff[❋,⋇] dénombrable et connexe, alors qu'un espace de dimension égale à zéro contient nécessairement des ensembles arbitrairement petits à la fois ouverts et fermés. Un espace de dimension égale à zéro est donc nécessairement non connexe. Les définitions alternatives de la dimension,

❋. Cet espace possède une base dénombrable mais n'est pas métrisable.

⋇. Un espace de Hausdorff est un espace topologique vérifiant l'axiome T_2 : si deux points sont distincts alors ils ont des voisinages distincts [14].

cessant d'être équivalentes si l'espace considéré n'est pas métrisable et séparable, possèdent d'autres propriétés aussi peu enviables [193]. □

Nous ne nous attarderons pas plus longuement sur les propriétés de la dimension topologique, ces dernières étant celles que lui accorde l'intuition. Le lecteur intéressé pourra notamment consulter les références suivantes [193, 277, 377, 379].

Mesure de Hausdorff

La mesure de Hausdorff est l'étape obligée pour accéder à la dimension de Hausdorff, qui nous permettra de proposer une notion complémentaire à la dimension topologique [59, 60, 92, 140, 180, 217]. Cette mesure a un intérêt propre. Elle permet notamment de comparer des ensembles de même dimension et est équivalente à la mesure de Lebesgue pour les dimensions entières.

Par souci de simplicité, nous nous limiterons aux espaces euclidiens, même si la plupart des considérations présentées ici peuvent être transposées aux espaces métrisables séparables. Pour un sous-ensemble X de $\mathbb{R}^n$ et $\varepsilon > 0$, on définit la quantité $\mathcal{H}^h_\varepsilon(X)$ comme une mesure du recouvrement optimum de X lorsque les éléments du recouvrement sont subordonnés à ε :

$$\mathcal{H}^h_\varepsilon(X) = \inf\Big\{ \sum_{i=1}^{+\infty} \mathrm{diam}(X_i)^h : X \subset \bigcup_{i=1}^{+\infty} X_i, \mathrm{diam}(X_i) \leqslant \varepsilon \Big\}, \tag{1.1}$$

Clairement, $\mathcal{H}^h_\varepsilon$ est une mesure extérieure sur $\mathbb{R}^n$. La *mesure extérieure de Hausdorff* de dimension h de X, $\mathcal{H}^h(X)$ peut alors être définie en faisant tendre ε vers 0,

$$\mathcal{H}^h(X) = \sup_{\varepsilon > 0} \mathcal{H}^h_\varepsilon(X). \tag{1.2}$$

On montre aisément que $\mathcal{H}^h$ est une mesure extérieure métrique[❋]. On peut alors introduire la mesure de Hausdorff par une technique classique.

Définition 1.5 La restriction de $\mathcal{H}^h$ à la σ-algèbre des ensembles $\mathcal{H}^h$-mesurables[∗] définit une mesure appelée *mesure de Hausdorff* de dimension h.

Comme tout ensemble est inclus dans un ensemble convexe de même diamètre, une définition équivalente est obtenue si on impose aux ensembles recouvrant X d'être convexes. De même, les ensembles peuvent aussi être choisis ouverts ou fermés (la mesure extérieure

❋. Une mesure extérieure sur un espace métrisable est métrique si $\mu(X_1 \cup X_2) = \mu(X_1) + \mu(X_2)$ lorsque $\mathrm{dist}(X_1, X_2) > 0$.

∗. Algèbre qui inclut les ensembles Boréliens, puisque la mesure extérieure est métrique [142].

diffère mais la limite est la même) [113]. Remarquons qu'en remplaçant les recouvrements quelconques intervenant dans (1.1) par des boules, on obtient une mesure différente, parfois appelée *mesure de Hausdorff-Besicovitch* [58].

Remarque 1.6 De la même manière que l'on définit la mesure Hausdorff, en posant

$$\mathcal{B}^h_\varepsilon(X) = \inf\Big\{ \sum_{j=1}^{+\infty} \mathrm{diam}(B_j)^h : X \subset \bigcup_{i=j}^{+\infty} B_j, \mathrm{diam}(B_j) \leqslant \varepsilon \Big\}, \tag{1.3}$$

où les ensembles B_j sont des boules, on obtient une mesure, appelée mesure de Hausdorff-Besicovitch, en posant

$$\mathcal{B}^h(X) = \sup_{\varepsilon>0} \mathcal{B}^h_\varepsilon(X). \tag{1.4}$$

Avec une telle définition, il est évident que l'on a $\mathcal{H}^h_\varepsilon(X) \leqslant \mathcal{B}^h_\varepsilon(X)$, mais l'égalité n'est pas toujours vérifiée [58]. En incluant les ensembles X_i intervenant dans (1.1) dans des boules B_i de diamètre deux fois plus grand, $\mathrm{diam}B_i = 2\,\mathrm{diam}X_i$, on obtient $\sum_i \mathrm{diam}(B_i)^h = 2^h \sum_i \mathrm{diam}(X_i)^h$ et donc $\mathcal{B}^h_{2\varepsilon}(X) \leqslant 2^h \mathcal{H}^h_\varepsilon(X)$. Les relations suivantes sont donc vérifiées,

$$\mathcal{H}^h(X) \leqslant \mathcal{B}^h(X) \leqslant 2^h \mathcal{H}^h(X). \tag{1.5}$$

Les deux mesures sont dite équivalentes. □

Donnons quelques propriétés de la mesure de Hausdorff. La mesure extérieure de Hausdorff est régulière* [61].

Proposition 1.7 *Pour tout sous-ensemble X de $\mathbb{R}^n$ il existe un ensemble G, intersection dénombrable d'ensembles ouverts, contenant X et tel que $\mathcal{H}^h(G) = \mathcal{H}^h(X)$. En particulier, $\mathcal{H}^h$ est une mesure extérieure régulière.*

Il existe une relation entre la mesure de Lebesgue et la mesure de Hausdorff, précisée par la proposition suivante [140].

Proposition 1.8 *Pour tout sous-ensemble X de $\mathbb{R}^n$, l'égalité suivante est vérifiée,*

$$\mathcal{L}^n(X) = \frac{\pi^{n/2}}{2^n \Gamma(n/2)} \mathcal{H}^n(X). \tag{1.6}$$

La mesure de Hausdorff de dimension naturelle n'apporte donc rien par rapport à la mesure de Lebesgue.

*. Une mesure μ sur $\mathbb{R}^n$ est régulière si pour tout ensemble X, il existe un ensemble borélien $B \supset X$ tel que $\mu(B) = \mu(X)$.

L'auto-similarité joue un rôle très important en géométrie fractale ; aussi est-il important que les mesures utilisées présentent des propriétés d'invariance par dilatation. Étant donné un ensemble X de $\mathbb{R}^n$ et un nombre positif λ, on définit l'ensemble λX comme suit,

$$\lambda X = \{\lambda x : x \in X\}. \tag{1.7}$$

Si $\{X_j\}$ est un recouvrement de X subordonné à ε, $\{\lambda X_j\}$ est un recouvrement de λX subordonné à $\lambda\varepsilon$. Ainsi,

$$\mathcal{H}^h_{\lambda\varepsilon}(\lambda X) \leqslant \sum_{j=1}^{+\infty} \operatorname{diam}(\lambda X_j)^h \leqslant \lambda^h \mathcal{H}^h_\varepsilon(X),$$

et $\mathcal{H}^h(\lambda X) \leqslant \lambda^h \mathcal{H}^h(X)$. Avec le même raisonnement mais en remplaçant λ par $1/\lambda$ et X par λX, on obtient la relation opposée, ce qui donne l'égalité

$$\mathcal{H}^h(\lambda X) = \lambda^h \mathcal{H}^h(X). \tag{1.8}$$

Nous verrons qu'une application de X dans $\mathbb{R}^m$ est dite hölderienne d'exposant α s'il existe des constantes C et $\alpha > 0$ telles que, si $x \in X$,

$$|f(x+l) - f(x)| \leqslant C|l|^\alpha$$

pour tout l tel que $x + l \in X$. Nous supposerons toujours implicitement que l'exposant α est strictement positif. Si $\{X_j\}$ est un recouvrement de X subordonné à ε, alors, puisque $\operatorname{diam}\big(f(X \cap X_j)\big) \leqslant C\operatorname{diam}(X_j)^\alpha$, $\{f(X \cap X_j)\}$ est un recouvrement de $f(X)$ subordonné à $C\varepsilon^\alpha$. On obtient

$$\sum_{j=1}^{+\infty} \operatorname{diam}\big(f(X \cap X_j)\big)^{h/\alpha} \leqslant C^{h/\alpha} \sum_{j=1}^{+\infty} \operatorname{diam}(X_j)^h.$$

La limite pour ε tendant vers 0 donne l'égalité

$$\mathcal{H}^{h/\alpha}\big(f(X)\big) \leqslant C^{h/\alpha} \mathcal{H}^h(X). \tag{1.9}$$

Pour les applications *lipschitz*, *i.e.* hölderiennes d'exposant 1, la dernière inégalité devient $\mathcal{H}^s\big(f(X)\big) \leqslant C\mathcal{H}^s(X)$.

Dimension de Hausdorff

La dimension topologique, qui à un espace associe un nombre entier, peut sembler contrintuitive lorsque l'on considère certains objets mathématiques, comme par exemple la courbe de Peano ou l'ensemble de Cantor, c'est-à-dire pour les objets que l'on a coutume d'appeler fractales. L'idée d'introduire une autre définition de la dimension, complémentaire à

la notion de dimension topologique, peut donc paraître utile. La dimension de Hausdorff [59, 60, 92, 140, 180, 217] d'un ensemble peut prendre des valeurs non entières et présente toutes les propriétés naturelles que l'on est en droit d'attendre d'une dimension.

Pour tout ensemble X de $\mathbb{R}^n$, la mesure de Hausdorff $\mathcal{H}^h(X)$ est décroissante lorsque h varie de 0 à $+\infty$. De plus, si $0 \leqslant h < t$, la relation suivante est vérifiée :

$$\mathcal{H}^h_\varepsilon(X) \geqslant \frac{\mathcal{H}^t_\varepsilon(X)}{\varepsilon^{t-h}},$$

ce qui montre que $\mathcal{H}^t(X)$ strictement positif entraîne $\mathcal{H}^h(X) = \infty$. Il existe donc une valeur unique $\dim_{\mathcal{H}}(X)$ pour laquelle $\mathcal{H}^h(X) = \infty$ lorsque $h < \dim_{\mathcal{H}}(X)$ et $\mathcal{H}^h = 0$ lorsque $h > \dim_{\mathcal{H}}(X)$.

Définition 1.9 La *dimension de Hausdorff* $\dim_{\mathcal{H}}(X)$ d'un sous-ensemble X de $\mathbb{R}^n$ est définie par l'égalité suivante :

$$\dim_{\mathcal{H}}(X) = \sup\{h : \mathcal{H}^h(X) = \infty\}. \tag{1.10}$$

Avec cette définition, la dimension de Hausdorff de l'ensemble vide est $\dim_{\mathcal{H}}(\varnothing) - \infty$. Il existe des définitions alternatives dont la seule différence porte sur la dimension de l'ensemble vide.

Remarque 1.10 Pour les ensembles non vides, la définition de la dimension de Hausdorff par l'égalité (1.10) est équivalente à celle reposant sur l'égalité suivante,

$$\dim_{\mathcal{H}}(X) = \inf\{h : \mathcal{H}^h(X) = 0\}. \tag{1.11}$$

Pour l'ensemble vide toutefois, en utilisant la relation (1.11), on obtient $\dim_{\mathcal{H}}(\varnothing) = 0$. Il est aussi envisageable de poser, comme pour la dimension topologique, $\dim_{\mathcal{H}}(\varnothing) = -1$. Cette convention a pour avantages de différencier l'ensemble vide des ensembles discrets[+], tout en associant un nombre réel à un ensemble, quel qu'il soit. Toutefois, l'ensemble vide est en général un cas dégénéré qu'il suffit de considérer comme particulier pour unifier ces différentes définitions. □

Si $X \subset X'$, alors $\dim_{\mathcal{H}}(X) \leqslant \dim_{\mathcal{H}}(X')$, puisque $\mathcal{H}^h(X) \leqslant \mathcal{H}^h(X')$ pour tout h. Pour tout X de $\mathbb{R}^n$, on a $\dim_{\mathcal{H}}(X) \leqslant n$. De fait, le cube unité C de $\mathbb{R}^n$ pouvant se découper en j^n sous-cubes de côté de longueur $1/j$, en prenant j tel que $\varepsilon \geqslant \sqrt{n}/j$, on obtient $\mathcal{H}^n_\varepsilon(C) \leqslant j^n(\sqrt{n}/j)^n = n^{n/2}$ et $\mathcal{H}^n(C) < \infty$. Ainsi $\mathcal{H}^h(C) = 0$ pour tout $h > n$. Il en est de même pour $\mathcal{H}^h(\mathbb{R}^n)$ puisque cet espace peut s'exprimer comme une union dénombrable

+. Nous verrons que la dimension d'un ensemble discret est nulle.

de ces cubes. Pour tout ouvert Ω de $\mathbb{R}^n$, on a $\dim_{\mathcal{H}}(\Omega) = n$, puisque Ω contient une boule dont le volume à n dimensions est fini positif. Il existe aussi une relation entre la dimension topologique et la dimension de Hausdorff donnée par l'inégalité suivante [357].

Théorème 1.11 *Pour tout ensemble X de $\mathbb{R}^n$, on a*[*] $\dim_{\mathcal{H}}(X) \geqslant \dim(X)$.

Pour tout ensemble X de $\mathbb{R}^n$, on a donc les inégalités :

$$\dim(X) \leqslant \dim_{\mathcal{H}}(X) \leqslant n. \tag{1.12}$$

La dimension de Hausdorff est stable[⁎]. Soit $\{X_j\}$ une suite d'ensembles ; on constate immédiatement que $\dim_{\mathcal{H}}(\cup_j X_j) \geqslant \dim_{\mathcal{H}}(X_i)$ quel que soit i. Inversement, si $s > \dim_{\mathcal{H}}(X_j)$ pour tout j, $\mathcal{H}^s(X_j) = 0$ et $\mathcal{H}^s(\cup_j X_j) = 0$, ce qui montre que

$$\dim_{\mathcal{H}}(\cup_j X_j) = \sup_j\{\dim_{\mathcal{H}}(X_j)\}. \tag{1.13}$$

Pour un point x, $\mathcal{H}^0(x) = 1$ et $\dim_{\mathcal{H}}(x) = 0$. Ainsi, un espace dénombrable possède une dimension de Hausdorff nulle, puisqu'il peut s'écrire comme une union dénombrable de points. Il existe aussi une propriété concernant la connexité [61].

Proposition 1.12 *Un ensemble de $\mathbb{R}^n$ dont la dimension de Hausdorff est strictement inférieure à 1 est totalement discontinu*[*].

Enfin, donnons une propriété importante de la dimension de Hausdorff. Étant donné une application f hölderienne d'exposant $\alpha > 0$ de X dans $\mathbb{R}^m$, si $s > \dim_{\mathcal{H}}(X)$, alors, par la relation (1.9), $\mathcal{H}^{s/\alpha}\big(f(X)\big) = 0$ et

$$\dim_{\mathcal{H}}\big(f(X)\big) \leqslant \frac{1}{\alpha}\dim_{\mathcal{H}}(X), \tag{1.14}$$

On est donc amené au corollaire suivant.

Corollaire 1.13 *Soit f une application de $X \subset \mathbb{R}^n$ dans $\mathbb{R}^m$. Si f est lipschitz, alors*

$$\dim_{\mathcal{H}}\big(f(X)\big) \leqslant \dim_{\mathcal{H}}(X).$$

Si f est bi-lipschitz, c'est-à-dire s'il existe des constantes non nulles C_1 et C_2 telles que

$$C_1|x_1 - x_2| \leqslant |f(x_1) - f(x_2)| \leqslant C_2|x_1 - x_2|,$$

alors

$$\dim_{\mathcal{H}}\big(f(X)\big) = \dim_{\mathcal{H}}(X).$$

[*]. Ce résultat reste vrai si on considère la mesure de Hausdorff définie sur les espaces métrisables.
[⁎]. Rappelons que la dimension est stable si elle vérifie une relation du type (1.13).
[*]. Autrement dit, tous les sous-ensembles de plus d'un élément sont non connexes.

Ainsi, la dimension de Hausdorff est invariante par transformation bi-lipschitz. De même qu'en topologie on peut affirmer que deux espaces sont équivalents s'il existe un homéomorphisme entre eux, deux ensembles sont équivalents vis-à-vis de la dimension de Hausdorff s'il existe une application bi-lipschitz les faisant correspondre.

La *dimension de Hausdorff-Besicovitch* est identique à celle de Hausdorff.

Remarque 1.14 On définit la dimension de Hausdorff-Besicovitch de manière analogue à celle de Hausdorff,

$$\dim_{\mathcal{B}}(X) = \sup\{h : \mathcal{B}^h(X) = \infty\}.$$

La remarque 1.6 permet d'affirmer que l'on a $\dim_{\mathcal{B}}(X) = \dim_{\mathcal{H}}(X)$. □

Nous pouvons à présent tenter de donner une définition d'un ensemble fractal. La définition est une de celles données par MANDELBROT [261, 262] ; elle correspond à l'idée que l'on se fait d'un ensemble fractal : un ensemble pour lequel la dimension topologique ne semble pas totalement adaptée à sa description géométrique.

Définition 1.15 Un ensemble non vide de $\mathbb{R}^n$ est appelé *ensemble fractal* ou *fractale* si dans la relation du théorème 1.11, l'inégalité est stricte. Autrement dit un ensemble fractal est un ensemble dont la dimension de Hausdorff est strictement supérieure à la dimension topologique.

Terminons en donnant le célèbre exemple de l'*ensemble de Cantor*.

Exemple 1.16 L'ensemble de Cantor se définit par étapes. Si $C_0 = [0,1]$, alors l'ensemble C_j $(j > 0)$ est obtenu à partir de C_{j-1} en retirant à ce dernier les ensembles ouverts de diamètre $1/3$, milieu de chaque intervalle définissant C_{j-1}. On obtient $C_1 = [0,1/3] \cup [2/3,1]$, $C_2 = [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1]$ et ainsi de suite. L'ensemble C_j est donc constitué de 2^j intervalles de longueur $1/3^j$. L'ensemble de Cantor est l'ensemble parfait* et dense nulle part* $C = \cap_j C_j$. Alternativement, C peut être défini comme l'ensemble des nombres de l'intervalle unité dont une des représentations* en base 3 ne contient pas le chiffre 1.

La dimension de Hausdorff de l'ensemble de Cantor est $h = \log 2/\log 3$ et pour ce nombre, $\mathcal{H}^h(C) = 1$. De fait, C peut être recouvert par les intervalles de C_j et ainsi $\mathcal{H}^s_{3^{-j}} \leqslant 2^j 3^{3-hj} \leqslant 1$. La limite sur j nous fournit $\mathcal{H}^h(C) \leqslant 1$. Supposons maintenant que

∗. Un ensemble parfait est un ensemble égal à l'ensemble de ses points limite. Rappelons que pour un ensemble topologique X, un point $x \in X$ est un point limite de $X_0 \subset X$ s'il existe un filtre de X_0 convergeant vers x.

⁎. Un ensemble n'est dense nulle part si l'intérieur de son adhérence est vide.

+. Ainsi le nombre 1 pouvant s'écrire $0.222\cdots$ en base 3, il appartient à C, ce qui fait de C un ensemble compact.

$\mathcal{I}$ est une collection d'intervalles recouvrant C. On cherche à montrer que

$$\sum_{I\in\mathcal{I}} \mathrm{diam}(I)^h \geqslant 1. \tag{1.15}$$

Puisque C est compact, on peut supposer que les éléments I sont fermés et en nombre fini. On peut aussi supposer que ces éléments peuvent s'écrire sous la forme $K_1 \cup \Omega \cup K_2$, où K_1 et K_2 sont deux intervalles définissant un des ensembles C_j et n'ayant pas nécessairement le même diamètre, et Ω est l'intervalle ouvert des nombres compris entre K_1 et K_3. On constate immédiatement que le diamètre de K_1 et K_2 est inférieur à celui de Ω et ainsi

$$\mathrm{diam}(I)^h \geqslant \Big(\frac{3}{2}\big(\mathrm{diam}(K_1)+\mathrm{diam}(K_2)\big)\Big)^h \geqslant \mathrm{diam}^h(K_1)+\mathrm{diam}^h(K_2).$$

L'inégalité (1.15) est vérifiée si l'on remplace I par les intervalles K_1 et K_2 et on peut supposer n'avoir que des intervalles de diamètre égal à $1/3^j$ pour un j. Pour ces intervalles de C_j, la relation (1.15) est vérifiée.

Si l'on définit maintenant l'ensemble C_j ($j > 0$) en demandant au diamètre de l'ensemble enlevé d'égaler $1-2s$, $s < 1/2$, on obtient un ensemble du même type mais avec une dimension de Hausdorff égale à $h = \log 2/\log 1/s$. Pour cette valeur, on a toujours $\mathcal{H}^h(C) = 1$. L'argument est le même que précédemment, à ceci près que le diamètre de K_1 et K_2 est majoré par $\mathrm{diam}(\Omega)\, s/(1-2s)$.

Enfin, on peut généraliser ce type d'ensemble pour les dimensions supérieures. Dans le plan, l'ensemble de départ est le carré unité $C_0 = [0,1]^2$ et l'ensemble C_j est construit à partir de C_{j-1} en ne laissant, dans chaque carré définissant C_{j-1}, que les quatre carrés de côté s^jpartageant un angle avec C_0. La dimension de Hausdorff de C est $h = \log 4/\log 1/s$ et on peut obtenir une borne inférieure et supérieure pour $\mathcal{H}^h(C)$, mais le calcul exact de cette valeur nécessite une démarche plus technique [142]. □

De plus amples développements sont proposés dans [141].

Dimension de Minkowski

Bien que très utile en théorie, la pratique révèle certaines limites de la dimension de Hausdorff. En effet, pour un ensemble donné, il peut être difficile de trouver un recouvrement permettant d'identifier la borne inférieure dans la relation (1.1). Il est toujours plus simple d'imposer la forme des ensembles de recouvrement et leur diamètre puis d'optimiser leur répartition. C'est la démarche adoptée ici [140, 226]. En contrepartie, cette nouvelle dimension peut avoir des comportements indésirables sur certains ensembles, comportements qui ne lui permettent pas d'occulter la dimension de Hausdorff.

Soit X un ensemble borné non vide de $\mathbb{R}^n$. Définissons $N_\varepsilon(X)$ comme le plus petit nombre d'ensembles de diamètre inférieur ou égal à ε requis pour recouvrir X. Intuitivement, une mesure du volume de X peut être donnée par $N_\varepsilon(X)\varepsilon^n$. Pour que cette notion ait un sens, il faut faire tendre ε vers zéro, en espérant que la limite existe. On peut essayer d'assurer la convergence en modifiant l'exposant relatif à ε pour obtenir $N_\varepsilon(X)\varepsilon^s$, où $s \leqslant n$ est la valeur pour laquelle le produit converge, c'est-à-dire celle où l'on obtient un saut entre ∞ et 0. On peut donc parler de la dimension s de l'ensemble X en écrivant

$$s = \lim_{\varepsilon \to 0} \frac{\log N_\varepsilon(X)}{-\log \varepsilon}. \tag{1.16}$$

Malheureusement, la convergence n'est toujours pas assurée et l'égalité (1.16) peut ne pas avoir de sens. On est donc amené à utiliser les limites inférieures et supérieures dans les développements qui suivent pour obtenir une définition rigoureuse.

Avant d'introduire la définition proprement dite, il nous faut établir quelques remarques relatives au type d'ensembles utilisable pour recouvrir X afin que la définition soit la plus « flexible » possible. Soient les cubes de $\mathbb{R}^n$ de côté ε formant un réseau

$$[k_1\varepsilon,(k_1+1)\varepsilon] \times [k_2\varepsilon,(k_2+1)\varepsilon] \times \cdots \times [k_n\varepsilon,(k_n+1)\varepsilon],$$

où $k_1,k_2,\ldots,k_n$ sont des entiers. Désignons par $N'_\varepsilon(X)$ le nombre minimum de ces cubes nécessaire pour recouvrir X ; ainsi, $N_{\varepsilon\sqrt{n}}(X) \leqslant N'_\varepsilon(X)$. En prenant ε assez petit, on obtient $\varepsilon\sqrt{n} < 1$ et donc

$$\frac{\log N_{\varepsilon\sqrt{n}}(X)}{-\log(\varepsilon\sqrt{n})} \leqslant \frac{\log N'_\varepsilon(X)}{-\log\sqrt{n} - \log\varepsilon}.$$

Il suffit alors de prendre les limites inférieure et supérieure pour ε tendant vers zéro afin d'obtenir

$$\underline{\lim_{\varepsilon \to 0}} \frac{\log N_\varepsilon(X)}{-\log \varepsilon} \leqslant \underline{\lim_{\varepsilon \to 0}} \frac{\log N'_\varepsilon(X)}{-\log \varepsilon}, \tag{1.17}$$

et

$$\overline{\lim_{\varepsilon \to 0}} \frac{\log N_\varepsilon(X)}{-\log \varepsilon} \leqslant \overline{\lim_{\varepsilon \to 0}} \frac{\log N'_\varepsilon(X)}{-\log \varepsilon}. \tag{1.18}$$

Pour démontrer que ces deux dernières relations sont des égalités, il suffit de remarquer que tout ensemble de diamètre inférieur ou égal à ε peut être recouvert par 3^n cubes de côté ε, ce qui donne l'inégalité $N'_\varepsilon(X) \leqslant 3^n N_\varepsilon(X)$. À partir de là, en procédant comme précédemment, on peut obtenir des relations du type (1.17) et (1.18) mais avec les signes d'inégalité opposés.

De la même manière, dans l'égalité (1.16) (en prenant éventuellement les limites inférieure et supérieure), on peut prendre $N_\varepsilon(X)$ comme étant le plus petit nombre de cubes

de côté ε nécessaire pour recouvrir X. Il suffit de remarquer que tout ensemble de diamètre ε est inclus dans un cube de côté ε et d'appliquer la même démarche. On peut aussi définir $N_\varepsilon(X)$ comme étant le plus petit nombre de boules fermées[⁂] de rayon ε nécessaire pour recouvrir X.

Signalons enfin que le fait de prendre le plus grand nombre de boules disjointes de rayon ε et de centre appartenant à X conduit à la même notion. Désignons ce nombre par $N'_\varepsilon(X)$ et prenons $N'_\varepsilon(X)$ de ces boules B_1,B_2,…,$B_{N'_\varepsilon(X)}$. Si x est un point de X, soit x est à une distance inférieure à ε d'une des boules, soit la boule de centre x peut être ajoutée à la collection. Ainsi les boules de même centre mais de rayon 2ε recouvrent X, donnant $N_{4\varepsilon}(X) \leqslant N'_\varepsilon(X)$. Si B_1,B_2,…,$B_{N'_\varepsilon(X)}$ sont des boules disjointes de rayon ε de centre appartenant à X, désignons par E_1,E_2,…,E_k une collection d'ensembles de diamètre inférieur ou égal à ε recouvrant X. Bien sûr, ces ensembles recouvrent les centres des boules disjointes considérées plus haut et ainsi chacune de ces boules doit contenir un ensemble E_j pour un certain j. Ainsi $k \geqslant N'_\varepsilon(X)$ et $N'_\varepsilon(X) \leqslant N_\varepsilon(X)$, ce qui permet de conclure.

On est donc amené à la définition suivante.

Définition 1.17 Les *dimensions de Minkowski inférieure* et *supérieure* d'un ensemble borné non vide X de $\mathbb{R}^n$ sont respectivement données par

$$\underline{\dim}_{\mathrm{M}}(X) = \underline{\lim}_{\varepsilon\to 0} \frac{\log N_\varepsilon(X)}{-\log\varepsilon}, \tag{1.19}$$

et

$$\overline{\dim}_{\mathrm{M}}(X) = \overline{\lim}_{\varepsilon\to 0} \frac{\log N_\varepsilon(X)}{-\log\varepsilon}. \tag{1.20}$$

Si ces limites sont égales, la *dimension de Minkowski*, encore appelée[⁎] *dimension de boîte* est donnée par l'égalité

$$\dim_{\mathrm{M}}(X) = \lim_{\varepsilon\to 0} \frac{\log N_\varepsilon(X)}{-\log\varepsilon}, \tag{1.21}$$

où $N_\varepsilon(X)$ est défini par une des assertions suivantes,

1. le plus petit nombre d'ensembles de diamètre inférieur ou égal à ε recouvrant X,
2. le plus petit nombre de boules fermées de rayon ε recouvrant X,
3. le plus petit nombre de cubes de côté ε recouvrant X,
4. le nombre de cubes d'un réseau de cubes de côté ε d'intersection non vide avec X,
5. le plus grand nombre de boules disjointes de rayon ε et de centre appartenant à X.

⁂. La boule fermée de rayon ε centrée en x_0 est l'ensemble $\{x : \mathrm{dist}(x_0,x) \leqslant \varepsilon\}$.
⁎. Aussi appelée *capacité* par les physiciens.

Les dimensions de Minkowski inférieure et supérieure sont monotones. La dimension supérieure est *finiment stable*, *i.e.* $\overline{\dim}_{\mathrm{M}}(X_1 \cup X_2) = \max\{\overline{\dim}_{\mathrm{M}}(X_1), \overline{\dim}_{\mathrm{M}}(X_2)\}$, mais pas la dimension inférieure. Ces dimensions sont aussi lipschitz-invariantes. Le raisonnement est le même que pour la dimension de Hausdorff. Si X peut être recouvert de $N_\varepsilon(X)$ ensembles de diamètre au plus ε, l'image de X par une application lipschitz peut être recouverte par le même nombre d'ensembles de diamètre au plus $C\varepsilon$.

Le principal désavantage de la dimension de Minkowski est qu'elle ne peut être présentée comme une mesure.

Remarque 1.18 La quantité

$$M^h(X) = \varliminf_{\varepsilon \to 0} N_\varepsilon(X)\, \varepsilon^h$$

ne permet pas de définir une mesure. Il suffit de montrer que M^h n'est pas dénombrablement stable. Ce résultat est établi par l'exemple 1.21. □

Le calcul de la dimension de Minkowski est plus direct que celle de Hausdorff.

Exemple 1.19 La dimension de Minkowski de l'ensemble de Cantor est égale à sa dimension de Hausdorff, $h = \log 2 / \log 3$. Pour la dimension supérieure, il suffit de remarquer que l'ensemble C_j possède 2^j intervalles de diamètre $1/3^j$ et $N_\varepsilon(C) \leqslant 2^j$ si $1/3^j < \varepsilon \leqslant 1/3^{j+1}$. Pour la dimension inférieure, tout intervalle de diamètre ε, $1/3^{j-1} \leqslant \varepsilon < 1/3^j$, intersecte au plus un intervalle de C_j et $N_\varepsilon(C) \geqslant 2^j$. □

Pour un ensemble X, posons $X_\varepsilon = \{x \in \mathbb{R}^n : \mathrm{dist}(x,X) \leqslant \varepsilon\}$. Si le volume de X_ε se comporte comme $\mathcal{L}^n(X_\varepsilon) \sim c\varepsilon^{n-h}$, on peut interpréter h comme étant la dimension de X. Cette méthode, attribuée à Minkowski [141], est étroitement reliée à la dimension de boîte.

Proposition 1.20 *Pour tout ensemble X de $\mathbb{R}^n$, on a*

$$\underline{\dim}_{\mathrm{M}}(X) = n - \varlimsup_{\varepsilon \to 0} \frac{\log \mathcal{L}^n(X_\varepsilon)}{\log \varepsilon},$$

et

$$\overline{\dim}_{\mathrm{M}}(X) = n - \varliminf_{\varepsilon \to 0} \frac{\log \mathcal{L}^n(X_\varepsilon)}{\log \varepsilon}.$$

Plus facile à calculer en pratique que la dimension de Hausdorff, les ensembles impliqués dans son calcul pouvant avoir le même diamètre, la dimension de Minkowski n'en reste pas moins un objet différent de la dimension de Hausdorff. Si X peut être recouvert de

$N_\varepsilon(X)$ ensembles de diamètre ε, on a $\mathcal{H}^h_\varepsilon(X) \leqslant N_\varepsilon(X)\varepsilon^h$. Si $\mathcal{H}^h(X) > 0$, on peut écrire $h \leqslant \varliminf_{\varepsilon\to 0} \log N_\varepsilon(X)/\log 1/\varepsilon$ et

$$\dim_{\mathcal{H}}(X) \leqslant \underline{\dim}_{\mathrm{M}}(X). \tag{1.22}$$

Même si ces dimensions sont identiques pour des ensembles suffisamment réguliers[❋], il existe nombre de cas où l'égalité n'est pas vérifiée, comme en atteste l'exemple 1.21.

La dimension de Minkowski peut avoir quelques fâcheuses propriétés, expliquant notamment pourquoi elle n'est pas utilisée en analyse multifractale. Si un ensemble X est recouvert par l'union finie de boules fermées de rayon ε, il en va de même pour son adhérence. Ainsi, le plus petit nombre de boules requis pour recouvrir X convient également pour $\bar{X}$, ce qui permet d'écrire

$$\underline{\dim}_{\mathrm{M}}(\bar{X}) = \underline{\dim}_{\mathrm{M}}(X),$$

et

$$\overline{\dim}_{\mathrm{M}}(\bar{X}) = \overline{\dim}_{\mathrm{M}}(X).$$

Ainsi, l'ensemble des nombres rationnels compris entre zéro et un a ainsi une de dimension de Minkowski égale à un. Les ensembles dénombrables peuvent donc avoir une dimension non nulle. Un second exemple est donné par l'ensemble des nombres de la forme $1/m$ uni au singleton zéro.

Exemple 1.21 La dimension de boîte de l'ensemble compact $E = \{0\} \cup \{1/m : m \in \mathbb{N}_0\}$ est $1/2$. De fait, définissons $E_j = \{1/m : 1 \leqslant m \leqslant j\}$ et, pour un $\varepsilon < 1/2$, soit j_ε tel que $1/\big(j_\varepsilon(j_\varepsilon+1)\big) \leqslant \varepsilon < 1/\big(j_\varepsilon(j_\varepsilon-1)\big)$. Un ensemble de diamètre ε recouvre au plus un des points de E_{j_ε} et il faut j_ε de ces ensembles pour recouvrir E_{j_ε}. Ainsi,

$$\frac{\log N_\varepsilon(E)}{-\log\varepsilon} \geqslant \frac{\log j_\varepsilon}{\log\big(j_\varepsilon(j_\varepsilon+1)\big)}.$$

Il suffit de $j_\varepsilon + 1$ intervalles de diamètre ε pour recouvrir $E_\varepsilon = [0, 1/j_\varepsilon]$. Pour ce même diamètre, $E \setminus E_\varepsilon$ peut être recouvert par $j_\varepsilon - 1$ intervalles et

$$\frac{\log N_\varepsilon(E)}{-\log\varepsilon} \leqslant \frac{\log 2j_\varepsilon}{\log\big((j_\varepsilon(j_\varepsilon-1)\big)}.$$

Si ε tend vers zéro, ces inégalités conduisent à $1/2 \leqslant \underline{\dim}_{\mathrm{M}}(E)$ et $\overline{\dim}_{\mathrm{M}}(E) \leqslant 1/2$. □

1.2 Ensembles auto-similaires

Les fractales auto-similaires (strictement [43, 116, 194, 312], ou statistiquement [195]) ont souvent été utilisées pour modéliser bon nombre de phénomènes physiques [7, 261,

❋. Comme les ensembles auto-similaires, voir la section 1.2.

262, 143, 154, 316]. C'est cette notion d'auto-similarité qui est étudiée ici. Elle fournit un moyen facile de générer des ensembles fractals jouissant de nombreuses propriétés. Cette section est basée sur l'approche systématique de HUTCHINSON [194], elle-même basée sur les travaux de MORAN [286].

Nous ne considérerons que les espaces euclidiens $\mathbb{R}^n$, même si la plupart des résultats peuvent être transposés aux espaces métrisables complets.

Ensembles invariants

Les systèmes de fonctions itérées permettent de générer des ensembles invariants uniques sous l'action d'une famille de contractions. L'application itérée de ces contractions sur tout ensemble compact converge, pour la métrique de Hausdorff, vers l'ensemble invariant.

Nous utiliserons la notation suivante.

Notation 1.22 La classe des compacts non vides de $\mathbb{R}^n$ est notée $\mathcal{K}$.

Rappelons d'abord ce qu'est la métrique de Hausdorff.

Définition 1.23 La *distance de Hausdorff* sur $\mathcal{K}$ est donnée par

$$\mathrm{dist}_{\mathcal{H}}(K,K') = \inf\{\varepsilon : K \subset K'_{\varepsilon},\, K' \subset K_{\varepsilon}\},$$

où, $X_{\varepsilon} = \{x \in \mathbb{R}^n : \mathrm{dist}(X,x) \leqslant \varepsilon\}$. Une manière équivalente de définir cette distance est la suivante,

$$\mathrm{dist}_{\mathcal{H}}(K,K') = \sup\{\mathrm{dist}(k,K'), \mathrm{dist}(K,k') : k \in K, k' \in K'\}.$$

Il est facile de voir que $\mathrm{dist}_{\mathcal{H}}$ définit une métrique sur $\mathcal{K}$. De plus $(\mathcal{K},\mathrm{dist}_{\mathcal{H}})$ est un espace métrisable complet [144].

Pour une application F définie sur $\mathbb{R}^n$, on définit la *constante de lipschitz*, ou *rapport de contraction* c, par

$$c = \sup_{x_1 \neq x_2} \frac{\mathrm{dist}\big(F(x_1),F(x_2)\big)}{\mathrm{dist}(x_1,x_2)}.$$

Une application est dite *contractante* si $c < 1$.

Notation 1.24 Sauf mention explicite du contraire, S_j $(j > 0)$ désignera une application contractante définie sur $\mathcal{K}$ de rapport c_j et $S = \{S_1, \ldots, S_m\}$ une famille de contractions.

On pose

$$S(X) = \bigcup_{j=1}^{m} S_j(X), \tag{1.23}$$

pour tout ensemble X de $\mathbb{R}^n$.

Le résultat fondamental sur lequel repose les développements qui suivent est qu'une famille de contractions, ou *système de fonctions itérées*, définit un ensemble compact non vide et unique, invariant par rapport à cette famille. L'application S est contractante pour la métrique de Hausdorff, puisque

$$\begin{aligned} \mathrm{dist}_{\mathcal{H}}(S(K_1),S(K_2)) &= \mathrm{dist}_{\mathcal{H}}\big(\cup_j S_j(K_1), \cup_j S_j(K_2)\big) \\ &\leqslant (\max_j c_j)\, \mathrm{dist}_{\mathcal{H}}(K_1,K_2). \end{aligned}$$

L'existence et l'unicité de l'ensemble invariant découlent alors du principe du point fixe.

Corollaire 1.25 *Étant donné une famille de contractions, il existe un ensemble compact non vide unique K^* satisfaisant*

$$K^* = S(K^*). \tag{1.24}$$

De plus, pour tout ensemble $K \in \mathcal{K}$,

$$S^p(K) \to K^*, \tag{1.25}$$

pour la métrique de Hausdorff.

Le résultat est en fait plus général.

Remarque 1.26 Le résultat précédent reste valable pour les espaces métrisables complets. L'application S est aussi contractante sur la collection des ensembles bornés fermés non vides. L'ensemble K^* est à la fois fermé, borné et compact. □

La notion d'invariance est alors claire :

Définition 1.27 L'ensemble K^* vérifiant l'égalité (1.24) est appelé *ensemble invariant* pour S.

Il est clair que l'ensemble de Cantor est un ensemble invariant.

Exemple 1.28 L'ensemble de Cantor est l'ensemble invariant pour les contractions $S_1(x) = x/3$ et $S_2(x) = 2/3 + x/3$. □

Notation 1.29 Dorénavant, K^* représentera toujours un ensemble invariant et nous écrirons $S_{j_1 j_2 \cdots j_p}(\,.\,)$ pour signifier $S_{j_1} \circ S_{j_2} \circ \cdots \circ S_{j_p}(\,.\,)$. De plus, $K^*_{j_1 \cdots j_p}$ désignera $S_{j_1 \cdots j_p}(K^*)$.

On peut quelque peu préciser la structure d'un ensemble invariant en fonction des contractions permettant de le définir. La définition 1.27 implique

$$K^* = \bigcup_{j=1}^{m} S_j(K^*) = \bigcup_{j_1,j_2} S_{j_1}(K^*_{j_2}) = \cdots = \bigcup_{j_1,j_2,\ldots,j_p} K^*_{j_1\cdots j_p}. \tag{1.26}$$

De plus,

$$K^*_{j_1\cdots j_p} = S_{j_1\cdots j_p}\Big(\bigcup_{j_{p+1}=1}^{m} K^*_{j_{p+1}}\Big) = \bigcup_{j_{p+1}=1}^{m} K^*_{j_1\cdots j_p j_{p+1}}, \tag{1.27}$$

donc

$$K^* \supset K^*_{j_1} \supset \cdots \supset K^*_{j_1\cdots j_p} \supset \cdots. \tag{1.28}$$

Notons aussi que

$$\operatorname{diam}(K^*_{j_1\cdots j_p}) \leqslant c_{j_1}\cdots c_{j_p}\operatorname{diam}(K^*). \tag{1.29}$$

Le membre de droite de la relation (1.29) tend vers zéro lorsque p croît vers l'infini. Ainsi, puisque l'espace est complet, $\cap_p K^*_{j_1\cdots j_p}$ est un singleton dont l'élément sera noté $K^*_{j_1\cdots j_p\cdots}$ et l'ensemble K^* est l'union de ces singletons. Finalement, si les unions de la relation (1.24) définissant l'ensemble invariant sont disjointes, alors il est totalement discontinu. Si deux points $K^*_{j_1\cdots j_p\cdots}$ et $K^*_{j_1\cdots j'_p\cdots}$ de K^* sont distincts, ils diffèrent pour un indice j_p. Ils sont donc respectivement inclus dans les ensembles $K^*_{j_1\cdots j_p}$ et $K^*_{j_1\cdots j'_p}$ d'intersection vide.

Le résultat suivant découle directement de l'inégalité triangulaire et des propriétés de la distance de Hausdorff. Il permet de voir si un ensemble est proche d'un ensemble invariant par rapport à une famille de contractions en donnant une borne supérieure à la distance entre ces deux ensembles.

Corollaire 1.30 *Soient $S_1,\ldots,S_m$ des contractions de $\mathbb{R}^n$. Pour tout sous-ensemble compact non vide K de $\mathbb{R}^n$, on a*

$$\operatorname{dist}_{\mathcal{H}}(S^k(K),K^*) \leqslant \frac{c^k}{1-c}\,\operatorname{dist}_{\mathcal{H}}\big(K,S(K)\big) \to 0, \tag{1.30}$$

pour la métrique de Hausdorff, où $c = \max_j\{c_j\}$.

En particulier, pour tout compact K,

$$\operatorname{dist}_{\mathcal{H}}(K,K^*) \leqslant \frac{1}{1-\max_j\{c_j\}}\,\operatorname{dist}_{\mathcal{H}}\big(K,S(K)\big). \tag{1.31}$$

Auto-similarités

Nous pouvons maintenant exhiber une classes d'ensembles fractals dont la dimension de Hausdorff est égale à la dimension de Minkowski. Cette dimension peut en outre être

calculée pratiquement et justifie, dans de nombreux cas, les méthodes heuristiques utilisées pour évaluer la dimension de Minkowski d'un ensemble.

Nous nous restreignons maintenant aux familles de similitudes de $\mathbb{R}^n$, *i.e.* aux applications F pour lesquelles il existe une constante non nulle c telle que

$$\operatorname{dist}\big(F(x_1),F(x_2)\big) = c\operatorname{dist}(x_1,x_2). \tag{1.32}$$

On montre aisément* qu'une telle égalité est vérifiée si et seulement si l'application est la composée d'une transformation orthonormale, d'une translation et d'une homothétie.

La dimension de similitude, dont le calcul est immédiat, permet dans bien des cas d'évaluer les dimensions de Hausdorff et Minkowski. Donnons une approche heuristique de cette dimension inspirée de celle définissant la dimension de Minkowski. Supposons que, à l'étape k, l'ensemble K^* soit « approximé » par les ensembles $K_j^{(k)}$, $1 \leqslant j \leqslant m$. La dimension de K^* peut être interprétée comme l'exposant s pour lequel la somme $\sum_j \operatorname{diam}(K_j^{(k)})^s$ converge avec k vers un nombre fini non nul, pouvant lui-même être interprété comme le volume de K^*. Ainsi, soit $\{S_1,\ldots,S_m\}$ une famille de similitudes. Une approximation en k étapes de K^* est donnée en partant d'un ensemble compact K et en calculant $S^k(K)$. Soit v_0 le volume de K. La dimension de similitude de K^* peut être définie comme l'exposant s pour lequel le volume s-dimensionnel de $S^{k+1}(K)$ défini par $v_{k+1} = \sum_j c_j^s v_k$ converge vers un nombre fini non nul. À la limite, on obtient $\sum_j c_j^s = 1$.

Définition 1.31 Soient $S_1,\ldots,S_m$ des similitude de $\mathbb{R}^n$. La quantité

$$\sum_{j=1}^{m} c_j^s \tag{1.33}$$

vaut m pour $s = 0$ et décroît vers zéro lorsque s croît vers l'infini ; la valeur de s pour laquelle cette somme vaut un est appelée *dimension de similitude* de K^* et est notée $\dim_{\mathrm{s}}(K^*)$.

Dans le cas particulier où les m similitudes ont le même rapport de contraction c, la dimension de similitude devient

$$\dim_{\mathrm{s}}(K^*) = \frac{\log 1/m}{\log c}. \tag{1.34}$$

Notation 1.32 Soit un ensemble invariant K^* défini par le contexte. Dans cette section, la dimension de Hausdorff d'un tel ensemble sera notée h, $h = \dim_{\mathcal{H}}(K^*)$ et sa dimension de similitude s, $s = \dim_{\mathrm{s}}(K^*)$.

∗. Il suffit de remarquer que $\big(F(\,\cdot\,) - F(0)\big)/c$ est une isométrie conservant le produit scalaire.

Les ensembles auto-similaires sont des ensembles invariants particulièrement intéressants, puisque omniprésents dans l'étude des fractales [42, 43, 142, 143, 261, 262].

Définition 1.33 L'ensemble K^* est *auto-similaire* par rapport à S si

- K^* est invariant pour S,
- $\mathcal{H}^h(K^*) > 0$ et $\mathcal{H}^h(K_j^* \cap K_k^*) = 0$ lorsque $j \neq k$.

Un ensemble auto-similaire est un ensemble invariant avec une condition de recouvrement minimum.

Sous certaines conditions, l'auto-similarité d'un ensemble invariant peut être caractérisée.

Proposition 1.34 *Soit $K^* \subset \mathbb{R}^n$ un ensemble invariant sous l'action de similitudes. On a $\mathcal{H}^s(K^*) < \infty$ et donc*

$$\dim_{\mathcal{H}}(K^*) \leqslant \dim_{\mathrm{s}}(K^*). \tag{1.35}$$

De plus, si l'on a $0 < \mathcal{H}^h(K^*) < \infty$, alors K^* est auto-similaire si et seulement si $\dim_{\mathcal{H}}(K^*) = \dim_{\mathrm{s}}(K^*)$.

On peut aussi obtenir des inégalités concernant la dimension de Minkowski.

Proposition 1.35 *Si K^* est un ensemble invariant sous l'action de similitudes, alors*

$$\dim_{\mathcal{H}}(K^*) \leqslant \underline{\dim}_{\mathrm{M}}(K^*) \leqslant \overline{\dim}_{\mathrm{M}}(K^*) \leqslant \dim_{\mathrm{s}}(K^*). \tag{1.36}$$

Preuve. Soient $S_1, \ldots, S_m$ des similitudes pour lesquelles K^* est invariant. On peut supposer que les rapports c_j $(1 \leqslant j \leqslant m)$ sont ordonnés : $c_1 \leqslant c_2 \leqslant \cdots \leqslant c_m$. Soit $r < 1$ et pour toute suite $\{j_k\}$ où les éléments de la suite sont des entiers compris entre 1 et m, choisissons n comme étant le plus petit entier pour lequel

$$c_1 r \leqslant c_{j_1} c_{j_2} \cdots c_{j_n} \leqslant r.$$

Enfin, désignons par N l'ensemble des suites (finies) obtenues de cette manière. On constate sans peine que $\sum_N (c_{j_1} c_{j_2} \cdots c_{j_n})^s \leqslant 1$ et ainsi, le nombre d'éléments de N ne peut excéder $(c_1 r)^{-s}$. De plus, on a

$$\operatorname{diam}(K^*_{j_1 j_2 \cdots j_n}) = c_{j_1} c_{j_2} \cdots c_{j_n} \operatorname{diam}(K^*) \leqslant r \operatorname{diam}(K^*).$$

Ainsi K^* peut être recouvert par au plus $(c_1 r)^{-s}$ boules de diamètre $r \operatorname{diam}(K^*)$. Par définition de la dimension de Minkowski, on a $\overline{\dim}_{\mathrm{M}}(K^*) \leqslant s$. □

Pour pouvoir améliorer la relation (1.36), il nous faut définir une notion supplémentaire permettant d'éviter des problèmes techniques.

Définition 1.36 Des similitudes $S_1,\ldots,S_m$ satisfont la *condition de l'ensemble ouvert* s'il existe un ensemble ouvert non vide Ω tel que

$$\Omega \supset \bigcup_{j=1}^{m} S_j(\Omega), \tag{1.37}$$

où l'union est disjointe.

On a alors un théorème reliant les différentes notions de dimension.

Théorème 1.37 *Si $S_1,\ldots,S_m$ sont des similitudes vérifiant la condition de l'ensemble ouvert, alors*

$$\dim_{\mathcal{H}}(K^*) = \dim_{\mathrm{M}}(K^*) = \dim_{\mathrm{s}}(K^*). \tag{1.38}$$

Pour ce nombre $\dim_{\mathcal{H}}(K^*)$, la mesure de Hausdorff est finie, $0 < \mathcal{H}^s(K^*) < +\infty$ et donc K^* est auto-similaire.

Exemple 1.38 L'ensemble de Cantor défini grâce à l'exemple 1.28 vérifie clairement la condition de l'ensemble ouvert. Il est donc auto-similaire et le calcul de sa dimension de similitude donne, grâce à l'égalité (1.34), $s = \log 2/\log 3$. □

Terminons en donnant une conséquence importante de la proposition 1.30 : tout ensemble compact peut être approché par des ensembles auto-similaires.

Corollaire 1.39 *Étant donnés un ensemble compact non vide K de $\mathbb{R}^n$ et $\varepsilon > 0$, il existe des similitudes contractantes $S_1,\ldots,S_m$ telles que*

$$\mathrm{dist}_{\mathcal{H}}(K,K^*) < \varepsilon.$$

Ce résultat donne lieu à une méthode de compression d'image, où l'idée est de coder non pas l'image elle-même, mais les similitudes donnant lieu à une approximation suffisamment bonne. Bien sûr cette technique est quelque peu plus évoluée que l'idée générale que nous en donnons ici. Le lecteur intéressé par l'application des systèmes de fonctions itérées pourra consulter les références [42, 43, 44, 312].

1.3 Formalisme multifractal

La théorie multifractale s'intéresse moins aux ensembles qu'aux mesures sur ces ensembles. Elle permet d'étudier la manière dont se répartissent les valeurs d'une mesure sur son support. C'est donc aussi cet ensemble support qui est étudié par l'intermédiaire de la mesure. Cette théorie a connu un grand essor il y a une vingtaine d'années grâce à des travaux relatifs aux systèmes dynamiques [54, 70, 106, 145, 167, 168, 176, 321] et à la turbulence [54, 309, 260, 276, 308] notamment. Les premiers exemples de mesure multifractale semblent être les cascades multiplicatives de Mandelbrot [260], pour modéliser la distribution d'énergie en turbulence pleinement développée. Pour une revue historique détaillée de cette théorie, nous renvoyons le lecteur aux références suivantes [7, 17, 132, 143, 178, 292, 312].

Spectre multifractal de grande déviation

Le spectre multifractal de grande déviation constitue une première approche intuitive et permet d'obtenir d'intéressants résultats théoriques. Le spectre quantifie l'importance de chaque valeur prise par la mesure.

Soit μ une mesure* définie sur les ensembles boréliens de $\mathbb{R}^n$, normée+. Pour éviter les problèmes de définition lorsque q est négatif, nous supposerons toujours que dans une somme du type $\sum_i \mu(E_i)^q$, seuls les ensembles E_i tels que $\mu(E_i) > 0$ sont pris en compte. Si $0 < \varepsilon < 1$, désignons par $\{C_i^{(\varepsilon)}\}$ les cubes d'un réseau de recouvrement de côté ε (dont la mesure est non nulle, $\mu(C_i^{(\varepsilon)}) > 0$) et par $N_\varepsilon(\alpha)$ le nombre de ces cubes dont la mesure est suffisamment grande, plus précisément

$$N_\varepsilon(\alpha) = \#\{i : \mu(C_i^{(\varepsilon)}) \geqslant \varepsilon^\alpha\}. \tag{1.39}$$

On définit également la *fonction de partition* en sommant sur tous les cubes du réseau de recouvrement de côté ε,

$$Z_\varepsilon(q) = \sum_{i\in\mathbb{N}} \mu(C_i^{(\varepsilon)})^q, \tag{1.40}$$

avec q réel. Bien sûr $Z_\varepsilon(0)$ donne le nombre de ces cubes nécessaire pour recouvrir le support de μ. Pour ε fixé, notons encore que $N_\varepsilon(\alpha)$ est croissant avec α, au contraire de $Z_\varepsilon(q)$ qui est décroissant lorsque q croît. Ainsi, en jouant sur cet exposant, on peut modifier les poids relatifs des valeurs $\mu(C_i^{(\varepsilon)})$ dans Z_ε.

*. Les mesures sont à valeur positives [105].
+. Une mesure μ est normée si elle est de support borné et si $\mu(\mathbb{R}^n) = 1$.

Le spectre $d(\alpha)$ que nous allons présenter joue le rôle d'une dimension. Donnons une approche heuristique avant de formuler les idées plus rigoureusement. Supposons que le nombre de cubes de réseau pour lesquels $\varepsilon^{\alpha+\delta} \leqslant \mu(C_i^{(\varepsilon)}) < \varepsilon^{\alpha}$ soit d'ordre $\varepsilon^{-d(\alpha)}$, pour des valeurs δ suffisamment petites,

$$Z_\varepsilon(q) \sim \int_{\mathbb{R}^+} \varepsilon^{q\alpha} \varepsilon^{-d(\alpha)} d\alpha. \tag{1.41}$$

La contribution dominante venant des valeurs de α pour lesquelles $q\alpha - d(\alpha)$ est le plus petit, quantité que l'on note $\tau(q)$, on obtient que

$$Z_\varepsilon(q) \sim \varepsilon^{\tau(q)}. \tag{1.42}$$

Bien sûr, rien ne nous assure que ces quantités sont bien définies. Pour simplifier les notations, nous omettrons dans la suite d'indiquer la longueur du côté des cubes de recouvrement $C_i^{(\varepsilon)}$ en écrivant simplement $C_i = C_i^{(\varepsilon)}$.

Pour pouvoir développer cette théorie plus en avant, nous devons supposer l'existence de la limite suivante.

Hypothèse de travail 1.40 Nous supposerons[⁂] que la double limite suivante existe[⁎] et la noterons $d(\alpha)$:

$$d(\alpha) = \lim_{\delta \to 0} \lim_{\varepsilon \to 0} \frac{\log\big(N_\varepsilon(\alpha+\delta) - N_\varepsilon(\alpha-\delta)\big)}{-\log \varepsilon}. \tag{1.43}$$

Cette limite[⁂] est de première importance en pratique, comme nous allons l'esquisser.

Définition 1.41 La courbe $d(\alpha)$ définie par l'égalité (1.43) est appelée le *spectre multifractal de grande déviation* de μ.

Étant donné η, pour des nombres δ et ε suffisamment petits, les inégalités suivantes devraient donc être vérifiées,

$$\varepsilon^{-d(\alpha)+\eta} \leqslant N_\varepsilon(\alpha+\delta) - N_\varepsilon(\alpha-\delta) \leqslant \varepsilon^{-d(\alpha)-\eta}. \tag{1.44}$$

L'approche heuristique (1.41) est bien formalisée par cette définition.

L'idée formulée par la relation (1.42) peut alors être écrite rigoureusement.

⁂. La démonstration de cette existence peut être un point délicat. Une méthode couramment utilisée fait appel au théorème de Chernoff [36, 100, 141].

⁎. Nous permettons à $d(\alpha)$ de ne pas être fini.

⁂. Si elle n'existe pas, il est possible d'utiliser les limites supérieures et inférieures pour obtenir diverses inégalités. Nous renvoyons à la référence [142] pour de plus amples développements.

Proposition 1.42 *En supposant que la quantité $N_\varepsilon(\alpha)$ introduite plus haut satisfait* (1.43) *et en définissant*

$$\tau(q) = \inf_{\alpha \geqslant 0} \{q\alpha - d(\alpha)\}, \tag{1.45}$$

l'égalité suivante est satisfaite,

$$\tau(q) = \lim_{\varepsilon \to 0} \frac{\log Z_\varepsilon(q)}{\log \varepsilon}. \tag{1.46}$$

En particulier, cette limite existe.

D'une manière générale, nous dirons que $\tau(q)$ existe si la limite (1.46) existe. L'intérêt de $\tau(q)$ est de permettre de calculer $d(\alpha)$ en inversant la relation (1.45), qui est une *transformée de Legendre*. Ainsi, la limite (1.46) permet, dans bien des cas, d'accéder au spectre multifractal $d(\alpha)$.

Terminons en donnant un exemple.

Exemple 1.43 Soient $0 < p < 1$ et $C = \cap_j C_j$ l'ensemble de Cantor. On associe à cet ensemble la mesure μ de la manière suivante. Chaque intervalle intervenant dans la construction de C à l'étape j va donner deux sous-intervalles à l'étape $j+1$. Sur chacun de ces sous-intervalles, on répartit la densité ρ de l'intervalle de départ avec les coefficients suivants : le premier intervalle se voit attribuer la densité $p\rho$ et le second la densité $(1-p)\rho$, traduisant l'importance relative d'un intervalle par rapport à l'autre. La masse de l'intervalle unité est posée égale à 1.

À l'étape j, pour chaque $i \leqslant j$, il y a $\binom{j}{i}$ intervalles de densité $p^i(1-p)^{j-i}$ et en appliquant la relation (1.40), $Z_{1/3^j} = (p^q + (1-p)^q)^j$. En posant $\varepsilon_j = 1/3^j$, on obtient

$$\tau(q) = \lim_{\varepsilon_j \to 0} \frac{Z_{\varepsilon_j}(q)}{\log \varepsilon_j} = \frac{\log\left(p^q + (1-p)^q\right)}{\log 1/3}, \tag{1.47}$$

en supposant que l'hypothèse de travail 1.40 est satisfaite.

Si $p = 1/2$, les densités sont équitablement réparties et la masse en un point x donné, *i.e.* la mesure de l'intervalle $[0,x]$, est la valeur de l'escalier du diable en ce même point x (voir l'exemple 2.45). À l'étape j, la masse de chaque intervalle constituant C_j est $\mu_j = \mu([0,1/3^j]) = 1/2^j$. Pour ce p, on trouve $\tau(q) = (q-1)\log 2/\log 3$. À partir de la fonction $\tau(q)$, la relation (1.45) nous permet de présumer que le spectre multifractal de grande déviation $d(\alpha)$ est égal à $-\infty$ pour tout α, excepté pour $\alpha = \log 2/\log 3$, valeur pour laquelle $d(\log 2/\log 3) = \log 2/\log 3$. Ainsi, une fonction $\tau(q)$ linéaire est associée à un spectre ne prenant une valeur finie qu'en un point. Un tel spectre est appelé monofractal (cf. relation (1.59)).

Excepté pour la valeur $p = 1/2$, la fonction $\tau(q)$ donnée par l'égalité (1.47) est non linéaire. L'ensemble des points pour lesquels la transformée de Legendre de τ ne prend pas la valeur $-\infty$ ne se réduit alors pas à un point. Nous dirons que la mesure associée à τ est *multifractale*. Remarquons que pour l'instant, rien ne nous assure que le « spectre » obtenu à partir de τ en inversant la relation (1.45) est le spectre défini par l'égalité (1.43). □

Calcul du spectre multifractal de grande déviation

Le développement des précédentes relations constitue la base de l'étude du spectre multifractal de grande déviation, en fournissant notamment une manière de calculer les valeurs $d(\alpha)$ par l'intermédiaire de $\tau(q)$.

Sous certaines hypothèses supplémentaires, nous pouvons préciser les relations entre d et τ.

Hypothèse de travail 1.44 Nous supposerons que d est une fonction dérivable de α, strictement positive et strictement concave.

Cette hypothèse peut poser problème, puisqu'il s'agit d'une hypothèse sur le comportement de d, spectre que l'on souhaite estimer ! Dans la pluspart des cas pratiques, cette hypothèse semble vérifiée ; nous verrons que c'est par exemple le cas pour les mesures auto-similaires.

Pour q donné, supposons que $\alpha_q > 0$ est la valeur (si elle existe) de α pour laquelle l'infimum intervenant dans la définition (1.45) de τ est réalisé. En ce point,

$$\big[D_\alpha\big(q\alpha - d(\alpha)\big)\big]_{\alpha=\alpha_q} = 0, \tag{1.48}$$

et donc

$$q = D_\alpha d(\alpha_q). \tag{1.49}$$

Par définition de α_q,

$$\tau(q) = q\alpha_q - d(\alpha_q) \tag{1.50}$$

et, si α_q est dérivable par rapport à q,

$$D_q\tau(q) = \alpha_q + qD_q\alpha_q - D_\alpha d(\alpha_q)D_q\alpha_q. \tag{1.51}$$

On obtient alors, grâce à (1.49),

$$D_q\tau(q) = \alpha_q. \tag{1.52}$$

En général, τ peut être calculé ou estimé lorsque q varie ; α_q et $d(\alpha_q)$ peuvent en être déduits par les relations (1.52) et (1.50).

Si $[\mu]$ désigne le support de μ, on a $Z_\varepsilon(0) = N_\varepsilon([\mu])$, où $N_\varepsilon([\mu])$ désigne le nombre de cubes nécessaire pour recouvrir $[\mu]$ et ainsi, par (1.46) et (1.50),

$$-\tau(0) = d(\alpha_0) = \dim_{\mathrm{M}}([\mu]). \tag{1.53}$$

Par (1.49), cette valeur correspond au maximum de d. Pour $q = 1$, $Z_\varepsilon(1) = 1$, puisque la mesure est normée, et donc $\tau(1) = 0$. En ce qui concerne d, on trouve $d(\alpha_1) = \alpha_1$ et $D_\alpha d(\alpha_1) = 1$.

Par la définition (1.40) de $Z_\varepsilon(q)$,

$$D_q \log Z_\varepsilon(q) = \frac{\sum_i \mu(C_i)^q \log \mu(C_i)}{\sum_i \mu(C_i)^q}. \tag{1.54}$$

En $q = 1$, on obtient directement

$$\left[D_q \frac{\log Z_\varepsilon(q)}{\log \varepsilon} \right]_{q=1} = \frac{\sum_i \mu(C_i) \log \mu(C_i)}{\log \varepsilon}. \tag{1.55}$$

Les égalités (1.46) et (1.52) entraînent alors, si les dérivées convergent lorsque ε tend vers zéro,

$$\alpha_1 = D_q \tau(1) = \lim_{\varepsilon \to 0} \frac{\sum_i \mu(C_i) \log \mu(C_i)}{\log \varepsilon}. \tag{1.56}$$

Le numérateur revêt la forme d'une entropie.

Définition 1.45 L'expression $\sum_i \mu(C_i) \log \mu(C_i)$ est appelée l'*entropie* de la partition de μ et α_1 est appelé la *dimension d'information* de μ.

En fait, nous pouvons introduire un continuum de valeurs $\dim_q(\mu)$ pour caractériser la mesure μ mieux que ne le ferait une seule valeur. On pose

$$\dim_q(\mu) = \frac{1}{1-q} \lim_{\varepsilon \to 0} \frac{\log Z_\varepsilon(q)}{-\log \varepsilon} = \frac{\tau(q)}{q-1}, \tag{1.57}$$

pour autant que cette expression ait un sens (cf. proposition 1.42). On obtient immédiatement que $\dim_0(\mu) = \dim_{\mathrm{M}}([\mu])$ et par le théorème de l'Hospital, $\dim_1(\mu) = \alpha_1$ (si bien sûr, les hypothèses sont vérifiées).

Si la mesure est répartie de manière homogène, *i.e.* si $\mu(C_i) = 1/N_\varepsilon([\mu])$ pour $1 \leqslant i \leqslant N_\varepsilon([\mu])$, on trouve par (1.40),

$$\frac{\log Z_\varepsilon(q)}{-\log \varepsilon} = (1-q) \frac{\log N_\varepsilon([\mu])}{-\log \varepsilon}. \tag{1.58}$$

On en conclut que, dans ces conditions, $\tau(q)$ est défini si et seulement si la dimension de Minkowski du support de μ est définie. De plus, $\dim_q(\mu)$ ne dépend pas de q : $\dim_q(\mu) = \dim_{\mathrm{M}}([\mu])$ pour tout q. On trouve alors $\tau(q) = (q-1)\dim_{\mathrm{M}}([\mu])$. La relation (1.45) implique que

$$d(\alpha) = \begin{cases} \dim_{\mathrm{M}}([\mu]) & \text{lorsque } \alpha = \dim_{\mathrm{M}}([\mu]), \\ -\infty & \text{lorsque } \alpha \neq \dim_{\mathrm{M}}([\mu]). \end{cases} \tag{1.59}$$

On parle de *spectre monofractal*. Cette démarche montre que la méthodologie multifractale n'apporte rien si la mesure est homogène (au mieux n'obtient-on que la dimension de Minkowski).

Reprenons l'exemple de la mesure associée à l'ensemble de Cantor.

Exemple 1.46 En supposant que les hypothèses de travail sont vérifiées, on peut maintenant poursuivre le développement de l'exemple 1.43 et déterminer le spectre de grande déviation. Le calcul de α_q à partir de τ est direct ; la relation (1.52) donne

$$\alpha_q = \frac{p^q \log p + (1-p)^q \log(1-p)}{\left(p^q + (1-p)^q\right) \log 1/3}.$$

L'égalité (1.50) permet d'affirmer que

$$d(\alpha_q) = \frac{1}{\log 3}\left(\log\left(p^q + (1-p)^q\right) - q\,\frac{p^q \log p + (1-p)^q \log(1-p)}{p^q + (1-p)^q}\right).$$

Le support du spectre est, si $p \leqslant 1/2$, $[-\log(1-p)/\log 3, -\log p/\log 3]$. Comme illustré par la figure 1.46, lorsque $p \neq 1/2$, la représentation du spectre est en « forme de cloche ».

Pour le cas particulier où $p = 1/2$, on retrouve $\alpha_q = \log 2/\log 3$ et

$$d(\log 2/\log 3) = \log 2/\log 3.$$

Notons que la démonstration de l'existence du spectre est assez technique et fait appel au théorème de Chernoff pour les grandes déviations. Cet outil est souvent utilisé pour démontrer que les hypothèses de travail sont vérifiées. Pour de plus amples détails concernant ces mesures et le *formalisme multifractal*, nous renvoyons aux références [70, 106, 176, 276, 292, 321]. □

Spectre multifractal de Hausdorff

Le spectre multifractal de grande déviation décrit le comportement global d'une mesure μ à une échelle ε et ne donne que peu d'information quant au comportement de la mesure en un point. Il existe d'autres définitions de spectre qui, à partir des valeurs de la mesure en un point, tentent de mesurer la répartition globale de ces valeurs.

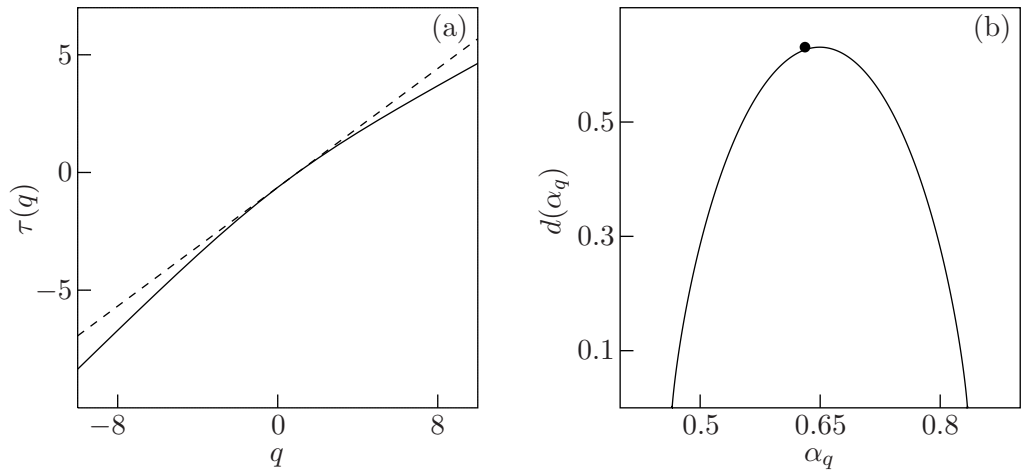

FIG. 1.1 – *En (a) sont représentées les fonctions τ pour la mesure associée à l'ensemble de Cantor, avec comme paramètre $p = 0.4$ (traits pleins) et $p = 1/2$ (pointillés). Pour $p = 1/2$, τ est linéaire et le spectre associé (représenté en (b)) se réduit en un point : le spectre est monofractal. En (b) est aussi représenté le spectre multifractal de la mesure lorsque $p = 0.4$.*

Commençons par introduire la notion de spectre multifractal dit de Hausdorff.

Définition 1.47 Étant donnée une mesure μ de $\mathbb{R}^n$ finie et régulière au sens de Borel, le *spectre multifractal de Hausdorff* est défini par l'égalité

$$d_{\mathcal{H}}(\alpha) = \dim_{\mathcal{H}}\{x \in \mathbb{R}^n : \underline{\lim}_{\varepsilon \to 0} \frac{\log \mu\big(B_\varepsilon(x)\big)}{\log \varepsilon} = \alpha\}, \tag{1.60}$$

où $B_\varepsilon(x)$ désigne la boule fermée de rayon ε centrée en x.

L'idée est donc de regarder le comportement de $\mu\big(B_\varepsilon(x)\big)$ lorsque ε tend vers 0, puis de regarder le comportement global des ensembles ainsi définis, alors que précédemment, on quantifiait les irrégularités globales de $\mu\big(B_\varepsilon(x)\big)$ avant de faire tendre ε vers 0. Dans ce qui suit, nous supposerons toujours que μ est une mesure finie et régulière.

D'après la définition, il est d'abord évident que l'on a

$$0 \leqslant d_{\mathcal{H}}(\alpha) \leqslant \dim_{\mathcal{H}}([\mu]), \tag{1.61}$$

mais il existe une relation plus forte [140],

$$0 \leqslant d_{\mathcal{H}}(\alpha) \leqslant \alpha, \tag{1.62}$$

pour tout α. En ce qui concerne la relation entre le spectre de Hausdorff et le spectre de grande déviation, il existe un résultat préliminaire.

Théorème 1.48 *Pour tout $\alpha \geqslant 0$,*

$$d_{\mathcal{H}}(\alpha) \leqslant d(\alpha) \tag{1.63}$$

si $d(\alpha)$ est positif et $d_{\mathcal{H}}(\alpha) = 0$ sinon.

Remarque 1.49 La limite sur ε peut ne pas exister dans la relation (1.43). On définit plutôt

$$\underline{d}(\alpha) = \lim_{\delta \to 0} \underline{\lim}_{\varepsilon \to 0} \max\{0, \frac{\log\big(N_\varepsilon(\alpha+\delta) - N_\varepsilon(\alpha-\delta)\big)}{-\log \varepsilon}\} \tag{1.64}$$

pour affirmer que

$$d_{\mathcal{H}}(\alpha) \leqslant \underline{d}(\alpha). \tag{1.65}$$

□

Il nous faut aussi introduire une quantité $\tau_{\mathcal{H}}$ analogue à celle définie par la relation (1.46). Cette définition est assez technique. Étant donné deux réels q et τ, on définit

$$\mathcal{H}_0^{q,\tau}(X) = \lim_{\varepsilon \to 0} \inf\{\sum_i \mu\big(B_{r_i}(x_i)\big)^q (2r_i)^\tau : X \subset \cup_i B_{r_i}(x_i), x_i \in X, r_i \leqslant \varepsilon\},$$

puis, pour que la mesure soit monotone,

$$\mathcal{H}^{q,\tau}(X) = \sup_{X' \subset X} \mathcal{H}_0^{q,\tau}(X').$$

Enfin, on pose

$$\tau_{\mathcal{H}}(q) = -\sup\{\tau : \mathcal{H}^{q,\tau}(\mathbb{R}^n) = +\infty\} = -\inf\{\tau : \mathcal{H}^{q,\tau}(\mathbb{R}^n) = 0\}. \tag{1.66}$$

On a alors le résultat suivant.

Proposition 1.50 *Si $\tau_{\mathcal{H}}$ est défini par l'égalité (1.66), l'inégalité*

$$d_{\mathcal{H}}(\alpha) \leqslant \inf_q \{q\alpha - \tau_{\mathcal{H}}(q)\} \tag{1.67}$$

est toujours vérifiée.

Pour les mesures ayant un comportement « suffisamment régulier », on peut espérer avoir l'égalité.

Malheureusement, il est parfois difficile d'évaluer $\tau_{\mathcal{H}}(q)$ en pratique. D'autres définition de $\tau_{\mathcal{H}}$ existent [81, 142, 323], souvent plus simples à manipuler, mais apportant moins de résultats théoriques.

Formalisme multifractal pour les mesures auto-similaires

Il est possible de montrer que le spectre multifractal de Hausdorff peut être obtenu à partir de τ pour certaines classes de mesures. C'est le cas des mesures dites auto-similaires, construites à partir de systèmes de fonctions itérées, et donc liées à la notion de fractale.

Supposons que le support de μ soit le compact $K^* = [\mu]$. Intuitivement, si cette mesure est auto-similaire, il devrait exister une famille de contractions $\{S_1, \ldots, S_m\}$ pour lesquelles les ensembles $S_j(K^*)$ soient disjoints, avec $S_j(K^*) \subset K^*$ et telle que $\cup_j S_j(K^*) = K^*$, donnant lieu à l'égalité $\mu(B) = \lambda_j \mu\big(S_j^{-1}(B)\big)$, pour tout sous-ensemble B de $S_j(K^*)$. Puisque ces ensembles sont disjoints, on peut sommer sur j dans la dernière égalité pour étendre la relation d'auto-similarité à K^*. Cette idée est précisée par la définition suivante.

Définition 1.51 Une mesure de probabilité satisfaisant l'égalité

$$\mu(B) = \sum_{j=1}^{m} \lambda_j \, \mu\big(S_j^{-1}(B)\big), \tag{1.68}$$

pour tout ensemble borélien B et telle que

$$S_i([\mu]) \cap S_j([\mu]) = \varnothing, \tag{1.69}$$

lorsque $i \neq j$ est appelée *mesure auto-similaire*.

Le résultat suivant est nécessaire pour prouver l'existence des mesures à étudier.

Théorème 1.52 *Soient $\{S_1, \ldots, S_m\}$ une famille de contractions sur un ensemble fermé F de $\mathbb{R}^n$ et $\{\lambda_1, \ldots, \lambda_m\}$ des réels positifs dont la somme vaut un. Il existe une mesure borélienne* unique telle que la relation (1.68) soit vérifiée, $\mu(F) = 1$ et*

$$\int_{\mathbb{R}^n} f(x)\, d\mu(x) = \sum_{j=1}^{m} \lambda_j \int_{\mathbb{R}^n} f\big(S_j(x)\big)\, d\mu(x), \tag{1.70}$$

pour toute fonction continue f de F vers $\mathbb{R}$.

De plus, le support de μ est l'ensemble invariant K^* pour la famille de contractions $\{S_j : 1 \leqslant j \leqslant m,\ \lambda_j \neq 0\}$ et si l'union $\cup_j S_j(K^*)$ sur toutes les contractions est disjointe, alors

$$\mu\big(S_{j_1 j_2 \cdots j_n}(K^*)\big) = \lambda_{j_1} \lambda_{j_2} \cdots \lambda_{j_n}. \tag{1.71}$$

∗. Une mesure telle que tout ensemble borélien soit μ-mesurable.

La condition d'intersection, parfois appelée *condition de séparation forte*, implique que le support de la mesure soit totalement discontinu et permet l'utilisation de la relation (1.71) [142].

Supposons avoir une mesure μ auto-similaire relative à une famille de contractions $\{S_1,\ldots,S_m\}$ et aux nombres λ_i $(1 \leqslant i \leqslant m)$ dont la somme vaut un. Étant donné un réel q, soit $\tau_{\mathcal{H}}(q)$ le nombre tel que

$$\sum_{i=1}^{m} \lambda_i^q {c_i}^{-\tau_{\mathcal{H}}(q)} = 1. \tag{1.72}$$

La fonction τ est analytique. En dérivant implicitement deux fois cette relation, on obtient

$$0 = \sum_{i=1}^{m} \lambda_i^q {c_i}^{-\tau_{\mathcal{H}}(q)} \left(D^2\tau_{\mathcal{H}}(q)\log c_i + (\log\lambda_i + D\tau_{\mathcal{H}}(q)\log c_i)^2\right)$$

et on constate que $\tau_{\mathcal{H}}$ est concave en q et même strictement concave si le rapport $\log\lambda_i/\log c_i$ n'est pas identique pour tous les i, ce que nous supposerons dorénavant pour éviter les cas dégénérés.

Si l'on définit $\alpha_{\min}$ et $\alpha_{\max}$ comme étant les pentes des asymptotes de $\tau_{\mathcal{H}}(q)$ et pose

$$\mathscr{L}(\alpha) = \inf_q\{q\alpha - \tau_{\mathcal{H}}(q)\}, \tag{1.73}$$

on constate que le support de $\mathscr{L}$ est $[\alpha_{\min},\alpha_{\max}]$. Pour un α donné, le minimum de $\mathscr{L}$ est atteint en un point q_α. En ce point

$$\alpha = D\tau_{\mathcal{H}}(q_\alpha), \tag{1.74}$$

et

$$\mathscr{L}(\alpha) = q_\alpha\alpha - \tau_{\mathcal{H}}(q_\alpha) = q_\alpha D\tau_{\mathcal{H}}(q_\alpha) - \tau_{\mathcal{H}}(q_\alpha). \tag{1.75}$$

En dérivant (1.72), on obtient

$$\alpha = \frac{\sum_i \lambda_i^q {c_i}^{-\tau_{\mathcal{H}}(q)}\log\lambda_i}{\sum_i \lambda_i^q {c_i}^{-\tau_{\mathcal{H}}(q)}\log c_i}. \tag{1.76}$$

De là, il peut être aisément montré que

$$\alpha_{\min} = \inf_{1\leqslant i\leqslant m}\frac{\log\lambda_i}{\log c_i}, \qquad \alpha_{\max} = \sup_{1\leqslant i\leqslant m}\frac{\log\lambda_i}{\log c_i} \tag{1.77}$$

En faisant appel à la géométrie de la transformée de Legendre, on constate que $\mathscr{L}$ est continu et que $\mathscr{L}(\alpha_{\min}) = \mathscr{L}(\alpha_{\max}) = 0$ si les quantités $\log\lambda_i/\log c_i$ sont toutes différentes*. Finalement, en dérivant (1.75), on obtient

$$D_\alpha\mathscr{L}(\alpha) = \alpha D_\alpha q_\alpha + q_\alpha - D_q\tau_{\mathcal{H}}(q_\alpha)D_\alpha q_\alpha = q_\alpha. \tag{1.78}$$

*. Dans ce cas, les asymptotes passent par l'origine, ce qui donne les deux égalités.

Puisque q décroît lorsque α croît, $\mathscr{L}$ est une fonction concave.

La théorie des auto-similarités nous permet d'affirmer, grâce à la relation (1.72), que $\tau_{\mathcal{H}}(0) = -\dim_{\mathcal{H}}([\mu]) = -\dim_{\mathrm{s}}([\mu])$. La même égalité implique que $\tau_{\mathcal{H}}(1) = 0$ et $\mathscr{L}(\alpha_1) = \alpha_1$, par la relation (1.75).

Le résultat suivant affirme que, pour les mesures auto-similaires, le spectre multifractal de Hausdorff peut être obtenu à partir de $\tau_{\mathcal{H}}$.

Théorème 1.53 *Soit μ une mesure auto-similaire. Si, avec les notations qui précèdent, α n'appartient pas à l'intervalle $[\alpha_{\min},\alpha_{\max}]$, alors l'ensemble*

$$\{x : \lim_{\varepsilon\to 0} \frac{\log \mu\big(B_\varepsilon(x)\big)}{\log \varepsilon} = \alpha\}$$

est vide. Sinon,

$$d_{\mathcal{H}}(\alpha) = \inf_q \{q\alpha - \tau_{\mathcal{H}}(q)\}. \tag{1.79}$$

Enfin, il existe des relations entre $d(\alpha)$ et $d_{\mathcal{H}}(\alpha)$.

Théorème 1.54 *Soit μ une mesure auto-similaire. Avec les mêmes notations que précédemment, en posant*

$$d^+(\alpha) = \max\{0,d(\alpha)\}, \tag{1.80}$$

l'inégalité suivante est satisfaite,

$$d^+(\alpha) \geqslant \inf_q \{q\alpha - \tau_{\mathcal{H}}(q)\}, \tag{1.81}$$

quel que soit α, avec l'égalité pour tout α correspondant à un q positif.

Autrement dit, en notant α_0 la valeur de α correspondant à $q = 0$, $d^+(\alpha) = d_{\mathcal{H}}(\alpha)$ si $\alpha \leqslant \alpha_0$.

On peut obtenir l'égalité dans la relation (1.81) pour toutes les valeurs de q, si l'on modifie légèrement la définition de N_ε, donnée par l'égalité (1.39).

Remarque 1.55 Dans la démonstration du théorème 1.54, si les cubes de recouvrement $\{C_i\}$ intervenant dans la définition de N_ε sont tels que $\mu(C_i') > 0$ pour tout i, où C_i' est le cube de recouvrement de même centre que C_i et de côté une demi-fois moins long, alors $d^+(\alpha) = d_{\mathcal{H}}(\alpha)$ pour tout α. □

La mesure associée à l'ensemble de Cantor est auto-similaire.

Exemple 1.56 La mesure associée à l'ensemble de Cantor peut être définie par la mesure

μ de support donné par [0,1] valant 1 sur cet intervalle et telle que

$$\mu([x_1,x_2]) = p\mu\big(S_1^{-1}([x_1,x_2])\big) + (1-p)\mu\big(S_2^{-1}([x_1,x_2])\big), \tag{1.82}$$

où $S_1(x) = x/3$ et $S_2(x) = 2/3 + x/3$ sont les contractions associées à l'ensemble de Cantor. Il s'agit clairement d'une mesure auto-similaire. La valeur $\tau_{\mathcal{H}}$ définie par la relation (1.72) vaut

$$\tau_{\mathcal{H}}(q) = \frac{\log\left(p^q + (1-p)^q\right)}{\log 1/3}$$

et les deux spectres sont égaux. □

Chapitre 2

Analyse et caractérisation de signaux irréguliers par la transformée en ondelettes continue

Les ondelettes constituent un outil devenu indispensable en analyse et traitement du signal permettant des formulations mathématiques des plus élégantes. Notre but ici est de présenter globalement la transformée en ondelettes continue dans l'espace des fonctions de carré intégrable avant de nous intéresser à la définition, la caractérisation et la détection des irrégularités au sens hölderien. Outre ces démarches, ce chapitre introduit également le formalisme multifractal pour les fonctions, utile pour caractériser globalement les fonctions présentant un grand nombre d'irrégularités. Le lecteur intéressé par des ouvrages généraux sur les ondelettes pourra consulter [13, 17, 101, 102, 112, 227, 257, 278, 305, 322, 365].

Mis à part quelques réflexions sur le formalisme multifractal, les théories exposées dans ce chapitre sont classiques ; nous ne référencerons ce texte qu'en début de sous-section, voire de section.

2.1 La transformée en ondelettes continue

La transformée en ondelettes continue, explicitement[*] introduite dans les travaux de Grossman et Morlet [173, 174] associe une représentation espace-échelle à une fonction, de la même manière que la transformée de Gabor lui associe une représentation epace-fréquence. La transformée en ondelettes permet une étude systématique des signaux [123, 257, 365] et a donné lieu à des avancées significatives tant en physique [17, 20, 25, 287, 292], qu'en mathématique [137, 203, 238, 278, 334].

Définitions

Nous présentons ici les bases de la théorie de la transformée en ondelettes en ne considérant que les fonctions mesurables de carré intégrable sur $\mathbb{R}$. Ses principales propriétés sont la conservation du module et l'inversibilité.

Notation 2.1 Nous travaillerons dans l'espace de Hilbert $L^2 = L^2(\mathbb{R})$ des fonctions mesurables de carré intégrable par rapport à la mesure de Lebesgue, équipé du produit scalaire naturel $\langle\,.\,,.\,\rangle$ et de la norme $\|\,.\,\|$ associée.

La transformée en ondelettes permet de représenter une fonction de multiples manières, dépendant essentiellement de l'ondelette choisie.

Définition 2.2 Une *ondelettes mère* est une fonction ψ de l'espace L^2 vérifiant la *condition d'admissibilité*

$$\int_{\mathbb{R}} \frac{|\hat{\psi}(\omega)|^2}{|\omega|}\, d\omega = \frac{C_\psi}{2\pi} < \infty. \tag{2.1}$$

Si ψ appartient à l'espace L^1, ce que nous supposerons implicitement, $\hat{\psi}$ est continu et pour que la condition d'admissibilité soit vérifiée, il faut que $\hat{\psi}(0) = \int_{\mathbb{R}} \psi(t)\, dt = 0$. Inversement, si cette dernière égalité est satisfaite et que $\int_{\mathbb{R}} |\psi(t)|(1+|t|^h)\, dt$ est fini[⁕] pour un $h > 0$, alors ψ vérifie la condition d'admissibilité (2.1). Ainsi dans les cas pratiques, où des propriétés de décroissance à l'infini sont exigées pour ψ, la condition d'admissibilité est satisfaite si $\hat{\psi}$ est nul à l'origine. La transformée en ondelettes d'une fonction est alors

[*]. Citons aussi les travaux de Calderòn [88]. Des bases d'ondelettes avaient déjà été implicitement utilisées dès le début du siècle passé [175, 352] sans être considérées comme telles.

[⁕]. Cette condition est légèrement plus contraignante que l'intégrabilité de ψ.

simplement définie par des convolutions avec l'ondelette mère translatée et dilatée. Par extension, cette dernière sera appelée *ondelette*.

Définition 2.3 La *transformée en ondelettes* d'une fonction f de l'espace L^2 par une ondelette ψ à l'*échelle* $a > 0$ et à la *position* (ou *temps*) $b \in \mathbb{R}$ est définie par

$$W_\psi f(b,a) = \langle f, \frac{1}{\sqrt{a}}\,\psi(\frac{\cdot - b}{a})\rangle = \frac{1}{\sqrt{a}}\int_{\mathbb{R}} f(t)\bar{\psi}(\frac{t-b}{a})\,dt. \tag{2.2}$$

On écrit parfois $Wf(b,a) = W_\psi f(b,a)$ lorsque le choix de l'ondelette est explicite ou sans conséquence.

La transformée en ondelettes peut être réécrite sous la forme d'un produit de convolution,

$$Wf(b,a) = \frac{1}{\sqrt{a}}[f * \bar{\psi}(-\frac{\cdot}{a})](b). \tag{2.3}$$

Cette transformée consiste donc à convoluer le signal avec une « fenêtre » qui se dilate avec le facteur échelle a. Le *demi-plan espace-échelle* est l'espace $\mathbb{R} \times \mathbb{R}_*^+$ et la *représentation espace-échelle* est la donnée du graphe $\{\big(b,a,Wf(b,a)\big) : (b,a) \in \mathbb{R} \times \mathbb{R}_*^+\}$. On peut aussi voir l'ondelette mère comme un filtre passe bande, puisque son intégrale est nulle. L'égalité (2.3) se réécrit, dans l'espace de Fourier,

$$\widehat{[Wf(\,.\,,a)]}(\omega) = \sqrt{a}\,\hat{f}(\omega)\,\bar{\hat{\psi}}(a\omega) \tag{2.4}$$

Comme nouas allons le voir, la transformée en ondelettes conserve la norme. De plus, dans une certaine mesure, on peut reconstruire une fonction à partir de sa transformée en ondelettes. Commençons par donner deux résultats préliminaires, pour lesquels on étend la définition de la transformée en ondelettes aux échelles a négatives en posant $Wf(b,a) = \langle f, |a|^{-1/2}\,\psi\big((\,\cdot\, - b)/a\big)\rangle$.

Théorème 2.4 *Si ψ vérifie la condition d'admissibilité, étant donné deux fonctions f_1 et f_2 de l'espace L^2, on a*

$$\iint_{\mathbb{R}^2} W_\psi f_1(b,a)\,\overline{W_\psi f_2}(b,a)\,db\frac{da}{a^2} = C_\psi\langle f_1, f_2\rangle. \tag{2.5}$$

En particulier, pour toute fonction f du même espace,

$$\|f\|^2 = \frac{1}{C_\psi}\int_{\mathbb{R}} \|Wf(\,.\,,a)\|^2\,\frac{da}{a^2}. \tag{2.6}$$

Grâce à cette égalité, on peut obtenir un résultat de convergence.

Corollaire 2.5 *Si ψ vérifie la condition d'admissibilité, pour toute fonction $f \in L^2$,*

$$\lim_{A_1,A_2\to\infty} \left\| f(t) - \frac{1}{C_\psi} \int\limits_{1/A_1 \leqslant |a| \leqslant A_2} \int_{\mathbb{R}} W_\psi f(b,a) \frac{1}{\sqrt{|a|}} \psi(\frac{t-b}{a})\,db\frac{da}{a^2} \right\| = 0. \tag{2.7}$$

Ces propositions prennent induisent la conservation de la norme et l'inversibilité, à ceci près que le paramètre d'échelle a varie sur $\mathbb{R}_*$ et non plus sur $\mathbb{R}_*^+$. À partir de ces relations, on peut établir des résultats équivalents pour la transformée en ondelettes sur $\mathbb{R} \times \mathbb{R}_*^+$, en imposant toutefois une condition d'admissibilité plus restrictive.

Définition 2.6 Une ondelette satisfait la *condition d'admissibilité restreinte* si

$$\int_{\mathbb{R}^-} \frac{|\hat{\psi}(-\omega)|^2}{\omega} d\omega = \int_{\mathbb{R}^+} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega = \frac{C_\psi}{2\pi} < +\infty. \tag{2.8}$$

Dans le cas d'une ondelette réelle, cette condition devient

$$\int_{\mathbb{R}^+} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega = \frac{C_\psi}{2\pi} < +\infty. \tag{2.9}$$

Grâce aux identités précédentes, le théorème 2.4 peut être adapté comme suit.

Corollaire 2.7 *Si ψ vérifie la condition d'admissibilité restreinte, étant donné deux fonctions f_1 et f_2 de l'espace L^2, on a*

$$\int_{\mathbb{R}^+} \int_{\mathbb{R}} W_\psi f_1(b,a) \overline{W_\psi f_2}(b,a)\, db \frac{da}{a^2} = C_\psi \langle f_1, f_2 \rangle. \tag{2.10}$$

En particulier, pour toute fonction f du même espace,

$$\|f\|^2 = \frac{1}{C_\psi} \int_{\mathbb{R}^+} \|Wf(\,.\,,a)\|^2 \frac{da}{a^2}. \tag{2.11}$$

On a aussi la convergence au sens faible.

Corollaire 2.8 *Si ψ vérifie la condition d'admissibilité restreinte, pour toute fonction $f \in L^2$, on a*

$$\lim_{A \to \infty} \left\| f(t) - \frac{1}{C_\psi} \int_{1/A}^{A} \int_{\mathbb{R}} W_\psi f(b,a) \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) db \frac{da}{a^2} \right\| = 0 \tag{2.12}$$

Noyau reproduisant

Les propriétés d'inversibilité et de conservation de la norme permettent de caractériser quelque peu l'espace relatif au demi-plan espace-échelle.

L'égalité (2.11) montre que W envoie les fonctions de l'espace $L^2(\mathbb{R})$ vers un espace de Hilbert, sous-espace de $L^2(\mathbb{R} \times \mathbb{R}_*^+; db\, da/a^2)$, que nous noterons H_W. Étant donné une

fonction f_W de H_W, en utilisant l'égalité (2.10), on trouve, si $f_W = W_\psi f$,

$$\begin{aligned} f_W(v,u) &= \langle f, \frac{1}{\sqrt{u}}\psi(\frac{\cdot - v}{u})\rangle \\ &= \frac{1}{C_\psi}\int_{\mathbb{R}^+}\int_{\mathbb{R}} K_\psi(b,a,v,u) f_W(b,a)\, db\frac{da}{a^2}, \end{aligned} \tag{2.13}$$

où on a posé

$$K_\psi(b,a,v,u) = \langle \frac{1}{\sqrt{a}}\psi(\frac{\cdot - b}{a}), \frac{1}{\sqrt{u}}\psi(\frac{\cdot - v}{u})\rangle. \tag{2.14}$$

Le *noyau reproduisant* K_ψ mesure en quelque sorte la corrélation existant entre les deux fonctions $\psi((\cdot - b)/a)$ et $\psi((\cdot - v)/u)$. Ainsi, il existe une certaine redondance dans la transformée en ondelettes continue.

Si l'on note P_ψ l'opérateur associé au noyau reproduisant K_ψ, alors P_ψ est auto-adjoint et H_W est défini par

$$H_W = \{f_W \in L^2(\mathbb{R}\times\mathbb{R}_*^+; db\, da/a^2) :\ P_\psi f_W = f_W\}. \tag{2.15}$$

Puisque H_W est associé par l'opérateur orthogonal P_ψ à un noyau reproduisant, il s'agit d'un sous-espace propre de $L^2(\mathbb{R}\times\mathbb{R}_*^+; db\, da/a^2)$.

La transformée en ondelettes en pratique

En pratique, il est bien sûr impossible de représenter la transformée en ondelettes continue $Wf(b,a)$ d'un signal f pour tout b et tout a. La définition (2.4) permet d'implémenter simplement cette transformée avec une complexité d'orde $N \log N$, grâce à la transformée de Fourier rapide.

Numériquement, l'ondelette, comme le signal, ne sont définis que par un nombre fini de points. Nous pouvons supposer, sans perte de généralité, que le signal f et l'ondelette ψ ne sont définis que pour l'ensemble des naturels $\mathbb{N}$ et que leur support est donné respectivement par $[f] = [0,t_M]$ et $[\psi] = [-b_0,b_0]$, où t_M et b_0 sont des naturels non-nuls.

Les échelles d'analyse, elles aussi, doivent prendre leurs valeurs dans un intervalle compact $[a_m,a_M]$, avec $a_m,a_M \in \mathbb{R}$ et $0 < a_m < a_M$. Une manière naturelle de procéder est de poser $a_m = 1$. Cela n'est en rien restrictif : si des échelles inférieures à a_m sont nécessaires, il suffit de redéfinir une nouvelle ondelette ψ' en posant $\psi'(t) = 1/\sqrt{a'_m}\,\psi(t/a'_m)$, avec $a'_m < a_m$. Ainsi, la plus petite échelle analysée est définie par la résolution de l'ondelette numérique utilisée. Remarquons aussi qu'il est illusoire de penser pouvoir définir l'ondelette pour des échelles arbitrairement petites. Par le théorème de Shanon-Whitaker [115, 296, 343, 388], la plus petite échelle possible est définie par l'intervalle de fréquence

	$t < 0$	$t > t_M$
plateau à 0	$f(t)=0$	$f(t)=0$
plateau	$f(t)=f(0)$	$f(t)=f(t_M)$
périodique	$f(t)=f\big(t \bmod (t_M+1)\big)$	$f(t)=f\big(t \bmod (t_M+1)\big)$
miroir	$f(t)=f(-t)$	$f(t)=f(2t_M-t)$

TAB. 2.1 – *Les valeurs à affecter au signal f hors de son support $[0,t_M]$ pour gérer les effets de bord en fonction de la méthode utilisée.*

où varie $\hat{\psi}$; pour éviter l'effet d'*aliasing* [94, 159], cet intervalle doit être inclus dans $[-\pi,\pi]$. L'échelle maximum a_M est en général fixée par la taille du signal f à analyser : il est inutile d'utiliser les échelles a pour lesquelles $2ab_0 > t_M$. Finalement, les valeurs que peut prendre le paramètre d'échelle $a \in [1,a_M]$ sont en nombre fini. On pose en général $a = 2^{o+v/v_M}$, où v varie par pas entiers entre 0 et $v_M - 1$, $v_M \in \mathbb{N}$ fixant la résolution et le second paramètre o varie par pas entiers entre 0 et $\log_2(a_M) - (v_M - 1)/v_M$. Lorsque a est égal à $2^{o+v/v_M}$, on parle d'échelle définie par la *voix* v et l'*octave* o. Vu ce qui précède, on peut faire varier b entre 0 et t_M.

Le calcul numérique de la convolution (2.3) peut nécessiter des valeurs du signal indéterminées. Typiquement, le support de l'ondelette numérique à une échelle a et une position b est donné par $[b - ab_0,b + ab_0]$. Les valeurs de b pour lesquelles $b - ab_0 < 0$ et $b + ab_0 > t_M$ posent problème puisque le signal f n'est pas défini en tous les points requis pour le calcul transformée en ondelettes. La manière la plus naturelle de procéder est certainement d'empêcher ces cas en interdisant à b de prendre ces valeurs critiques. De cette manière, le nombre de points définissant la transformée en ondelettes à une échelle fixée diminue lorsque l'échelle augmente. Une autre méthode consiste à « élargir le signal » en extrapolant les valeurs nécessaires de $f(t)$ lorsque t est négatif ou supérieur à t_M. Nous exposons ici les quatre manières les plus usitées pour définir ces valeurs. La première, appelée méthode *plateau nul*, consiste à poser $f(t) = 0$ lorsque $t \notin [0,t_M]$. Pour la méthode *plateau*, on pose $f(t) = f(0)$ lorsque $t < 0$ et $f(t) = f(t_M)$ lorsque $t > t_M$, alors que pour l'analyse en fréquence, on utilise en général la méthode *périodique* en redéfinissant f par $f(t) = f\big(t \bmod (t_M + 1)\big)$. Enfin, la méthode *miroir* est caractérisée par les égalités $f(t) = f(-t)$ pour $t < 0$ et $f(t) = f(2t_M - t)$ pour $t > t_M$. Pour la détection de singularités, la meilleure méthode semble être celle du plateau. Les différents procédés sont récapitulés dans le tableau 2.1.

2.2 Caractéristiques des ondelettes

Les critères guidant le choix d'une ondelette mère pour réaliser la transformée dépendent de l'utilisation que l'on souhaite en faire. Pour l'analyse fréquentielle, on choisira une ondelette complexe relativement bien localisée dans le demi-plan temps-fréquence, au sens où elle minimise l'inégalité de Heisenberg [102]. Pour l'analyse de singularités, le choix est tout autre. On choisira des ondelettes possédant suffisamment de moments nuls et bien localisée dans l'espace, c'est-à-dire à décroissance suffisamment rapide. Implicitement, nous supposerons toujours avoir affaire à des signaux dont l'image est incluse dans $\mathbb{R}$.

Dans ce qui suit, par soucis de généralité, nous utiliserons la transformée en ondelettes continue sur $\mathbb{R}^n$, qui généralise naturellement celle sur $\mathbb{R}$.

Notation 2.9 La transformée en ondelettes sur $\mathbb{R}^n$ est définie par

$$W_\psi f(b,a) = \langle f, \frac{1}{a^n}\,\psi(\frac{\cdot - b}{a})\rangle, \tag{2.16}$$

où ψ est une fonction radiale possédant suffisamment de moments nuls.

Le cas $n = 1$ reste l'espace sur lequel nous travaillerons de manière privilégiée, mais certains résultats se formalisent naturellement dans les espaces $\mathbb{R}^n$. Dans $\mathbb{R}$, le facteur de normalisation est $1/a$ et non plus $1/\sqrt{a}$. Les raisons de ce changement sans conséquence sont simples. Cette normalisation est d'abord la plus naturelle pour les changements de variable du type $t' = t/a$. La seconde raison est en rapport avec l'analyse fréquentielle. La transformée de Fourier de $\frac{1}{a}\,\psi(t/a)$ est donnée par $\hat{\psi}(a\omega)$. À la dilatation par un facteur a correspond une multiplication de la fréquence par ce même facteur. Remarquons que la transformée en ondelettes définie par l'égalité (2.16) peut aussi avoir un sens dans un cadre plus général que $L^2(\mathbb{R}^n)$. Ainsi, si l'ondelette est suffisamment régulière, la transformée peut porter sur les distributions. Enfin, on pose $L^2 = L^2(\mathbb{R}^n)$.

Ondelettes adaptées à la détection de singularités

Pour l'analyse de singularités d'un signal, nous verrons qu'il est primordial de pouvoir disposer d'ondelettes ayant une « régularité suffisante ». Nous précisons ici cette notion de régularité.

La première caractéristique que nous imposerons aux ondelettes concerne le nombre de moments nuls.

Définition 2.10 Une ondelette ψ possède m *moments nuls* ($m > 0$) si la relation suivante

est satisfaite,

$$\int_{\mathbb{R}^n} t^k \, \psi(t) \, dt = 0, \tag{2.17}$$

pour tout $k \in \mathbb{N}^n$ tel que $|k| < m$.

Illustrons cette propriété sur $\mathbb{R}$. Soit P une polynôme de degré strictement inférieur à m. Soit aussi une ondelette ψ possédant m moments nuls. On a tôt fait de constater que pour une telle ondelette,

$$W_\psi P(b,a) = \frac{1}{a} \int_{\mathbb{R}} P(t) \; \psi(\frac{t-b}{a}) \; dt = 0. \tag{2.18}$$

La vitesse de décroissance est elle aussi primordiale dans la caractérisation des singularités.

Définition 2.11 Une fonction f est 0-*régulière* si elle vérifie

$$|f(t)| \leqslant \frac{C_k}{1+|t|^k}, \quad \forall \, k \in \mathbb{N} \tag{2.19}$$

pour des constantes C_k et tout t.

Dans $\mathbb{R}$, si l'ondelette est suffisamment régulière, la transformée correspondante peut s'écrire comme un *opérateur différentiel multi-échelles*.

Théorème 2.12 *Une ondelette* 0*-régulière sur* $\mathbb{R}$ *possède* m *moments nuls si et seulement s'il existe une fonction* θ *à décroissance rapide vérifiant*

$$\psi(t) = (-1)^m D^m \theta(t). \tag{2.20}$$

De plus, ψ *ne possède pas plus de* m *moments nuls si et seulement si* $\int_{\mathbb{R}} \theta(t) dt \neq 0$.

Sous les hypothèses de la proposition précédente,

$$W_\psi f(b,a) = a^{m-1} D^m [f * \theta(-\frac{\cdot}{a})](b). \tag{2.21}$$

Nous aurons besoin d'imposer cette décroissance aux dérivées successives de l'ondelette.

Définition 2.13 Une fonction est m-*régulière* ($m > 0$) si

$$|\partial^\alpha f(t)| \leqslant \frac{C_k}{1+|t|^k}, \quad \forall \, k \in \mathbb{N} \tag{2.22}$$

pour tout multi-indice α tel que $|\alpha| \leqslant m$.

En fait, nous supposerons toujours avoir affaire à de telles ondelettes.

Hypothèse de travail 2.14 Nous supposerons que les ondelettes envisagées pour l'étude des irrégularités d'un signal possèdent au moins $m \geqslant \lfloor h \rfloor + 1$ moments nuls et sont m-régulières, h étant l'exposant de Hölder défini par le contexte (voir la section 2.3).

En pratique, pour l'étude de la régularité, nous utiliserons les fonctions dérivées de la gaussienne, présentées dans la sous-section suivante.

Lignes de maxima du module de la transformée en ondelettes et dérivées de la fonction gaussienne

Les lignes de maxima constituent un outil puissant, tant pour la caractérisation ponctuelle que globale des irrégularités présentent dans un signal. Elles sont naturellement associées aux ondelettes obtenues en dérivant la fonction gaussienne. Dans cette sous-section, les ondelettes sont supposées réelles.

Donnons d'abord la définition des lignes de maxima du module.

Définition 2.15 Une *fonction de maxima du module* ℓ associée à la transformée en ondelettes Wf est une fonction continue définie sur un intervalle $[a_m,a_M]$,

$$\ell : [a_m,a_M] \to \mathbb{R} \quad a \mapsto b, \tag{2.23}$$

telle que $b = \ell(a)$ soit un maximum local de $|Wf(\,.\,,a)|$ pour tout a du domaine de définition. L'extremum doit être strict à gauche ou à droite (ou les deux). La courbe (simple) définie par le chemin $(\gamma,[a_m,a_M])$, où γ est l'application donnée par l'égalité $\gamma(a) = (\ell(a),a)$, est appelée *ligne de maxima du module*. L'ensemble des lignes de maxima du module associée à une transformée en ondelettes est appelé *squelette de la transformée en ondelettes*.

Numériquement, puisque les échelles utilisées sont discrètes, les maxima du module sont reliés entre eux à travers les échelles de proche en proche soit par un segment, pour assurer la continuité, soit par des fonctions splines pour obtenir une plus grande régularité. Si l'on utilise cette méthode d'interpolation, les valeurs prises par la ligne de maxima ne sont plus nécessairement entières, alors que le signal est défini sur $\mathbb{N}$.

Si le squelette de la transformée en ondelettes Wf (et la connaissance des valeurs de Wf aux points appartenant à une ligne de maxima du module) ne permet pas de reconstruire le signal f, il permet en général d' en obtenir une bonne approximation [259].

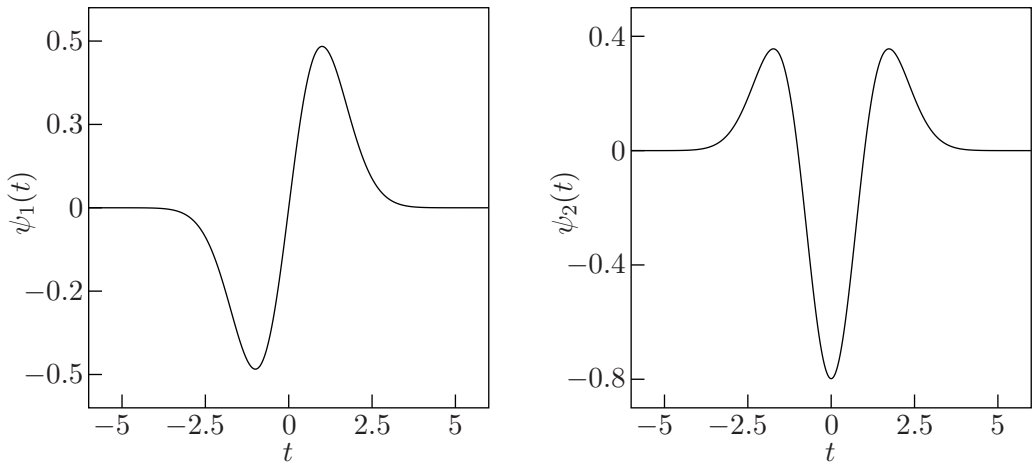

FIG. 2.1 – *Les ondelettes ψ_1 et ψ_2 obtenues à partir de la dérivée première et seconde de la gaussienne respectivement.*

Dans $\mathbb{R}$, si un maximum du module se situe dans le demi-plan espace-échelle, la proposition suivante [190, 395] affirme qu'avec le bon choix d'ondelette, il existe au moins une ligne de maxima passant par ce point et se prolongeant vers les petites échelles.

Théorème 2.16 *Si l'ondelette ψ définie sur $\mathbb{R}$ est la dérivée m-ième de la gaussienne,*

$$\psi_m(t) = \frac{1}{\sqrt{2\pi}} D^m \exp(-\frac{t^2}{2}) \quad (m > 0), \tag{2.24}$$

toute les lignes de maxima sont définies sur un intervalle du type $]0,a_0]$ ($a_0 \in \mathbb{R}_*^+$) ou $]0, +\infty[$.

Les ondelettes obtenue à partir de la dérivée première et seconde de la gaussienne sont représentée par la figure 2.1.

Remarquons que, par la proposition 2.12, la dérivée m-ième de la gaussienne possède exactement m moments nuls.

En pratique, ce sont ces ondelettes que nous utiliserons pour la caractérisation des singularités ou la détermination du spectre de Hölder d'un signal. La transformée en ondelettes par la dérivée m-ième de la fonction gaussienne à l'échelle a peut être interprétée, grâce à l'égalité (2.21), comme la dérivée m-ième du signal convolué par la fonction gaussienne à l'échelle a, cette dernière jouant le rôle de fenêtre lissante. En particulier, en utilisant les notations définies par l'égalité (2.24) et en notant θ la fonction gaussienne, les extrema locaux de $W_{\psi_1} f(\,.\,,a)$ selon b correspondent aux points d'inflection du signal lissé $W_\theta f(\,.\,,a)$ [⋇] et aux points où la transformée $W_{\psi_2} f(\,.\,,a)$ change de signe. La détection de tels extrema correspond à une détection de bords de Canny [90]. La détection

⋇. Puisque θ n'est pas de moyenne nulle, $W_\theta f$ ne désigne pas une transformée en ondelettes.

des valeurs pours lesquelles $W_{\psi_2} f$ change de signe correspond à une détection de bords de Marr-Hildreth [267, 268]. Les maxima (resp. minima) de $|W_{\psi_1} f|(\,.\,,a)$ représentent les points où le signal $W_\theta f(\,.\,,a)$ varie rapidement (resp. lentement) ; l'étude des points où la transformée en ondelettes $W_{\psi_2} f$ change de signe ne permet pas d'obtenir directement ce type d'information.

Enfin, pour les ondelettes du type (2.24), nous utiliserons la définition suivante, précisant la notion de taille.

Définition 2.17 Étant donné une fonction gaussienne $f_\sigma(t) = \exp(-t^2/2\sigma^2)/\sqrt{2\pi}\sigma$, la *taille caractéristique* $S_\psi(a)$ de l'ondelette $\psi_m(t) = D^m f_\sigma(t)$ à l'échelle a vaut $S_\psi(a) = 2a\sigma$.

Pour des problèmes plus spécifiques, il est parfois utile d'adopter d'autres définitions de la taille de l'ondelette [300].

Concernant l'étude fréquentielle

La transformée en ondelettes continue permet aussi une étude du type temps-fréquence. Si l'on considère l'ondelette comme une fenêtre, la taille de celle-ci varie selon l'échelle à laquelle on se trouve. Ainsi, contrairement à la transformée de Gabor, la fenêtre de la transformée en ondelettes est adaptative : elle s'élargit pour les basses fréquences de manière à conserver « la même quantité d'information ». Nous travaillerons ici exclusivement sur $\mathbb{R}$.

Pour une étude du type temps-fréquence, il est plus efficace de prendre une ondelette progressive pour séparer amplitude et phase.

Définition 2.18 Une ondelette ψ est dite *progressive* si la fonction de quadrature associée est nulle,

$$(I - iH)\psi(t) = 0, \tag{2.25}$$

où $H\psi$ désigne la transformée de Hilbert de l'ondelette[❋],

$$H\psi(t) = \frac{1}{\pi}\,\mathrm{PV}\!\int_{\mathbb{R}} \frac{\psi(x)}{t-x}\,dx = \frac{1}{\pi}\lim_{\varepsilon\to 0^+}\Big(\int_{-\infty}^{t-\varepsilon} \frac{\psi(x)}{t-x}dx + \int_{t+\varepsilon}^{\infty} \frac{\psi(x)}{t-x}dx\Big). \tag{2.26}$$

Puisque $\widehat{H\psi}(\omega) = -i\,\mathrm{sign}(\omega)\,\hat{\psi}(\omega)$, l'égalité (2.25) est vérifiée si et seulement si la transformée de Fourier est causale,

$$\hat{\psi}(\omega) = 0, \quad \forall\,\omega < 0. \tag{2.27}$$

❋. Autrement dit, l'ondelette appartient au second espace de Hardy complexe.

La transformée de Hilbert d'une fonction réelle étant réelle, une telle ondelette est nécessairement complexe. Cependant, la transformée de Fourier d'une ondelette progressive étant causale, elle est entièrement déterminée par la transformée de Fourier de sa partie réelle,

$$\hat{\psi}(\omega) = \begin{cases} 2\widehat{\Re\psi}(\omega) & \text{si } \omega \geqslant 0, \\ 0 & \text{si } \omega < 0. \end{cases} \tag{2.28}$$

La transformée de Hilbert permet de définir la "partie analytique" d'un signal*.

Remarque 2.19 La *partie analytique* f_a d'un signal réel f est définie par l'égalité $f_a(t) = f(t) + iHf(t)$. Cette "représentation analytique" permet d'obtenir un signal complexe sans fréquence négative [76, 103]. Si f vérifie une identité du type (2.25), alors bien sûr $f_a = 2f$. □

L'ondelette mère que nous utiliserons pour effectuer les études temps-fréquence sera celle de Morlet.

Définition 2.20 L'*ondelette de Morlet* ψ_M est définie par sa transformée de Fourier :

$$\hat{\psi}_M(\omega) = \exp(-\frac{(\omega-\Omega)^2}{2}) - \exp(-\frac{\omega^2}{2})\exp(-\frac{\Omega^2}{2}), \tag{2.29}$$

où Ω est appelé la *fréquence centrale* de l'ondelette. Il s'agit d'une gaussienne translatée, le second terme assurant que cette fonction est nulle à l'origine. L'ondelette peut s'écrire explicitement

$$\psi_M(t) = \Big(\exp(-i\Omega t) - \exp(-\frac{\Omega^2}{2})\Big)\ \exp(-\frac{t^2}{2}). \tag{2.30}$$

En général, l'expression $\exp(-\Omega^2/2)$ dans la relation (2.29) peut être négligée* pour obtenir

$$\hat{\psi}_M(\omega) = \exp(-\frac{(\omega-\Omega)^2}{2}) \tag{2.31}$$

et

$$\psi_M(t) = \exp(i\Omega t)\ \exp(-\frac{t^2}{2}). \tag{2.32}$$

Si l'on choisit $\Omega = \pi\sqrt{2/\ln 2}$, le rapport entre les deux plus hauts maxima est $1/2$. En effet, pour cette valeur, la période de $\cos(\Omega t)$ est $\sqrt{\ln 4}$ et $\psi_M(\sqrt{\ln 4}) = 1/2$.

∗. cette dénomination est quelque peu malheureuse : la partie analytique d'un signal n'est pas à mettre en relation avec les fonctions analytiques.

∗. Si Ω est supérieur à 5, ce terme est inférieur à $4\,10^{-6}$.

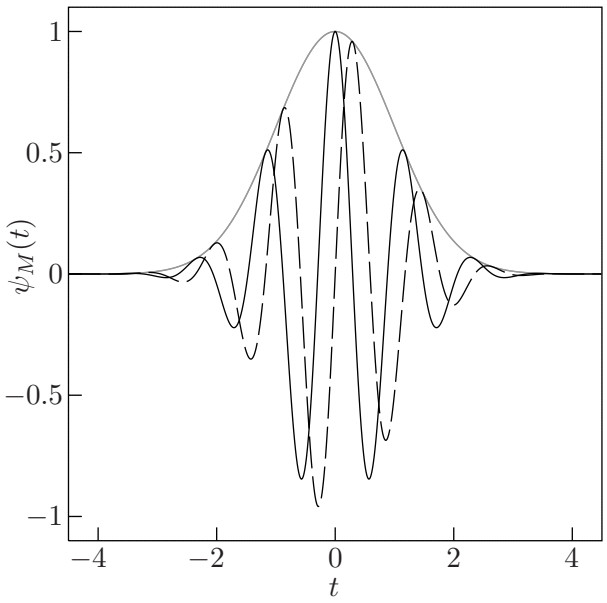

FIG. 2.2 – *Les parties réelle et imaginaire (en traits interrompus) de l'ondelette de Morlet. Son module est représenté en gris.*

L'égalité (2.32) ne définit pas à proprement parler une ondelette progressive, mais ici aussi, les valeurs de $\hat{\psi}_M$ sont en pratique considérées comme négligeables pour les fréquences négatives*. L'ondelette de Morlet est représentée par la figure 2.2

Nous sommes intéressés par le problème de la détection de fréquences caractéristiques. Illustrons l'intérêt des ondelettes progressives en considérant un exemple simple, où la fonction à étudier est $f(t) = \cos(\omega_0 t)$. En supposant que la transformée est définie, on obtient, grâce à l'égalité (2.3),

$$W_\psi f(b,a) = \frac{1}{2} \exp(i\omega_0 b)\, \hat{\bar{\psi}}(a\omega_0). \tag{2.33}$$

Si $\hat{\psi}$ est à valeurs réelles, le module de la transformée en ondelettes $|Wf(b,a)|$ est donné par la composante $\hat{\psi}(a\omega_0)$. La phase de la transformée $\exp(i\omega_0 b)$, quant à elle, décrit le comportement de la phase de la fonction étudiée.

À partir des relations (2.31) et (2.33), on peut détecter la fréquence ω_0 de la fonction $\cos(\omega_0 t)$ en cherchant le maximum de $\hat{\psi}_M(\,.\,\omega_0)$. Ce dernier est atteint pour $a = \Omega/\omega_0$. Dans le demi-plan espace-échelle, la fréquence ω_0 de la fonction $\cos(\omega_0 t)$ vaut donc Ω/a_ω, où a_ω est l'échelle à laquelle le module de la transformée en ondelettes atteint son maximum (le

*. Si Ω est supérieur à 5, les valeurs correspondant à des fréquences négatives sont aussi inférieures à $4\,10^{-6}$.

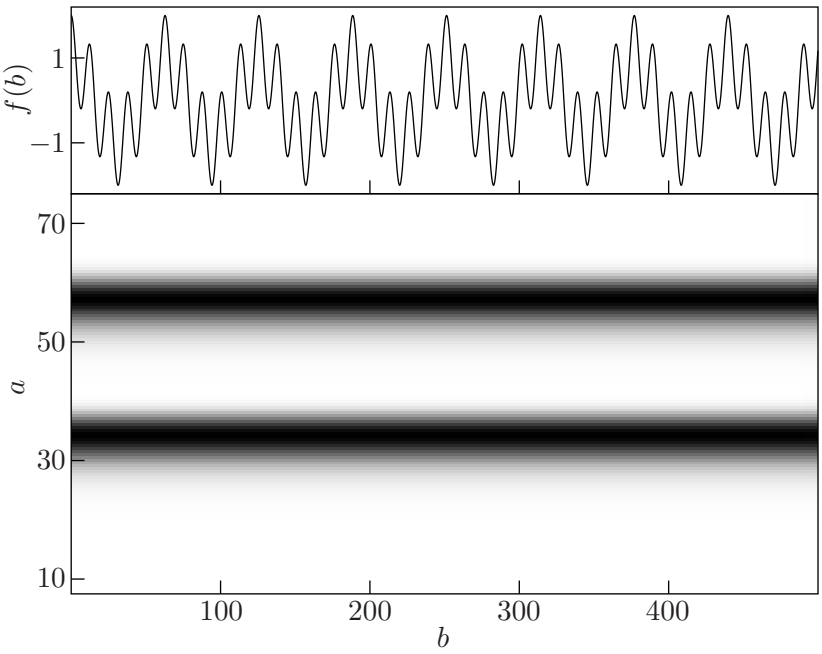

FIG. 2.3 – *Représentation du module de la transformée en ondelettes d'un signal du type* $f(t) = \big(\cos(0.5\,t) + \cos(0.1\,t)\big)\chi_{[-T,T]}(t)$ *en utilisant l'ondelette mère de Morlet. Les couleurs utilisées pour cette représentation vont du blanc (pour les valeurs les plus faibles) au noir (pour les valeurs les plus fortes).*

module de la transformée en ondelettes ne dépend pas de b).

Exemple 2.21 Considérons une fonction du type[*] $f(t) = \big(\cos(\omega_1\,t)+\cos(\omega_2\,t)\big)\chi_{[-T,T]}(t)$ ($T \in \mathbb{R}^+$), avec $\omega_1 = 1/2$ et $\omega_2 = 10^{-1}$. Les grandes valeurs de $|W_{\psi_M} f(b,.)|$ se répartissent en deux zones (cf. figure 2.3) permettant d'estimer ω_1 et ω_2. Il suffit de sommer à chaque échelle a les valeurs de la transformée en ondelettes selon b et de renormaliser par la longueur de l'intervalle $b_2 - b_1$ sur lequel sont évalués les coefficients $Wf(b,a)$, $b \in [b_1,b_2] \subset [-T,T]$. On obtient ainsi un signal présentant deux maxima (cf. figure 2.4) en $a_1 = 10.5 \pm 1/2$ et $a_2 = 53.5 \pm 1/2$, permettant d'estimer les fréquences $\tilde{w}_1 = \Omega/a_1 = 0.51 \pm 0.025$ et $\tilde{w}_2 = 0.1 \pm 10^{-3}$. □

Les déductions qui précèdent sont toujours d'application si l'on module de manière douce la fréquence ou l'amplitude [365]. Pour une fonction du type $f(t) = C(t)\cos(\omega_0 t)$, où C est une fonction suffisamment régulière, un développement polynomial de cette dernière permet d'approcher la transformée en ondelettes de f par une expression du type de (2.33), si la dérivée de C est suffisamment petite, en particulier par rapport à l'amplitude. Si l'on module la fréquence, l'algorithme associé se complique légèrement.

[*]. Cette fonction est à support compact et on suppose que ce support est suffisamment grand pour ne pas influencer l'intervalle sur lequel est évalué la transformée en ondelettes.

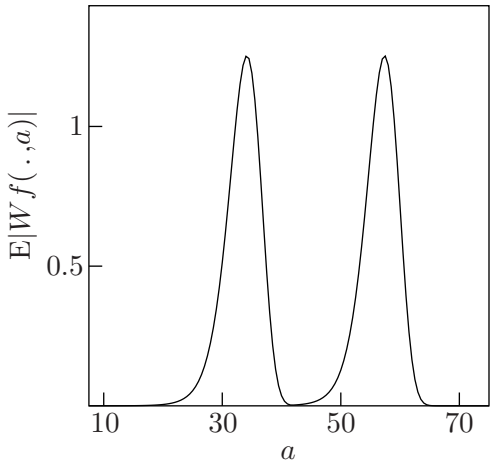

FIG. 2.4 – *La moyenne des valeurs du module de la transformée en ondelettes de f en fonction de l'échelle. Les deux périodes de f apparaîssent clairement.*

2.3 Exposants de Hölder

Notre but est de définir et caractériser l'irrégularité d'un signal, tant globalement, *via* les espaces de Hölder, que localement, grâce aux exposants de Hölder. Nous montrerons ensuite que ces exposants ne peuvent entièrement décrire les singularités de type oscillant et introduirons les exposants d'oscillation. Dans cette section, nous supposerons avoir affaire à des applications réelles.

Espaces de Hölder

Les espaces de Hölder raffinent les espaces des fonctions s fois continûment dérivables, en ce sens qu'ils permettent à l'exposant s de varier sur l'ensemble des réels. Ces espaces peuvent facilement être caractérisés par la transformée en ondelettes et seront ensuite généralisés aux exposants négatifs grâce aux espaces de Besov. Les espaces de Hölder, Zygmund et Besov sont largement considérés dans les références [371, 372, 373, 374]. Les relations entre ces espaces et la transformée en ondelettes sont exposées dans les références [112, 256, 278, 281, 282, 365] notamment.

Les développements qui suivent portent sur les fonctions définies sur l'espace euclidien $\mathbb{R}^n$. Pour introduire les espaces de Hölder, nous aurons besoin de la notion suivante.

Notation 2.22 Nous noterons Δ_l^m la *différence d'ordre* $m \in \mathbb{N}_0$,

$$\Delta_l^m f(t) = \sum_{j=0}^{m} \binom{m}{j} (-1)^{m-j} f(t + jl), \tag{2.34}$$

où $l \in \mathbb{R}^n$. Pour éviter toute confusion avec la décomposition de Littlewood-Paley, nous écrirons Δ_l^1, et non Δ_l, pour désigner la différence d'ordre 1. En particulier, on a $\Delta_l^1 f(t) = f(t+l) - f(t)$ et $\Delta_l^2 f(t) = f(t+2l) - 2f(t+l) + f(t)$.

Afin de simplifier les définitions, nous utiliserons aussi les espaces C_∞^m.

Notation 2.23 On pose

$$C_\infty^0 = C^0(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n), \tag{2.35}$$

et équipe cet espace de la norme $\|\,.\,\|_{C_\infty^0} = \|\,.\,\|_{L^\infty}$. Les espaces C_∞^m, $m \in \mathbb{N}$, se définissent naturellement,

$$C_\infty^m = \{f : D^\alpha f \in C_\infty^0, |\alpha| \leqslant m\}. \tag{2.36}$$

On munit ces espaces de la norme

$$\|f\|_{C_\infty^m} = \sum_{|\alpha| \leqslant m} \|D^\alpha f\|_{L^\infty}. \tag{2.37}$$

Commençons par introduire formellement les espaces de Hölder. Pour un exposant s strictement compris entre zéro et un, l'idée est de définir l'espace des fonctions f telles qu'il existe une constante C pour laquelle

$$|f(t+l) - f(t)| \leqslant C|l|^s, \tag{2.38}$$

quelque soient les points t et l de l'espace $\mathbb{R}^n$. Pour pouvoir équiper cet espace d'une norme, il nous faut considérer les fonctions bornées. Pour de telles fonctions, seules les petites valeurs de l sont importantes. On étend cette notion aux valeurs de s positives non-entières de manière analogue aux espaces C_∞^m.

Définition 2.24 Étant donné $s \in \mathbb{R}_*^+ \setminus \mathbb{N}_0$, l'*espace de Hölder* $\mathcal{C}^s$ est défini comme suit,

$$\mathcal{C}^s = \{f \in C_\infty^{\lfloor s \rfloor} : \|f\|_{\mathcal{C}^s} = \|f\|_{C_\infty^{\lfloor s \rfloor}} + \sum_{|\alpha| = \lfloor s \rfloor} \sup_{0<|l|\leqslant 1} \frac{|\Delta_l^1 D^\alpha f|}{|l|^{s - \lfloor s \rfloor}} < \infty\}. \tag{2.39}$$

Si $s \in \mathbb{N}_0$, $\mathcal{C}^s$ est défini comme suit[⁎],

$$\mathcal{C}^s = \{f \in C_\infty^{\lfloor s \rfloor^-} : \|f\|_{\mathcal{C}^s} = \|f\|_{C_\infty^{\lfloor s \rfloor^-}} + \sum_{|\alpha| = \lfloor s \rfloor^-} \sup_{0<|l|\leqslant 1} \frac{|\Delta_l^2 D^\alpha f|}{|l|^{s - \lfloor s \rfloor^-}} < \infty\}, \tag{2.40}$$

où $\lfloor s \rfloor^- = \lfloor s \rfloor$ si $s \notin \mathbb{N}$ et $\lfloor s \rfloor^- = \lfloor s \rfloor - 1$ sinon.

⁎. Pour définir les espaces $\mathcal{C}^s$, avec $s \in \mathbb{N}_0$, il suffit de poser $\lfloor s \rfloor^- = -1$, mais l'égalité (2.40) définit les espaces de Zygmund de manière générale [374].

Ces espaces sont parfois appelés espaces de Hölder-Zygmund. La raison pour laquelle nous avons défini différemment les espaces de Hölder pour les exposants entiers est que nous voulons assurer les mêmes propriétés à tous les espaces de Hölder, comme nous le verrons dans la suite. Les fonctions vérifiant une relation du type (2.38) pour $s = 1$ appartiennent à $\mathcal{C}^1$, mais la réciproque est fausse[❋]. La fonction $f(t) = t \log |t|$ vérifie $|\Delta_l^2 f(t)| \leqslant C|l|$, mais pas $|t| \log |t|$. Les espaces $\mathcal{C}^s$ sont des espaces de Banach.

Ces espaces ne peuvent être caractérisés par la transformée en ondelettes[∗], en raison de la composante L^∞. Toutefois, comme nous souhaitons utiliser ce type d'espace pour définir la régularité d'une fonction, il n'est pas nécessaire de pouvoir discerner deux fonctions différant par une fonction de classe C^∞. Dans la suite, nous allons donc considérer la version homogène de ces espaces.

Définition 2.25 Étant donné $s \in \mathbb{R}_*^+ \setminus \mathbb{N}_0$, l'*espace de Hölder homogène* $\dot{\mathcal{C}}^s$ est défini comme suit,

$$\dot{\mathcal{C}}^s = \{f \in \tilde{C}^{\lfloor s \rfloor} : \|f\|_{\dot{\mathcal{C}}^s} = \sum_{|\alpha|=\lfloor s \rfloor} \sup_{|l|\neq 0} \frac{|\Delta_l^1 D^\alpha f|}{|l|^{s-\lfloor s \rfloor}} < \infty\}, \tag{2.41}$$

où $\tilde{C}^{\lfloor s \rfloor}$ désigne l'espace quotient $C^{\lfloor s \rfloor}$ modulo les polynômes de $\mathbb{R}^n$. Si $s \in \mathbb{N}_0$, $\dot{\mathcal{C}}^s$ est défini comme suit,

$$\dot{\mathcal{C}}^s = \{f \in \tilde{C}^{\lfloor s \rfloor^-} : \|f\|_{\dot{\mathcal{C}}^s} = \sum_{|\alpha|=\lfloor s \rfloor^-} \sup_{|l|\neq 0} \frac{|\Delta_l^2 D^\alpha f|}{|l|^{s-\lfloor s \rfloor^-}} < \infty\}. \tag{2.42}$$

Une fonction appartenant à un espace $\dot{\mathcal{C}}^s$, $s \in \mathbb{R}_*^+$, est appelée *fonction uniformément hölderienne* d'exposant s.

Pour une fonction f hölderienne d'exposant $s > \lfloor s \rfloor$, on peut écrire,

$$f(t + l) = P_t(l) + R(l), \tag{2.43}$$

avec $|R(l) \leqslant C|l|^s$, où P_t est un polynôme de $\mathbb{R}^n$. Inversement, si une telle égalité est vérifiée, $f \in \dot{\mathcal{C}}^s$. Lorsque $s = \lfloor s \rfloor$, il faut remplacer le reste R par R', où l'inégalité $|R'(l)| \leqslant C|l|^s \log_Z |l|$ est vérifiée, avec $\log_Z(t) = \max\{\big|\log |t|\big|, \log 2\}$ [278].

Pour pouvoir généraliser les espaces de Hölder aux exposants négatifs, mais aussi les caractériser par la transformée en ondelettes, nous allons introduire les espaces de Besov, *via* la décomposition de Littlewood-Paley.

❋. D'une manière générale, les espaces de Hölder, comme définis par la relation (2.39), sont inclus dans les espaces de Zygmund, définis par la relation (2.40), ces espaces étants différents pour les exposants entiers.

∗. L'introduction de la fonction d'échelle permet de résoudre ce problème [278].

Notation 2.26 L'espace sur lequel agissent les opérateurs est celui des distributions tempérées. Soit une fonction $\hat{\varphi}$ de l'espace de Schwartz S telle que

$$\hat{\varphi}(\omega) = \begin{cases} 1 & \text{si } |\omega| \leqslant 1/2, \\ 0 & \text{si } |\omega| \geqslant 1. \end{cases} \tag{2.44}$$

On définit le filtre passe-bas S_j qui, dans l'espace de Fourier, consiste à multiplier la transformée de Fourier de la distribution par $\hat{\varphi}(\omega/2^j)$ et on pose $\Delta_j = S_{j+1} - S_j$. Le support de la transformée de Fourier de $\Delta_j f$ est contenu dans $\{\omega : 2^{j-1} \leqslant |\omega| \leqslant 2^{j+1}\}$ et $I = S_0 + \sum_{j \geqslant 0} \Delta_j$.

Les espaces de Besov peuvent être définis à partir de cette décomposition.

Définition 2.27 Étant donnés $s \in \mathbb{R}$ et $1 \leqslant p,q \leqslant \infty$, l'*espace de Besov* $B^s_{p,q}$ est défini comme suit,

$$B^s_{p,q} = \{f \in S' : \|f\|_{B^s_{p,q}} = \|S_0 f\|_{L^p} + \|\{2^{js}\|\Delta_j f\|_{L^p}\}_{j\in\mathbb{N}}\|_{l^q} < \infty\}, \tag{2.45}$$

où S désigne l'espace de Schwartz.

Il s'agit d'un espace de Banach* pour la norme $\|\,.\,\|_{B^s_{p,q}}$. Les espaces de Hölder sont des espaces de Besov [372].

Proposition 2.28 *Pour tout $s > 0$, on a*

$$\mathcal{C}^s = B^s_{\infty,\infty}. \tag{2.46}$$

On définit naturellement ces espaces sur les sous-ensembles de $\mathbb{R}^n$. Étant donné un sous-ensemble E de $\mathbb{R}^n$, nous dirons qu'une fonction f appartient à $\mathcal{C}^s(E)$ si elle est la restriction sur E d'une fonction f_0 de $\mathcal{C}^s$. On peut aussi définir les espaces $\mathcal{C}^s$ localement.

Pour considérer le cas des espaces homogènes, il nous faut brièvement rappeler ce qu'est l'espace quotient S'_0 des distributions tempérées modulo les polynômes.

Notation 2.29 On pose $S_0 = \{\phi \in S : (D^\alpha \hat{\phi})(0) = 0, \forall \alpha\}$, où α est un multi-indice. La restriction d'une distribution tempérée f à S_0 appartient à S'_0 et on peut considérer S'_0 comme l'espace quotient des distributions tempérées modulo les polynômes. En particulier, si P est un polynôme de $\mathbb{R}^n$, $(f + P)(\phi) = f(\phi)$, pour tout $\phi \in S_0$.

*. On peut étendre la définition 2.27 aux indices p et q inférieurs à 1, mais les espaces obtenus ne sont plus de Banach. Ce sont cependants des espaces métriques complets (*i.e.* polonais) [372].

La définition des espaces de Besov homogènes devient alors naturelle.

Définition 2.30 Étant donné $s \in \mathbb{R}$ et $1 \leqslant p,q \leqslant \infty$, l'*espace de Besov homogène* $\dot{B}^s_{p,q}$ est défini comme suit,

$$\dot{B}^s_{p,q} = \{f \in S'_0 : \|f\|_{\dot{B}^s_{p,q}} = \|\{2^{js}\|\Delta_j f\|_{L^p}\}_{j\in\mathbb{Z}}\|_{l^q} < \infty\}. \tag{2.47}$$

L'analogue de la proposition 2.28 est aussi vérifiée[+].

Proposition 2.31 *Pour tout $s > 0$, on a*

$$\dot{\mathcal{C}}^s = \dot{B}^s_{\infty,\infty}. \tag{2.48}$$

Ces identités permettent de définir les espaces de Hölder pour tous les exposants s en posant $\mathcal{C}^s = B^s_{\infty,\infty}$ et $\dot{\mathcal{C}}^s = \dot{B}^s_{\infty,\infty}$, pour tout s réel. Ces espaces obéissent à diverses relations d'inclusion.

Proposition 2.32 *Soient $s \in \mathbb{R}$ et $1 \leqslant p,q \leqslant \infty$. On a les inclusions suivantes :*

- *si $s > 0$, alors $B^s_{\infty,\infty} \subset \dot{B}^s_{\infty,\infty}$,*
- *si $s_1 \leqslant s_2$, alors $B^{s_2}_{p,q} \subset B^{s_1}_{p,q}$, cette inclusion n'étant pas vérifiée pour les espaces homogènes.*
- *si $1 \leqslant q_1 \leqslant q_2 \leqslant \infty$, alors $\dot{B}^s_{p,q_1} \subset \dot{B}^s_{p,q_2}$ et $B^s_{p,q_1} \subset B^s_{p,q_2}$,*
- *si $1 \leqslant p_1 \leqslant p_2 \leqslant \infty$ et $s_1 = s_2 + n(1/p_1 - 1/p_2)$, alors $\dot{B}^{s_1}_{p_1,q} \subset \dot{B}^{s_2}_{p_2,q}$ et $B^{s_1}_{p_1,q} \subset B^{s_2}_{p_2,q}$.*

Ces résultats sont classiques. Ainsi, on a $\mathcal{C}^s \subset \dot{\mathcal{C}}^s$ et $\mathcal{C}^{s_2} \subset \mathcal{C}^{s_1}$, si $s_1 \leqslant s_2$. Dans ce cas, $\dot{\mathcal{C}}^{s_2} \not\subset \dot{\mathcal{C}}^{s_1}$. Finalement, cette identification permet de caractériser les espaces de Hölder homogènes par la transformée en ondelettes[❋] *via* les espaces de Besov [278].

Proposition 2.33 *On a $f \in \dot{B}^s_{p,q}$ si et seulement si*

$$\|\frac{1}{a^s}\|Wf(\,.\,,a)\|_{L^p}\|_{L^q(\mathbb{R}^+_*)} < \infty. \tag{2.49}$$

En particulier, $f \in \dot{B}^s_{\infty,\infty}$ si et seulement s'il existe une constante C telle que

$$|Wf(b,a)| \leqslant Ca^s. \tag{2.50}$$

+. Il existe d'autres relations de ce genre. Ainsi, $\dot{H}^s = \dot{B}^s_{2,2}$, où $\dot{H}^s$ est l'espace de Sobolev homogène $\{f \in \mathcal{S}' : |\omega|^s|\hat{f}(\omega)| \in L^2\}$.

❋. Rappelons une dernière fois que l'ondelette est supposée suffisamment régulière ; ici $(\lfloor\alpha\rfloor+1)$-régulière au moins.

Nous verrons que la transformée en ondelettes permet aussi de définir les espaces deux-microlocaux grâce à une relation analogue à l'inégalité (2.50).

Terminons en faisant deux remarques. La première concerne l'appellation « homogène ». Un espace E est dit homogène si sa norme vérifie l'égalité $\|f(\lambda\,.\,)\| = \lambda^r \|f\|$, pour un $r \in \mathbb{R}$ et tout $\lambda > 0$. Pour les espaces de Besov homogènes $\dot{B}^s_{p,q}$, on constate sans peine que $\|f(\lambda\,.\,)\|_{\dot{B}^s_{p,q}} = \lambda^{s-n/p}\|f\|_{\dot{B}^s_{p,q}}$ et notamment $\|f(\lambda\,.\,)\|_{\dot{\mathcal{C}}^s} = \lambda^s \|f\|_{\dot{\mathcal{C}}^s}$. Les espaces de Besov homogènes sont définis sur S'_0. Cela signifie que, pour n'importe quel polynôme P de $\mathbb{R}^n$, $\|f + P\|_{\dot{B}^s_{p,q}} = \|f\|_{\dot{B}^s_{p,q}}$; c'est en particulier vrai pour les espaces de Hölder homogènes. Pourtant, il est plus naturel de quotienter ces espaces $\dot{\mathcal{C}}^s$ non pas par tous les polynômes, mais uniquement par ceux de degré au plus $\lfloor s \rfloor$. D'une manière générale, on peut souhaiter quotienter l'espace $\dot{B}^s_{p,q}$ par les polynôme de degré inférieur à $s - n/p$ uniquement [282]. Dans ce cas, le polynôme intervenant dans la relation (2.43) est de degré $\lfloor s \rfloor$. Notre but étant d'obtenir des estimations locales de la régularité d'une fonction, nous adopterons implicitement cette convention.

Régularité hölderienne ponctuelle

Nous allons maintenant définir une version ponctuelle des espaces de Hölder. L'exposant lié à ce type d'espace en un point définit, d'une certaine manière, la régularité de la fonction en ce point. Pour pouvoir relier ces considérations à la transformée en ondelettes, nous devrons considérer les espaces deux-microlocaux introduit par Bony [72]. Les développements qui suivent sont essentiellement dûs à Jaffard [201]. Le lecteur pourra se référer aux travaux [199, 201, 214, 281, 282] pour de plus amples développements.

Par analogie avec la caractérisation (2.43) des espaces de Hölder homogènes, on introduit les espaces $\mathcal{C}^{s,*}$ comme suit.

Définition 2.34 Nous écrirons $f \in \mathcal{C}^{s,*}(t)$ s'il existe un polynôme P_t de degré inférieur à s et une constante C_t tels que

$$|f(t+l) - P_t(l)| \leqslant C_t |l|^s, \tag{2.51}$$

pour tout $l \in \mathbb{R}^n$. Nous écrirons $f \in \mathcal{C}^{s,*}$ si la relation (2.51) est vérifiée pour tout $t \in \mathbb{R}^n$, avec une constante C indépendante de t.

Ainsi, $f \in \mathcal{C}^{1,*}$ signifie que f est lipschitz.

Les espaces deux-microlocaux nous permettront d'étudier les espaces $\mathcal{C}^{s,*}$ *via* la transformée en ondelettes. La notation utilisée est celle relative à la décomposition de Littlewood-

Paley (notation 2.26).

Définition 2.35 Étant donné deux nombres réels s et s', l'espace deux-microlocal $\mathcal{C}^{s,s'}(t)$ est l'espace défini comme suit,

$$\mathcal{C}^{s,s'}(t) = \{f \in S_0' : \|f\|_{\mathcal{C}^{s,s'}(t)} = \|\{2^{js}\|(1+2^j|l|)^{s'}\Delta_j f(t+l)\|_{L^\infty}\}_{j\in\mathbb{Z}}\|_{l^\infty} < \infty\}. \quad (2.52)$$

Cet espace, muni de la norme $\|\,.\,\|_{\mathcal{C}^{s,s'}(t)}$, est un espace de Banach. On constate immédiatement que $\mathcal{C}^{s,0}(t) = \dot{B}^s_{\infty,\infty}$. La caractérisation de ces espaces par la transformée en ondelettes[⋇] est aisément démontrée.

Proposition 2.36 *Pour tous s et s', $f \in \mathcal{C}^{s,s'}(t)$ si et seulement si*

$$|Wf(b,a)| \leqslant Ca^s(1+\frac{|t-b|}{a})^{-s'}, \quad (2.53)$$

pour une constante C, quel que soit a et b.

Par des considérations élémentaires sur des opérateurs de convolution, on peut montrer que les espaces deux-microlocaux sont stables pour la dérivation.

Proposition 2.37 *Pour tous s et s', $f \in \mathcal{C}^{s,s'}(t)$ si et seulement si $\partial_j f \in \mathcal{C}^{s-1,s'}(t)$ pour tout j ($1 \leqslant j \leqslant n$).*

Concernant les relations entre $\mathcal{C}^{s,*}$ et les espaces deux-microlocaux, on peut facilement montrer les relations suivantes.

Proposition 2.38 *Si $s > -n$, les inclusions suivantes sont vérifiées,*

$$\mathcal{C}^{s,*}(t) \subset \mathcal{C}^{s,-s}(t), \quad (2.54)$$

et, si s' est tel que $s > s'$,

$$\mathcal{C}^{s,-s'}(t) \subset \mathcal{C}^{s,*}(t). \quad (2.55)$$

Le théorème suivant, reliant les espaces $\mathcal{C}^{s,*}$ et $\mathcal{C}^{s,s'}$ est plus ardu à obtenir.

Théorème 2.39 *Soit s un nombre strictement positif. Si f un élément de $\mathcal{C}^{s,-s}(t) \cap \dot{\mathcal{C}}^\varepsilon$ pour un $\varepsilon > 0$, il existe un polynôme P_t de degré moindre que s tel que, si $|l| \leqslant 1$,*

$$|f(t+l) - P_t(l)| \leqslant C|l|^s \log\frac{2}{|l|}, \quad (2.56)$$

ce résultat étant optimal.

⋇. Rappelons une dernière fois que l'ondelette est supposée suffisament régulière (*cf.* hypothèse de travail 2.14).

Si $f \in \mathcal{C}^{s,*}(t)$, la transformée en ondelettes de f vérifie l'inégalité (2.53). Inversement, si cette inégalité est vérifiée, le théorème 2.39 donne, sous des hypothèses de régularité supplémentaires, un résultat légèrement plus faible que l'appartenance locale à $\mathcal{C}^{s,*}(t)$. Ce type d'espace ne peut donc être caractérisé par la transformée en ondelettes. Toutefois, comme nous allons le constater dans la suite, les versions locales de ces résultats sont d'une importante portée pratique.

Il peut être intéressant de pouvoir accéder aux exposants négatifs ou nuls (présents en turbulence notamment). Il est tentant de définir $\mathcal{C}^{s,*}(t)$ lorsque s est négatif à partir de la relation (2.51) en demandant à f de vérifier l'inégalité

$$|f(t+l)| \leqslant C|l|^s. \tag{2.57}$$

Cette définition est assez restreinte, puisque si $s \leqslant -n$, f peut ne pas être assimilable à une distribution, rien n'assurant l'intégrabilité locale (dans un tel cas, les outils tels que la transformée en ondelettes ne sont pas définis). Le problème peut être réglé en demandant que f appartienne à $\dot{B}^s_{\infty,\infty}$ et que la restriction[❋] de f sur $\mathbb{R}^n \setminus \{t\}$ soit associée à une fonction vérifiant l'inégalité (2.57). Pour de telles valeurs de s, $\mathcal{C}^{s,*}(t) \subset \dot{B}^s_{\infty,\infty}$. Cette inclusion est dans la direction opposée de celle ayant lieu pour les exposants positifs. On a de plus les relations suivantes.

Proposition 2.40 *Si $s \leqslant n$, les inclusions suivantes sont vérifiées,*

$$\mathcal{C}^{s,*}(t) \subset \mathcal{C}^{s,-s}(t), \tag{2.58}$$

et, pour tout s' tel que $s > s'$,

$$\mathcal{C}^{s,-s'} \subset \mathcal{C}^{s,*}(t). \tag{2.59}$$

Pour pouvoir étudier la régularité ponctuelle de f, il nous faut considérer les versions locales des précédents résultats.

Définition 2.41 Étant donné une fonction $f \in L^\infty_{\text{loc}}$, nous écrirons $f \in \mathcal{C}^s(t)$ s'il existe un polynôme P_t de degré inférieur à s, une constante C et un voisinage $\mathcal{V}_0$ de 0 tels que

$$|f(t+l) - P_t(l)| \leqslant C|l|^s, \tag{2.60}$$

pour tout l appartenant à $\mathcal{V}_0$. Un tel f est dit *hölderien* d'exposant s en t.

Si f est $\lfloor s \rfloor$ fois continûment dérivable au voisinage de t, le polynôme P_t est le développement de Taylor de la fonction f au point t. Clairement, $\mathcal{C}^s(t) \subset \mathcal{C}^{s'}(t)$ lorsque $s' \leqslant s$. Le

❋. Remarquons que toute fonction définie sur $\mathbb{R}^n \setminus \{t\}$ est la restriction d'une distribution de $\mathbb{R}^n$ appartenant à $\dot{B}^s_{\infty,\infty}$.

plus grand s tel que f soit hölderien d'exposant s en t traduit la régularité de la fonction en ce point.

Définition 2.42 L'*exposant de Hölder* $h(t;f)$ de f en t est la borne supérieure des s tels que f soit hölderien d'exposant s en t,

$$h(t;f) = \sup\{s : f \in \mathcal{C}^s(t)\}. \tag{2.61}$$

On omet souvent la référence à f en écrivant $h(t) = h(t;f)$. Une fonction ayant un exposant de Hölder constant est appelée *fonction mono-Hölder*.

Le fait que l'exposant de Hölder d'une fonction soit égal à h n'implique pas que cette fonction soit hölderienne d'exposant h. La fonction $|t|^h \log|t|$ appartient à $\mathcal{C}^s(t)$ pour tout $s < h$, mais pas à $\mathcal{C}^h(t)$.

Les fonctions dont l'exposant de Hölder est inférieur à 1 présentent un intérêt particulier : si f est hölderien d'exposant $0 < s < 1$ en t, alors f n'est pas dérivable en ce point et s caractérise le *type de singularité* de f en t. Dans ce cas, la relation (2.60) devient

$$|f(t+l) - f(t)| \leqslant C|l|^s. \tag{2.62}$$

Le résultat suivant découle directement des propositions exposées dans cette section.

Corollaire 2.43 *Si f appartient à $\mathcal{C}^s(t)$, alors il existe une constante C telle que*

$$|Wf(b,a)| \leqslant Ca^s(1 + \frac{|t-b|}{a})^s, \tag{2.63}$$

dans un voisinage de $(t,0^+)$. Inversement, si f appartient à $\dot{\mathcal{C}}^\varepsilon$ pour un exposant $\varepsilon > 0$ et si l'inégalité précédente est vérifiée, alors f appartient à $\mathcal{C}^{s-\delta}(t)$ pour tout δ tel que $0 < \delta \leqslant s$.

Donnons un exemple.

Exemple 2.44 Soient $\phi < 1$ et ω deux nombres positifs tels que $\phi\omega > 1$. La *fonction de Weierstraß* [182, 191]

$$\mathfrak{W}_{\phi,\omega}(t) = \sum_{j=1}^{+\infty} \phi^j \cos(\omega^j t), \tag{2.64}$$

n'est dérivable en aucun point. L'exposant de Hölder de $\mathfrak{W}_{\phi,\omega}$ est une fonction constante égale à $s = -\log\phi / \log\omega$ [210].

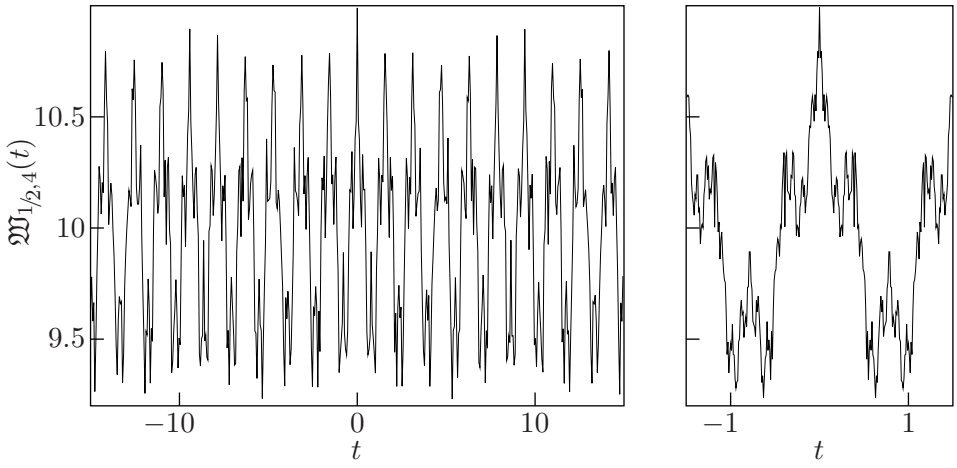

FIG. 2.5 – *La fonction de Weierstraß, avec* $\phi = 1/2$ *et* $\omega = 4$. *Son exposant de Hölder est donc* $1/2$.

Montrons d'abord que la fonction appartient à $\mathcal{C}^s(t)$ quel que soit t. Dans l'égalité

$$\mathfrak{W}_{\phi,\omega}(t+l) - \mathfrak{W}_{\phi,\omega}(t) = \sum_{j=1}^{+\infty} \phi^j \big(\cos\big(\omega^j(t+l)\big) - \cos(\omega^j t)\big),$$

où l est suffisamment petit, la différence entre les cosinus peut être majorée par 2 ou, grâce au théorème de la moyenne, par $\omega^j|l|$. Soit $j_0 = \lfloor -\log|l| / \log\omega \rfloor$; en utilisant l'une ou l'autre majoration selon que j est supérieur ou inférieur à j_0, on obtient

$$|\mathfrak{W}_{\phi,\omega}(t+l) - \mathfrak{W}_{\phi,\omega}(t)| \leqslant \sum_{j=1}^{j_0} \phi^j \omega^j |l| + 2\sum_{j=j_0}^{+\infty} \phi^j.$$

Par la définition même de j_0, chaque somme peut être majorée par une constante que multiplie $|l|^s$, ce qui prouve la régularité de $\mathfrak{W}_{\phi,\omega}$.

Pour l'irrégularité, on peut utiliser la proposition 2.43. Pour simplifier, choisissons une ondelette de classe C^2 telle que $[\psi] \subset [1/\omega, \omega]$, de manière à ne sélectionner qu'une fréquence et posons $\hat{\psi}(1) = 2$ par simple souci d'élégance. À l'échelle $a = 1/\omega^{j_0}$,

$$\begin{aligned} W_\psi \mathfrak{W}_{\phi,\omega}(b, \omega^{-j_0}) &= \sum_{j=1}^{+\infty} \phi^j \int_{\mathbb{R}} \cos(\omega^j t)\ \psi\Big(\frac{t-b}{\omega^{-j_0}}\Big)\ \frac{dt}{\omega^{-j_0}} \\ &= \sum_{j=1}^{+\infty} \phi^j \int_{\mathbb{R}} \cos(\omega^{j-j_0}\, t + \omega^j b)\ \psi(t)\ dt. \end{aligned}$$

Avec le choix particulier de $\hat{\psi}$, la seule intégrale non nulle correspond à $j = j_0$ et

$$W_\psi \mathfrak{W}_{\phi,\omega}(b, \omega^{-j_0}) = \phi^{j_0} \exp(i\omega^{j_0} b),$$

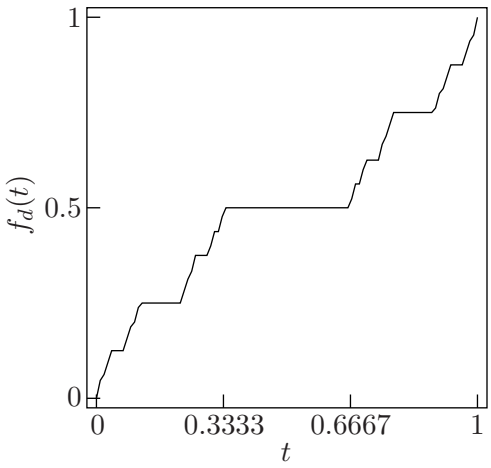

FIG. 2.6 – *L'escalier du diable.*

ce qui implique, par la proposition 2.43, que l'exposant de Hölder ne peut être plus grand que s, car on a

$$|W_\psi \mathfrak{W}_{\phi,\omega}(b, \frac{1}{\omega^{j_0}})| = (\frac{1}{\omega^{j_0}})^s.$$

□

Présentons l'*escalier du diable*, dont l'étude des propriétés de singularité se révèle élémentaire.

Exemple 2.45 Désignons par C l'ensemble de Cantor. Soit un nombre $t = \sum_j a_j/3^j$, $a_j \in \{0,1,2\}$, n'appartenant pas à C et j_0 le plus petit indice tel que $a_{j_0} = 1$. L'escalier du diable f_d est défini par

$$f_d : [0,1] \setminus C \to [0,1] \quad t = \sum_{j=1}^{+\infty} \frac{a_j}{3^j} \mapsto \sum_{j=1}^{j_0-1} \frac{a_j}{2^{j+1}} + \frac{1}{2^{j_0}}. \tag{2.65}$$

Cette fonction est définie presque partout sur [0,1] et peut être étendue par continuité sur tout l'ensemble [0,1] : si t appartient à C, on pose, avec les mêmes notations que précédemment,

$$f_d(t) = \sum_{j=1}^{+\infty} \frac{a_j}{2^{j+1}}. \tag{2.66}$$

L'escalier du diable est une fonction croissante d'exposant de Hölder $\log 2/\log 3$ sur C et ∞ sur le complémentaire. Pour simplifier les notations, nous écrirons, pour tout nombre t de l'intervalle [0,1],

$$f_d(t) = \sum_{j=1}^{j_0} \frac{a'_j}{2^j}, \tag{2.67}$$

avec $a'_j = a_j/2$ pour $j < j_0$, $a'_{j_0} = 1$ si $j_0 < \infty$ et où nous avons posé $j_0 = \infty$ si $t \in C$. Soient deux nombres différents t_1 et t_2 de C. Ils diffèrent pour un premier indice a_k. Comme $j_0 = \infty$, $|f_d(t_1) - f_d(t_2)| \leqslant 1/2^{k-1}$. De plus, $|t_1 - t_2| \leqslant 1/3^{k-1}$. Pour conclure, remarquons que les inégalités peuvent être des égalités. Pour les nombres n'appartenant pas à C, il suffit de les prendre assez proches (tels que $k > j_0$).

Puisque f_d est une fonction croissante de [0,1] vers [0,1], sa dérivée μ_d est une mesure de probabilité. De plus, le support de μ_d est C, puisque f_d est localement constant dans le complémentaire de C. Pour un sous-intervalle $[t,t+h]$ de [0,1], la régularité hölderienne de f_d permet d'écrire

$$\mu_d([t,t+l]) = f_d(t+l) - f_d(t) \leqslant 2|l|^{\log 2/\log 3}. \tag{2.68}$$

Il existe de nombreuses généralisations possibles [5, 41, 68, 74, 144, 332]. □

Nous allons maintenant reformuler le corollaire 2.43 de manière à caractériser l'exposant de Hölder d'une fonction en un point. Pour ce faire, introduisons la notation suivante.

Notation 2.46 Si f_1 et f_2 sont deux fonctions, nous écrirons $f_1 = \bar{\mathcal{O}}(f_2)$ si

$$\underline{\lim} \frac{\log|f_1|}{\log|f_2|} \geqslant 1, \tag{2.69}$$

et $f_1 = \widetilde{\mathcal{O}}(f_2)$ si l'inégalité dans la relation (2.69) devient

$$\lim \frac{\log|f_1|}{\log|f_2|} = 1. \tag{2.70}$$

L'égalité (2.70) signifie que les fonctions sont du même ordre de grandeur à une correction logarithmique près. Avec ces conventions, on peut énoncer le résultat suivant.

Corollaire 2.47 *Soit f une fonction appartenant à $\dot{\mathcal{C}}^\varepsilon$ pour un $\varepsilon > 0$. L'exposant de Hölder de f en t est h si et seulement si les conditions suivantes sont vérifiées,*

- *on a, au voisinage de $(t,0^+)$,*

$$|Wf(b,a)| = \bar{\mathcal{O}}(a^h + |t-b|^h), \tag{2.71}$$

- *il existe une suite $\{(b_j,a_j)\}$ convergeant vers $(t,0^+)$ telle que*

$$|Wf(b,a)| = \widetilde{\mathcal{O}}(a_j^h + |t-b_j|^h). \tag{2.72}$$

Une suite $\{(b_j,a_j)\}$ vérifiant l'égalité (2.72) est appelée *suite de minimisation.*

Singularités oscillantes

Les exposants de Hölder ponctuels présentent une propriété peu désirable : l'instabilité par rapport à la dérivation et vis-à-vis des opérateurs pseudo-différentiels classiques (même des opérateurs très simples, tels la transformée de Hilbert [211]). En fait, cet exposant ne permet pas de décrire tous les types de singularité de manière satisfaisante. Il faut pour cela introduire un second exposant. Ce sujet, reposant en partie sur les travaux de Tchamitchian et Torresani [361], est notamment traité dans les références suivantes [22, 23, 24, 29, 32, 38, 99, 206, 214, 285].

Contrairement aux espaces de Hölder, les espaces de Hölder ponctuels définis au paragraphe précédent ne sont pas stables par dérivation. Si une fonction f appartient à $\mathcal{C}^s(t)$ pour un certain s, rien n'implique que $\partial^m f$ appartienne à $\mathcal{C}^{s-m}(t)$. Un exemple classique est donné par la fonction $t^2 \sin(1/t)$: elle appartient à $\mathcal{C}^2(0)$, mais sa dérivée n'est pas continue à l'origine et n'est donc pas dans $\mathcal{C}^1(0)$. L'exposant de Hölder ponctuel n'est lui-même pas stable en ce qui concerne la l'intégration. La fonction $f(t) = t \sin(1/t)$ possède un exposant de Hölder en zéro égal à $h(0) = 1$, alors que sa primitive a pour exposant 3.

L'étude du comportement de l'exposant de Hölder des primitives de fonctions possédant une composante oscillante du type $\sin(1/t)$ permet de mieux caractériser les singularités associées. La primitive de la fonction $t^s \sin(1/t^\beta)$ (nous choisirons toujours la primitive s'annulant à l'origine) possède un exposant de Hölder ponctuel en zéro égal à $h_1(0) = s + \beta + 1$. En fait, la m-ième primitive de cette fonction possède un exposant de Hölder donné par $h_m(0) = s + m(\beta + 1)$. Ces considérations mènent à la notion de chirp.

Définition 2.48 Une fonction $f \in L^\infty_{\mathrm{loc}}(\mathbb{R})$ est un *chirp* de type (α,β) si

$$f_m \in \mathcal{C}^{\alpha+m(1+\beta)}(t) \quad \forall\, m \in \mathbb{N}, \tag{2.73}$$

où f_m désigne une primitive d'ordre m de f. Si $h_m(t)$ est l'exposant de Hölder de la m-ième primitive de f, l'*exposant de chirp* de f est donné par

$$\beta_{\mathrm{c}}(t) = \underline{\lim}_{m\to\infty} \frac{h_m(t)}{m} - 1. \tag{2.74}$$

La définition (2.74) met bien en évidence le caractère oscillant de la fonction $t^\alpha \sin(1/t^\beta)$ en faisant apparaître le terme β. Cependant, cet exposant ne permet pas de décrire tous les comportements oscillants comme on le voudrait. La fonction $t \sin(1/t^\beta) + |t|^{3/2}$ possède un exposant de Hölder égal à un à l'origine [23]. Pour $\beta \neq 0$ et $m \geqslant 1$, on a $1 + m(\beta+1) > 3/2 + m$ si m est tel que $m\beta > 1/2$ et donc $h_m(0) = 3/2 + m$, ce qui donne $\beta_{\mathrm{c}}(0) = 0$ et non β.

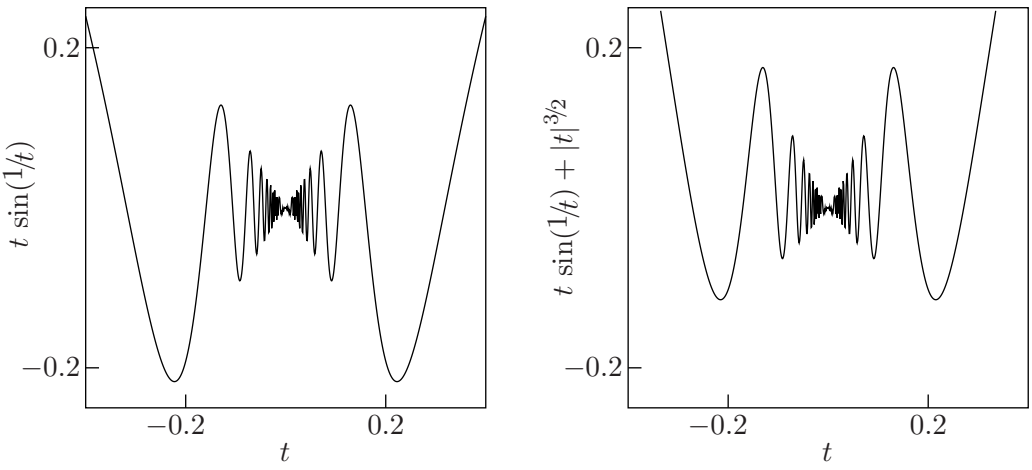

FIG. 2.7 – *Deux fonctions du type* $t \sin(t^{-\beta})$ *et* $t \sin(t^{-\beta}) + |t|^{\alpha}$ *respectivement. La première est bien caractérisée par la notion de chirp. Pour la seconde, il faut utiliser l'exposant d'oscillation.*

Il faut raffiner cette définition en tenant compte du comportement de la fonction pour des intégrations fractionnaires.

Notation 2.49 Étant donné une fonction f de l'espace L^{∞}_{loc}, nous noterons f_r la primitive fractionnaire d'ordre r de cette dernière. Plus précisément, on pose $f_r = (I - \Delta)^{-r/2}(\phi f)$, où ϕ est une fonction de D telle que $\phi(t) = 1$ et où Δ représente le laplacien*. L'exposant de Hölder associé à f_r sera, quant à lui, noté $h_r(t)$.

Pour la fonction $f(t) = t \sin(1/t^{\beta}) + |t|^{3/2}$, f_r a pour exposant de Hölder $1 + r(\beta + 1)$ à l'origine lorsque r est suffisamment petit et non pas $1 + r$ [23]. Ainsi, avec cette manière de procéder, on peut faire apparaître le β dans l'évolution de l'exposant. La définition relative à ces développements est sous-tendue par le résultat suivant.

Proposition 2.50 *Étant donné* $t \in \mathbb{R}^n$*, la fonction* $r \mapsto h_r(t)$ *est concave et sa dérivée à droite en zéro existe si* $h(t) < \infty$*.*

Ainsi, la définition suivante est licite.

Définition 2.51 L'*exposant d'oscillation* de f en $t \in \mathbb{R}^n$ est donné par

$$\beta(t) = [\partial_r h_r(t)]_{r=0^+} - 1, \tag{2.75}$$

où $[\partial_r h_r(t)]_{r=0^+}$ représente la dérivée à droite au point 0 de la fonction $h_r(t)$ par rapport à r. Par extension, nous dirons que les exposants d'oscillation de f en t sont donnés par

∗. L'opérateur $(I - \Delta)^{-r/2}$ est donc l'opérateur de convolution consistant à multiplier la transformée de Fourier de la fonction par $(1 + |\omega|^2)^{-r/2}$.

le couple $\big(h(t),\beta(t)\big)$.

Ces exposants prennent leurs valeurs dans $[0,\infty]\times[0,\infty]$ et si $h(t)=\infty$, l'exposant associé $\beta(t)$ n'est pas défini.

On peut donner une interprétation à l'exposant $\beta(t)$; la suite de minimisation est définie par le corollaire 2.47.

Proposition 2.52 *On peut définir l'exposant d'oscillation $\beta(t)$ de f de la manière suivante,*

$$\beta(t)=\sup\big\{0,\inf\{\beta:\exists \text{ une suite de minimisation pour } f \text{ dans } \Lambda(\beta)\}\big\}, \tag{2.76}$$

où $\Lambda(\beta)=\{(b,a):a\geqslant|t-b|^{\beta+1}\}$.

Nous nous attarderons plus longuement sur ce résultat dans la prochaine section.

Pratiquement, le résultat suivant permet de déterminer les exposants d'oscillation d'une fonction en un point. Nous utilisons les notations 2.46.

Corollaire 2.53 *Si f appartient à $\dot{\mathcal{C}}^{\varepsilon}$, pour un $\varepsilon>0$, alors les exposants de singularité de f en t sont (h,β) si et seulement si la transformée en ondelettes de f satisfait les conditions suivantes*

- *on a, au voisinage de $(t,0^+)$*
$$|Wf(b,a)|=\bar{\mathcal{O}}(a^h+|t-b|^h) \text{ avec }, \tag{2.77}$$
- *il existe une suite $\{(b_j,a_j)\}$ convergeant vers $(t,0^+)$ telle que*
$$|Wf(b_j,a_j)|=\widetilde{\mathcal{O}}(a_j^h+|t-b_j|^h) \tag{2.78}$$
et
$$(a_j+|t-b_j|)^{1+\beta}=\widetilde{\mathcal{O}}(a_j), \tag{2.79}$$
- *β est la borne inférieure des nombres pour lesquels la relation (2.79) est vérifiée.*

Nous donnerons une interprétation de ce résultat dans la prochaine section.

2.4 Étude de la régularité d'une fonction par la transformée en ondelettes

Nous possédons maintenant tous les outils pour pouvoir caractériser les singularités isolées d'une fonction grâce aux ondelettes. Le problème est donc, étant donné une fonction f

vérifiant certaines conditions, de pouvoir déterminer sa régularité en un point quelconque, autrement dit, son exposant de Hölder (et éventuellement son exposant d'oscillation) en chaque point. Pour simplifier les développements, nous supposerons que le signal étudié appartient à l'espace $L^2(\mathbb{R})$.

Remarques concernant la mesure de la régularité d'une fonction

Nous faisons ici deux remarques concernant l'étude pratique de singularités d'un signal. La première montre les effets indésirables qui peuvent se manifester si l'ondelette mère utilisée ne possède pas suffisamment de moments nuls. La seconde introduit le cône d'influence associé à une ondelette et discute de la position des lignes de maxima relatives à une singularité oscillante.

Une ondelette mère ne respectant pas les hypothèses 2.14 ne peut donner toute l'information sur la régularité hölderienne d'une fonction. Soit f une fonction uniformément hölderienne définie sur $\mathbb{R}$ et d'exposant h supérieur au nombre de moments nuls m d'une ondelette ψ. La proposition 2.12 permet d'écrire

$$\lim_{a\to 0} \frac{Wf(b,a)}{a^m} = \lim_{a\to 0} \frac{1}{a} D^m[f * \theta(-\frac{\cdot}{a})](b) = \lim_{a\to 0} \int_{\mathbb{R}} D^m f(b+at)\theta(t)\ dt. \tag{2.80}$$

La fonction θ étant 0-régulière, le théorème de convergence dominée permet d'affirmer qu'il existe une constante C telle que

$$\lim_{a\to 0} \frac{Wf(b,a)}{a^m} = CD^m f(b). \tag{2.81}$$

Pour une telle fonction, le comportement aux petites échelles vérifie l'inégalité (2.50), avec $s = m < h$, malgré la plus grande régularité de f.

L'exposant de Hölder permet de caractériser les singularités du type $|t|^h$, appelées *cusp*, où l'exposant d'oscillation β est nul, mais pas les singularités oscillantes, du type $|t|^h \sin(1/t^\beta)$. Soit f une fonction hölderienne d'exposant s en t dont on veut étudier la régularité par une ondelette mère ψ de support compact, $[\psi] = [-b_0,b_0]$. À une échelle a donnée, la transformée en ondelettes $W_\psi f(b,a)$ dépend de la valeur de $f(t)$ si $|t-b| \leqslant ab_0$. L'ensemble des points (b,a) vérifiant cette condition est appelé *cône d'influence* de t. En pratique, pour les ondelettes rapidement décroissantes, on peut toujours définir un cône d'influence effectif, déterminant les points influançant numériquement $f(t)$. Les valeurs de la transformée en ondelettes au voisinage de t permettent d'obtenir, grâce à au corollaire 2.43, l'exposant de Hölder $h(t)$ de f en t. En ce qui concerne les singularités oscillantes,

le corollaire 2.53 nous apprend qu'il est nécessaire de considérer des points en dehors du cône : pour ces singularités, les fortes valeurs de la transformée en ondelettes se situent sur des « crêtes* » d'équation du type $a = C|b-t|^{1+\beta}$ ($\beta > 0$) dans le demi-plan espace-échelle, en dehors de tout cône d'influence.

Détection de singularités isolées dans un signal

Nous pouvons maintenant donner un algorithme pratique de détection et de caractérisation des singularités de type « cusp ». Avec une légère modification dans l'approche, on peut aussi détecter des variation douces, où la fonction est infiniment continûment dérivable. Nous supposerons ici que les singularités rencontrées peuvent être étudiées en ne considérant que les points situés dans le cône d'influence. Les signaux envisagés ne comportent donc pas de singularité oscillante.

Sous des hypothèses de régularité globale suffisantes, le corollaire 2.47 affirme que les irrégularités ponctuelles peuvent être caractérisées par l'étude des maxima du module de la transformée en ondelettes aux petites échelles. En supposant que les ondelettes utilisées sont du type (2.24), il suffit d'étudier les lignes de maxima du module convergeant vers la singularité pour pouvoir déterminer l'exposant de Hölder ponctuel [258]. En pratique, la relation (2.72) conduit à évaluer l'exposant h *via* une régression linéaire. Supposons que les hypothèses de la proposition 2.47 et la relation (2.71) soient vérifiées. Pour une ondelette du type (2.24), à chaque ligne du maxima du module convergeant vers $(t,0^+)$, associons la fonction f_ℓ, exprimant le logarithme du module de la transformée en ondelettes le long de cette ligne en fonction du logarithme de l'échelle,

$$f_\ell(t) = \log_2 |Wf(\ell(2^t),2^t)|. \tag{2.82}$$

S'il s'agit d'une ligne de maxima du module définie par une suite du type (2.72), une régression linéaire sur f_ℓ devrait permettre d'estimer h en évaluant la pente de cette fonction aux petites échelles a.

Exemple 2.54 Le signal $f(t) = \exp(-(t-500)^2\, 10^4)\chi_{[0,1300]}(t)+\sqrt{|t-2500|}\chi_{]1300,3000]}(t)$ présente deux types de singularité. Le premier est une discontinuité (en $t = 1300$) et le second résulte d'un comportement du type $\sqrt{|t|}$ (en $t = 2500$). Ainsi, $f \in \mathcal{C}^0(1300) \cap \mathcal{C}^{1/2}(2500)$. La transformée en ondelettes de ce signal, avec l'ondelette mère dérivée première de la gaussienne ψ_1, fait apparaître cinq lignes de maxima du module (figure 2.8). Les deux premières lignes pointent sur les changements de concavité de l'exponentielle. En effectuant une régression linéaire selon la méthode définie par l'égalité (2.82), on mesure

*. Traduisant la variation de la fréquence instantanée [117].

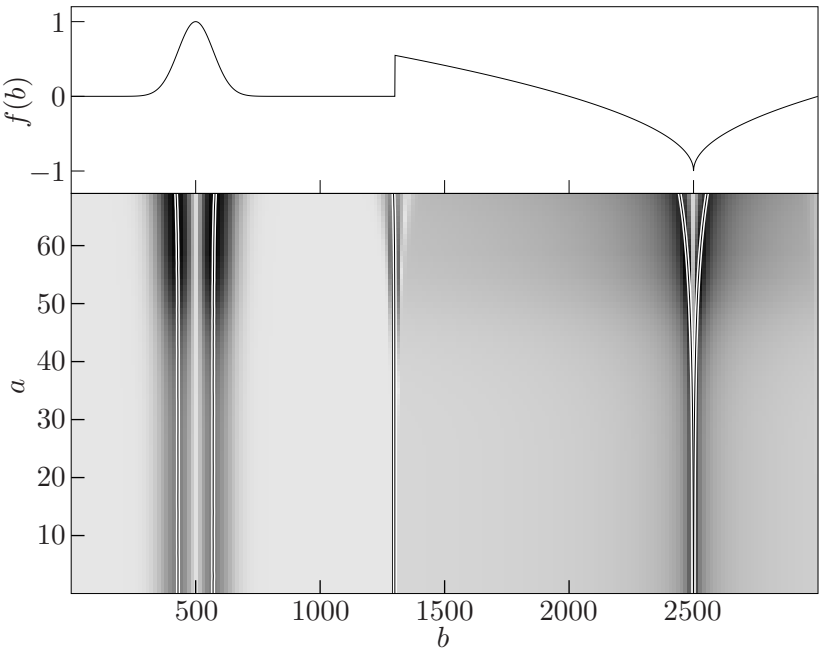

FIG. 2.8 – *Le signal $f(t)$ défini dans l'exemple 2.54 et la représentation du module de sa transformée en ondelettes en utilisant l'ondelette mère dérivée première de la gaussienne ψ_1. Les couleurs utilisées vont du blanc (pour les valeurs les plus faibles) au noir (pour les valeurs les plus fortes). Les lignes de maxima du module, représentées par des lignes noires sur fond blanc, sont surimposées à la représentation de la transformée.*

un coefficient angulaire égal à 1. Vu que ψ_1 ne possède qu'un seul moment nul, on pourrait en conclure que l'exposant de Hölder de f est un. Toutefois, la transformée en ondelettes avec la dérivée seconde de la gaussienne ψ_2 donne une pente égale à 2. D'une manière générale, en utilisant l'ondelette dérivée m-ième de la gaussienne ψ_m, la pente trouvée est égale à m, ce qui confirme que la fonction est infiniment continûment dérivable en ce point. La troisième ligne pointe sur la discontinuité. Une régression linéaire aux petites échelles permet de trouver une pente proche de $s = 0$. Enfin, les deux dernières lignes pointe sur la singularité du type $\sqrt{|t|}$. Pour chacune d'entre elles, on mesure un coefficient angulaire égal à $s = 1/2$. La figure 2.9 illustre le comportement du module des valeurs de la transformée en ondelettes le long des lignes de maxima en fonction de l'échelle. □

Cette méthode peut être adaptée pour l'analyse de *variations douces* [259] modélisées par l'égalité suivante. Supposons qu'une fonction f soit de la forme

$$f = f_s * f_\sigma, \tag{2.83}$$

au voisinage d'un point $t_0 \in \mathbb{R}$, où f_s est hölderien d'exposant s dans ce voisinage et f_σ

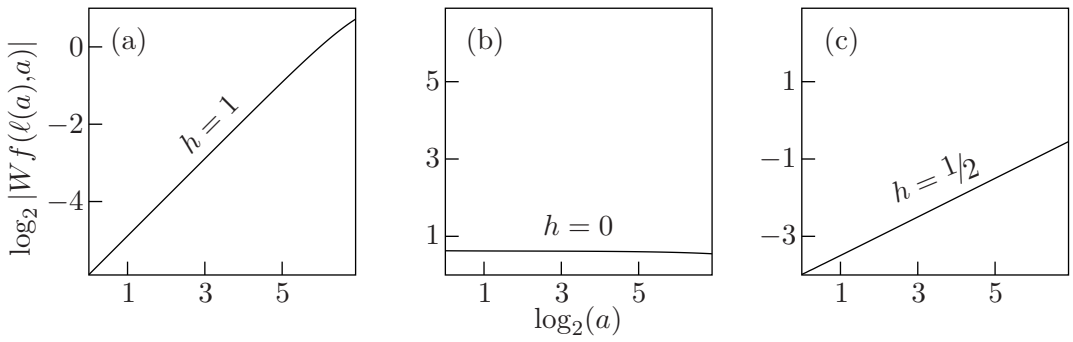

FIG. 2.9 – *Comportement du module des valeurs de la transformée en ondelettes le long des lignes de maxima du module de cette transformée. (a) La première pointe sur le changement de concavité de l'exponentielle, $h = m = 1$; (b) la deuxième sur la discontinuité, $h = 0$; (c) la troisième sur la singularité du type $\sqrt{|t|}$, $h = 1/2$.*

est une gaussienne d'écart-type σ,

$$f_\sigma(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-t^2}{2\sigma^2}). \tag{2.84}$$

Le résultat suivant permet d'étudier la singularité de f_s.

Proposition 2.55 *Soit $\psi(t) = (-1)^m D_t^m \exp(-t^2/2\sigma_\psi^2)$. Si f peut s'écrire sous la forme (2.83), alors il existe une constante C telle que*

$$|Wf(b,a)| \leqslant Ca^s(1 + \frac{\sigma}{\sigma_\psi a})^{s-m}, \tag{2.85}$$

dans un voisinage de $(t_0,0^+)$.

Ainsi, aux grandes échelles par rapport à σ/σ_ψ, la transformée en ondelettes « décroît » comme a^s. Par contre aux plus petites échelles, on observe une décroissance en a^m à cause de la plus grande régularité de f. La fonction se comporte donc comme une fonction singulière en t_0 à grande échelle, alors que l'on peut constater la régularité de f au voisinage de ce point si l'on considère les petites échelles. Traitons un exemple.

Exemple 2.56 Considérons la fonction $f = \sqrt{|\,.\,|} * f_{\sigma=1}$, où $f_{\sigma=1}$ est la gaussienne définie par la relation (2.84). La transformée en ondelettes de ce signal par l'ondelette mère dérivée de la gaussienne ψ_1 présente deux lignes de maxima du module au voisinage du minimum de f (figure 2.10). Si l'on représente ces lignes comme une fonction de l'échelle suivant l'égalité (2.82) (*cf.* figure 2.11), pour les petites échelles, une régression linéaire nous permet d'estimer la pente égale à 1, le nombre de moments nuls de ψ_1. Pour les échelles plus grandes, elle vaut $1/2$, et permet donc d'estimer l'exposant de Hölder de la singularité lissée. L'inconvénient majeur de cette méthode est qu'il faut recourir à des

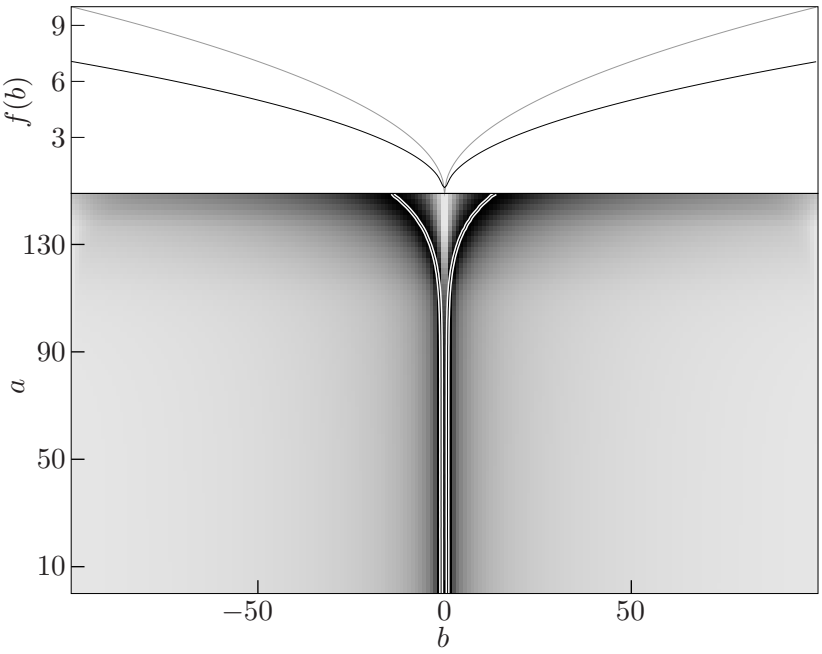

FIG. 2.10 – *Le signal $f = \sqrt{|.|} * f_{\sigma=1}$ et la représentation du module de sa transformée en ondelettes avec l'ondelette mère ψ_1. Les couleurs utilisées vont du blanc (pour les valeurs les plus faibles) au noir (pour les valeurs les plus fortes). En gris est représentée la fonction $\sqrt{|t|}$. Les lignes de maxima du module (lignes noires sur fond blanc) permettent d'estimer l'exposant de Hölder de la singularité lissée.*

valeurs de la transformée en ondelettes pour de grandes échelles, ce qui implique que la singularité ne soit pas proche, relativement à la taille de l'ondelette, d'autres singularités qui viendraient perturber le comportement de la transformée à proximité du point étudié.

□

Ces considérations sont à mettre en relation avec le phénomène de dissipation en turbulence pleinement développée.

Remarque 2.57 La turbulence pleinement développée [47, 154, 362] concerne l'étude des écoulements turbulents (supposés incompressibles) aux très grands nombres de Reynolds (c'est-à-dire pour une faible viscosité). Selon les hypothèses de Kolmogorov (K41) [225], le taux de transfert de l'énergie cinétique ϵ, représentant le taux auquel l'énergie est transférée des grandes vers les petites échelles, est constant lorsque le nombre de Reynolds tend vers l'infini. Cette hypothèse fut rapidement remise en cause par les observations expérimentales (où les nombres de Reynolds sont bien sûr finis) et depuis lors, nombre de travaux ont été consacrés à la description des fluctuations du taux de transfert de l'énergie cinétique. MANDELBROT [260] réunifia diverses approches en proposant un modèle général

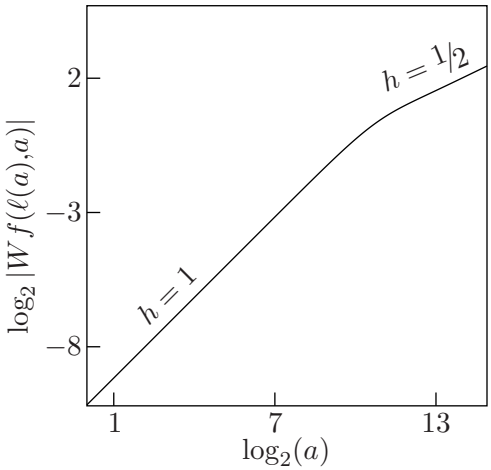

FIG. 2.11 – *Comportement du module des valeurs de la transformée en ondelettes de f le long des lignes de maxima du module (en utilisant ψ_1 comme ondelette mère) en fonction du logarithme de l'échelle. On distingue clairement deux comportements linéaires. Le premier, à petite échelle, exprime le comportement oscillant de l'ondelette mère et le second, à plus grande échelle, rend compte de la singularité lissée présente au voisinage de ces lignes.*

de cascade multiplicative[*] et en introduisant la notion de mesure fractale. Dans la plupart des modèles proposés, y compris ceux postérieurs aux travaux de MANDELBROT, l'énergie contenue dans les tourbillons à grande échelle va, par fractionnements successifs, être transférée dans des tourbillons de plus en plus petits jusqu'à atteindre ce que l'on appelle l'échelle de dissipation[⁂] η. Pour des échelles inférieures à η, il n'y a plus d'invariance d'échelle et les singularités présentes dans un signal turbulent « disparaissent » car lissées par les effets visqueux, le signal à ces échelles étant régulier. Il existe une analogie évidente entre ce phénomène et la proposition 2.55. □

2.5 Formalisme multifractal pour les fonctions

Pour les fonctions hautement irrégulières, vouloir analyser localement les singularités n'a plus beaucoup de sens. Le formalisme multifractal permet de caractériser globalement ces singularités en quantifiant leur importance. Les fondements d'une telle démarche ont été introduits bien avant leur justification mathématique. Les analogies avec le formalisme multifractal pour les mesures sont flagrantes. Nous introduisons ici les diverses méthodes actuelles d'estimation en les justifiant et donnant leurs limites. Les concepts de base et la

[*]. Les modèles de cascades multiplicativess ont notamment été introduits par NOVIKOV & STEWART [303] et YAGLOM [394].

[⁂]. Cette échelle est aussi appelée échelle de Kolmogorov. Selon K41, si Re représente le nombre de Reynolds, η se comporte comme $Re^{-3/4}$.

méthode des fonctions de structure ont étés introduits par Parisi et Frisch [309]. Les résultats mathématiques exposés ici concernant le formalisme multifractal sont principalement dûs à Jaffard [203, 204, 214]. Le formalisme multifractal basé sur la transformée en ondelette continue, la méthode des maxima du module de la transformée en ondelettes et les considérations sur le nombre de moments nuls se trouvent dans les travaux de l'équipe d'Arneodo (voir les références [18, 19, 28, 185] pour la méthode transformée en ondelettes intégrale et [17, 25, 37, 292, 293, 294] pour la méthode des maxima du module de la transformée en ondelettes). Le formalisme multifractal grand canonique, prenant en compte les singularités cusp et oscillantes, est exposé dans les références [22, 24]. Enfin, nous tenons à remercier Alain Arneodo et Philippe StJean pour les discussions que nous avons tenues concernant la caractérisation des singularités oscillantes *via* le formalisme multifractal.

Spectre de Hölder et méthodes d'estimation

Nous introduisons ici la notion de spectre de Hölder d'une fonction irrégulière et donnons deux méthodes d'estimation de ce spectre. Nous en introduirons de nouvelles dans la suite. En toute généralité, le formalisme multifractal ne fournit qu'une borne supérieure du spectre de Hölder.

Une caractérisation globale des singularités peut se faire en considérant la fonction suivante.

Définition 2.58 Le *spectre de Hölder* (aussi appelé *spectre de singularités*) d'une fonction $f \in \mathcal{C}^\varepsilon$, pour un $\varepsilon > 0$, est la fonction définie par

$$d(h; f) = \dim_{\mathcal{H}}(\{t : h(t) = h\}). \tag{2.86}$$

En général, on omet la référence à f en écrivant $d(h) = d(h; f)$. Nous nommerons fonction *monofractale* toute fonction f n'admettant qu'un seul exposant de Hölder fini, *i.e.* pour laquelle il n'existe qu'une seule valeur h finie telle que $d(h) \neq -\infty$.

On peut distinguer de types de fonctions monofractales.

Remarque 2.59 Il y a lieu de faire remarquer qu'une fonction monofractale peut présenter des exposants de Hölder $h = \infty$. Rappelons qu'une fonction présentant un seul exposant de Hölder est appelée fonction mono-Hölder [210]. Ainsi, l'escalier du diable (*cf.* exemple 2.45) est une fonction monofractale et la fonction de Weierstraß (*cf.* exemple 2.44) est une fonction mono-Hölder. □

L'escalier du diable est un exemple simple où le spectre multifractal peut être déterminé analytiquement.

Exemple 2.60 Selon l'exemple 2.45, l'exposant de Hölder de l'escalier du diable vaut $\log 2/\log 3$ aux points appartenant à l'ensemble de Cantor (voir l'exemple 1.16) et l'infini ailleurs. La dimension de Hausdorff de l'ensemble de Cantor et de l'ensemble complémentaire dans [0,1] étant $\log 2/\log 3$ et 1 respectivement, l'escalier du diable possède donc le spectre d de support $\{\log 2/\log 3\} \cup \{\infty\}$ suivant,

$$d(h) = \begin{cases} \log 2/\log 3 & \text{si } h = \log 2/\log 3, \\ 1 & \text{si } h = \infty. \end{cases} \tag{2.87}$$

□

De manière analogue au formalisme multifractal pour les mesures, l'estimation du spectre par une transformée de Legendre inverse donne lieu à ce que l'on appelle le *formalisme multifractal pour les fonctions*. Dans un cadre général, ce dernier ne peut être directement lié au premier. Toutefois, si μ est une mesure de probabilité définie sur [0,1], en posant $f(t) = \mu([0,t])$, on montre sans difficulté que les deux formalismes peuvent être unifiés dans ce cas précis⁎.

Une méthode d'estimation du spectre fût d'abord proposée par Parisi et Frish. Cette dernière, appelée *méthode des fonctions de structure* repose sur le raisonnement heuristique suivant. Soit f la fonction dont on veut estimer le spectre de Hölder. Pour une singularité d'exposant h, il y a lieu de penser que, pour l assez petit, la quantité $|f(t+l) - f(t)|^q$ va se comporter comme $|l|^{hq}$. De plus, les singularités d'exposant h devraient être au nombre de $|l|^{-d(h)}$, chacune contribuant pour un volume $|l|^n$. Dans l'intégrale $\int_{\mathbb{R}^n} |f(t+l) - f(t)|^q \, dt$, la plus grande contribution est donnée par le terme de plus petit exposant, que l'on appellera ζ. Donc, en postulant $\zeta(q) = \inf_h\{hq - d(h)\} + n$, nous pouvons espérer trouver le spectre en calculant une transformée de Legendre inverse, $d(h) = \inf_q\{hq - \zeta(q)\} + n$. Remarquons que si d n'est pas concave⁕, l'infimum ne peut donner que l'enveloppe concave du spectre réel. Ainsi, nous devons supposer, et ce sera le cas pour toutes les méthodes envisagées, que le spectre de Hausdorff est une fonction concave*.

⁎. Ainsi, si E_α désigne l'ensemble intervenant dans la définition du spectre multifractal de Hausdorff (1.60), on constate que $t \in E_\alpha$ peut se lire $|f(t+l) - f(t)|$ tend vers $|l|^\alpha$, à une correction logarithmique près, ce qui est à mettre en parallèle avec la relation (2.89).

⁕. Une application de la formule de Young permet de montrer que ζ est toujours concave [330].

*. Signalons que les espaces S^ν récemment introduits [30, 31, 121] permettent de s'affranchir de l'hypothèse de concavité dans le calcul du spectre. Cependant, il n'existe à ce jour aucune mise en oeuvre de ce formalisme.

Suivant ces développement, on définit la *fonction de partition*,

$$S(l,q) = \int_{\mathbb{R}^n} |f(t+l) - f(t)|^q \, dt, \tag{2.88}$$

où $f \in L^q$ est une fonction à valeur réelles. Ceci étant, on pose

$$\zeta(q) = \underline{\lim}_{l \to 0} \frac{\log S(l,q)}{\log |l|}. \tag{2.89}$$

On estime le spectre de Hölder en calculant*

$$d(h) = \inf_q \{qh - \zeta(q)\} + n. \tag{2.90}$$

Une autre manière de procéder dans l'estimation du spectre consiste à remplacer la fonction de structure par une intégrale sur la transformée en ondelettes. Au vu du corollaire 2.43, pour une singularité d'ordre h, la quantité $|Wf(b,a)|^q$ devrait se comporter comme $|a|^{hq}$ pour les petites échelles a, aux positions b voisines de cette singularité. On pose donc

$$\tilde{Z}(a,q) = \int_{\mathbb{R}^n} |Wf(b,a)|^q \, db, \tag{2.91}$$

et on définit

$$\tilde{\eta}(q) = \underline{\lim}_{a \to 0} \frac{\log \tilde{Z}(a,q)}{\log a}, \tag{2.92}$$

pour estimer le spectre en calculant

$$d(h) = \inf_q \{qh - \tilde{\eta}(q)\} + n. \tag{2.93}$$

Cette méthode, appelée *méthode transformée en ondelettes intégrale*, devrait être plus robuste à la présence de bruit dans le signal étudié, puisque l'on moyenne le signal *via* la transformée en ondelettes.

Lorsque $q > 0$, ces méthodes d'estimation ne fournissent qu'une majoration du spectre de Hölder. Pour le montrer, donnons d'abord la définition de l'espace de Nikol'skiĭ H_p^s [4, 302].

Définition 2.61 Soit $s \geqslant 0$ et $p \geqslant 1$. L'espace de Nikol'skiĭ H_p^s est défini comme suit,

$$H_p^s = \{f \in S' : \|f\|_{H_p^s} = \|f\|_{L^p} + \sup_{|l|>0} \frac{\|\Delta_l^k \nabla^{\lfloor s \rfloor^-} f\|_{L^p}}{|l|^{s-\lfloor s \rfloor^-}} < \infty\}, \tag{2.94}$$

où ∇ désigne le gradient et $k \in \{1,2\}$ est tel que $k > s - \lfloor s \rfloor^-$.

*. Nous reviendrons plus tard sur le cas où la dimension du support du spectre de Hölder n'est pas égale à n (*cf.* relation (2.101)).

Si s n'est pas un nombre entier, $f \in H_p^s$ si $f \in L^p$ et si, pour tout multi-indice α tel que $|\alpha| = \lfloor s \rfloor$,

$$\int_{\mathbb{R}^n} \frac{|\partial^\alpha f(t+l) - \partial^\alpha f(t)|^p}{|l|^{(s-\lfloor s \rfloor)p}}\, dt \leqslant C, \tag{2.95}$$

pour une constante C. Comme $\zeta(q)$ est la limite des nombre ξ tels que $S(l,q) \leqslant C|l|^\xi$, pour les petites valeurs de l, il est normal de redéfinir ζ comme suit,

$$\zeta'(q) = \sup\{s : f \in H_{q,\mathrm{loc}}^{s/q}\}. \tag{2.96}$$

Lorsque $q > 1$ et $\zeta(q) < q$, on a $\zeta(q) = \zeta'(q)$. Dans le cas contraire, il convient d'utiliser les différences d'ordre deux de f dans la définition (2.88) de S. Dans la suite, nous supposerons que la méthode a été modifiée comme indiqué. Si ce n'est fait, seule une portion de la partie croissante du spectre pourra être déterminée. Nous allons maintenant obtenir une relation du même type pour $\tilde{\eta}$. Par la proposition 2.33, $f \in \dot{B}_{q,\infty}^s$ si et seulement si

$$\| \frac{1}{a^s} \|Wf(\,.\,,a)\|_{L^q} \|_{L^\infty(\mathbb{R}_*^+)} < \infty. \tag{2.97}$$

Puisque $\tilde{\eta}(q)$ est l'infimum de ξ vérifiant $\tilde{Z}(a,q) \leqslant Ca^\xi$ pour a suffisamment petit, on a

$$\tilde{\eta}(q) = \{s : f \in \dot{B}_{q,\infty,\mathrm{loc}}^{s/q}\}, \tag{2.98}$$

lorsque $q > 0$. Pour montrer que ces méthodes coïncident, nous utiliserons le résultat suivant. Il peut être prouvé en utilisant divers théorèmes d'inclusion [4, 214].

Proposition 2.62 *Si $q \geqslant 1$, alors*

$$H_q^{s+\varepsilon} \subset \dot{B}_{q,\infty}^s \subset H_q^{s-\varepsilon}, \tag{2.99}$$

pour tout $\varepsilon > 0$.

On a alors le résultat voulu, affirmant l'équivalence des deux méthodes pour la plupart des valeurs de q.

Corollaire 2.63 *Si $q > 1$, $\zeta'(q) = \tilde{\eta}(q)$.*

Les méthodes exposées ne peuvent permettre d'obtenir le spectre de Hausdorff en toute généralité, mais fournissent une borne supérieure.

Théorème 2.64 *Si $s > n/q$, avec $q > 0$ et $f \in \dot{B}_{q,\infty,loc}^s$, alors $d(h) \leqslant n - (s-h)q$. En particulier, si $\tilde{\eta}$ est tel que $\tilde{\eta}(q) > n$ quel que soit q,*

$$d(h) \leqslant \inf_q \{qh - \tilde{\eta}(q)\} + n. \tag{2.100}$$

Terminons par quelques remarques. Il n'est pas envisageable d'étendre la fonction $\tilde{\eta}(q)$ aux valeurs de q négatives. La caractérisation (2.97) pour de tels q n'a pas beaucoup de sens. Il est clair que le calcul de $\tilde{Z}(a,q)$ pour de telles valeurs est hautement instable. La méthode des maxima du module de la transformée en ondelettes permettra de contourner ces difficultés. Les espaces d'oscillation autoriserons aussi l'accès aux valeurs de q négatives, en proposant une version « raisonnable » d'espaces généralisant les espaces de Besov $B^s_{q,\infty}$ pour tout indice q réel. Dans la formule (2.93) permettant d'obtenir le spectre de Hausdorff, ou plutôt son enveloppe concave, le terme n provient de la contribution en volume de l'erreur. Si la dimension de Hausdorff $\dim_{\mathcal{H}}[d]$ du support du spectre n'est pas égale à n, il est naturel de remplacer (2.93) par la formule

$$d(h) = \inf_q \{qh - \tilde{\eta}(q)\} + \dim_{\mathcal{H}}[d]. \tag{2.101}$$

Le problème est que cette dimension n'est pas connue *a priori* , ce qui constitue clairement un handicap lorsqu'elle ne vaut pas n. Ici aussi, la méthode des maxima du module de la transformée en ondelettes palliera ce problème. Enfin mentionnons que l'égalité dans la relation (2.100) est vérifiée pour de nombreuses fonctions [53, 207], comme les fonctions auto-similaires que nous allons maintenant présenter.

Fonctions auto-similaires et méthode des maxima du module de la transformée en ondelettes

Pour le cas particulier des fonctions auto-similaires, les méthodes proposées dans la section précédente sont exactes, dans le sens où elles permettent d'obtenir la partie croissante du spectre de Hölder. Dans cette section, nous présentons également la méthode des maxima du module de la transformée en ondelettes qui donnera accès à l'entièreté du spectre.

Commençons par introduire la notion de fonction auto-similaire, basée sur l'analogie avec les mesures auto-similaires (voir section 1.3). Supposons d'abord que f soit une fonction continue de support compact. Soit Ω l'ensemble ouvert borné de $\mathbb{R}^n$ tel que $\bar{\Omega} = [f]$. Intuitivement, pour une fonction auto-similaire, il devrait exister une famille de sous-ensembles de Ω disjoints $\Omega_1, \ldots, \Omega_m$ et une famille de similitudes[*] $\{S_1, \ldots, S_m\}$ pour lesquelles $S_i(\Omega) = \Omega_i$ avec $1 \leqslant i \leqslant m$, et telles que

$$f(t) = \lambda_i f\big(S_i^{-1}(t)\big), \tag{2.102}$$

lorsque $t \in \Omega_i$. On peut introduire dans ces égalités une « erreur suffisamment régulière » g_i, dont la nature sera précisée par la définition. Cette notion est généralisée et formalisée

[*]. Une similitude est le produit d'une isométrie et d'une homothétie de rapport inférieur à 1.

dans la définition suivante.

Définition 2.65 Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est *auto-similaire* d'ordre $k \in \mathbb{R}^+$ si

- il existe un ensemble ouvert borné Ω et des similarités contractantes $S_1, \ldots, S_m$ telles que[*]
$$S_i(\Omega) \subset \Omega \qquad (1 \leqslant i \leqslant m), \tag{2.103}$$
et
$$S_i(\Omega) \bigcap S_j(\Omega) = \varnothing \qquad (1 \leqslant i,j \leqslant m, i \neq j), \tag{2.104}$$
- il existe une fonction $g \in \dot{\mathcal{C}}^k$, $\lfloor k \rfloor$-régulière telle que f satisfasse
$$f(t) = \sum_{i=1}^{m} \lambda_i f\big(S_i^{-1}(t)\big) + g(t), \tag{2.105}$$
où les λ_i $(1 \leqslant i \leqslant m)$ sont des nombres complexes,
- la fonction f n'est pas uniformément hölderienne d'exposant k dans un sous-ensemble fermé de Ω.

La fonction g peut être interprétée comme la correction douce (la fonction est hölderienne) à l'auto-similarité de f. Nous chercherons à déterminer en quels points t la fonction f appartient à $\mathcal{C}^h(t)$, avec $h < k$. Dans cette section, f représentera une telle fonction auto-similaire.

Le spectre de Hölder de f peut s'obtenir analytiquement. Pour déterminer le support de ce spectre, il nous faut introduire les nombres $h_{\min}$ et $h_{\max}$ relativement aux similarités $\{S_i\}$, comme nous l'avons fait pour le formalisme multifractal concernant les mesures grâce aux égalités (1.77). Si c_i désigne le rapport de contraction de S_i, posons

$$h_{\min} = \inf_{1 \leqslant i \leqslant m} \frac{\log \lambda_i}{\log c_i}, \qquad h_{\max} = \sup_{1 \leqslant i \leqslant m} \frac{\log \lambda_i}{\log c_i}, \tag{2.106}$$

La quantité $\tau(q)$ est définie par analogie avec la dimension de similarité. C'est le nombre vérifiant l'égalité

$$\sum_{i=1}^{m} \lambda_i^q (c_i)^{-\tau(q)} = 1. \tag{2.107}$$

Le spectre $d(h)$ peut être déterminé à partir de τ.

Théorème 2.66 *Le spectre $d(h)$ est donné sur $[0,k[$ par*

$$d(h) = \inf_q \{qh - \tau(q)\}. \tag{2.108}$$

[*]. Il s'agit en fait de la condition de l'ensemble ouvert. Voir section 1.2.

Si $h_{\min} > 0$*, la fonction* $d(h)$ *est nulle hors de* $[h_{\min},h_{\max}] \cup [k, +\infty[$ *et concave sur* $[h_{\min},h_{\max}]$*. Sa valeur maximum* $d_{\max}$ *sur cet intervalle est telle que*

$$\sum (c_i)^{d_{\max}} = 1. \tag{2.109}$$

La méthode reposant sur la transformée en ondelettes intégrale (il en est donc de même si l'on utilise les fonctions de structure, avec les restrictions mentionnées plus haut) permet d'accéder à la partie croissante du spectre de Hölder.

Théorème 2.67 *Si* $h_{\min} > 0$ *et* g *est de classe* C^∞*, alors, pour tout* $h \leqslant h_0$*, où* h_0 *est la valeur pour laquelle le maximum de* $d(h)$ *est atteint,*

$$d(h) = \inf_{q>0} \{qh - \tilde{\eta}(q)\} + n. \tag{2.110}$$

Si g *n'est pas de classe* C^∞*, soit* q_k *la valeur telle que* $\tilde{\eta}(q_k) = kq_k$ *et* $h_{q_k} < h_0$ *la valeur pour laquelle l'infimum dans la relation (2.110) est atteint en* q_k*. Quel que soit* $h < h_{q_k}$*, la relation (2.110) est vérifiée.*

En fait, la deuxième partie du théorème s'écrit comme suit. L'équation (2.110) est vérifiée pour les valeurs de q telles que $\tau(q) \leqslant kq - n$.

Les méthodes d'obtention du spectre de Hausdorff envisagées ici présentent le défaut majeur de ne fournir aucune information sur les valeurs de $d(h)$ correspondant à un q négatif, c'est-à-dire sur la partie décroissante du spectre. La *méthode des maxima du module de la transformée en ondelettes* pallie ce problème. L'idée est de ne considérer que les valeurs de la transformée en ondelettes correspondant à une singularité, en ne retenant que les valeurs sur les lignes de maxima du module[⁎]. On définit la fonction de partition comme suit,

$$Z(a,q) = \sum_{\ell} \sup_{a' \leqslant a} |Wf(\ell(a'),a')|^q, \tag{2.111}$$

où la somme est prise sur toutes les lignes de maxima définies sur un intervalle du type $[a_0,a]$ et

$$\eta(q) = \lim_{a \to 0} \frac{\log Z(a,q)}{\log a}. \tag{2.112}$$

Ici aussi, l'idée est de calculer

$$d(h) = \inf_{q} \{qh - \eta(q)\}. \tag{2.113}$$

Cette méthode doit cependant être corrigée. En effet, telle que définie plus haut, elle peut donner lieu à des valeurs $\eta(q)$ beaucoup plus petites que $\tilde{\eta}(q) - n$. Dans $\mathbb{R}$, $\int_{\mathbb{R}} |Wf(b,a)|^q db$

⁎. Bien sûr, en pratique il peut exister des lignes de maxima ne correspondant pas à une singularité.

et $a\sum_\ell \sup|Wf(\ell(a),a)|^q$ sont de même amplitude si les lignes sont espacées d'une distance a, puisqu'il s'agit alors d'une somme de Riemann. Pour bâtir un contre-exemple, on utilise une fonction où les maxima sont espacés d'une distance bien inférieure à a. Sans surprise, un tel contre-exemple fait intervenir les singularités oscillantes[*]. De tels cas pathologiques sont de peu d'importance numérique, d'autant que ces méthodes, comme nous le verrons, ne sont pas adaptées à l'analyse de singularités oscillantes. En pratique, il suffit de supprimer les lignes de maxima ne se prolongeant pas au-delà d'une échelle a minimum. On peut toutefois s'affranchir de ce comportement et raffiner la méthode des maxima de la transformée en ondelettes en imposant une distance minimum entre les maxima, pour éviter la prolifération de maxima locaux dûs à des oscillations rapides. Pour ce faire, on divise $\mathbb{R}^n$ en cubes de longueur $L > 1$ et, pour chaque cube de Ω, on garde uniquement le plus grand maximum local. Nous utiliserons toujours η pour désigner l'estimateur relatif à cette méthode modifiée.

Pour les fonctions auto-similaires, on peut montrer que lorsque $q > 0$, $\int_{\mathbb{R}} |Wf(b,a)|^q db$ et $a\sum_\ell \sup|Wf(\ell(a),a)|^q$ sont de même amplitude. La méthode reposant sur les maxima présente l'avantage de fournir un spectre $d(h_q)$ pour toutes les valeurs de q.

Théorème 2.68 *Si $h_{\min} > 0$ et $\overline{\cup S_i(\Omega)} \subset \bar{\Omega}$, alors*

$$d(h) = \inf_{q\in\mathbb{R}} \{qh - \eta(q)\}. \tag{2.114}$$

Si la condition $\overline{\cup S_i(\Omega)} \subset \bar{\Omega}$ n'est pas vérifiée, l'égalité (2.114) reste valable pour les mêmes hypothèses que celles énoncées au théorème 2.67.

Puisque $\mathcal{L}^n(\Omega_i) = c_i^n \mathcal{L}^n(\Omega)$, la condition $\overline{\cup S_i(\Omega)} \subset \bar{\Omega}$ équivaut à demander $d_{\max} = n$, par la relation (2.109). Cette égalité peut être plus facile à vérifier.

Remarques sur les méthodes d'estimation du spectre de Hölder

Nous allons maintenant faire quelques remarques sur les méthodes d'estimation du spectre de Hölder. La première met en évidence les limites du formalisme multifractal. Nous discuterons ensuite sur le nombre de moments nuls que doit posséder une ondelette pour pouvoir procéder à l'estimation du spectre de Hölder et terminerons sur des considérations concernant les singularités oscillantes.

[*]. La transformée en ondelettes d'une fonction du type $f(t) = t^s \sin(t^{-s'})g(t)$, où $g \in C^\infty(]0,1])$, $[g] \subset [0,1]$ et $g(t) = 1$ si $t \in [0,1/2[$, avec une ondelette bien choisie possède une infinité de lignes de maxima pour $a \in [1/2,1]$

Le résultat suivant montre que le formalisme multifractal, basé sur la transformée de Legendre inverse, ne peut fournir en toute généralité qu'une borne supérieure du type (2.100) pour le spectre de Hausdorff.

Théorème 2.69 *Étant donné une fonction d intégrable au sens de Riemann* sur $\mathbb{R}^+$, il existe deux fonctions f_1 et f_2 associées à une même fonction $\tilde{\eta}$ telles que le spectre de f_1 soit d et que f_2 soit de classe C^∞ sauf à l'origine.*

Le spectre de la fonction f_2 est donc égal à 0 en tous les points sauf un. Ainsi, il est clair que l'information contenue dans $\tilde{\eta}$ n'est pas suffisante pour déterminer d. De plus, il suffit qu'une fonction d soit intégrable au sens de Riemann pour être le spectre de Hölder d'une fonction f. Pourtant, force est de constater que le formalisme multifractal s'applique à des fonctions beaucoup plus générales que les fonctions strictement auto-similaires. Parmi les méthodes d'estimation du spectre, la plus stable est la méthode des maxima du module de la transformée en ondelettes : choisir les maxima du module comme estimateurs et non l'ensemble de valeurs de la transformée conduit naturellement à une méthode plus stable. Aussi, c'est cette dernière que nous emploierons en pratique.

Il convient bien sûr d'utiliser une ondelette possédant suffisamment de moments nuls pour estimer le spectre de Hölder d. Cependant, si la fonction est de classe C^∞, il n'est pas souhaitable d'utiliser une ondelette mère possédant une infinité de moments nuls et, de manière générale, on évite de recourir directement à des ondelettes possédant un nombre de moments nuls trop élevé [291]. En effet, une ondelette avec un nombre de moments nuls trop important possède un comportement très oscillant, ce qui augmente le nombre de lignes de maxima associées à un signal par la transformée. Pour la plupart des ondelettes utilisées, comme les dérivées successives de la gaussienne, le nombre de lignes de maxima du module augmente linéairement avec le nombre de moments nuls. Or, nous avons vu qu'il est primordial que la distance entre deux lignes de maxima du module successives soit suffisante pour que la méthode associée soit efficace. En pratique, il suffit d'observer comment le spectre évolue lorsque l'on augmente le nombre de moments nuls de l'ondelette mère : on effectue la transformée en ondelettes avec un nombre de moments nuls croissant jusqu'à ce que le spectre ne soit plus modifié par ce changement.

Soit une fonction $f = s + r$ telle que $r \in C^\infty$ et l'exposant de Hölder de la fonction s en un point quelconque soit fini. Formellement, les lignes de maxima du module de la transformée en ondelettes peuvent être classées en deux familles : celles se comportant comme a^h, reflétant l'irrégularité de s et celles se comportant comme a^m, où m est le nombre de moments nuls de l'ondelette mère, traduisant la plus grande régularité de r.

∗. Le résultat de Jaffard [203] est légèrement plus général.

Pour les petites échelles et les valeurs de q positives, $Z(a,q)$ devrait se comporter comme $a^{\eta(q)}$, alors que pour les valeurs de q négatives, il devrait se comporter comme a^{qm}. La fonction numériquement estimée η' associée à f va donc notablement différer de la fonction η associée à s : $\eta'(q)$ se comportera comme qm pour les valeurs de q négatives. Pour ces valeurs, le spectre de Hölder induit se comportera donc linéairement. Cette perturbation est facilement détectable, puisque la pente observée dépend du nombre de moments nuls de l'ondelette mère. On élimine ces lignes de maxima « parasites » en ne considérant pas les lignes de maxima en dessous d'un seuil fixé ou en choisissant m suffisamment grand pour pouvoir identifier les lignes de maxima du module « à décroissance très rapide », se comportant comme a^m.

Concernant les singularités oscillantes, le formalisme multifractal canonique est plutôt inefficace. En fait, on peut généraliser les fonctions auto-similaires en considérant les fonctions dont la transformée en ondelettes est auto-similaire. Ceci mène à un formalisme multifractal « *grand canonique** », où l'on cherche à estimer le spectre

$$d(h,\beta) = \dim_{\mathcal{H}}\{t : h(t) = h \text{ et } \beta(t) = \beta\}, \tag{2.115}$$

relatif à une fonction f dont la transformée en ondelettes est auto-similaire. Il peut être montré que l'estimation de la fonction $\tilde{\eta}$ relative à un signal comportant des singularités oscillantes ne prend pas en compte tous les paramètres caractérisant ce type de singularité et conduit notamment à une mauvaise évaluation du spectre de Hausdorff d. À notre connaissance, les tentatives de mise en oeuvre numérique d'un formalisme multifractal grand canonique se sont toutes révélées infructueuses, et il n'existe à ce jour aucune méthode efficace permettant d'estimer le spectre de Hausdorff d'une fonction comportant des singularités oscillantes en toute généralité (bien que des résultats « théoriques » existent [205]).

D'une manière générale, nous devons marquer notre scepticisme quant à la capacité des formalismes multifractals actuels[+] de caractériser numériquement les singularités oscillantes. Remarquons d'abord que numériquement les points au voisinage de telles singularités sont extrêmement mal définis, puisque la fréquence instantanée est très grande. On ne peut donc envisager une étude trop locale. Deuxièmement, les exposants d'oscillation β nécessaires pour caractériser les singularités oscillantes ne peut être estimé par les valeurs de la transformée en ondelettes situées dans un cône d'influence quelconque. Cependant, il va de soi que les oscillations à l'extérieur d'un cône d'influence sont très dépendantes des autres singularités. Il n'est pas difficile d'imaginer un signal comportant deux singulari-

*. Par analogie avec la thermodynamique.

+. Notre point de vue concerne également le formalisme multifractal basé sur les coefficients en ondelettes dominants, qui sera présenté dans la prochaine section.

tés oscillantes dont les oscillations se parasiteraient l'une l'autre à une distance suffisante des dites singularités. Tout comme les singularités lissées, il semble que les singularités oscillantes ne puissent être caractérisées uniquement *via* les petites échelles de la transformée en ondelettes, alors que nous sommes limités dans le choix des grandes échelles par la présence d'autres singularités. À nos yeux, ces limitations, spatiales et en échelle des valeurs numériquement accessibles de la transformée en ondelettes rend difficile toute tentative de caractérisation basée sur le formalisme multifractal. On pourrait arguer que ce formalisme repose sur l'estimation d'un grand nombre de singularités et qu'un « effet statistique » pourrait permettre d'évaluer correctement le spectre. C'est peut être oublier que l'on procède à partir d'estimations locales avant tout et que nous doutons même de la possibilité de caractériser numériquement un signal avec un faible nombre de singularités oscillantes en toute généralité.

Nous ne pouvons être aussi catégorique quant à l'inexistence d'autres méthodes de caractérisation des singularités oscillantes. Comme nous venons de le faire remarquer, une des raisons pour lesquelles, selon nous, le formalisme multifractal est inefficace pour les singularités oscillantes est le caractère non-local de ce type de singularité, en ce sens qu'il faut aussi considérer le signal en dehors de tout cône d'influence. On pourrait cependant espérer analyser ce caractère oscillant en choisissant une famille d'ondelettes mères $\{\psi_j\}_{j\in\mathbb{N}}$ telles que le cône d'influence de ψ_j soit inclus dans celui de ψ_{j+1} et en étudiant le comportement des valeurs de la transformée en ondelettes aux points appartenant au cône d'influence de ψ_{j+1} et n'appartenant pas à celui de ψ_j, comme pour constater la présence d'un « flux » engendré par la singularité oscillante traversant les cônes. On pourrait, par exemple, procéder de la manière suivante. Supposons que l'on souhaite étudier la singularité d'une fonction f en un point t. Si $C_j(a)$ désigne l'ensemble des points b tels que (b,a) appartiennne au cône d'influence de ψ_j, soit $I_j(a)$ $(j>0)$ l'ensemble des points b appartenant à $C_j(a)$ et pas à $C_{j-1}(a)$ tels que $b<t$. Pour chacun de ces intervalles, on peut construire les fonctions

$$\gamma(j,a) = \frac{1}{\mathcal{L}^n\big(I_j(a)\big)} \int_{I_j(a)} |Wf(b,a)|\, db. \tag{2.116}$$

L'étude de l'évolution suivant j et a de ces fonctions pourrait apporter des informations sur la nature de la singularité. Il n'en reste pas moins que rien ne sous-tend de telles affirmations. L'obtention d'une famille d'ondelettes $\{\psi_j\}$ est chose aisée. En pratique il n'est bien sûr pas nécessaire de considérer des ondelettes mères à support compact. Si l'on pose $f_j(t) = \exp(-t^2/2\sigma_j^2)$, où $\{\sigma_j\}$ est une suite strictement croissante de $\mathbb{R}_*^+$, il suffit de définir la famille d'ondelettes $\psi_j = D^m f_j$, avec $m \in \mathbb{N}_0$. Une idée similaire consiste à utiliser en tant qu'ondelettes des fonctions dépendant explicitement du paramètre d'échelle a, de façon à faire évoluer le cône de manière non-linéaire en fonction de l'échelle. On peut

par exemple définir la fonction

$$\psi_a(t) = -t \exp(\frac{-t^2}{2a\sigma^2}). \tag{2.117}$$

Nous nous proposons d'implémenter cette méthode et d'en interpréter les résultats dans un avenir proche.

Paramétrage du spectre de Hölder

Les développements qui suivent permettent de représenter le spectre de Hölder comme une courbe paramétrée par q. Cette méthode est numériquement une des plus simples et permet de représenter entièrment le spectre, mais nécessite quelques hypothèses sur le spectre lui-même. Par souci de simplicité, nous travaillerons sur $\mathbb{R}$.

Les développements qui suivent sont analogues à ceux utilisés pour le formalisme multifractal des mesures. Nous utiliserons les mêmes notations et ferons les mêmes hypothèses (cf. hypothèse 1.44).

Hypothèse de travail 2.70 Nous supposerons que le spectre de Hölder d est une fonction dérivable de h, strictement positive et strictement concave.

Pour q fixé, supposons que $h_q > 0$ est la valeur (en supposant qu'elle existe) de h pour laquelle l'infimum dans la relation (2.113) est réalisé. On peut obtenir les relations analogues aux égalités (1.49) et (1.52),

$$q = D_h d(h_q) \tag{2.118}$$

et, si h_q est dérivable par rapport à q,

$$D_q \eta(q) = h_q. \tag{2.119}$$

On peut alors écrire

$$d(h_q) = q D_q \eta(q) - \eta(q). \tag{2.120}$$

En supposant que $Z(a,q)$ est dérivable par rapport à q lorsque a tend vers 0^+, d'après la définition (2.112) de η, on a[❋]

$$\partial_q \log Z(a,q) \sim D_q \eta(q) \log(a). \tag{2.121}$$

En posant

$$h_a(q) = \partial_q \log Z(a,q), \tag{2.122}$$

❋. La limite porte bien sûr sur a tendant vers 0. S'il n'y a pas de convergence, on prend la limite inférieure.

l'égalité suivante est vérifiée,

$$h_q = \underline{\lim}_{a\to 0} \frac{h_a(q)}{\log a}. \tag{2.123}$$

De même, définissons

$$d_a(q) = q\partial_q \log Z(a,q) - \log Z(a,q), \tag{2.124}$$

pour obtenir, à partir de (2.120),

$$d(h_q) = \underline{\lim}_{a\to 0} \frac{d_a(q)}{\log a}. \tag{2.125}$$

Les relations (2.123) et (2.125) permettent donc de représenter le spectre de Hölder comme une courbe paramétrée par q. En pratique, pour simplifier les notations, on pose

$$Z(a,q) = \sum_{\ell} |Wf(\ell(a),a)|^q, \tag{2.126}$$

ce qui permet d'obtenir

$$h_a(q) = \sum_{\ell(a)} \frac{|Wf(\ell(a),a)|^q}{Z(a,q)} \log |Wf(\ell(a),a)|, \tag{2.127}$$

et

$$d_a(q) = \sum_{\ell(a)} \frac{|Wf(\ell(a),a)|^q}{Z(a,q)} \log \frac{|Wf(\ell(a),a)|^q}{Z(a,q)}. \tag{2.128}$$

Une méthode pour caractériser numériquement le spectre de singularités peut être définie grâce à ce paramétrage. La première étape consiste à calculer la transformée en ondelettes du signal à étudier en utilisant des ondelettes du type (2.24). La définition des lignes de maxima du module se fait en repérant d'abord les maxima du module de la transformée dont les valeurs ne sont pas inférieures à un seuil donné. Il faut ensuite relier ces maxima du module pour obtenir les lignes de maxima du module et supprimer celles ne pouvant se prolonger à la plus petite échelle accessible numériquement a_0. Pour chaque ligne, la valeur d'un maximum du module à une échelle donnée est remplacée par la valeur maximum atteinte par les maxima du module aux échelles inférieures le long de cette même ligne. Ces étapes sont regroupée sous l'appellation *chaînage*. Il faut ensuite, q étant fixé, calculer $Z(a,q)$, défini par l'égalité (2.126), pour toutes les échelles numériquement disponibles, en ayant préalablement ordonné les valeurs intervenant dans chaque somme. On obtient ainsi un signal Z_q qui à une échelle a fait correspondre la valeur $Z(a,q)$. Le calcul des quantités définies par (2.127) et (2.128) peut alors s'effectuer. Les valeurs h_q et $d(h_q)$ peuvent enfin être obtenues en construisant les signaux qui, au logarithme de l'échelle a, font correspondre $h_a(q)$ et $d_a(q)$ respectivement, puis en faisant une régression linéaire sur un intervalle du type $[a_0,a_\varepsilon[$ où a_ε est une valeur de l'échelle suffisamment petite. Ces étapes sont résumées au tableau 2.2.

Calcul du spectre de singularités d'un signal

1. définir la valeur de q et un seuil ε,
2. calculer la transformée en ondelettes,
3. relier les maxima du module dont le module est supérieur à ε,
4. supprimer les maxima ne se prolongeant pas aux petites échelles,
5. prendre le maximum atteint aux échelles inférieures le long de chaque ligne,
6. calculer $Z(a,q)$,
7. calculer $h_a(q)$ et $d_a(q)$ pour toutes les valeurs de a,
8. effectuer une régression linéaire sur $\{\big(\log a, h_a(q)\big)\}$ et $\{\big(\log a, d_a(q)\big)\}$,
9. la valeur de la pente aux petites échelles permet d'estimer h_q et $d(q)$.

TAB. 2.2 – *L'algorithme de la méthode des maxima du module de la transformée en ondelettes. La partie délicate réside dans la détermination de l'intervalle sur lequel va être effectuée la régression.*

Les méthodes reposant sur la transformée en ondelettes sont souvent numériquement efficaces, peut-être parce que l'ondelette moyenne le bruit présent dans le signal. La méthode reposant sur les maxima du module est plus stable que la transformée en ondelettes intégrale. On peut conjecturer que la raison réside dans le fait que les erreurs numériques sont de moindre importance aux maxima. Le spectre est accessible pour toutes les valeurs de q, mais la méthode est bien moins stable pour les valeurs négatives. Ici aussi, une explication peut être donnée. Pour ces valeurs, ce sont les petites valeurs des maxima, plus sujettes aux erreurs, qui sont prépondérantes. C'est pour cette raison qu'il convient de ne considérer que les valeurs de la transformée supérieures à un seuil fixé[⁎]. De plus, ces petites valeurs sont moins nombreuses que les grandes le long des lignes de maxima.

Terminons en donnant un exemple de fonction multifractale.

Exemple 2.71 Remarquons d'abord que la méthode des maxima du module de la transformée en ondelettes appliquée à l'escalier du diable (figure 2.12) permet de retrouver la partie finie du spectre (2.87) obtenu dans l'exemple 2.60. Pour obtenir une fonction multifractale, nous allons généraliser l'escalier du diable. Si μ désigne une mesure associée à l'ensemble de Cantor, comme définie dans l'exemple 1.43, on peut construire l'escalier du diable correspondant en posant $f(t) = \mu([0,t])$, $t \in [0,1]$. Cette fonction est localement constante sur le complémentaire de l'ensemble de Cantor C et on montre sans peine que la fonction associée au poids $1/2$ est l'escalier du diable présenté dans l'exemple 2.45. Pour un escalier du diable généralisé, associé à un poids $p \neq 1/2$, le spectre est celui de la mesure associée, donné dans l'exemple 1.46. La méthode des maxima du module de la transformée en ondelettes donne des valeurs du spectre en bon accord avec la fonction théorique comme en atteste la figure 2.15. L'ondelette mère utilisée ici est la dérivée première de la

⁎. Typiquement, ce seuil est en général fixé à 10^{-2}.

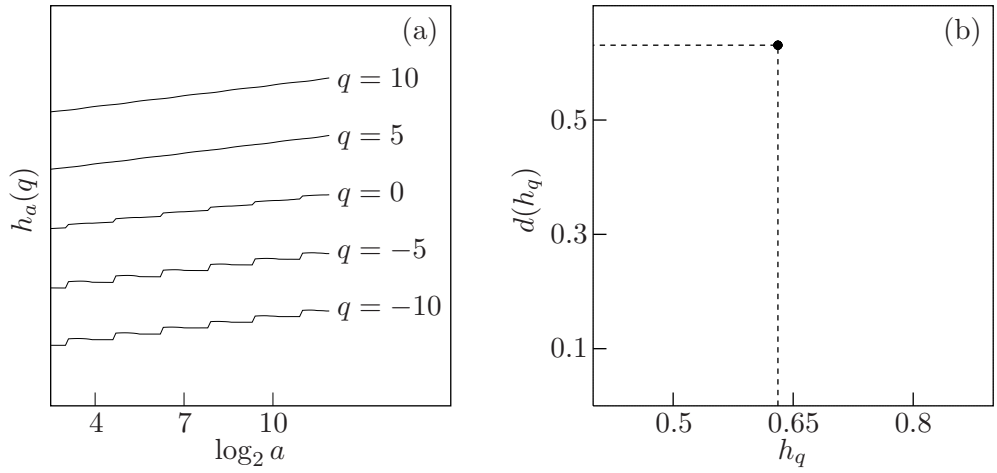

FIG. 2.12 – *Détermination, par la méthode des maxima du module de la transformée en ondelettes, du spectre de l'escalier du diable (associé au poids $p = 1/2$). Les fonctions $h_a(q)$ pour différentes valeurs de q sont représentées en (a). Notons que les valeurs de l'ordonée sont arbitraires, car les fonctions ont été décalées pour plus de clarté. Un même comportement linéaire est observé indépendamment des valeurs de q, à des oscillations de période $\log_2 3$ près, dues à l'invariance discrète par une dilatation d'un facteur 3. (b) On mesure bien un spectre monofractal. L'ondelette mère utilisée est la dérivée première de la gaussienne.*

gaussienne. □

2.6 Coefficients en ondelettes dominants et formalisme multifractal associé

Le formalisme multifractal basé sur les coefficients dominants est, à notre connaissance, la plus récente des méthodes d'estimation du spectre de Hölder d'une fonction f. Elle repose sur la transformée en ondelettes discrète et conserve dans une large mesure l'efficacité numérique de la méthode des maxima du module de la transformée en ondelettes tout en possédant une base théorique satisfaisante [213]. Après un bref rappel de l'analyse multirésolution, nous présenterons les notions fondamentales avant d'exposer la méthode proprement dite. Nous terminerons en formulant quelques remarques. Le méthode des coefficients dominants a été essentiellement introduite[❋] par JAFFARD [205, 211, 212] et notre présentation est basée sur ses travaux. L'analyse multirésolution, due à MALLAT [256] et MEYER [278], est notamment traitée dans les références [112, 278].

❋. La première mise en oeuvre est due à LASHERMES et ABRY.

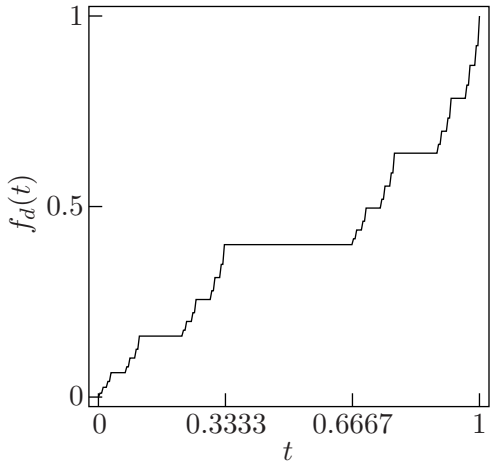

FIG. 2.13 – *L'escalier du diable associé au poids 0.4 (voir l'exemple 2.71).*

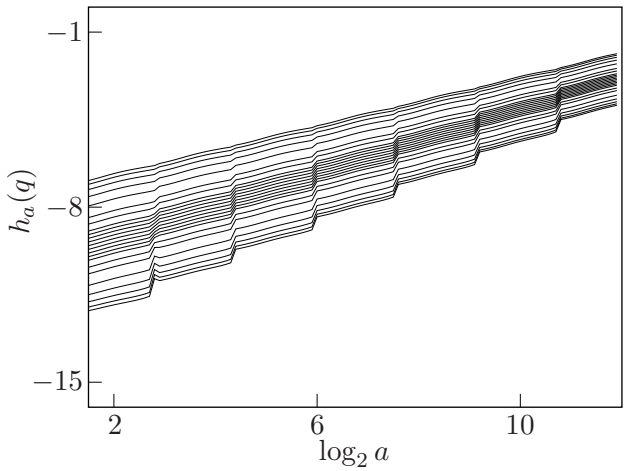

FIG. 2.14 – *Les fonctions h_a obtenues par la méthode des maxima du module de la transformée en ondelettes pour l'escalier du diable, avec un poids $p = 0.4$. Les valeurs de q s'étalent de $q = -7$ (en bas) à $q = 7$ (en haut). L'ondelette mère est la dérivée première de la gaussienne.*

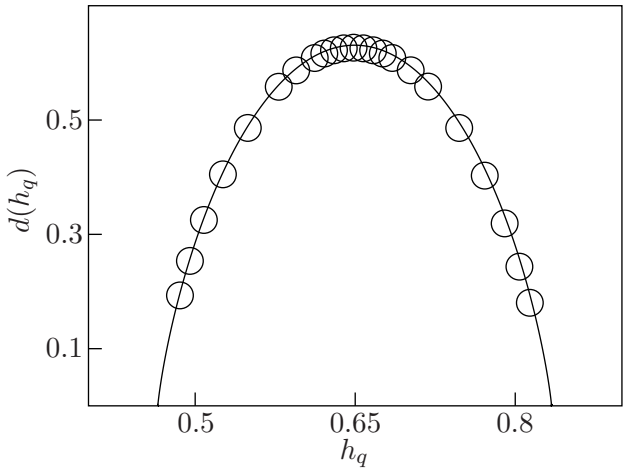

FIG. 2.15 – *Les mesures expérimentales du spectre de l'escalier du diable associé au poids* $p = 0.4$ *en utilisant la méthode des maxima du module de la transformée en ondelettes et l'ondelette mère dérivée première de la gaussienne. Ces valeurs ont été calculées pour* q *allant de* -7 *à* 7*. Entre* -2 *et* 2*, les pas sont de* 0.5*, et entre* -1 *et* 1 *de* 0.25*. Ces mesures (cercles) sont en bon accord avec la courbe théorique (trait continu).*

Analyse multirésolution de l'espace L^2

La transformée en ondelettes discrète peut être vue comme une transformée en ondelettes continue où on supprime la redondance en imposant au paramètre d'échelle a d'être de la forme $a = 2^j$ ($j \in \mathbb{Z}$) et à la position de pouvoir s'écrire $b = 2^j k$ ($j \in \mathbb{Z}$, $k \in \mathbb{Z}^n$). Après un bref rappel de la théorie de l'analyse multirésolution, nous introduirons la notion de coefficient dominant.

La théorie de l'analyse multirésolution permet d'obtenir une base orthonormée constituée d'ondelettes. Outre son intérêt théorique, elle fournit de puissants algorithmes en traitement du signal [123, 257, 296].

Définition 2.72 Une *analyse multirésolution* est la donnée d'une suite d'espaces $\{V_j\}_{j\in\mathbb{Z}}$ de $L^2 = L^2(\mathbb{R}^n)$ vérifiant les propriétés suivantes :

- pour tout j, on a $V_j \subset V_{j+1}$,
- pour tout j, $f \in V_j$ si et seulement si $f(2\,.\,) \in V_{j+1}$,
- il existe une fonction $\varphi \in V_0$ telle que la famille $\{\varphi(\,.\,-k)\}_{k\in\mathbb{Z}}$ forme une base orthonormée de V_0,
- on a $\cap_j V_j = \{0\}$ et $\cup_j V_j$ est dense dans L^2.

La fonction φ est appelé *fonction d'échelle*.

Nous noterons W_j le complémentaire orthogonal de V_j dans V_{j+1}, $V_{j+1} = V_j \oplus W_j$. On peut montrer qu'il existe $2^n - 1$ fonctions $\psi^{(i)}$ ($i \in \{1,2,\ldots,2^n-1\}$), de même m-régularité que φ, telles que les fonctions $\psi^{(i)}(\,.\,-k)$ ($1 \leqslant i < 2^n$, $k \in \mathbb{Z}^n$) forment une base orthonormée de W_0. Dès lors, les fonctions $2^{nj/2}\psi^{(i)}(2^j\,.\,-k)$ définissent une base orthonormée de W_j.

Les espaces W_j sont orthogonaux et l'espace L^2 peut être décomposé en somme directe de deux manières. Symboliquement,

$$L^2 = \bigoplus_{j\in\mathbb{Z}} W_j \qquad \text{ou} \qquad L^2 = V_0 + \bigoplus_{j\in\mathbb{N}} W_j. \tag{2.129}$$

Ainsi, si $f \in L^2$, on peut écrire

$$f(t) = \sum_{j\in\mathbb{Z},\, k\in\mathbb{Z}^n,\, i} c_{i,j,k}\, \psi^{(i)}(2^j t - k), \tag{2.130}$$

où

$$c_{i,j,k} = 2^{nj} \int_{\mathbb{R}^n} f(t)\, \psi^{(i)}(2^j t - k)\, dt, \tag{2.131}$$

et

$$f(t) = \sum_{k\in\mathbb{Z}^n} c_k\, \varphi(t-k) + \sum_{j\in\mathbb{N},\, k\in\mathbb{Z}^n,\, i} c_{i,j,k}\, \psi^{(i)}(2^j t - k), \tag{2.132}$$

avec

$$c_k = \int_{\mathbb{R}^n} f(t)\, \varphi(t-k)\, dt. \tag{2.133}$$

Les valeurs $\{c_{i,j,k}\}$ sont appelées les *coefficients en ondelettes** de f. L'indice j représente l'échelle et k la position.

L'égalité (2.131) peut être vue comme la version discrète de la transformée en ondelettes continue. L'analogie avec la décomposition de Littlewood-Paley est évidente ; la plupart des espaces présentés précédemment ($\dot{B}^s_{p,q}$ et $\mathcal{C}^{s,s'}$ notamment) peuvent être directement caractérisés par la transformée en ondelettes discrète [201, 278]. Quant à la décomposition du type (2.132), elle caractérise entre autres les espaces de Besov non-homogènes $B^s_{p,q}$ [278]. Les algorithmes associés à l'analyse multirésolution sont souvent très rapides. En pratique, on projette la fonction étudiée f sur l'un des espaces V_j. La fonction f est ainsi approchée par une fonction $f_j \in V_j$. On obtient ensuite une expression du type (2.132) en décomposant V_j par un algorithme pyramidal très simple, de complexité d'ordre N. Les espaces V_j sont appelés les *espaces d'approximation*, tandis que les espaces W_j constituent les *espaces des détails*.

Remarque 2.73 Si ϕ et ψ sont associés à une analyse multirésolution de $L^2(\mathbb{R})$, on peut obtenir une analyse multirésolution de L^2 par produit tensoriel de la manière suivante.

∗. Remarquons que, comme pour la transformée en ondelettes continue, nous n'utilisons plus la normalisation de l'espace L^2.

Si $t = (t_1, \ldots, t_n) \in \mathbb{R}^n$, on pose $\varphi(t) = \phi(t_1) \cdots \phi(t_n)$ et, pour $t \in \mathbb{R}$, $\psi_1 = \psi$, $\psi_0 = \phi$. Des ondelettes sur $\mathbb{R}^n$ sont alors obtenues en définissant $\psi^{(i)}(t) = \psi_{i_1}(t_1) \cdots \psi_{i_n}(t_n)$, où $i \in \{0,1\}^n \setminus \{0\}^n$ (ainsi, il n'existe pas de i tel que $\psi^{(i)} = \varphi$). □

Le formalisme multifractal repose sur la notion de coefficient dominant. Nous utiliserons les notations suivantes.

Notation 2.74 Nous allons indexer les coefficients de la transformée en ondelettes avec des *cubes dyadiques* $\lambda_{j,k} = {}^{k}\!/_{2^j} + [0,{}^{1}\!/_{2^j}[^n$ et non plus les indices j et k. Puisque i peut prendre $2^n - 1$ valeurs différentes, on peut supposer qu'il prend ses valeurs sur $\{0,1\}^n \setminus \{0\}^n$. Ainsi, on pose

$$\lambda_{i,j,k} = \frac{k}{2^j} + \frac{i}{2^{j+1}} + [0,\frac{1}{2^{j+1}}[^n. \tag{2.134}$$

Nous omettrons dorénavant les indices du cube dyadique en écrivant $\lambda = \lambda_{i,j,k}$. Ainsi, nous avons $c_\lambda = c_{i,j,k}$ et $\psi_\lambda(t) = \psi^{(i)}(2^j t - k)$.

Les raisons d'un recours aux cubes dyadiques sont simples, si on fait l'analogie avec la transformée en ondelettes continue et le cône d'influence : les ondelettes ψ_λ sont essentiellement localisées autour du cube λ. Plus précisément, lorsque l'ondelette est à support compact, il existe une constante C telle que, pour tous les indices i, j et k, $[\psi_\lambda] \subset C\lambda$, où $C\lambda$ est le cube dyadique C fois plus large que λ. Nous appellerons c_λ un *coefficient dyadique*. Nous allons maintenant introduire les notions utiles pour le formalisme multifractal basé sur les coefficients dyadiques. La définition suivante est inspirée des lignes de maxima du module de la transformée en ondelettes continue.

Définition 2.75 Étant donné un cube dyadique λ, les *coefficients dominants* de λ sont les coefficients dyadiques suivants,

$$d_\lambda = \sup_{\lambda' \subset \lambda} |c_{\lambda'}|. \tag{2.135}$$

Si $f \in L^\infty$, les coefficients dominants sont finis, puisque $|c_\lambda| \leqslant C\|f\|_{L^\infty}$. Si l'on se place dans l'espace $\mathbb{R}$, les cubes dyadiques sont de la forme $\lambda_{j,k} = [{}^{k}\!/_{2^j}, k + {}^{1}\!/_{2^j}[$ et pavent le demi-plan espace-échelle. Ce dernier est représenté à la figure 2.16. Si le cube λ est celui marqué par un disque noir, d_λ revient à considérer tous les cubes dyadiques situés en-dessous (l'échelle j' associée à ces cubes est donc telle que $j' \geqslant j$), pour choisir le plus grand coefficient associé. Nous aurons également besoin de considérer les cubes voisins.

Définition 2.76 Deux cubes dyadiques λ_1 et λ_2 sont *adjacents* s'ils sont à la même échelle j et $\mathrm{dist}(\lambda_1,\lambda_2) = 0$. Un cube est adjacent à lui-même. Le cube dyadique à l'échelle j

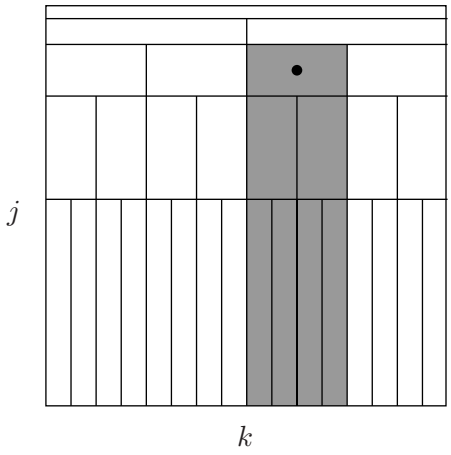

FIG. 2.16 – *Représentation du demi-plan espace-échelle pavé de cubes dyadiques. Les cubes intervenant dans le calcul du coefficient dominant du cube marqué par un disque sont représentés en gris.*

contenant le point t sera noté $\lambda_j(t)$. Les 3^n cubes adjacents à λ seront représentés par $A(\lambda)$. Enfin, on pose

$$d_j(t) = \sup_{\lambda' \in A(\lambda_j(t))} d_{\lambda'} \tag{2.136}$$

Il s'agit simplement d'élargir le domaine aux cubes voisins du cube recouvrant le point considéré, à une échelle donnée, comme illustré par la figure 2.17. À partir de ces considérations, on peut définir le cône d'influence pour les coefficients dyadiques. Cette définition fait naturellement appel aux cubes adjacents.

Définition 2.77 Le *cône d'influence* d'un point t est l'ensemble des cubes dyadiques λ tels qu'il existe une échelle j pour laquelle $\lambda \in A\big(\lambda_j(t)\big)$.

Ces notations permettent de formuler élégamment les résultats précédents. Par exemple, on a la proposition suivante*.

Proposition 2.78 *Si une fonction f appartient à $\mathcal{C}^s(t)$ ($s \geqslant 0$), alors il existe une constante C telle que, pour tout $j \geqslant 0$,*

$$d_j(t) \leqslant C 2^{-sj}. \tag{2.137}$$

Inversement si l'inégalité précédente est vérifiée et qu'il existe $\varepsilon > 0$ tel que $f \in \dot{\mathcal{C}}^\varepsilon$, alors $f \in \mathcal{C}^{s-\delta}(t)$ pour tout δ tel que $0 < \delta \leqslant s$.

Ceci étant, nous avons maintenant les notations nécessaires pour définir le formalisme multifractal basé sur les coefficients dominants.

*. Remarquons que les conditions (2.63) et (2.137) sont équivalentes

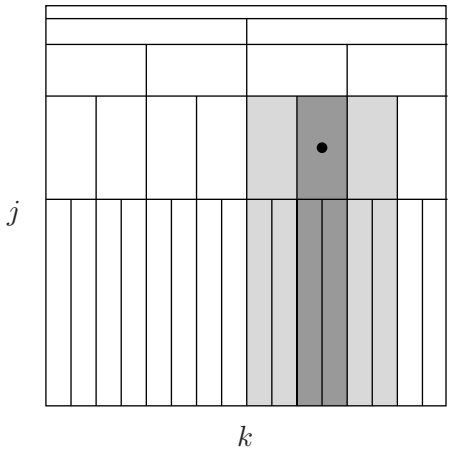

FIG. 2.17 – *Dans cette représentation espace-échelle, $d_j(t)$ est défini à partir des cubes dyadiques grisés, où le cube marqué par un disque représente le cube recouvrant le point t à l'échelle j. Si on le note λ, les cubes intervenant dans le calcul de d_λ sont représentés en gris plus foncé.*

Le formalisme multifractal associé aux coefficients dominants

Le formalisme multifractal relatif aux coefficients dominants définit une méthode implémentable, pouvant donner accès à l'entièreté du spectre de Hölder. Elle repose sur les espaces d'oscillation $\mathcal{O}^s_q$, ce qui la différencie de la méthode des maxima du module de la transformée en ondelettes, qui n'a pas pu être associée à un espace fonctionnel.

Comme précédemment, nous supposerons qu'il existe $\varepsilon > 0$ tel que $f \in \dot{\mathcal{C}}^\varepsilon$. Si $h(t) = h$, la proposition 2.78 affirme l'existence d'une infinité de cubes dyadiques λ adjacents à des cubes contenant t tels que, en utilisant la notation 2.46, $d_\lambda = \widetilde{\mathcal{O}}(2^{-j})$. Suivant la méthode des maxima du module de la transformée en ondelettes, il est naturel de poser

$$S(q,j) = 2^{-nj} \sum_{\lambda \in \Lambda_j} d^q_\lambda, \tag{2.138}$$

où $\Lambda_j = \{\lambda : \mathrm{diam}(\lambda) = 2^{-j}\}$. On remplace la fonction à évaluer η par la fonction ω définie comme suit,

$$\omega(q) = \varliminf_{j \to \infty} \frac{\log S(q,j)}{\log 2^{-j}}, \tag{2.139}$$

pour espérer trouver le spectre multifractal en posant

$$d(h) = \inf_q \{qh - \omega(q)\} + n. \tag{2.140}$$

Comme pour la transformée en ondelettes intégrale, ω peut être défini *via* un espace

fonctionnel.

Définition 2.79 Soient $s \in \mathbb{R}$ et $q > 0$. L'espace $\mathcal{O}_q^s$ est défini comme suit[+],

$$\mathcal{O}_q^s = \{f \in S' : \|f\|_{\mathcal{O}_q^s} = \|\{c_k\}\|_{l^q} + \|\{2^{(sq-n)j}\|\{d_\lambda\}_k\|_{l^q}\}_{j\geqslant 0}\|_{l^\infty} < \infty\}. \tag{2.141}$$

Il peut être montré que cette définition ne dépend pas de la base d'ondelettes choisie. Si $q \geqslant 1$, ces espaces sont de Banach, et si $0 < q < 1$, ces espaces sont métrisables et complets[*].

Il est clair que l'on a, si $q > 0$,

$$\omega(q) = \sup\{s : f \in \mathcal{O}_{q,\mathrm{loc}}^{s/q}\}. \tag{2.142}$$

Ces espaces sont reliés aux espaces de Besov par l'identité suivante.

Proposition 2.80 *Si $s > n/q$, alors $\mathcal{O}_q^s = B_{q,\infty}^s$.*

On peut donc obtenir l'analogue du théorème 2.64 pour le formalisme multifractal basé sur les coefficients dyadiques. Il est aussi possible de définir $\mathcal{O}_q^s$ lorsque q est négatif. On peut ainsi associer w à ces espaces pour tout $q \neq 0$ et étendre cette fonction sur $\mathbb{R}$ par une fonction concave. Enfin, on a le résultat suivant.

Théorème 2.81 *L'inégalité suivante est vérifiée,*

$$d(h) \leqslant \inf_{q\in\mathbb{R}}\{qh - \omega(q)\} + n. \tag{2.143}$$

Apport du formalisme multifractal basé sur les coefficients dominants

Nous formulons ici quelques remarques concernant les avantages et les nouveautés apportées par ce nouveau formalisme multifractal.

Les espaces d'oscillation prennent mieux en compte la disposition géométrique des coefficients dyadiques que ne le font les espaces de Besov. La borne supérieure (2.143) donnée par les espaces $\mathcal{O}_q^s$ est plus petite que celle basée sur les espaces $B_{q,\infty}^s$. Si, lorsque $s > n/q$, ces deux estimations sont équivalentes, les espaces de Besov et d'oscillation diffèrent notablement pour les autres valeurs de s. L'apport théorique est donc indéniable.

+. Ces espaces sont des cas particuliers des espaces $\mathcal{O}_q^{s,s'} : \mathcal{O}_q^s = \mathcal{O}_q^{s,0}$.

*. Il suffit de munir $\mathcal{O}_q^s$ de la distance $\mathrm{dist}_{\mathcal{O}_q^s}(f^{(1)},f^{(2)}) = \|f^{(1)} - f^{(2)}\|_{B_{\infty,\infty}^s} + \|\{c_k^{(1)} - c_k^{(2)}\}\|_{l^p} + \|\{2^{(sq-n)j}\|\{\sup_{\lambda'\subset\lambda}|c_{\lambda'}^{(1)} - c_{\lambda'}^{(2)}|\}_k\|_{l^q}\}_{j\geqslant 0}\|_{l^\infty}$.

Sur le plan pratique, une question importante concerne les singularités oscillantes. Ce nouveau formalisme améliore-t-il leur détection ou leur caractérisation par rapport à la méthode des maxima du module de la transformée en ondelettes ? La réponse est pour le moins mitigée. Les premiers résultats numériques ne penchent pas en la faveur d'une des deux méthodes [235] ; pour les singularités de type chirp, l'exposant d'oscillation semble toujours inaccessible. Il est cependant difficile de dire s'il s'agit de restrictions théoriques ou, avant tout, comme nous le pensons, de limitations pratiques, comme il en a été discuté précédemment. On ne peut cependant pas exclure que les premiers bilans numériques soient biaisés par des erreurs ou des problèmes de mise en oeuvre.

Il est aussi possible que les espaces d'oscillation $\mathcal{O}_q^{s,s'}$, que l'on peut voir comme un espace d'oscillation $\mathcal{O}_q^s$ avec un paramètre s' supplémentaire, puissent permettre une caractérisation du spectre grand canonique $d(h,\beta)$, en définissant des quantités du type $\omega(q,s') = \sup\{s : f \in \mathcal{O}_{q,\mathrm{loc}}^{s/q,s'/q}\}$. Toutefois, pour les méthodes reposant sur ces fonctions, le premier problème à régler est celui de la mise en oeuvre pratique.

Dans ce mémoire, c'est la méthode des maxima du module de la transformée en ondelettes qui a été systématiquement utilisée, des tests de fiabilité restant à effectuer concernant la méthode relative aux coefficients dominants. Nous nous proposons toutefois de mettre en oeuvre cette méthode dans un avenir proche.

Chapitre 3

Marches aléatoires browniennes

Les marches aléatoires de type brownien sont utilisées pour la modélisation dans une large variété de domaines [261, 262, 313], allant de la finance [73, 135, 187, 265, 331] à la biologie [17, 20, 314, 315, 351, 387]. Par marche brownienne, nous entendons mouvement brownien tel que défini par Wiener [390] mais aussi mouvement brownien fractionnaire, introduit par Mandelbrot et Van Ness [263] comme une généralisation à variance finie du mouvement brownien traditionnel. La propriété majeure du mouvement brownien fractionnaire est la présence de corrélations à longue portée, qui introduit une faible dépendance entre les points d'une réalisation. De telles dépendances sont constatées dans de nombreuses observations expérimentales (telles celles concernant l'ADN [34, 35]), faisant du mouvement brownien fractionnaire un des modèles les plus efficaces pour leur modélisation. Après la présentation d'une méthode de construction numérique d'un bruit gaussien fractionnaire, nous donnerons un modèle de marche discrète présentant des corrélations à longue portée, basée sur le mouvement brownien fractionnaire. Concernant le mouvement brownien, le lecteur pourra consulter les références [127, 183, 218, 240, 246, 376, 385] ; pour le mouvement brownien fractionnaire, nous renvoyons aux contributions [129, 136, 144, 263, 336, 360]. La section concernant les marches binaires est originale.

3.1 Le mouvement brownien

Le mouvement brownien, d'abord observé empiriquement en 1828 par le biologiste Brown [82] qui étudiait des grains de pollen au microscope a permis l'introduction de la mesure de Wiener, mais aussi la modélisation d'un bon nombre de phénomènes naturels. Ainsi, en 1909, Perrin [315] utilisait des développements théoriques, notamment amorcés par Einstein en 1905 [133, 134], en relation avec des mesures expérimentales pour calculer avec précision le nombre d'Avogadro. Depuis longtemps, le mouvement brownien est un bel exemple d'interaction entre théorie et expérimentation.

Définitions

La modélisation par Wiener du mouvement brownien est bien postérieure à sa découverte [390]. Le mouvement brownien s'introduit facilement, mais les discussions sur la mesure associée sont plus délicates.

Pour simplifier l'écriture, nous adopterons la convention suivante,

Notation 3.1 Pour un processus stochastique $\{X(t,\omega) : t \geqslant 0\}$, la dépendance de $X(t,\omega)$ en ω sera dorénavant supposée implicitement ; nous écrirons donc $X(t)$ en lieu et place de $X(t,\omega)$.

Commençons par donner la définition « probabiliste » d'un mouvement brownien.

Définition 3.2 Un processus stochastique $\{B(t) : t \geqslant 0\}$ défini sur un espace de probabilité $(\Omega,\mathcal{B},\mathrm{P})$ est appelé *mouvement brownien* (ou encore *processus de Wiener*) s'il satisfait les trois conditions suivantes :

1. $B(0) = 0$ presque sûrement,
2. le système $\{B(t) : t > 0\}$ est gaussien sur $(\Omega,\mathcal{B},\mathrm{P})$,
3. si t et $t+\delta$ sont positifs, l'incrément $B(t+\delta) - B(t)$ est de moyenne nulle et de variance $|\delta|$.

En particulier, ce processus est de moyenne nulle.

Suivant la condition 3, la fonction de distribution d'un mouvement brownien est associée à une gaussienne.

Remarque 3.3 De la définition du mouvement brownien, on obtient, pour tout Δ et

tout δ positif,

$$P(B(\,.\,+\delta)-B(\,.\,)\leqslant\Delta)=\frac{1}{\sqrt{2\pi\delta}}\int_{-\infty}^{\Delta}\exp(-\frac{t^2}{2\delta})\,dt. \tag{3.1}$$

□

Pour un tel processus*, la connaissance de la distribution de $\{B(t):t\geqslant 0\}$ se résume à celle de la fonction de covariance et cette dernière peut se déterminer aisément. On trouve, en utilisant le point 3 de la définition 3.2,

$$\begin{aligned}\mathrm{E}\big(B(t_1)B(t_2)\big) &= \frac{1}{2}\big(\mathrm{E}B^2(t_1)+\mathrm{E}B^2(t_2)-\mathrm{E}\big(B(t_1)-B(t_2)\big)^2\big)\\ &= \frac{1}{2}(t_1+t_2-|t_1-t_2|) \qquad (3.2)\\ &= \min\{t_1,t_2\}. \qquad (3.3)\end{aligned}$$

Comme $\mathrm{E}B(t)=0$, l'égalité (3.2) entraîne la condition 3 dans cette même définition.

L'égalité (3.2) montre clairement le caractère non stationnaire du mouvement brownien, puisque $\mathrm{E}\big(B(t+\delta)B(t)\big)=t$ pour tous $t,\delta\geqslant 0$. On a de plus les propriétés suivantes.

Proposition 3.4 *Si $\delta>0$ et $a\neq 0$ sont des nombres réels fixés,*

- *les processus $\{B(t+\delta)-B(\delta):t\geqslant 0\}$ et $\{B(a^2t)/a:t\geqslant 0\}$ sont des mouvements browniens ; c'est en particulier le cas pour $\{-B(t):t\geqslant 0\}$,*
- *les processus $\{B(t):t>0\}$ et $\{tB(1/t):t>0\}$ ont la même distribution.*

Le théorème fondamental suivant permet d'introduire la mesure associée à un mouvement brownien, grâce notamment au théorème d'extension de Kolmogorov et ainsi de prouver l'existence du mouvement brownien.

Théorème 3.5 *Définissons le cylindre*

$$C_{\mathcal{B}_n}(x_1,...,x_n)=\{f\in C^0([0,+\infty[):\big(f(x_1),...,f(x_n)\big)\in\mathcal{B}_n\},$$

pour tout ensemble borélien $\mathcal{B}_n$ à n dimensions, tous les réels $0\leqslant x_1<x_2<...<x_n$ et notons $\mathfrak{B}$ la plus petite σ-algèbre contenant le sous-ensemble

$$\{C_{\mathcal{B}_n}(x_1,...,x_n):0\leqslant x_1<x_2<...<x_n\}$$

de $C^0([0,+\infty[)$. Il existe une mesure de probabilité unique $\mathcal{W}$ sur $(C^0([0,+\infty),\mathfrak{B})$ satisfaisant

$$P\big((B(t_1),...,B(t_n))\in\mathcal{B}_n\big)=\mathcal{W}\big(C_{\mathcal{B}_n}(t_1,...,t_n)\big),$$

pour tout ensemble borélien $\mathcal{B}_n$ n-dimensionnel et les réels $0\leqslant t_1<t_2<...<t_n$.

*. Un processus gaussien de moyenne nulle

Ce résultat implique une importante propriété, à savoir la continuité du mouvement brownien au sens stochastique[⁂],

Corollaire 3.6 *Soit* $\{B(t,\omega) : t \geqslant 0\}$ *un mouvement brownien défini sur* $(\Omega,\mathcal{B},P)$*. Pour presque tout* $\omega \in \Omega$*, la réalisation* $t \mapsto B(t,\omega)$ *est une fonction continue.*

L'étape suivante concerne naturellement la dérivabilité d'un tel processus. Ici la réponse est négative.

Théorème 3.7 *Soit* $\{B(t,\omega) : t \geqslant 0\}$ *un mouvement brownien défini sur* $(\Omega,\mathcal{B},P)$*. Pour presque tout* $\omega \in \Omega$*, la réalisation* $t \mapsto B(t,\omega)$ *n'est nulle part dérivable.*

La réalisation d'un mouvement brownien possède un exposant de Hölder égal à $1/2$ presque sûrement. On qualifie le mouvement brownien de processus monofractal[⁑].

Théorème 3.8 *Pour tout* α *appartenant à l'intervalle* $]0,1/2[$*, il existe deux constantes* δ_α *et* C_α *telles que la réalisation d'un mouvement brownien défini dans* $[0,1]$ *et d'image* $\mathbb{R}$ *satisfasse la relation*

$$|B(t+\delta) - B(t)| \leqslant C_\alpha |\delta|^\alpha \tag{3.4}$$

lorsque $|\delta| < \delta_\alpha$ *presque sûrement.*

Finalement, le résultat suivant donne la dimension du graphe d'une réalisation d'un mouvement brownien.

Théorème 3.9 *Une réalisation d'un mouvement brownien défini dans* $[0,1]$ *et d'image* $\mathbb{R}$ *possède des dimensions de Hausdorff et Minkowski égales à* $3/2$ *presque sûrement.*

Ces deux derniers résultats seront généralisés par la suite.

Finalement, notons qu'il est parfois utile de relaxer l'hypothèse 3 de la définition 3.2 du mouvement brownien en permettant à la variance des incréments $B(t+\delta) - B(t)$ d'être proportionnelle à δ :

3′. si t et $t+\delta$ sont positifs, l'incrément $B(t+\delta) - B(t)$ est de moyenne nulle et de variance $|\delta|\mathrm{E}B^2(1)$.

⁂. Plus précisément, le critère de Kolmogorov permet de montrer qu'il existe une modification du mouvement brownien, *i.e.* un second processus qui en chaque point est égal au premier presque sûrement, dont les réalisations sont continues [183].

⁑. Plus précisément, il s'agit d'un processus mono-Hölder (*cf.* remarque 2.59).

Processus stochastiques auto-similaires

Pour des processus stochastiques, l'auto-similarité déterministe n'a pas beaucoup de sens. L'auto-similarité statistique consiste à imposer cette propriété aux distributions associées et non plus aux réalisations. Le mouvement Brownien est un exemple de processus auto-similaire.

Avant toute chose, introduisons la notation suivante.

Remarque 3.10 Dans ce chapitre, nous ne considérerons que des processus définis sur $\mathbb{R}^+$, mais les développements qui suivent restent pour la plupart valables si l'on travaille sur $\mathbb{R}$ ou $\mathbb{R}_*^+$. □

La notion d'auto-similarité que nous allons maintenant présenter nous permettra d'introduire le *mouvement brownien fractionnaire* dans la prochaine section.

Définition 3.11 Un processus stochastique à valeurs réelles $\{X(t) : t \geqslant 0\}$ est *auto-similaire* d'exposant $H > 0$ si, pour tout $a > 0$, les distributions de dimension finie de $\{X(at) : t \geqslant 0\}$ sont identiques aux distributions de dimension finie de $\{a^H X(t) : t \geqslant 0\}$. Si le processus est à valeurs complexes, il est dit *auto-similaire* d'exposant $H > 0$ si ses parties réelle et imaginaire le sont.

Faisons deux remarques. Un processus auto-similaire non-dégénéré ne peut être stationnaire. Si le processus $\{X(t) : t \geqslant 0\}$ est auto-similaire d'exposant H, alors pour tout $a > 0$, $X(0)$ possède les mêmes distributions que $a^H X(0)$. On en déduit donc que $X(0)$ doit valoir 0 presque sûrement.

Exemple 3.12 Le mouvement brownien est auto-similaire d'exposant $1/2$. De fait,

$$\mathrm{E}\big(B(at_1)B(at_2)\big) = a\min\{t_1,t_2\} = \mathrm{E}\big(\sqrt{a}B(t_1)\sqrt{a}B(t_2)\big), \tag{3.5}$$

pour tous $a > 0$ et t_1, $t_2 \geqslant 0$. □

3.2 Le mouvement brownien fractionnaire

Le mouvement brownien fractionnaire généralise le mouvement brownien traditionnel grâce à des considérations sur la fonction d'auto-covariance. Cette classe de marches aléatoires est d'une grande importance dans la modélisation de toute une série de phénomènes présentant des corrélations.

Du mouvement brownien au mouvement brownien fractionnaire

Le mouvement brownien est un processus possédant une fonction d'auto-covariance typique. Une généralisation naturelle, qui consiste à permettre aux exposants de cette fonction de varier, fournit ce que l'on appelle le mouvement brownien fractionnaire.

Cette section repose sur les deux observations suivantes.

Proposition 3.13 *Soit $0 < H \leqslant 1$. Il existe un processus gaussien $\{X(t) : t \geqslant 0\}$ de moyenne nulle et de fonction d'auto-covariance*

$$\mathrm{E}\big(X(t_1)X(t_2)\big) = \frac{1}{2}(t_1^{2H} + t_2^{2H} - |t_1 - t_2|^{2H})\,\mathrm{E}X^2(1). \tag{3.6}$$

On a tôt fait de vérifier que ces processus sont auto-similaires d'exposant H et à incréments stationnaires. En fait il n'existe pas d'autre processus gaussien auto-similaire d'exposant $H < 1$. De plus, le résultat suivant montre aussi que le cas $H = 1$ peut être considéré comme un cas dégénéré.

Proposition 3.14 *Soit $\{X(t) : t \geqslant 0\}$ un processus gaussien auto-similaire d'exposant H à incréments stationnaires non-dégénéré. Pour $0 < H \leqslant 1$, $X(0) = 0$ presque sûrement et la fonction de covariance d'un tel processus est définie par l'égalité (3.6). De plus, $0 < H < 1$ implique $\mathrm{E}X(t) = 0$ et pour le cas $H = 1$, on a $X(t) = tX(1)$ presque sûrement.*

Remarque 3.15 Cette propriété est fausse en dimension plus élevée. Il existe des champs anisotropes auto-similaires à incréments stationnaires [71]. □

Pour un H fixé, ces processus diffèrent donc par une constante multiplicative (et par leur moyenne si $H = 1$). La définition suivante est donc licite.

Définition 3.16 Un processus gaussien auto-similaire d'exposant H à incréments stationnaires est appelé *mouvement brownien fractionnaire* d'exposant H. Un tel processus sera noté $\{B_H(t) : t \geqslant 0\}$. On parlera de *mouvement brownien fractionnaire standard* lorsque $\mathrm{E}B_H^2(1) = 1$.

Pour $H = 1/2$, on trouve immédiatement $\mathrm{E}\big(B_{1/2}(t_1)B_{1/2}(t_2)\big) = \min\{t_1,t_2\}\,\mathrm{E}X^2(1)$, pour tout $t_1,t_2 \geqslant 0$. Le mouvement brownien fractionnaire généralise donc le mouvement brownien traditionnel. La proposition 3.14 implique aussi que $\{B_1(t) : t \geqslant 0\}$ ait la même

distribution que $\{t(Z+m) : t \geqslant 0\}$, où $m \in \mathbb{R}$ et Z est une variable aléatoire gaussienne de moyenne nulle.

De manière équivalente, on peut introduire le mouvement brownien fractionnaire par une définition du type 3.2,

Définition 3.17 Soit $0 < H < 1$. Un processus stochastique $\{B_H(t) : t \geqslant 0\}$ est appelé *mouvement brownien fractionnaire* d'indice H s'il satisfait les trois conditions suivantes :

1. $B_H(0) = 0$ presque sûrement,
2. le système $\{B_H(t) : t \geqslant 0\}$ est gaussien,
3. si t et $t+\delta$ sont positifs, l'incrément $B_H(t+\delta) - B_H(t)$ est de moyenne nulle et de variance $|\delta|^{2H}\mathrm{E}B_H^2(1)$.

Concernant la fonction de distribution, la remarque 3.3 se généralise comme suit.

Remarque 3.18 Les incréments $B_H(t+\delta) - B_H(t)$ d'un mouvement brownien fractionnaire standard d'exposant H sont de moyenne nulle et de variance δ^{2H}. Ainsi,

$$P(B_H(\,.\,+\delta) - B_H(\,.\,) \leqslant \Delta) = \frac{1}{\sqrt{2\pi}\delta^H} \int_{-\infty}^{\Delta} \exp(-\frac{t^2}{2\delta^{2H}})\, dt. \tag{3.7}$$

□

Concernant la régularité, le résultat suivant généralise ceux obtenus pour le mouvement brownien.

Théorème 3.19 *Soit B_H la réalisation d'un mouvement brownien fractionnaire d'exposant H défini dans* [0,1] *et d'image* $\mathbb{R}$. *Pour tout α appartenant à l'intervalle*]0,H[, *il existe des constantes δ_α et C_α telles que la relation*

$$|B_H(t+\delta) - B_H(t)| \leqslant C_\alpha |\delta|^\alpha \tag{3.8}$$

soit satisfaite pour tout $|\delta| < \delta_\alpha$ presque sûrement.

Ainsi, l'exposant de Hölder local d'une réalisation d'un mouvement brownien fractionnaire d'indice H est H presque sûrement. Tout comme le mouvement brownien, le mouvement brownien fractionnaire est monofractal.

Théorème 3.20 *Une réalisation d'un mouvement brownien fractionnaire défini de l'intervalle* [0,1] *vers* $\mathbb{R}$ *possède des dimensions de Hausdorff et Minkowski égales à $2-H$ presque sûrement.*

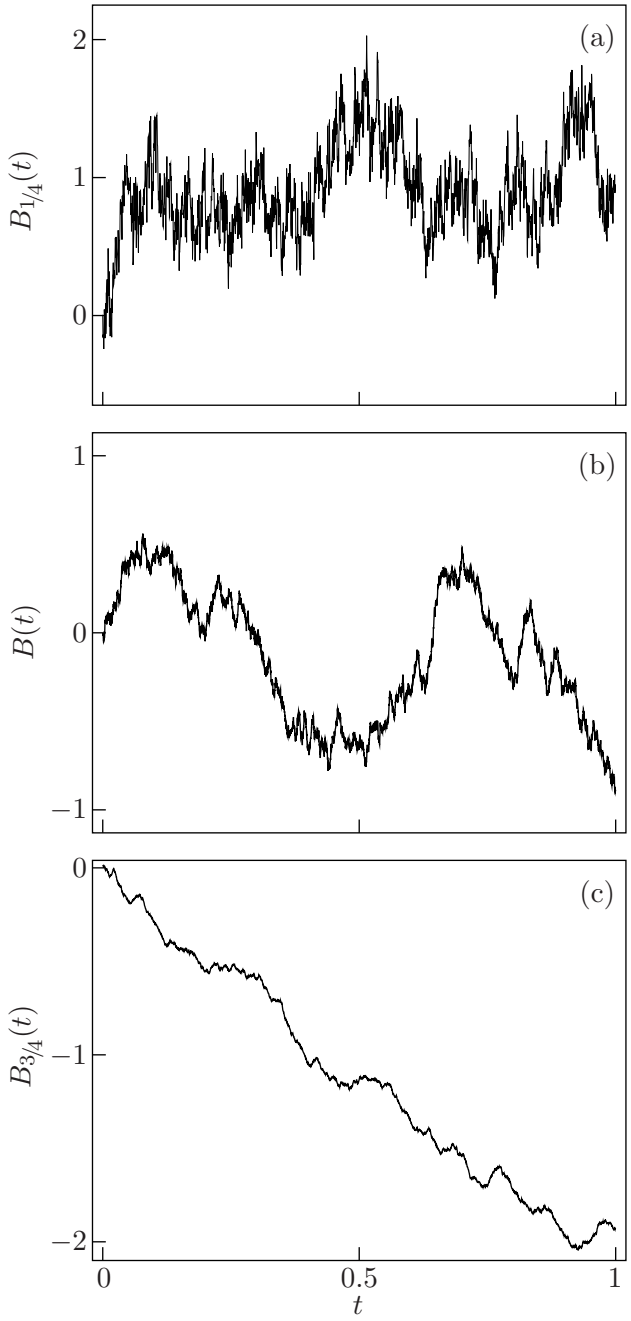

FIG. 3.1 – *Simulations numériques de mouvements browniens fractionnaires avec $H = 1/4$ (a), $H = 1/2$ (b) (il s'agit donc d'un mouvement brownien) et $H = 3/4$ (c), en utilisant la méthode des matrices circulantes. On constate clairement que la régularité augmente avec l'indice H.*

Corrélations à longue portée

Les corrélations à longue portée sont la principale caractéristique du mouvement brownien fractionnaire. Le mouvement brownien traditionnel se démarque des autres en ce sens qu'il ne présente pas de corrélations.

Présentons d'abord quelques notions essentielles.

Définition 3.21 Comme un mouvement brownien fractionnaire $\{B_H(t) : t \geqslant 0\}$ est à incréments stationnaires, si l'on pose

$$\Delta_H(j) = \Delta_1^1 B_H(j) = B_H(j+1) - B_H(j) \qquad (j \in \mathbb{N}), \tag{3.9}$$

le processus $\{\Delta_H(j) : j \in \mathbb{N}\}$ définit un processus stationnaire appelé *bruit gaussien fractionnaire*. Il est appelé *bruit gaussien fractionnaire standard* si $\mathrm{E}\Delta_H^2(j)$ vaut 1.

Un bruit gaussien fractionnaire est donc une suite gaussienne stationnaire de moyenne nulle et de variance $\mathrm{E}\Delta_H^2(j) = \mathrm{E}B_H^2(1)$, pour tout j naturel. Intéressons-nous maintenant à la fonction de covariance d'une telle suite.

Notation 3.22 La fonction de corrélation d'un bruit gaussien stationnaire sera notée

$$\gamma(j) = \mathrm{E}\big(\Delta_H(0)\Delta_H(j)\big). \tag{3.10}$$

Pour simplifier les notations, nous supposerons avoir affaire à un bruit gaussien fractionnaire standard.

On peut expliciter la fonction de corrélation en termes de j :

$$\gamma(j) = \mathrm{E}\Big(B_H(1)\big(B_H(j+1) - B_H(j)\big)\Big) = \frac{1}{2}\big((j+1)^{2H} - 2j^{2H} + |j-1|^{2H}\big). \tag{3.11}$$

Si $H = 1/2$, $\gamma(j) = 0$ pour tout j non-nul et donc $\{\Delta_{1/2}(j) : j \in \mathbb{N}\}$ est une suite gaussienne indépendante. Lorsque H est différent de $1/2$, on a le résultat suivant, aisément démontré en utilisant la règle de l'Hospital.

Proposition 3.23 *Si* $H \neq 1/2$,

$$\gamma(j) \sim \mathrm{E}B_H^2(1)H(2H-1)j^{2H-2}, \tag{3.12}$$

lorsque j *tend vers l'infini.*

Lorsque $H \neq 1/2$, la suite $\{\gamma(j)\}$ tend vers zéro, mais lorsque $1/2 < H < 1$, la série associée diverge. On dit généralement que le bruit gaussien fractionnaire possède des *dépendances à longue portée.*

Définition 3.24 Un processus à incréments stationnaires $\{X(t) : t \geqslant 0\}$ présente des *corrélations à longue portée* si la fonction de corrélation du bruit associé possède un comportement asymptotique du type

$$E\Big(\big(X(j+1)-X(j)\big)\big(X(1)-X(0)\big)\Big) \sim cj^k, \tag{3.13}$$

où c est strictement positif et $-1 < k < 0$.

Lorsque $0 < H < 1/2$, la série associée converge, mais comme le coefficient $H(2H-1)$ est strictement négatif, $\gamma(j)$ est lui aussi, strictement négatif pour les grandes valeurs de j. On parle de *dépendance négative*.

Remarque 3.25 Le mouvement brownien fractionnaire est auto-similaire mais pas stationnaire. En définissant le bruit gaussien fractionnaire, on définit un processus stationnaire mais qui n'est plus auto-similaire. □

Les dépendances à longue portée et les dépendances négatives peuvent être considérées comme une des propriétés majeures du mouvement brownien fractionnaire, comme en atteste la proposition suivante. Ce résultat explique aussi, d'une certaine manière, pourquoi le mouvement brownien fractionnaire est rencontré couramment dans la nature.

Proposition 3.26 *Soit $0 < H < 1$. Si $\{Z_j\}_{j\in\mathbb{N}}$ est une suite gaussienne stationnaire de moyenne nulle et de fonction d'auto-covariance $\gamma(j) = \mathrm{E}(Z_0 Z_j)$ satisfaisant*

- *$\gamma(j) \sim cj^{2H-2}$, avec $c > 0$ si $1/2 < H < 1$,*
- *$\sum_j \gamma(j) = 0$ et $\gamma(j) \sim cj^{2H-2}$, avec $c < 0$ si $0 < H < 1/2$,*
- *$\sum_j |\gamma(j)| < \infty$ et $\sum_j \gamma(j) = c > 0$ si $H = 1/2$,*

alors

$$\Big\{\frac{1}{k^H}\sum_{j=0}^{\lfloor kt\rfloor} Z_j : 0 \leqslant t \leqslant 1\Big\}$$

converge en distribution vers

$$\{\sigma^2 B_H(t) : 0 \leqslant t \leqslant 1\},$$

où $\{B_H(t) : 0 \leqslant t \leqslant 1\}$ est un mouvement brownien fractionnaire standard et où

$$\sigma^2 = \begin{cases} \big(H(2H-1)\big)^{-1} c & \text{si } 1/2 < H < 1, \\ -\big(H(2H-1)\big)^{-1} c & \text{si } 0 < H < 1/2, \\ c & \text{si } H = 1/2. \end{cases} \tag{3.14}$$

Réalisation numérique d'un mouvement brownien fractionnaire

Nous exposons dans cette section une méthode exacte d'ordre $N \log N$ pour générer un mouvement brownien fractionnaire. La méthode exposée est quelque peu plus générale puisqu'elle permet la synthèse d'un processus gaussien stationnaire dont la matrice de covariance est donnée. Il est ainsi possible de construire un bruit gaussien fractionnaire, qu'il restera à sommer pour obtenir le processus désiré. Aucune méthode, même approchée, n'est de complexité plus faible. Pour cette raison, elle est la seule à être présentée ici. Cette méthode à été initialement proposée par Chan et Wood [95, 96, 393]. Afin de simplifier les notations, Z désignera une suite gaussienne centrée et réduite dont les éléments sont indépendants.

Nous voulons obtenir, à partir d'une suite gaussienne Z, une réalisation de m points $\{X(j/m) : 0 \leqslant j < m\}$ d'un processus gaussien stationnaire dont la fonction de covariance est spécifiée par γ. La matrice de covariance associée est une matrice de Toeplitz symétrique définie positive[❋] mais pas nécessairement circulante. L'idée [95, 393] est de considérer la matrice de dimension $m' \times m'$ avec $m' \geqslant 2(m-1)$ définie comme suit,

$$\Sigma = \begin{pmatrix} \sigma_0 & \sigma_1 & \cdots & \sigma_{m'-1} \\ \sigma_{m'-1} & \sigma_0 & \cdots & \sigma_{m'-2} \\ \vdots & \vdots & & \vdots \\ \sigma_1 & \sigma_2 & \cdots & \sigma_0 \end{pmatrix}, \tag{3.15}$$

où

$$\sigma_j = \begin{cases} \gamma(\frac{j}{m}) & \text{si } 0 \leqslant j \leqslant m'/2, \\ \gamma(\frac{m'-j}{m}) & \text{si } m'/2 \leqslant j \leqslant m'-1. \end{cases} \tag{3.16}$$

Si la matrice Σ est définie positive, elle constitue la matrice de covariance d'un processus gaussien stationnaire en m' points dont m éléments consécutifs forment une réalisation de structure de covariance désirée.

Si Λ représente une matrice diagonale associée à Σ, cette dernière peut se mettre sous la forme $\Sigma = Q\Lambda Q^*$. Pour une matrice circulante [114, 169], Q est unitaire, les valeurs propres $\{\lambda_k\}$ de Σ sont les transformées de Fourier discrètes d'une colonne de Σ,

$$\lambda_k = \sum_{j=0}^{m'-1} (\Sigma)_{j+1,1} \exp(-2i\pi \frac{jk}{m'}) \tag{3.17}$$

et pour tout vecteur $\vec{v}$, $Q\vec{v}$ est la transformée de Fourier de ce même vecteur. On constate immédiatement que le processus $Q\Lambda^{1/2}Q^*Z$ possède Σ comme matrice de covariance.

❋. Entendons par là hermitienne et de valeurs propres non-négatives

Réalisation d'un mouvement brownien fractionnaire
1. définir la fonction de covariance γ et la taille m du processus voulue,
2. calculer le plus petit naturel k tel que $2^k \geqslant 2(m-1)$,
3. définir la première colonne $\Sigma_1 \in \mathbb{R}^{2^k}$ de la matrice Σ définie par l'égalité (3.15),
4. calculer les valeurs propres de Σ en calculant la transformée de Fourier de Σ_1,
5. calculer les racines de ces valeurs propres,
6. générer deux processus gaussiens centrés réduits Z_1 et Z_2,
7.. simuler Q^*Z en construisant un signal du type $Z_1 + iZ_2$,
8. définir le vecteur $\vec{v} = \Lambda^{1/2}Q^*Z$, où Λ est la matrice diagonale associée à Σ,
9. définir X en calculant la transformée de Fourier rapide de $\vec{v}$,
10. extraire m éléments consécutifs de la partie réelle de X,
11. sommer ces éléments.

TAB. 3.1 – *L'algorithme des matrices circulantes permettant de générer un mouvement brownien fractionnaire sur l'intervalle* [0,1[. *Cette méthode est exacte d'ordre* $2^k \log 2^k$ *et permet de générer n'importe quel processus gaussien stationnaire.*

La réalisation désirée peut alors facilement s'obtenir. Les valeurs propres de Σ (et donc de Λ) se calculent aisément grâce à la relation (3.17). Si Z_1 et Z_2 représentent deux processus gaussiens centrés réduits indépendants, on peut construire $Z_0 = Z_1 + iZ_2$, puis $\Lambda^{1/2}Z_0$. Finalement, la transformée de Fourier appliquée à ce vecteur nous donne une solution de la forme* $Q\Lambda^{1/2}Q^*Z$. Il reste à prendre m éléments consécutifs dans $Q\Lambda^{1/2}Q^*Z$.

En règle générale, rien ne nous assure que la matrice circulante Σ soit définie positive, ce qui peut induire des valeurs propres négatives. Une première méthode consiste à ne pas prendre ces valeurs négatives, en les remplaçant par zéro. Une méthode plus élégante consiste à prendre une dimension m' plus élevée et à recommencer le processus. Toutefois, ici, nous sommes intéressés par la construction d'un bruit gaussien fractionnaire. Les éléments de la matrice de variance-covariance sont donc donnés par l'égalité (3.11). Dans ce cas, on a le résultat suivant [120].

Proposition 3.27 *La matrice* Σ *définie par l'égalité (3.15) construite à partir de la fonction de corrélation d'un bruit gaussien fractionnaire est une matrice définie positive.*

Cette méthode de réalisation ne pose donc jamais problème. Le tableau 3.1 récapitule l'algorithme décrit dans cette section.

∗. Il y a en fait deux solutions : il suffit de prendre la partie réelle et la partie imaginaire

3.3 Marches binaires construites à partir de mouvements browniens fractionnaires

Le but de la présente section est l'élaboration d'un modèle de marche discrète binaire, *i.e.* de marche aléatoire discrète constituée de sauts d'amplitude unité, dont le bruit associé présente des corrélations à longue portée. Pour ce faire, nous partons du bruit gaussien fractionnaire. Cette construction nous permettra, dans la suite, de générer des séquences ADN artificielles dont les codages seront corrélés à longue portée, ce qui justifiera *a posteriori* , dans une certaine mesure, l'appellation « mouvement brownien fractionnaire » donnée à certaines marches binaires rencontrées en biologie. Nous tenons à remercier Alain Arneodo et Benjamin Audit pour les discussions concernant ce sujet.

Marches binaires discrètes de moyenne nulle

Nous montrons ici que les corrélations à longue portée d'un bruit gaussien fractionnaire résident essentiellement dans le signe des incréments. Fort de ce résultat, nous introduisons ensuite une marche binaire issue du mouvement brownien fractionnaire.

Nous allons considérer la suite $\{\delta_H(j)\}_j$ obtenue à partir d'un bruit gaussien fractionnaire $\{\Delta_H(j)\}_j$ dont on ne conserve que les signes, $\delta_H(j) = \text{sign}\Delta_H(j)$. Si la normalité n'est pas préservée, il n'en va pas de même pour les corrélations. La démonstration de cette propriété repose sur le résultat suivant.

Proposition 3.28 *Soit un vecteur bi-normal* $\vec{z} = (z_1,z_2)$ *de moyenne* $(0,0)$ *dont la matrice de covariance est donnée par*

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

avec $|\rho| < 1$. *On a*

$$\text{corr}(\text{sign}(z_1),\text{sign}(z_2)) = \frac{2}{\pi}\arcsin\rho. \tag{3.18}$$

Preuve. Remarquons d'abord que si z est une variable gaussienne de moyenne nulle et de

variance unité, alors $\mathrm{E}\mathrm{sign}(z) = 0$. En utilisant les coordonnées polaires, on obtient alors

$$\begin{aligned}
\mathrm{E}\big(\mathrm{sign}(v_1)\mathrm{sign}(v_2)\big) &= \iint_{\mathbb{R}^2} \mathrm{sign}(x_1)\,\mathrm{sign}(x_2)\,\frac{\exp(-\frac{x_1^2+x_2^2-2\rho x_1 x_2}{2(1-\rho^2)})}{2\pi\sqrt{1-\rho^2}}\,dx_1\,dx_2 \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_0^{\pi/2}\int_{\mathbb{R}^+} \exp(-\frac{r(1-\rho\sin 2\theta)}{2(1-\rho^2)}) \\
&\quad - \exp(-\frac{r(1+\rho\sin 2\theta)}{2(1-\rho^2)})\,dr\,d\theta \\
&= \frac{2\sqrt{1-\rho^2}}{\pi}\int_0^{\pi/2}\frac{\rho\sin 2\theta}{1-\rho^2\sin^2 2\theta}\,d\theta \\
&= \frac{2}{\pi}\mathrm{arctg}\frac{\rho}{\sqrt{1-\rho^2}},
\end{aligned}$$

ce qui permet de conclure. □

Par définition d'un mouvement brownien fractionnaire, $\mathrm{E}\big(\Delta_H(0)\Delta_H(j)\big) \sim cj^{2H-2}$, pour une constante c. Supposons maintenant que le bruit gaussien fractionnaire soit standard. Puisque $\arcsin\rho = \rho + \rho^3/6 + \mathcal{O}(\rho^5)$, on a aussi, grâce à la proposition 3.28, $\mathrm{E}\big(\delta_H(0)\delta_H(j)\big) \sim c'j^{2H-2}$, pour une constante $c' = 2c/\pi$. Puisque la fonction sign est insensible à la multiplication de l'argument par une constante positive, nous venons d'obtenir le corollaire suivant,

Corollaire 3.29 *La suite $\{\delta_H(j) = \mathrm{sign}\Delta_H(j)\}_j$ présente les même corrélations à longue portée que le bruit gaussien fractionnaire $\{\Delta_H(j)\}_j$ dont elle est issue, $\mathrm{E}\big(\delta_H(0)\delta_H(j)\big) \sim cj^{2H-2}$, pour une constante c.*

La question naturelle qui se pose maintenant est de savoir si la même propriété se répète pour les modules des éléments d'un bruit gaussien fractionnaire. Ici, la réponse est négative.

Proposition 3.30 *Soit un vecteur bi-normal $\vec{z} = (z_1,z_2)$ de moyenne $(0,0)$ dont la matrice de covariance est donnée par*

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

avec $|\rho| < 1$. On a

$$\mathrm{corr}(|z_1|,|z_2|) = \frac{2}{\pi-2}(\rho\arcsin\rho + \sqrt{1-\rho^2} - 1). \tag{3.19}$$

Preuve. Si z est une variable gaussienne de moyenne nulle et de variance unité, alors $\mathrm{E}|z| = \sqrt{2/\pi}$ et $\mathrm{E}|z|^2 = 1$. On obtient alors, en passant en coordonnées polaires,

$$\begin{aligned}
\mathrm{E}|z_1 z_2| &= \frac{1}{2\pi\sqrt{1-\rho^2}} \iint_{\mathbb{R}^2} |x_1|\,|x_2| \,\exp(-\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2(1-\rho^2)})\, dx_1\, dx_2 \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_0^{\pi/2} \int_{\mathbb{R}^+} \frac{r\sin 2\theta}{2} \exp(-\frac{r - \rho r \sin 2\theta}{2(1-\rho^2)})\, dr\, d\theta \\
&\quad + \frac{1}{2\pi\sqrt{1-\rho^2}} \int_0^{\pi/2} \int_{\mathbb{R}^+} \frac{r\sin 2\theta}{2} \exp(-\frac{r + \rho r \sin 2\theta}{2(1-\rho^2)})\, dr\, d\theta \\
&= \int_0^{\pi/2} \frac{\sin 2\theta}{\pi\sqrt{1-\rho^2}} \frac{(1-\rho^2)^2}{(1-\rho\sin 2\theta)^2}\, d\theta \\
&\quad + \int_0^{\pi/2} \frac{\sin 2\theta}{\pi\sqrt{1-\rho^2}} \frac{(1-\rho^2)^2}{(1+\rho\sin 2\theta)^2}\, d\theta \\
&= \frac{2\rho}{\pi} \mathrm{arctg} \frac{\rho}{\sqrt{1-\rho^2}} + \frac{2\sqrt{1-\rho^2}}{\pi},
\end{aligned}$$

ce qui est suffisant pour conclure. □

Puisque $\rho \arcsin \rho + \sqrt{1-\rho^2} - 1 = \rho^2/2 + \mathcal{O}(\rho^4)$, nous n'avons pas l'analogue de la proposition 3.29.

Corollaire 3.31 *La suite $\{|\Delta_H(j)|\}_j$ ne présente pas les mêmes corrélations à longue portée que le bruit gaussien fractionnaire dont elle est issue,* $\mathrm{E}(|\Delta_H(0)||\Delta_H(j)|) \sim cj^{4H-4}$ *pour une constante c. Si $1/2 < H \leqslant 3/4$, $\{|\Delta_H(j)|\}_j$ ne présente pas de corrélation à longue portée. Par contre, si $H > 3/4$, la suite présente des corrélations à longue portée de nature différente : si $k = 2(H-1)$ désigne l'exposant associé aux corrélations du bruit gaussien fractionnaire dans la relation (3.13), l'exposant k' de la suite $\{|\Delta_H(j)|\}_j$ est $k' = 2k = 2(H'-1)$, où $H' = 2H - 1$.*

Ainsi, la suite des modules d'un bruit gaussien fortement corrélé à longue portée reste corrélée à longue portée, alors que ce n'est pas le cas pour les corrélations faibles.

Étant donné un mouvement brownien fractionnaire $\{B_H(t) : t \geqslant 0\}$, nous pouvons donc construire une marche discrète à pas entiers $\{b_H(j) : j \in \mathbb{N}\}$ dont le bruit associé prend ses valeurs sur l'ensemble $\{-1,1\}$ et présente des corrélations à longue portée, en posant

$$b_H(j) = \sum_{k=0}^{j} \delta_H(k) = \sum_{k=0}^{j} \mathrm{sign}\big(B_H(k+1) - B_H(k)\big). \tag{3.20}$$

Cette marche sera appelée *marche binaire*. La remarque suivante permet de considérer une marche binaire comme une approximation d'un mouvement brownien fractionnaire.

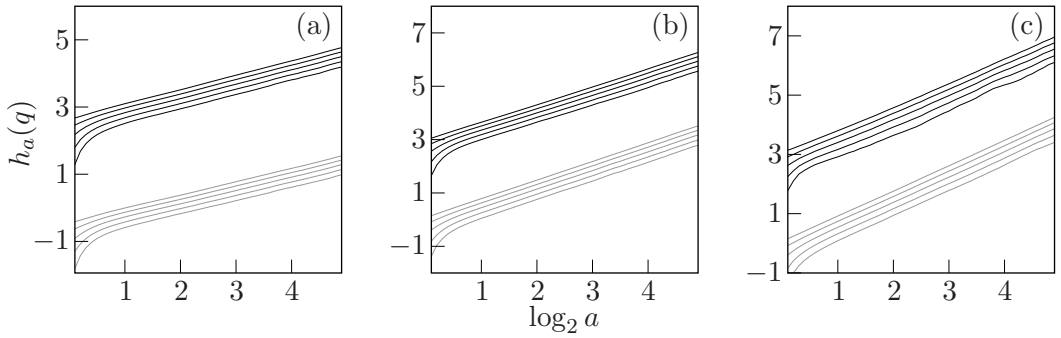

FIG. 3.2 – *Les fonctions $h_a(q)$ (en noir) obtenues par la méthode des maxima du module de la transformée en ondelettes appliquée à des réalisations de marches binaires b_H d'indice $H = 0.4$ (a), $H = 0.7$ (b) et $H = 0.85$ (c) respectivement. On constate clairement un comportement linéaire avec un coefficient angulaire constant pour chaque réalisation proche de l'indice H. Ce comportement est donc celui des fonctions $h_a(q)$ relatives aux mouvements browniens fractionnaires dont sont issues les marches binaires. Ces dernières sont représentées en gris et ont été décalées verticalement pour plus de clarté. L'ondelette mère utilisée est la dérivée seconde de la gaussienne et $h_a(q)$ est représenté pour les valeurs de q suivantes, -1, -0.5, 0, 0.5 et 1.*

Remarque 3.32 La marche définie par l'égalité (3.20) normalisée de manière appropriée converge en distribution vers un mouvement brownien fractionnaire de même indice H. Il peut être montré [359] que si la fonction f est telle que $f\big(\Delta_H(j)\big)$ est de moyenne nulle et de variance finie, alors le processus

$$F_N(t) = \frac{1}{d_N} \sum_{j=1}^{\lfloor Nt \rfloor} f\big(\Delta_H(j)\big), \tag{3.21}$$

avec $0 \leqslant t \leqslant 1$ et où d_N^2 est asymptotiquement proportionnel à $\operatorname{var} \sum_{j=1}^{N} f\big(\Delta_H(j)\big)$, converge en distribution avec N vers un mouvement brownien fractionnaire d'indice H si le rang d'Hermite de f vaut 1 (*cf.* remarque 3.38). □

Illustrons la conservation des corrélations à longue portée par un exemple.

Exemple 3.33 La méthode des maxima du module de la transformée en ondelettes a été appliquée à la réalisation de trois marches binaires d'indice respectif $H = 0.4$, $H = 0.7$ et $H = 0.85$. Pour chacune des réalisations, on obtient des fonctions h_a linéaires en fonction de $\log_2 a$ et de même pente, comme l'illustre la figure 3.2. Le coefficient angulaire mesuré est similaire à l'indice H de la réalisation du brownien fractionnaire dont le signal est issu (légèrement surestimé pour $H = 0.4$ et sous-estimé pour $H = 0.7$ et $H = 0.85$). □

L'analyse multifractale d'une marche binaire b_H permet donc d'obtenir l'indice H (ceci s'explique notamment grâce à la remarque 3.32), traduisant la présence de corrélations

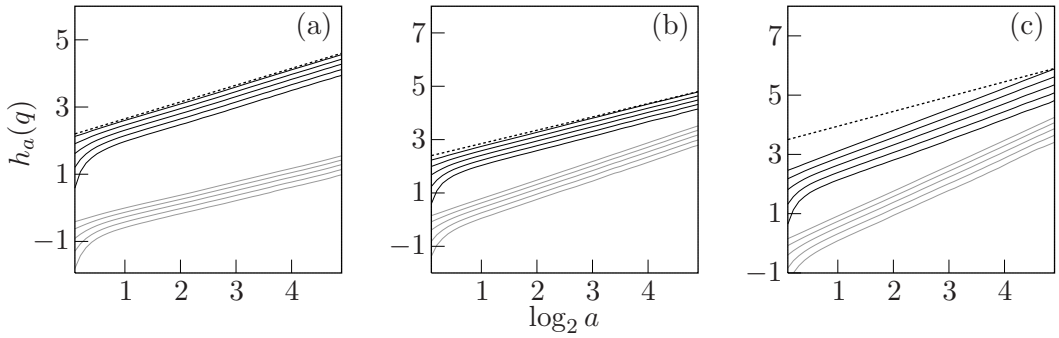

FIG. 3.3 – *Les fonctions $h_a(q)$ (en noir) obtenues par la méthode des maxima du module de la transformée en ondelettes appliquée à des signaux Θ_H d'indice $H = 0.4$ (a), $H = 0.7$ (b) et $H = 0.85$ (c) respectivement. On constate clairement un comportement linéaire avec un coefficient angulaire constant pour chaque réalisation. Pour les deux indices $H = 0.4$ et $H = 0.7$, le coefficient angulaire mesuré vaut 0.5, alors qu'il vaut 0.7 pour l'indice $H = 0.85$. Les fonctions $h_a(q)$ relatives aux mouvements browniens fractionnaires dont sont issues les signaux Θ_H sont représentées en gris et ont été décalées verticalement pour plus de clarté. Une droite de coefficient angulaire 0.5 est représentée en pointillés. L'ondelette mère utilisée est la dérivée seconde de la gaussienne et $h_a(q)$ est représenté pour les valeurs de q suivantes, -1, -0.5, 0, 0.5 et 1.*

à longue portée dans le signal. Remarquons que dans la littérature concernant la mesure expérimentale d'exposants de Hölder de signaux monofractals, l'exposant obtenu est presque systématiquement appelé « indice de corrélation ». Cet amalgame est tellement répendu que nous l'emploierons nous-même par la suite. Cependant, nous apporterons d'autres évidences de l'existence de corrélations à longue portée dans le signal étudié ; de plus, nous éviterons toute interprétaion physique concernant l'exposant de Hölder mesuré reposant sur la présence de telles corrélations.

Terminons par un exemple montrant que le corollaire 3.31 est vérifié numériquement.

Exemple 3.34 Notons $\Theta_H(j) = \sum_{k=0}^{j} |B_H(k+1) - B_H(k)|$. Pour trois réalisations de mouvements browniens fractionnaires d'indice respectif $H = 0.4$, $H = 0.7$ et $H = 0.85$ (il s'agit des même réalisations que celles utilisées dans l'exemple 3.33), le signal Θ_H correspondant a été construit. La méthode des maxima du module de la transformée en ondelettes appliquée à ces signaux permet d'obtenir des fonctions $h_a(q)$ linéaires de même coefficient angulaire pour chaque réalisation. La mesure de ces coefficients donne des valeurs en excellent accord celles attendues, comme le montre la figure 3.3 : pour les indices $H = 0.4$ et $H = 0.7$, le coefficient angulaire est proche de $1/2$, alors qu'il vaut approximativement 0.7 si $H = 0.85$. Remarquons que l'ondelette mère utilisée doit nécessairement posséder plus d'un moment nul, puisque le signal Θ_H est strictement croissant. □

Marches binaires discrètes de moyenne non nulle

Nous pouvons donc obtenir une marche discrète d'incrément unité (c'est-à-dire d'incrément égal à ± 1) corrélée à longue portée à partir d'un mouvement brownien fractionnaire. Nous allons maintenant modifier cette marche afin de pouvoir faire varier le rapport entre le nombre de sauts positifs et négatifs. Bien sûr cette démarche n'a d'intérêt que si l'on conserve les corrélations à longue portée.

Commençons par simplifier les notations.

Notation 3.35 Pour les incréments discrets δ_H, l'indice H sera dorénavant supposé explicite, $\delta(j) = \delta_H(j)$.

Si l'on souhaite pouvoir obtenir une suite binaire de moyenne non-nulle δ_α à partir d'un bruit gaussien fractionnaire, il est naturel de considérer la construction suivante,

$$\delta_\alpha(j) = \begin{cases} -1 & \text{si } \Delta_H(j) < \alpha, \\ +1 & \text{si } \Delta_H(j) \geqslant \alpha, \end{cases} \tag{3.22}$$

pour $j \in \mathbb{N}$ et où $\alpha \in \mathbb{R}$ est un paramètre définissant le rapport entre les incréments positifs et négatifs, *i.e.* la moyenne de la suite. La valeur à attribuer à α s'obtient facilement. En effet, la probabilité qu'un élément du bruit discret $\delta_\alpha(j)$ soit positif vaut $\mathrm{P}(\delta_\alpha(j) = 1) = 1 - d(\alpha)$, où d représente la fonction de distribution relative à la loi gaussienne d'écart-type σ,

$$d(t) = \frac{1}{2}\big(1 + \mathrm{erf}(\frac{t}{\sqrt{2}\sigma})\big). \tag{3.23}$$

La fonction de distribution d_δ associée à ce bruit discret est la suivante,

$$d_\delta(j) = \begin{cases} d(\alpha) & \text{si } j = -1, \\ 1 & \text{si } j = 1. \end{cases} \tag{3.24}$$

De là, on obtient

$$\mathrm{E}\delta_\alpha(j) = 1 - 2d(\alpha) \quad \text{et} \quad \mathrm{var}\delta_\alpha(j) = 4d(\alpha)\big(1 - d(\alpha)\big). \tag{3.25}$$

La *marche binaire de moyenne non nulle* associée à un bruit gaussien fractionnaire $\{\Delta_H(j) : j \in \mathbb{N}\}$ est ainsi donnée par l'égalité

$$b_\alpha(j) = \sum_{k=0}^{j} \delta_\alpha(k) = \sum_{k=0}^{j} \mathrm{sign}(\Delta_H(k) - \alpha). \tag{3.26}$$

Il nous faut maintenant montrer que cette construction préserve les corrélations à longue portée.

Proposition 3.36 *Le bruit discret défini par l'égalité (3.22) présente les mêmes corrélations à longue portée que le bruit gaussien fractionnaire dont il est issu.*

Preuve. Soit $\{\Delta_H(j) : j \in \mathbb{N}\}$ un bruit gaussien fractionnaire standard. Nous utiliserons la notation suivante, $m_\alpha = \mathrm{E}\,\mathrm{sign}(\Delta_H(j) - \alpha)$. En utilisant l'égalité (3.23), on obtient $D_\alpha m_\alpha = -\sqrt{2/\pi}\exp(-\alpha^2/2)$. Il est suffisant de montrer que, étant donné un vecteur binormal $\vec{z} = (z_1,z_2)$, le développement de Taylor de $\mathrm{E}\big(\mathrm{sign}(z_1 - \alpha)\mathrm{sign}(z_2 - \alpha)\big)$ possède un terme linéaire et que le terme constant est égal à m_α^2. Nous évaluerons aussi les termes quadratiques en vue de pouvoir contrôler le comportement numérique.

L'intégrale

$$\iint_{\mathbb{R}^2} \mathrm{sign}(x_1)\,\mathrm{sign}(x_2)\,\frac{\exp(-\frac{(x_1+\alpha)^2+(x_2+\alpha)^2-2\rho(x_1+\alpha)(x_2+\alpha)}{2(1-\rho^2)})}{2\pi\sqrt{1-\rho^2}}\,dx_1\,dx_2 \tag{3.27}$$

ne peut être évaluée par passage aux coordonnes polaires. En posant

$$g(x_1,x_2) = \frac{1}{2\pi}\exp\big(-\frac{(x_1+\alpha)^2+(x_2+\alpha)^2}{2}\big),$$

on obtient

$$\begin{aligned}
&\frac{1}{2\pi\sqrt{1-\rho^2}}\exp(-\frac{(x_1+\alpha)^2+(x_2+\alpha)^2-2\rho(x_1+\alpha)(x_2+\alpha)}{2(1-\rho^2)})\\
&= g(x_1,x_2) + g(x_1,x_2)(x_1+\alpha)(x_2+\alpha)\,\rho\\
&\quad + \Big(\frac{1}{2}g(x_1,x_2) + \frac{1}{2}g(x_1,x_2)\big((x_1+\alpha)^2(x_2+\alpha)^2 - (x_1+\alpha)^2 - (x_2+\alpha)^2\big)\Big)\,\rho^2\\
&\quad + \mathcal{O}(\rho^3).
\end{aligned}$$

Par le théorème de la convergence dominée, il est suffisant d'évaluer l'intégrale portant sur ces termes pour obtenir les termes constants, linéaires et quadratiques.

Par définition,

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{sign}(x)\,\exp(-\frac{1}{2}(x+\alpha)^2)\,dx = m_\alpha.$$

De plus,

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} |x|\,\exp(-\frac{1}{2}(x+\alpha)^2)\,dx &= \frac{\exp(-\alpha/2)}{\sqrt{2\pi}}\int_{\mathbb{R}} |x|\,\exp\big(-\frac{1}{2}(x^2+2\alpha x)\big)\,dx\\
&= \frac{\exp(-\alpha/2)}{\sqrt{2\pi}}(2 - \alpha\exp(\frac{\alpha^2}{2})\sqrt{2\pi}m_\alpha).
\end{aligned}$$

De la même manière,

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{sign}(x)\,x^2\,\exp(-\frac{1}{2}(x+\alpha)^2)\,dx = \frac{\exp(-\alpha/2)}{\sqrt{2\pi}}\big((1+\alpha^2)\exp(\frac{\alpha^2}{2})\sqrt{2\pi}m_\alpha - 2\alpha\big).$$

Pour les termes linéaires, on trouve,

$$\begin{aligned}
&\iint_{\mathbb{R}^2} \operatorname{sign}(x_1)\operatorname{sign}(x_2)\,(x_1+\alpha)(x_2+\alpha)\,g(x_1,x_2)\,dx_1\,dx_2 \\
&= \sqrt{\frac{8}{\pi}}\,m_\alpha\alpha\exp(-\frac{\alpha^2}{2}) + \frac{1}{2\pi}\exp(-\alpha^2)\,(2-\alpha\exp(\frac{\alpha^2}{2})\sqrt{2\pi}m_\alpha)^2 - m_\alpha^2\alpha^2 \\
&= \frac{2}{\pi}\exp(-\alpha^2).
\end{aligned}$$

Pour les termes quadratiques,

$$\begin{aligned}
&\iint_{\mathbb{R}^2} \operatorname{sign}(x_1)\operatorname{sign}(x_2)\,\Big(\frac{1}{2} - \frac{1}{2}\big((x_1+\alpha)^2 + (x_2+\alpha)^2 \\
&- (x_1+\alpha)^2(x_2+\alpha)^2\big)\Big)\,g(x_1,x_2)\,dx_1\,dx_2 \\
&= \frac{m_\alpha^2}{2} - \frac{m_\alpha}{\pi}\exp(-\frac{\alpha^2}{2})\,(\sqrt{2\pi}\alpha + \exp(\frac{\alpha^2}{2})\pi m_\alpha) \\
&\quad - \frac{1}{4\pi^2}\exp(-\alpha^2)\,(2\sqrt{\pi}\alpha + \exp(\frac{\alpha^2}{2})\sqrt{2}\pi m_\alpha)^2 \\
&= \frac{\alpha^2}{\pi}\exp(-\alpha^2).
\end{aligned}$$

Si le bruit gaussien fractionnaire n'est pas standard mais que $\mathrm{var}\Delta_H(j) = \sigma^2$, alors $\operatorname{sign}(\Delta_H(j) - \alpha) = \operatorname{sign}((\Delta_H(j) - \alpha)/\sigma)$ et il suffit d'appliquer la démonstration précédente à $\Delta_H(j)/\sigma$, en posant $\alpha' = \alpha/\sigma$. □

Terminons cette section par deux remarques. La proposition précédente peut être généralisée.

Remarque 3.37 Le résultat précédent peut aisément se généraliser. Au lieu de calculer les premiers termes du développement de $\mathrm{E}\big(\operatorname{sign}(z_1-\alpha)\operatorname{sign}(z_2-\alpha)\big)$, on peut évaluer ceux de $\mathrm{E}\big(\operatorname{sign}(z_1-\alpha)\operatorname{sign}(z_2-\beta)\big)$, β étant un second paramètre, sans beaucoup plus d'efforts. Les développements deviennent toutefois moins clairs et un tel résultat n'a que peu de portée pratique. □

Il existe une démonstration alternative du résultat 3.36, mais ne elle ne permet pas de décrire le comportement numérique de la fonction de corrélation.

Remarque 3.38 Le résultat précédent peut être démontré en utilisant le développement de la fonction d'erreur en termes de polynômes d'Hermite [122, 359], puisque [2]

$$D^n\mathrm{erf}(t) = (-1)^{n-1}\,\frac{2}{\sqrt{\pi}}\,H_{n-1}(t)\,\exp(-t^2). \tag{3.28}$$

En fait, il peut être montré que si $\{X(j) : j \in \mathbb{N}\}_j$ est processus gaussien de moyenne nulle et de variance unité dont la fonction d'auto-covariance satisfait $\gamma(j) \sim c|j|^{2H-2}$, alors, si f

est une fonction telle que $\mathrm{E}f^2\big(X(0)\big) < \infty$, en posant $Y(j) = f\big(X(j)\big)$, la relation suivante est satisfaite : $\mathrm{corr}\big(Y(0),Y(j)\big) \sim c'|j|^{r_H(2H-2)}$, où r_H représente le rang d'Hermite, *i.e.* l'index du plus petit coefficient non nul dans le développement d'Hermite de $f - \mathrm{E}f\big(X(0)\big)$ [122]. En particulier, le rang d'Hermite de $f(t) = \mathrm{sign}(t)$ vaut un et celui de $f(t) = |t|$ vaut deux, ce qui implique les propositions 3.29 et 3.31. Le rang d'Hermite de la fonction $\mathrm{sign}(t-\alpha)$ est aussi un, le coefficient correspondant valant $-2\exp(-\alpha^2/2)$. La proposition 3.36 a toutefois l'avantage de donner les premiers termes du développement de Taylor de la fonction de corrélation, ce qui permet de contrôler le comportement numérique. □

Deuxième partie

L'ADN

Chapitre 1

Description de l'ADN : structure et fonctions

L'acide désoxyribonucléique, ou *ADN* [89, 184, 391], est le support de la vie dans la plupart des organismes vivants*, en ce sens que c'est lui qui détient le patrimoine génétique, renfermant les particularités de chaque individu. De composition chimique relativement simple, sa configuration moléculaire resta longtemps un mystère. Les mécanismes dont il est le siège sont de plus très complexes et encore mal connus. Nous n'exposerons dans ce chapitre que les principes de base nécessaires à la compréhension, sans entrer plus en avant dans les détails.

1.1 Composition de l'ADN

Il n'est pas suffisant de donner la composition chimique de l'ADN pour pouvoir aborder son étude, aussi succincte soit-elle. En effet, la découverte de la configuration chimique de cette macro-molécule fût certainement à l'origine de l'essor que connut la biologie moléculaire. Tous les mécanismes que nous allons décrire sont intimement liés à cette configuration.

⁎. Il l'est dans toutes les cellules vivantes et dans beaucoup de virus.

Composition chimique

Il nous faut d'abord introduire les éléments de base composant l'ADN : les nucléotides [89, 184, 391].

L'ADN, tout comme l'*ARN* (l'acide ribonucléique), est un polymère dont les monomères sont appelés *nucléotides*. Chaque nucléotide est constitué par trois unités :

- une base azotée,
- un sucre : un pentose, le *désoxyribose* pour l'ADN, et le *ribose* pour l'ARN (représentés par la figure 1.1),
- un ou plusieurs groupes phosphate (PO_4).

La base peut être soit une *purine*, soit une *pyrimidine*. Les purines sont des bases à deux cycles, contrairement aux pyrimidines qui ne sont formées que d'un seul cycle. Comme la même numérotation chimique ne peut être utilisée pour les bases et le pentose, ce dernier est numéroté avec un prime en exposant pour éviter toute confusion. Le désoxyribose est un ribose dans lequel il manque un groupement OH en position $2'$. Le groupe phosphate est attaché au pentose à la position $5'$ par un lien phospho-diester. La base est attachée à ce même sucre à la position $1'$ par un lien N-glycosidique. Le complexe sucre-phosphate constitue le squelette de l'ADN ; les bases azotées, quant à elles, déterminent la séquence génétique.

Les purines *adénine* (A), *guanine* (G) et les pyrimidines *cytosine* (C) et *thymine* (T) constituent les bases de l'ADN (figure 1.2). Dans l'ARN, la pyrimidine *uracile* (U) remplace la thymine. L'autre différence entre l'ADN et l'ARN est que le sucre pentotique de l'ARN est le ribose, alors que c'est le désoxyribose pour l'ADN. Ces différences suffisent pour conférer des propriétés différentes à ces polymères : par exemple, l'ARN est sensible à la dégradation par alcaloïde à cause de la présence du groupe OH à la position $2'$ du cycle du pentose, alors que l'ADN est résistant à cette dégradation.

Les groupements base-sucre sont liés entre eux par liens covalents grâce à un phosphate et à chacun de ces groupements (appelés *nucléosides*) est lié une base. Les nucléotides s'organisent donc en chaîne. Par convention, on lit toujours un acide nucléique dans le sens de l'extrémité $5'$ (comportant un groupement phosphate) vers l'extrémité $3'$ (qui possède un groupe OH libre).

Dans la suite, nous ferons souvent l'amalgame entre *base* et nucléotide. Un *di-nucléotide* est une séquence de deux nucléotides tandis qu'une séquence de trois nucléotides est

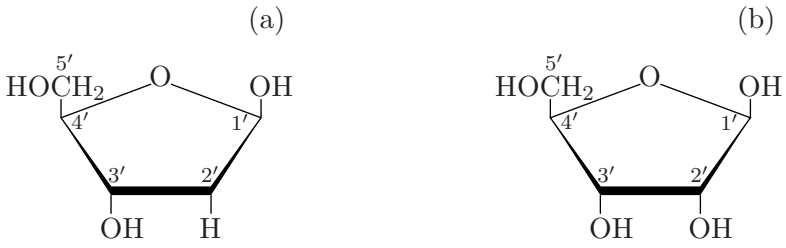

FIG. 1.1 – *Représentation du désoxyribose (a) et du ribose (b). Le désoxyribose est une forme beaucoup plus stable que le ribose.*

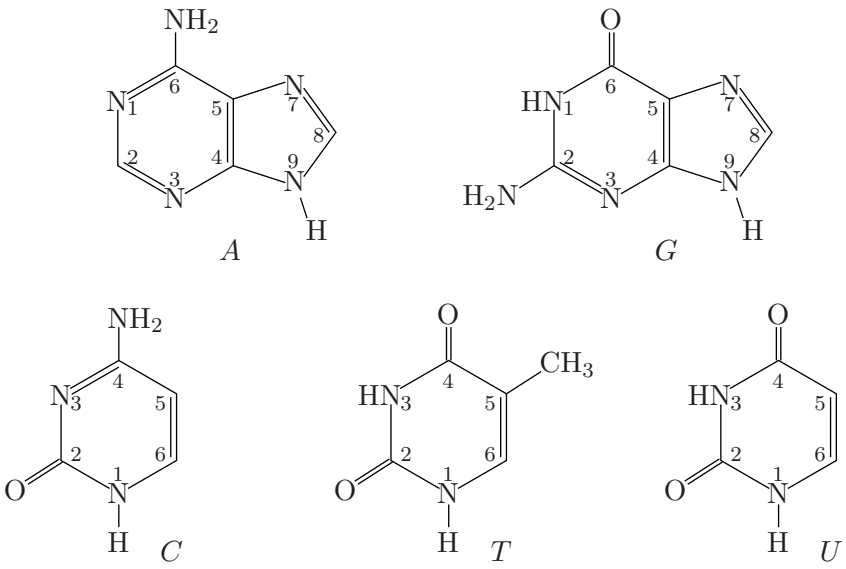

FIG. 1.2 – *Représentation des bases azotées. Dans l'ordre : les purines adénine et guanine, les pyrimidines cytosine, thymine et uracile.*

FIG. 1.3 – *On représente schématiquement une portion d'une double hélice d'ADN par une échelle renversée. Les échelons correspondent aux bases appariées et les montants représentent les nucléosides.*

appelée *tri-nucléotide*.

L'ADN forme une double hélice

La configuration en double hélice est une particularité essentielle de l'ADN. Sa découverte allait assurer le prix Nobel à WATSON et CRICK [386] et permettre toute une série d'expérimentations afin de révéler peu à peu les mécanismes de l'hérédité.

La molécule d'ADN est composée de deux chaînes, ou *brins polynucléotides*, allant en directions opposées (on dit aussi *anti-parallèles*). Les nucléotides sur un brin sont appareillés avec ceux de l'autre brin de manière très spécifique : la purine guanine ne s'appareille qu'avec la pyrimidine cytosine, tout comme la purine adénine ne peut s'appareiller qu'avec la pyrimidine thymine. Les deux brins sont donc *complémentaires* en ce sens que la séquence de nucléotides sur un brin détermine la séquence sur le brin complémentaire (figure 1.3). Ces deux chaînes sont enroulées autour d'un axe commun sous la forme d'une *double hélice* (figure 1.4).

Cette structure en double hélice est essentiellement stabilisée par des liaisons hydrogènes entre les paires de bases. La paire A:T présente deux liaisons hydrogène tandis que G:C en présente trois. Cette dernière liaison est donc plus difficile à briser. Toutefois, des interactions de type van der Waals et des interactions hydrophobes interviennent également dans la stabilisation de cette structure hélicoïdale. Les bases hydrophobes sont empilées à l'intérieur de l'hélice et l'enchaînement sucre-phosphate hydrophile est dirigé vers l'extérieur. La séparation de l'ADN en molécules à simple brin est appelée *dénaturation*, l'opération inverse est appelée *renaturation* [89, 184, 391].

Dans la cellule, sous les conditions physiologiques les plus courantes, l'ADN adopte une configuration très spécifique appelée *forme B* (figure 1.4). Sous cette forme, chaque brin fait un tour complet tout les 3.4 nm. L'hélice est à pas droit et chaque base, située dans un plan quasi perpendiculaire à l'axe de l'hélice, est espacée de 0.34 nm, ce qui signifie que chaque tour complet possède dix paires de nucléotides. Il existe d'autres configurations

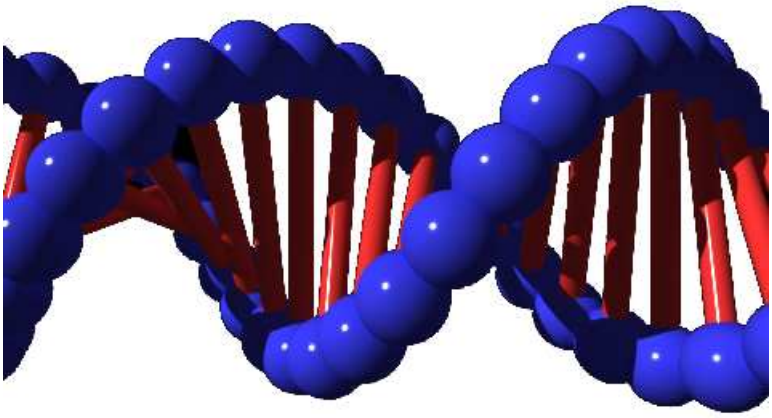

FIG. 1.4 – *Représentation de la double hélice d'ADN sous la forme B.*

que nous ne ferons que citer. La *forme Z* est observée sur des portions riches en guanine-cytosine et possède une hélice à pas gauche. Quant à la *forme A*, elle est parfois observée dans certaines régions d'ADN naturel en présence de fortes concentrations en cations comme Mg^{++} et Ca^{++}, ou lorsque le degré d'hydratation est faible (typiquement inférieur à soixante-cinq pour cent). Elle possède onze paires de nucléotides par tour. Quant aux sous-classe du type B, elles sont rarement rencontrées. La *forme C* est observée lorsque le degré d'hydratation est très faible (inférieur à quarante-cinq pour cent) et la *forme D* n'existe qu'avec des ADN artificiels.

Nous utiliserons la notation suivante, courante en biologie moléculaire.

Notation 1.1 La longueur d'une molécule d'ADN est généralement donnée en *paires de bases* (azotées). Par extension, nous emploierons aussi ce terme pour désigner la longueur d'une séquence de nucléotides, bien qu'une telle séquence ne représente qu'un seul brin d'ADN. Ainsi, une molécule d'ADN d'une longueur de n paires de bases, ou encore n pdb, est constituée de deux brins de longueur n pdb chacun. La paire de base (pdb) représente donc une unité de longueur.

Les chromosomes

Pour permettre la division des cellules, l'ADN s'organise en macro-molécules appelées chromosomes [89, 184, 391].

Chez les *eucaryotes*, *i.e.* les organismes possédant un noyau, l'ADN est situé dans le noyau et partitionné sous sa forme la plus condensée en *chromosomes* (figure 1.5). Un

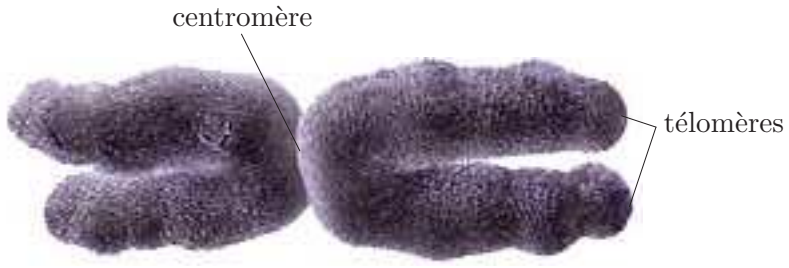

FIG. 1.5 – *Représentation d'un chromosome (mitotique). Cette macro-molécule constitue la forme la plus condensée de l'ADN.*

chromosome est constitué de deux molécules d'ADN identiques appelées *chromatides*, disposées symétriquement et jointes par un *centromère*. Chaque chromosome comporte donc quatre extrémités ou *télomères*. Chez les *procaryotes* (c'est-à-dire les bactéries) et les virus, le matériel génétique est en général composé d'une copie de l'ADN en double hélice.

Les organismes *diploïdes*, comme ceux des mammifères, possèdent deux jeux de chromosomes, chaque jeu provenant d'un parent. Ainsi, l'humain (*Homo sapiens*) possède quarante-six chromosomes répartis en vingt-trois paires (nombre *haploïdique*) numérotées de 1 (correspondant à la séquence de nucléotides la plus longue) à 22 (la séquence la plus courte), les chromosomes sexuels - définissant le sexe du porteur, X pour la femelle (XX) et Y pour le mâle (XY) - étant considérés comme la vingt-troisième paire.

On parle parfois aussi de chromosome chez les procaryotes. On peut le définir comme la molécule ADN contenant le génome.

Dans la suite, lorsque nous réfèrerons à un chromosome, ce sera pour désigner la partie de l'ADN formant un chromosome lorsqu'elle est sous sa forme la plus condensée. Il s'agit donc d'une partie de l'ADN d'un organisme, sans rapport direct avec le niveau de compaction[⁕]. Pour désigner un chromosome sous sa forme la plus condensée, nous parlerons maintenant de chromosome *mitotique*.

1.2 Le mécanisme de réplication

Lors de chaque division cellulaire, la totalité de l'ADN doit être dupliquée. La duplication d'une molécule d'ADN parent en deux molécules d'ADN filles est appelée *réplication*.

⁕. Il s'agit d'un abus de langage courant dans la littérature.

Cette étape se produit durant la phase S du cycle cellulaire et n'est pas encore totalement comprise pour tous les organismes [89, 184, 391].

Principes

Les principes de la réplication de l'ADN sont assez simples, mais les détails ne sont pas tous connus et certains mécanismes sont encore hypothétiques.

L'ADN se réplique en s'utilisant lui-même comme patron, chaque brin permettant de synthétiser un nouveau brin. La réplication est *semi-conservative* en ce sens que les deux molécules d'ADN issues de la réplication contiennent chacune un brin de l'ADN original. Les nouveaux brins complémentaires sont synthétisés par une enzyme appelée l'*ADN polymérase III* [179, 391].

L'unité d'ADN où se produit la réplication est appelée *réplicon* [198]. Ce réplicon possède une *origine*, où est initiée la réplication, et au moins une *terminaison*, où est arrêtée la réplication. L'ADN bactérien est généralement circulaire et constitue à lui seul un réplicon [198] : il existe un site unique de départ sur le chromosome, ce site ne variant pas d'un cycle de réplication à l'autre, présentant une séquence bien conservée de taille supérieure à 245 paires de base [87, 304]. Chez les eucaryotes, où la quantité d'ADN à répliquer est plus importante, les origines de réplication sont multiples et il n'est pas clair que les sites de départ soient invariants. En général, et c'est toujours le cas chez les procaryotes, à partir d'une origine, la réplication progresse dans les deux sens : la réplication est *bi-directionnelle*. Chez l'humain, la réplication progresse de cinquante nucléotides par seconde, ce taux pouvant atteindre cinq cent nucléotides par seconde pour certains organismes.

L' ADN polymérase ne peut synthétiser que dans le sens $5' \rightarrow 3'$. La réplication doit donc se faire de manière discontinue sur un des deux brins (figures 1.6). On parle de *brin retardé*. Le brin synthétisé continûment est appelé *brin avancé*. Les petits fragments discontinus synthétisés sont de taille $1000 - 2000$ paires de base et sont appelés *fragments d'Okazaki* [272]. Si la réplication est bi-directionnelle, un brin retardé à gauche d'une origine est un brin avancé à droite et inversement.

Un segment d'ARN appelé *amorce* doit initier la synthèse. En effet, la polymérisation de l'ADN requiert une amorce pour démarrer un fragment précurseur. Le fragment synthétisé continûment n'a besoin que d'une amorce. Pour l'autre fragment, il faut donc une amorce tous les $1000 - 2000$ paires de base.

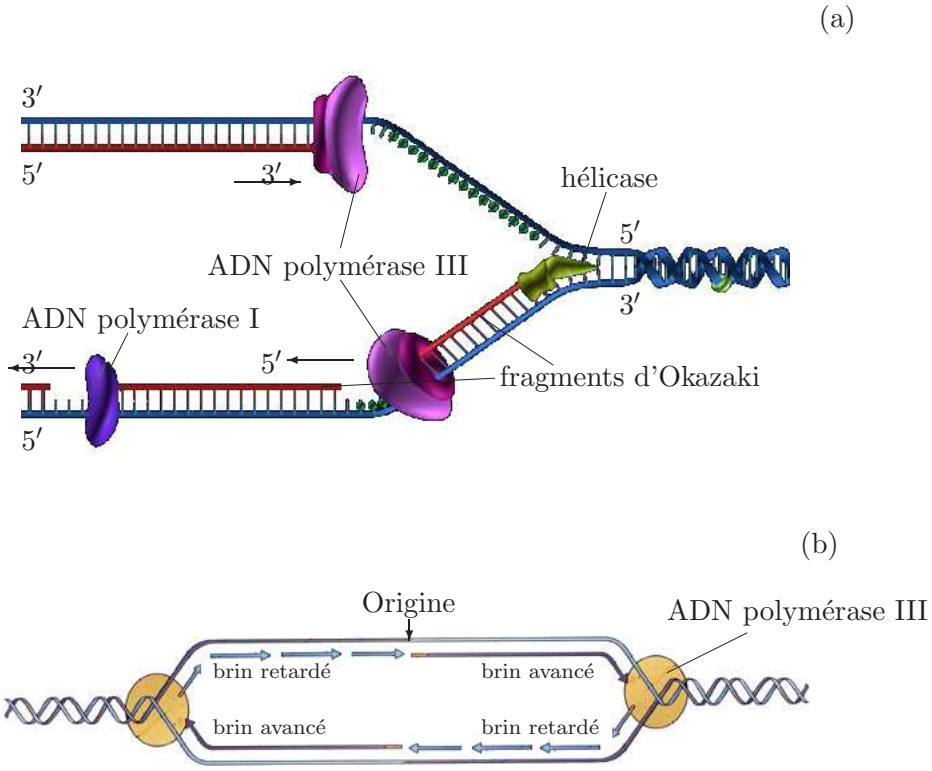

FIG. 1.6 – *Représentation schématique de la réplication. (a) L'hélicase ouvre la double hélice pour permettre à l'ADN polymérase III d'accéder aux nucléotides. Le long de chacun des deux brins, un nouveau brin est synthétisé, de manière continue pour l'un et par morceaux pour l'autre. (b) Chez la plupart des organismes, la réplication est bi-directionnelle ; dans ce cas, pour un brin donné, deux ADN polymérases III agissent de part et d'autre de l'origine de réplication formant une bulle de réplication. Si un nouveau brin est synthétisé continûment d'un côté de l'origine, il le sera de manière discontinue de l'autre et vice-versa.*

Durant la réplication, la double hélice d'ADN est déroulée localement et les deux brins sont séparés. On voit donc apparaître une *bulle de réplication* (figure 1.6), formée par les deux brins séparés, qui s'étend de part et d'autre de l'origine au fur et à mesure que la réplication progresse. À l'endroit où les brins se rejoignent, c'est-à-dire aux extrémités de la bulle de réplication, on dit que l'ADN forme une *fourche de réplication*. Chez les procaryotes, les deux fourches de réplication progressent dans des directions opposées sensiblement à la même vitesse.

Dans l'appariement des bases, le taux d'erreur est de 1 sur 10^5 bases ajoutées. Si une telle erreur se produit, un mécanisme de correction entre en jeu en utilisant notamment l'*activité exonucléase*⁂ $3' \rightarrow 5'$ de l'ADN polymérase III. Ce processus est appelé *relecture*.

Contrairement aux procaryotes [272, 366], l'initiation de la réplication chez les eucaryotes est encore mal connue. Pour certains organismes eucaryotes tels *Saccharomyces cerevisae*, où toutes les origines de réplication sont annotées, il existe une séquence consensus de 11 paires de base, appelée ARS❋ [51, 66, 320], qui peut initier la réplication. L'existence d'une telle séquence initiant systématiquement la réplication ne semble toutefois pas être une généralité, en particulier pour les eucaryotes supérieurs. Chez les organismes eucaryotes multi-cellulaires [162, 274], on estime que le nombre d'origines de réplication est de l'ordre de 30 000. Seule une vingtaine ont été identifées dans différents organismes et un plus petit nombre ont été caractérisées en détail [364]. Ces origines sont très hétérogènes en taille et ont des activations différentes dans le temps et suivant le type cellulaire.

Activités enzymatiques

Les éléments permettant de réaliser la réplication sont les enzymes. Celles-ci sont nombreuses est possèdent parfois un comportement complexe.

L'activité enzymatique* principale, grande consommatrice d'énergie, est dévolue aux *ADN polymérases*. Ces enzymes sont chargées de recruter les nucléotides libres et de les apparier aux nucléotides du brin parent.

Pour permettre à l'ADN polymérase III d'accéder au brin parent et de commencer la réplication, un certain nombre de protéines, dites secondaires, sont nécessaires [228].

– les *topoisomérases* sont chargées d'introduire des supertours négatifs aux fourches de

⁂. Une enzyme exonucléase est une enzyme capable de dégrader l'ADN et l'ARN
❋. Pour Autonomous Replication Sequence.
∗. Rappelons qu'une enzyme est un type de protéine permettant la catalyse d'une réaction chimique.

réplication ; elles diminuent donc le degré de surenroulement de l'hélice et permettent ainsi l'ouverture et la progression de la bulle ;

- les *hélicases* déroulent l'hélice et séparent les deux brins parents, en rompant les liaisons hydrogènes les unissant ;
- les *protéines SSB* (pour Single Stranded Binding) stabilisent les brins séparés, prévenant leur réunification et la formation de boucles et bourrelets ;
- à chaque extrémité de la bulle, une *primase* (ARN polymérase) associée à d'autres protéines synthétise de petites portions d'ARN constituant les amorces, reconnues par l'ADN polymérase III pour initier la réplication ;
- l'*ADN polymérase I* suit la progression de la polymérase III et détruit les amorces d'ARN pour les remplacer par des séquences d'ADN correspondantes ;
- les ADN ligases sont chargées de souder les fragments d'Okazaki.

L'ensemble des activités enzymatiques localisées au sein d'une fourche de réplication constitue un *réplisome*.

1.3 Le mécanisme de transcription

Le transfert de l'information contenue dans un gène sous forme d'une séquence d'ARN est appelé transcription. Ce processus est la première étape de la traduction du code génétique. Le mécanisme de la transcription est complexe et nous en donnons ici une description simplifiée [89, 179, 243].

Rôle de la transcription

Dans un brin d'ADN, il existe certaines régions contenant l'information nécessaire à la construction, la régulation de protéines et d'autres molécules déterminant la croissance et le fonctionnement de l'organisme, appelées gènes.

La *transcription* est le processus par lequel une partie de la séquence ADN est copiée (par une ARN polymérase) pour former un brin d'ARN complémentaire. Les segments d'ADN transcrits sont appelés *gènes*. Ce sont les gènes qui codent la structure chimique des protéines, constituants fondamentaux des cellules. Plus précisément, une succession de nucléotides peut déterminer une séquence d'acide aminés formant la protéine. Le codage de tri-nucléotides, appelés *codons*, en acides aminés est un principe essentiellement commun à tous les organismes, de la bactérie à l'homme.

La totalité du matériel génétique d'une cellule ou d'un individu est appelé *génome*. Le génome humain comprend approximativement trois milliards de paires de nucléotides représentant 30 000 gènes environ [189]. La taille des gènes peut varier de quelques centaines à plusieurs milliers (voire dizaines de milliers) de nucléotides, avec une taille moyenne avoisinant 30 000 nucléotides. Cependant, même les gènes les plus longs n'utilisent qu'une faible portion de leur séquence pour coder l'information nécessaire à l'expression d'une protéine. Ces régions codantes sont appelées *exons* et les régions non-codantes *introns*. Chez l'homme, la taille moyenne des exons est de l'ordre de 150 nucléotides et celle des introns de 800 nucléotides ; ces derniers constituent environ 80% des gènes. Plus l'organisme est complexe, plus la quantité et la taille des introns semblent importantes.

Description du mécanisme de la transcription

Comme la réplication, la transcription obéit à une série de principes mis en évidence par l'expérimentation.

La transcription fait intervenir une activité enzymatique nommée *ARN polymérase* holoenzyme. Cet énorme complexe enzymatique déroule et disjoint les deux brins de l'ADN hélicoïdal. Elle recrute aussi les nucléotides du futur brin d'ARN, pour les assembler par complémentarité avec les bases de la séquence du brin d'ADN (figure 1.7).

Contrairement à la réplication, qui intéresse la totalité du génome à chaque cycle, le programme de transcription n'est pas fixe : seules de petites portions du génome sont transcrites à une époque donnée de la vie de la cellule et ces portions varient en fonction de nombreux facteurs, tels le développement et l'environnement.

Le processus de transcription est assez similaire chez les procaryotes et chez les eucaryotes. Toutefois les eucaryotes possèdent trois types d'ARN polymérase au lieu d'un seul chez les procaryotes, chaque variété étant responsable de la synthèse d'une classe d'ARN. La transcription s'effectue en trois étapes successives : l'*initiation*, l'*élongation* et la *terminaison*.

L'initiation débute lorsque l'ARN polymérase s'associe à une région spécifique d'un brin d'ADN en amont des gènes, appelée *promoteur*.

L'élongation de la chaîne d'ARN s'effectue par polymérisation successive de nucléotides dans le sens $5' \rightarrow 3'$. Un seul brin d'ADN, le *brin modèle*, est transcrit ; l'autre brin, avec la même orientation $5' \rightarrow 3'$, est appelé *brin codant*.

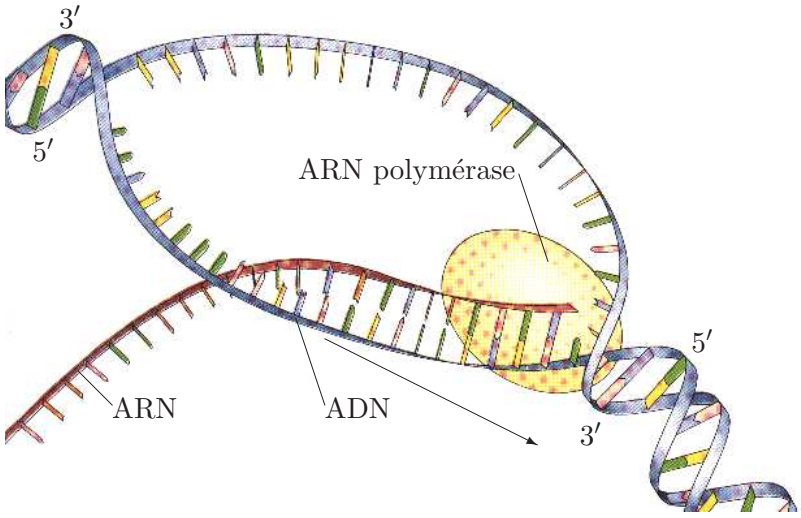

FIG. 1.7 – *Représentation schématique de la transcription. Une polymérase progresse sur le brin transcrit dans le sens* $3' \rightarrow 5'$*, synthétisant un brin d'ARN. Au fur et à mesure de la progression de la polymérase, les deux brins d'ADN sont désappariés, le brin transcrit est alors temporairement apparié avec l'ARN nouvellement synthétisé, tandis que le brin non-transcrit est libre.*

L'étape de terminaison intervient lorsqu'un signal indiquant la fin du gène est lu. Cette *séquence de terminaison* cause la formation d'un bourrelet sur la chaîne d'ARN, provoquant la dissociation du complexe ADN - polymérase - ARN. Dans certains cas, la terminaison fait également intervenir un facteur protéique appelé *facteur* ρ.

La transcription repose sur les mêmes étapes de base, que ce soit pour les procaryotes ou les eucaryotes, mais ces étapes en elles-mêmes présentent d'importantes différences selon que l'organisme possède un noyau ou pas. De nombreuses raisons peuvent expliquer ces différences, à commencer par le nombre de gènes : de l'ordre de 4000 chez les procaryotes et de 30 000 à 60 000 chez les eucaryotes.

Chez les procaryotes, la transcription a lieu directement dans le cytoplasme et est catalysée par une seule polymérase, l'*ARN polymérase*, alors que chez les eucaryotes, la transcription a lieu dans le noyau et utilise trois polymérases différentes : les ARN polymérases I, II et III. L'information sur les terminaisons est beaucoup moins bien comprise que chez les procaryotes. Il est même possible que ces terminaisons soient floues [49, 131] et que la polymérase se détache mille nucléotides après la fin spécifiée.

1.4 L'empaquetage de l'ADN

Typiquement, la diamètre d'un noyau de cellule humaine est de 5–8 μm, alors que 2 m d'ADN doivent pouvoir se lover dans ce noyau. L'ADN est donc empaqueté suivant des niveaux d'organisation hiérarchisés par une série d'enroulements et de boucles, mettant en jeu des protéines, appelées *protéines de structure* [89, 184, 391].

Organisation de la chromatine

Lors de la division cellulaire, l'ADN est sous sa forme la plus dense, les chromosomes mitotiques (figure 1.5). Pourtant, il ne peut rester aussi compact durant tout un cycle cellulaire, car les nucléotides doivent rester accessibles pour les complexes protéiques impliqués dans des mécanismes fonctionnels tels la transcription et la réplication. En général, l'ADN présente plusieurs niveaux de compaction, coexistants dans un même chromosome.

Le complexe formé par l'ADN et les protéines de structure est appelé la *chromatine*. Chez les procaryotes, la manière dont est empaqueté l'ADN est encore très mal comprise ; nous nous concentrerons ici sur les organismes eucaryotes, où l'ADN est empaqueté dans le noyau des cellules. La chromatine n'est dans son état le plus condensé que durant une brève période, au moment où la cellule se divise. Durant cette étape, les protéines impliquées dans la transcription et la réplication ne peuvent plus avoir accès à l'ADN. Durant les autres étapes du cycle cellulaire, les chromosomes sont plus ou moins décondensés dans le noyau sous forme de longs fils fins et enchevêtrés.

Lorsque les chromosomes ne sont pas sous leur forme la plus empaquetée, on parle de chromosome *interphasique* et de chromatine *interphasique*. À ce moment, la condensation n'est pas uniforme : certaines régions sont plus dépliées que d'autres. La forme la plus condensée de la chromatine interphasique, l'*hétérochromatine*, est quasiment inactive au plan transcriptionnel. Chez l'homme, elle constitue environ 10% d'un chromosome interphasique. Le reste de la chromatine interphasique est appelée l'*euchromatine*. La forme la moins condensée est appelée la chromatine *active* et constitue typiquement 10% d'un chromosome interphasique chez l'homme. La chromatine active est transcrite ou disponible pour la transcription. D'une manière générale, les régions du chromosome transcrites sont plus déroulées, tandis que celles qui sont inactives au plan transcriptionnel sont plus condensées. Ainsi la structure d'un chromosome interphasique diffère d'un type cellulaire à l'autre, selon la distribution spatiale des gènes qui y sont exprimés.

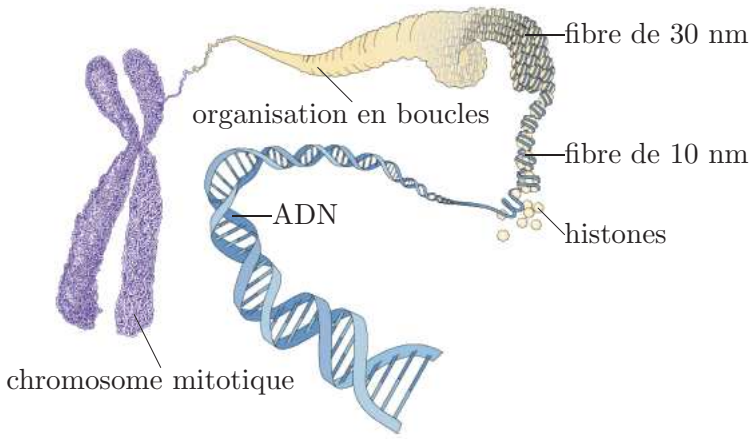

FIG. 1.8 – *Schéma représentant les diverses étapes de l'empaquetage de l'ADN.*

Les étapes de l'empaquetage

Dans le noyau des cellules eucaryotes, l'ADN est empaqueté en plusieurs étapes révélant plusieurs niveaux d'organisation (figure 1.8). Dans le chromosome mitotique, l'ADN est 50 000 fois plus court que la molécule déroulée. La première étape de repliement de l'ADN nu dans le chapelet nucléosomal est l'étape la mieux connue, les autres étapes étant plus hypothétiques [89, 184, 391].

Le premier niveau d'empaquetage de la chromatine, et le plus fondamental, est celui aboutissant à la formation de *nucléosomes*. Le coeur du nucléosome est formé de huit protéines, quatre paires de protéines identiques, appelées *histones* (les histones sont les histones H2A, H2B, H3 et H4). Le double brin d'ADN s'enroule 1.6 fois autour de cet octamère d'histones sur une longueur d'environ 146 pdb. La taille d'un nucléosome est approximativement de 200 pdb, les noyaux protéiques étant espacés entre eux d'environ 50 pdb. Les histones existent chez tous les eucaryotes, alors que les bactéries n'en possèdent pas. Cependant, les archæbactéries contiennent des protéines homologues aux histones eucaryotes. Ces protéines sont peut-être les précurseurs, sur le plan de l'évolution, des histones. À l'issue de cette première étape de compaction, la longueur de la séquence d'ADN est divisée par trois et constitue ce que l'on appelle la fibre de 10 nm, en raison de l'épaisseur du chapelet nucléosomal.

Les nucléosomes sont ensuite empilés les uns sur les autres pour constituer une structure plus compacte, la fibre de 30 nm. La compaction des nucléosomes en cette fibre dépend de la présence éventuelle d'une cinquième histone (l'histone H1, aussi appelée H5). Les

étapes suivantes de condensation sont encore hypothétiques, mais on pense que la fibre de 30 nm s'organise en boucles, dites *boucles de chromatine*, attachées à un squelette protéique (figure 1.8).

Chapitre 2

Codages mono- ou multi-nucléotidiques de l'ADN : de l'analyse à la synthèse des séquences ADN

Notre but étant d'appliquer les outils de l'analyse multifractale à l'ADN, nous devons d'abord disposer d'une méthode permettant de construire un signal à partir d'une séquence ADN. Pour ce faire, nous utiliserons le principe du codage mono- ou multi-nucléotidique. Nous commencerons par définir ces notions, avant de présenter quelques codages couramment rencontrés dans la littérature. L'étude spectrale des signaux ainsi obtenus à partir du génome de l'homme nous permettra de mettre en évidence l'existence de corrélations à longue portée jusqu'à des distances de l'ordre de 20 à 30 kpb et la présence de rythmes de basse fréquence correspondant à des périodicités caractéristiques de l'ordre de 100, 400 et 800 kpb. Après avoir rappelé les résultats concernant l'étude des corrélations à longue portée au sein des signaux construits à partir de séquences d'ADN, nous donnerons une méthode de construction de séquences nucléotidiques artificielles prenant en compte l'existence de ces corrélations.

L'originalité de la première section réside dans la formalisation des notions bien connues de codage des séquences ADN et leur étude fréquentielle, qui a donné lieu aux publications suivantes [297, 299]. L'approche spectrale des résultats obtenus dans les références [34, 35] est aussi originale, de même que la méthode de synthèse de séquences artificielles, qui fera l'objet d'une publication [298].

2.1 Construction de signaux ADN par codage

Grâce au codage, une séquence nucléotidique peut être associée à une multitude de signaux, appelés marche ADN, chacun reflétant plus particulièrement une propriété spécifique de la séquence étudiée (configuration spatiale de la double hélice, distribution de la composition nucléotidique, *et cætera*). Après avoir exposé les principes de base, nous donnerons quelques exemples de codage, conduisant à l'obtention de marches ADN. Nous montrerons que ces marches semblent posséder une composante mouvement brownien fractionnaire corrélé à longue portée associée à la structure nucléosomale de la fibre de chromatine de 30 nm. Cette composante invariante d'échelle se superpose à la présence de rythmes de basse fréquence, véritable signature d'un niveau d'organisation supérieur de la chromatine.

Codages et marches ADN

Le codage permet d'associer un signal à un brin d'ADN en considérant une séquence de nucléotides comme un mot construit sur un alphabet nucléotidique [26, 35, 242, 314, 383]. On associe un signal à un tel mot *via* une application définie sur l'alphabet et à valeur dans $\mathbb{R}$. Nous formalisons ici cette notion.

Les deux brins d'une macro-molécule d'ADN étant complémentaires, l'étude des éléments constitutifs d'une séquence ADN peut se ramener à l'étude des mots construits sur un alphabet de quatre lettres $\{A, C, G, T\}$, représentant les quatre bases nucléotidiques constitutives d'un des deux brins de l'ADN.

Notation 2.1 Nous désignerons l'alphabet représentant les bases nucléotidiques par $L = \{A, C, G, T\}$. Cet alphabet sera appelé *alphabet nucléotidique*. Sauf mention contraire, nous utiliserons l'ordre lexicographique pour ordonner cet alphabet et le munirons implicitement de l'opération concaténation.

Une séquence ADN est donc un mot construit sur l'alphabet L. Nous considèrerons parfois

ces mots comme construits sur l'alphabet $L^3 = \{AAA, AAC, \ldots, TTT\}$. Nous aurons aussi besoin d'extraire une suite de lettres hors d'un mot.

Notation 2.2 Si m est un mot construit sur l'alphabet L, nous désignerons par $\pi(m,n,j)$ le sous-mot de m de taille j commençant à la n-ième lettre de m, l'indexation débutant à zéro.

Nous souhaitons pouvoir associer une valeur réelle à une lettre nucléotidique ou à une suite de lettres. L'application réalisant cette opération est appelée codage.

Définition 2.3 Un *codage* ADN k-nucléotidique est la donnée d'une application τ définie sur L^k ($k \in \mathbb{N}_0$) et à valeurs dans l'ensemble des réels. Si le codage n'est pas injectif, il sera dit *dégénérescent*.

On peut étendre un codage défini sur l'alphabet L à l'alphabet L^q, $q \in \mathbb{N}_0$. Il suffit de considérer τ comme un morphisme défini sur L, muni de l'opération concaténation et à valeur dans $\mathbb{R}$, que l'on munit de l'opération addition. Pratiquement, si $m \in L^q$, on pose $\tau(m) = \sum_{j=0}^{q-1} \tau\big(\pi(m,j,1)\big)$. Par extension, si τ est un codage k-nucléotidique ($k > 1$), on pose $\tau(m) = \sum_{j=0}^{q-k} \tau\big(\pi(m,j,k)\big)$. Un codage défini sur L^k, $k > 1$, est dit *fondamental* s'il n'existe pas deux codages τ_1 et τ_2 définis respectivement sur L^{q_1} et L^{q_2}, avec $q_1 + q_2 = k$, tels que pour tous mots $m_1 \in L^{q_1}$ et $m_2 \in L^{q_2}$, on ait $\tau(m_1 m_2) = \tau_1(m_1) + \tau_2(m_2)$.

Pour que ces notions soient rigoureusement définies, nous adopterons la convention suivante.

Remarque 2.4 Pour éviter tout problème de définition, il faut considérer le mot vide ε en posant $\tau(\varepsilon) = 0$ et poser, pour tout mot m appartenant à L^k, $\pi(m,n,j) = \varepsilon$ lorsque $n < 0$ ou $n > k - j$. De plus, $\pi(m,n,0) = \varepsilon$ pour tout n. □

Définition 2.5 Un *bruit ADN* [26, 314] associé au codage k-nucléotidique τ est une fonction définie sur $\mathbb{N}$ comme suit,

$$f_\tau(n;m) = \tau\big(\pi(m,nr,k)\big), \tag{2.1}$$

où $r \in \mathbb{N}$ est appelé le *pas* ou l'*incrément élémentaire* et m est un mot construit sur l'alphabet nucléotidique. On omet souvent la référence à m en écrivant simplement $f_\tau(n) = f_\tau(n;m)$. En général, le pas r est choisi égal à 1 où à k.

Définition 2.6 La *marche ADN* associée au bruit f_τ [26, 314] est la fonction définie en posant $F_\tau(0) = 0$ et

$$F_\tau(n+1) = F_\tau(n) + f_\tau(n). \tag{2.2}$$

Si $r = 1$, on a $F_\tau(n) = \tau\big(\pi(m,0,n+k-1)\big)$. Si $r < k$, les valeurs de f_τ sont corrélées : $k - r$ lettres sont communes aux deux mots définissant respectivement $\tau\big(\pi(m,nr,k\big)$ et $\tau\big(\pi(m,n(r+1),k\big)$. On peut ainsi lisser le signal en diminuant la valeur de r.

Signalons qu'il est possible de définir des codages à valeur dans $\mathbb{R}^{n}$ *. Toutefois l'utilité de tels codages s'est jusqu'à présent révélée limitée et nous ne les considérerons pas dans la suite. Une séquence ADN contient des informations de différentes natures (structurelles, comme les propriétés de courbure locale de la double hélice, ou fonctionnelles, comme la localisation des gènes), cette abondance d'information pouvant *a priori* rendre difficile l'analyse d'une propriété particulière. En utilisant un signal unidimensionnel, le but est de s'affranchir, dans la limite du possible, des informations « parasites » pour se focaliser sur un signal plus spécifique. Ainsi, le signal obtenu sera généralement insuffisant pour ré-obtenir la séquence de départ, le codage utilisé étant dégénérescent.

Exemples de codage nucléotidique

La construction d'un signal à partir d'une séquence de nucléotides se résume par la donnée d'un codage. Il en existe de nombreux et le choix du codage dépend de la propriété que l'on souhaite mettre en évidence [34, 35, 242, 314, 345, 383].

Codages du type purine-pyrimidine

Les codages les plus simples sont certainement les codages binaires qui séparent les nucléotides en deux familles, en attribuant à chacune de ces familles une valeur différente.

Le codage défini explicitement par les relations suivantes,

$$\begin{cases} \tau(A) = 1 \\ \tau(G) = 1 \\ \tau(C) = -1 \\ \tau(T) = -1 \end{cases}, \tag{2.3}$$

permet de faire la distinction entre les purines A et G et les pyrimidines C et T. Il est naturellement appelé *codage purine-pyrimidine* [314]. De la même manière, on définit le *codage amino-keto* [383], qui distingue les bases thymine et guanine possédant un groupe fonctionnel « C–O », des bases adénine et cytosine, où l'oxygène est remplacé par le groupe fonctionnel « NH_2 », en associant A et C à 1 et G et T à -1. Finalement, le *codage faible-fort* [383] fait la distinction entre les bases C et G, liées entre elles par trois

*. Un exemple classique est le codage défini par les relations suivantes : $\tau(A) = (1,1)$, $\tau(C) = (-1,1)$, $\tau(G) = (1,-1)$ et $\tau(T) = (-1,-1)$.

ponts hydrogènes au sein de l'ADN, et les bases A et T, liées par seulement deux ponts. Il associe donc A et T à 1, et C et G à -1.

Codages mono-nucléotidiques

Un *codage mono-nucléotidique* est un codage permettant d'analyser la distribution de la position d'une des bases A, C, G ou T dans un mot, en associant trois des bases à la même valeur [242, 314, 383]. Il existe donc principalement quatre codages mono-nucléotidiques différents. Si l'on souhaite, par exemple, étudier la répartition en adénine d'une séquence ADN, il suffit de définir le morphisme τ de la manière suivante,

$$\begin{cases} \tau(A) = \ \ 1 \\ \tau(G) = -1/3 \\ \tau(C) = -1/3 \\ \tau(T) = -1/3 \end{cases}, \tag{2.4}$$

où la valeur $-1/3$ est choisie telle que, si les concentrations en A, C, G et T sont égales, le bruit associé possède une moyenne nulle.

Si les concentrations en nucléotides s'écartent notablement de l'équi-partition, il y a lieu de modifier le codage comme suit. Supposons que les concentration en A, C, G et T sont respectivement c_A, c_C, c_G et c_T, et posons $c = c_C + c_G + c_T$. Il suffit de redéfinir le codage de la manière suivante, $\tau(A) = 1$ et $\tau(l) = -c_l/c$, pour l appartenant à $\{C,G,T\}$.

Codage pourcentage en GC

La concentration en bases C et G est une des caractéristiques du génome les plus étudiées, en particulier à travers les notions très débattues d'isochores [56, 57, 104, 189, 177, 310]. Par exemple, la densité en gènes semble être corrélée à la concentration en GC, les régions riches en gènes étant des régions à haute concentration en GC [15, 107].

Pour représenter la concentration en bases C et G le long d'un brin d'ADN, on utilise le codage suivant, appelé *codage pourcentage en GC*,

$$\begin{cases} \tau(C) = 1 \\ \tau(G) = 1 \\ \tau(A) = 0 \\ \tau(T) = 0 \end{cases}. \tag{2.5}$$

En observant le bruit associé au codage pourcentage en GC chez l'homme, on constate l'existence de zones, appelées *isochores*, de moyenne différente. Cette remarque est particulièrement évidente pour le chromosome 21, comme l'illustre la figure 2.1, où le bruit

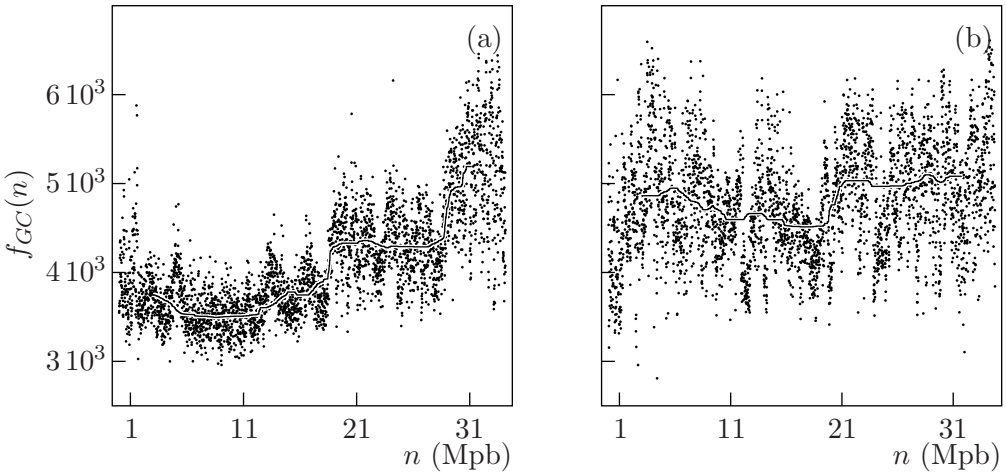

FIG. 2.1 – *Le pourcentage en GC pour les chromosomes 21 (a) et 22 (b) de l'homme, calculé dans des fenêtres non chevauchantes de largeur* 10 *kpb. Pour le chromosome 21 (a), on constate clairement l'existence de zones différant par leur moyenne. Pour mettre cette observation en évidence, la médiane a été surimposée.*

f_{GC} a été calculé dans des fenêtres non chevauchantes de largeur 10 kpb. Ainsi, en notant m_{21} le mot associé au chromosome 21, $f_{GC}(n) = \tau\big(\pi(m_{21},10^4 n,10^4)\big)$.

Le *codage pourcentage en TA* est bien sûr le codage associant la valeur un aux nucléotides A et T et zéro aux nucléotides C et G.

Les codages de courbure

Les codages de courbure permettent de caractériser localement les propriétés structurales et mécaniques de la double hélice d'ADN. Nous considérerons essentiellement les codages PNuc [165] et DNase [84, 85, 155].

L'ADN génomique est souvent considéré comme un polymère constitué de quatre différents types de monomères (les nucléotides). La manière dont ceux-ci se succèdent influence la configuration spatiale et les propriétés mécaniques de la molécule d'ADN. La courbure locale est définie comme l'angle entre le plan défini par une paire de base et le plan défini par la paire suivante, dans le sens de lecture $5' \to 3'$. Un *codage de courbure locale* est un codage qui tente de rendre compte de la configuration spatiale des deux brins d'ADN en associant à une suite de nucléotides une valeur représentant la courbure locale. On ne dispose pas de mesure physique directe de la courbure. Pour cette raison, il peut exister plusieurs codages s'y rapportant. L'hypothèse sous-jacente aux deux codages que nous allons présenter est la suivante : pour une position donnée, le rayon de courbure est une propriété locale qui ne dépend que des deux nucléotides voisins les plus proches. Les co-

Tri-base	PNuc	Tri-base	PNuc	Tri-base	PNuc	Tri-base	PNuc
AAA	0	CAA	3.3	GAA	3	TAA	2
AAC	3.7	CAC	6.5	GAC	5.4	TAC	3.7
AAG	5.2	CAG	4.2	GAG	5.4	TAG	2.2
AAT	0.7	CAT	6.7	GAT	5.3	TAT	2.8
ACA	5.2	CCA	5.4	GCA	6	TCA	5.4
ACC	5.4	CCC	6	GCC	10	TCC	3.8
ACG	5.4	CCG	4.7	GCG	7.5	TCG	8.3
ACT	5.8	CCT	5.4	GCT	7.5	TCT	3.3
AGA	3.3	CGA	8.3	GGA	3.8	TGA	5.4
AGC	7.5	CGC	7.5	GGC	10	TGC	6
AGG	5.4	CGG	4.7	GGG	6	TGG	5.4
AGT	5.8	CGT	5.4	GGT	5.4	TGT	5.2
ATA	2.8	CTA	2.2	GTA	3.7	TTA	2
ATC	5.3	CTC	5.4	GTC	5.4	TTC	3
ATG	6.7	CTG	4.2	GTG	6.5	TTG	3.3
ATT	0.7	CTT	5.2	GTT	3.7	TTT	0

TAB. 2.1 – *Table expérimentale obtenue (après un rééchelonnement linéaire entre* 0 *et* 10*) à partir du positionnement de nucléosomes sur des séquences ADN [165].*

dages considérés seront ainsi définis sur un alphabet tri-nucléotide. La complémentarité des deux brins de l'ADN, c'est-à-dire le fait qu'une base ne puisse s'appareiller qu'avec une et une seule base spécifique, implique que la valeur associée à un tri-nucléotide doit aussi être celle associée à son tri-nucléotide complémentaire miroir, les deux brins étant anti-parallèles. Ainsi, la valeur associée à ACG sera identique à celle associée à CGT.

Le *codage PNuc*, défini par la table 2.1, résulte de l'observation du mode de positionnement préférentiel des nucléosomes sur des séquences ADN et tente de rendre compte de la courbure spontanée de la double hélice dans le complexe nucléosomal. Ce codage ne peut être simplifié.

Proposition 2.7 *Le codage PNuc défini sur les tri-nucléotides est un codage fondamental.*

Preuve. Soit τ le codage PNuc. Montrons qu'il ne peut se décomposer en deux codages τ_1 et τ_2 respectivement définis sur L^2 et L. Ce problème conduit à considérer un système linéaire. Supposons donc que $\tau(l_1l_2l_3) = \tau_1(l_1l_2) + \tau_2(l_3)$, avec $l_j \in L$. En prenant $l_1 = l_2 = A$, les valeurs du tableau 2.1 permettent d'obtenir $\tau_2(C) - \tau_2(A) = 3.7$. En prenant $l_1 = l_2 = T$, ce même tableau donne $\tau_2(C) - \tau_2(A) = 1$, d'où la contradiction. Par symétrie, il n'existe pas de décomposition de τ en deux codages τ_1 et τ_2, respectivement définis sur L et L^2.

Tri-base	DNase I	Tri-base	DNase I	Tri-base	DNase I	Tri-base	DNase I
AAA	0.1	CAA	6.2	GAA	5.1	TAA	7.3
AAC	1.6	CAC	6.8	GAC	5.6	TAC	6.4
AAG	4.2	CAG	9.6	GAG	6.6	TAG	7.8
AAT	0	CAT	5.2	GAT	3.6	TAT	9.7
ACA	5.8	CCA	0.7	GCA	7.5	TCA	10
ACC	5.2	CCC	5.7	GCC	8.2	TCC	6.2
ACG	5.2	CCG	3	GCG	4.3	TCG	5.8
ACT	2	CCT	4.7	GCT	6.3	TCT	6.5
AGA	6.5	CGA	5.8	GGA	6.2	TGA	10
AGC	6.3	CGC	4.3	GGC	8.2	TGC	7.5
AGG	4.7	CGG	3	GGG	5.7	TGG	0.7
AGT	2	CGT	5.2	GGT	5.2	TGT	5.8
ATA	9.7	CTA	7.8	GTA	6.4	TTA	7.3
ATC	3.6	CTC	6.6	GTC	5.6	TTC	5.1
ATG	8.7	CTG	9.6	GTG	6.8	TTG	6.2
ATT	0	CTT	4.2	GTT	1.6	TTT	0.1

TAB. 2.2 – *Table de courbure obtenue (après un simple rééchelonnement linéaire entre* 0 *et* 10*) par expérience de type désoxyribonucléase I [84, 85, 155].*

De la même manière, on peut montrer qu'il n'existe pas de décomposition du codage PNuc τ en deux codages τ_1 et τ_2 tels que $\tau(l_1l_2l_3) = \tau_1(l_1l_3) + \tau_2(l_2)$. □

Remarque 2.8 Le codage PNuc est corrélé au codage pourcentage en GC défini par les égalités (2.5). Pour s'en persuader, il suffit de construire la table GC obtenue en associant à un tri-nucléotide $l \in L^3$ son codage GC, $\tau_{GC}(l)$. Le coefficient de corrélation entre ces deux tables vaut $r = 0.732 \pm 0.001$. ▫

Le *codage DNase I*, donné par la table 2.2, est obtenu à partir de la digestion de la désoxyribonucléase I (DNase I). La DNase I est une enzyme sans affinité pour une séquence de nucléotides spécifique, qui coupe l'ADN en le pliant. Cette table peut-être interprétée, avec prudence, comme étant l'énergie nécessaire pour déformer localement la double hélice à partir de sa position d'équilibre. Elle tente donc de rendre compte de la souplesse locale de l'ADN. Comme le codage PNuc, le codage DNase I est un codage fondamental.

Proposition 2.9 *Le codage DNase défini sur les tri-nucléotides est un codage fondamental.*

Preuve. Il suffit de suivre les mêmes étapes que pour la démonstration de la proposition 2.7. □

Remarque 2.10 Le codage DNase est faiblement corrélé au pourcentage en GC ($r = 0.102 \pm 0.001$) et au codage PNuc ($r = 0.276 \pm 0.001$). □

Les codage biais

Les codages biais sont des codages mettant en évidence les dissymétries locales dans la composition nucléotidique [248, 249]. Ces dissymétries sont riches en informations : en comprenant pourquoi elles existent, nous serons à même de détecter et d'analyser des zones de la séquence ADN de première importance fonctionnelle. Nous remercions Philippe StJean pour les discussions concernant l'indépendance des codages.

Commençons par introduire les *codages sélectifs*. Pour toutes lettres l_0 et l de l'alphabet L, on introduit le codage τ_{l_0} comme suit,

$$\tau_{l_0}(l) = \begin{cases} 1 & \text{si } l_0 = l, \\ 0 & \text{si } l_0 \neq l. \end{cases} \tag{2.6}$$

Ainsi, $\tau_A(A)$ vaut 1 et $\tau_A(C) = \tau_A(G) = \tau_A(T) = 0$.

Soit m un mot construit sur l'alphabet nucléotidique contenant au moins un exemplaire de chacune des lettres A, C, G et T. Le *codage biais* TA de m est défini comme suit,

$$\Delta_{TA}(m) = \frac{\tau_T(m) - \tau_A(m)}{\tau_T(m) + \tau_A(m)}. \tag{2.7}$$

De la même manière, pour le *codage biais* GC, on pose

$$\Delta_{GC}(m) = \frac{\tau_G(m) - \tau_C(m)}{\tau_G(m) + \tau_C(m)}. \tag{2.8}$$

On introduit aussi le *codage biais total* [297, 299, 368, 369],

$$\Delta(m) = \Delta_{TA}(m) + \Delta_{GC}(m). \tag{2.9}$$

Ces codages ne sont pas associés à une taille de mot en particulier.

Les bruits associés à ces codages sont définis par

$$S_{TA}(n) = \Delta_{AT}\big(\pi(m,nN,N)\big), \tag{2.10}$$

$$S_{GC}(n) = \Delta_{GC}\big(\pi(m,nN,N)\big), \tag{2.11}$$

et

$$S(n) = S_{TA} + S_{GC} = \Delta\big(\pi(m,nN,N)\big). \tag{2.12}$$

Le signal S est appelé le *signal biais*. Le pas N représente la largeur de la fenêtre utilisée pour calculer les biais. Pour que la variable n représente le milieu de cette fenêtre, nous translaterons ces fonctions en posant $S'(n) = S(n + \lceil N/2 \rceil)$. Dans la suite, nous utiliserons presque exclusivement une largeur de fenêtre N égale à 1 kpb pour analyser les génomes. Cette taille correspond à un bon compromis entre résolution nécessaire à l'échelle des chromosomes et statistique suffisante dans le calcul du signal biais. Cette échelle permet de distinguer une grande partie des gènes humains (dont la taille caractéristique est de quelques dizaines de milliers de paires de bases), mais ne permet pas d'identifier d'éventuelles séquences fonctionnelles (exons, ou promoteurs de gènes), dont la taille caractéristique est en général inférieure à quelques centaines de paires de bases.

Le codages biais TA (resp. GC) évalue la différence relative de concentrations entre les bases T et A (resp. G et C). Si la fenêtre centrée en n présente autant de nucléotides A que de nucléotides T, $S_{TA}(n)$ est nul. Il prend des valeurs extrêmes (1 ou -1) si un des deux nucléotides n'est pas présent dans la sous-séquence considérée. Nous nous attarderons plus longuement sur l'interprétation de ces codages dans les prochains chapitres.

Concernant les relations entre les biais GC et TA ou entre le biais GC et le pourcentage en GC, il peut être montré que ces codages sont indépendants.

Remarque 2.11 Le biais en GC et le pourcentage en GC sont deux codages différents. Pour un mot m, si le pourcentage en GC vaut α, $\tau_{GC}(m) = \alpha > 0$, le biais en GC peut prendre $\alpha + 1$ valeurs différentes de la forme $S_{GC}(m) = \gamma/\alpha$, avec $\gamma \in \{-\alpha, -\alpha + 2, \ldots, \alpha - 2, \alpha\}$. Plus précisément, si α désigne le pourcentage en GC d'un mot m et δ le nombre de lettres G moins le nombre de lettres C dans ce même mot, alors, si l'on considère α et δ comme des variables aléatoires non-triviales indépendantes, le biais en GC et le pourcentage en GC sont eux aussi indépendants. Le même genre de considérations s'applique pour les relations entre S_{TA} et S_{GC}. □

Exemples de signaux ADN obtenus par divers codages

Nous illustrons ici les signaux obtenus avec différents codages appliqués sur une partie du chromosome 21 de l'homme. Selon la taille de la séquence considérée, le signal obtenu revêt la forme d'un bruit corrélé (la marche ADN correspondante ressemblant à un mouvement brownien fractionnaire persistant) ou semble osciller lentement (comportements basse fréquence).

La figure 2.2 représente, de haut en bas, les signaux associés aux codages purine-pyrimidine, mono-nucléotidique A, GC, PNuc et DNase pour le chromosome 21 de l'homme.

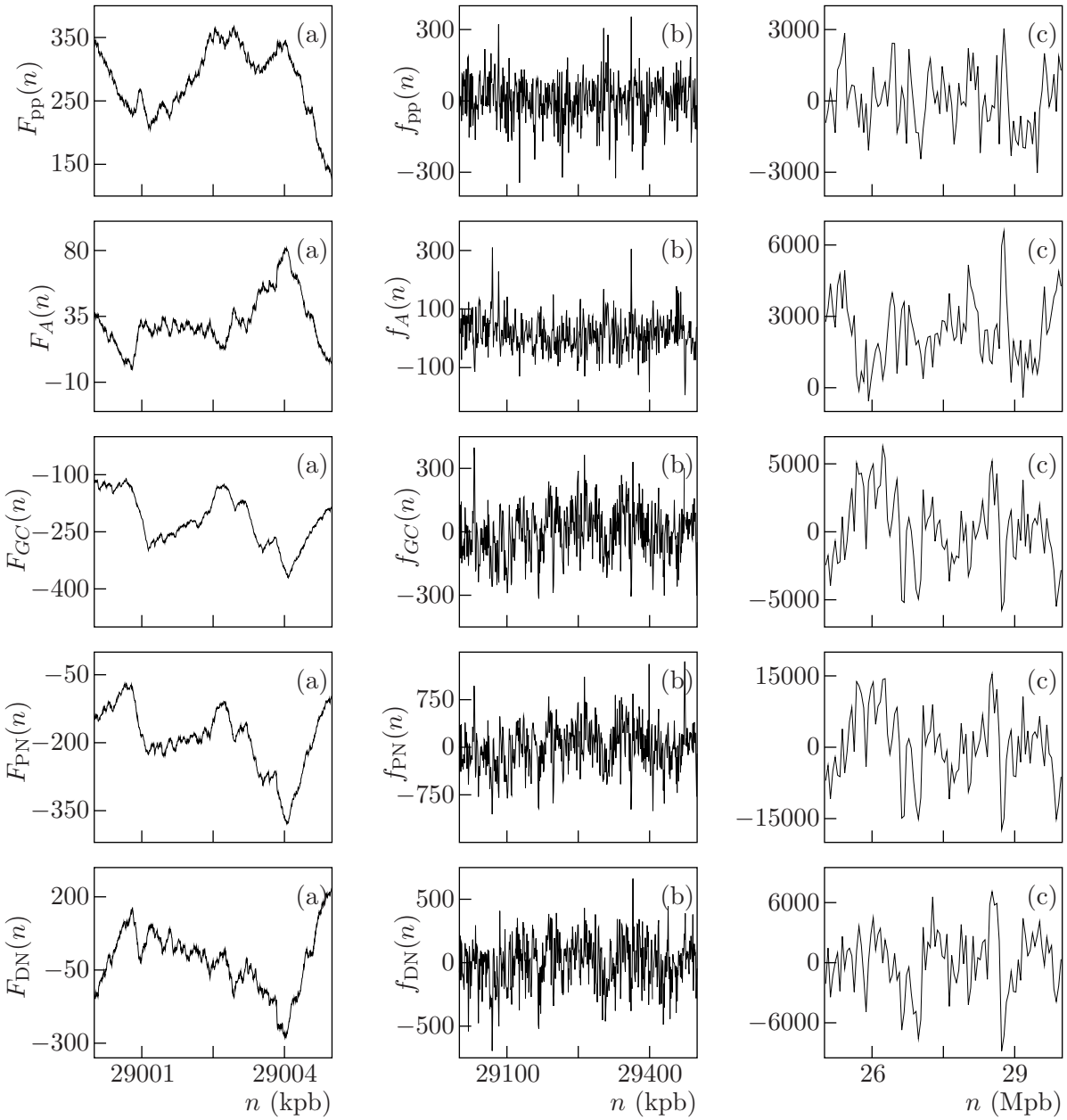

FIG. 2.2 – *Les marches associées à une partie du chromosome 21 de l'homme calculées avec les codages (de haut en bas) purine-pyrimidine, mono-nucléotidique A, GC, PNuc et DNase (a) et le bruit correspondant pour des fenêtres de largeur* 1 *kpb (b) et* 50 *kpb (c). Le comportement de la marche (a) ressemble à celui d'un mouvement brownien fractionnaire dont les pas sont donnés par le bruit correspondant (b). Pour des tailles de fenêtre plus grandes (c), le bruit semble osciller lentement.*

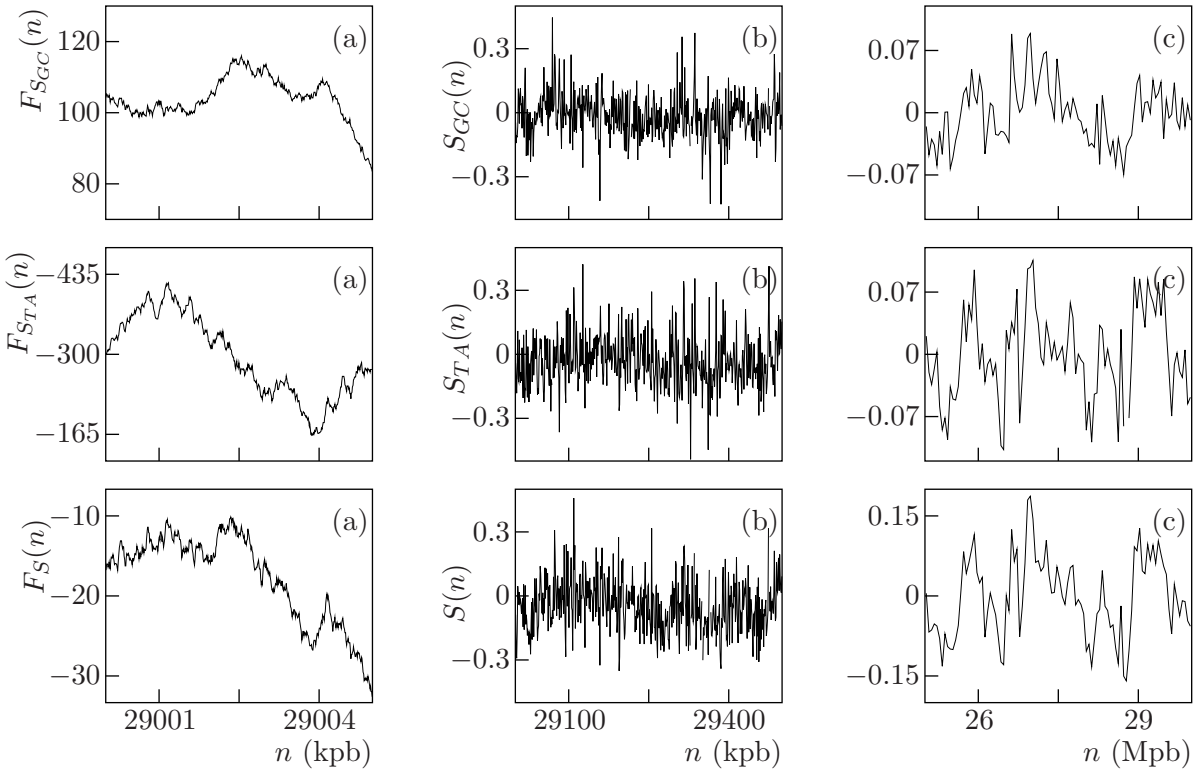

FIG. 2.3 – *Les marches associées aux codages biais GC, TA et total pour une partie du chromosome 21 de l'homme (a) et le bruit associé aux mêmes codages calculé avec des fenêtres de largeur* 1 *kpb (b) et* 50 *kpb (c). Le comportement de ces signaux est qualitativement le même que celui observé dans la figure 2.2.*

Pour les codages à valeurs uniquement positives, la moyenne du bruit a été retranchée. Les fenêtres (a) correspond à la marche ADN calculée sur une séquence m de taille 5 kpb, $F_\tau(n) = \tau(m,0,n)$, où τ est le codage considéré. Le bruit associé à chaque codage a été calculé dans des fenêtres de largeur 1 kpb, $f_\tau(n) = \tau(m,10^3 n,10^3)$ (b) et 50 kpb, $f_\tau(n) = \tau(m,5\,10^4 n,5\,10^4)$ (c), pour des mots m de taille 500 kpb et 5 Mpb respectivement. Si l'on considère les marches ADN (et les bruits correspondants calculés dans des fenêtres de largeur 1 kpb), les profils obtenus ressemblent qualitativement à des mouvements browniens fractionnaires, discutés et illustrés dans la section 3.2 de la première partie (figure 3.1). Pour des fenêtres de taille plus grande, le bruit semble présenter des comportements oscillatoires plus réguliers, avec des fréquences caractéristiques correspondant à des tailles de 100 kpb, voire quelques centaines de kpb (figure 2.2 (c)). Nous tenterons de préciser l'existence de tels rythmes « basse fréquence » dans la prochaine section.

En ce qui concerne les codages biais, que nous étudierons plus en détail dans les prochains chapitres, les profils obtenus présentent les mêmes caractéristiques (figure 2.3). Les marches ont toutes le comportement d'un mouvement brownien fractionnaire (figure 2.3 (a)) dont les pas, *i.e.* le bruit correspondant (figure 2.3 (b)), sont corrélés positivement. À nouveau, les bruits présentent un caractère oscillant lorsque des tailles suffisantes sont considérées (figure 2.3 (c)). Le biais total S (*cf.* relation (2.12)) présente le même type d'oscillations que les biais GC et TA. En fait, on peut même discerner la présence d'oscillations dans le biais total calculé dans des fenêtres de largeur 1 kpb (figure 2.3 (b)). Cette observation justifiera l'étude privilégiée que nous ferons du codage biais total dans la suite par rapport aux codages biais TA et GC.

Étude fréquentielle des signaux ADN

L'analyse des bruits, obtenus avec les divers codages envisagés, par l'ondelette de Morlet (définition 2.20 de la première partie) va nous permettre de mettre en évidence l'existence d'une composante gaussienne fractionnaire qui se superpose à l'existence de rythmes de basse fréquence, dont nous donnerons ici une interprétation de nature structurelle.

Afin d'explorer le contenu fréquentiel des signaux issus de codages des séquences ADN du génome humain, nous avons procédé de la manière suivante. Étant donné un codage τ, pour chacun des 22 chromosomes asexués de l'homme, la transformée en ondelettes du bruit associé à τ calculé dans des fenêtres de largeur 1 kpb, $f_\tau(n) = \tau(m_j,1000n,1000)$, est effectuée avec l'ondelette mère de Morlet. Pour un chromosome, toutes les échelles numériquement accessibles sont explorées ; par contre, les valeurs de la transformée affectées

par les effets de bord ne sont pas considérées. Ensuite, la fonction

$$\Lambda(a) = \mathrm{E}|W_{\psi_M} f_\tau(\,.\,,a)|, \tag{2.13}$$

appelée *spectre d'échelles* (*scalogramme*) de f_τ est calculée. Finalement, le *spectre d'échelles moyen* $\tilde{\Lambda}$, défini comme la moyenne des signaux Λ sur tous les chromosomes, est évalué.

Nous avons vu dans le chapitre 2 de la première partie que la présence d'oscillations de basse fréquence devrait se manifester dans le spectre moyen par l'existence de pics (ou de bosses) aux valeurs de l'échelle a correspondant aux périodes caractéristiques de ces oscillations (*cf.* figure 2.4 de la première partie). Si le signal comporte une composante bruit fractionnaire à haute fréquence, le spectre moyen devrait se comporter en loi de puissance aux petites échelles. En effet, un tel comportement est caractéristique des distributions auto-similaires ; ici, nous gardons volontairement floue la notion d'auto-similarité. Pour le voir, donnons le raisonnement heuristique suivant, reposant sur la relation (2.21) de la première partie. Supposons qu'une fonction $F \in L^2$ vérifie la relation $F(\lambda t) \approx \lambda^H F(t)$ pour tout t, sans que nous ne définissions la signification du symbole d'égalité $\approx$; on peut par exemple demander l'égalité des distributions. Dans ce cas, il est naturel de supposer que la relation suivante est aussi satisfaite,

$$\mathrm{E}|W[DF](\,.\,,a)| \approx C a^{H-1}. \tag{2.14}$$

En pratique, la question de la régularité de F ne se pose pas car le bruit $f = DF$ associé à F est considéré uniquement par l'intermédiaire de sa transformée en ondelettes, avec des ondelettes mères suffisamment régulières (*cf.* relation (2.21) de la première partie). Ainsi, le comportement auto-similaire d'exposant H d'une marche ADN doit se manifester par un comportement linéaire de coefficient angulaire $H - 1$ dans le logarithme du spectre $\log_2 \Lambda$ associé au bruit correspondant en fonction du logarithme de l'échelle $\log_2 a$. C'est par exemple le cas si la fonction considérée est un mouvement brownien fractionnaire $B_{H=0.72}$ dont les incréments définissant le bruit associé sont corrélés positivement.

La figure 2.4 représente le logarithme du spectre moyen en fonction du logarithme de l'échelle calculé pour les bruits associés aux divers codages jusqu'ici considérés (à l'exception du codage biais, qui sera, rappelons-le, spécialement considéré dans la suite). Un comportement linéaire est apparent pour les échelles a inférieures à $2^5 = 32$ kpb dans la plupart des signaux. Remarquons que les plus petites échelles sont affectées par la discrétisation du signal à 1 kpb. Le coefficient angulaire (égal à $H - 1$) mesuré dans les gamme d'échelles correspondant au comportement auto-similaire nous permet d'estimer un exposant d'auto-similarité voisin de $H = 0.72$ pour les marches ADN obtenues en cumulant chaque bruit (un mouvement brownien fractionnaire avec un tel exposant est représenté dans la figure 2.4 (a)).

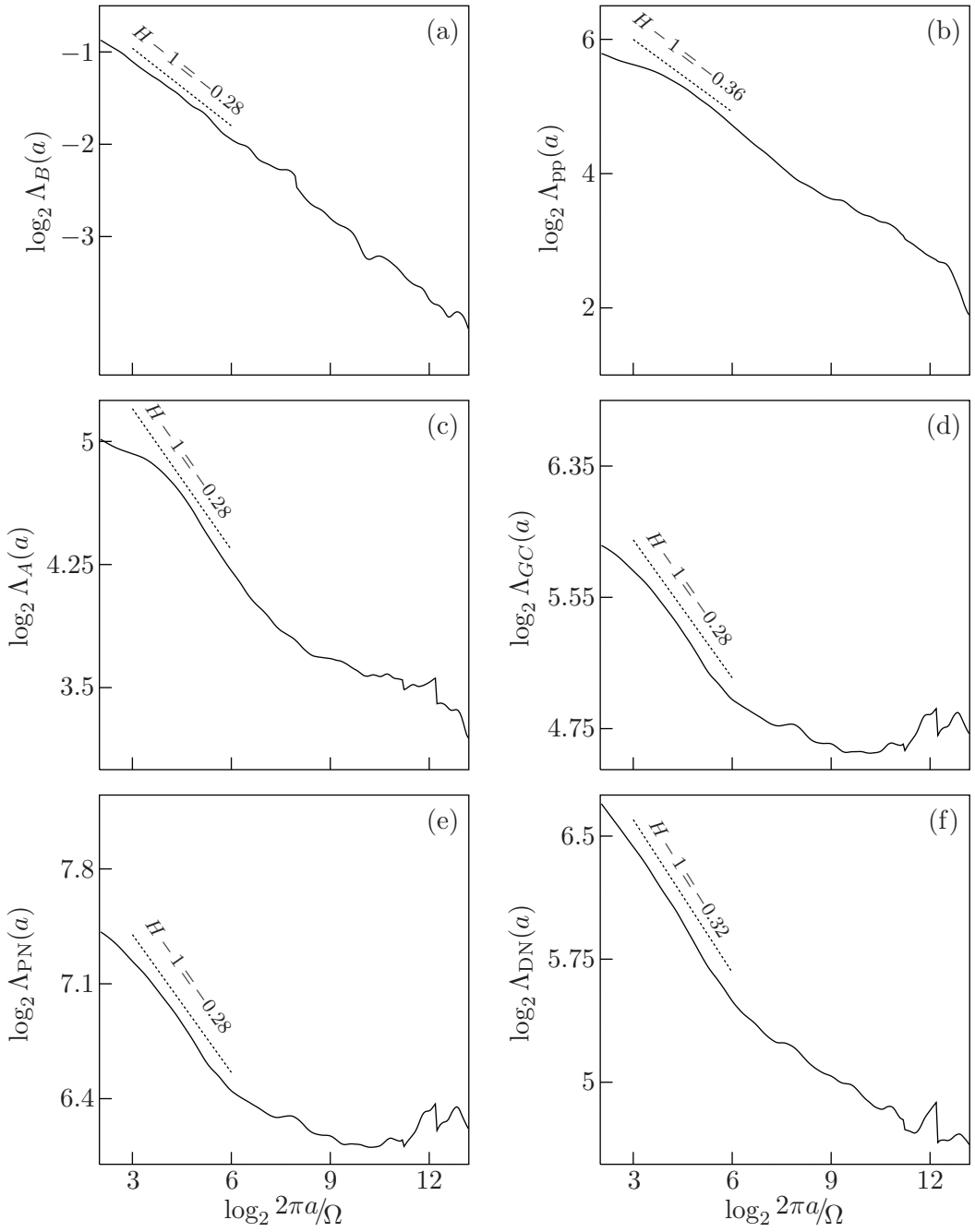

FIG. 2.4 – *Le logarithme du spectre d'échelles* Λ *en fonction du logarithme de l'échelle pour les incréments d'un mouvement brownien fractionnaire* ($H = 0.72$) *(a) et les bruits associés aux codages purine-pyrimidine (b), mono-nucléotidique A (c), GC (d), PNuc (e) et DNase (f) des* 22 *chromosomes asexués de l'homme, calculés dans des fenêtres de largeur* 1 *kpb. Les profils (c), (d), (e) et (f) sont représentés dans le même repère (et peuvent donc être comparés). Vu l'absence d'oscillation visible, une autre gamme d'échelles a dû être choisie pour les profils (a) et (b).*

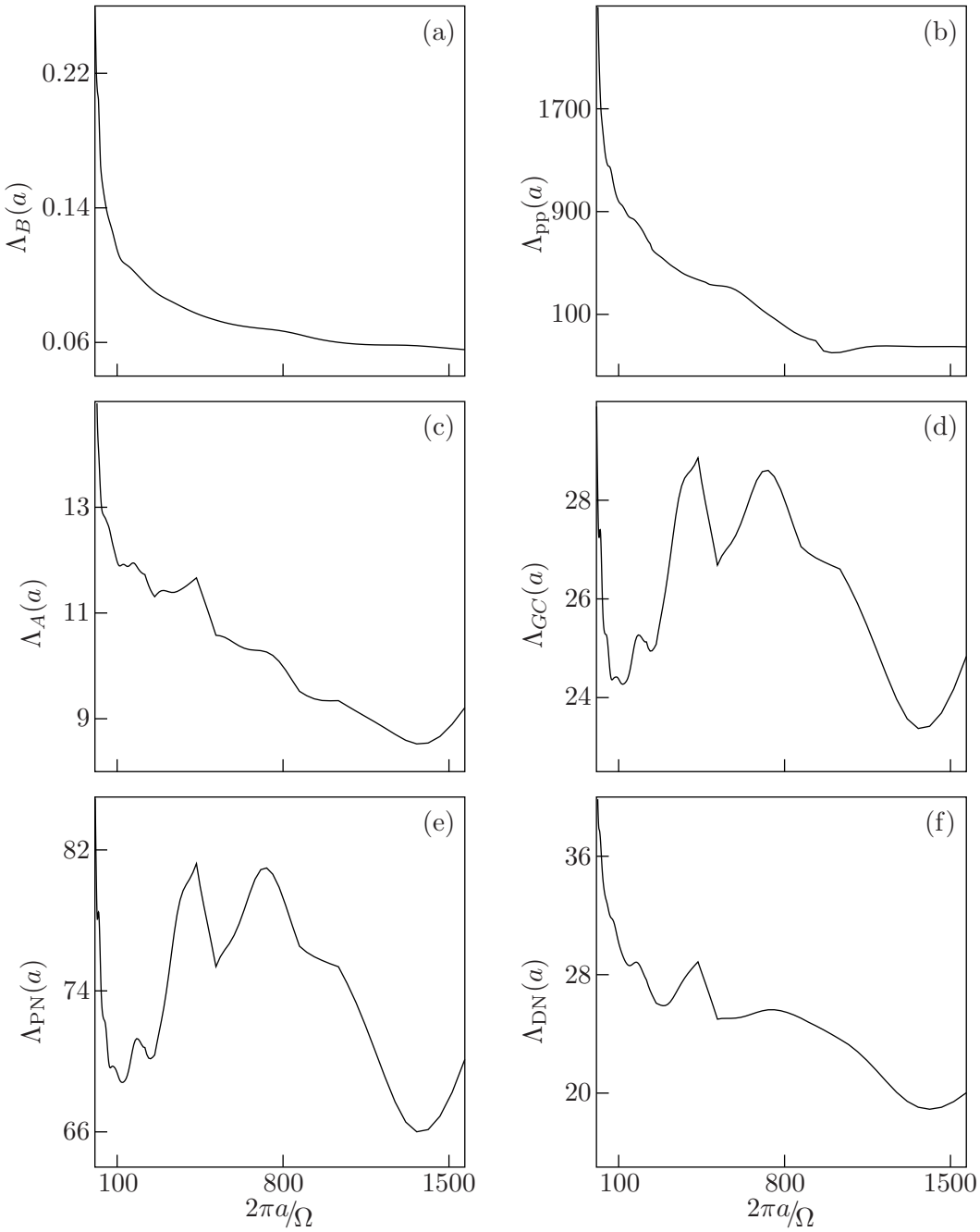

FIG. 2.5 – *Le spectre en fonction de l'échelle pour les incréments d'un mouvement brownien fractionnaire ($H = 0.72$) (a) et les bruits calculés dans des fenêtres de largeur* 1 *kpb associés aux codages purine-pyrimidine (b), mono-nucléotidique A (c), GC (d), PNuc (e) et DNase (f) relatifs aux* 22 *chromosomes asexués de l'homme.*

Aux échelles supérieures à 32 kpb, on observe la présence de pics ou de bosses brisant l'invariance d'échelle dans quasiment tous les spectres d'échelles, représentés en échelles logarithmiques dans la figure 2.4. L'examen en échelles linéaires de ces spectres dans la figure 2.5 montre que pour tous les codages, excepté le codage purine-pyrimidine, des maxima bien définis correspondant aux périodes 170 ± 50, 420 ± 90 et 730 ± 70 kpb sont observés. Si on exclut le codage DNase, la période 965 ± 50 kpb est aussi détectée. Remarquons que les spectres des codages exclus présentent aussi des maxima correspondant aux mêmes périodes, mais les bosses sont beaucoup plus étalées. Les maxima sont particulièrement présents pour les spectres concernant les codages PNuc et GC. Ces deux codages étant corrélés, ces spectres sont d'ailleurs similaires. Puisque le codage PNuc est un codage relatif à la structure nucléosomale de l'ADN, on peut penser que les périodes observées sont en relation avec les niveaux supérieurs d'empaquetage de l'ADN dans le noyau des cellules eucaryotes. La première période, proche de 150 kpb, n'est pas très éloignée de la taille des boucles d'ADN observée par diverses techniques expérimentales [52, 78, 108, 160, 232, 319]. Les autres périodes peuvent donc correspondre à la taille de boucles de chromatine plus grandes ou à plusieurs boucles formant des domaines structurels relativement autonomes [289].

2.2 Existence de corrélations à longue portée au sein des séquences ADN : de l'analyse à la synthèse

Vu l'analogie évidente entre les marches ADN présentées dans la section précédente et les marches aléatoires, en particulier les mouvements browniens fractionnaires, il est naturel d'étudier ces signaux, issus de séquences nucléotidiques, du point de vue du formalisme multifractal. L'existence de corrélations à longue portée a été établie pour la plupart des codages [21, 26, 34, 35]. Après avoir rappelé brièvement ces résultats, nous présenterons un algorithme permettant de construire des séquences nucléotidiques présentant des corrélations à longue portée *via* leur codage [298].

Existence de corrélations à longue portée dans les marches ADN

L'existence de corrélations dans les séquences est un résultat établi. Nous rappelons brièvement les conclusions obtenues lors des précédentes études [34, 35], en les illustrant avec le spectre d'échelles.

Commençons d'abord par un raisonnement heuristique concernant l'auto-similarité d'un signal et le formalisme multifractal.

Remarque 2.12 Supposons que la fonction $F \in L^2$ vérifie, comme précédemment, une relation du type $F(\lambda t) \approx \lambda^H F(t)$ pour tout t. On s'attend à observer

$$Z(a,q) \approx C_q N_a \, a^{qH}, \tag{2.15}$$

pour une constante C_q, où N_a représente le nombre de lignes de maxima du module[*] se prolongeant à l'échelle a, Z étant la fonction de partition obtenue par la méthode des maxima du module de la transformée en ondelettes et définie par l'égalité (2.111) de la première partie. Si tel est le cas, les fonctions $h_a(q)$ (*cf.* équation (2.127) de la première partie) devraient se comporter linéairement en fonction de $\log_2 a$, avec un coefficient angulaire égal à H pour tous les q. Une telle propriété est par exemple observée dans l'étude du mouvement brownien fractionnaire (*cf.* figures 3.2 et 3.3 de la première partie). Ces raisonnements sont sous-tendus par les résultats théoriques de l'équipe d'Arneodo [37] et de Jaffard [204] (et esquissés au chapitre 2 de la première partie). □

En pratique, si un signal semble posséder des propriétés d'auto-similarité, l'exposant d'auto-similarité peut être calculé en évaluant les coefficients angulaires relatifs aux fonctions $h_a(q)$. Le fait que ces coefficients angulaires soient identiques (*i.e.* indépendant de q) constitue un argument en faveur de l'auto-similarité. Si ce comportement est seulement observé pour une gamme d'échelle $[a_1,a_2]$, on peut penser que le signal est auto-similaire lorsqu'il est moyenné par des fenêtres dont la largeur correspond à la largeur de l'ondelette aux échelles $a_1 \leqslant a \leqslant a_2$.

Venons-en maintenant aux signaux issus de séquences nucléotidiques. Comme nous l'avons vu (figures 2.2 et 2.4), les signaux de type marche ADN semblent très similaires aux mouvements browniens lorsque l'on ne considère pas les plus grandes échelles [21, 26]. Dans cette optique, les corrélations mesurées grâce à l'ondelette de Morlet (figure 2.4) indiquent que ces marches sont corrélées à longue portée, avec des valeurs de l'exposant de Hölder voisines de $H = 0.72$. Toutefois, vu la largeur des fenêtres utilisées pour calculer le bruit (1 kpb), nous n'avons pas la résolution pour évaluer ces corrélations aux plus petites échelles. L'utilisation de la méthode des maxima du module de la transformée en ondelettes a donné lieu à d'importantes observations à ce sujet [34, 35]. Il faut distinguer deux types de résultats, selon la gamme d'échelles utilisée pour effectuer la régression linéaire sur les fonctions $h_a(q)$ en fonction de $\log_2 a$. Dans tous les cas, le coefficient angulaire mesuré ne dépend pas de q, confirmant la nature monofractale auto-similaire des marches ADN. Dans cette section, H représentera l'exposant d'auto-similarité mesuré. Nous illustrerons

[*]. Typiquement, N_a est proportionnel à 2^{-a}.

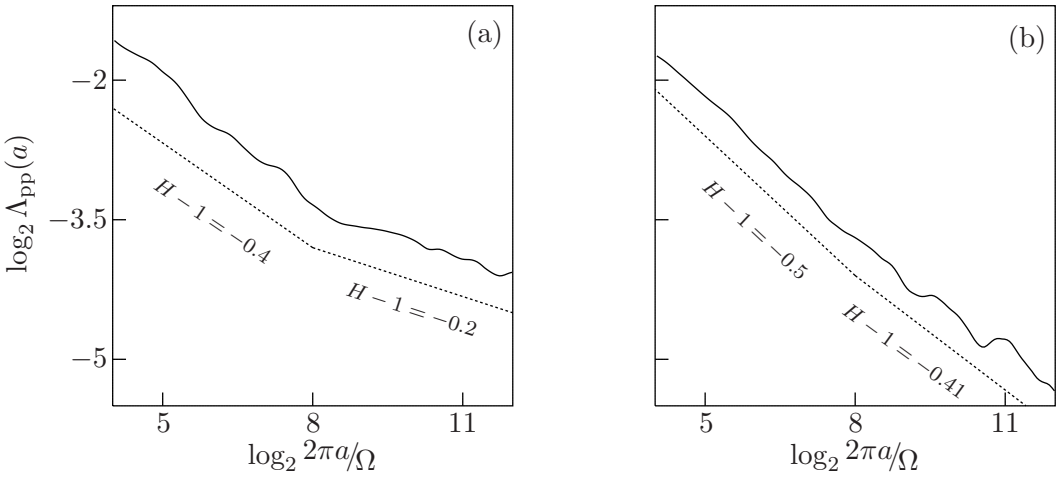

FIG. 2.6 – *Le spectre d'échelles (en représentation logarithmique) de la marche associée au codage purine-pyrimidine calculée dans des fenêtres de largeur* 1 *pb pour les 22 chromosomes asexués de l'homme (a) et pour le génome d'Escherichia coli K12 (b). L'existence de deux régimes d'invariance d'échelles séparés par l'échelle correspondant à* 200 *pb est claire. Pour le petit régime, aucune corrélation significative n'est observée chez E. coli* ($H = 0.5$)*, alors que pour l'homme on mesure une valeur* $H = 0.6$.

les résultats grâce au spectre d'échelles. Remarquons que, comme précédemment, le codage biais ne fait pas partie des codages envisagés ici ; il sera traité plus spécialement dans les chapitres qui suivent.

Si l'on se limite à des tailles inférieures ou de l'ordre de 200 pb, ce que l'on appelle le *petit régime* (pour régime des petites distances), l'exposant H dépend du type d'organisme considéré. Les organismes eucaryotes sont associés à une valeur de H voisine de 0.6 pour le codage PNuc, comme l'illustre la figure 2.6 (a). Suivant le modèle brownien, ces organismes présentent[⋇] donc des corrélations à longue portée. Le codage PNuc est le codage qui maximise la valeur de cet exposant, mais les autres codages présentent tous des exposants H significativement supérieurs à $1/2$[❋]. Pour les bactéries, les valeurs de H trouvées ne sont pas significativement différentes de $1/2$ (figure 2.6 (b)). Dans ce cas, on ne peut donc pas mettre en évidence l'existence de corrélations. Certains génomes archæbactériens présentent aussi des corrélations. Comme les bactéries, ces organismes sont des organismes procaryotes, mais certains d'entre eux possèdent une particularité, à savoir que le mécanisme d'empaquetage de leur ADN est semblable à celui des organismes eucaryotes, en ce sens qu'il fait intervenir des protéines semblables aux histones. Ces résultats suggèrent donc que les corrélations à longue portée dans cette gamme d'échelles sont la signature de la présence de nucléosomes [34, 35]. Signalons aussi que, parmi tous les or-

⋇. Lorsque nous écrivons qu'un organisme présente des corrélations à longue portée, il faut entendre qu'un des codages présente ce type de corrélations.

❋. Rappelons que $1/2$ est la valeur qui traduit l'absence de corrélation.

ganismes analysés, il n'y eut quasiment aucune exception à ces observations. La mesure de cet exposant H permet donc de déterminer si un organisme donné est un organisme eucaryote ou une bactérie (le cas des archæbactéries étant un cas particulier).

Pour le *grand régime*, c'est-à-dire pour des distances supérieures à 200 pb, tous les organismes, qu'ils soient eucaryotes ou procaryotes, présentent un H supérieur à la valeur observée pour le petit régime et significativement supérieur à $1/2$ (figure 2.6 (a) et (b)) [20, 34, 35]. Pour l'homme, les exposants H mesurés dans les 22 chromosomes asexués sont supérieurs à 0.7 et ce, quelque soit le codage utilisé.

Un modèle pour l'ADN reposant sur le mouvement brownien fractionnaire

Tout comme le mouvement brownien fractionnaire, les séquences ADN présentent des corrélations à longue portée *via* les codages simples. Nous allons ici construire des séquences ADN artificielles présentant de telles corrélations à partir des marches binaires discrètes, introduites dans la section 3.3 de la première partie [298]. Avec une légère modification dans l'algorithme, nous serons capable de synthétiser des séquences présentant deux régimes de corrélations à longue portée, ce comportement ayant été révélé pour les marches issues de codages ADN dans les références [34, 35] et illustré dans le paragraphe précédent. Nous remercions Benjamin Audit pour les discussions concernant les méthodes pour de générer des signaux présentant deux indices d'auto-similarité.

Il pourrait être tentant de générer une séquence ADN artificielle directement à partir des incréments d'un mouvement brownien fractionnaire, en choisissant la lettre nucléotidique suivant la valeur de l'incrément. La remarque suivante montre qu'il est impossible de contrôler les corrélations de cette manière.

Remarque 2.13 En s'inspirant de la méthode définissant une marche binaire à partir de mouvements browniens fractionnaires suivant l'égalité (3.22) de la première partie, on peut construire une séquence nucléotidique directement à partir d'un bruit gaussien fractionnaire $\{\Delta_H(j)\}_j$ en posant, si l_j représente la j-ième lettre ($j \in \mathbb{N}$) de la séquence désirée,

$$l_j = \begin{cases} A & \text{si } \Delta_H(j) \leqslant \alpha_1 \\ C & \text{si } \alpha_1 < \Delta_H(j) < \alpha_2 \\ G & \text{si } \alpha_2 \leqslant \Delta_H(j) < \alpha_3 \\ T & \text{si } \Delta_H(j) \geqslant \alpha_3 \end{cases}, \tag{2.16}$$

où α_1, α_2 et α_3 sont trois nombres réels définissant les concentrations en bases nucléotidiques.

Toutefois, avec cette méthode, les corrélations à longue portée pour les valeurs de H inférieures à $3/4$ ne sont pas toujours conservées, alors que cette valeur est critique pour les séquences ADN. Pour le montrer, posons $\alpha_2 = 0$ et $\alpha_3 = -\alpha_1 = \alpha$, α est un nombre positif. Si l'on applique le codage faible-fort à la séquence ainsi construite, on obtient une suite discrète binaire $\{\delta(j)\}_j$ qui peut être directement obtenue de la manière suivante,

$$\delta(j) = \begin{cases} 1 & \text{si } |\Delta_H(j)| \geqslant \alpha, \\ -1 & \text{si } |\Delta_H(j)| < \alpha. \end{cases}$$

La proposition 3.31 de la première partie nous apprend notamment que les corrélations à longue portée ne sont pas conservées lorsque l'on passe d'un bruit gaussien fractionnaire à la suite de ses modules lorsque $H \leqslant 3/4$. Cette méthode de génération de séquence nucléotidique ne permet donc pas de contrôler ce type de corrélations. □

Méthode de synthèse Pour contourner ce problème, nous allons utiliser deux mouvements browniens fractionnaires pour construire une suite nucléotidique. Notre approche repose sur la simple remarque suivante : en base deux, if faut deux bits pour énumérer les lettres de l'alphabet L, et il y a $4! = 24$ manières différentes d'ordonner cet alphabet. Par exemple, en ordonnant les lettres de L de la manière suivante, $C < T < G < A$, chaque nucléotide est associé de manière univoque à un mot de l'alphabet $\{0,1\}^2$ comme suit : $C = 00$, $T = 01$, $G = 10$ et $A = 11$. Si on remplace l'alphabet $\{0,1\}^2$ par $\{-1, +1\}^2$, on obtient l'association

$$A = +1 + 1, \quad G = +1 - 1, \quad C = -1 - 1, \quad T = -1 + 1. \tag{2.17}$$

En faisant correspondre L et $\{-1, +1\}^2$ de cette manière, si on définit le codage qui à un nucléotide associe le bit de poids faible de la représentation précédente, c'est-à-dire le second élément de la représentation binaire des nucléotides (2.17), on obtient le codage faible-fort. De la même manière, le codage purine-pyrimidine détermine le bit de poids fort.

Nous appellerons *suite binaire* tout mot construit sur l'alphabet $\{-1, +1\}$. Comme l'illustre l'exemple précédent, à une séquence nucléotidique correspond deux suites binaires, c'est-à-dire deux codages simples. Nous pouvons ainsi donner une méthode d'obtention de suite nucléotidique présentant des corrélations à longue portée par l'intermédiaire de leurs codages simples, grâce aux marches binaires discrètes. À partir de deux mouvements browniens fractionnaires d'indice respectif H_1 et H_2, on construit deux bruits binaires discrets associés δ_1 et δ_2, suivant l'algorithme donné dans la section 3.3 de la première partie. Pour une abscisse j donnée, le couple $\{\delta_1(j),\delta_2(j)\}$ détermine une lettre nucléotidique, grâce à une représentation du type (2.17). On définit ainsi un mot sur l'alphabet L dont la j-ième lettre est déterminée par ce couple $\{\delta_1(j),\delta_2(j)\}$. Donnons un

exemple. Supposons avoir deux réalisations de bruit gaussien fractionnaire Δ_{H_1} et Δ_{H_2}, d'indice respectif H_1 et H_2. En adoptant la convention (2.17), si $\Delta_{H_1}(j)$ et $\Delta_{H_2}(j)$ sont tous deux positifs, la j-ème lettre de la suite nucléotidique construite sera A. Si par-contre, $\Delta_{H_2}(j)$ est négatifs (Δ_{H_1} est toujours positif), cette lettre sera G, *et cætera*. On associe donc à la convention (2.17) la règle suivante,

$$\begin{array}{ll} \left.\begin{array}{l}\Delta_{H_1}(j)>0\\ \Delta_{H_2}(j)>0\end{array}\right\} \longrightarrow A, & \left.\begin{array}{l}\Delta_{H_1}(j)>0\\ \Delta_{H_2}(j)<0\end{array}\right\} \longrightarrow G, \\ \left.\begin{array}{l}\Delta_{H_1}(j)<0\\ \Delta_{H_2}(j)>0\end{array}\right\} \longrightarrow T, & \left.\begin{array}{l}\Delta_{H_1}(j)<0\\ \Delta_{H_2}(j)<0\end{array}\right\} \longrightarrow C. \end{array} \tag{2.18}$$

Par construction, deux des codages simples associés à la séquence, ceux utilisés pour indexer l'alphabet L (les codages purine-pyrimidine et faible-fort si l'on utilise la convention (2.17)), présenteront des corrélations à longue portée d'indice H_1 et H_2 respectivement. Concernant le troisième codage possible, les corrélations sont déterminées par les deux premiers codages et on ne peut imposer son indice de corrélation. La séquence est ainsi définie par deux indices, H_1 et H_2.

Modification des concentrations nucléotidiques Cette méthode permet aussi de modifier les concentrations des nucléotides dans la séquence artificielle. En effet, ces concentrations déterminent le rapport entre le nombre d'éléments positifs et négatifs dans chaque codage simple, et donc la moyenne des suites binaires correspondantes. Supposons que l'on veuille, dans la règle de construction (2.18), imposer à la séquence artificielle un pourcentage en GC égal à c. Une telle concentration impose bien-sûr que, dans la marche associée au codage faible-fort, le pourcentage d'incréments négatifs soit égal à c. Pour obtenir une telle suite binaire, il suffit de choisir α_c tel que $\mathrm{P}(\Delta_{H_2}(j) < \alpha_c) = c$. La proposition 3.36 de la première partie nous apprend que la suite ainsi construite présentera les mêmes corrélations à longue portée que le bruit gaussien fractionnaire Δ_{H_2}. En remplaçant les inégalités $\Delta_{H_2}(j) > 0$ et $\Delta_{H_2}(j) < 0$ par $\Delta_{H_2}(j) > \alpha_c$ et $\Delta_{H_2}(j) < \alpha_c$ respectivement dans la règle de construction (2.18), on obtient une séquence nucléotidique avec les concentrations en GC désirées.

Synthèse de séquences n-nucléotidiques Avec le même raisonnement, il est possible de générer des séquences ADN artificielles définies par plus de deux indices. Il suffit de ne plus considérer une séquence nucléotidique comme un mot construit sur l'alphabet L, mais sur l'alphabet L^n. Par exemple, si les lettres utilisées sont les éléments de L^3, *i.e.* les tri-nucléotides, ce ne sont plus deux codages qu'il faut utiliser mais $\log_2(4^3) = 6$. Ainsi, en supposant qu'une séquence ADN est constituée de tri-nucléotides, elle est définie par six indices. Nous pouvons donc augmenter la complexité des séquences artificielles en augmentant le nombre d'indices associés à la séquence, ce qui se fait en définissant les

séquences sur un alphabet L^n, avec n de plus en plus grand.

Synthèse de séquences avec un second régime de corrélations Il est aussi possible de générer des signaux présentant deux exposants d'auto-similarité, apparaissant à des échelles différentes, c'est-à-dire d'imposer deux régimes différents de corrélations à longue portée. Pour ce faire, nous allons modifier la méthode de Sellan [341] pour générer un mouvement brownien fractionnaire à partir d'une analyse multirésolution. Dans la référence [341], la relation suivante est démontrée,

$$\begin{aligned} B_H(t) &= a_0 + \sum_k a_{H+1/2}(k)\varphi'(t-k) \\ &\quad + \sum_{j\leqslant 0,k} \gamma(k) 4^{-(H+1/2)} 2^{-(H+1/2)j} 2^{-j/2} \psi'\big(2^{-j}(t-k)\big), \end{aligned} \tag{2.19}$$

presque sûrement et uniformément sur tout compact, où $\gamma(k)$ sont des variables gaussiennes indépendantes de même distribution et $a_H(k)$ est un processus ARIMA* fractionnaire d'exposant H, que l'on peut voir comme la version discrète du mouvement brownien fractionnaire [75, 186, 336] ; φ' et ψ' sont des fonctions adéquates, jouant le rôle de la fonction d'échelle et de l'ondelette mère [284, 341]. Le terme a_0 est là pour assurer que le signal obtenu est nul à l'origine. Pour plus de détails sur cette méthode que nous ne faisons qu'esquisser, le lecteur pourra consulter les références suivantes, [119, 284, 317, 341]. Grâce à la relation (2.19), la réalisation d'un mouvement brownien fractionnaire peut être générée à l'aide d'un processus ARIMA fractionnaire et de variables gaussiennes, jouant le rôle des détails. On peut contrôler la variance σ^2 du mouvement brownien grâce à la variance σ_γ^2 des variables gaussiennes $\gamma(k)$, car on a $\sigma_\gamma^2 = \sigma^2 \Gamma(2H+1)\sin(\pi H)$ [3]. Une manière naturelle pour construire une marche présentant deux exposants d'auto-similarité consiste à modifier l'exposant du processus ARIMA fractionnaire sans modifier les détails gaussiens. Si l'on souhaite obtenir une marche aléatoire d'exposant H_1 aux petites échelles et H_2 à partir de l'octave J, on définit d'abord la fonction f_J appartenant à l'espace V_J de l'analyse multirésolution dont les coefficients sont donnés par la réalisation numérique d'un ARIMA fractionnaire d'exposant $H_2 + 1/2$. La fonction f_J constitue donc l'approximation de f à l'octave J. On construit ensuite la fonction f à partir de f_J et des détails selon la relation (2.19), mais en définissant ces détails par rapport à l'exposant H_1. On obtient ainsi la relation

$$\begin{aligned} f(t) &= a_0 + \sum_k a_{H_2+1/2}(k)\varphi'(t-k) \\ &\quad + \sum_{j\leqslant 0,k} \gamma(k) 4^{-(H_1+1/2)} 2^{-(H_1+1/2)j} 2^{-j/2} \psi'\big(2^{-j}(t-k)\big) \\ &= a_0 + f_J(t) + \sum_{j\leqslant 0,k} \gamma(k) 4^{-(H_1+1/2)} 2^{-(H_1+1/2)j} 2^{-j/2} \psi'\big(2^{-j}(t-k)\big). \end{aligned} \tag{2.20}$$

∗. ARIMA pour AutoRegressive Moving Average Integrated.

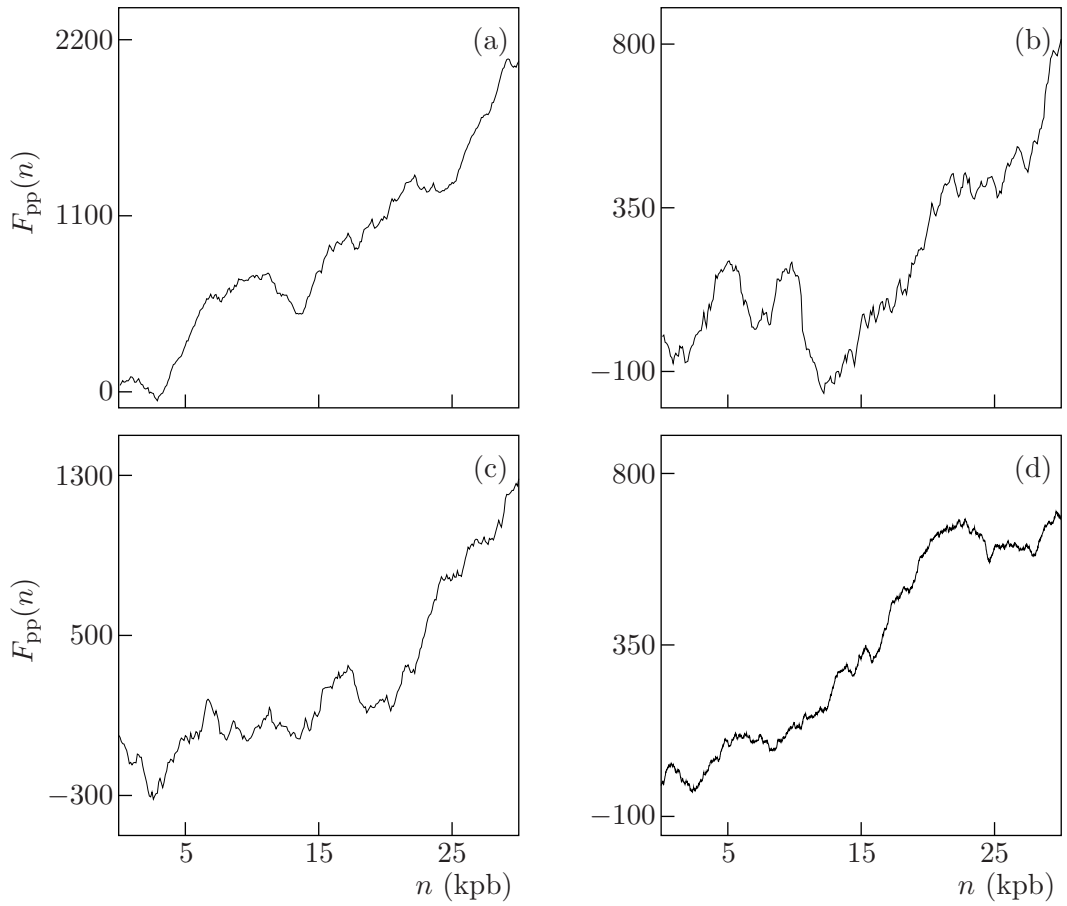

FIG. 2.7 – *Comparaison qualitative entre des marches ADN associées au codage purine-pyrimidine et des marches artificielles. (a) et (b) Les marches ADN associées au chromosome 21 et 22 de l'homme respectivement. (c) Une marche artificielle présentant un seul régime de corrélations ($H = 0.62$). (d) Une marche artificielle présentant deux régimes de corrélations, reproduisant le petit ($H_1 = 0.6$) et le grand ($H_2 = 0.8$) régimes.*

Même si cette construction est des plus formelles, le signal f obtenu présente des propriétés de corrélation satisfaisantes, comme en attestent les figures 2.7 (d) et 2.8 (b). Dans les figures 2.7 (c) et 2.8 (a) sont illustrés, pour comparaison, une marche ADN simulant un seul régime de corrélations et son spectre d'échelles.

Il est bien sûr illusoire de penser qu'un modèle aussi simple puisse décrire fidèlement les séquences ADN. L'ADN possède de nombreuses zones fonctionnelles aux propriétés particulières (boîtes $TATA$, îlots CpG, gènes, introns, codons, isochores en GC, ...) coexistant dans une même séquence. Il existe d'autres facteurs importants que le modèle ne prend pas en compte. Ainsi, le di-nucléotide CG semble fortement exposé aux mutations. Il est beaucoup moins présent que les autres di-nucléotides, et ce, quelques soient les pourcentages de mono-nucléotides dans la séquence. Cette particularité ne se retrouve évidemment pas dans les séquences générées numériquement. Finalement, le modèle ne

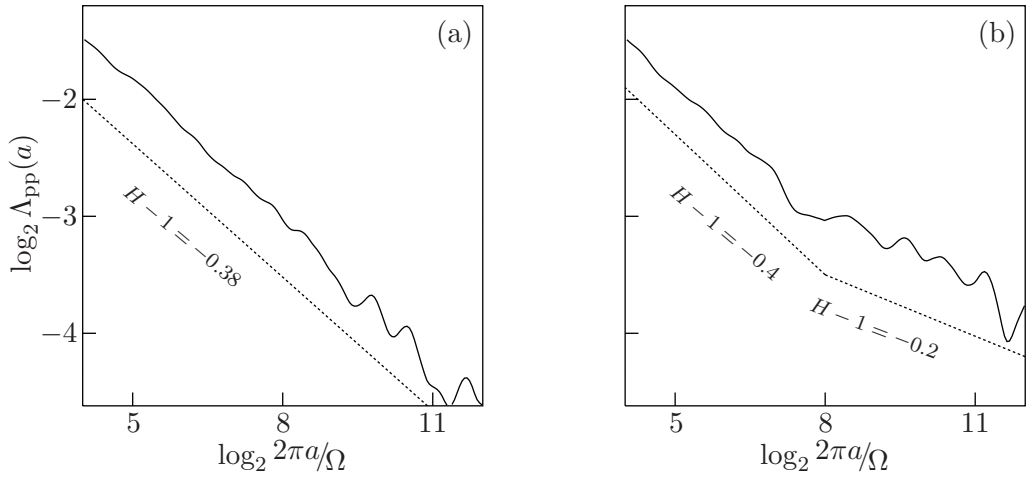

FIG. 2.8 – *Le spectre d'échelles (en représentation logarithmique) des marches aléatoirement générées données dans la figure 2.7. (a) La marche simulant un seul régime de corrélations à longue portée avec un exposant* $H = 0.62$ *similaire aux valeurs observées chez l'homme. (b) La marche simulant les deux régimes de corrélations à longue portée d'exposant* $H_1 = 0.6$ *et* $H_2 = 0.8$*. On constate l'existence d'un changement de pente au voisinage de l'échelle* 200 *pb, comme cela est observé dans les marches ADN des chromosomes asexués de l'homme (cf. figure 2.6).*

permet pas non-plus de simuler la présence de rythmes dans les codages. En conclusion, les séquences ADN présentent bien des particularités qui ne peuvent être expliquées par les corrélations à longue portée. Avant de pouvoir élaborer un modèle plus réaliste (un tel modèle serait d'une extrême complexité), il incombe de mieux connaître les zones fonctionnelles de l'ADN, qui sont pour la plupart inconnues ou dont les mécanismes de fonctionnement sont mal compris. Toutefois, les signaux artificiellement générés présentent des similitudes frappantes avec les marches ADN, comme en atteste la figure 2.7, et ils pourraient s'avérer utiles dans l'étude des conséquences sur la séquence de la présence de corrélations dans certains codages ainsi que sur la manière dont ces codages s'influencent mutuellement.

Chapitre 3

Analyse multifractale du biais de composition dans le génome humain

Grâce à l'approche multifractale, les diverses composantes de la marche ADN associée au biais de composition vont pouvoir être mises en évidence et quantifiées. Outre le comportement de type brownien fractionnaire à haute fréquence et l'existence de rythmes basses fréquences dans le bruit associé, déjà observés avec les autres codages dans le chapitre précédent, la méthode des maxima du module de la transformée en ondelettes va nous permettre de révéler l'existence de sauts dans les incréments de la marche ADN associée au biais. Avant de procéder à cette analyse, nous nous attarderons sur les mécanismes biologiques susceptibles d'influencer les valeurs que prend le signal biais de composition et d'expliquer l'existence de sauts ascendants et descendants dans ce signal.

L'étude sur la bifractalité est originale et une publication est en cours de rédaction. Les résultats concernant la dissymétrie entre le nombre de sauts ascendants et le nombre de sauts descendants constituent une partie des sujets abordés dans les publications [79, 369].

3.1 Le biais de composition

Par un argument simple de symétrie, on peut postuler que les concentrations entre nucléotides complémentaires sur un même brin d'ADN sont égales. Toutefois, des écarts, à l'échelle de quelques nucléotides, peuvent survenir du fait des mutations du matériel génétique, créant ce que l'on appelle un biais de composition.

Les mutations du matériel génétique

Les mutations sont considérées comme le moteur de l'évolution [243]. En effet, elles permettent une modification lente mais certaine du génome. Les mutations défavorables sont éliminées du patrimoine héréditaire par sélection naturelle, alors que les mutations bénéfiques tendent à s'accumuler.

Généralement, le vocable *mutation* est utilisé pour les modifications permanentes de l'ADN (ou de l'ARN), donnant lieu à une séquence nucléotidique différente de l'originale. Ces mutations peuvent notamment résulter d'erreurs de copie du matériel génétique, d'exposition à des radiations ou des virus. Pour les organismes multicellulaires, on distingue les *mutations germinales*, transmises à la descendance, des *mutations somatiques*, pouvant donner lieu à la mort d'une cellule ou provoquer le cancer [243]. Nous nous intéresserons tout particulièrement aux conséquences des mutations au cours de l'évolution en se focalisant sur les mutations germinales.

Nous ne discuterons ici que des mutations ponctuelles, n'affectant qu'un seul nucléotide, appelées *substitutions*. Les plus courantes sont les *transitions*, échangeant une purine par une purine ou une pyrimidine par une pyrimidine ($A \leftrightarrow G$ ou $C \leftrightarrow T$), tandis que les *transversions* échangent une purine par une pyrimidine ou inversement (C,$T \leftrightarrow A$,G). La *pression de sélection*, c'est-à-dire le degré d'intolérance à la mutation, varie en fonction de la zone du génome considérée. Il est évident qu'une mutation au niveau d'un exon n'aura pas les mêmes conséquences qu'une mutation au niveau d'un intron ou d'une région intergénique. Ainsi, il existe généralement plusieurs codons associés à un même acide aminé, ces codons différant principalement par la troisième base. Une mutation sur cette dernière peut donc n'avoir aucune conséquence au niveau de l'acide aminé, alors qu'elle peut changer la signification même du codon si elle porte sur l'une des deux autres bases.

Définition du biais de composition

La découverte expérimentale par Chargaff [97] de l'égalité, au sein de l'ADN, des concentrations entre les nucléotides A et T d'une part et C et G de l'autre a certainement influencé Watson et Crick pour la mise au point de leur modèle de la structure en double hélice de l'ADN [386]. Cette règle de parité est en effet trivialement vérifiée si les bases sont associées par paire spécifique, chaque nucléotide ayant un nucléotide complémentaire ; ce principe d'égalité est appelé *règle de Chargaff* ou *règle de parité de type 1*. Il existe aussi une autre observation expérimentale concernant les concentrations nucléotidiques : lorsque l'on considère des séquences suffisamment longues (c'est par exemple vrai pour des tailles de l'ordre du chromosome chez l'homme), les écarts à l'égalité de concentration entre nucléotides de type complémentaire sur un seul et même brin semble toujours rester relativement faibles [245, 248, 333]. Cette égalité semble résulter de la symétrie entre les deux brins.

En supposant que les deux brins constituant l'ADN sont symétriques, dans le sens où ils n'ont aucune raison de posséder des propriétés physiques ou chimiques différentes, le nombre de nucléotides d'un type sur un brin doit égaler le nombre de nucléotides du même type sur l'autre. Suivant cette hypothèse et le modèle de la double hélice, le nombre de nucléotides d'un type doit égaler le nombre de nucléotides du type complémentaire (A est le complémentaire de T et C celui de G) sur le même brin. Cette observation porte le nom de *règle de parité de type 2* [97, 98, 333]. Si cette règle est bien vérifiée lorsque l'on considère de grandes séquences, comme en atteste la table 3.1, elle est régulièrement violée sur des séquences de plus petites tailles. Ce sont ces écarts à l'équilibre que l'on appelle *asymétrie de composition* ou *biais*. Pour détecter ces asymétries de composition, il existe principalement deux méthodes. La première, reposant sur l'alignement de séquences homologues[+], ne sera pas abordée ici. Il existe encore trop peu de séquences disponibles pour pouvoir recourir systématiquement à cette méthode. La seconde consiste à étudier l'asymétrie de composition *via* le codage biais, défini dans la section 2.1. Ce codage permet d'obtenir un signal quantifiant les écarts locaux à l'equi-concentration, c'est-à-dire à la règle de parité de type 2.

Les mécanismes biologiques ayant pu créer des écarts à la symétrie de composition sont ceux susceptibles d'affecter différemment les deux brins, en particulier en les exposant de façon différente aux mutations. Ainsi, les mécanismes sous-jacents à la transcription et à la réplication possèdent toutes les caractéristiques pour contribuer à l'apparition d'une asymétrie de composition, puisqu'ils changent les environnements respectifs des deux brins

+. Deux séquences ADN sont homologues si elles possèdent un ancêtre commun.

chr	A	T	C	G	taille
1	26.42	26.46	18.94	18.93	245522847
2	29.18	29.24	19.65	19.66	243018229
3	29.41	29.43	19.36	19.36	199505737
4	30.2	30.21	18.68	18.69	191411218
5	29.68	29.75	19.4	19.43	180857866
6	29.56	29.55	19.37	19.38	170975699
7	28.9	28.94	19.87	19.86	158628139
8	29.19	29.15	19.58	19.58	146274826
9	24.95	24.95	17.6	17.59	138429268
10	28.37	28.4	20.21	20.2	135413628
11	28.49	28.5	20.26	20.28	134452384
12	29.1	29.12	20.07	20.06	132449811
13	25.70	25.77	16.13	16.12	114142980
14	24.44	24.63	16.95	16.99	106368585
15	23.44	23.41	17.12	17.1	100338915
16	24.46	24.57	19.85	19.93	88827254
17	26.87	26.93	22.5	22.46	78774742
18	29.51	29.55	19.49	19.53	76117153
19	22.54	22.6	21.11	21.17	63811651
20	26.46	26.79	20.99	21.06	62435964
21	21.6	21.43	14.86	14.89	46944323
22	18.3	18.21	16.83	16.81	49554710

TAB. 3.1 – *Le pourcentage de nucléotides A, C, G et T pour chaque chromosome asexué du génome humain. Les nucléotides non identifiés ont été pris en compte dans le calcul de la taille des chromosomes. On constate clairement que les nombres de A et T d'une part et de C et de G d'autre part sont voisins, et ce quelque soit le chromosome considéré.*

[50, 149, 150, 152, 171, 220, 288, 326, 335, 355, 363]. Lors de la transcription, la double hélice s'ouvre localement (figure 1.7), le brin non transcrit étant laissé à nu et donc exposé de façon différente aux mutations. Pour la réplication, l'asymétrie des deux brins résulte dans le fait que pour le brin retardé, la réplication ne se fait pas de façon continue mais par fragments (figure 1.6). Remarquons que pour la transcription, seules les parties représentant un gène sont sujettes au biais, voire même presque exclusivement les introns, alors que c'est l'entièreté de la séquence qui peut être affectée lors de la réplication. *A priori*, les biais dûs à ces deux processus essentiels au bon fonctionnement de la cellule se superposent pour contribuer à l'asymétrie de composition observée.

Séquences répétées

Il existe, dans les génomes eucaryotes, des séquences nucléotidiques apparues plus récemment dans l'évolution dont il faudra s'affranchir dans l'étude systématique du biais.

Les *séquences répétées* sont des séquences nucléotidiques apparaissant de nombreuses fois dans le génome d'un organisme eucaryote [189, 243]. Parmi ces séquences d'origines diverses [48, 189, 347, 348, 353], on distingue les LINE[❋], qui sont des séquences de longueur généralement supérieure à 5 kpb trouvées en plus de 10^4 copies dans les génomes des eucaryotes multicellulaires. Ces séquences codent pour des protéines assurant leur multiplication. Les SINE[⋇] constituent la deuxième grande famille de séquences répétées. Ils ont une taille caractéristique de 500 pb et sont trouvés en plus de 10^5 copies dans les génomes des eucaryotes multicellulaires. Ces régions se caractérisent par des fragments très riches en A ou T.

Ces séquences répétées sont apparues récemment à l'échelle de l'évolution et ont vraisemblablement été moins soumises aux mutations. Elles peuvent donc présenter une asymétrie de composition propre. Pour cette raison, sauf mention du contraire, nous ne considèrerons que les séquences nucléotidiques dont ont été masquées les séquences répétées[❋]. On peut remarquer que le signal biais revêt une apparence beaucoup plus bruitée sur les séquences natives que lorsque les séquences répétées ont été masquées, comme en atteste la figure 3.1. Cette observation confirme l'existence d'un biais spécifique aux séquences répétées de plus forte variance que le biais résiduel accumulé lors de l'évolution. Remarquons que chez l'homme, les séquences répétées constituent plus de 40% du génome [139, 189].

❋. LINE pour *Long Interspersed Nuclear Elements*.

⋇. SINE pour *Short Interspersed Nuclear Elements*. Plus généralement, les SINE appartiennent essentiellement à la famille des Alu.

❋. Pour ce faire, le logiciel *RepeatMasker* a été utilisé [349].

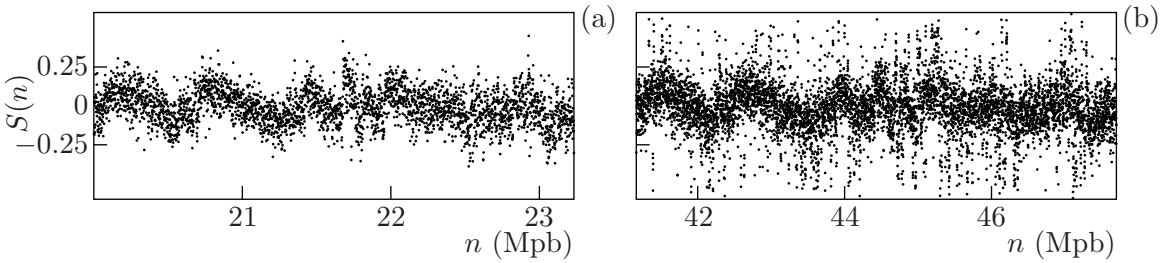

FIG. 3.1 – *Le signal biais total S (cf. égalité (2.12)) calculé dans des fenêtres de largeur* 1 *kpb sur un morceau du chromosome 12 de l'homme. Le signal a été calculé en masquant les séquences répétées (a) et sur la séquence native (b).*

3.2 Analyse multifractale du biais à l'aide de la méthode des maxima du module de la transformée en ondelettes

Jusqu'à présent, le signal biais n'a pas été considéré dans l'étude des corrélations à longue portée. Nous allons ici mettre en évidence le caractère *bifractal** de ces signaux : la marche ADN associée au biais diffère d'un mouvement brownien fractionnaire par la présence de sauts dans ses incréments, dont l'importance fonctionnelle sera discutée dans les prochains chapitres. Nous révèlerons ensuite la présence de rythmes de basse fréquence dans ce signal, que l'on rapprochera de ceux observés dans les codages structurels étudiés dans le chapitre 2.

Comportement statistique du signal biais dans le génome de l'homme

Nous nous proposons dans un premier temps de comparer la statistique des fluctuations du signal biais relatif au génome humain à une statistique gaussienne. Nous allons mettre en évidence l'existence d'écarts à une telle statistique, dans la mesure où les marches ADN correspondantes diffèrent d'un mouvement brownien fractionnaire par les valeurs extrêmes de leurs incréments, qui sont bien plus nombreuses qu'escomptées pour une distribution gaussienne.

La figure 3.2 représente l'histogramme des valeurs du signal biais total S (*cf.* égalité (2.12)) calculé sur les 22 chromosomes asexués de l'homme, dans des fenêtres de largeur 1 kpb. Sur les séquences natives, on constate dans la figure 3.2 (b) que la distribution des valeurs de S s'écarte notablement d'une statistique gaussienne ; en particulier, les

∗. La notion de bifractalité est détaillée dans la section 8.5.2 dans la référence [154]

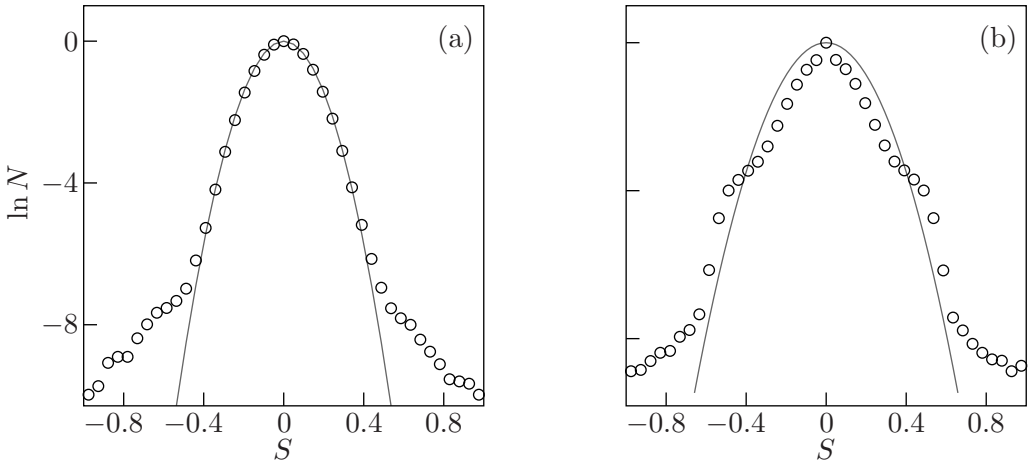

FIG. 3.2 – *Histogramme des valeurs du signal biais (l'ordonnée correspond au logarithme naturel des valeurs mesurées) calculées dans des fenêtres de largeur* 1 *kpb sur les séquences des* 22 *chromosomes asexués de l'homme. (a) Le signal est calculé sur les séquences masquées, i.e. sans les séquences répétées. La courbe en trait plein est une parabole correspondant à une densité de probabilité gaussienne de même variance. (b) Le signal biais est calculé sur les séquences natives. On constate dans ce cas un écart important à la distribution gaussienne.*

suites de lettres identiques (spécialement A et T) induisent des valeurs du biais de grande amplitude. Lorsque les séquences répétées ne sont pas prises en compte, une inspection visuelle de la distribution obtenue dans la figure 3.2 (a) montre que les valeurs du signal biais ne s'écartent notablement de la distribution gaussienne que dans les queues de la distribution, *i.e.* pour les valeurs de module supérieur à $|S| = 0.4$, témoignant d'une dissymétrie conséquente entre les deux brins. La moyenne de la distribution associée au signal est nulle et la variance mesurée vaut $\sigma_S^2 = 0.014$. Ces résultats confirment *a posteriori* la pertinence du choix de ne pas considérer les séquences répétées, malgré la proportion importante de nucléotides qu'elles représentent.

Mise en évidence de la nature bifractale du signal biais aux petites échelles (inférieures à 40 kpb)

Puisque, pour les petites valeurs, les incréments semblent suivre une loi comparable à une loi gaussienne (*cf.* figure 3.2 (a)), il est naturel de se poser la question de l'existence de corrélations à longue portée dans la marche ADN associée au biais (figure 3.3). Comme nous allons le voir ici, l'analyse multifractale va confirmer la présence de corrélations à longue portée dans cette marche ADN, mais aussi révéler la présence de deux types de singularités. En plus des singularités d'exposant de Hölder $h = 0.78$, similaires aux singularités rencontrées dans les marches générées avec d'autres codages (*cf.* chapitre 2),

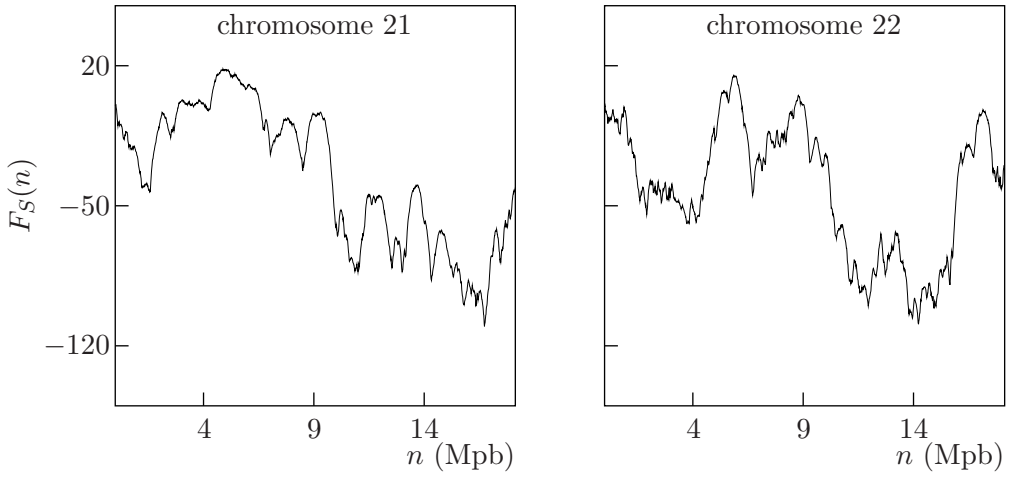

FIG. 3.3 – *Marches ADN associées au biais, calculées pour les chromosomes 21 et 22 de l'homme, dans des fenêtres de largeur* 1 *kpb, sans considérer les séquences répétées.*

il existe des singularités d'exposant $h = 1$ dans ces marches, qui attestent de la présence de sauts dans les incréments, c'est-à-dire dans le signal biais lui-même.

Lorsque l'on représente, pour des échelles a correspondant à des tailles inférieures à 40 kpb, les fonctions $h_a(q)$ (*cf.* égalité (2.127) de la première partie) obtenues par la méthode des maxima du module de la transformée en ondelettes appliquée à la marche ADN associée au biais des 22 chromosomes asexués de l'homme, en fonction du logarithme de l'échelle (figure 3.4 (a)), on constate pour certaines valeurs de q un comportement linéaire. Toutefois, le coefficient angulaire de ces fonctions dépend du paramètre q. Pour les valeurs de q inférieures à $-1/4$*, on obtient un coefficient proche de $h = 0.78$, alors que pour les valeurs de q plus grandes, $q > 1$, le coefficient angulaire est plutôt proche de la valeur $h = 1$. Pour ces deux types de singularité (d'exposant de Hölder $h = 0.78$ et $h = 1$), on trouve la dimension (*cf.* égalité (2.128) de la première partie) $d(0.78) = d(1) = 1$ (figure 3.4 (d)). Pour les valeurs de q intermédiaires, $-1/4 < q < 1$, le comportement linéaire de $\log_2 h_a(q)$ en fonction de $\log_2 a$ n'est plus aussi évident et l'estimation de l'exposant h par régression linéaire devient questionnable. Comme nous le discuterons par la suite, on a plutôt l'impression de voir, sur la gamme d'échelles considérée, une transition entre deux pentes de coefficient angulaire $h = 0.78$ (aux petites échelles) et $h = 1$ (aux grandes échelles) respectivement. Ces observations suggèrent que la marche ADN pourrait présenter un caractère bifractal [154], ce qui semble être corroboré par le comportement du spectre d'exposant $\eta(q)$ (*cf.* égalité (2.112) de la première partie) de la fonction de partition $Z(a,q)$ (figure 3.4 (c)) : la présence de deux types singularités, d'exposant $h = 0.78$ et $h = 1$, devrait se manifester par l'existence d'une transition

*. Puisque, dans ce chapitre, les mesures ne peuvent être effectuées que numériquement, il va de soi que toutes les valeurs données ici sont approximatives.

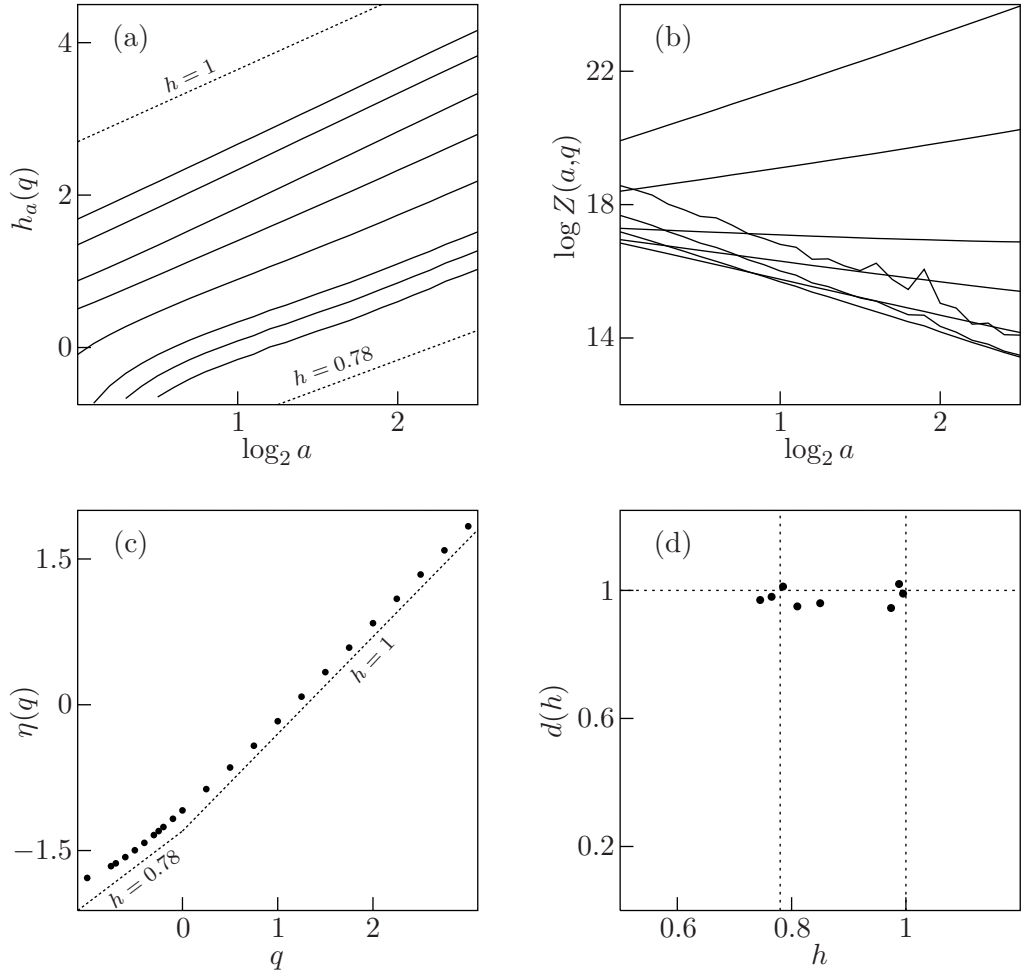

FIG. 3.4 – *Analyse multifractale, par la méthode des maxima du module de la transformée en ondelettes, de la marche associée au signal biais des* 22 *chromosomes asexués de l'homme aux échelles a inférieures à* 40 *kpb. L'ondelette mère utilisée est la dérivée première de la gaussienne. (a) Les fonctions* $h_a(q)$ *(cf. égalité (2.127) de la première partie) en fonction du logarithme de l'échelle, pour différentes valeurs de* q*. En pointillés sont représentées les droites correspondant aux exposants* $h = 0.78$ *(en bas) et* $h = 1$ *(en haut). (b) Le logarithme des fonctions de partition* $\log_2 Z(a,q)$ *(cf. équation (2.111) de la première partie) en fonction de* $\log_2 a$*. Dans (a) et (b), les courbes correspondent aux valeurs de* q *suivantes (de bas en haut) :* $q = -1, -0.75, -0.5, 0, 0.5, 1, 2$ *et* 3*. (c) La fonction* η *(cf. équation (2.112) de la première partie) présente deux comportements linéaires selon la gamme de valeurs de* q *considérée :* $\eta_1(q) = 0.78q - 1$ *et* $\eta_2 = q - 1$*. (d) Le spectre* $d(h)$ *des singularités se réduit à deux valeurs d'exposant :* $d(0.78) = 1$ *et* $d(1) = 1$ *(représentées par les deux droites verticales en pointillés).*

entre deux comportement linéaires de coefficients angulaires proche de $h = 0.78$ et $h = 1$ respectivement, à savoir $\eta_1(q) = 0.78q - 1$ et $\eta_2(q) = q - 1$.

En effet, si seulement deux types de singularités, d'exposant h_1 et h_2 respectivement, avec $h_1 < h_2$, sont présentes dans le signal, chacune devrait dominer le comportement des fonctions $h_a(q)$ pour $q > q_c$ et $q < q_c$ respectivement, où q_c est une valeur critique de transition. En mécanique statistique, ce phénomène s'appelle *transition de phase* [25, 70, 109, 292]. La fonction de partition $Z(a,q)$ (*cf.* équation (2.111) de la première partie) peut alors être décomposée en deux contributions de la façon suivante,

$$Z(a,q) = C_1(q)\, a^{h_1 q - 1} + C_2(q)\, a^{h_2 q - 1}, \tag{3.1}$$

où $C_1(q)$ et $C_2(q)$ sont des préfacteurs. Dans la mesure où $h_1 < h_2$, lorsque a tend vers zéro, la fonction de partition devrait se comporter comme $Z(q,a) \sim C_1(q)\, a^{qh_1 - 1}$ pour $q > 0$ et comme $Z(q,a) \sim C_2(q)\, a^{qh_2 - 1}$ pour $q < 0$, avec une transition entre ces deux comportements à la valeur critique $q_c = 0$. Les singularités les plus fortes, c'est-à-dire celles d'exposant h_1, domineraient donc dans le calcul de la fonction de partition pour les valeurs de q positives et les singularités les plus faibles domineraient aux valeurs de q négatives. Cependant, on constate dans les figures 3.4 (a) et (c) que ce sont les singularités d'exposant $h_2 = 1$ qui dominent pour les valeurs de q positives, alors qu'elles sont plus faibles que les singularités d'exposant $h_1 = 0.78$, qui elles dominent aux valeurs de q négatives. L'explication d'une telle inversion de comportement réside dans le fait que l'on ne peut numériquement pas accéder à des échelles a suffisamment petites, le biais étant calculé dans des fenêtres de largeur 1 kpb, pour pouvoir observer le comportement attendu. Les préfacteurs $C_1(q)$ et $C_2(q)$ dans l'égalité (3.1) caractérisent le nombre relatif de lignes de maxima dans le squelette de la transformée en ondelettes correspondant aux singularités $h_1 = 0.78$ et $h_2 = 1$ respectivement, mais aussi l'amplitude des maxima du module de la transformée le long de ces lignes. Ainsi, si sur la gamme d'échelles accessible à l'analyse, les maxima du module le long des lignes associées à h_2 sont beaucoup plus importants que ceux le long des lignes associées à h_1, cette inégalité va s'amplifier pour les puissances q positives, ce qui peut donner lieu à d'importantes différences entre $C_1(q)$ et $C_2(q)$ et en particulier une inégalité du type $C_2(q) \gg C_1(q)$, lorsque $q > 0$. Cette inégalité va contribuer à ce que le second terme dans la fonction de partition $Z(a,q)$ donnée par l'égalité (3.1) soit le terme dominant aux échelles a explorées lorsque q est positif. Pour les valeurs de q négatives, c'est l'effet inverse qui se produit et c'est le premier terme de l'égalité (3.1) qui domine. Le fait que les fortes valeurs du biais soient anormalement nombreuses par rapport à une distribution gaussienne (figure 3.2 (a)) peut parfaitement engendrer un tel phénomène. Pour le type de signal de biais que nous analysons, il y a tout lieu de penser que les points associés aux singularités du type $h = 1$ dans la marche

ADN (figure 3.3) correspondent à des « sauts[*] » dans le bruit associé, c'est-à-dire des zones où l'amplitude du signal varie fortement.

Pour asseoir notre démonstration numérique de la bifractalité [154] de la marche ADN associée au biais, nous avons procédé comme suit. Pour un signal autosimilaire d'exposant H, nous avons vu (*cf.* remarque 2.12) que la relation suivante devrait être vérifiée, $|WF(b,\lambda a)| \approx \lambda^H |WF(b,a)|$ ($\lambda > 0$). Ainsi, le comportement du module de la transformée en ondelettes à l'échelle 1, $|WF(b,1)|$, devrait aussi être celui de la transformée en ondelettes à une échelle a quelconque, moyennant une renormalisation par le facteur a^H, $|WF(b,a)|/a^H$. Plus précisément, l'histogramme des valeurs du module de la transformée en ondelettes à l'échelle 1 devrait être identique à celui des valeurs du module de la transformée à l'échelle a, renormalisée par a^H. Symboliquement, si N désigne l'histogramme renormalisé par le nombre d'éléments,

$$N(|WF(b,1)|) = N(|WF(b,a)|/a^H). \tag{3.2}$$

Par contre, si le signal est multifractal, *i.e.* s'il n'existe pas qu'un seul exposant d'invariance d'échelle H, il est impossible de trouver un facteur du type a^H permettant de superposer les histogrammes des modules des valeurs de la transformée en ondelettes en fonction de l'échelle. Toutefois, pour un signal bifractal, une première partie des histogrammes devrait se superposer avec un facteur de renormalisation a^{h_1} et une seconde avec un facteur a^{h_2}. La figure 3.5 montre clairement que tel est bien le cas pour la marche ADN associée au biais : aux petites valeurs des coefficients en ondelettes, ces histogrammes se superposent relativement bien avec le facteur de renormalisation $a^{h_1=0.78}$ (figure 3.5 (a)), alors que pour les grandes valeurs, ils se superposent beaucoup mieux avec le facteur $a^{h_2=1}$. Ainsi, tout laisse à penser que les marches ADN construites à l'aide du codage biais sont bifractales, dans le sens où aux singularités d'exposant $h_1 = 0.78$ caractéristiques de marches browniennes fractionnaires s'ajoutent des singularités d'exposant $h_2 = 1$ correspondant à des sauts de grande amplitude dans les incréments de la marche.

Étude du signal biais aux grandes échelles (supérieures à 40 kpb)

Aux échelles plus grandes (à partir de 50 kpb), la marche ADN associée au biais ne présente plus de caractère bifractal : on n'observe plus de comportement invariant d'échelle d'exposant $h = 0.78$, signature de l'existence de corrélations à longue portée dans le signal biais S. Par contre, tout comme avec les autres codages étudiés dans le chapitre 2, on

[*]. Bien sûr, on ne peut pas à proprement parler de saut, le signal étant discret.

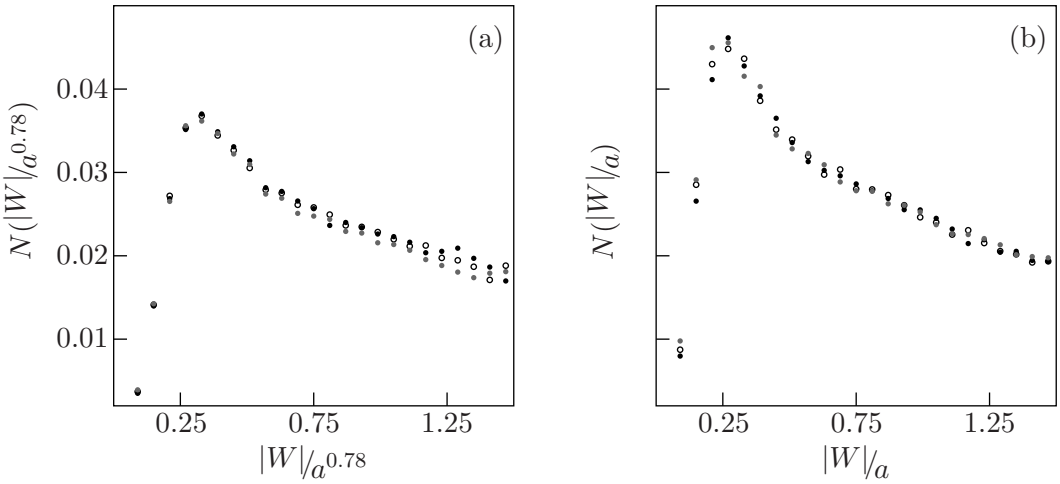

FIG. 3.5 – *Histogrammes des valeurs des maxima du module de la transformée en ondelettes de la marche ADN associée au biais calculée sur les* 22 *chromosomes asexués de l'homme, aux petites échelles* $a = 10$ *kpb* (•), 20 *kpb* (○) *et* 40 *kpb* (•). *(a) Les maxima sont renormalisés par* $a^{0.78}$. *On constate une superposition des histogrammes aux petites valeurs de* $|W|a^{-0.78}$ *et une nette séparation aux plus grandes valeurs. (b) Les maxima sont renormalisés par* a. *Les histogrammes se superposent maintenant aux grandes valeurs de* $|W|a^{-1}$ *et diffèrent aux petites valeurs. L'ondelette mère utilisée est la dérivée première de la gaussienne.*

détecte la présence d'oscillations de relaxation de basse fréquence, suggérant l'existence de rythmes.

Comme cela est illustré dans la figure 3.6, aux échelles $a > 50$ kpb, les fonctions $h_a(q)$ associées à la marche ADN du biais représentées en fonction de $\log_2 a$ possèdent toutes le même coefficient angulaire $h = 1$, et ce quelque soit la valeur de q. En fait, à des échelles a suffisamment grandes, les singularités d'exposant $h_1 = 0.78$ sont lissées par l'ondelette et seules les singularités de type $h_2 = 1$ restent visibles. Les points associés aux singularités d'exposant $h = 1$ représentent les positions où le bruit associé, c'est-à-dire le signal biais, varie brusquement sur une distance de l'ordre de l'échelle a (d'après la relation (2.14), les points associés aux singularités d'exposant $h = 1$ dans la marche sont les points associés aux singularités d'exposant $h = 0$ dans le bruit, que l'on peut interpréter comme des discontinuités). Ainsi, aux grandes échelles, les sauts du signal biais sont les seules singularités détectées par la transformée en ondelettes, qui comme nous allons le voir révèle plutôt la présence d'oscillations non linéaires de basse fréquence dont le caractère relaxationnel est relié à la présence de ces sauts [297, 299].

Pour compléter cette étude de la marche ADN et du bruit associé, à savoir le biais, nous allons nous interroger sur l'organisation à grande échelle de ces signaux. En effet, le signal biais semble présenter des oscillations de basse fréquence, comme l'illustre la figure

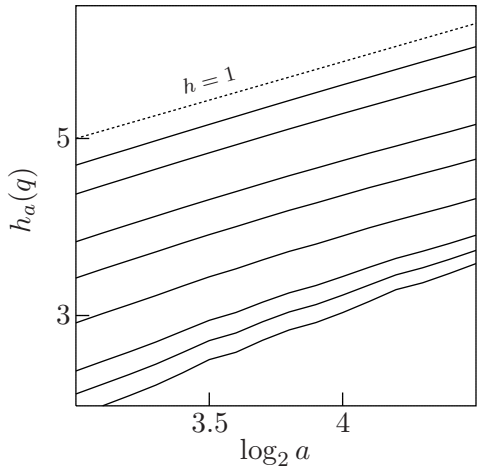

FIG. 3.6 – *Représentations des fonctions* $h_a(q)$ *en fonction de* $\log_2 a$ *obtenues en appliquant la méthode des maxima du module de la transformée en ondelettes à la marche associée au signal biais des* 22 *chromosomes asexués de l'homme. Les valeurs de* q *représentées sont, de bas en haut,* $q = -1$, -0.75, -0.5, 0, 0.5, 1, 2 *et* 3. *La gamme d'échelles représentée correspond à des tailles comprises entre* 50 *et* 150 *kpb. On constate le caractère quasiment linéaire des différentes courbes. Le coefficient angulaire est constant et approximativement égal à* $h = 1$ *quelque soit* q. *En pointillés est représentée une droite de pente* $h = 1$ *pour comparaison.*

3.1 (a). Pour quantifier la présence de telles oscillations, nous avons procédé comme dans le chapitre précédent, en calculant le spectre d'échelle $\Lambda(a)$ associé au signal biais (*cf.* égalité (2.13)), calculé pour des fenêtres de largeur 1 kpb. Le signal $\Lambda(a)$ est représenté en coordonnées logarithmiques dans la figure 3.7 (a). Sans surprise, le spectre semble d'abord décroître en loi de puissance, comportement caractéristique d'un bruit gaussien fractionnaire d'exposant $h = 0.78 - 1 = -0.22$. Peu après 60 kpb, on peut constater l'apparition de plusieurs bosses centrées autour de 400 et 600 kpb environ, ceci étant particulièrement évident dans la figure 3.7 (b), où $\Lambda(a)$ est cette fois représenté en fonction de a et non plus en coordonnées logarithmiques. Il semble donc exister des fréquences caractéristiques, voire des périodicités locales, dans le signal biais, suggérant la présence de rythmes [297, 299]. Ici, le codage utilisé, à savoir le biais, étant de nature fonctionnelle, on peut imaginer que ces rythmes reflètent simplement l'organisation des réplicons observée chez les vertébrés à sang chaud [55, 197, 244, 253, 396]. En effet, si la taille d'un réplicon est variable (de 100 kpb jusqu'à plusieurs Mpb), la taille moyenne d'un réplicon chez les mammifères a été estimée à 500 kpb [55] ; cette taille est en surprenant accord avec les périodicités caractéristiques relevées dans la figure 3.7. On peut en outre montrer qu'il existe un faisceau de présomptions en faveur d'un modèle chaotique donnant lieu à une telle organisation [297, 299].

La présente étude permet donc de tirer plusieurs conclusions. Aux petites échelles, la

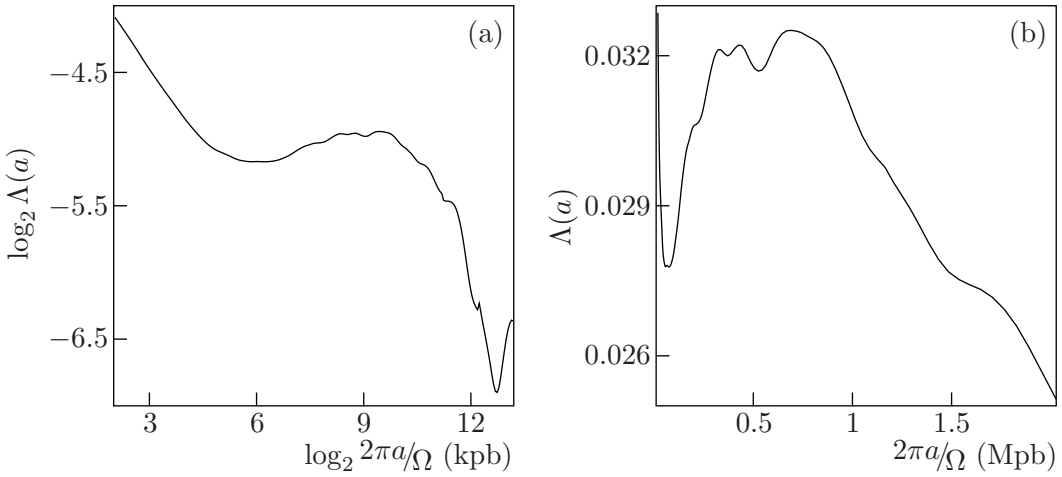

FIG. 3.7 – *Spectre d'échelles* $\Lambda(a)$ *calculé sur le signal biais évalué sur les* 22 *chromosomes asexués de l'homme.(a) Représentation en échelles logarithmiques. (b) Représentation en échelles linéaires.*

marche ADN associée au signal biais comporte une « composante » brownien fractionnaire. Les corrélations à longue portée associées à ce mouvement correspondent à un exposant d'auto-similarité $h_1 \approx 0.78$. Cependant, les fortes valeurs des incréments s'éloignent considérablement d'une distribution gaussienne et on a pu mettre en évidence la présence d'un deuxième type de singularités d'exposant $h_2 \approx 1$ dans la marche ADN associée, qui présente ainsi des propriétés de bifractalité. Autrement dit, les incréments, correspondant au signal biais, se comportent comme un bruit gaussien fractionnaire auquel viennent s'ajouter de nombreux sauts. Aux grandes échelles, ce sont ces sauts qui dominent l'analyse multifractale. En fait, comme pour les précédents codages étudiés dans le chapitre 2, nous avons réussi à révéler la présence de rythmes de basse fréquence dans le signal biais. Il est important de remarquer que le caractère fortement relaxationnel des oscillations non linéaires observées est étroitement relié à la présence de sauts détectés grâce à la transformée en ondelettes. Dans les chapitres qui suivent, nous allons approfondir l'étude du biais S en se focalisant plus particulièrement sur la détection des sauts qui, comme nous le verrons, vont nous apprendre beaucoup sur la position des origines de réplication, des gènes et de façon plus générale, sur l'existence d'un biais de réplication et d'un biais de transcription dans les génomes des mammifères [79, 369].

3.3 Mise en évidence d'une dissymétrie entre sauts ascendants et sauts descendants à grande échelle dans le signal biais

La présence de sauts dans le signal biais, révélée par notre analyse en ondelettes dans la section précédente, constitue une caractéristique majeure de ce signal. Dans la perspective d'identifier les mécanismes fonctionnels pouvant induire de tels sauts, nous allons étudier la manière dont ceux-ci sont distribués le long des chromosomes. Cette démarche nous conduira à mettre à jour l'existence d'une dissymétrie dans la répartition des sauts ascendants et des sauts descendants visibles à grande échelle : le nombre de sauts ascendants de grande amplitude est bien plus important que le nombre de sauts descendants d'amplitude comparable [79, 369].

C'est la présence de sauts de grande amplitude qui différencie le signal biais des autres bruits ADN obtenus par les codages présentés au chapitre 2. Si l'étude multifractale du signal biais chez l'homme, réalisée dans la section précédente, montre l'importance de ces sauts et leur omniprésence à grande échelle, il est important d'en caractériser leur répartition suivant leur amplitude, ainsi qu'en fonction de l'échelle d'analyse. Pour obtenir plus d'informations à ce sujet, nous avons procédé comme suit. Le signal biais S a d'abord été lissé pour obtenir le signal $\tilde{S}$ suivant :

$$\tilde{S}(n) = \frac{1}{20} S * \chi_{[-10,10]}(n), \tag{3.3}$$

qui permet de s'affranchir le plus possible de la composante bruit présente dans le signal, sans toutefois dénaturer l'amplitude des sauts. L'amplitude du saut $\Delta S(n)$ en un point n du signal est définie par l'égalité suivante,

$$\Delta S(n) = \tilde{S}(n + 20) - \tilde{S}(n - 20). \tag{3.4}$$

Nous ne prenons donc pas en compte les points situés à une distance inférieure à 10 kpb du point n considéré. On évite de cette manière de biaiser la mesure en considérant les points situés à la transition du saut, où une variabilité plus importante des valeurs de S est observée. Précisons que les résultats restent qualitativement inchangés lorsque moins de précautions sont prises [79, 369].

Pour détecter les sauts visibles à une échelle donnée, l'ondelette mère dérivée première de la gaussienne est tout à fait appropriée. En effet, la transformée en ondelettes d'un signal avec cette ondelette peut être vue comme la dérivée du signal lissé par une gaussienne dont la variance dépend de l'échelle (*cf.* égalité (2.21) de la première partie). À une échelle d'analyse donnée, les positions correspondant aux grandes valeurs de la transformée en

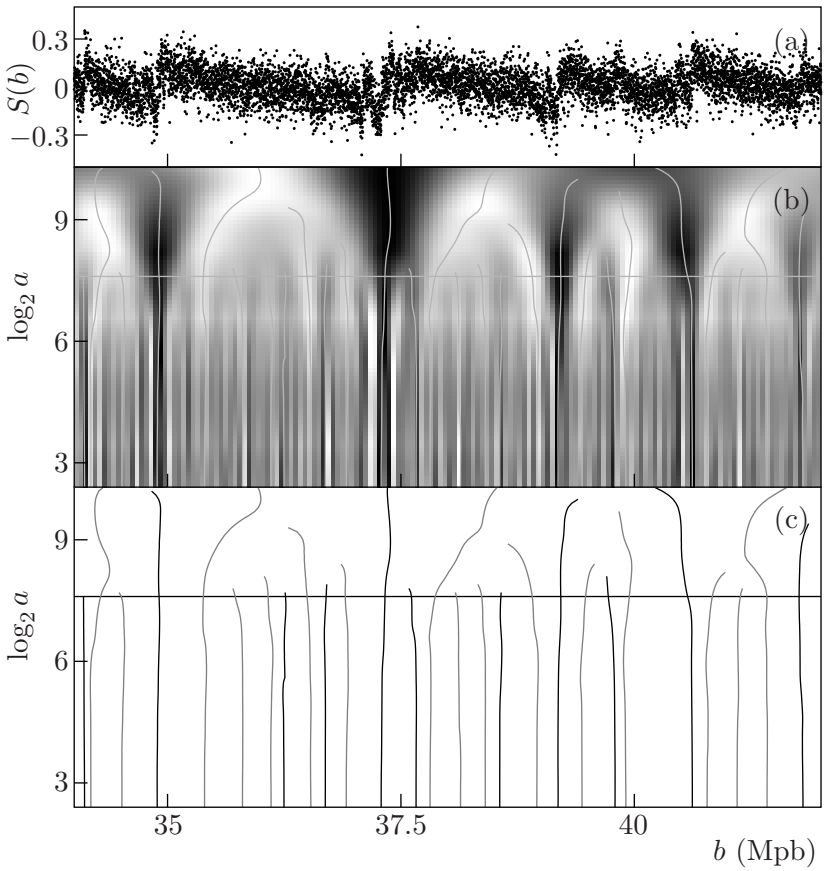

FIG. 3.8 – *Illustration de la méthode utilisée pour détecter les sauts ascendants de grande amplitude dans le signal biais S du chromosome 12 de l'homme. (a) Signal biais évalué dans des fenêtres de largeur* 1 *kpb le long d'un fragment de longueur* 10 *Mpb. (b) Représentation espace-échelle donnée par la transformée en ondelettes calculée avec l'ondelette mère dérivée première de la gaussienne. Les niveaux de gris utilisés vont du blanc (pour les valeurs les plus faibles) au noir (pour les valeurs les plus fortes). (c) Sélection des lignes de maxima du module de la transformée en ondelettes correspondant à un saut visible à l'échelle caractéristique de* 200 *kpb, matérialisée par un trait noir horizontal ; les lignes correspondant à un saut descendant ($\Delta S < 0$) sont représentées en gris et celles correspondant à un saut ascendant ($\Delta S > 0$) en noir. La position pointée par une ligne de maxima à petite échelle donne la position du saut qui lui est associée.*

ondelettes peuvent donc être associées à de fortes variations du signal lissé à cette échelle ; les valeurs positives correspondent à des sauts ascendants et les valeurs négatives à des sauts descendants.

La répartition des sauts à grande échelle (200 kpb) du signal biais a été étudiée à partir du squelette de la transformée en ondelettes de la façon suivante. Pour chaque chromosome asexué de l'homme, la transformé en ondelettes du signal biais S (calculé dans des fenêtres de largeur 1 kpb) a été effectuée, puis les lignes de maxima du module de la transformée en ondelettes ont été déterminées. Pour étudier les sauts visibles à grande échelle, seules les lignes de maxima se prolongeant aux échelles supérieures à la taille 200 kpb ont été retenues. Les positions des sauts dans le signal analysé sont alors déterminées par les positions pointées par les lignes de maxima ainsi sélectionnées aux plus petites échelles. En pratique, ce n'est pas la plus petite échelle numériquement accessible que nous avons utilisée pour préciser la position des sauts, mais celle correspondant à 25 kpb ; en effet, le bruit étant omniprésent aux échelles inférieures, il ne sert à rien de descendre plus bas le long des lignes de maxima, cela n'améliorant pas de façon significative la précision des positions détectées. Cette méthode de détection des sauts est illustrée dans la figure 3.8

En appliquant cette méthodologie, 2415 sauts ascendants ($\Delta S > 0$) et 2686 sauts descendants ($\Delta S < 0$) ont été détectés dans le génome humain. La figure 3.9 montre que les amplitudes des sauts ascendants n'ont pas la même distribution statistique que celles des sauts descendants. En effet, le nombre de sauts ascendant de grande amplitude est beaucoup plus important que le nombre de sauts descendants d'amplitude comparable, cette tendance s'inversant pour les sauts de faible amplitude. En particulier, le rapport de ces deux nombres peut atteindre des valeurs supérieures à 5 pour des amplitudes $|\Delta S| \geqslant 0.2$ (figure 3.9 (b)).

Pour savoir si cette asymétrie dans la répartition de l'amplitude des sauts est présente à toutes les échelles, nous avons effectué la même étude en sélectionnant cette fois les lignes de maxima se prolongeant aux échelles $a \geqslant 20$ kpb et non plus 200 kpb. Comme rapporté dans la figure 3.10, avec un tel seuil, 53789 sauts ascendants et 57879 sauts descendants ont été obtenus. L'amplitude des sauts est toujours distribuée de façon dissymétrique, mais cette différence entre sauts ascendants et sauts descendants est désormais beaucoup moins marquée : le rapport entre le nombre de sauts ascendants et de sauts descendants d'amplitude $|\Delta S| \geqslant 0.2$ n'est plus que de 2 dans la figure 3.10 (b). Cette tendance se confirme lorsque l'on considère des échelles encore plus petites. La forte dissymétrie observée dans la distribution des amplitudes des sauts ascendants et descendants semble donc être essentiellement due aux sauts ascendants de forte amplitude visibles à grande échelle, qui n'ont quasiment pas d'équivalent descendant.

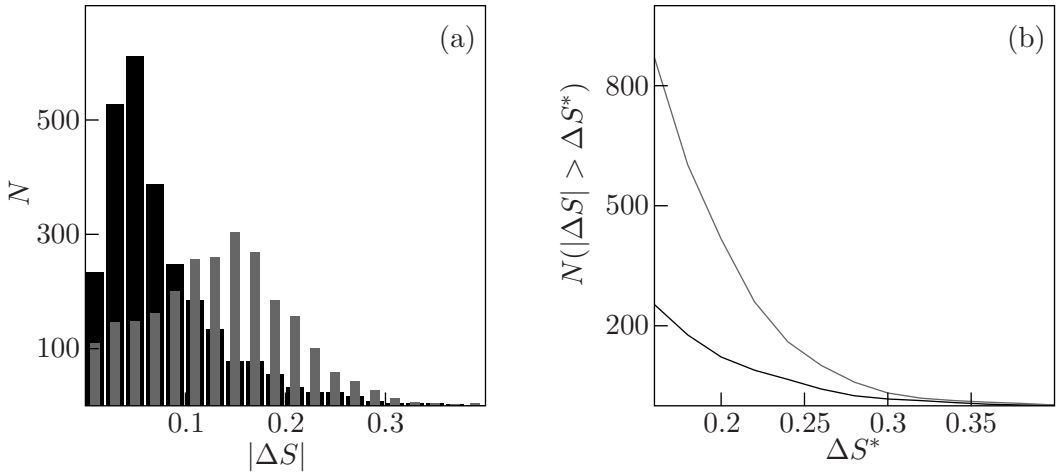

FIG. 3.9 – *Analyse statistique de l'amplitude $|\Delta S|$ des sauts pointés par une ligne de maxima du module se prolongeant à des échelles $a \geqslant 200$ kpb dans le squelette de la transformée en ondelettes du signal biais des 22 chromosomes asexués du génome humain. (a) Les histogrammes montrent clairement qu'il existe une dissymétrie dans la distribution des amplitudes des sauts ascendants (gris) et des sauts descendants (noir). (b) Ce résultat est plus clair lorsque l'on représente le nombre de sauts dont l'amplitude $|\Delta S|$ est supérieure à un seuil ΔS^* donné en fonction de ce seuil ; la courbe noire (resp. grise) correspond aux sauts descendants (resp. ascendants).*

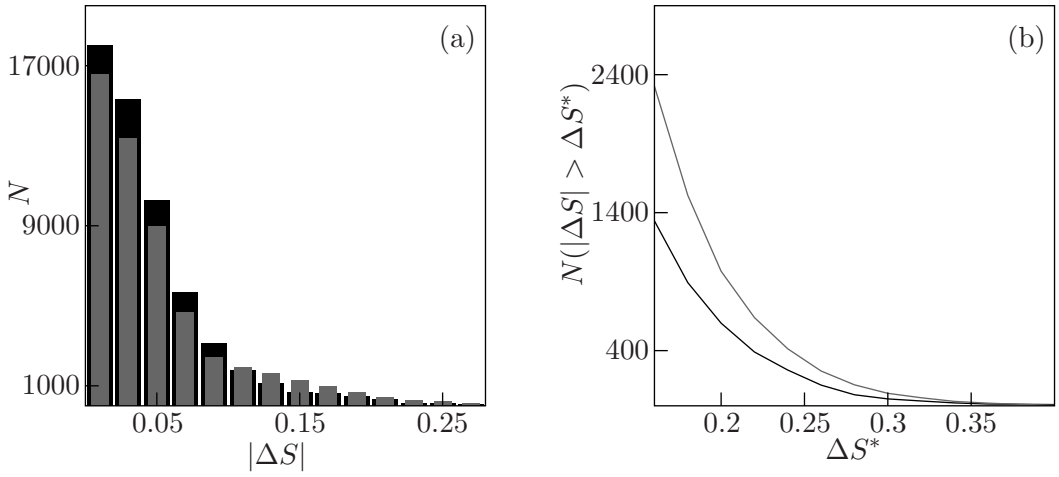

FIG. 3.10 – *Analyse statistique de l'amplitude $|\Delta S|$ des sauts pointés par une ligne de maxima du module se prolongeant à des échelles $a \geqslant 20$ kpb dans le squelette de la transformée en ondelettes du signal biais des 22 chromosomes asexués du génome humain. (a) Les histogrammes montrent clairement qu'il existe une dissymétrie dans la distribution des amplitudes des sauts ascendants (gris) et des sauts descendants (noir). (b) Ce résultat est plus clair lorsque l'on représente le nombre de sauts dont l'amplitude $|\Delta S|$ est supérieure à un seuil ΔS^* donné en fonction de ce seuil ; la courbe noire (resp. grise) correspond aux sauts descendants (resp. ascendants).*

Nous avons défini une méthodologie de détection de sauts présents dans le signal biais de l'homme et montré qu'à grande échelle, la distribution des amplitudes varie selon que l'on considère les sauts ascendants ou descendants : il existe beaucoup plus de sauts ascendants de grande amplitude que de sauts descendants. Cette tendance s'estompe pour les plus petites échelles. Comme nous allons le voir par la suite, ces observations seront d'une grande importance tant au niveau des implications biologiques sur les mécanismes potentiels ayant pu engendrer les sauts observés, qu'au niveau méthodologique pour la mise en oeuvre d'une méthode multi-échelle de prédiction des origines de réplication chez l'homme et plus généralement chez les mammifères.

Chapitre 4

Mise en évidence d'un biais de transcription et d'un biais de réplication dans les séquences d'ADN chez les mammifères

Les déséquilibres entre les concentrations de nucléotides A et T d'une part et C et G de l'autre observés dans le génome humain sont en tout ou en partie la trace laissée au cours de l'évolution par les processus de réplication et de transcription. Nous avons vu que ces mécanismes ont pour caractéristique de briser la symétrie naturellement présente entre les deux brins d'une molécule d'ADN. À cet égard, on peut donc s'interroger sur l'éventuelle relation entre ces mécanismes et l'une des propriétés caractéristiques du codage biais que nous avons mise en évidence dans le chapitre 3, à savoir la présence de nombreux sauts de plus ou moins forte amplitude. Chez les procaryotes et les génomes viraux, les études concernant le biais de composition sont nombreuses [152, 171, 220, 248, 250, 251, 254, 255, 273, 288, 326, 327, 328, 335, 363] ; En particulier, il a été utilisé ainsi que les marches ADN associées pour détecter les origines de réplication dans certains génomes [150, 152, 171, 251, 288, 328, 335, 363]. Chez les eucaryotes, l'étude de cette asymétrie n'a pas vraiment donné de résultat clair, malgré certaines évidences expérimentales [128,

161, 170, 344, 355]. Dans ce chapitre, la recherche de l'origine des sauts observés dans le signal biais (*cf.* chapitre 3) va nous conduire à mettre en évidence l'existence d'un biais lié à la transcription et d'un biais lié à la réplication chez les mammifères. Le premier de ces biais va nous permettre d'interpréter les sauts ascendants et descendants d'amplitude comparable détectés à petite échelle. La réplication par contre fournira une explication de la dissymétrie entre le nombre de sauts ascendants et le nombre de sauts descendants d'amplitude comparable observée à plus grande échelle.

Ces études ont donné lieu à plusieurs publications, la première concernant le rôle de la transcription [368] et deux autres traitant du rôle de la réplication [79, 369].

4.1 Étude du biais de composition chez l'homme lié à la transcription

Dans cette section, nous allons montrer l'existence d'un biais mutationnel lié à la transcription dans le génome de l'homme [368]. Ce biais se traduit par la présence de profils en forme de « créneau » dans le signal biais, la largeur du créneau dépendant de la taille du gène et sa hauteur reflétant l'amplitude du biais. Chaque créneau étant bordé par un saut ascendant et un saut descendant, la transcription sera associée à l'existence d'un même nombre de sauts ascendants et descendants et ne permettra pas d'expliquer la prédominance des sauts ascendants de grande amplitude observée dans le chapitre 3.

Influence de la transcription sur le signal biais

Pour déterminer si le mécanisme de transcription joue un rôle dans l'observation d'un biais de composition chez l'homme, nous allons comparer le signal biais sur des séquences relatives à des introns (les exons étant soumis à une pression de sélection toute particulière) avec le signal biais sur des séquences intergéniques. Les différences significatives observées permettront de mettre en évidence la présence d'un « biais transcriptionnel » [367, 368].

Pour réaliser cette étude, les séquences d'introns d'un même gène ont été concaténées en une seule séquence. Pour éviter de prendre en compte le biais induit par les mécanismes d'épissage[+], nous avons éliminé 560 kpb aux deux extrémités de chaque intron. Les séquences répétées ont aussi été éliminées[⁕]. Le nombre de gènes[⁂] pris en compte est

+. L'épissage consiste en l'excision des introns et le « raboutage » des exons, pour permettre à l'ARNm mature d'être traduit en protéine.

⁕. Pour ce faire, le logiciel *RepeatMasker* a été utilisé [349].

⁂. La banque utilisée est *RefGene* (avril 2003).

Sens	Seq. rép.	$\overline{S_{TA}}$	$\overline{S_{GC}}$
sens	avec	$0.0472 \pm 7\,10^{-4}$	$0.0297 \pm 7\,10^{-4}$
sens	sans	$0.0452 \pm 7\,10^{-4}$	$0.0369 \pm 7\,10^{-4}$
anti-sens	avec	$-0.0456 \pm 7\,10^{-4}$	$-0.0305 \pm 7\,10^{-4}$
anti-sens	sans	$-0.0425 \pm 7\,10^{-4}$	$-0.0363 \pm 7\,10^{-4}$
unique	avec	$0.0464 \pm 5\,10^{-4}$	$0.0301 \pm 5\,10^{-4}$
unique	sans	$0.0439 \pm 5\,10^{-4}$	$0.0366 \pm 7\,10^{-4}$

TAB. 4.1 – *Valeurs moyennes des biais $\overline{S_{TA}}$ et $\overline{S_{GC}}$, calculées sur 12469 gènes introniques de l'homme. Pour les quatre premières lignes, la distinction est faite entre les gènes sens (sens +,* 6312 *gènes) et anti-sens (sens −,* 6157 *gènes). Pour les deux dernières, tous les gènes sont replacés dans le même sens.*

de 12469. La première observation est que lorsque le gène a la même orientation que le brin Watson, c'est-à-dire quand il est transcrit dans le même sens que le brin (sens +), les biais S_{TA} (égalité (2.10)) et en S_{GC} (égalité (2.11)) sont positifs (table 4.1), alors que si le gène a l'orientation contraire (anti-sens ou sens −), les valeurs obtenues sont négatives et quasiment opposées (table 4.1 et figure 4.1). Si l'on examine les signaux obtenus avec les séquences répétées, les biais en GC et en TA ne sont que peu affectés. Lorsque l'on replace tous les gènes dans le sens de la séquence, la moyenne sur toutes les séquences révèle un excès de T par rapport à A, $\overline{S_{TA}} = 0.0439 \pm 0.0005$ et un excès de G par rapport à C, $\overline{S_{GC}} = 0.0366 \pm 0.0007$.

Si les deux types de biais observés (en TA et en GC) résultent du même mécanisme, ils devraient être corrélés ; pourtant, cette corrélation est très faible ($r = 0.09$) lorsque tous les gènes sont considérés. En fait, les biais observés sur les petits gènes sont fortement bruités. Lorsque l'on ne considère plus que les grands gènes, la valeur du coefficient de corrélation augmente significativement ; on obtient $r = 0.61$ lorsque la longueur du gène (introns concaténés) est supérieure à 25 kpb (table 4.2). Notons aussi que les biais présentent une faible corrélation avec le pourcentage en GC (inférieure à 10^{-1}) et qu'il en va de même pour la corrélation entre les biais et la longueur des gènes. Cela n'est pas étonnant, puisque le niveau d'expression des gènes ne semble pas être corrélé avec la taille des gènes [392] ni avec le pourcentage en GC [342].

Afin d'évaluer le nombre de gènes présentant un biais statistiquement significatif, nous avons comparé les biais mesurés dans les régions introniques à ceux attendus pour des séquences aléatoires de même longueur L et de même pourcentage en GC. Si c désigne ce pourcentage, estimer la probabilité P pour qu'une séquence de longueur L possède une valeur du biais en GC supérieure à s revient à évaluer la probabilité pour que le nombre de nucléotides G présents dans la séquence soit supérieur à $\lfloor cL(s+1)/2 \rfloor$. À titre d'exemple,

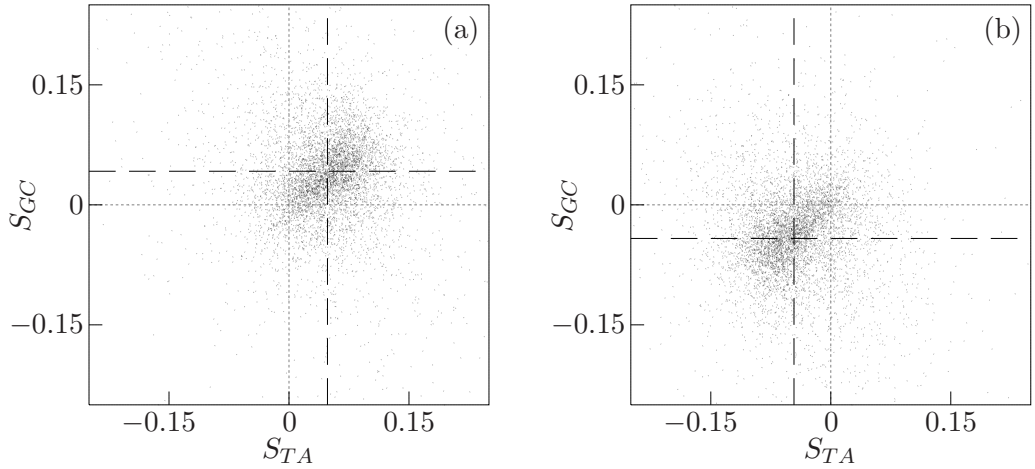

FIG. 4.1 – *Asymétries de composition en TA et en GC dans les régions introniques des gènes humains. Chaque point correspond à un des* 12469 *gènes analysés ; les séquences répétées et les* 560 *kpb localisées à chaque extrémités d'introns ont été éliminées. (a) Gènes orientés dans le sens du brin Watson (sens +). (b) Gènes orientés dans le sens du brin Crick (sens −). On trouve $\overline{S_{TA}} = 0.0452$, $\overline{S_{GC}} = 0.0369$ et $\overline{S_{TA}} = -0.0425$, $S_{GC} = -0.0363$ respectivement dans (a) et (b). Les lignes en pointillés noirs représentent les biais moyens mesurés et celles en gris représentent les droites horizontales et verticales s'intersectant à l'origine.*

Long. min.	Corr. entre les biais	% gènes biaisés
0	0.09	64
10	0.45	82
25	0.61	86

TAB. 4.2 – *Coefficient de corrélation (deuxième colonne) entre les biais S_{TA} et S_{GC} calculé sur les gènes (sans séquence répétée) dont les introns concaténés ont une longueur minimum (en kpb), et pourcentage de ces gènes (troisième colonne) présentant un biais significatif en TA et GC ($P < 10^{-2}$). Le coefficient de corrélation et le pourcentage sont forts lorsque les petits gènes ne sont pas pris en compte. Pour la longueur minimum de* 25 *kpb, la population est de* 4185 *gènes.*

si $c = 0.55$ et $L = 10000$, la probabilité pour que le biais en GC soit supérieur à 0.032 est inférieure à 10^{-2}. Lorsque les plus grands gènes sont seuls pris en compte, le pourcentage de gènes dont les biais en GC et TA sont significatifs ($P < 10^{-2}$) est largement supérieur à 80% (table 4.2).

La comparaison dans la figure 4.2 des valeurs des biais S_{TA} et S_{GC} dans les régions transcrites à celles obtenues dans les régions intergéniques voisines met en évidence le rôle de la transcription dans les asymétries de composition observées chez l'homme : la valeur moyenne des biais en TA et GC le long du génome, calculé dans des fenêtres de largeur 1 kpb, montre clairement que, juste avant le gène, *i.e.* à l'extrémité $5'$ de ce gène, il existe une variation brusque et croissante des biais depuis zéro dans la région intergénique jusqu'à des valeurs positives de l'ordre de 0.04 à 0.06 pour $\overline{S_{TA}}$ et 0.03 à 0.05 pour $\overline{S_{GC}}$ (figure 4.2). De même, à l'extrémité $3'$ du gène, on observe un saut décroissant du biais vers une valeur nulle dans l'intergénique (le fait que ce saut décroissant soit plus ou moins « arrondi » traduit la nature relativement mal définie de la terminaison de la réplication qui ne se termine pas forcément à un site donné mais peut « baver »). Des biais de composition sont donc spécifiquement observés dans les régions transcrites, indiquant qu'ils résultent bien de processus liés à la transcription dans les cellules de lignée germinale [367, 368].

Évaluation des taux de substitution pouvant engendrer un biais transcriptionnel

Puisque la transcription induit des asymétries de composition notables, certaines mutations doivent donc affecter de façon différente les séquences géniques et les séquences intergéniques. Nous allons maintenant essayer de quantifier ces différences.

Pour qu'un biais transcriptionnel puisse s'établir, des mutations doivent se produire à des fréquences différentes sur le brin codant et le brin non-codant. Il faut, par exemple, que la substitution $T \to C$ et son complémentaire $A \to G$, se produisent à un taux plus élevé sur un des deux brins. Pour déterminer les types de substitution pouvant générer les asymétries mesurées, il faut connaître les taux des 12 substitutions possibles entre nucléotides, appelé *matrice de substitution*. Dans une étude portant sur quelques gènes humain, Green *et al.* [170] ont mis en évidence l'existence de taux de transition différents selon que l'on considère une séquence (non-codante) transcrite ou une région intergénique. Ce travail révèle en particulier la présence d'un excès de substitutions $A \to G$ par rapport aux substitutions $T \to C$ sur le brin codant, ce qui entraîne un excès de nucléotides G et T sur le brin codant de la plupart des gènes. Ce résultat est en accord avec les observations

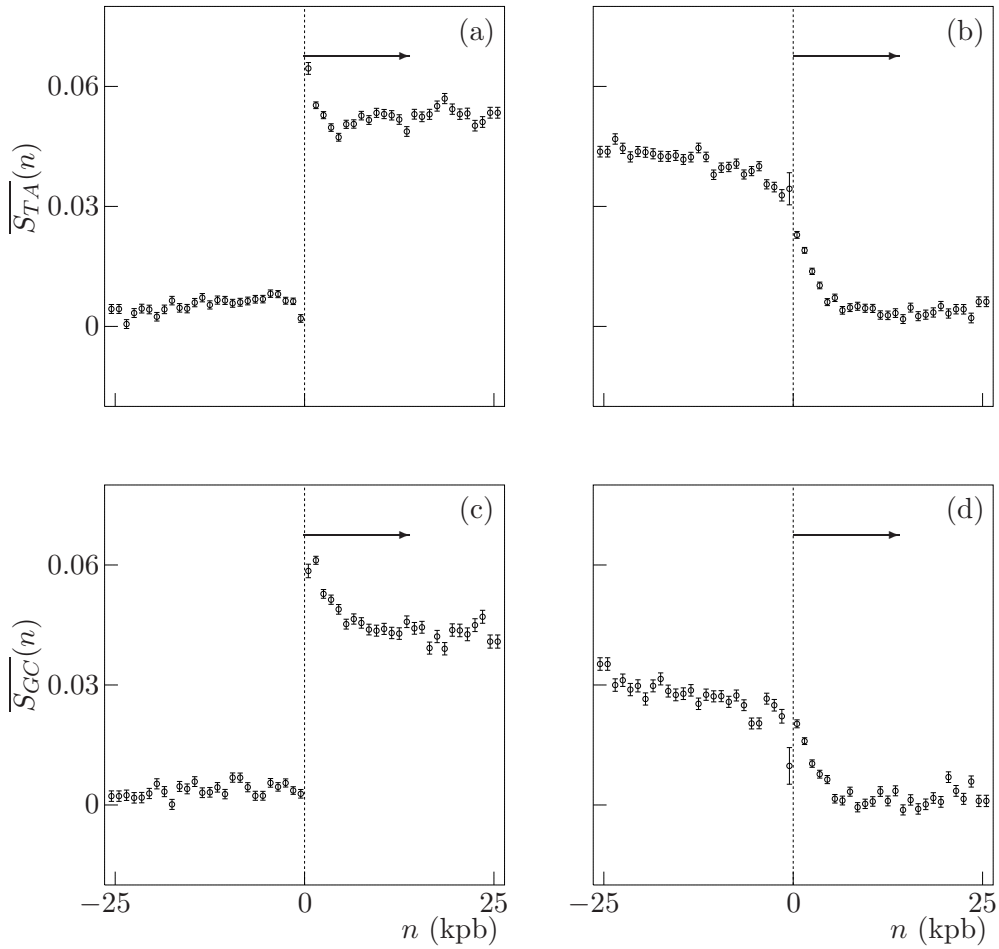

FIG. 4.2 – *Les biais moyen en TA ((a) et (b)) et GC ((c) et (d)) au voisinage des extrémités* $5'$ *((a) et (c)) et* $3'$ *((b) et (d)) des gènes. En abscisse est représentée la distance (en kpb) à l'extrémité du gène, zéro correspondant à l'extrémité* $5'$ *pour les panneaux de gauche et à l'extrémité* $3'$ *pour ceux de droite. En ordonnée est reporté* $\overline{S_{TA}}$ *pour les panneaux du haut et* $\overline{S_{GC}}$ *pour ceux du bas, calculé dans des fenêtres d'une largeur de* 1 *kpb. On constate clairement qu'il existe une transition entre les régions intergéniques (où* $\overline{S_{TA}}$ *et* $\overline{S_{GC}}$ *sont proches de zéro) et géniques (où* $\overline{S_{TA}}$ *et* $\overline{S_{GC}}$ *sont positifs). Les flèches correspondent au sens de la transcription des gènes.*

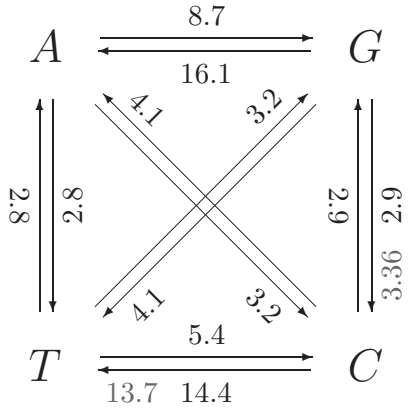

FIG. 4.3 – *Taux de substitution utilisés pour le calcul de la composition en nucléotides à l'équilibre. En noir les valeurs (en pourcentage) correspondant aux taux calculés dans la référence [138], qui ont été modifiées selon leur valeurs estimées dans les régions transcrites [170]. Pour obtenir des valeurs de biais S_{TA} et S_{GC} compatibles avec les valeurs moyennes obtenues sur les introns humains dans la table 4.1, les taux de substitution $G \to C$ et $C \to T$ ont été modifiés pour obtenir les valeurs en gris.*

ici obtenues (table 4.1).

Pour vérifier si ces transitions peuvent expliquer les asymétries de composition observées, nous avons calculé les compositions nucléotidiques à l'équilibre, en utilisant les taux de substitutions d'une étude antérieure [138] donnés dans la figure 4.3, mais modifiés pour être en accord avec les observations de GREEN *et al.* [170]. Un biais en TA comparable aux valeurs calculées a été ainsi obtenu : $S_{TA} = 0.047$, alors que la valeur du biais en GC s'avère être bien supérieure à celle observée : $S_{GC} = 0.078$. L'étude de GREEN *et al.* porte sur de petits segments (1.5 Mpb) à l'échelle du génome et ne reflète peut être que partiellement la réalité. Pour atteindre des valeurs du biais comparables à celles trouvées dans la table 4.1, il nous est apparu nécessaire de modifier les taux de transversion par rapport à leurs valeurs dans les régions non-transcrites, possibilité qui n'avait pas été envisagée dans les travaux précédents [170]. Il existe d'autres modifications de la matrice de substitution que celles que nous proposons ici, mais la solution que nous présentons a l'avantage de ne faire intervenir que deux changements dans la matrice originelle. Si l'on ne considère pas les taux de transition, les transversions relatives à G et C donnent la plus grande asymétrie de brin. Nous avons donc augmenté le taux de substitution $G \to C$ de 2.9% à 3.36% et diminué le taux de transition $C \to T$ de 14.4% à 13.7% pour ajuster la valeur du biais en GC observée $\overline{S_{GC}} = 0.0366$ (figure 4.3). À notre connaissance, ces transversions liées à la transcription n'ont pas été observées chez les eucaryotes, mais semblent exister chez certains génomes bactériens [327]. Ces observations impliquent donc un raffinement du modèle présenté par GREEN *et al.* [170].

Profils caractéristiques induits par les mécanismes de transcription : présentation des diverses configurations possibles

Les différences entre séquences géniques et intergéniques induites par le biais transcriptionnel se manifestent par l'apparition de « créneaux » dans le signal biais, correspondant à la présence de gènes. Nous donnons ici les diverses configurations géniques pouvant être rencontrées et schématisons la forme des profils associés dans le signal biais.

Nous avons mis en évidence l'existence d'un biais mutationnel associé à la transcription. Comme illustré dans la figure 4.2, ce biais se manifeste en faisant apparaître un profil en forme de créneau dans les signaux S_{TA} et S_{GC} : toutes les parties non-codantes des régions transcrites présentent des biais non nuls relativement constants, alors qu'il sont quasiment nuls dans les régions intergéniques. Les limites du créneau marquent ainsi les extrémités du gène et il est naturel de penser que la hauteur du plateau, c'est-à-dire la valeur du biais, dépend du niveau d'expression du gène (dans la lignée germinale), puisque ce biais est directement lié à la fréquence de transcription. Cependant, ces biais tendent au cours de l'évolution vers des valeurs maximales fixées par les taux de substitution (*cf.* figure 4.3). Il est donc possible que certains gènes fortement exprimés aient atteint une valeur du biais maximale.

Détaillons les formes différentes de profils que peuvent induire les mécanismes de transcription. Un saut dû à la transcription en un point dépend essentiellement de la présence ou non de gènes de part et d'autre de ce point et de leur sens, c'est-à-dire de la *configuration génique* autour du point. Les diverses configurations géniques possible sont résumées dans la table 4.3. Pour symboliser ces configurations, nous utiliserons la convention suivante : le symbole « $\rightarrow$ » représentera un gène sens, « $\leftarrow$ » un gène anti-sens et « $-$ » une région intergénique. Selon la table 4.4, les configurations géniques induisant un saut descendant sont symbolisées par $\rightarrow\leftarrow$, $-\leftarrow$ et $\rightarrow-$, tandis que les configurations géniques associées aux symboles $\leftarrow\rightarrow$, $-\rightarrow$ et $\leftarrow-$ correspondent à un saut ascendant. Les trois autres configurations représentent les cas où un gène succède à un gène de même orientation, ou l'absence de gène. De tels cas peuvent tout de même induire un saut, vu que la valeur du biais attribuée à la transcription peut fluctuer d'un gène à l'autre (en fonction du niveau d'expression notamment). Remarquons que les configurations géniques $\rightarrow\leftarrow$, appelée *convergente*, et $\leftarrow\rightarrow$, appelée *divergente*, engendrent les sauts de plus grandes amplitudes (en fait d'amplitude double) dans le signal biais.

Le mécanisme de transcription se manifestant dans le signal biais par la présence de

	côté 5′		
	anti-sens	intergène	sens
côté 3′			
anti-sens			
intergène			
sens			

TAB. 4.3 – *Illustration des différentes configurations géniques possibles pouvant générer un saut dans le signal biais.*

	côté 5′		
	anti-sens	intergène	sens
côté 3′			
anti-sens	←←	—←	→←
intergène	←—	——	→—
sens	←→	—→	→→

TAB. 4.4 – *Symboles utilisés pour représenter les diverses configurations géniques possibles. Le symbole « → » représente un gène sens, « ← » un gène anti-sens et « — » une région intergénique.*

profils en forme de créneau, les sauts mis en jeu sont symétriquement distribués : à un saut ascendant correspond un saut descendant d'amplitude voisine. La transcription peut expliquer bon nombre des sauts ascendants et descendants détectés dans le signal biais total S (section 3.3) à des échelles de quelques dizaines de milliers de paires de base, c'est-à-dire à des échelles plus petites ou égales à la taille caractéristique des gènes (de l'ordre de 30 000 pb) dans le génome humain [189].

4.2 Étude du biais de composition chez l'homme lié à la réplication

Si les études mettant en évidence le rôle de la transcription dans l'apparition du biais de composition chez les eucaryotes sont peu nombreuses, celles portant sur le rôle de la réplication n'ont pas réussi à mettre clairement en évidence l'existence d'un biais de réplication à grande échelle dans les génomes eucaryotes. Ainsi, l'étude locale autour de l'origine de réplication β-globine menée dans les références [86, 151] pour plusieurs primates n'a pas révélé de différence de taux de mutation autour de l'origine sur les brins avancés et retardés. La situation est plus contrastée chez la levure *Saccharomyces cerevisae* où une asymétrie de composition très nette a été observée aux extrémités des chromosomes, asymétrie que l'on ne retrouve pas dans le reste du génome [161]. Diverses tentatives d'explication ont été proposées, dont la plus réaliste est basée sur la remarque qu'aux extrémités des chromosomes, les brins sont toujours dans le même état avancé ou retardé, alors que sur le reste du chromosome, l'utilisation aléatoire à chaque cycle de réplication des origines identifiées par les séquences consensus ARS ferait que les brins seraient alternativement avancés et retardés, annulant toute éventuelle asymétrie de composition. Récemment, une étude [344] de l'asymétrie de composition de longs segments des chromosomes de l'homme a révélé l'existence de structures qui rappèlent les profils du biais observé chez les procaryotes, mais sans apporter de conclusion sur l'existence d'un biais réplicatif. Dans cette section, nous allons montrer l'existence d'une asymétrie de composition au voisinage des origines de réplication connues chez l'homme, qui ne peut pas être entièrement expliquée par les mécanismes de transcription [79, 369]. Comme pour la transcription, les mécanismes de réplication se manifestent dans le signal biais par la présence de profils caractéristiques en forme de « toit d'usine » différant de la forme en « créneau » induite par la transcription. Ces profils permettent en particulier d'expliquer le grand nombre de sauts ascendants de grande amplitude relativement au nombre de sauts descendant d'amplitude comparable observé à grande échelle dans le chapitre 3 et nous conduiront à proposer un modèle de la réplication chez les mammifères dans le chapitre 5.

Biais de réplication chez les procaryotes : le modèle réplicon

Comme nous l'avons déjà mentionné au début de ce chapitre, l'étude du biais de composition dû à la réplication chez certains procaryotes a révélé, dans de nombreux cas [150, 152, 171, 251, 288, 328, 335, 363], des profils similaires à celui observé pour le chromosome de *Bacillus subtilis* dans la figure 4.4.

Le biais S_{GC} (figure 4.4 (a)) est un signal bruité présentant un saut ascendant clair (sur quelques milliers de paires de base) depuis une valeur négative vers une valeur positive au niveau de l'origine de réplication. Il existe une asymétrie de composition opposée de part et d'autre de l'origine de réplication, signature d'un écart à la règle de parité de type 2. Le chromosome de *Bacillus subtilis* étant circulaire, la terminaison de la réplication où les fourches de réplication se rencontrent, est située à l'opposé de l'origine dans le chromosome et correspond à un saut décroissant du biais d'une valeur positive à une valeur négative. On retrouve donc un profil de biais en forme de créneau mais avec une différence par rapport au profil de biais induit par la transcription (figure 4.2), à savoir qu'à chaque saut, le biais change de signe. Comme cela est illustré dans la figure 4.4 (b), la présence de ces sauts respectivement ascendant (origine) et descendant (terminaison) dans le biais se manifeste sur le profil biais cumulé par un comportement caractéristique en forme de V (origine) ou de Λ (terminaison). Remarquons que les comportements observés pour S_{GC} et F_{GC} dans la figure 4.4 sont aussi observés pour S_{TA} et F_{TA}, avec toutefois une différence d'amplitude moyenne ($|S_{GC}| > |S_{TA}|$) [326, 328]. Ces comportements sont caractéristiques du modèle réplicon [198] pour les eubactéries[❋]. En particulier, l'hypothèse selon laquelle les « points de singularité » des profils de biais cumulé coïncident avec les positions des origines (et des terminaisons) de réplication a permis de prédire *in silico* les positions d'origines de réplication non connues expérimentalement [150, 152, 171, 251, 288, 335, 363].

Il est important de remarquer que comme cela est illustré dans la figure 4.4, lorsque l'on examine l'organisation des gènes autour de l'origine de réplication de *Bacillus subtilis*, la majorité des gènes sens (resp. antisens) sont localisés préférentiellement à la droite (resp. à la gauche) de l'origine de réplication. Cela suggère que la progression des fourches de réplications est orientée avec la transcription de façon à minimiser le risque de collision frontale entre ADN polymérase et ARN polymérase [77, 252, 325, 329].

❋. Les eubactéries sont une subdivision majeure des procaryotes ne comprenant pas les archæbactéries.

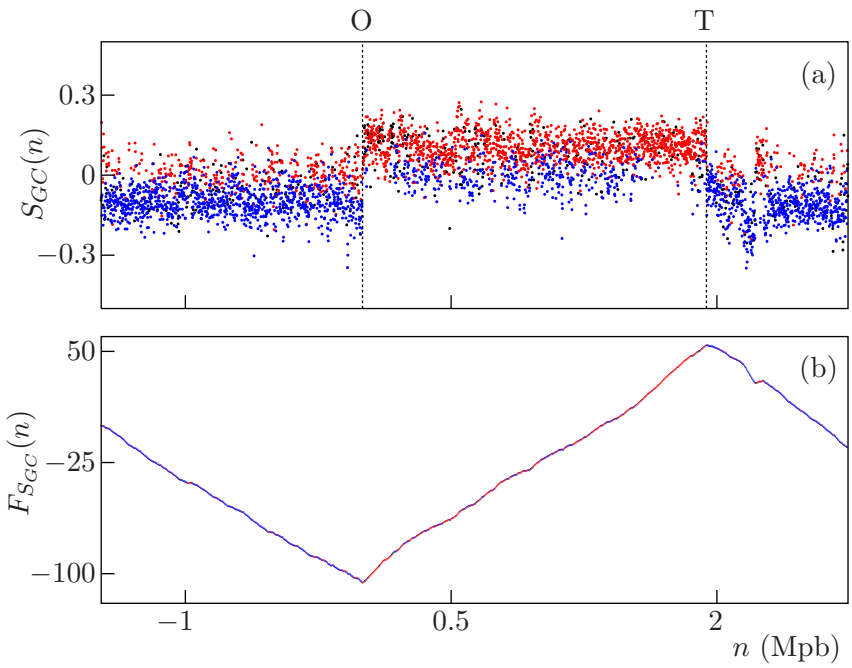

FIG. 4.4 – *(a) Le biais GC le long du génome circulaire de Bacillus subtilis calculé dans des fenêtres de largeur* 1 *kpb. L'origine de l'axe des abscisses a été placée au niveau de l'unique origine de réplication. (b) La marche associée au signal biais GC. Les lignes verticales correspondent aux positions de l'origine (O) et du terminus (T) de réplication. Les points rouges (resp. bleus) correspondent aux gènes sens (resp. antisens).*

Évidences de l'existence d'un biais dû à la réplication

Seulement neuf origines de réplication sont connues expérimentalement chez l'homme. Sur la figure 4.5 est représenté le comportement du signal biais total $S = S_{TA} + S_{GC}$ au voisinage des six origines situées près des gènes MCM4 [231], HSPA4 [358], TOP1 [222], MYC [375], SCA7 [295] et AR [295]. Comme précédemment observé chez *Bacillus subtilis* (*cf.* figure 4.4), le biais présente un saut ascendant à ces origines de réplication (figure 4.5), qui se traduit dans la marche associée par un profil en forme de V pointant sur la position des origines (figure 4.6). Pour les six origines, S_{TA}, S_{GC} et S transitent de valeurs négatives à gauche des origines à des valeurs positives à droite.

Comme précédemment, l'amplitude ΔS d'un saut dans le signal biais S, calculé dans des fenêtres de largeur 1 kpb, est définie par l'égalité (3.4). La figure 4.7 illustre la méthodologie sous-jacente et son utilité pour la détection des origines de réplication : le signal moyenné sur une distance de 20 kpb permet une estimation satisfaisante de l'amplitude des sauts car il minimise les erreurs dues à l'extrême variabilité des valeurs du biais aux points situés à la transition. Remarquons que les conclusions obtenues dépendent peu de

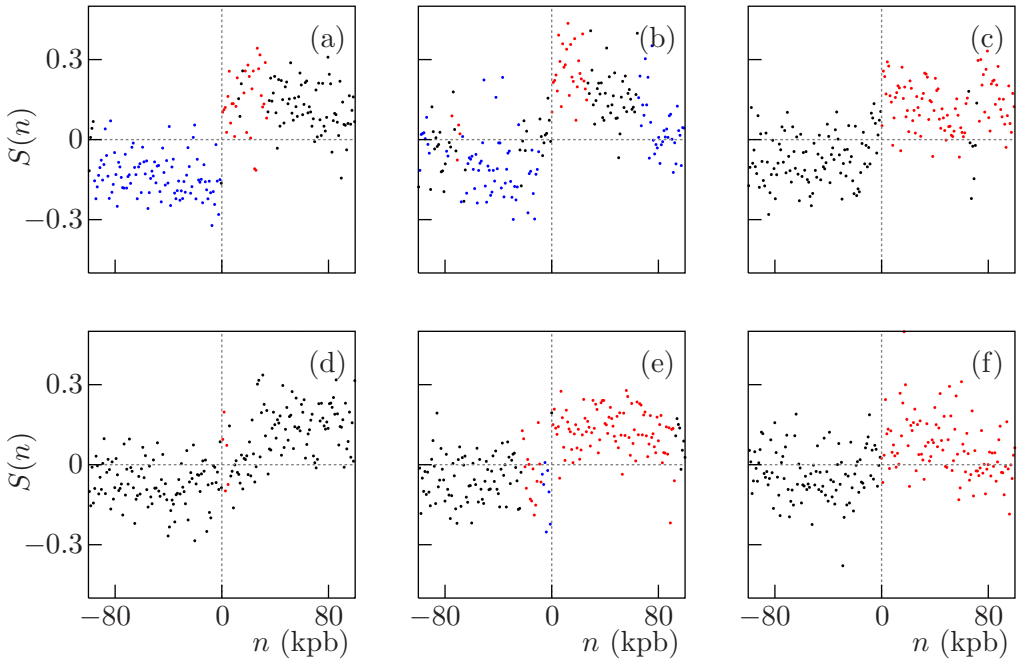

FIG. 4.5 – *Signal biais $S = S_{TA} + S_{GC}$ calculé dans des fenêtres de largeur 1 kpb (après avoir éliminé les séquences répétées) au voisinage de six origines de réplication connues expérimentalement chez l'homme : MCM4 (a), HSPA4 (b), TOP1 (c), MYC (d), SCA7 (e) et AR (f). L'origine de l'axe des abscisses désigne la position de l'origine de réplication. Les gènes sens situés sur le brin Watson sont représentés en rouge et ceux sur le brin Crick sont représentés en bleu.*

l'échelle de lissage [369]. Les amplitudes ΔS évaluées pour chacune des six origines sont données dans la table 4.5. Remarquons que ces valeurs sont systématiquement supérieures à l'amplitude de saut maximale induite par la transcription dans les situations de gènes convergents, à savoir $|\Delta S| = 0.14$ (*cf.* section 4.1). Ainsi les amplitudes des sauts du biais observés dans la figure 4.5 ne peuvent pas être entièrement expliquées par une orientation favorable des gènes de part et d'autre de l'origine, qui reviendrait à additionner les valeurs extrêmes de la table 4.1.

Si les résultats rapportés dans les figures 4.5 et 4.6 suggèrent fortement l'existence d'un biais de réplication, nous ne pouvons toutefois pas totalement exclure qu'une orientation favorable des gènes de part et d'autre de l'origine de réplication soit responsable des sauts observés. Pour rejeter cette possibilité, nous avons calculé les biais moyens dans des fenêtres de largeur 100 kpb dans les régions intergéniques des brins avancés autour des six origines de réplication de la table 4.5 (table 4.6). Les régions intergéniques considérées sont celles ne contenant aucune annotation relative à des données de transcription, de manière à s'affranchir au mieux du biais de composition dû à la transcription [369]. La mesure du biais total dans ces régions donne une valeur moyenne de $\overline{S} = \overline{S_{TA}} + \overline{S_{GC}} =$

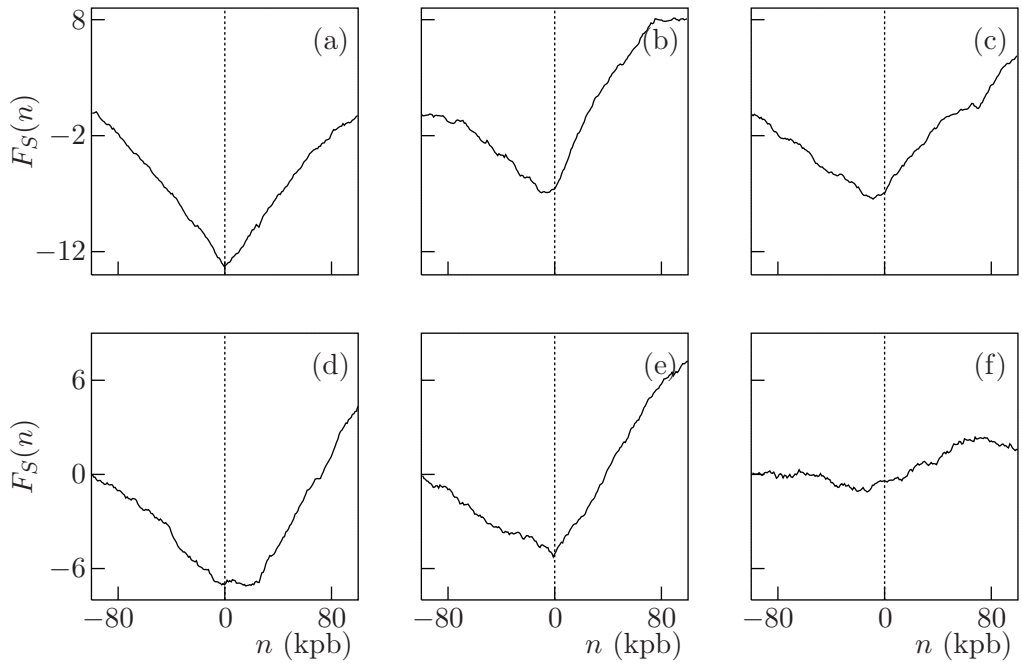

FIG. 4.6 – *Marche associée au signal biais S calculé dans des fenêtres de largeur* 1 *kpb (après avoir éliminé les séquences répétées) au voisinage de six origines de réplication connues expérimentalement : MCM4 (a), HSPA4 (b), TOP1 (c), MYC (d), SCA7 (e) et AR (f). L'origine de l'axe des abscisses désigne la position de l'origine de réplication. Comme pour Bacillus subtilis (figure 4.4), la marche adopte un profil caractéristique en forme de V pointant à proximité de l'origine.*

origine	ΔS	ΔS_{TA}	ΔS_{GC}
MCM4	0.37	0.17	0.20
HSPA4	0.38	0.13	0.25
TOP1	0.22	0.16	0.06
MYC	0.15	0.05	0.10
SCA7	0.16	0.04	0.12
AR	0.18	0.12	0.06

TAB. 4.5 – *Le saut* ΔS *évalué selon la méthode définie par l'égalité (3.4) de part et d'autre de chaque origine de réplication. L'amplitude de ces sauts ne peut être entièrement expliquée par un biais de composition dû à la transcription.*

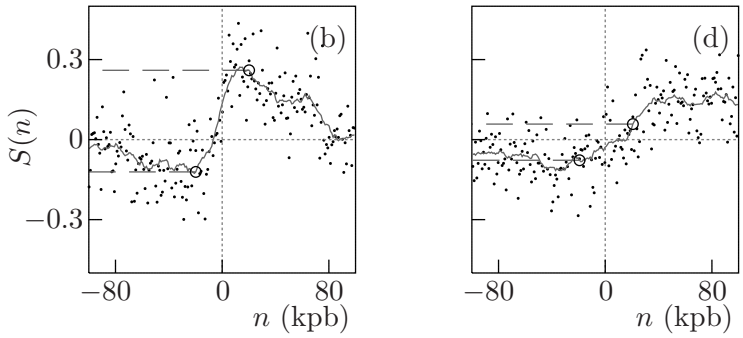

FIG. 4.7 – *Illustration de la méthode définie par l'égalité (3.4) pour évaluer l'amplitude du saut $|\Delta S|$ de part et d'autre des origines HSPA4 (b) et MYC (d) (cf. figure 4.5). Le signal biais total est moyenné dans des fenêtres de largeur* 20 *kpb (courbe en gris foncé), puis les valeurs à une distance de* 20 *kpb de l'origine du signal ainsi obtenu sont sélectionnées pour estimer ΔS. On constate que la faible amplitude du saut obtenue avec MYC (d), à savoir $\Delta S = 0.15$, traduit simplement l'existence d'une longue zone de transition autour de l'origine.*

0.07, correspondant à une amplitude moyenne de saut $\overline{\Delta S} = 0.14$, c'est-à-dire une valeur significativement différente de zéro pour que l'on puisse conclure à l'existence d'un biais de composition dû à la réplication [80, 369].

Un examen attentif des résultats rapportés dans la table 4.6 révèle que la valeur du biais moyen est beaucoup plus forte ($\overline{S}$ est voisin de 0.14) lorsque la transcription et la réplication s'effectuent dans le même sens, *i.e.* lorsque l'orientation des gènes contribue à l'amplitude du saut, et est proche de zéro dans le cas contraire. Les longueurs des séquences transcrites dans le même sens que la réplication sont plus grandes que celles des séquences transcrites dans le sens opposé. La transcription a donc tendance à s'effectuer dans le même sens que la réplication, comme cela est généralement observé dans les organismes procaryotes (*cf.* figure4.4) [325, 326].

Conservation du biais de réplication chez les mammifères

Dans le génome, les régions de deux organismes différents contenant la même succession de gènes sont appelées *synthéniques* [179]. Dans ces régions, les éléments fonctionnels ont beaucoup de chance d'être conservés d'un organisme à l'autre, l'absence de remaniement chromosomique dans ces régions n'excluant cependant pas la possibilité qu'il y ait eu des mutations de façon indépendante dans le génome de chaque organisme depuis leur divergence. Les origines de réplication sont donc susceptibles d'être conservées dans les régions synthéniques. Si une origine ne l'est pas, on s'attend à ce que l'asymétrie s'annule progressivement dans la région correspondante. Nous avons analysé le signal biais S et la

	biais TA	biais GC	biais total	taille
intergénique	$0.039 \pm 4\,10^{-3}$	$0.03 \pm 4\,10^{-3}$	$0.069 \pm 4\,10^{-3}$	487
brin transcrit	$0.075 \pm 3\,10^{-3}$	$0.068 \pm 4\,10^{-3}$	$0.143 \pm 4\,10^{-3}$	358
brin non-transcrit	$-0.019 \pm 1\,10^{-2}$	$-0.003 \pm 14\,10^{-3}$	$-0.022 \pm 13\,10^{-3}$	49
intergénique s.r.c.	$0.04 \pm 4\,10^{-3}$	$0.03 \pm 5\,10^{-3}$	$0.07 \pm 5\,10^{-3}$	461
souris	$0.036 \pm 4\,10^{-3}$	$0.022 \pm 5\,10^{-3}$	$0.058 \pm 5\,10^{-3}$	441

TAB. 4.6 – *Asymétries de composition moyennes $\overline{S_{TA}}$, $\overline{S_{GC}}$ et $\overline{S}$ associées aux six origines de réplication étudiées dans la table 4.5. Les séquences intergéniques sont toujours considérées dans le sens de la réplication. Les régions non-conservées entre l'homme et la souris ont aussi été analysées spécifiquement (intergénique sans région conservée (s.r.c.) et souris respectivement). L'asymétrie dans les introns est considérée sur le brin non-transcrit quand la transcription est dans le même sens que la réplication et sur le brin transcrit dans le second cas. La longueur des séquences étudiées est donnée en kpb dans la dernière colonne.*

marche correspondante dans les régions synthéniques des séquences contenant les origines humaines, dans les génomes de la souris (*Mus musculus*) et du chien (*Canis familiaris*). Comme le montre la figure 4.8, dans les deux cas on observe toujours un profil du biais cumulé en forme de V pointant à la position originelle de l'origine de réplication dans la séquence humaine. Pour la souris, on obtient une valeur du biais moyen associé à la réplication voisine de $\overline{S} = 0.058$ (table 4.6). La faible proportion de séquences conservées entre l'homme et la souris semble pourtant indiquer que les séquences ont fortement divergé. Notons qu'une étude expérimentale récente a confirmé l'existence de l'origine MYC chez la souris [163]. Tout ceci porte à croire que les origines de réplication et le biais associé sont conservés chez les mammifères.

Il existe des segments d'ADN fonctionnels conservés entre l'homme et la souris. Ces séquences, soumises à des pressions de sélection, pourraient avoir une asymétrie de composition propre et ainsi contribuer au biais observé dans les régions intergéniques. Pour les six origines concernées, ces séquences correspondent au plus à 5.3% du total des séquences intergéniques et l'élimination de ces régions ne change pas les résultats rapportés dans la figure 4.8 et dans la table 4.6.

Trois des origines connues expérimentalement chez l'homme ne montrent pas de saut net dans le signal biais ; il s'agit de Lamine B2 [1, 307], β-globine [8, 9, 10] et DNMT1 [16] (figure 4.9). Les marches associées ne présentent pas non plus de profil en forme de V permettant d'identifier la position de l'origine (figure 4.10). Plusieurs explications peuvent être avancées. Ces origines de réplication peuvent ne pas être actives dans la lignée germinale. Elles peuvent aussi ne pas être actives à chaque cycle cellulaire dans la lignée germinale, ce qui réduirait le biais de réplication associé. Le biais de transcription peut

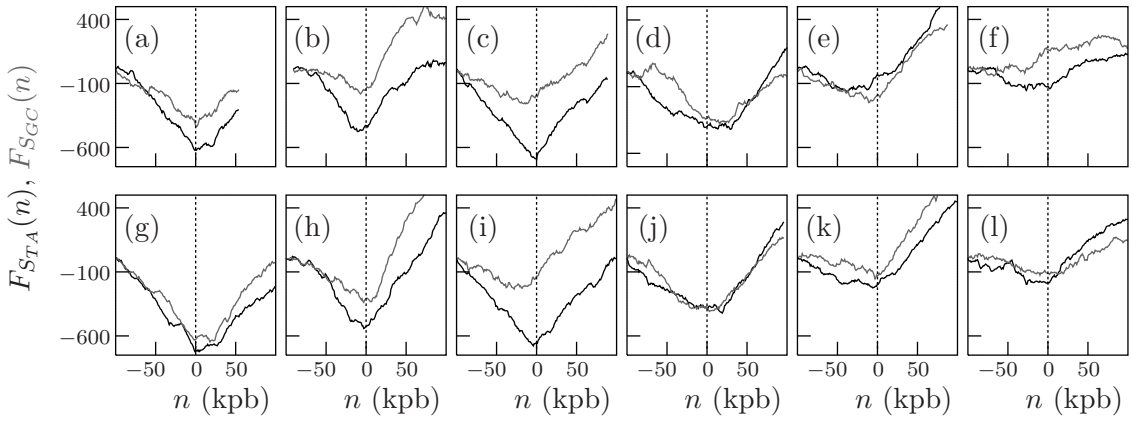

FIG. 4.8 – *Marches $F_{S_{TA}}$ (en noir) et $F_{S_{GC}}$ (en gris) associées aux signaux biais S_{TA} et S_{GC} respectivement, calculées dans les six régions du génome de la souris ((a)–(f)) et du chien ((g)–(l)) qui sont synthéniques aux régions de l'homme contenant une origine de réplication connue expérimentalement : MCM4 ((a) et (g)), HSPA4 ((b) et (h)), TOP1 ((c) et (i)), MYC ((d) et (j)), SCA7 ((e) et (k)) et AR ((f) et (l)).*

éventuellement compenser un biais de réplication. On peut en effet imaginer un environnement génique convergent (→←) compensant le biais de réplication. Il est aussi possible que les régions contenant ces origines aient subit de nombreux remaniement chromosomiques ou que certaines régions ne soient actives que depuis récemment dans l'évolution. Finalement, certains profils peuvent aussi s'expliquer par la présence de plusieurs origines dans une même zone.

Pour résumer, nous avons observé dans le signal biais du génome humain la présence de sauts ascendants au voisinage de la plupart des origines de réplication connues. Nous avons vu que chez certains génomes bactériens, mitochondriaux ou viraux, de brusques variations du signal sont associées à la réplication et utilisées pour détecter les origines de réplication [249, 288, 363]. Les résultats rapportés dans cette section suggèrent qu'une telle association existe aussi chez l'homme, la présence de gènes de part et d'autre des origines n'étant pas suffisante pour expliquer l'amplitude des sauts mesurée.

Mise en évidence d'un profil caractéristique en forme de « toit d'usine » dans le signal biais : une signature du mécanisme de réplication chez les mammifères

Nous avons vu au chapitre précédent qu'il existe, à grande échelle, une dissymétrie importante entre le nombre de sauts ascendants de grande amplitude et le nombre de sauts descendants de même amplitude dans le signal biais des chromosomes humains. Ces sauts ascendants ayant des amplitudes comparables à celles rapportées dans la table 4.5 pour les

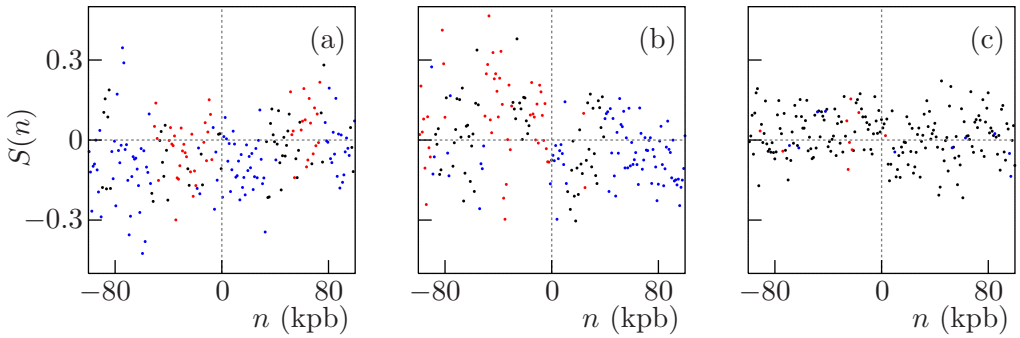

FIG. 4.9 – *Signal biais $S = S_{TA} + S_{GC}$ calculé dans des fenêtres de largeur 1 kpb (après avoir éliminé les séquences répétées) au voisinage de trois origines connues expérimentalement chez l'homme : DNMT1 (a), lamine B2 (b) et β-globine (c). L'origine de l'axe des abscisses désigne la position de l'origine de réplication. Les gènes sens situés sur le brin Watson sont représentés en rouge et les gènes anti-sens sur le brin Crick en bleu.*

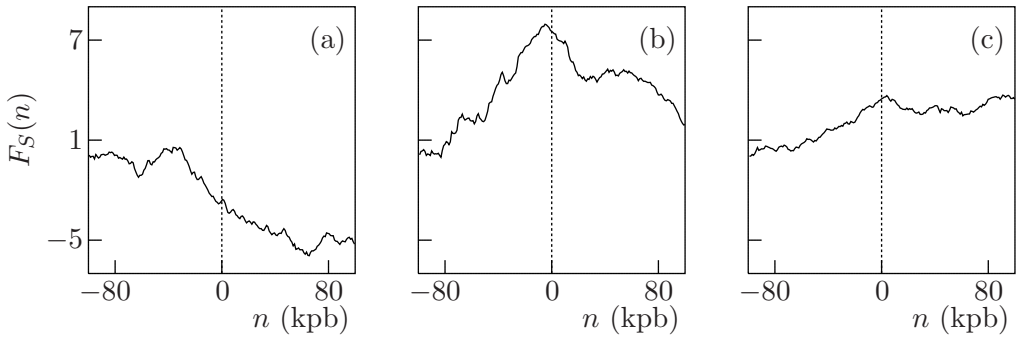

FIG. 4.10 – *Marche associée au signal biais S calculé dans des fenêtres de largeur 1 kpb au voisinage de trois origines connues expérimentalement chez l'homme : DNMT1 (a), lamine B2 (b) et β-globine (c). L'origine de l'axe des abscisses désigne la position de l'origine de réplication.*

origines de réplication connues expérimentalement, on pourrait donc être tenté d'associer ces sauts ascendants détectés à grande échelle à une signature de la présence d'une origine de réplication.

Pour comprendre d'où pourrait provenir l'asymétrie dans la répartition des sauts présente dans le signal biais à grande échelle (*cf.* figure 3.9), il suffit d'observer le comportement de ce signal sur des fragments de plusieurs millions de paires de base. Dans la figure 4.11, on constate clairement la présence d'un profil caractéristique en forme de « toit d'usine » sur trois régions des chromosomes 2, 9 et 18 de l'homme. Dans chacun des trois profils, on observe des sauts ascendants de grande amplitude séparés par un comportement linéairement décroissant du biais sans aucun saut descendant d'amplitude comparable. Ces motifs expliquent donc très bien l'excès de sauts ascendants observé aux échelles de quelques centaines de milliers paires de base dans le chapitre 3.

Les profils de biais observés en forme de toit d'usine aussi marqué ne peuvent pas être expliqués par la transcription qui, comme nous l'avons vu dans la figure 4.4 (a), induit un profil en forme de créneau caractéristique du modèle réplicon pour les eubactéries. Le comportement linéairement décroissant entre deux sauts ascendants du biais peut donc être attribué au mécanisme de réplication. Cette hypothèse est corroborée par le fait que, comme on peut la voir dans la figure 4.11, des profils en toit d'usine sont observés dans des régions majoritairement intergéniques, excluant la possibilité que cette forme très caractéristique puisse résulter d'un arrangement particulier des gènes. Il est aussi important de remarquer que si l'on regarde à plus grande échelle, le comportement du biais autour des origines de réplication connues, comme cela est fait dans la figure 4.12, on réalise que celles-ci correspondent à des sauts ascendant dans un profil en forme de toit d'usine. Dans la marche ADN associée au biais, la décroissance linéaire du biais modifie légèrement la forme en V pointant sur l'origine observée chez certains procaryotes : « les branches du V » ne varient plus linéairement, comme chez *Bacillus subtilis* (figure 4.4 (b)), mais quadratiquement, comme cela est observé dans la figure 4.13 pour les origines de réplication connues expérimentalement.

Nous avons identifié un profil particulier du biais de réplication en forme de « toit d'usine » expliquant la dissymétrie observée dans le signal biais entre les nombres de sauts ascendants et descendants de grande amplitude. Ce profil semble être conservé dans les génomes des mammifères, comme cela est illustré dans la figure 4.14. En examinant les profils obtenus pour les 22 chromosomes asexués de l'homme, on remarque que la distance entre deux sauts ascendants successifs est très variable, depuis la centaine de milliers de paires de base jusqu'à plusieurs millions de paires de base. Cette gamme de distances entre origines de réplication putatives correspond très bien à ce qui a été estimé dans

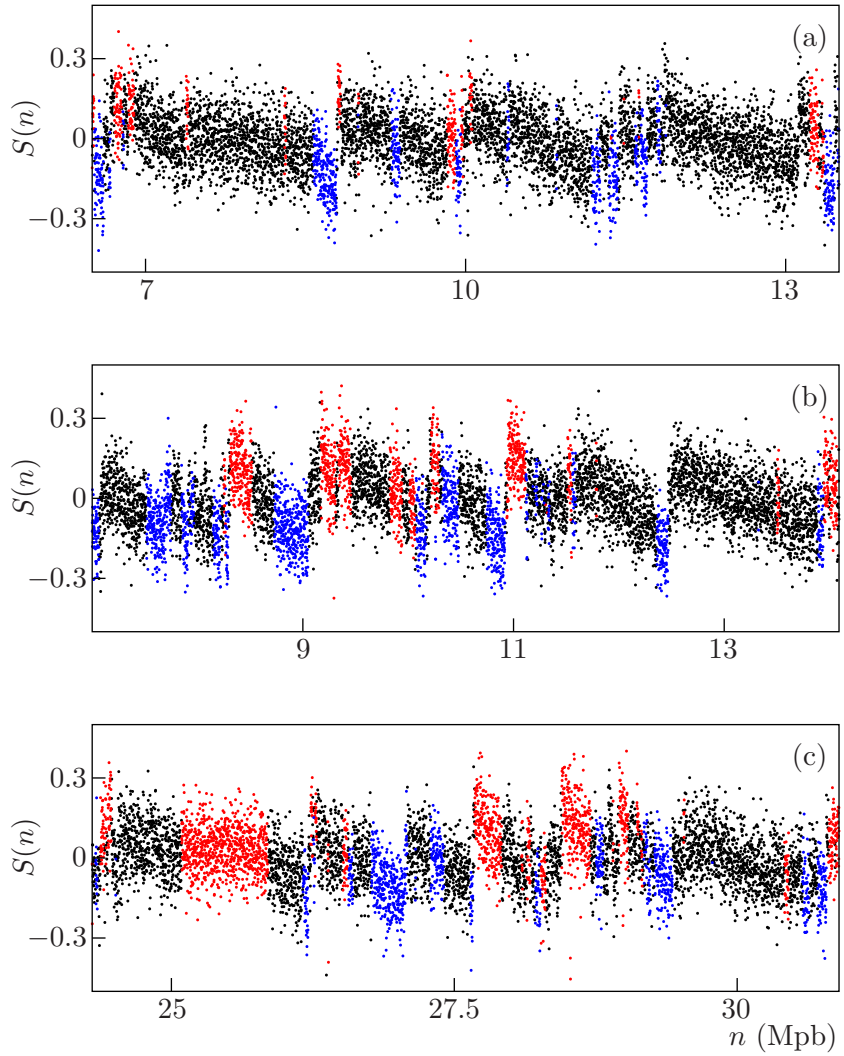

FIG. 4.11 – *Mise en évidence de profils de biais en forme de « toit d'usine ». Représentation du biais total $S = S_{TA} + S_{GC}$ sur des portions des chromosomes 2 (a), 9 (b) et 18 (c), calculé dans des fenêtres de largeur 1 kpb (après avoir éliminé les séquences répétées). Les gènes sens situés sur le brin Watson sont colorés en rouge et les gènes anti-sens sur le brin Crick en bleu.*

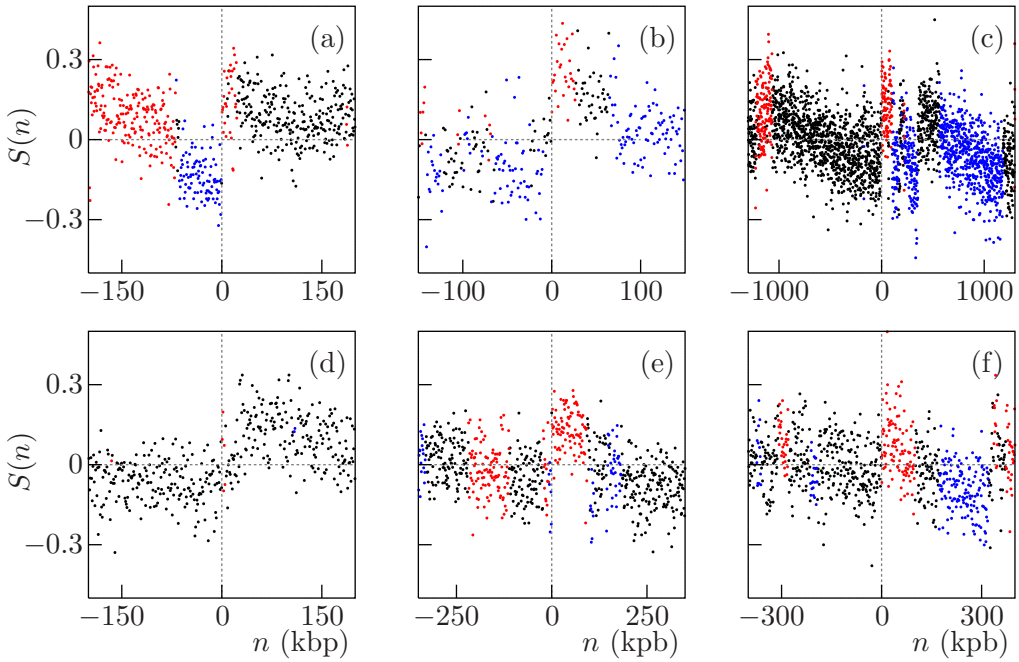

FIG. 4.12 – *Signal biais $S = S_{TA} + S_{GC}$ calculé dans des fenêtres de largeur 1 kpb (après avoir éliminé les séquences répétées) au voisinage de six origines de réplication connues expérimentalement chez l'homme : MCM4 (a), HSPA4 (b), TOP1 (c), MYC (d), SCA7 (e) et AR (f). L'origine de l'axe des abscisses désigne la position de l'origine de réplication. Les gènes sens situés sur le brin Watson sont représentés en rouge et les gènes anti-sens sur le brin Crick en bleu. Les origines connues se placent au bord de profils en forme de « toit d'usine ».*

la référence [55] comme taille moyenne des réplicons chez les mammifères, à savoir de l'ordre de 400–500 kpb dans les séquences natives, soit environ 250 kpb sans les séquences répétées, avec des valeurs extrêmes de l'ordre de quelques millions de paires de base. Il est aussi important de remarquer que cette gamme de distances entre sauts ascendants de grande amplitude est tout à fait consistante avec la gamme de fréquences caractéristiques mise en évidence dans la figure 3.7 (b) et qui caractérise l'existence de rythmes basses fréquences dans l'asymétrie de composition chez les mammifères (chapitre 3) [297, 299].

Étude statistique des profils de biais dûs à la réplication

Avec comme objectif de prédire la position de certaines origines de réplication dans le génome humain, nous allons maintenant procéder à une étude statistique des caractéristiques des profils en toit d'usine présents dans les signaux de biais des 22 chromosomes asexués de l'homme et fournir des informations sur l'organisation génique au voisinage de ces origines putatives.

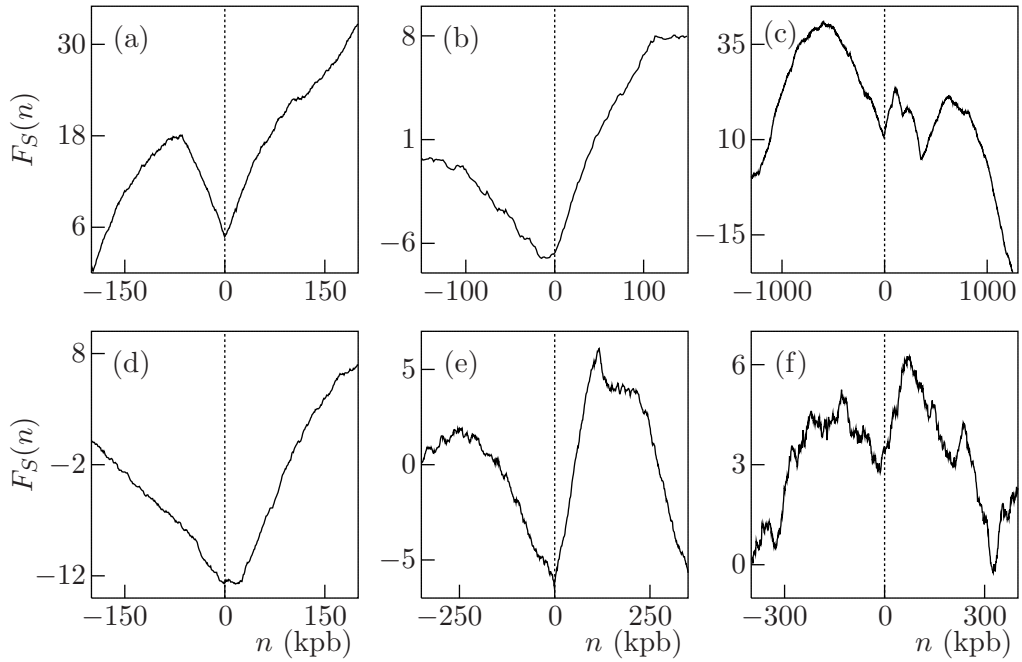

FIG. 4.13 – *La marche associée au signal biais $S = S_{TA} + S_{GC}$ calculé dans des fenêtres de largeur* 1 *kpb (après avoir éliminé les séquences répétées) au voisinage de six origines de réplication connues expérimentalement chez l'homme : MCM4 (a), HSPA4 (b), TOP1 (c), MYC (d), SCA7 (e) et AR (f). L'origine de l'axe des abscisses désigne la position de l'origine de réplication. Par rapport au profil en forme de V caractéristique du modèle réplicon (figure 4.4), on observe un profil en forme de « ϒ ».*

Pour sélectionner les profils à étudier dans le signal biais, nous avons d'abord choisi les sauts d'amplitude suffisante, suivant la méthodologie décrite dans la figure 3.8. L'échelle de sélection des maxima choisie de 200 kpb permet de s'affranchir autant que possible des sauts induits par la transcription, la taille moyenne des gènes (environ 30 kpb sur les séquences natives, ce qui équivaut approximativement à 15 kpb sur les séquences masquées) étant significativement inférieure au seuil de 200 kpb. Remarquons que ce seuil est par contre plus petit que la distance moyenne entre deux origines de réplication putatives successives [55]. Comme précédemment, nous estimerons l'amplitude d'un saut ΔS à l'aide de l'égalité (3.4).

Pour les sauts d'amplitude $|\Delta S|$ supérieure à 0.125, le rapport entre le nombre de sauts ascendants et le nombre de sauts descendants est supérieur à 3. Toutefois, ce rapport dépend fortement de la concentration en GC : pour les régions ayant un pourcentage en GC supérieur à 42%, ce rapport est égal à 1.9 [369], ce qui traduit simplement le fait que dans ces régions riches en gènes, les profils de biais en toit d'usine sont moins apparents et plus difficilement détectables, probablement à cause d'une diminution significative de la distance moyenne entre origines de réplication, qui deviendrait comparable à la taille des

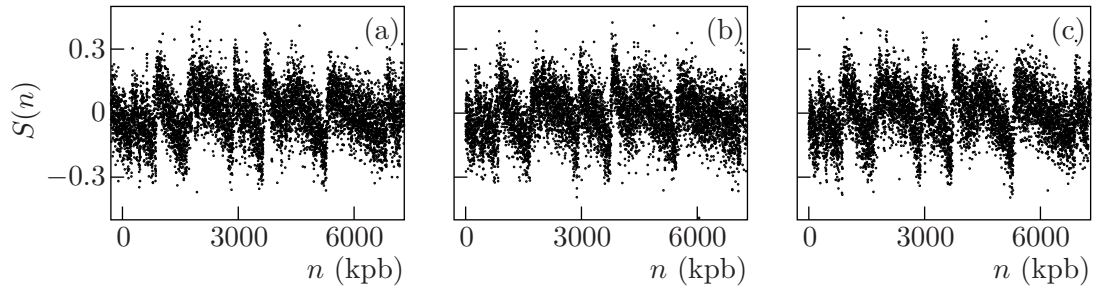

FIG. 4.14 – *Signal biais $S = S_{TA} + S_{GC}$ calculé dans des fenêtres de largeur 1 kpb sur des morceaux de génomes de mammifères (après avoir éliminé les séquences répétées). (a) Morceau du chromosome 9 de l'homme. (b) Morceau du chromosome 4 de la souris. (c) Morceau du chromosome 5 du chien. Les Morceaux des chromosomes de la souris et du chien ((b) et (c)) sont synthéniques au morceau du chromosome 9 de l'homme (a).*

gènes. Pour cette raison, nous n'avons travaillé que dans les régions dont le pourcentage en *GC* est inférieur à 42%, régions pour lesquelles le nombre de sauts ascendants excède d'un facteur supérieur à 5 le nombre de sauts descendants. Nous avons utilisé ce critère d'amplitude pour décider si un saut pouvait être attribué à une origine ou non : seuls les sauts ascendants dont l'amplitude $|\Delta S|$ est supérieure à 0.125 ont donc été retenus. Ainsi, 1012 sauts ascendants ont été sélectionnés. Dans la figure 4.15 est représentée l'organisation génique autour de ces 1012 origines de réplication putatives, en fonction de la distance à celles-ci. Les profils de composition génique obtenus confirment que les gènes sens (resp. anti-sens) ont tendance à être positionnés à droite (resp. à gauche) des origines et donc à être co-orientés avec le sens de progression des fourches de réplication. Toutefois, à partir de la détection des sauts descendants de forte amplitude, on estime qu'environ 20% des sauts ascendants parmi les 1012 pourraient être dus à la transcription.

Parmi les sauts ascendants sélectionnés, un nombre non négligeable peut donc provenir d'une orientation favorable des gènes, *i.e.* d'une configuration génique divergente (←→). Afin de disposer d'une banque plus fiable, nous avons sélectionné, parmi les origines précédemment retenues, les paires d'origines successives telles que l'amplitude des sauts dans le signal biais ne peut pas s'expliquer par une orientation favorable des gènes et pour lesquelles le profil biais bruité entre ces deux sauts ascendants décroit linéairement sans présenter de saut, descendant comme ascendant, d'amplitude notable. Avec ces critères supplémentaires, 287 paires d'origines définissant des régions dans le signal biais vérifiant rigoureusement le profil caractéristique en toit d'usine ont été finalement retenues ; ces paires correspondent à une prédiction de 486 origines putatives différentes, nombre à comparer avec la dizaine d'origines identifiées expérimentalement [79].

En étudiant la composition génique à proximité des 486 origines ainsi prédites (table

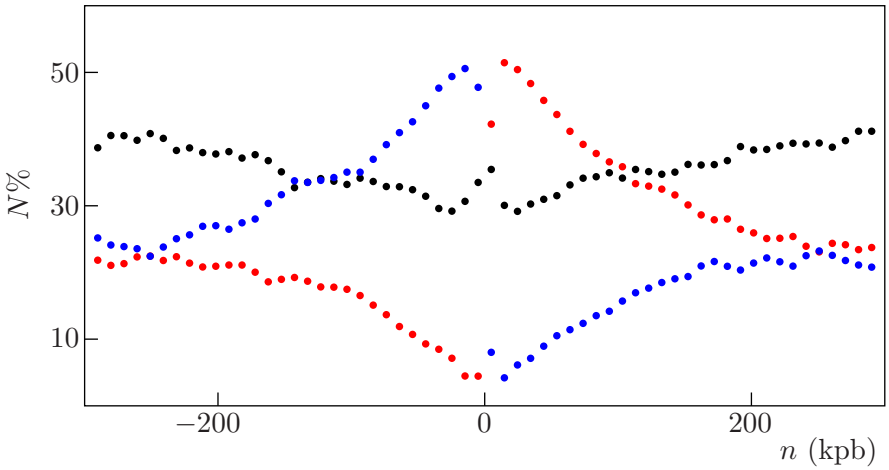

FIG. 4.15 – *Profils de composition génique et intergénique calculés dans des fenêtres de largeur* 10 *kpb de part et d'autre des* 1012 *origines prédites, en fonction de la distance (en kpb) à l'origine (positionnée en zéro). Dans chacune des fenêtres est calculé le pourcentage moyen de zones intergéniques (●), ou contenant des introns de gènes sens (●), ou anti-sens (●).*

4.7), dont rappelons-le, les sauts associés dans le signal biais ne peuvent s'expliquer uniquement par un biais de transcription, on remarque l'absence ou presque de configuration génique convergente (→←), qui aurait tendance à introduire un saut dans le signal biais dû à la transcription qui compenserait en partie le saut ascendant dû à la réplication. Cette observation n'est pas surprenante, puisque nous avons voulu minimiser le nombre de faux positifs dans notre banque de candidats en imposant un critère contraignant sur l'amplitude du saut. Par contre, le faible nombre d'origines associées à la configuration divergente (←→) comparativement aux organisations intergène – sens (— →) et gène anti-sens – intergène (← —) rassure quant à la pertinence de la sélection effectuée parmi les 1012 initialement détectées. Ainsi, près de 40% des 486 origines prédites sont associées à des organisations géniques très nettes avec une transcription orientée dans le même sens que le sens de progression de la fourche de réplication. Remarquons que des résultats identiques sont obtenus pour les origines situées du côté $5'$ ou du côté $3'$ des 287 domaines sélectionnés.

À l'aide de ces 287 domaines, nous pouvons maintenant caractériser le profil typique du signal biais induit par la réplication, à savoir un comportement linéairement décroissant bordé par deux sauts ascendants pointant sur les origines. Les résultats rapportés dans la figure 4.16 ont été obtenus après normalisation à 1 des longueurs des domaines de réplication sélectionnés. De cette manière, nous pouvons comparer les profils du signal biais entre eux et en particulier les coefficients angulaires des pentes observées, qui ne semblent dépendre que de la distance entre deux origines.

côté 5′			
anti-sens	intergène	sens	
			côté 3′
8	1	0	anti-sens
92	48	0	intergène
17	83	5	sens

TAB. 4.7 – *Composition génique majoritaire mesurée dans des fenêtres de largeur* 40 *kpb du côté* 5′ *et* 3′ *des* 486 *origines prédites, délimitant* 287 *domaines de réplication identifiés. Une fenêtre de largeur* 1 *kpb est considérée comme intergénique, gène sens ou anti-sens par la règle de majorité, en utilisant la banque de donnée KnownGene. Chaque saut dans le signal biais est associé aux contenus géniques dans les fenêtres de* 40 *kpb situées respectivement aux côtés* 5′ *et* 3′ *; un saut n'est retenu que si les deux fenêtres qui l'entourent comportent* 80% *de points intergéniques, gènes sens ou anti-sens.*

Le profil moyen, du signal biais dans les régions de réplication identifiées (figure 4.16) a été calculé comme suit. Pour chaque domaine renormalisé, le profil moyen à été calculé dans de fenêtres de taille $1/10$ centrées aux positions $t = j/10$ ($0 \leqslant j \leqslant 10$). Les deux fenêtres du bord (en $t = 0$ et $t = 1$) ont été retirées de l'étude pour éviter les problèmes de mesure de biais dûs à l'imprécision avec laquelle les bords des profils sont estimés. Deux types de fenêtres ont été retenus : les fenêtres intergéniques à plus de 90% et celles géniques à plus de 90%. Le profil moyen calculé uniquement à l'aide des fenêtres intergéniques à plus de 90% (figure 4.16 (a)) montre un comportement linéaire remarquable coupant l'axe des abscisses au milieu ($t = 1/2$) du domaine de réplication. On peut estimer les valeurs du biais au bord de ces domaines. On obtient $\overline{S} = 0.085 \pm 0.004$ en $t = 0$ et $\overline{S} = -0.085 \pm 0.004$ en $t = 1$, valeurs qui peuvent être considérées comme caractéristiques du biais de réplication. Le biais de réplication $\overline{S} = 0.085$ serait donc légèrement plus élevé que le biais moyen de transcription, voisin de $\overline{S} = 0.07$ (table 4.1). Le profil issu des fenêtres géniques à plus de 90% (figure 4.16 (b)) est tout aussi remarquablement linéaire et présente des valeurs du signal biais au bord plus importantes, $\overline{S} = 0.14 \pm 0.005$ et $\overline{S} = -0.137 \pm 0.005$ en $t = 0$ et $t = 1$ respectivement. Ces valeurs s'expliquent par l'orientation préférentielle des gènes autour des origines évoquées précédemment : ceux-ci ont tendance à s'orienter dans le même sens que la progression de la fourche de réplication. Au voisinage immédiat des origines, 50% des gènes sont orientés coopérativement, alors que 10% seulement le sont dans le sens opposé (figure 4.16 (c)). Le biais dû à la transcription a donc tendance à s'ajouter au biais dû à la réplication autour des origines et donc à augmenter l'amplitude des sauts observés dans le signal biais.

Les coefficients angulaires de chaque profil du signal biais dans les domaines renormali-

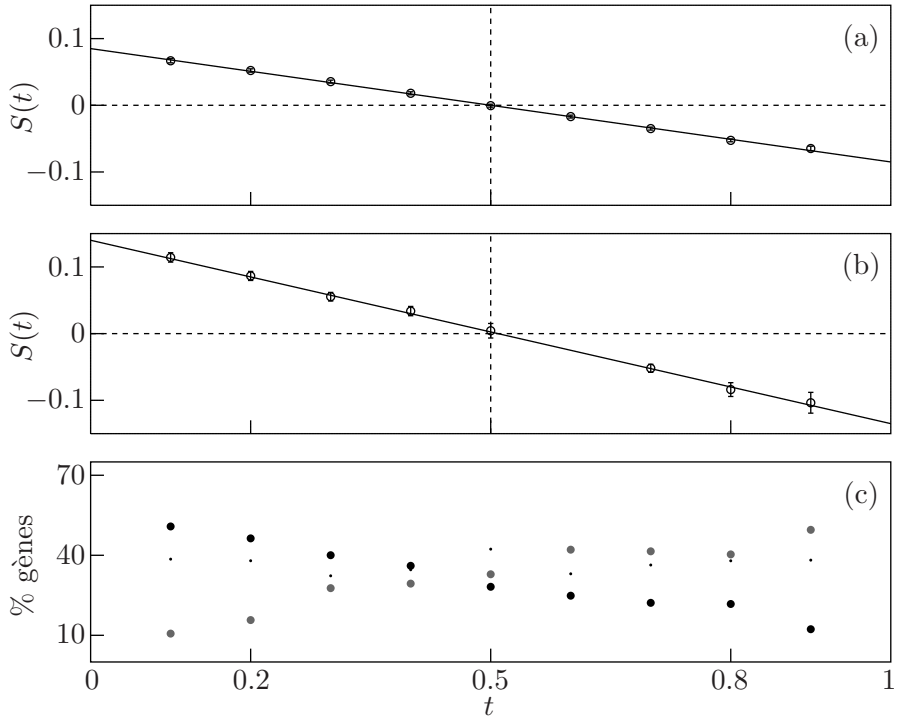

FIG. 4.16 – *Profil moyen du biais $S = S_{TA} + S_{GC}$ obtenu à partir des 287 domaines de réplication identifiés. La taille de chaque domaine a été ramenée à 1 et la valeur moyenne du biais a été calculée dans des fenêtres de taille 10^{-1} centrées aux positions $t = j\,10^{-1}$ ($0 \leqslant j \leqslant 10$), les fenêtres au bord ayant été omises. (a) Profil moyen du biais calculé à partir des fenêtres contenant plus de 90% de zones intergéniques. (b) Profil moyen calculé à partir des fenêtres contenant plus de 90% de zones géniques. Les droites sont issues d'une simple régression linéaire. (c) Pourcentage de gènes sens (•), gènes anti-sens (•) et de zones intergéniques (·) dans les fenêtres considérées en (a) et (b).*

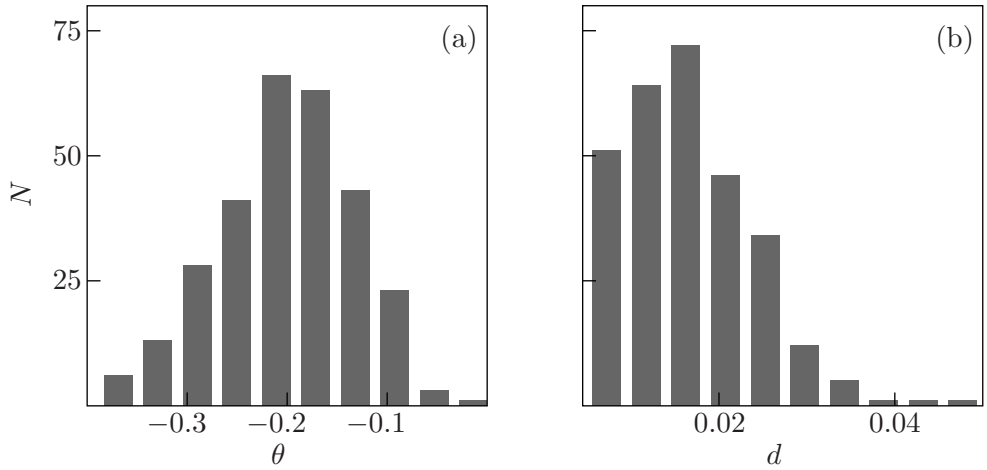

FIG. 4.17 – *(a) Histogramme des coefficients angulaires des* 287 *domaines de réplication identifiés, après avoir renormalisé leur longueur à l'unité. La méthode utilisée est celle de la médiane. (b) Histogramme des déviations absolues moyennes aux pentes estimées en (a).*

sés ont été calculés grâce à la méthode de la médiane* ; l'histogramme des valeurs obtenues est représenté dans la figure 4.17 (a). Le coefficient angulaire correspond à l'opposé de l'amplitude totale des sauts ascendants du biais observé aux deux origines de réplication délimitant le domaine. Dans la figure 4.17 (b) nous avons calculé l'histogramme des écarts à un comportement linéaire décroissant de chaque profil, donnés par les déviations absolues moyennes, afin de vérifier que chaque profil pris séparément présente bien un comportement linéaire caractéristique des profils de biais en forme de toit d'usine.

Nous avons ainsi mis au point une méthodologie de sélection de domaines qui a permis de localiser 486 origines de réplication délimitant ces domaines. Nous avons pu estimer la valeur du biais moyen dû à la réplication dans le signal biais total, $\overline{S} = 0.085$. Ce biais correspond à des sauts ascendants d'amplitude $\overline{\Delta S} = 0.17$ aux positions des origines, soit près de trois fois l'amplitude caractéristique des sauts observés aux positions des promoteurs des gènes, dûs aux mécanismes de transcription. Nous avons vérifié quantitativement le comportement linéaire du signal biais entre deux sauts ascendants consécutifs sélectionnés ; ce comportement ne dépend que de la distance entre les sauts. Les observations suggèrent aussi que la réplication et la transcription ont tendance à s'effectuer dans le même sens, ce qui n'est pas sans rappeler l'organisation génique autour des origines de réplication chez les procaryotes (figure 4.4) [325, 326]. Si les 486 origines prédites représentent incontestablement une avancée quantitative par rapport aux 9 origines de réplication connues expérimentalement, cela ne constitue toujours qu'une petite partie de la dizaine (voire quelques dizaines) de milliers d'origines attendues. Un effort méthodolo-

∗. Cette méthode est brièvement décrite dans l'annexe B.

gique supplémentaire est donc nécessaire. La méthode développée dans ce chapitre n'est qu'une première étape dont l'objectif était de prédire avec une grande confiance de nouvelles origines de réplication. Basée sur la détection de sauts de grande amplitude, cette méthode ne permet pas de détecter les origines se situant dans un environnement génique défavorable (telles les situations convergeantes ($\rightarrow\leftarrow$)) en raison des comportements antagonistes des biais dûs à la réplication et à la transcription (dans de telles situations, le mécanisme de transcription pourrait engendrer un saut descendant dans le signal biais qui contrebalancerait le saut ascendant dû au mécanisme de réplication). De plus, cette méthode n'est efficace que s'il existe une nette séparation entre la taille caractéristique des domaines de réplication et celle des gènes, ce qui ne semble pas être le cas dans les régions à forte concentration en *GC* du génome humain (et plus généralement des mammifères). Ainsi, les 486 origines de réplication prédites ne concernent que des régions du génome de l'homme où le pourcentage en *GC* est inférieur à 42%.

Chapitre 5

Modélisation de la réplication chez les mammifères : mise au point d'une méthodologie de prédiction des origines de réplication

Contrairement aux profils en forme de créneau induits par la transcription (*cf.* figure 4.2), nous ne possédons pas de modèle permettant d'expliquer la forme en toit d'usine caractéristique des profils de biais induit par les mécanismes de réplication chez l'homme. Dans cette section, nous allons proposer un modèle de la réplication chez les mammifères qui suppose que les origines de réplication sont bien positionnées, alors que les terminaisons sont aléatoires. Nous utiliserons ensuite les prédictions de ce modèle pour développer une nouvelle méthodologie plus performante de détection des origines de réplication dans le génome humain qui nous permettra de multiplier par plus de deux le nombre d'origines prédites. Enfin, nous proposerons une dernière amélioration de notre méthodologie qui permettra de séparer les biais dûs à la transcription et à la réplication.

Nous renvoyons le lecteur aux références [79, 369] où les résultats concernant le modèle de réplication chez les mammifères ont été publiés. La nouvelle méthodologie de détection

des origines de réplication après séparation des biais de transcription et de réplication est en cours de rédaction pour publication.

5.1 Un modèle pour la réplication chez les mammifères générant des profils de biais en forme de « toit d'usine »

Au chapitre précédent, nous avons montré qu'il existe un biais de composition dû à la réplication chez les mammifères. Dans cette section, nous allons nous attacher à généraliser le modèle de réplicon introduit pour les eubactéries (figure 4.4) afin de reproduire et d'interpréter les profils de biais en forme de toit d'usine observés chez les mammifères (*cf.* figure 4.11). Nous allons ainsi proposer un « modèle réplicon pour les mammifères » dans lequel les origines de réplication sont fixes, alors que les terminaisons sont aléatoirement et uniformément réparties.

Modélisation de la réplication chez les mammifères

Si les sites d'initiation et de terminaison de la réplication étaient toujours positionnés au même endroit dans la séquence d'un cycle cellulaire à l'autre, les asymétries de composition engendrées par le mécanisme de réplication devraient se manifester par un profil en forme de créneau dans le signal biais, comme cela est observé chez les bactéries, entre deux origines de réplication actives (figure 4.4). Pour l'homme, ce n'est visiblement pas le cas. Toutefois, les terminaisons ne sont pas fixes chez tous les organismes : si la terminaison de réplication a été expérimentalement identifiée en des sites spécifiques chez *Schizosaccharomyces pombe*, elle semble être aléatoire entre les origines actives chez *Saccharomyces cerevisiae* et dans des extraits d'oeuf du *Xénopus laevis* [247, 337]. Il est donc possible que cette caractéristique s'applique aux cellules des lignées germinales des mammifères. Nous ne connaissons pas de résultat expérimental concernant les positions des terminaisons de la réplication chez l'homme. Il est donc raisonnable de penser qu'il n'existe pas de position spécifique aux sites de terminaison.

Le modèle que nous proposons [79, 369] pour expliquer le profil de biais en forme de toit d'usine observé chez les mammifères repose sur les deux hypothèses suivantes. Tout d'abord les positions des origines de réplication dans les lignées germinales sont supposées fixes et bien déterminées d'un cycle à l'autre. Par contre, les sites de terminaison sont positionnés au hasard à chaque cycle, avec une probabilité uniforme le long de la séquence entre deux origines voisines.

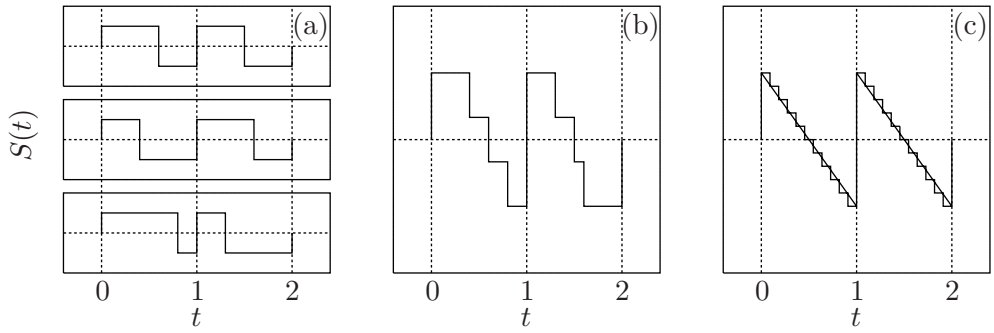

FIG. 5.1 – *Illustration du modèle de réplication chez les mammifères. Le signal biais* $S = S_{TA} + S_{GC}$ *est schématiquement représenté avec trois origines de réplication en* $t = 0$, $t = 1$ *et* $t = 2$. *Lors de chaque cycle cellulaire, la réplication entre deux origines introduit un biais de composition visible sur le signal par l'apparition de deux profils en forme de créneau dont les moyennes respectives sont de signe opposé (a). D'un cycle cellulaire à l'autre, ces profils se cumulent (b) et si les terminaisons (à savoir les sauts descendants dans le profil en creneaux) sont aléatoirement et uniformément réparties entre deux origines voisines, on obtient à la limite un profil linéairement décroissant (c).*

Durant un cycle cellulaire, on suppose qu'une « asymétrie » de composition constante est introduite lors du passage des complexes effectuant la réplication. Dans le signal biais, cela se traduit par l'apparition d'un profil en forme de créneau à chacun de ces cycles, le signe de biais moyen dans cette zone dépendant du sens de progression de la fourche de réplication (le sens $5' \rightarrow 3'$ donnant lieu à un biais positif et inversement). La largeur du créneau, pour un cycle donné, est bien sûr déterminée par la distance séparant l'origine de la terminaison. L'asymétrie de composition engendrée par un passage des complexes effectuant la réplication se superpose à l'asymétrie déjà présente, résultant de la réplication lors de précédents cycles cellulaires, jusqu'à constituer l'asymétrie observée. Puisque la position des terminaisons de réplication est supposée aléatoire avec une probabilité uniforme entre deux origines de réplication voisines (fixes), on voit se superposer dans le signal biais des profils en forme de créneau de tailles différentes, dépendant de la position de la terminaison lors des différents cycles cellulaires. À la limite, après un très grand nombre de cycles, le profil de biais en escalier converge vers un profil linéairement décroissant entre les deux origines voisines, expliquant ainsi l'absence de saut décroissant d'amplitude comparable à celle des sauts ascendants observés aux origines. Ce modèle est illustré dans la figure 5.1.

Suivant ce modèle, les valeurs du biais aux extrémités du profil, c'est-à-dire aux origines de réplication, sont fonction du taux d'activation de chaque origine, *i.e.* du pourcentage de cycles où elles sont actives, tandis que le coefficient angulaire de la pente décroissante est conditionné par la distance séparant les deux origines. En observant le biais à grande échelle (figure 4.11) sur l'ensemble des chromosomes humains, on constate que le taux

d'activation des origines de réplication varie peu et semble être homogène le long du génome, tandis que les distances séparant deux origines sont très variables, typiquement entre 200 kpb* et plusieurs Mpb.

Ce simple modèle permet d'expliquer les profils de biais en forme de toit d'usine observés. Remarquons qu'il est possible de raffiner ce modèle si nécessaire, par exemple en ne supposant plus que la distribution de probabilité de positionnement des terminaisons est uniforme. On pourrait par exemple imaginer que les positions possibles pour les terminaisons de réplication sont en nombre restreint ; typiquement, si deux origines sont séparées de quelques centaines de milliers de paires de base, il est possible que seulement une centaine de sites puissent jouer le rôle de terminaison au lieu des centaines de milliers du modèle. Il est aussi envisageable que les terminaisons aient une probabilité réduite de se trouver à proximité immédiate des origines. Dans ce cas, le profil caractéristique ne prendrait plus l'apparence d'une pente décroissante, mais d'une forme sigmoïde [369].

Discussion du modèle de réplication

Il y a lieu de faire quelques remarques concernant le modèle de réplication proposé ci-dessus et notamment envisager les mécanismes susceptibles de donner lieu à des terminaisons aléatoires.

Le modèle illustré dans la figure 5.1 repose implicitement sur l'hypothèse que le biais de composition introduit par chaque passage des complexes réplicatifs est constant, ce qui revient essentiellement à supposer que ceux-ci avancent à vitesse constante ou sans grande variation de celle-ci. *A priori*, rien n'interdit de supposer que les taux d'erreur et de réparation qui engendrent les asymétries de composition dépendent de la vitesse de passage des complexes. Nous ne savons d'ailleurs pas si l'amplitude des profils augmenterait ou au contraire diminuerait avec la vitesse. On pourrait ainsi être tenté d'expliquer la forme des profils observés par des variations de la vitesse des complexes effectuant la réplication. Toutefois, en imaginant que la vitesse de progression du complexe se fasse en fonction de la distance à l'origine de départ ou du temps écoulé depuis ce départ, on devrait systématiquement observer un profil anti-symétrique dans le signal biais de part et d'autre des origines. De plus, il apparaît difficile d'expliquer comment la vitesse pourrait être dépendante des distances séparant deux origines successives, alors que les pentes des profils de biais observés semblent varier en fonction de cette distance. Il faut aussi noter que les complexes effectuant la réplication sont des moteurs moléculaires consommant de l'ATP dont la progression est assez robuste, à condition de disposer de ressources

*. Ce qui correspond approximativement à 400 kpb pour les séquences natives.

suffisantes. De grandes variations de la vitesse de ces complexes semblent donc fort peu probables.

Chez les procaryotes ayant un chromosome circulaire, il existe deux possibilités exclusives de terminaison de la réplication dans un organisme. Soit la terminaison est définie par une position particulière sur la séquence où un complexe protéique se lie pour servir de point d'arrêt aux complexes effectuant la réplication, soit, en l'absence de site spécifique, la terminaison s'effectue au point de rencontre des deux fourches de réplication. Chez les mammifères, on peut s'interroger sur les mécanismes possibles pouvant engendrer des terminaisons de réplication aléatoires. Nous pouvons donner deux explications, sans qu'aucune observation ne puisse corroborer l'une comme l'autre. Il est possible que le point de terminaison soit bien la rencontre de deux complexes effectuant la réplication. Dans l'hypothèse où la vitesse de progression des complexes est quasiment constante, la distribution uniforme résulterait de la dynamique particulière des temps d'allumage des origines de réplication. Une seconde possibilité est l'existence de nombreux sites de terminaison uniformément répartis entre deux origines voisines. Un complexe protéique se lierait avec l'un d'entre eux, au hasard, pour arrêter les complexes effectuant la réplication. Avec une telle explication, il resterait à comprendre pourquoi un et un seul site par domaine de réplication serait sélectionné à chaque cycle.

5.2 Nouvelle méthodologie multi-échelle de prédiction des origines de réplication

Au chapitre précédent, une série de 486 origines de réplication putatives a été obtenue grâce à une double sélection. Premièrement, les sauts d'amplitude suffisante ont été détectés dans le signal biais. Ensuite, seules les paires de sauts entourant un profil de moyenne linéairement décroissante ont été sélectionnées. L'étude des profils ainsi observés a permis la mise au point d'un modèle de réplication chez les mammifères reposant sur des terminaisons aléatoires. L'objectif est maintenant d'augmenter substantiellement le nombre d'origines prédites. Pour ce faire, nous allons mettre au point un algorithme de prédiction d'origines de réplication se basant directement sur la détection de profils caractéristiques linéairement décroissants. Nous tenterons ensuite de séparer la composante du biais due à la réplication de celle due à la transcription.

Détection de profils linéairement décroissants dans un signal bruité

Sans prendre en compte l'origine du signal, le problème envisagé ici est la détection de profils décroissants linéairement dans un signal fortement bruité. Pour ce faire, nous allons utiliser une méthodologie basée sur la transformée en ondelettes (chapitre 2 de la première partie), avec une ondelette mère spécialement adaptée.

Dans l'identification des profils dûs à la réplication, le problème majeur consiste à s'affranchir du bruit omniprésent dans le signal biais. Si les origines précédemment sélectionnées présentent un profil typique clair, il n'en va pas de même dans de nombreuses zones du génome, où la forme en toit d'usine semble noyée dans le bruit ambiant. L'idée est de construire un algorithme de détection basé sur la transformée en ondelettes pour identifier les profils caractéristiques à grande échelle et minimiser la contribution du bruit.

La méthode que nous allons ici mettre en oeuvre ne va plus simplement consister à utiliser la transformée en ondelettes pour détecter les sauts ascendants de grande amplitude, mais plutôt à profiter de la décomposition espace-échelle fournie par celle-ci pour directement déceler, *via* un choix d'ondelette mère approprié, la présence d'un profil en forme de toit d'usine, à savoir deux sauts ascendants séparés par un comportement linéairement décroissant. Un choix naturel consiste à prendre comme ondelette mère la solution de notre modèle de réplication (figure 5.1), c'est-à-dire la fonction

$$\psi_s(t) = -t\chi_{[-1,1]}(t), \tag{5.1}$$

illustrée dans la figure 5.2. Le comportement de la transformée associée est particulièrement simple pour une fonction f présentant un profil semblable à celui de la réplication entre deux origines positionnées en r_1 et r_2 respectivement,

$$f(t) = -\theta t\chi_{[r_1,r_2]}(t). \tag{5.2}$$

En effet, à une échelle fixée, si le support de l'ondelette $\psi_s((\,\cdot\, - b)/a)$ est recouvert par celui de la fonction f, c'est-à-dire pour les valeurs b de la transformée telles que $b \in [b_1(a),b_2(a)]$, avec $b_1(a) = r_1 + a$ et $b_2(a) = r_2 - a$, on obtient

$$W_{\psi_s} f(b,a) = \int_{-1}^{1} \theta(at + b)\, t\, dt = \frac{2\theta}{3}a, \tag{5.3}$$

ce qui montre que la transformée en ondelettes en un tel point b varie linéairement en fonction de l'échelle. Ce résultat reste valable pour les fonctions du type $f(t) = (-\theta t + c)\chi_{[r_1,r_2]}(t)$, puisqu'une ondelette est par définition de moyenne nulle. Si le coefficient angulaire de la fonction est négatif, *i.e.* si $\theta > 0$, ce que nous supposerons implicitement,

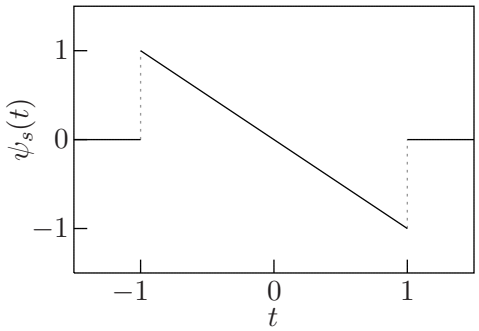

FIG. 5.2 – *Représentation de l'ondelette mère* $\psi_s(t) = -t\chi_{[-1,1]}(t)$.

on constate sans peine que la transformée en ondelettes est maximum pour ces valeurs de $b \in [b_1(a),b_2(a)]$ et que la fonction n'est pas dérivable (selon b) en $b = b_1(a)$ et $b = b_2(a)$ (*cf.* remarque 5.1).

La longueur de l'intervalle $[b_1(a),b_2(a)]$ diminue lorsque l'échelle a augmente, jusqu'à ce que celui-ci se réduise en un point, à une échelle que nous noterons a^*. À cette échelle a^*, $b_1(a^*) = b_2(a^*)$ et les supports de l'ondelette $\psi_s(\big(\,.\,-b_1(a^*)\big)/a^*)$ et de la fonction f sont identiques. Pour les échelles plus grandes, on remarque que la transformée en ondelettes prend des valeurs inférieures : $W_{\psi_s}f(b_1(a^*),a) < W_{\psi_s}f(b_1(a^*),a^*)$, si $a > a^*$ (*cf.* remarque 5.1). Ainsi, pour toutes les échelles a telles que $a \leqslant a^*$ et les valeurs de b telles que $b \in [b_1(a),b_2(a)]$, la transformée en ondelettes se comporte linéairement en fonction de l'échelle, selon l'égalité (5.3), ce qui n'est pas vrai pour les échelles $a > a^*$.

La fonction $W_{\psi_s}f(b,a)$ s'obtient explicitement lorsque f est une fonction du type (5.2).

Remarque 5.1 La transformée en ondelettes de la fonction $f = (\theta t + c)\chi_{[r_1,r_2]}(t)$ par l'ondelette ψ_s est donnée par la fonction

$$W_{\psi_s}f(b,a) = \begin{cases} 0 & \text{si } b+a \leqslant r_1, \\ -\frac{\theta b+c}{2}(1-\frac{(r_1-b)^2}{a^2}) - \frac{\theta\, a}{3}(1-\frac{(r_1-b)^3}{a^3}) & \text{si } b-a < r_1 \text{ et } b+a > r_1, \\ -2/3\,\theta\, a & \text{si } b-a \geqslant r_1 \text{ et } b+a \leqslant r_2, \\ -\frac{\theta b+c}{2}(\frac{(r_2-b)^2}{a^2}-1) - \frac{\theta\, a}{3}(\frac{(r_2-b)^3}{a^3}+1) & \text{si } b-a < r_2 \text{ et } b+a > r_2, \\ 0 & \text{si } b-a \geqslant r_2, \end{cases} \tag{5.4}$$

pour les échelles inférieures à la valeur a^*. Pour les échelles supérieures, il faut considérer le cas où c'est le support de l'ondelette qui recouvre celui de la fonction. En fait, en posant

$r_1^* = \sup\{b-a,r_1\} - b$ et $r_2^* = \inf\{b+a,r_2\} - b$, on peut écrire

$$W_{\psi_s} f(b,a) = \begin{cases} 0 & \text{si } r_1^* > r_2^* \\ -\frac{\theta b+c}{2a^2}\left((r_2^*)^2 - (r_1^*)^2\right) - \frac{\theta}{3a^2}\left((r_2^*)^3 - (r_1^*)^3\right) & \text{sinon} \end{cases} . \tag{5.5}$$

La constantes c et les extrémités du support r_1 et r_2 apparaîssent pour les positions intermédiaires où l'ondelette n'est pas entièrement recouverte par la fonction f. Pour les échelles plus grandes que la valeur a^*, l'égalité (5.5) montre que l'ondelette décroît quadratiquement en fonction de l'échelle a. □

La taille et la position du profil linéairement décroissant (5.2) peuvent être déterminés grâce à l'ondelette mère ψ_s (*cf.* égalité (5.1)) comme suit. Pour chaque position b, on repère la plus grande échelle $a_M(b)$ où le comportement de la transformée est linéaire. Ainsi, $a_M(b)$ représente l'échelle telle que $Wf(b,a) = Ca$, si $a < a_M(b)$ et $Wf(b,a) < Ca$ si $a > a_M(b)$, pour une constante C non nulle. La plus grande échelle $\sup_b\{a_M(b)\}$ ainsi atteinte définit la taille du support de f. Il suffit de remarquer que $a^* = \sup_b\{a_M(b)\}$ et qu'à cette échelle la longueur du support de l'ondelette vaut $\text{diam}[\psi_s((\,\cdot - b)/a^*)] = \text{diam}[f] = 2a^*$, par définition de ψ_s. Ainsi,

$$a^* = \frac{r_2 - r_1}{2}. \tag{5.6}$$

Le milieu du support $[r_1,r_2]$ peut ainsi être déterminé grâce à la position b^* où l'échelle maximum est atteinte. Par définition de b^*, $[\psi_s((\,\cdot - b^*)/a^*)] = [f] = [b^* - a^*,b^* + a^*]$ et donc

$$b^* = \frac{r_1 + r_2}{2}. \tag{5.7}$$

Cette méthode permet de déterminer entièrement le support $[r_1,r_2]$ de la fonction f. Finalement, le coefficient angulaire θ peut aussi être déterminé *via* la relation (5.3).

Illustrons ces résultats par un exemple concret.

Exemple 5.2 Soit la fonction $f(t) = 10^{-2}t\chi_{[-100,100]}(t)$. La transformée en ondelettes de cette fonction en utilisant l'ondelette mère ψ_s est représentée dans la figure 5.3. Elle montre clairement que les grandes valeurs de la transformée se concentrent dans le demi-plan espace-échelle au voisinage du point $(b^*,a^*) = (0,100)$, alors que les plus petites valeurs se situent au voisinage de $(-100,1)$ et $(100,1)$. Comme cela est illustré dans la figure 5.4, la longueur du plateau centré à l'origine diminue lorsque l'échelle augmente, jusqu'à ce qu'il se réduisent au seul point $(b^*,a^*) = (0,100)$. Pour une position b donnée, la transformée en ondelettes varie linéairement en fonction de l'échelle jusqu'à ce que le support de l'ondelette $\psi_s((\,\cdot - b)/a)$ ne soit plus recouvert par celui de f. Les valeurs de la transformée décroissent alors brusquement (en $1/a^2$), comme cela est illustré dans la figure 5.5. La

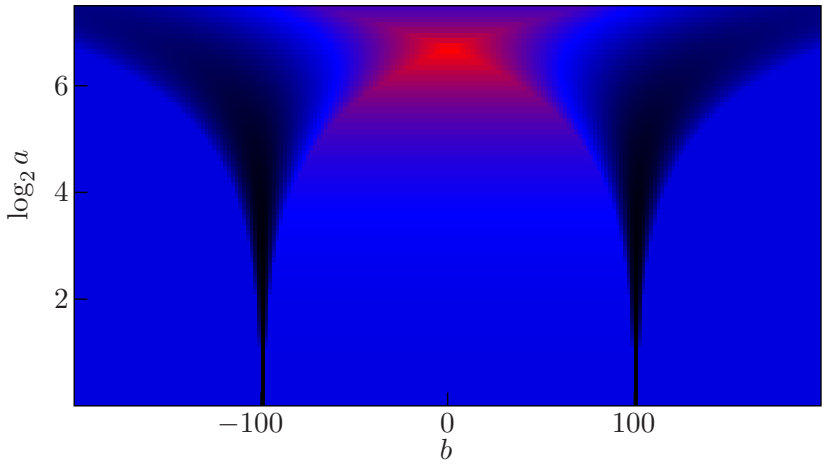

FIG. 5.3 – *Représentation espace-échelle de la transformée en ondelettes de la fonction* $f(t) = 10^{-2}t\chi_{[-100,100]}(t)$. *La transformée est maximum en* $(b^*,a^*) = (0,100)$ *et minimum en* $(-100,1)$ *et* $(100,1)$, *soit aux extrémités du support de* f. *Les couleurs utilisées vont du noir (pour les valeurs les plus faibles de la transformée en ondelettes) au rouge (pour les valeurs les plus fortes), en passant par le bleu (pour les valeurs intermédiaires).*

position b^*, correspondant au plus grand intervalle dans les échelles du type $]0,a_M(b)]$, définit le milieu du support du signal f et l'échelle $a^* = a_M(b^*)$ où ce comportement cesse d'être linéaire, permet de déterminer la longueur du support $r_2 - r_1 = 2a^*$. Ici, on obtient $b^* = 0$ et $a^* = 100$ (figure 5.5). □

Avec cette nouvelle méthodologie de détection de profil en dent de scie, nous pouvons quelque peu relaxer la contrainte sur l'amplitude des sauts ascendants qui bordent la décroissance linéaire. Le signal moyen du biais $\tilde{S}$ est calculé selon l'égalité (3.3). Un point n de ce signal sera associé à un saut acceptable comme origine de réplication si, en se basant sur le calcul du saut donné par l'expression (3.4),

$$\tilde{S}(n-20) \leqslant -\varepsilon \quad \text{et} \quad \tilde{S}(n+20) \geqslant \varepsilon, \tag{5.8}$$

où ε est un seuil ajustable. Désormais, le but est moins de détecter les sauts ascendants de grande amplitude que d'exclure les zones ne correspondant pas à une transition biais négatif – biais positif. C'est pourquoi nous avons décidé de fixer la valeur de ε à $\varepsilon = 0.03$, soit une valeur systématiquement supérieure à l'écart-type estimé du signal $\tilde{S}$ pour les chromosomes humains (pour le plus grand chromosome, il est inférieur à 0.015). Pour chacun des 22 chromosomes asexués de l'homme, nous avons ainsi construit le dictionnaire D des points n vérifiant les conditions données par les inégalités (5.8).

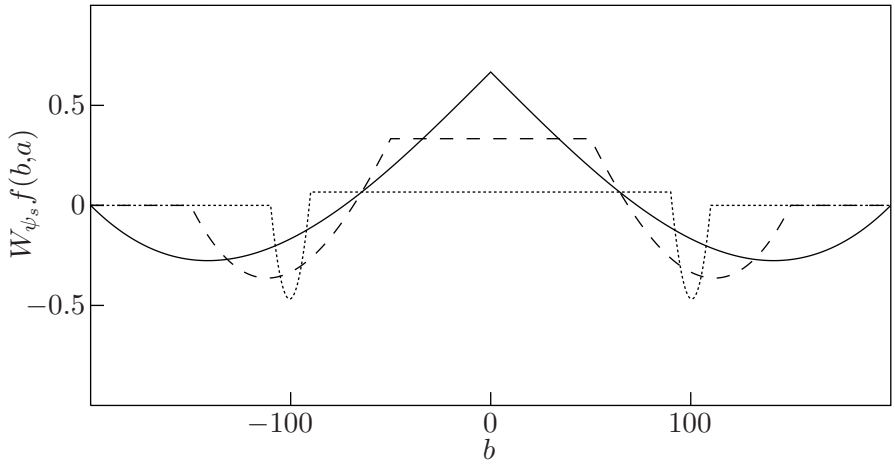

FIG. 5.4 – *La transformée en ondelettes de* $f(t) = 10^{-2} t \chi_{[-100,100]}(t)$ *pour les échelles* $a = 10$ *(pointillés),* 50 *(traits discontinus) et* 100 *(trait continu), en utilisant l'ondelette mère* ψ_s *(égalité (5.1)). L'échelle* $a = a^* = 100$ *est celle où le plateau centré en* $b = b^* = 0$ *se réduit à un point.*

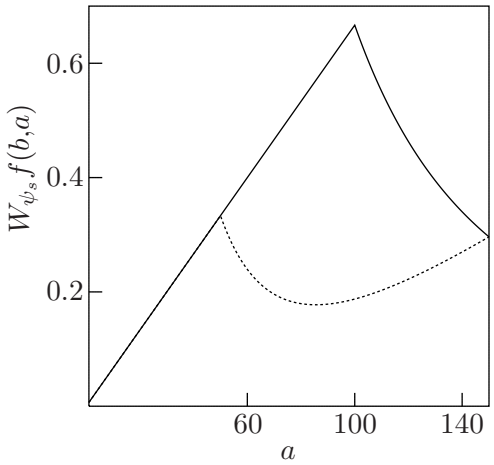

FIG. 5.5 – *La transformée en ondelettes de* $f(t) = 10^{-2} t \chi_{[-100,100]}(t)$ *en utilisant l'ondelette mère* ψ_s *(égalité (5.1)), pour les positions* $b = 50$ *(pointillés) et* $b = b^* = 0$ *(trait continu) en fonction de l'échelle. On observe un comportement linéaire puis une décroissance nette. Pour la position* $b = b^* = 0$*, ce changement de comportement a lieu à une échelle* a^* *plus élevée que pour les autres positions, ce qui permet d'affirmer que le centre du support de* f *se trouve dans cet exemple en* $b^* = 0$*. L'échelle* a^* *permet de définir la longueur* $r_2 - r_1 = 2a^* = 200$ *de ce support.*

Application test sur des profils synthétiques en forme de toit d'usine bruité

Dans cette sous-section, nous allons tester notre méthodologie sur des profils synthétiques en forme de toit d'usine bruités de la forme

$$f(t) = \sum_j f_j(t) + g(t), \tag{5.9}$$

où $f_j(t)$ est une fonction du type de celle considérée dans la sous-section précédente,

$$f_j(t) = \Big(- \theta_j\big(t - (r_j + \delta_j/2) + c_j\big)\Big)\, \chi_{[r_j, r_j+\delta_j]}, \tag{5.10}$$

avec $r_j + \delta_j \leqslant r_{j+1}$ et où $g(t)$ est une fonction inconnue jouant le rôle d'un bruit. Au voisinage du milieu du support de f_j, la transformée en ondelettes avec l'ondelette mère ψ_s (égalité (5.1)) devrait se comporter linéairement en fonction de a, suivant l'égalité (5.3). Pour chaque position b, on évalue la plus grande échelle $a_M(b)$ jusqu'à laquelle la transformée se comporte bien linéairement en fonction de l'échelle. Pour ce faire, on sélectionne l'intervalle $[a_1,a_2]$ minimisant, au sens des moindres carrés, la distance moyenne entre les points d'une droite et la fonction $Wf(b,.)$. Le paramètre a_1 est pris tel que la longueur du support de l'ondelette associée $\psi_s(\,\cdot/a_1)$ soit supérieure à 40 kpb, pour minimiser les effets du bruit et de la discrétisation de l'ondelette, et inférieure à 50 kpb, pour éviter de sélectionner une partie linéaire ne s'étendant pas aux petites échelles. Cet intervalle étant identifié, on pose $a_M(b) = a_2$.

L'étape suivante consiste, pour chaque paire de points (n_1,n_2) appartenant au dictionnaire D des sauts vérifiant la condition (5.8), à s'assurer que l'intervalle $[n_1,n_2]$ définit un profil linéairement décroissant. La taille associée à ce profil hypothétique $\delta = 2a_M((n_1+n_2)/2)$ est comparée à la taille réelle $n_2 - n_1$. Si la différence entre ces deux valeurs est inférieure à $\min\{\delta/10,30\}$ (les erreurs autorisées sont légèrement supérieures à l'erreur maximum due à la discrétisation de la transformée en ondelettes), l'intervalle est considéré comme un profil candidat et on lui associe un score égal à la distance des moindres carrés moyenne calculée à l'étape précédente. Lorsque toutes les paires ont été passées en revue, on élimine les intervalles se superposant en comparant les distances associées. Si, par exemple, les intervalles $I_1 = [n_1,n_2]$, $I_2 = [n_2,n_3]$ et $I_3 = [n_1,n_3]$ sont tous les trois des profils candidats, on retiendra soit I_1 et I_2, soit I_3 selon que la moyenne des distances associées à I_1 et I_2 est inférieure ou non à celle associée à I_3.

Cette méthode peut être raffinée comme suit. Premièrement, pour un b fixé, lorsque l'intervalle $[a_1,a_2]$ minimisant la distance a été sélectionné, on peut poser non plus $a_M(b) = a_2$, mais

$$a_M(b) = \sup\{a : Wf(b,a) > Wf(b,a_2)\}, \tag{5.11}$$

à condition que la différence entre a_M(b) (défini par l'égalité (5.11)) et a_2 soit inférieure à $a_2/10$. La raison de ce raffinement est que la valeur de l'échelle pour laquelle on passe d'un comportement linéaire en fonction de l'échelle à un comportement non-linéaire peut être mal localisée, à cause du bruit. Le pic de transition observé dans la figure 5.5 entre un comportement linéaire croissant et un comportement parabolique décroissant peut devenir une bosse dont le maximum est plus ou moins bien défini. Enfin, la modification la plus importante consiste à améliorer la précision sur la détermination de l'intervalle $[b^* - a^*, b^* + a^*]$ correspondant au support d'un motif du profil en toit d'usine. À l'échelle $a_M(b^*)$, on repère d'abord les minima de la transformée en ondelettes situés à une distance inférieure à $a_M(b^*)$, *i.e.* la moitié de la longueur estimée de l'intervalle I que l'on cherche à déterminer, de b^* (qui correspond au milieu estimé de I). On calcule alors les lignes d'extrema de la transformée en ondelettes en ne retenant que celles passant par un minimum précédemment sélectionné. Plutôt que de définir la taille du support comme étant $r = 2a_M(b^*)$, on sélectionne parmi les lignes de minima les deux lignes de part et d'autre du segment $\{(b^*, a) : 40 < a < a_M(b^*)\}$ pointant, à l'échelle de 40 kpb, sur le saut ascendant de plus grande amplitude. Si b_1 et b_2 désignent ces positions, on pose $b^* = (b_1 + b_2)/2$ et $a_M(b^*) = (b_2 - b_1)/2$.

Pour tester cette méthode de détection, nous avons créé un signal du type $f(t) = \sum_j f_j(t) + g(t)$, où les fonctions f_j sont définies par l'égalité (5.10) et où $g(t)$ est un bruit blanc. Pour simuler la réplication, nous avons posé $c_j = 0.07$ et choisi la variance du bruit égale à 0.08. Les longueurs des supports des fonctions f_j ont été tirées au hasard selon une loi normale de moyenne 550 kpb et d'écart-type 300 kpb. Des fonctions créneaux simulant le biais de transcription ont également été surajoutées avec une longueur moyenne de 30 kpb et un écart-type de 50 kpb. Une réalisation d'un tel signal est illustrée dans la figure 5.6 (a). La transformée en ondelettes, en utilisant l'ondelette mère ψ_s (égalité (5.1)), est représentée dans la même figure 5.6. On remarque directement que les caractéristiques principales de la transformée en ondelettes d'un signal f_j, illustrées dans les figures 5.3, 5.4 et 5.5, se retrouvent dans la figure 5.6 : un maximum de la transformée se situe en général dans le demi-plan espace-échelle au voisinage d'un point du type $(r_j + \delta_j/2, a^*)$, c'est-à-dire qu'il est positionné au centre du support de f_j et à l'échelle a^* correspondant à la moitié de la taille du support. Remarquons que la présence du bruit se manifeste surtout aux petites échelles, où cela perturbe la détection d'éventuels motifs f_j de petite taille.

Sur les 2201 motifs de type f_j présents dans notre échantillon statistique du signal synthétique, 1997 ont été retrouvés par notre méthodologie, soit plus de 90%. Concernant la taille des profils ainsi détectés, il y a, comme on pouvait s'y attendre, une disparité

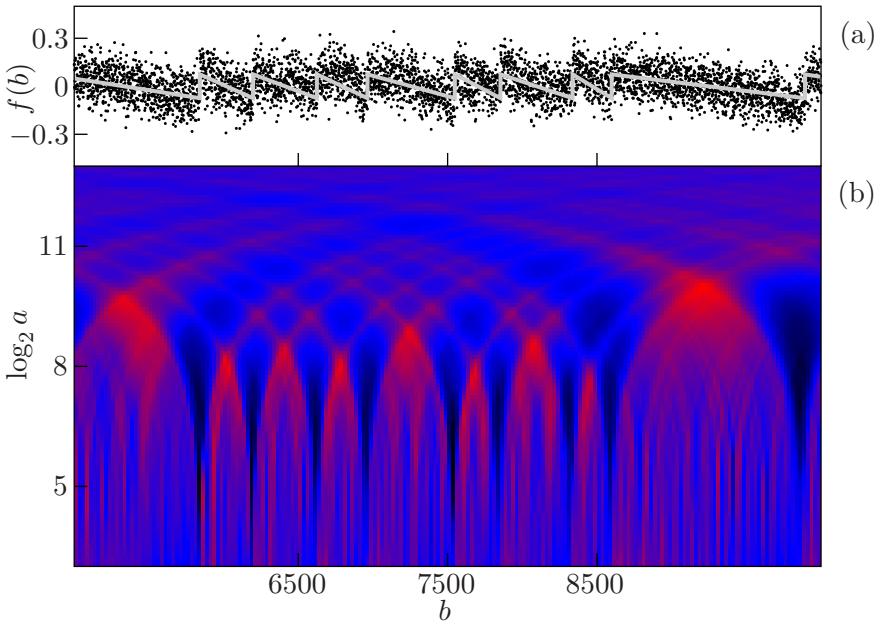

FIG. 5.6 – *(a) Représentation d'un signal synthétique simulant le signal biais en forme de toit d'usine (voir texte). Le signal profil non bruité (lignes grises) est superposé au signal bruité (points noirs). (b) Transformée en ondelettes du signal bruité synthétique représenté en (a) en utilisant l'ondelette mère ψ_s (figure 5.2). On distingue clairement l'existence de domaines dans le demi-plan espace-échelle caractéristiques de la représentation en ondelettes de fonctions du type f_j (cf. figure 5.3). La transformée en ondelettes avec l'ondelette mère adaptée permet la détection multi-échelle des profils en toit d'usine via une détection de forme dans le demi-plan espace-échelle. Le code couleur est identique à celui utilisé dans la figure 5.3.*

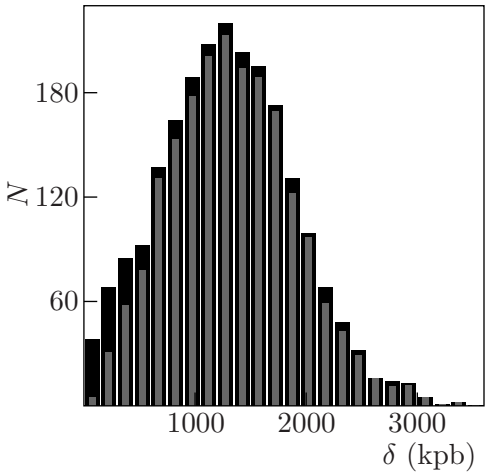

FIG. 5.7 – *Histogramme des tailles des domaines détectés (en gris) dans le signal synthétique en forme de toit d'usine (illustré dans la figure 5.6 (a)), comparativement à l'histogramme des tailles des domaines effectivement générés (en noir). Les tailles de ces domaines ont été tirées selon une loi gaussienne ; des fonctions en forme de créneau et un bruit blanc ont été additionnés au signal, comme décrit dans le texte.*

nette, illustrée dans la figure 5.7. Moins de 70% des motifs de longueur inférieure à 450 kpb sont détectés, pourcentage qui tombe à 13% si l'on ne considère que les motifs dont la longueur du support est inférieure à 150 kpb. À ces échelles, les maxima sont noyés au milieu de ceux induits par le bruit, omniprésent ; celui-ci peut même conduire à la « détection » de profils non existants (8 faux positifs dont 6 d'une taille inférieure à 150 kpb). Signalons aussi que 7% des profils non-détectés ont étés concaténés : là où il y avait deux profils successifs, l'algorithme a considéré qu'il ne s'agissait que d'un seul. Un autre point qu'il nous a semblé nécessaire d'améliorer est la précision avec laquelle sont détectés les extrémités des supports, l'erreur étant en moyenne légèrement supérieure à 15 kpb. Vu le caractère irrégulier de l'ondelette mère utilisée (*cf.* figure 5.2), il est difficilement envisageable d'acquérir une plus grande précision en utilisant uniquement la fonction ψ_s.

Le problème d'imprécision sur la position des sauts ascendants aux bords des domaines peut être amélioré comme suit. Une fois que les extrémités des supports $\{n_k\}$ ont étés estimés grâce à l'ondelette mère ψ_s, la transformée en ondelettes du signal biais est effectuée à nouveau en utilisant cette fois l'ondelette mère dérivée première de la gaussienne (*cf.* figure 2.1 (a) de la première partie). Pour chaque extrémité n_k, le maximum présent à l'échelle correspondant à la taille caractéristique de 200 kpb le plus proche de n_k est sélectionné. On chaîne ensuite les maxima du module à travers les échelles pour ne garder que les lignes passant par un point du type (n_k,200). La position pointée par chaque ligne à l'échelle 40 kpb définit une nouvelle position n'_k, pointant plus précisément sur le saut que ne le faisait n_k. Avec cette modification, l'erreur moyenne sur la localisation des

extrémités des domaines est réduite à 5 kpb.

Nous pouvons supposer que la méthodologie ainsi développée permet de détecter efficacement les profils associés au mécanisme de réplication dans le signal biais avec une précision de l'ordre de 5 kpb, pour autant que la taille du domaine soit supérieure à 200 kpb.

Détections des origines de réplication dans le génome humain

L'application de notre méthodologie de détection de profil linéairement décroissant bordé par deux sauts ascendants de grande amplitude au signal biais $S = S_{TA} + S_{GC}$ des 22 chromosomes asexués de l'homme va nous permettre de sélectionner de nouveaux domaines de réplication. Parmi tous les domaines ainsi détectés, nous allons ensuite uniquement retenir les domaines où les biais de transcription et de réplication pourront être séparés. Nous obtiendrons ainsi une banque de 1153 orgines putatives dont les domaines associés dans le signal biais S recouvriront 40% du génome humain.

La première étape de notre méthodologie de prédiction des origines de réplication dans le génome humain est l'application de la méthode multi-échelle présentée dans la section précédente pour détecter les profils en toit d'usine aux signaux biais S des 22 chromosomes asexués de l'homme. Cette étape est illustrée dans la figure 5.8.

Ensuite, parmi tous les domaines ainsi détectés, seuls ceux vérifiant notre modèle pour la réplication (*cf.* figure 5.1) sont sélectionnés. Selon ce modèle, la réplication induit un profil linéairement décroissant entre deux origines de réplication dans le signal biais S. Ainsi, le fait de retrancher ce comportement linéaire à un domaine délimité par deux origines de réplication est suffisant pour éliminer la composante du biais due à la réplication entre ces deux origines. Une fois cette composante ôtée du signal biais S, seule la composante du biais due aux mécanismes de transcription devrait subsister. Le profil de biais restant devrait donc être en forme de créneaux, les créneaux ascendants (positifs) correspondant aux gènes situés sur le brin Watson et les créneaux descendants (négatifs) correspondant aux gènes situés sur le brin Crick. Le comportement linéaire induit par les mécanismes de la réplication est donc celui qui une fois retranché du signal biais S donne lieu à un profil en forme de créneaux, les gènes étant décalés verticalement (positivement pour les gènes sens sur le brin Watson et négativement pour les gènes anti-sens sur le brin Crick) par rapport aux régions intergéniques de moyenne nulle.

Puisque la forme en toit d'usine doit être symétrique dans le signal biais, le profil linéai-

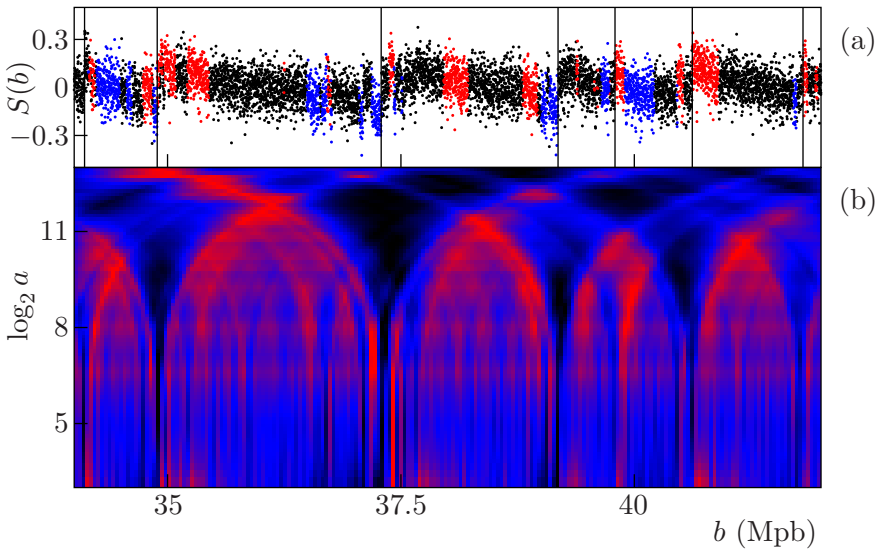

FIG. 5.8 – *Application de la méthodologie développée dans la section précédente aux chromosomes asexués de l'homme. (a) Le signal biais représenté est issu du chromosome 12. Les gènes sens situés sur le brin Watson sont colorés en rouge et les gènes anti-sens sur le brin Crick en bleu. (b) La transformée en ondelettes du signal biais avec l'ondelette mère ψ_s (égalité (5.1)), permettant de délimiter les bords des domaines caractéristiques du signal biais. Le code des couleurs est identique à celui utilisé dans la figure 5.3.*

rement décroissant doit être nul au milieu du domaine, c'est-à-dire au point équidistant des deux origines bordant le domaine. La composante linéaire induite par les mécanismes de réplication est donc seulement déterminée par son coefficient angulaire θ. Ce coefficient peut être estimé comme suit. Comme précédemment, la taille de chaque profil est renormalisée de telle sorte qu'elle soit égale à l'unité. Suivant notre modèle de la réplication (*cf.* section 5.1), les points du signal biais d'un domaine correspondant à des zones intergéniques doivent se placer le long d'une droite intersectant l'axe des abscisses au point $t = 1/2$, *i.e.* le long d'une droite du type $-\theta(t - 1/2)$. Il n'en va pas de même pour les points correspondant aux zones géniques sens + (resp. sens −), puisque le biais de transcription peut décaler les points vers des valeurs plus grandes (resp. plus petites). Ces points doivent donc se placer sur une droite du type $-\theta t + c_1$ (resp. $-\theta t + c_2$), les constantes c_1 et c_2 dépendant du gène. Si T_i désigne les abscisses des points correspondant à des zones intergéniques et T_s (resp. T_a) celles des points correspondant à des zones de gènes sens (resp. anti-sens), le signal biais d'un profil devrait être approximé par une fonction de la forme

$$\omega(t;\theta,c_1,c_2) = -\theta(t-\frac{1}{2})\chi_{T_i}(t) + \big(-\theta t + c_1(t)\big)\chi_{T_s}(t) + \big(-\theta t + c_2(t)\big)\chi_{T_a}(t), \quad (5.12)$$

avec $t \in [0,1]$. Cette fonction est définie par un coefficient angulaire $-\theta$ unique, définissant le comportement linéaire induit par les mécanismes de réplication. Le coefficient θ associé

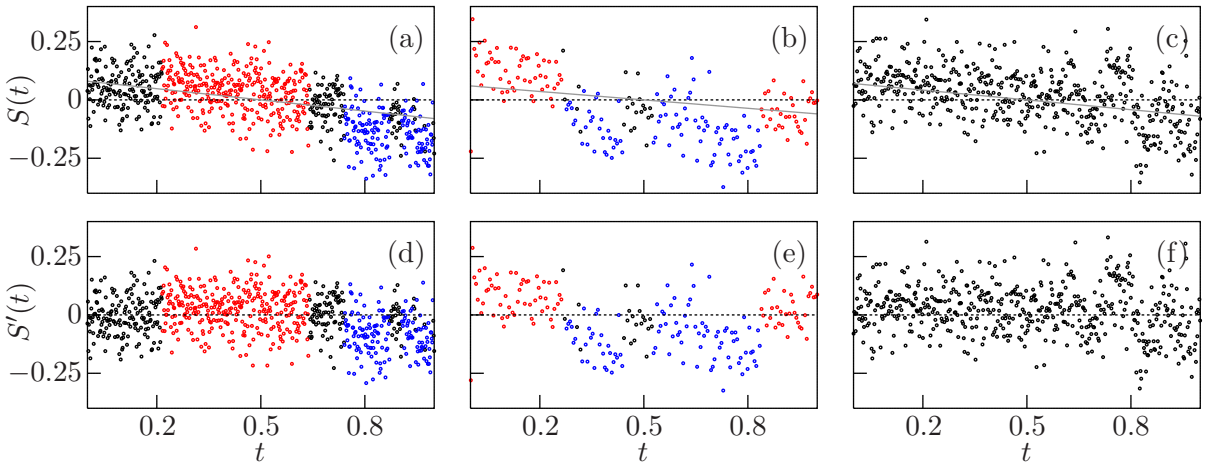

FIG. 5.9 – *Illustration de la méthode utilisée pour séparer le biais de réplication du biais de transcription dans les signax biais S des 22 chromosomes asexués du génome humain. (a) Exemple d'un profil comportant des zones géniques et intergéniques. (b) Exemple d'un profil avec peu de zones intergéniques. (c) Exemple d'un profil intergénique. Pour chaque domaine ((a), (b) et (c) respectivement), le coefficient angulaire $-\theta$ est estimé, en minimisant la distance (par un test de chi-deux) entre S et la fonction $\omega(t,\omega,c_1,c_2)$ (égalité 5.12). Le profil redressé ((d), (e) et (f) respectivement) caractérise un domaine désormais dépourvu de biais induit par la réplication. Le signal redressé S' représente donc le seul biais induit par les mécanismes de transcription, comme peuvent en témoigner les profils en forme de créneaux obtenus dans (d), (e) et (f).*

à un domaine peut ainsi être déterminé en minimisant la distance (par un test de type chi-deux portant sur θ et les constantes c_1 et c_2 de chaque gène) entre le signal biais associé et la fonction $\omega(t;\theta,c_1,c_2)$. Une fois le coefficient angulaire estimé, le biais au bord du profil dû à la réplication peut en être déduit : $S = \theta/2$. Le signal redressé $S'(t) = S(t) + \theta(t - 1/2)$ ne présente donc plus de biais réplicationnel. Cette méthode est illustrée dans la figure 5.9.

Appliquée au génome humain, cette méthodologie permet de sélectionner 1153 origines possibles, délimitant 759 domaines. Le nombre d'origines prédites est donc multiplié par un facteur supérieur à deux par rapport à la précédente banque d'origines putatives (chapitre 4). Les nouveaux domaines couvrent près de 41% du génome, leur taille moyenne étant de 747 kpb dans les séquences masquées, soit 1271 kpb dans les séquences natives. L'histogramme des tailles trouvées est représenté dans la figure 5.10. Notre méthodologie permet d'estimer une valeur du biais de réplication moyen égale à $\overline{S} = 0.06$. L'histogramme des valeurs du biais de réplication pour chaque domaine sélectionné, vérifiant notre modèle pour la réplication, est donné dans la figure 5.11. Remarquons que cet histogramme est en très bon accord avec le biais de réplication estimé dans les régions intergéniques autour des origines de réplication connues expérimentalement (table 4.6).

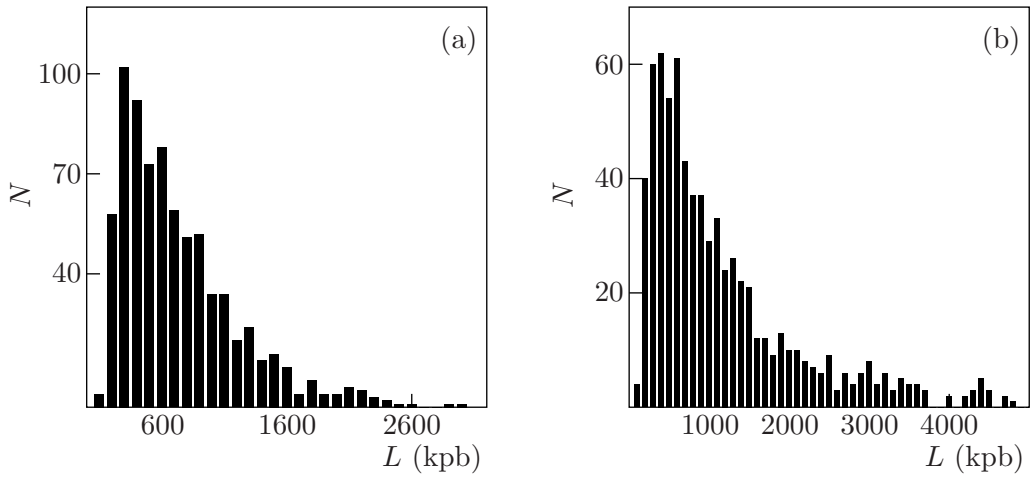

FIG. 5.10 – *Histogrammes des tailles (en kpb) des domaines de réplication détectés. En (a) sont représentées les tailles sans prendre en compte les séquences répétées. Pour les séquences natives, on obtient en (b) un histogramme présentant le même profil, mais où les tailles ont été multipliées par un facteur voisin de deux.*

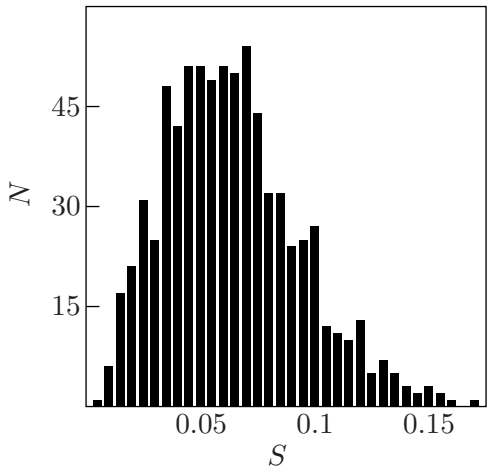

FIG. 5.11 – *Histogramme des biais de réplication mesurés relativement aux domaines détectés dans le génome humain (chromosomes asexués).*

Annexe A

La bijection de Cantor et la courbe de Peano

Les articles de Cantor [91] et Peano [311] permettent de mettre en lumière les limites de la dimension topologique, chacun amenant une question fondamentale sur la notion de dimension. En effet, suivant l'article de Cantor, il existe une bijection entre l'intervalle et le carré unité, laissant à penser qu'un plan n'est pas plus riche qu'une droite. L'article de Peano quant à lui exhibe une courbe continue remplissant ce même carré, remettant en cause l'idée que la dimension représente le plus petit nombre de paramètres nécessaire pour décrire le mouvement d'un point matériel.

A.1 La bijection de Cantor

Dans un article fondateur [91], Cantor introduit la notion de puissance et esquisse sa philosophie des ensembles . Il y introduit aussi ce que l'on appelle la *bijection de Cantor*, qui établit une bijection entre l'intervalle et le carré unité, que nous allons introduire ici.

Préliminaires

Il nous faut commencer par présenter quelques éléments (se trouvant dans l'article de Cantor) sur l'infini et le fait d'être dénombrable. Le résultat est qu'il existe une bijection entre l'intervalle unité et l'ensemble des irrationnels de cet intervalle.

Avant de donner la définition de l'application de Cantor, il nous faut faire quelques rappels. On peut ordonner les nombres rationnels de l'intervalle [0,1] de la manière suivante : tout nombre rationnel peut s'écrire sous forme irréductible p/q. On pose alors $N = p + q$. À chaque nombre rationnel p/q peut donc être associée un entier N unique et réciproquement, à chaque entier N correspond un nombre fini de rationnels p/q. On ordonne alors les rationnels par N croissant, les rationnels associés au même N étant eux-mêmes rangés par ordre croissant.

Nous identifierons l'ensemble des irrationnels de l'intervalle unité par $E = [0,1] \setminus \mathbb{Q}$. Montrons qu'il existe une bijection entre E et [0,1]. Soit $\{x_j\}_{j>0}$ une suite de nombres irrationnels distincts entre eux*. Si l'on note X l'ensemble des valeurs prises par la suite $\{x_j\}_{j>0}$ et H l'ensemble des irrationnels de l'intervalle unités distincts des x_j,

$$H = \{x \in [0,1] : x \notin \mathbb{Q}, x \neq x_j \ \forall j > 0\} = [0,1] \setminus (\mathbb{Q} \cup X), \tag{A.1}$$

on a

$$[0,1] = H \cup X \cup (\mathbb{Q} \cap [0,1]), \tag{A.2}$$

et

$$E = H \cup X_p \cup X_i, \tag{A.3}$$

où X_p (resp. X_i) désigne l'ensemble des valeurs prises par la suite irrationnelle $\{x_{2j}\}_{j>0}$ (resp. $\{x_{2j-1}\}_{j>0}$). Comme il existe une bijection entre X et X_p d'une part et entre X_i et l'ensemble des rationnels de l'intervalle unité d'autre part, les égalités (A.2) et (A.3) permettent d'affirmer la proposition suivante,

Proposition A.1 *Il existe une bijection entre* [0,1] *et* $[0,1] \setminus \mathbb{Q}$.

Définition

Fort du résultat précédent, la construction de la bijection de Cantor devient alors simple. La présente section est entièrement basée sur l'article original, excepté la proposition sur la continuité.

*. Par exemple $x_j = \sqrt{2}/2^j$.

Rappelons d'abord [118, 223, 306] que tout nombre irrationnel t peut être identifié à une suite infinie $\{a_j\}_{j>0}$ de nombres naturels, cette suite déterminant le développement en fraction continue ordinaire de t. Inversement, une fraction continue ordinaire est toujours convergente et la limite est un nombre irrationnel si la suite associée est infinie. Si t est un nombre irrationnel de l'intervalle [0,1], nous écrirons $t = [0,a_1,a_2,\ldots]$. On définit l'application de Cantor pour tout nombre irrationnel

$$t = [0,a_1,a_2,\ldots]$$

de l'intervalle [0,1], en lui associant le couple d'irrationnels $(x(t),y(t))$ défini de la manière suivante,

$$x(t) = [0,a_1,a_3,\ldots,a_{2j-1},\ldots], \qquad y(t) = [0,a_2,a_4,\ldots,a_{2j},\ldots]. \tag{A.4}$$

Inversement, à tout couple d'irrationnels (x,y) de $[0,1]^2$, on peut associer l'irrationnel $t(x,y)$ de [0,1] en posant, si

$$x = [0,a_{1,1},a_{1,2},\ldots,a_{1,j},\ldots], \qquad y = [0,a_{2,1},a_{2,2},\ldots,a_{2,j},\ldots], \tag{A.5}$$

$$t(x,y) = [a'_1,a'_2,\ldots,a'_j,\ldots], \tag{A.6}$$

avec

$$a'_{2(j-1)+k} = a_{k,j} \qquad (j > 0, k \in \{1,2\}). \tag{A.7}$$

Il existe donc une bijection entre l'ensemble des irrationnels de l'intervalle unité et l'ensemble des irrationnels du carré unité.

La proposition A.1 nous permet d'étendre l'application de Cantor à l'intervalle [0,1] tout entier.

Théorème A.2 *Il existe une bijection entre l'intervalle unité et le carré unité.*

Donnons dès à présent une propriété fondamentale,

Proposition A.3 *L'application de Cantor n'est pas continue.*

Preuve. Soit la suite $\{t_j\}_{j>0}$ de terme général

$$t_j = \frac{1}{\sqrt{2} + 10^j}.$$

On peut supposer que t_j est un élément de H (où H est défini par la relation (A.1)) quelque soit j. Ainsi, si $\mathfrak{C}$ désigne la bijection de Cantor, on trouve directement

$$\mathfrak{C}(t_j) = (t_j,\sqrt{2} - 1).$$

Si maintenant, on considère la suite $\{t'_j\}_{j>0}$ de terme général

$$t'_j = \frac{1}{1/(2\sqrt{2}) + 10^j},$$

on trouve, en supposant que $t'_j \in H$,

$$\mathfrak{C}(t'_j) = (1/(\phi + 10^j), \sqrt{2/10}),$$

où $\phi = (1-\sqrt{5})/2$ est l'inverse du nombre d'or. Les suites t_j et t'_j convergent vers la même limite, mais la deuxième composante de $\mathfrak{C}(t_j)$ vaut $\sqrt{2}-1$ alors que celle de $\mathfrak{C}(t'_j)$ vaut $\sqrt{2/10}$. □

Formulons maintenant quelques remarques. Ce résultat peut se généraliser à $\mathbb{R}^n$.

Remarque A.4 On peut sans difficulté montrer qu'il existe une bijection entre l'intervalle unité et $[0,1]^n$. Il suffit de remplacer les égalités (A.5), (A.6) et (A.7) par les suivantes,

$$\begin{aligned} e_1 &= [0,a_{1,1},a_{1,2},\ldots,a_{1,j},\ldots], \\ &\vdots \\ e_n &= [0,a_{n,1},a_{n,2},\ldots,a_{n,j},\ldots], \end{aligned} \tag{A.8}$$

et de définir

$$t(e_1,\ldots,e_n) = [a'_1,a'_2,\ldots,a'_j,\ldots], \tag{A.9}$$

avec

$$a'_{n(j-1)+k} = a_{k,j} \qquad (j>0, k\in\{1,\ldots,n\}), \tag{A.10}$$

puis d'appliquer la proposition A.1. □

On ne peut obtenir une telle application si on remplace le développement en fraction continue d'un nombre par son expression décimale.

Remarque A.5 Si l'on essaie d'appliquer directement au réel

$$t = 0.d_1d_2\cdots$$

des relations du type[*]

$$x(t) = 0.d_1d_3\cdots d_{2j-1}\cdots, \qquad y(t) = 0.d_2d_4\cdots d_{2j}\cdots$$

on ne définit pas une fonction du fait de la multiplicité des représentations de t : si l'on prend $t_1 = 0.199\cdots$ et $t_2 = 0.2$, on obtient $y(t_1) = 0.99\cdots$ et $y(t_2) = 0$, alors que $t_1 = t_2$. □

*. Voir l'annexe A.2 pour les détails sur la notation.

A.2 La courbe de Peano

La courbe dite de Peano constitue le premier exemple de courbe continue remplissant le carré unité et fait partie des exemples qui mirent à la lumière du jour les lacunes de la notion de dimension telle qu'envisagée au début du siècle passé. La présente section est entièrement basée sur l'article original de Peano [311].

Définition

La définition de la *courbe de Peano* peut sembler peu naturelle mais elle présente l'avantage d'être simple, comme en témoignent les résultats trivialement déductibles de la construction.

Définissons d'abord l'application K qui à un naturel j inférieur ou égal à 2 associe $Kj = 2 - j$. On pose $K^0 j = j$ et $K^l j = KK^{l-1} j$ $(l > 0)$. La courbe de Peano se construit comme suit. En base de numération 3, base dans laquelle nous nous placerons dans toute cette section, tout nombre t de l'intervalle [0,1] peut être défini par une suite $\{d_j\}_{j\in\mathbb{N}_0}$ et représenté* sous la forme $t = 0.d_1 d_2 d_3 \cdots$, où $d_j \in \{0,1,2\}$, $j \in \mathbb{N}_0$. Cette représentation n'est éventuellement pas unique. À ce nombre t, on associe le couple de points $(x(t),y(t))$ de la manière suivante :

$$x(t) = 0.d_1^{(x)} d_2^{(x)} d_3^{(x)} \cdots , \qquad y(t) = 0.d_1^{(y)} d_2^{(y)} d_3^{(y)} \cdots ,$$

avec

$$d_j^{(x)} = K^{\sum_{k=1}^{j-1} d_{2k}} d_{2j-1}, \qquad d_j^{(y)} = K^{\sum_{k=1}^{j} d_{2k-1}} d_{2j}.$$

On constate que le j-ième chiffre de $x(t)$, $d_j^{(x)}$ est égal soit à d_{2j-1}, soit à $2 - d_{2j-1}$ selon que la somme des chiffres d'indice pair $\sum_{k=1}^{j-1} d_{2k}$ qui le précède est paire ou impaire. Il en va de même pour $y(t)$. Puisque la parité est conservée,

$$d_j^{(x)} = K^{\sum_{k=1}^{j-1} d_k^{(y)}} d_{2j-1}, \qquad d_j^{(y)} = K^{\sum_{k=1}^{j} d_k^{(x)}} d_{2j}$$

et

$$d_{2j-1} = K^{\sum_{k=1}^{j-1} d_k^{(y)}} d_j^{(x)}, \qquad d_{2j} = K^{\sum_{k=1}^{j} d_k^{(x)}} d_j^{(y)}.$$

Cette construction définit deux fonctions de t (à savoir x et y). En effet, on montre aisément que $x(t)$ et $y(t)$ ne dépendent pas de la représentation choisie pour le nombre t.

*. Nous amalgamons donc un nombre réel avec sa représentation. Cet abus de langage est conforté par la remarque A.6

Remarque A.6 Pour que la définition soit licite, il est à montrer que si $t \in [0,1]$ possède deux représentations différentes en base 3, explicitement $t = 0.d_1d_2d_3\cdots$ et $t = 0.d'_1d'_2d'_3\cdots$, avec $d_j \neq d'_j$ pour au moins un j, les couples correspondants $(x(t),y(t))$ et $(x'(t),y'(t))$ définis respectivement à partir de $\{d_j\}_{j\in\mathbb{N}_0}$ et $\{d'_j\}_{j\in\mathbb{N}_0}$ sont égaux (mais ne possèdent éventuellement pas la même représentation). Autrement dit $(x(t),y(t))$ dépend bien de t et non de sa représentation. De fait, supposons d'abord que les deux formes de t soient $t = 0.d_1d_2\ldots d_{2j-1}d_{2j}222\ldots$ avec $d_{2j} \neq 2$ et $t = 0.d_1d_2\ldots d_{2j-1}d'_{2j}000\cdots$, avec $d'_{2j} = d_{2j} + 1$. On a alors

$$\sum_{k=1}^{j-1} d_{2k} + d'_{2j} = \sum_{k=1}^{j} d_{2k} + 1$$

et ainsi les éléments $d_l^{(x)}$ de $x(t)$, lorsque $l \geqslant j+1$ sont donnés par $K^{\sum_{k=1}^{j} d_{2k}}2$, quelque soit la représentation de t choisie. Les deux autres formes possibles pour t sont $t = 0.d_1d_2\ldots d_{2j-1}222\ldots$ avec $d_{2j-1} \neq 2$ et $t = 0.d_1d_2\ldots d'_{2j-1}000\cdots$, avec $d'_{2j-1} = d_{2j-1}+1$. Il suffit de constater que si l'on prend $t_1 = 0.d_{2j-1}222\cdots$ et $t_2 = 0.d'_{2j-1}000\cdots$, les nombres $x(t_1)$ et $x(t_2)$ correspondants ont même valeur (mais pas la même représentation). Le même argument s'applique pour $y(t)$. □

Propriétés

Les principales caractéristiques de cette courbe sont sa continuité, sa surjectivité et sa non-injectivité.

À partir du paragraphe précédent, on peut affirmer que $x(t)$ et $y(t)$ sont des fonctions de t. Ces fonctions sont de plus continues. De fait, si la suite $\{t_n\}_{n\in\mathbb{N}_0}$ tend vers la limite t_0, pour tout $j \in \mathbb{N}$, il existe un nombre n suffisamment grand tel que les $2j$ premiers chiffres du développement de $t_{n'}$ et de l'une des représentations de t_0 coïncident quelque soit $n' \geqslant n$. Par définition, les j premiers chiffres du développement de $x(t_{n'})$ (resp. $y(t_{n'})$) coïncident avec les j premiers chiffrent du développement de $x(t_0)$ (resp. $y(t_0)$).

On constate aisément que l'application qui à tout point t de l'intervalle $[0,1]$ fait correspondre le point $(x(t),y(t))$ du carré $[0,1]^2$ par la méthode définie plus haut est surjective. On peut donc construire une courbe continue « remplissant » la carré unité. Cette dernière n'est cependant pas injective. Pour s'en convaincre, il suffit de prendre un point de l'image dont l'une des coordonnées admet deux représentations en base trois. À ce point du carré correspondent alors au moins deux points de l'intervalle unité (voire quatre si chacune des coordonnées admet une double représentation).

Terminons ce paragraphe par quelques remarques. D'abord sur la généralisation à l'es-

pace $\mathbb{R}^n$ de la construction.

Remarque A.7 Cette construction se généralise aisément si l'on souhaite prendre le cube unité de l'espace $\mathbb{R}^3$ comme image[❋]. Il suffit de faire correspondra à $t \in [0,1]$,

$$x(t) = 0.d_1^{(x)}d_2^{(x)}d_3^{(x)}\cdots, \qquad y(t) = 0.d_1^{(y)}d_2^{(y)}d_3^{(y)}\cdots, \qquad z(t) = 0.d_1^{(z)}d_2^{(z)}d_3^{(z)}\cdots,$$

avec

$$\begin{aligned} d_j^{(x)} &= K^{(\sum_{k=1}^{j-1} d_k^{(y)} + \sum_{k=1}^{j-1} d_k^{(z)})} d_{3n-2}, \\ d_j^{(y)} &= K^{(\sum_{k=1}^{j} d_k^{(x)} + \sum_{k=1}^{j-1} d_k^{(z)})} d_{3n-1}, \\ d_j^{(z)} &= K^{(\sum_{k=1}^{j} d_k^{(x)} + \sum_{k=1}^{j} d_k^{(y)})} d_{3n}. \end{aligned}$$

□

Ensuite sur la généralisation à une autre base que la base 3.

Remarque A.8 Les mêmes conclusions peuvent être obtenues si l'on prend pour base de numération un autre nombre impair que 3. Pour les nombre pairs, la construction décrite ici doit être quelque peu modifiée et devient moins simple. □

❋. Et même $\mathbb{R}^n$, comme on le constate trivialement.

Annexe B

Régression linéaire par la méthode de la médiane

La régression linéaire est un outil de base pour l'analyse et la modélisation de données. La méthode des moindre carrés est une méthode quasiment universelle : peu de personnes se soucient d'indiquer la manière dont elles ont procédé pour modéliser des données par une relation linéaire. Pourtant la régression des moindres carrés repose sur une distribution des erreurs gaussienne, ce qui est loin d'être le cas pour bon nombre de données expérimentales. Aussi, il peut être bon de disposer de méthodes plus générales, permettant des estimations plus robustes.

B.1 La méthode des moindres carrés

Nous rappelons ici les concepts de base sous-jacents aux méthodes de régression linéaire en exposant brièvement la méthode des moindres carrés [64, 83].

Le principe de la régression linéaire peut être exposé comme suit. Étant donnés un ensemble de points $\{(x_i,y_i)\}_{1\leqslant i\leqslant n}$, on souhaite trouver le coefficient θ et la constante c

décrivant le mieux ces données par une relation linéaire,

$$y_i \approx \theta x_i + c, \tag{B.1}$$

où nous restons volontairement flou sur le sens de « mieux ». Une méthode intuitive consiste à déterminer θ et c tels qu'ils minimisent une certaine distance entre les points $\{(x_i,y_i)\}$ et les points $\{(x_i,\theta x_i + c)\}$. Le résultat obtenu peut être très différent selon la fonction distance choisie.

Supposons que les données $\{(x_i,y_i)\}_{1\leqslant i\leqslant n}$ soient issues d'un modèle linéaire $y(x) = \theta^* x + c^*$ et que des erreurs soient commises lors de la mesure des points y_i. Supposons de plus que ces erreurs soient identiquement et indépendamment distribuées selon une loi normale d'écart type σ autour du modèle $y(x)$, $y(x_i) = y_i + \varepsilon_i$, $\varepsilon_i \in N(0,\sigma^2)$. Une manière de choisir θ^* et c^* consiste à maximiser la probabilité que, étant donnés le modèle linéaire défini par $y(x) = \theta x + c$, les points mesurés soient $\{(x_i,y_i)\}$, plus ou moins une certaine valeur[*] δ pour les y_i. La fonction de vraisemblance à maximiser est la suivante [64],

$$(\frac{1}{\sqrt{2\pi}\sigma})^n \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta x_i - c)^2), \tag{B.2}$$

ce qui revient à minimiser

$$\sum_{i=1}^{n}(y_i - \theta x_i - c)^2. \tag{B.3}$$

selon θ et c. Cette méthode, appelée *méthode des moindres carrés*. Les expressions de θ^* et c^* en fonction des points $\{(x_i,y_i)\}$ peuvent être obtenues directement en dérivant l'expression (B.3) par rapport à θ et c. On obtient

$$c^* = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \tag{B.4}$$

et

$$\theta^* = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \tag{B.5}$$

B.2 La méthode de la médiane

Si les erreurs ne sont pas distribuées selon une loi normale[⁕], la régression linéaire des moindres carrés peut devenir trop sensible aux petites perturbations dans la mesure des points y_i. La méthode de la médiane[188, 215] est une méthode plus robuste que celle des moindres carrés, car elle repose sur une distribution des erreurs plus large.

[*]. Sinon, cette probabilité serait zéro.

[⁕]. On peut relaxer l'hypothèse d'homoscédasticité.

Supposons que les erreurs selon y_i dans la mesures de points $\{(x_i,y_i)\}$ ne soient pas distribuées selon une loi normale. On peut tenter de généraliser l'expression à maximiser (B.2) par celle qui suit,

$$\exp\big(-\sum_{i=1}^{n} f(y_i - \theta x_i - c)\big), \tag{B.6}$$

où f est une fonction jouant le rôle d'un logarithme négatif de fonction de densité. Cela revient à minimiser

$$\sum_{i=1}^{n} f(y_i - \theta x_i - c) \tag{B.7}$$

selon θ et c. Si l'on pose $f(t) = t^2/2$, on retrouve le cas où les erreurs sont distribuées selon une loi normale. Une manière de rendre la régression plus robuste consiste à supposer que les erreurs sont distribuées selon une double exponentielle en posant $f(t) = |t|$. Dans ce cas, la fonction à minimiser est

$$\sum_{i=1}^{n} |y_i - \theta x_i - c|, \tag{B.8}$$

appelée *déviation absolue*.

La médiane m d'un ensemble de données $\{x_i\}$ est aussi une valeur[⁂] minimisant la déviation absolue $\sum_i |x_i - m|$. Ainsi, pour un θ fixé, une valeur de c minimisant l'expression (B.8) est

$$c(\theta) = \underset{1 \leqslant i \leqslant n}{\text{médiane}}\{y_i - \theta x_i\}. \tag{B.9}$$

Pour trouver les extrema de l'expression (B.8) selon θ, on considère l'égalité suivante,

$$\sum_{i=1}^{n} \text{sign}(y_i - \theta x_i - c)\ x_i = 0, \tag{B.10}$$

où on a posé $\text{sign}(0) = 0$. En remplaçant c dans cette équation par la fonction $c(\theta)$ définie par l'égalité (B.9), on obtient une équation d'une seule variable. Cette dernière peut être résolue par une méthode du type bissection[❋]. L'estimation des paramètres θ^* et c^* par minimisation de la déviation absolue est appelée *méthode de la médiane*.

La recherche des paramètres optimaux θ^* et c^* selon d'autres critères, *i.e.* en utilisant une autre fonction f, peut se révéler coûteuse en temps. Il s'agit alors d'un problème d'optimisation. Pour une distribution avec de plus grandes queues de distribution, on peut prendre une distribution du type Cauchy ou Lorentz, $(1 + (y_i - \theta x_i - c)^2/2\sigma^2)^{-1}$, en posant $f(t) = \log(1 + t^2/2)$.

⁂. Avec cette fonction, la solution n'est pas unique ; ainsi tout point compris entre x_1 et x_2 minimise la déviation absolue associée à ces deux points.

❋. Il faut être prudent si l'on utilise d'autres méthodes pour trouver les racines, en raison des discontinuités de l'équation (B.10).

Bibliographie

[1] G. Abdurashidova, M. Deganuto, R. Klima, S. Riva, G. Biamonti, M. Giacca, and A. Falaschi. Start sites of bidirectional DNA synthesis at the human lamin B2 origin. *Science*, 287:2023–2026, 2000.

[2] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, New-York, 1972.

[3] P. Abry and F. Sellan. The wavelet-based synthesis for fractional Brownian motion. *Applied and Computational Harmonic Analysis*, 3:377–383, 1996.

[4] R.A. Adams. *Sobolev Spaces*. Academic Press, New York, 1978.

[5] W.W. Adams. A remarkable class of continued fractions. *Proceedings of the American Mathematical Society*, 65:194–198, 1977.

[6] R. Adler. *The Geometry of Random Fields*. John Wiley & Sons, New-York, 1981.

[7] A. Aharony and J. Feder, editeurs. *Fractals in Physics, Essays in honour of B.B. Mandelbrot*, Amsterdam, 1989. North Holland.

[8] M.I. Aladjem. The mammalian beta globin origin of DNA replication. *Frontiers in Bioscience: A Journal and Virtual Library*, 9:2540–2547, 2004.

[9] M.I. Aladjem, M. Groudine, L.L. Brody, E.S. Dieken, R.E. Fournier, G.M. Wahl, and E.M. Epner. Participation of the human beta-globin locus control region in initiation of DNA replication. *Science*, 270:815–819, 1995.

[10] M.I. Aladjem, L.W. Rodewald, J.L. Kolman, and G.M. Wahl. Genetic dissection of a mammalian replicator in the human beta-globin locus. *Science*, 281:1005–1009, 1998.

[11] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *L'essentiel de la Biologie Cellulaire*. Flammarion, Paris, 1998.

[12] R.A. Alberty. Use of Legendre transforms in chemical thermodynamics. *Pure and Applied Chemistry*, 73:1349–1380, 2001.

[13] A. Aldroubi and M. Unser. *Wavelets in Medecine and Biology*. CRC Press, Boca Raton, 1996.

[14] P.S. Aleksandrov and P. Alexandrov. *Combinatorial Topology*. Dover Publications, New-York, 1998.

[15] F. Antequera. Structure, function and evolution of CpG island promoters. *Cellular and Molecular Life Sciences*, 60:1647–1658, 2003.

[16] F.D. Araujo, J.D. Knox, S. Ramchandani, R. Pelletier, P. Bigey, G. Price, M. Szyf, and M. Zannis-Hadjopoulos. Identification of initiation sites for DNA replication in the human dnmt1 (DNA-methyltransferase) locus. *The Journal of Biological Chemistry*, 274:9335–9341, 1999.

[17] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J.F. Muzy. *Ondelettes, Multifractales et Turbulence : de l'ADN aux Croissances Cristallines*. Art et Sciences, Paris, 1995.

[18] A. Arneodo, F. Argoul, J. Elezgaray, and G. Grasseau. Wavelet transform analysis of fractals: application to nonequilibrium phase transitions. Dans G. Turchetti, editeur, *Nonlinear Dynamics*, Singapour, 1989. World Scientific.

[19] A. Arneodo, F. Argoul, and G. Grasseau. Transformation en ondelettes et renormalisation. Dans P.G. Lemarié, editeur, *Les Ondelettes en 1989*, Berlin, 1990. Springer-Verlag.

[20] A. Arneodo, B. Audit, N. Decoster, J.F. Muzy, and C. Vaillant. Climate disruptions, market crashes, and heart attacks. Dans A. Bunde and H.J. Schellnhuber, editeurs, *The Science of Disaster*, pages 27–102, Berlin, 2002. Springer.

[21] A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Physical Review Letters*, 74:3293–3296, 1995.

[22] A. Arneodo, E. Bacry, S. Jaffard, and J.F. Muzy. Oscillating singularities on Cantor sets: a grand-canonical multifractal formalism. *Journal of Statistical Physics*, 87:179–209, 1997.

[23] A. Arneodo, E. Bacry, S. Jaffard, and J.F. Muzy. Singularity spectrum of multifractal functions involving oscillating singularities. *Journal of Fourier Analysis and Applications*, 4:159–174, 1998.

[24] A. Arneodo, E. Bacry, S. Jaffard, and J.F. Muzy. Oscillating singularities and fractal functions. *CRM Proceeding & Lecture Notes*, 18:315–329, 1999.

[25] A. Arneodo, E. Bacry, and J.F. Muzy. The thermodynamics of fractals revisited with wavelets. *Physica A*, 213:232–275, 1995.

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

[27] A. Arneodo, N. Decoster, and S.G. Roux. A wavelet-based method for multifractal image analysis. I. Methodology and test applications on isotropic and anisotropic random rough surfaces. *The European Physical Journal B*, 15:567–600, 1999.

[28] A. Arneodo, G. Grasseau, and M. Holschneider. Wavelet transform of multifractals. *Physical Review Letters*, 61:2281–2284, 1988.

[29] J.M. Aubry. Representation of the singularities of a function. *Applied and Computational Harmonic Analysis*, 6:282–286, 1999.

[30] J.M. Aubry, F. Bastin, and S. Dispa. Prevalence of multifractal functions in S^ν spaces. Soumis pour publication, 2006.

[31] J.M. Aubry, F. Bastin, S. Dispa, and S. Jaffard. Topological properties of the sequence spaces S^ν. À paraître dans *Journal of Mathematical Analysis and Applications*, 2006.

[32] J.M. Aubry and S. Jaffard. Random wavelet series. *Communications in Mathematical Physics*, 227:483–514, 2002.

[33] B. Audit. *Analyse statistique des séquences d'ADN par l'intermédiaire de la transformée en ondelettes*. thèse de doctorat, Université de Paris VI, 1999.

[34] B. Audit, C. Thermes, Y. d'Aubenton Carafa, C. Vaillant, J.F. Muzy, and A. Arneodo. Long-range correlations in genomic DNA: a signature of nucleosomal DNA. *Physical Review Letters*, 86:2471–2474, 2001.

[35] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes. Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *Journal of Molecular Biology*, 316:903–918, 2002.

[36] K. Azuma. Weighted sums of certain dependant random variables. *Tôhoku Mathematical Journal*, 19:357–367, 1967.

[37] E. Bacry, A. Arneodo, and J.F. Muzy. Singularity spectrum of fractal signals from wavelet analysis: exact results. *Journal of Statistical Physics*, 70:635–674, 1993.

[38] E. Bacry, J.F. Muzy, and A. Arneodo. Oscillating singularities in locally self-similar functions. *Physical Review Letters*, 74:4823–4826, 1995.

[39] R. Badii and A. Politi. Hausdorff dimension and uniformity of strange attractors. *Physical Review Letters*, 52:1661–1664, 1984.

[40] R. Badii and A. Politi. Statistical description of chaotic attractors: the dimension function. *Journal of Statistical Physics*, 40:725–750, 1985.

[41] D.H. Bailey and R.E. Crandall. On the random character of fundamental constant expansions. *Experimental Mathematics*, 10:175–190, 2001.

[42] M.F. Barnsley. *Fractals Everywhere*. Academic Press, Orlando, 1988.

[43] M.F. Barnsley and S.G. Demko. Iterated function schemes and the global construction of fractals. *Proceedings of the Royal Society. London. Series A*, 399:243–275, 1985.

[44] M.F. Barnsley and A.D. Sloan. A better way to compress images. *Byte*, 13:215–233, 1988.

[45] F. Bastin and S. Nicolay. A general recurrence relation between the moments of a scaling function. Dans J.-P. Gazeau, R. Kerner, J.-P. Antoine, S. Métens, and J.-Y. Thibon, editeurs, *Group 24: Physical and Mathematical Aspects of Symmetries*, pages 921–924, Bristol, 2003. IoP.

[46] F. Bastin and S. Nicolay. A note on moments of scaling functions. *Rocky Mountain Journal of Mathematics*, 34:1197–1206, 2004.

[47] G.K. Batchelor. *The Theory of Homogeneous Turbulence*. Cambridge University Press, Cambridge, 1960.

[48] M.A. Batzer and P.L. Deininger. Alu repeats and human genomic diversity. *Nature Review. Genetics*, 3:370–379, 2002.

[49] E. Beaudoing and D. Gautheret. Identification of alternate polyadenilation sites and analysis of their tissue distribution using EST data. *Genome Research*, 11:1520–1526, 2001.

[50] A. Beletskii and A.S. Bhagwat. Correlation between transcription and C to T amination in the non-transcribed DNA strand. *Biological Chemistry*, 379:549–551, 1998.

[51] S.P. Bell and A. Dutta. DNA replication in eukaryotic cells. *Annual Review of Biochemistry*, 71:333–374, 2002.

[52] A.S. Belmont, S. Dietzel, A.C. Nye, Y.G. Strukov, and T. Tumbar. Large-scale chromatin structure and fonction. *Current Opinion in Cell Biology*, 11:307–311, 1999.

[53] M. Ben Slimane. Multifractal formalism for self-similar functions under the action of nonlinear dynamical systems. *Constructive Approximation*, 15:209–240, 1999.

[54] R. Benzi, G. Paladin, G. Parisi, and A. Vulpani. On the multifractal nature of fully developped turbulence and chaotic systems. *Journal of Physics. A: Mathematical and General*, 17:3521–3531, 1984.

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

[56] G. Bernardi. The isochore organization of the human genome. *Annual Review of Genetics*, 23:637–661, 1989.

[57] G. Bernardi. Misunderstanding about isochores. Part 1. *Gene*, 276:3–13, 2001.

[58] A.S. Besicovitch. On the fundamental geometrical properties of linearly measurable plane sets of points. *Matematische Annalen*, 98:422–464, 1928.

[59] A.S. Besicovitch. On linear sets of points of fractional dimension. *Matematische Annalen*, 101:161–193, 1929.

[60] A.S. Besicovitch. On the sum of digits of real numbers represented in dyadic systems. *Matematische Annalen*, 110:321–330, 1935.

[61] A.S. Besicovitch. On the fundamental geometrical properties of linearly measurable plane sets of points II. *Matematische Annalen*, 115:296–329, 1938.

[62] A.S. Besicovitch. On the fundamental geometrical properties of linearly measurable plane sets of points III. *Matematische Annalen*, 116:349–357, 1939.

[63] O.V. Besov. S.M. Nikol'skiĭ's work on the theory of function spaces and its applications. *Proceeding of the Steklov Institute of Mathematics*, 232:19–24, 2001.

[64] P.R. Bevington. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, New-York, 1969.

[65] G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms. *Communications on Pure and Applied Mathematics*, 44:141–183, 1991.

[66] A.K. Bielinsky, H. Blitzblau, E.L. Beal, M. Ezrokhi, H.S. Smith, M.R. Botchan, and S.S. Gerbi. Origin recognition complex binding to a metazoan replication origin. *Current Biology*, 11:1427–1431, 2001.

[67] P. Billingsley. The singular function of bold play. *American Scientist*, 71:392–397, 1983.

[68] P.E. Böhmer. Über die Transcendenz gewisser dyadischer Brüche. *Matematische Annalen*, 96:367–377, 1926.

[69] P.E. Böhmer. Über die Transcendenz gewisser dyadischer Brüche. Erratum. *Matematische Annalen*, 96:735, 1926.

[70] T. Bohr and T. Tèl. The thermodynamics of fractals. Dans B.L. Hao, editeur, *Direction in Chaos*, page 16, Singapour, 1988. World Scientific.

[71] A. Bonami and A. Estrade. Anisotropic analysis of some gaussian models. *Journal of Fourier Analysis and Applications*, 9:215–239, 2003.

[72] J.M. Bony. Second microlocalization and propagation of singularities for semi-linear hyperbolic equations. Dans *Proceeding of Tanaguchi Symposium*, pages 11–49. HERT. Katata, 1984.

[73] J.P. Bouchaud and M. Potters. *Théorie des Risques Financiers*. Eyrolles, Aléa-Saclay, 1997.

[74] D. Bowman. Approximation of $\lfloor n\alpha + s \rfloor$ and the zero of $\{n\alpha + s\}$. *Journal of Number Theory*, 50:128–144, 1995.

[75] G. Box and G. Jenkins. *Time Serie Analysis: Forecasting and Control*. Holden-Day, Oakland, 1976.

[76] R. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill, New-York, 1995.

[77] B.J. Brewer. When polymerases collide: replication and the transcriptional organisation of E. coli chromosome. *Cell*, 53:679–686, 1998.

[78] J.M. Bridger and W.A. Bickmore. Putting the genome on the map. *Trends in Genetics*, 14:403–409, 1998.

[79] E.B. Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton Carafa, C. Thermes, and A. Arneodo. From DNA sequence analysis to modelling replication in the human genome. *Physical Review Letters*, 94:248103, 2005.

[80] L.E.J. Brouwer. Beweis der invarianz der Dimensionenzahl. *Matematische Annalen*, 70:161–165, 1911.

[81] G. Brown, G. Michon, and J. Peyrière. On the multifractal analysis of measures. *Journal of Statistical Physics*, 66:775–790, 1992.

[82] R. Brown. A brief account of microscopical observations made in the months on June, July, and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philosophycal Magazine*, 4:161–173, 1828.

[83] K.A. Brownlee. *Statistical Theory and Methodology*. John Wiley & Sons, New-York, 1965.

[84] I. Brukner, R. Sanchez, D. Suck, and S. Pongor. Sequence-dependent bending properties of DNA as revealed by DNase I: parameters of trinucleotides. *The EMBO Journal*, 14:1812–1818, 1995.

[85] I. Brukner, R. Sanchez, D. Suck, and S. Pongor. Trinucleotide models for DNA bending propensity: comparison of models based on DNase I digestion and nucleosome packaging data. *Journal of Biomolecular Structure & Dynamics*, 13:309–317, 1995.

[86] M. Bulmer. Strand asymmetry of mutation rates in the beta-globin region. *Journal of Molecular Evolution*, 33:305–310, 1991.

[87] V. Burland, G. Plunkett, D.L. Daniels, and F.R. Blattner. DNA sequence and analysis of 136 kilobases of Escherichia coli genome: organizational symmetry around the origin of replication. *Genomics*, 16:551–561, 1993.

[88] A.P. Calderòn. Intermediate spaces and interpolation, the complex method. *Studia Mathematica*, 24:113–190, 1964.

[89] C.R. Calladine and H.R. Drew. *Understanding DNA*. Academic Press, San Diego, 1999.

[90] J. Canny. A computational approach to edge detection. *Institute of Electrical and Electronic Engineers. Transactions on Pattern Analysis and Machine Intelligence*, 8:679–698, 1886.

[91] G. Cantor. Ein beitrag zur mannigfaltigkeitslehre. *Journal für die reine und angewandte Mathematik*, 84:242–258, 1878.

[92] C. Carathéodory. Über das lineare maß von punktmengen - eine verallgemeinerung des längenbegriffs. *Nachrichten der K. Gesellschaft de Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*, pages 404–426, 1914.

[93] D.R. Chalice. A characterization of the cantor function. *The American Mathematical Monthly*, 98:255–258, 1991.

[94] D.C. Champeney. *Fourier Transform and their Physical Applications*. Academic Press, New-York, 1973.

[95] G. Chan and A.T.A. Wood. An algorithm for simulating stationary Gaussian random fields. *Applied Statistics*, 46:171–181, 1997.

[96] G. Chan and A.T.A. Wood. Simulation of multifractional Brownian motion. Dans R. Payne and P. Green, editeurs, *Procceedings in Computational Statistics*, pages 233–238. Physica-Verlag, 1998.

[97] E. Chargaff. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experimentia*, 6:201–240, 1950.

[98] E. Chargaff. Structure and function of nucleic ascids as cell constituents. *Federation Proceedings*, 10:654–659, 1951.

[99] E. Chassande-Motin and P. Flandrin. On the time-frequency detection of chirps. *Applied and Computational Harmonic Analysis*, 6:252–281, 1999.

[100] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[101] C. K. Chui. *Wavelets: a Mathematical Tool for Signal Analysis*. SIAM, Philadelphia, 1997.

[102] C.K. Chui. *An Introduction to Wavelets*. Academic Press, New-York, 1992.

[103] L. Cohen. *Time-Frequency Analysis: Theory and Applications*. Physica A, New Jersey, 1995.

[104] N. Cohen, T. Dagan, L. Stone, and D. Graur. GC omposition of the human genome: in search of isochores. *Molecular Biology and Evolution*, 22:1260–1272, 2004.

[105] D.L. Cohn. *Measure Theory*. Birkhäuser, Stuttgart, 1980.

[106] P. Collet, J. Lebovitz, and A. Porzio. The dimension spectrum of some dynamical systems. *Journal of Statistical Physics*, 47:609–644, 1987.

[107] J. Craig and W.A. Bickmore. The distribution of CpG islands in mammalian chromosomes. *Nature Genetics*, 7:376–382, 1994.

[108] T. Cremer and C. Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews. Genetics*, 2:292–301, 2001.

[109] P. Cvitanovic. Hausdorff dimension of irrational windings. Dans R. Gilmore, editeur, *Proceedings of the XV International Colloquium on Group Theoretical Methods in Physics*, pages 184–198, Singapore, 1987. World Scientific.

[110] D. Dacunha-Castelle and M. Duflo. *Probabilités et Statistiques*, volume I: Problèmes à Temps Fixe. Masson, Paris, 1982.

[111] D. Dacunha-Castelle and M. Duflo. *Probabilités et Statistiques*, volume II: Problèmes à Temps Mobile. Masson, Paris, 1983.

[112] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, PA, 1992.

[113] R.O. Davies. A property of Hausdorff measure. *Proceedings of the Cambridge Philosophical Society*, 52:30–34, 1956.

[114] P.J. Davis. *Circulant Matrices*. John Wiley & Sons, New-York, 1979.

[115] N. Decoster. *Analyse multifractale d'images de surfaces rugueuses à l'aide de la transformation en ondelettes*. thèse de doctorat, Université de Bordeaux I, 1999.

[116] F.M. Dekking. Recurrent sets. *Advances in Mathematics*, 44:78–104, 1982.

[117] N. Delprat, B. Escudié, P. Guillemain, R. Krolabd-Martinet, P. Tchamitchian, and B. Torresani. Asymptotic wavelet and Gabor analysis: extraction of instantaneous frequencies. *Institute of Electrical and Electronic Engineers. Transactions on Information Theory*, 38:644–664, 1992.

[118] B. Démidovitch and I. Maron. *Éléments de Calcul Numérique*. MIR, Moscou, 1979.

[119] T. Dieker. Simulation of fractional Brownian motion. Dissertation de DEA, Vrije Universiteit Amsterdam, 2002.

[120] C. Dietrich and G. Newsam. Fast and exact simulation of stationnary processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18:1088–1107, 1997.

[121] S. Dispa. *Beyond Besov spaces, S^ν spaces : topology and prevalent properties*. thèse de doctorat, Université de Paris VI, 1999.

[122] R.L. Dobrushin and P. Major. Non-central limit theorem for nonlinear functionals of Gaussian fields. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 50:27–52, 1979.

[123] D.L. Donoho. De-noising by soft-thresholding. Technical report, Department of Statistics, Stanford University, 1992.

[124] D.L. Donoho. Wavelet shrinkage and w.v.d.: a 10-minute tour. Technical report, Department of Statistics, Stanford University, 1993.

[125] D.L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. Technical report, Department of Statistics, Stanford University, 1994.

[126] D.L. Donoho and I.M. Johnstone. Ideal denoising in a orthogonal basis chosen from a library of bases. Technical report, Department of Statistics, Stanford University, 1994.

[127] J.L. Doob. The brownian movement and stochastic equations. *Annals of Mathematics*, 43:351–369, 1942.

[128] L. Duret. Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, 12:640–649, 2002.

[129] M.E. Dury. *Identification et simulation d'une classe de processus stables autosimilaires à accroissements stationnaires*. thèse de doctorat, Université Blaise Pascal, Clermont-Ferrand, 2001.

[130] A. Dvir. Promoter escape by RNA polymerase II. *Biochimica et Biophysica Acta*, 1577:208–223, 2002.

[131] M.J. Dye and N.J. Proudfoot. Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II. *Cell*, 105:669–681, 2001.

[132] G.A. Edgar. *Integral, Probability and Fractal Measures*. Springer-Verlag, New-York, 1998.

[133] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 17:549, 1905.

[134] A. Einstein. *Investigations on the Theory of Brownian Movements*. Dover Publications, New-York, 1956.

[135] R.J. Elliot and L. Chan. Perpetual american options with fractional Brownian motion. *Quantitative Finance*, 4:123–128, 2004.

[136] P. Embrechts and M. Maejima. *Selfsimilar Processes*. Princeton University Press, London, 2002.

[137] G. Erlebacher, M.Y. Hussaini, and L.M. Jameson, editeurs. *Wavelets: Theory and Applications*, Oxford, 1996. Oxford University Press.

[138] E.S. Lander *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[139] J.C. Venter *et al.* . The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[140] K.J. Falconer. *The Geometry of Fractal Sets*. Cambridge University Press, Cambridge, 1985.

[141] K.J. Falconer. *Fractal Geometry*. John Wiley & Sons, Chichester, 1990.

[142] K.J. Falconer. *Techniques in Fractal Geometry*. John Wiley & Sons, New-York, 1997.

[143] J. Feder. *Fractals*. Plenum Press, New-York, 1988.

[144] H. Federer. *Geometric Measure Theory*. Springer-Verlag, New-York, 1969.

[145] M.J. Feingenbaum, M.H. Jensen, and I. Procaccia. Time ordering and the thermodynamics of strange sets: theory and experimental tests. *Physical Review Letters*, 57:1503–1506, 1986.

[146] W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. John Wiley & Sons, New-York, deuxième edition, 1966.

[147] W. Feller. *An Introduction to Probability Theory and its Applications*, volume I. John Wiley & Sons, New-York, troisième edition, 1968.

[148] G. Fix and G. Strang. Fourier analysis of the finite element method in Ritz-Galerkin theory. *Studies in Applied Mathematics*, 18:265–274, 1969.

[149] M.P. Francino, L. Chao, M.A. Riley, and H. Ochman. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science*, 272:107–109, 1996.

[150] M.P. Francino and H. Ochman. Strand assymmetries in DNA evolution. *Trends in Genetics*, 13:240–245, 1997.

[151] M.P. Francino and H. Ochman. Strand symmetry around the *β-globin* origin of replication in primates. *Molecular Biology and Evolution*, 17:416–422, 2000.

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

[153] M. Fréchet. Les dimensions d'un ensemble abstrait. *Matematische Annalen*, 68:145–168, 1910.

[154] U. Frisch. *Turbulence*. Cambridge University Press, Cambridge, 1995.

[155] A. Gabrelian, A. Simoncsits, and S. Pongor. Distribution of bending propensity in DNA sequences. *Federation of European Biochemical Society Letters*, 393:124–130, 1996.

[156] F.B. Galleco. Invariants on compact spaces. *Rendiconti dell'Istituto dell'Università di Trieste*, 30:31–44, 1999.

[157] F.R. Gantmacher. *The Theory of Matrices*, volume 1. Chelsea Publishing Co., New-York, 1959.

[158] F.R. Gantmacher. *The Theory of Matrices*, volume 2. Chelsea Publishing Co., New-York, 1959.

[159] G. Gasquet and P. Witomski. *Analyse de Fourier et Applications*. Dunod, Paris, 2003.

[160] S.M. Gasser. Visualizing chromatin dynamics in interphase nuclei. *Science*, 296:1412–1416, 2002.

[161] A. Gierlik, M. Kowalczuk, P. Mackiewicz, M.R. Dudek, and S. Cebrat. Is there replication-associated mutational pressure in the *saccharomyces cerevisae* genome. *Journal of Theoretical Biology*, 202:305–314, 2000.

[162] D.M. Gilbert. Making sense of eukaryotic DNA replication origins. *Science*, 294:96–100, 2001.

[163] C. Girard-Reydet, D. Gregoire, Y. Vassetzky, and M. Mechali. DNA replication initiates at domains overlapping with nuclear matrix attachment regions in the xenopus and mouse c-myc promoter. *Gene*, 332:129–138, 2004.

[164] J.A. Glazier, H.H. Jensen, A. Libchaber, and J. Stavans. Structure of Arnold tongues and the $f(\alpha)$ spectrum for period doubling: experimental results. *Physical Review A*, 34:1621–1624, 1986.

[165] D.S. Goodsell and R.E. Dickerson. Bending and curvature calculations in B-DNA. *Nucleic Acid Research*, 22:5497–5503, 1994.

[166] P. Grassberger. Generalized dimensions of strange attractors. *Physics Letters. A.*, 97:227–230, 1983.

[167] P. Grassberger, R. Badii, and A. Politi. Scaling laws for invariant measures on hyperbolic and nonhyperbolic attractors. *Journal of Statistical Physics*, 51:135–178, 1988.

[168] P. Grassberger and I. Procaccia. Dimensions and entropies of strange attractors from fluctuating dynamics approach. *Physica D*, 13:34–54, 1984.

[169] R.M. Gray. Toeplitz and circulant matrices: a review. Technical report, Information Systems Laboratory, Stanford University, California, 2002.

[170] P. Green, B. Ewing, W. Miller, P.J. Thomas, and E.D. Green. Transcription-associated mutational asymmetry in mamalian evolution. *Nature Genetics*, 33:514–517, 2003.

[171] A. Grigoriev. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research*, 26:2286–2290, 1998.

[172] G. Gripenberg. Approximations of wavelet projections. *Applied and Computational Harmonic Analysis*, 2:257–264, 1995.

[173] A. Grossman and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, 15:723–736, 1984.

[174] A. Grossman and J. Morlet. Decomposition of functions into wavelets of constant shape and related transforms. Dans L. Streit, editeur, *Mathematics and Physics, Lectures on Recent Results*, pages 135–165. World Scientific, 1985.

[175] A. Haar. Zur Theorie der orthogonalen Funktionensysteme. *Matematische Annalen*, 69:331–371, 1910.

[176] T.C. Halsey, M.H. Jensen, L.P. Kadanoff, I. Procaccia, and B.I. Shraiman. Fractal measures and their singularities: the characterization of strange sets. *Physical Review A*, 33:1141–1151, 1986.

[177] D. Häring and J. Kypr. No isochore in the human chromosomes 21 and 22? *Biochemical and Biophysical Research Communications*, 280:567–573, 2001.

[178] D. Harte. *Multifractals*. CRC Press, Boca Raton, 2001.

[179] D.L. Hartl and E.W. Jones. *Genetics: Analysis of Genes & Genomes*. Jones and Bartlett, Sudbury, 2001.

[180] F. Hausdorff. Dimension und Äußeres maß. *Matematische Annalen*, 79:157–179, 1919.

[181] H.G.E. Hentschel and I. Procaccia. The infinite number of generalized dimensions of fractals and strange attractors. *Physica D*, 8:435–444, 1983.

[182] Y. Heurteaux. Weierstrass functions in Zygmund's class. *Proceedings of the American Mathematical Society*, 133:2711–2720, 2005.

[183] T. Hida. *Brownian Motion*. Springer-Verlag, New-York, 1980.

[184] K. van Holde. *Chromatin*. Springer, Berlin, 1989.

[185] M. Holschneider. *L'analyse d'objets fractals et leur transformation en ondelettes*. thèse de doctorat, Université d'Aix-Marseille II, 1998.

[186] J.R.M. Hosking. Fractionnal differencing. *Biometrika*, 68:165–176, 1981.

[187] Y. Hu, B. Øskendal, and D.M. Salopek. Weighted local time for fractional brownian motion and applications to finance. *Stochastic Analysis and Applications*, 23:15–30, 2005.

[188] P.J. Huber. *Robust Statistcs*. John Wiley & Sons, New-York, 1981.

[189] International human genome sequencing consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[190] B. Hummel and R. Moniot. Reconstruction from zero-crossings in scale-space. *Institute of Electrical and Electronic Engineers. Transactions on Acoustic, Speech and Signal Processing*, 37:2111–2130, 1989.

[191] B.R. Hunt. The hausdorff dimension of graphs of weierstrass functions. *Proceedings of the American Mathematical Society*, 126:791–800, 1998.

[192] W. Hurewicz and K. Menger. Dimension und Zusammenhangsstufe. *Matematische Annalen*, 100:618–633, 1928.

[193] W. Hurewicz and H. Wallman. *Dimension Theory*. Princeton University Press, London, 1948.

[194] J.E. Hutchinson. Fractals and self-similarity. *Iniana University Mathematics Journal*, 30:713–747, 1981.

[195] J.E. Hutchinson and L. Rüschendorf. Selfsimilar fractals and selfsimilar random fractals. Dans C. Bandt, S. Graf, and M. Zähle, editeurs, *Fractal Geometry and Stochastics II*, pages 109–123, Basel, 2000. Birkhäuser.

[196] J. Istas and G. Lang. Quadratic variations and estimation of the local hölder index of a Gaussian process. *Annales de l'Institut Henri Poincarré*, 33:407–436, 1997.

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

[198] F. Jacob and S. Brenner. On the regulation of DNA synthesis in bacteria: the hypothesis of the replicon. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 256:298–300, 1963.

[199] S. Jaffard. *Construction et propriétés des bases d'ondelettes. Remarques sur la contrôlabilité exacte*. thèse de doctorat, École Polytechnique, 1989.

[200] S. Jaffard. Exposants de Hölder en des points donnés et coefficients d'ondelettes. *Comptes Rendus Mathématiques. Académie des Sciences. Paris*, 308:79–81, 1989.

[201] S. Jaffard. Pointwise smoothness, two-microlocalization and wavelet coefficients. *Publicacions Matemàtiques*, 35:155–168, 1991.

[202] S. Jaffard. Théorème de trace et « dimensions négatives ». *Comptes Rendus Mathématiques. Académie des Sciences. Paris*, 320:409–413, 1995.

[203] S. Jaffard. Multifractal formalism for functions part I: results valid for all functions. *SIAM Journal on Mathematical Analysis*, 28:945–970, 1997.

[204] S. Jaffard. Multifractal formalism for functions part II: self-similar functions. *SIAM Journal on Mathematical Analysis*, 28:971–998, 1997.

[205] S. Jaffard. Oscillation spaces: properties and applications to fractal and multifractal functions. *Journal of Mathematical Physics*, 39:4129–4141, 1998.

[206] S. Jaffard. Construction of functions with prescribed Hölder and chirp exponents. *Revista Mathemática Iberoamericana*, 16:331–349, 2000.

[207] S. Jaffard. On the Frisch-Parisi conjecture. *Journal de Mathématiques Pures et Appliquées*, 79:525–552, 2000.

[208] S. Jaffard. Communication personnelle, 2003.

[209] S. Jaffard. Beyond Besov spaces I: distributions of wavelet coefficients. *Journal of Fourier Analysis and Applications*, 10:221–246, 2004.

[210] S. Jaffard. Power laws in probability and statistics: introduction to multifractal analysis. Conférence du CIRM à Luminy, mars 2004.

[211] S. Jaffard. Wavelet techniques for pointwise regularity. Dans M. Lapidus and M. van Frankenhuysen, editeurs, *Fractal Geometry and Applications: a Jubilee of Benoit Mandelbrot*, pages 91–151. Proceedings of Symposia in Pure Mathematics, 2004.

[212] S. Jaffard. Beyond Besov spaces II: oscillation spaces. *Constructive Approximation*, 1:29–61, 2005.

[213] S. Jaffard, B. Lashermes, and P. Abry. Wavelet leaders in multifractal analysis. Soumis pour publication, 2006.

[214] S. Jaffard and Y. Meyer. Wavelet methods for pointwise regularity and local oscillations of functions. *Memoirs of the American Mathematical Society*, 123:1–102, 1996.

[215] A.H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New-York, 1970.

[216] M.H. Jensen, L.P. Kadanoff, A. Libchaber, I. Procaccia, and J. Stavans. Global universality at the onset chaos: results of a forced Rayleigh-Bénard experiment. *Physical Review Letters*, 55:2798–2801, 1985.

[217] J.P. Kahane. Ensembles aléatoires et dimensions. Dans I. Peral and J.L. Rubio di Francia, editeurs, *Recent Progress in Fourier Analysis*, pages 65–121, Amsterdam, 1985. North Holland.

[218] M. Kak. Random walks and the thory of Brownian motion. *The American Mathematical Monthly*, 54:369–391, 1947.

[219] I. Karatzas and S.E. Shreve. *Brownnian Motion and Stochastic Calculus*. Springer-Verlag, New-York, 1997.

[220] S. Karlin. Bacterial DNA strand compositional asymmetry in bacterial and large viral genomes. *Trends in Microbiology*, 7:305–308, 1999.

[221] D. Katzen and I. Procaccia. Phase transitions in the thermodynamic formalism of multifractals. *Physical Review Letters*, 58:1169–1172, 1987.

[222] C. Keller, E.M. Ladenburger, M. Kremer, and R. Knippers. The origin recognition complex marks a replication origin in the human top1 gene promoter. *The Journal of Biological Chemistry*, 22:1036–1048, 2002.

[223] A.Y. Khintchine. *Continued Fractions*. Noordhoff, Groningen, 1963.

[224] A.N. Khovanskii. *The Application of Continued Fractions and their Generalizations to Problems in Approximation Theory*. Noordhoff, Groningen, 1963.

[225] A.N. Kolmogorov. The local structure of turbulence in imcompressible viscious fluid for very large Reynolds number. *Doklady Akademii Nauk SSR*, 30:9–13, 1941.

[226] A.N. Kolmogorov. New metric invariant of transitive dynamical systems and automorphisms of Lebesgue spaces. *Doklady Akademii Nauk SSR*, 119:861–684, 1958.

[227] T.H. Koornwinder. *Waveletes: an Elementary Treatment of Theory and Applications*. World Scientific, Singapore, 1993.

[228] A. Kornberg and T.A. Baker. *DNA Replication*. W.H. Freeman and Co., New-York, 1992.

[229] R. Kronland-Martinet, J. Morlet, and A. Grossmann. Analysis of sound pattern through wavelet transforms. *International Journal of Pattern Recognition and Artificial Intelligence*, 1:273–301, 1987.

[230] C. Lacaux. *Champs de Lévy multifractionnaires*. thèse de doctorat, Toulouse III, 2004.

[231] E.M. Ladenburger, C. Keller, and R. Knippers. Identification of a binding region for human origin recognition complex protein 1 and 2 that coincide with an origin of DNA replication. *Molecular and Cellular Biology*, 22:1036–1048, 2002.

[232] U.K. Laemmli, E. Kas, L. Poljak, and Y. Adachi. Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. *Current Opinion in Genetics and Development*, 2:275–285, 1992.

[233] P. Lancaster. *Theory of Matrices*. Academic Press, New-York, 1969.

[234] F. Larsen, G. Gundersen, and R. Lopez H. Prydz. CpG islands as gene markers in the human genome. *Genomics*, 13:1095–1107, 1992.

[235] B. Lashermes. *Analyse multifractale pratique : coefficients dominants et ordres critiques. Application à la turbulence pleinement développée. Effets de nombre de Reynolds fini*. thèse de doctorat, École Normale Supérieure de Lyon, 2005.

[236] P. Laubin. Analyse numérique, problèmes non linéaires. Université de Liège, 1998–1999.

[237] H. Lebesgue. Sur la non-applicabilité de deux domaines appartenant respectivement à des espaces à n et $n+p$ dimensions. *Matematische Annalen*, 70:166–168, 1911.

[238] P.G. Lemarié, editeur. *Les Ondelettes en 1989*, Berlin, 1990. Springer-Verlag.

[239] P.G. Lemarié and Y. Meyer. Ondelettes et bases hilbertiennes. *Revista Mathemática Iberoamericana*, 2:1–18, 1986.

[240] P. Levy. *Processus Stochastiques et Mouvement Brownien*. Gauthier-Villars, Paris, 1948.

[241] P. Levy. *Théorie de l'Addition des Variables Aléatoires*. Gauthier-Villars, Paris, 1954.

[242] W. Li and K. Kaneko. Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhysics Letters*, 17:655–660, 1992.

[243] W.H. Li and D. Graur. *Fundamental of Molecular Evolution*. Sinauer, Sunderland, 1991.

[244] N.A. Liapunova. Organization of replication units and DNA replication in mammalian cells as studied by DNA fiber radioautography. *International Review of Cytology*, 154:261–308, 1994.

[245] H.J. Lin and E. Chargaff. On the denaturation of deoxyribonucleic acid. II: effects of concentration. *Biochimica et Biophysica Acta*, 145:398–409, 1967.

[246] S.J. Lin. Stochastic analysis of fractional Brownian motion, fractional noises and applications. *SIAM Reviews*, 10:422–437, 1995.

[247] R.D. Little, T.H. Platt, and C.L. Schildkraut. Initiation and termination of DNA replication in human rRNA genes. *Molecular and Cellular Biology*, 13:6600–6613, 1993.

[248] J.R. Lobry. Properties of a general model of DNA evolution under no-strand bias conditions. *Journal of Molecular Evolution*, 40:326–330, 1995.

[249] J.R. Lobry. Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, 13:660–665, 1996.

[250] J.R. Lobry. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, 78:323–326, 1996.

[251] J.R. Lobry and N. Sueoka. Asymmetric directional mutation pressures in bacteria. *Genome Biology*, 3:58.1–58.14, 2002.

[252] P. Lopez and H. Philippe. Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. *Comptes Rendus de l'Académie des Sciences. Série III (Sciences de la vie)*, 324:2001–2008, 2001.

[253] H. Ma, J. Samarabandu, R.S. Devdhar, R. Acharya, P.C. Cheng, C. Meng, and R. Berezney. Spatial and temporal dynamics of DNA replication sites in mammalian cells. *The Journal of Cell Biology*, 143:1415–1425, 1998.

[254] P. Mackiewicz, A. Gierlik, M. Kowalczuk, M.R. Dudek, and S. Cebrat. Asymmetry of nucleotide composition of prokaryotic chromosomes. *Journal of Applied Genetics*, 40:1–14, 1999.

[255] P. Mackiewicz, A. Gierlik, M. Kowalczuk, M.R. Dudek, and S. Cebrat. How does replication-associated mutational pressure influence amino acid composition of proteins? *Genome Research*, 9:409–416, 1999.

[256] S.G. Mallat. Multiresolutions approximations and wavelets orthonormal bases of $L^2(\mathbb{R})$. *Transactions of the American Mathematical Society*, 315:67–87, 1989.

[257] S.G. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, New-York, 1999.

[258] S.G. Mallat and W.L. Hwang. Singularity detection and processing with wavelets. *Institute of Electrical and Electronic Engineers. Transactions on Information Theory*, 38:617–643, 1992.

[259] S.G. Mallat and S. Zhong. Characterization of signals from multiscale edges. *Institute of Electrical and Electronic Engineers. Transactions on Pattern Analysis and Machine Intelligence*, 14:710–732, 1992.

[260] B.B. Mandelbrot. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *Journal of Fluid Mechanics*, 62:331–358, 1974.

[261] B.B. Mandelbrot. *Fractals, Form, Chance and Dimension*. Freeman, San Francisco, 1977.

[262] B.B. Mandelbrot. *The Fractal Geometry of Nature*. Freeman, San Francisco, 1982.

[263] B.B. Mandelbrot and J. Van Ness. Fractional Brownian motion, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

[264] P. Manneville. Systèmes dynamiques et chaos. Cours de DEA de physique des liquides et de mécanique de l'École Polytechnique, 1998–1999.

[265] R.N. Mantegna and H.E. Stanley. *An Introduction to Econophysics*. Cambridge University Press, Cambridge, 2000.

[266] E. Marczewski. *Collected Mathematical Papers*. Polish Academy of Sciences, Warszawa, 1996.

[267] D. Marr. *Vision*. Freeman, San Francisco, 1982.

[268] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society. London. Series A*, 207:187–217, 1980.

[269] F. Martìnez-Lopez, M.A. Cabrerizo-Vìlchez, and R. Hidalgo-Àlvarez. On the self-similarity of fractal colloidal aggregates in two dimensions. *Journal of Physics. A: Mathematical and General*, 34:7393–7398, 2001.

[270] P. Matilla. *Geometry of Set and Measures in Euclidian Spaces: Fractals and Rectificability*. Cambridge University Press, Cambridge, 1995.

[271] H.P. McKean. *Stochastic Integrals*. Academic Press, New-York, 1969.

[272] T. McKee and J.R. McKee. *Biochemistry: an Introduction*. McGraw-Hill, Boston, 1999.

[273] M.J. McLean, K.H. Wolfe, and K.M. Devine. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *Journal of Molecular Evolution*, 47:691–696, 1998.

[274] M. Méchali. DNA replication origins: from sequence specificity to epigenetics. *Nature Reviews. Genetics*, 2:640–645, 2001.

[275] C. Melot. *Sur les singularités oscillantes et le formalisme multifractal*. thèse de doctorat, Paris XII, 2002.

[276] C. Meneveau and K.R. Screenivasan. The multifractal nature of turbulent energy dissipation. *Journal of Fluid Mechanics*, 224:429–484, 1991.

[277] K. Menger. Zur Dimensions und Kurventheorie. *Monatshefte für Mathematik und Physik*, 36:411–432, 1922.

[278] Y. Meyer. *Ondelettes et Opérateurs I : Ondelettes*. Hermann, Paris, 1990.

[279] Y. Meyer. *Ondelettes et Opérateurs II : Opérateurs de Calderòn-Zygmund*. Hermann, Paris, 1990.

[280] Y. Meyer, editeur. *Wavelets and Applications*, Berlin, 1992. Springer-Verlag.

[281] Y. Meyer. *Wavelets, Vibrations and Scalings*. American Mathematical Society, Providence, 1997.

[282] Y. Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*. American Mathematical Society, Providence, 2001.

[283] Y. Meyer and S. Roques, editeurs. *Progress in Wavelets Analysis and Applications*, Gif-sur-Yvettes, 1993. Éditions frontières.

[284] Y. Meyer, F. Sellan, and M.S. Taqqu. Wavelets, generalized white nois and fractional integration: the synthesis of fractional Brownian motion. *The Journal of Fourier Analysis and Applications*, 5:465–494, 1999.

[285] Y. Meyer and H. Xu. Wavelet analysis and chirps. *Applied and Computational Harmonic Analysis*, 4:366–379, 1997.

[286] P.A.P. Moran. Additive functions of intervals and Haudorff measure. *Proceedings of the Cambridge Philosophical Society*, 42:15–23, 1946.

[287] J. Morlet. Sampling theory and wave propagation. Dans C.H. Chen, editeur, *NATO ASI Series, Issues in Acoustic Signal/Image Processing and Recognition*, pages 203–261, Berlin, 1983. Springer-Verlag.

[288] J. Mràzek and S. Karlin. Strand compositional asymmetry in bacterial and large viral genomes. *Proceedings of the National Academy of Sciences of the USA*, 95:3720–3725, 1998.

[289] C. Munkel, R. Elis, S. Dietzel, D. Zink, C. Mehring, G. Wedermann, T. Cremer, and J. Langowski. Comparmentalization of interphase chromosomes observed in simulation and experiment. *Journal of Molecular Biology*, 285:1053–1065, 1999.

[290] M. Münster. *Mesure et intégration dans les ensembles abstraits et dans les espaces topologiques*. thèse de doctorat, Université de Liège, 1969–1970.

[291] J.F. Muzy. Analyse de distributions fractales à partir de la transformée en ondelettes. *Academic Press*, 20:63–232, 1995.

[292] J.F. Muzy, A. Arneodo, and E. Bacry. A multifractal formalism revisited with wavelets. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 4:245–302, 1994.

[293] J.F. Muzy, E. Bacry, and A. Arneodo. Wavelets and multifractal formalism for singular signals: application tot turbulence data. *Physical Review Letters*, 67:3515–3518, 1991.

[294] J.F. Muzy, E. Bacry, and A. Arneodo. Multifractal formalism for fractal signals: the structure-function approach versus the wavelet-transform modulus-maxima method. *Physical Review E*, 47:875–884, 1993.

[295] T. Nenguke, M.I. Aladjem, J.F. Gusella, N.S. Wexler, and N. Arnheim. Candidate DNA replication initiation regions at human trinucleotide repeat dicease loci. *Human Molecular Genetics*, 12:1021–1028, 2003.

[296] S. Nicolay. Analyse multirésolution et applications. Dissertation de DEA, Université de Liège, 2000–2001.

[297] S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton Carafa, C. Thermes, and A. Arneodo. Low frequency rythms in human DNA sequences: a key to the organization of gene location and orientation? *Physical Review Letters*, 93:108101, 2004.

[298] S. Nicolay, B. Audit, and A. Arneodo. Synthetizing artificial DNA sequences that display long range dependence. *Bioinformatics*, 2006. Soumis pour publication.

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

[300] S. Nicolay and E.B. Brodie of Brodie. Échelles et fréquences caractéristiques des ondelettes dérivées de la gaussienne. Note interne, 2004.

[301] S.M. Nikol'skiĭ. Extension of functions of several variables preserving differential properties. *American Mathematical Society Translations (2)*, 83:159–188, 1969.

[302] S.M. Nikol'skiĭ. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer-Verlag, Berlin, 1975.

[303] E.A. Novikov and R. Stewart. Intermittency of turbulence and spectrum of fluctuations in energy-dissipation. *Izvestia Akademii Nauk SSR. Seria Geologia i Geofizika*, 3:408–412, 1964.

[304] N. Ogasawara and H. Yoshikawa. Genes and their organization in the replication origin region of the bacterial chromosome. *Molecular Microbiology*, 6:629–634, 1992.

[305] R.T. Ogden. *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Boston, 1997.

[306] C.D. Olds. *Continued Fractions*. Random House, New-York, 1963.

[307] S. Paixao, I.N. Colaluca, M. Cubells, F.A. Peverali, A. Destro, S. Giadrossi, M. Giacca, A. Falaschi, S. Riva, and G. Biamonti. Modular structure of the human lamin B2 replicator. *Molecular and Cellular Biology*, 24:2958–2967, 2004.

[308] G. Paladin and A. Vulpiani. Anomalous scaling laws in multifractal objects. *Physics Reports*, 156:147–225, 1987.

[309] G. Parisi and U. Frisch. Fully developed turbulence and intermittency. Dans M. Ghil, R. Benzi, and G. Parisi, editeurs, *Proc. of the International Summer School on Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*, pages 84–88, Amsterdam, 1985. North Holland.

[310] A. Pavlíček, J. Pačes, O. Clay, and G. Bernardi. A compact view of isochores in the draft human genome sequence. *Federation of European Biochemical Society Letters*, 511:165–169, 2002.

[311] G. Peano. Sur une courbe, qui remplit toute une aire plane. *Matematische Annalen*, 36:157–160, 1890.

[312] H.O. Peitgen, H. Jürgens, and D. Saupe. *Chaos and Fractals: New Frontiers of Science*. Springer, New-York, 2004.

[313] H.O. Peitgen and D. Saupe, editeurs. *The Science of Fractal Images*, New-York, 1987. Springer-Verlag.

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

[315] J.B. Perrin. Mouvement brownien et réalité moléculaire. *Annales de Chimie et de Physique*, 18:5–114, 1909.

[316] L. Pietronero and E. Tosatti, editeurs. *Fractals in Physics*, Amsterdam, 1986. North Holland.

[317] V. Pipiras. Wavelet-based simulation of fractional Brownian motion revisited. *Applied and Computational Harmonic Analysis*, 19:49–60, 2005.

[318] H. Poincaré. Pourquoi l'espace a trois dimensions. *Revue de Métaphysique et de Morale*, 20:489–504, 1912.

[319] M.G. Poirier, A. Nemani, P. Gupta, S. Eroglu, and J.F. Marko. Probing chromosome structure with dynamic force relaxation. *Physical Review Letters*, 86:360–363, 2001.

[320] M.K. Raghuraman, E.A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D.J. Lockhart, R.W. Davis, B.J. Brewer, and W.L. Fangman. Replication dynamics of the yeast genome. *Science*, 294:115–121, 2001.

[321] D.A. Rand. The singularity spectrum $f(\alpha)$ for cookie-cutters. *Ergodic Theory and Dynamical Systems*, 3:527–541, 1989.

[322] H.L. Resnikoff and R.O. Wells jr. *Wavelet Analysis*. Springer, New-York, 1998.

[323] R. Riedi. An improved multifractal formalism and self-similar measures. *Journal of Mathematical Analysis and Applications*, 189:462–490, 1995.

[324] B.D. Ripley. *Stochastic Simulation*. John Wiley & Sons, New-York, 1987.

[325] E.P. Rocha. Is there a role for replication fork asymmetry in the distribution of genes in the bacterial genomes? *Trends in Microbiology*, 10:393–395, 2002.

[326] E.P. Rocha. The replication-related organization of bacterial genomes. *Microbiology*, 150:1609–1627, 2004.

[327] E.P. Rocha and A. Danchin. Ongoing evolution of strand composition in bacterial genomes. *Molecular Biology and Evolution*, 18:1789–1799, 2001.

[328] E.P. Rocha, A. Danchin, and A. Viari. Universal replication biases in bacteria. *Molecular Microbiology*, 32:11–16, 1999.

[329] E.P. Rocha, P. Guerdoux-Jamet, I. Moszer, A. Viari, and A. Danchin. Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *Journal of Biotechnology*, 78:209–219, 2000.

[330] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, London, 1996.

[331] L.C.G. Rogers. Arbitrage with fractional Brownian motion. *Mathematical Finance*, 7:95–105, 1997.

[332] H.L. Royden. *Real Analysis*. Prentice-Hall, New Jersey, 1988.

[333] R. Rudner, J.D. Karkas, and E. Chargaff. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proceedings of the National Academy of Sciences of the USA*, 60:921–922, 1968.

[334] M.B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, editeurs. *Wavelets and their Applications*, Boston, 1992. Jones and Bartlett.

[335] S.L. Salzberg, A.J. Salzberg, A.R. Kerlavage, and J.F. Tomb. Skewed oligomers and origins of replication. *Gene*, 217:57–67, 1998.

[336] G. Samorodnitsky and M. S. Taqqu. *Stable Non-Gaussian Random Processes*. Chapman & Hall, New-York, 1994.

[337] D. Santamaria, E. Viguera, M.L. Martinez-Robles, O. Hyrien, P. Hernandez, D.B. Krimer, and J.B. Schvartzman. Bi-directional replication and random termination. *Nucleic Acids Research*, 28:2099–2107, 2000.

[338] S.C. Satchwell, H.R. Drew, and A.A. Travers. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*, 191:659–675, 1986.

[339] J. Schmets. Analyse mathématique, introduction aux espaces fonctionnels. Université de Liège, 1998–1999.

[340] J. Schmets. Théorie de la mesure. Université de Liège, 1998–1999.

[341] F. Sellan. Synthèse de mouvements browniens fractionnaires à l'aide de la transformation par ondelettes. *Comptes Rendus Mathématiques. Académie des Sciences. Paris*, 321:351–358, 1995.

[342] M. Semon, D. Mouchiroud, and L. Duret. Relationship between gene expression and *GC*-content in mammals: statistical significance and biological relevance. *Human Molecular Genetics*, 14:421–427, 2005.

[343] C. E. Shanon. Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, 37:10–21, 1949.

[344] C. Shioiri and N. Takahata. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *Journal of Molecular Evolution*, 53:364–376, 2001.

[345] B.D. Silberman and R. Linsker. A measure of DNA periodicity. *Journal of Theoretical Biology*, 118:295–300, 1986.

[346] Y.G. Sinai. Gibbs measure in ergodic theory. *Uspehi*, 27:21–64, 1972.

[347] A.F. Smit. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics & Development*, 6:743–748, 1996.

[348] A.F. Smit. Intersperced repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics & Development*, 9:657–663, 1999.

[349] A.F. Smit, R. Hubley, and P. Green. Repeatmasker open 3.0. http://www.repeatmasker.org.

[350] N.G. Smith, M.T. Webster, and H. Ellegren. Deterministic mutation rate variation in the human genome. *Genome Research*, 12:1350–1356, 2002.

[351] H.E. Stanley, S.V. Buldyrev, A.L. Goldberg, S. Havlin, S.M. Ossadnik, C.K. Peng, and M. Simmons. Fractal landscapes in biological systems. *Fractals*, 1:283–301, 1993.

[352] J.O. Strömberg. A modified Franklin system and higher order spline systems on $\mathbb{R}^n$ as unconditional bases for Hardy spaces. Dans W. Beckner, A.P. Calderòn, R. Fefferman, and P.W. Jones, editeurs, *Conference in Harmonic Analysis in Honor of Antoni Zygmund, Wadworth Math. Series*, pages 475–493. Wadsworth, 1983.

[353] P. Sudbery. *Human Molecular Genetics*. Addison Wesley, Singapour, 1998.

[354] N. Sueoka. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *Journal of Molecular Evolution*, 40:318–325, 1995.

[355] J.Q. Svejstrup. Mechanisms of transcription-coupled DNA repair. *Nature Reviews. Molecular Cell Biology*, 3:21–29, 2002.

[356] W. Sweldens and R. Piessens. Calculation of the wavelet decomposition using quadrature formulae. *Centrum voor Wiskunde en Informatica. Quaterly*, 5:33–52, 1992.

[357] E. Szpilrajn. La dimension et la mesure. *Fundamenta Mathematicae*, 28:81–89, 1937.

[358] T. Taira, S.M. Iguchi-Ariga, and H. Ariga. A novel DNA replication origin identified in the human heat shock protein 70 gene promoter. *Molecular and Cellular Biology*, 14:6386–6397, 1994.

[359] M.S. Taqqu. Weak convergence to fractional Brownian motion and to the Rosenblatt process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31:287–302, 1975.

[360] M.S. Taqqu. A bibliographical guide to self-similar processes and long-range dependence. Dans E. Eberlain and M.S. Taqqu, editeurs, *Dependence in Probability and Statistics*, pages 137–165, Boston, 1985. Birkhäuser.

[361] P. Tchamitchian and B. Torresani. Ridge and skeleton extraction from the wavelet transform. Dans M.B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, and Y. Meyer L. Raphael, editeurs, *Wavelets and Their Applications*, pages 123–151, Boston, 1992. Jones and Bartlett.

[362] H. Tennekes and J.L. Lumley. *A First Course in Turbulence*. MIT Press, Cambridge, 1960.

[363] E.R. Tillier and R.A. Collins. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *Journal of Molecular Evolution*, 50:249–257, 2000.

[364] V. Todorovic, A. Falaschi, and M. Giacca. Replication origins of mammalian chromosomes: the happy few. *Frontiers in Bioscience*, 4:859–868, 1999.

[365] B. Torresani. *Analyse Continue par Ondelettes*. CNRS Éditions, Paris, 1995.

[366] M. Touchon. *Biais de composition chez les mammifères : rôles de la transcription et de la réplication*. thèse de doctorat, Université Paris VII, 2005.

[367] M. Touchon, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acid Research*, 32:4969–4978, 2004.

[368] M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes. Transcription-coupled TA and GC strand asymmetries in the human genome. *Federation of European Biochemical Society Letters*, 555:579–582, 2003.

[369] M. Touchon, S. Nicolay, B. Audit, E.B. Brodie, Y. d'Aubenton Carafa, A. Arneodo, and C. Thermes. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proceedings of the National Academy of Sciences of the USA*, 102:9836–9841, 2005.

[370] C. Tricot. *Courbes et Dimensions Fractales*. Springer-Verlag, Berlin, 1999.

[371] H. Triebel. Characterization of Besov-Hardy-Sobolev spaces via harmonic functions, temperatures and related means. *Journal of Approximation Theory*, 35:275–297, 1982.

[372] H. Triebel. *Theory of Function Spaces*. Birkhäuser Verlag, Basel, 1983.

[373] H. Triebel. Characterization of Besov-Hardy-Sobolev spaces: a unified approach. *Journal of Approximation Theory*, 52:162–203, 1988.

[374] H. Triebel. *Theory of Function Spaces II*. Birkhäuser Verlag, Basel, 1992.

[375] A. Trivedi, S.E. Waltz, S. Kamath, and M. Leffak. Multiple initiation in the c-myc replication origin independant of chromosomal location. *DNA and Cell Biology*, 17:885–896, 1998.

[376] G.E. Uhlenbeck and L.S. Ornstein. On the theory of Brownian motion. *Physical Review*, 36:823–841, 1930.

[377] P. Urysohn. Mémoire sur les multiplicités cantoriennes. *Fundamenta Mathematicae*, 7:30–137, 1925.

[378] P. Urysohn. Über die Mächtigkeit zusammenhängender Mengen. *Matematische Annalen*, 94:262–295, 1925.

[379] P. Urysohn. Mémoire sur les multiplicités cantoriennes (suite). *Fundamenta Mathematicae*, 8:225–351, 1926.

[380] T. Vicsek. *Fractal Growth Phenomena*. World Scientific, Singapour, 1989.

[381] J. Ville. Théorie et applications de la notion de signal analytique. *Cables et Transmission*, 2A:61–74, 1948.

[382] R.F. Voss. Random fractals: characterization and measurement. Dans R. Pynn and A. Skjeltorp, editeurs, *Scaling Phenomena in Disordered Systems*, pages 1–11, New-York, 1985. Plenum Press.

[383] R.F. Voss. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Physical Review Letters*, 68:3805–3808, 1992.

[384] H.S. Wall. *Analytic Theory of Continued Fractions*. Chelsea Publishing, New-York, 1973.

[385] M.C. Wang and G.E. Uhlenbeck. On the theory of brownian motion II. *Review of Modern Physics*, 36:323–342, 1945.

[386] J.D. Watson and F.C.H. Crick. A structure of deoxyribonucleic acid. *Nature*, 171:737–738, 1953.

[387] B.J. West and W. Deering. Fractal physiology for physicists: Lévy statistics. *Physics Reports*, 246:1–100, 1994.

[388] J. Whittaker. Interpolatory function theory. *Cambridge Tracts on Mathematics and Mathematical Physics*, 33:107, 1935.

[389] D.V. Widder. *The Heat Equation*. Academic Press, New-York, 1975.

[390] N. Wiener. Differential-space. *Journal of Mathematics and Physics*, 2:131–174, 1923.

[391] A.P. Wolfe. *Chromatin Structure and Function*. Academic Press, London, 1995.

[392] G.K. Wong, D.A. Passey, and J. Yu. Is "junk" DNA mostly intron DNA? *Genome Research*, 10:1672–1678, 2000.

[393] A.T.A. Wood and G. Chan. Simulation of stationary Gaussian processes in $[0,1]^d$. *Journal of Computational and Graphical Statistics*, 3:409–432, 1994.

[394] A.M. Yaglom. The influence of the fluctuation in energy dissipation on the shape of turbulent characteristics in the inertial interval. *Soviet Physics–Doklady*, 2:26–30, 1966.

[395] A.L. Yuille and T.A. Poggio. Scaling theorems for zero crossings. *Institute of Electrical and Electronic Engineers. Transactions on Pattern Analysis and Machine Intelligence*, 8:15–25, 1986.

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

Index

D

E

F

G

H

I

L

M

N

O

P

R

S

T

U

V