N° d'ordre :

N° attribué par la bibliothèque :

HABILITATION À DIRIGER DES RECHERCHES

CNRS - École Normale Supérieure de Lyon

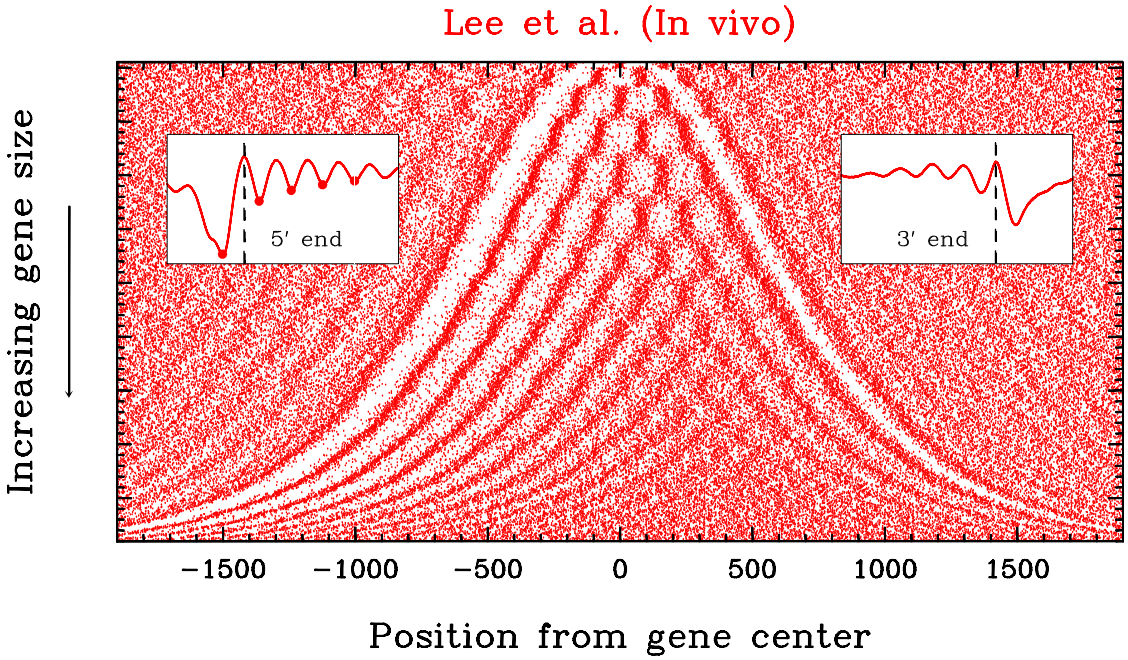
LABORATOIRE JOLIOT CURIE

Section CNU 64

Cédric VAILLANT

Titre :

CHAPELET NUCLÉOSOMAL ET ORGANISATION DU GÉNOME



Positionnement des nucléosomes et organisation du génome

Cédric Vaillant

31 août 2011

TABLE DES MATIÈRES

1 Introduction

En guise d'introduction, je commencerai par une briève présentation de mon parcours. Après des classes préparatoires en physique à Orléans, j'ai intégré Supélec. J'ai pu tout de même faire un DEA de physique théorique en troisième année d' Ecole d'ingénieur (DEA "Champ, particules, matières" à Orsay). Après Supélec et ce DEA j'ai eu l'opportunité de faire un service de coopération scientifique en Mongolie, à Oulan-Bator, au sein de centre de sismologie où je travaillais pour le compte du CEA. Mon travail consistait à la veille scientifique du réseau de capteurs sismique mis en place par le CEA. A mon retour j'ai intégré l'équipe d'Alain Arnéodo au Centre de Recherche Paul Pascal à Pessac pour faire une thèse sous sa direction ; l'objectif était d'étudier les propriétés de "corrélation à longue portée" dans les génomes fraichement séquencés comme celui de *S. Cerevisiae* et les génomes bactériens comme *E. Coli* et ainsi de prendre le relais de Benjamin Audit qui s'était plus spécifiquement focalisé sur l'étude de corrélations dans les introns/exons. L'interprétation chromatinienne que nous en fimes me poussa à réorienter mon travail sur l'étude théorique de l'influence de la séquence et notamment de l'influence de corrélation à longue portée dans la séquence sur les propriétés élastiques de longues chaîne ADN. Ce fut une très bonne expérience, tant scientifique que personnelle ; le CRPP était une unite propre avec un spectre "scientifique" très large et une certaine forme de mutualisation des moyens, notamment financiers. Je suis allé ensuite en post-doctorat à l'EPFL, à Lausanne dans le groupe de J. Maddocks où j'ai commencé à travailler sur les problèmes d'effets de séquences et de corrélations sur des systèmes de boucles d'ADN. Après deux ans et quelques échec au concours du CNRS j'ai rejoint le Laboratoire Statistique et Génome de Bernard Prum à la génopole d'Evry où j'ai débuté la modélisation de l'effet de la séquence sur le positionnement des nucléosemes. J'ai été recu au concours de la section 44 (c'était la commission interdisciplinaire "modélisation du vivant") en tant que CR1 pour intégrer le Laboratoire Joliot-Curie. C'était donc en Septembre 2006. Depuis mes recherches ont principalement tourné autour de l'organisation du chapelet nucléosomal et, pour faire original, autour des effets de séquence. C'est ce travail que je vais présenter ici avec un cheminement qui intégrera la contribution d'autres groupes et laboratoires. L'objectif était de faire un état de l'art dans ce domaine mais il reste du travail de clarification et de synthèse avant que ce soit vraiment satisfaisant. Je compte bien l'améliorer donc et toutes suggestion à l'occasion de cette HDR seront les bienvenues. Une grande partie du travail présenté ici est issue de la thèse de G. Chevereau que j'ai eu l'opportunité de diriger de 2007 à 2010 ("Thermodynamique du positionnement des nucléosomes", Université de Lyon-ENS de Lyon). Mes recherches scientifiques, de ma thèse jusqu'à maintenant se sont faits en collaboration avec A. Arneodo, B. Audit, F. Argoul, C. Thermes et Y. d'Aubenton-Carafa, travaux de collaboration qui nous avons compilés et publiés très récemment dans *Physics Reports*, **498**, 45-188 (2011) : "Multi-scale coding of genomic information : From DNA sequence to genome structure and function."

L'ADN chromosomique des cellules eucaryotes est fortement empaqueté au sein d'un complexe nucléoprotéïque, la chromatine (Fig. 1.1). Le premier niveau de compaction, le nucléosome, correspond à un enroulement de 146 paires de bases autour d'un octamère d'histone. L'arrangement linéaire de ces nucléosomes le long de la chaine ADN forme le chapelet nucléosomal ou "fibre de 10 nm". Cette fibre, notamment grâce à la fixation d'histones de liaison peut se condenser en une structure plus compacte, "la fibre de 30 nm", qui elle-même, à plus grande échelle, peut adopter une organisation en boucles stabilisée par des pontages protéiques (par ex., CTCF, cohésine). Par ailleurs que ce soit au niveau de l'ADN, avec la méthylation, ou au niveau des histones, avec les modifications covalentes des queues ou l'insertion de variants, la chromatine se caractérise par des signatures biochimiquex (épigénétiques) le long des génomes. Or, ces modifications biochimiques sont impliquées soit directement dans la structuration de la fibre (par exemple en modulant la stabilité des nucléosomes, ou l'interaction entre nucléosomes..) soit dans le recrutement de facteurs auxiliaires comme les remodelleurs et/ou des protéïnes de type HP1/Polycomb. Il apparait ainsi, qu'aussi bien l'organisation spatiale que la composition biochimique de la chromatine, en modulant l'accessibilité des différents complexes enzymatiques à leurs sites nucléiques joue un rôle fondamental dans la régulation (spécification et maintenace) des programmes transcriptionnels et réplicationnels ainsi que dans les processus de recombinaison, transposition ou insertion virale... Identifier, modéliser et tester expérimentalement certains mécanismes "chromatiniens" de régulation de ces métabolismes nucléaires en relation avec l'organisation des génomes fut l'objectif principal de mon projet de recherche à mon entrée au CNRS en 2006 au sein du Laboratoire Joliot-Curie.

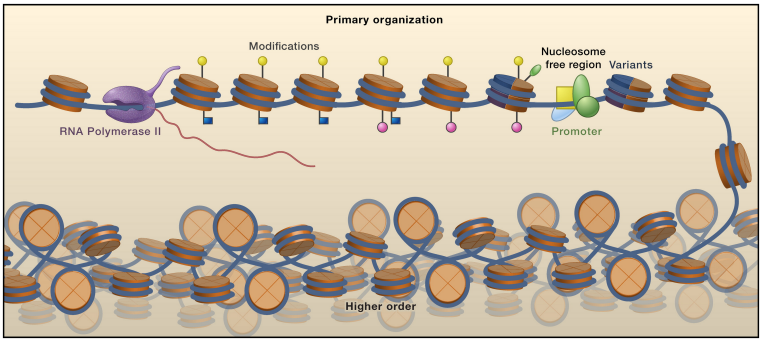


FIGURE 1.1 : *La structure primaire de la chromatine peut être vue comme un collier de perles avec des nucléosomes espacés périodiquement en aval du site d'initiation de la transcription. A l'exception de sites spécifiques de régulation, les nucleosomes sont absebts du couer du promoteur, là où la "machinerie" de transcription s'assemble. Ces regions "libres" de nucléosomes (NFR, pour "Nucleosome Free Regions") offre la possibilité de réguler l'expression des gènes à un niveau qui dépasse le simple accès du promoteur, par exemple en contôlant l'élongation de la polymerase ARN II ("pol II") (Core and Lis, 2008). Le coeur protéïque des nucléosomes est composé d'histones, qui sont souvent sujets à des modifications biochimiques "post"-traductionnelles à certains acides aminés spécifiques et qui peuvent être remplacés par des varaiants d'histones au cours de la redéposition suite à l'élongation par pol II (bleu foncé et violet). Comme représenté, la chromatine peut se condenser en une structure plus compacte grâce à certaines modifications, notamment l'acétylations des queues des histones H3,H4.*

2 FIBRE DE 11 NM

2.1 CHAPELET : POSITIONNEMENT, DISTANCE INTER-NUCLÉOSOME

Ces dernières années ont vu émerger des techniques expérimentales permettant la cartographie de plus en plus précise de la structure et composition de la chromatine le long des génomes et en particulier du positionnement et de l' occupation en nucléosome, canoniques et variants, de l'enrichissement en marques épigénétiques, en protéines chromatiniennes structurales telles HP1/PolyComb, insulateurs, cohesines, en régulateurs chromatiniens "actifs" tels les remodeleurs et en complexes fonctionnels tels

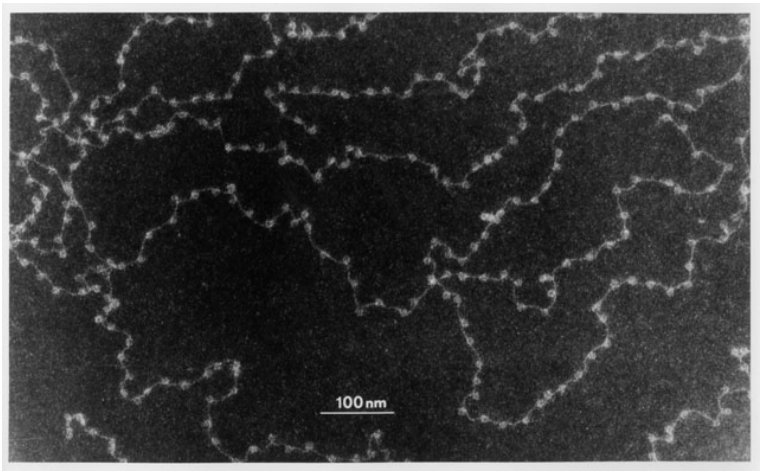


FIGURE 2.1 : *Image de microscopie électronique du chapelet nucléosomal.*

les polymerases... Dans ce manuscrit, il est essentiellement question de positionnement de nucleosome et donc de structure primaire du chapelet nucléosomal : localement le nucléosome réduit l'accessibilité des protéines à leurs sites de fixation et donc une première question qu'on peut se poser est connaissant un site particulier quelle est la probabilité pour qu'il soit occupé ? On s'interessera donc à l'occupation en nucléosome le long des génomes ; par ailleurs la distance entre nucléosomes successifs agit sur la compaction spatiale du chapelet et donc sur l'accessibilité à une plus grande échelle : on peut donc aussi s'intéresser à la distance typique entre deux nucléosomes successifs le long des génomes. On verra comment théoriquement on peut extraire ces deux types de cartes génomiques.

2.1.1 Méthodes expérimentales

Cette partie traite des méthodes d'extraction et d'analyse des cartes chromatiniennes et est extraite de (Zhang et Pugh (2011) (Zhang and Pugh, 2011)) et décrit en détail ce qui est illustré en figure 2.2 :

"**Initial Preparation of Mononucleosomes**

From an operational perspective, there are two types of starting material for mapping nucleosomes : tissue excised from a multicellular eukaryotic organism and minimally aggregated cells, including those drawn from blood, grown in tissue culture, or cultured as free-living microorganisms. Excised tissue may contain a heterogeneous mixture of cells, which may obscure chromatin patterns specific to a cell type. Cellular heterogeneity may be minimized by highly selective and precise tissue excision, which may necessitate acquisition of less material. Although the minimum amount of excised material required to generate nucleosome maps is not known, a lower limit of not, vert, similar $\sim$ 10000 cells drawn from blood may provide a guide (Adli et al., 2010 M. Adli, J. Zhu and B.E. Bernstein, Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors, Nat. Methods 7 (2010), pp. 615). More commonly, large numbers of cells are easily collected from tissue culture or microorganisms, such as yeast for which $10^7 - 10^8$ cells are used. More sophisticated methods may be used to isolate tissue-specific nuclei (Deal and Henikoff, 2010).

The production of genome-wide nucleosome maps has variously used or avoided formaldehyde crosslinking (Fig. 2.2, step 1). Formaldehyde essentially "freezes" existing protein-protein and protein-nucleic acid interactions in place, thereby preserving the in vivo status of interactions, without adverse effects on nucleosomes (Fragoso and Hager, 1997). Without crosslinking, nucleosomes may reorganize during cell harvesting, chromatin preparation, and chromatin fragmentation. However, for most genes in yeast, we and other laboratories have found that nucleosome organization is largely the same in the presence or absence of formaldehyde when MNase is used for chromatin fragmentation (Kaplan et al., 2009). Nevertheless, we have also found genomic regions where nucleosome organization varies in the absence of formaldehyde, and thus, we recommend a simple formaldehyde crosslinking step.

Yeast and plants have cell walls, which require disruption through mechanical breakage (e.g., vigorous vortexing with glass beads) or enzymatic digestion ([Albert et al., 2007] and [Rando, 2010]) (Fig. 2.2, step 2). Tissue or small whole animals, such as worms, may be disrupted by grinding of frozen material (Kolasinska-Zwierz et al., 2009). Tissue culture cells in sufficient quantities may be disrupted by douncing cells in a hypotonic buffer. If the amount of material is low, then it may be more practical to lyse with an ionic detergent, such as SDS, in combination with a freeze-thaw cycle (Adli et al., 2010 M. Adli, J. Zhu and B.E. Bernstein, Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors, Nat. Methods 7 (2010), pp. 615). However, because SDS disruption is not compatible with subsequent MNase digestion, the chromatin must be fragmented by low-resolution sonication.

Overview of Available Strategies

Data from MNase ChIP-Seq provide population averages from a large number of cells. To map nucleosome configurations on individual DNA molecules, ectopic expression of DNA methyltransferases may be used in vivo or added to nuclei. Nucleosomal DNA is identified because its sequence is eventually altered, whereas linker DNA is not. For example, the M.CviPI methyltransferase methylates cytosine in 5′-GC-3′ dinucleotides when it is present in linker DNA (Pardo et al., 2010). This methyl-cytosine, unlike cytosine, is protected against bisulfite conversion to uracil (and ultimately thymine) in vitro. After traditional Sanger sequencing, the configuration of a nucleosomal array on the original DNA molecule is inferred from the sequence. GC dinucleotides are inferred to be nucleosome-free, whereas $\sim$ 150 base-pair spans of GTs that are GCs in the reference (i.e., untreated) genome are interpreted as nucleosomal.

The value of mapping an array of nucleosomes on a single DNA molecule is that adjacent nucleosome positions may appear to overlap in a population average but are actually mutually exclusive when examined on a single-molecule basis. Currently, this strategy has not been applied on a genomic scale, as it optimally requires high-throughput long-read (>1000 nucleotides) sequencing.

Regions depleted of nucleosomes are candidates for regulatory regions. Therefore, if the primary purpose of chromatin mapping is to screen for nucleosome-depleted regions in many cell types or under various conditions, then FAIRE (formaldehyde-assisted isolation of regulatory elements) may be the appropriate method (Giresi and Lieb, 2009). FAIRE simply depends upon the differential partitioning of nucleosomal and nucleosome-free DNA in

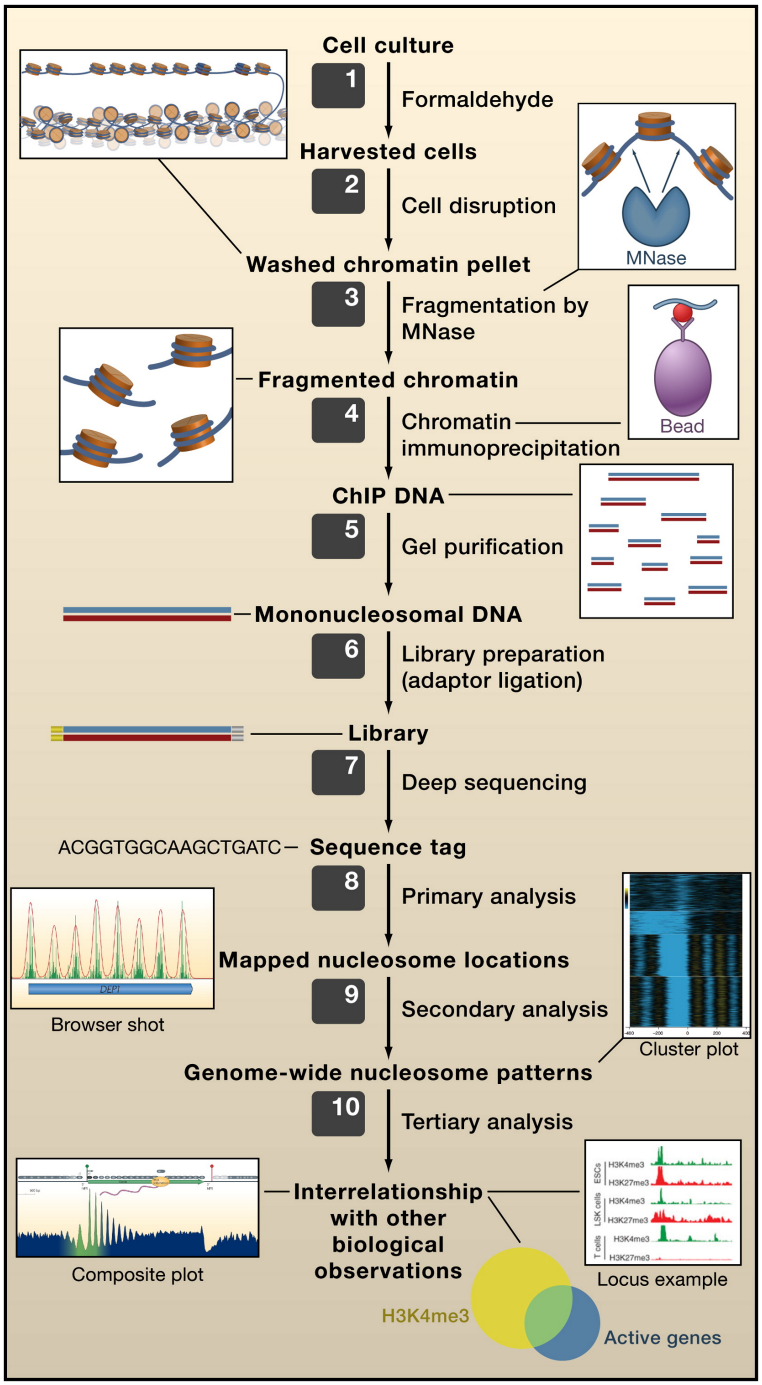


FIGURE 2.2 : *Schéma illustrant les différentes étapes associé à la cractérisation du chapelet nucléosomal : de la préparation des nucléosomes à l'analyse en passant par la cartographie*

phenol-chloroform and aqueous phases. Thus, the major advantages of FAIRE are its simplicity and cost efficiency. However, its resolution is low compared to mapping the location of individual nucleosomes.

DNase I hypersensitivity has been a classical means of mapping regions of accessible chromatin. Like FAIRE, it is a strategy used on a genomic scale in the ENCODE project (Hesselberth et al., 2009). However, due to frequent cleavages within nucleosomal DNA, nucleosome positions may be more difficult to discern when DNase I is used instead of MNase. Moreover, DNase I involves many sample handling steps and is complicated by technical variation in DNA digestion.

Chromatin Fragmentation

The method of chromatin fragmentation is critical to producing nucleosome maps of a desired resolution (Fig. 2.2, step 3). Sonication produces DNA fragments ranging from $\sim$ 200 to $\sim$ 700 base pairs. The heterogeneity of fragment size and cleavage sites makes sonication suitable for characterizing chromatin states over wide regions encompassing many nucleosomes, but it is not optimal for mapping individual nucleosomes. MNase digestion, on the other hand, produces DNA fragments with ends that correspond to the ends of nucleosomes and, thus, produces maps with very high resolution.

One potential limitation of MNase digestion is its bias toward cleaving at A or T more frequently than at G or C. However, extensive MNase digestion that predominantly produces mononucleosomes largely, but not entirely, overcomes this bias because even unfavorable cleavage sites become cleaved. Furthermore, residual bias can be computationally compensated (Albert et al., 2007). A limitation of extensive MNase digestion is the production of subnucleosomal-sized DNA fragments, particularly at highly transcribed genes where the DNA on the surface of remodeled or partially disassembled nucleosomes may be more exposed (Weiner et al., 2010 A. Weiner, A. Hughes, M. Yassour, O.J. Rando and N. Friedman, High-resolution nucleosome mapping reveals transcription-dependent promoter packaging, Genome Res. 20 (2010), pp. 90). The lack of nucleosome-sized DNA fragments in such regions may be interpreted as being entirely nucleosome free, as opposed to the presence of remodeled or partial nucleosomes that escape detection.

Different chromatin samples and preparations of MNase (i.e., commercial lots) may yield different degrees of MNase digestion. Therefore, it is prudent to titrate the MNase to achieve $\sim 80\%$ of the DNA as mononucleosomal, which is detected by electrophoresis as a band at $\sim$ 150 base pairs (Fig. 2.2, step 3) (Rando, 2010). In addition, pooling chromatin that has been fragmented to various extents from an MNase titration may help avoid biased isolation of mononucleosome subpopulations that differ in accessibility.

Fragmentation by sonication releases insoluble chromatin fragments from the pellet to the supernatant. MNase treatment solubilizes mononucleosomes in yeast but is often less efficient in fly and mammalian systems. Therefore, a brief sonication in these latter two systems improves solubilization, without creating additional fragmentation. Alternatively, salt extractions of increasing strength can be used to selectively solubilize "active" chromatin (Henikoff et al., 2009). Gel analysis of histones and DNA released to the supernatant versus that retained in the pellet can be conducted to confirm full extraction.

Chromatin Immunoprecipitation

Perhaps the most frequent use of nucleosome mapping is to characterize the distribution of histone modification states or histone variants. In these cases, immobilized antibodies against the particular modification or variant are necessary to immunoprecipitate (or "ChIP") chromatin fragments possessing the specific modification or variant (Fig. 2.2, step 4) (Liu et al., 2005). Because only a small percentage of DNA becomes crosslinked to histones by formaldehyde, immunoprecipitation should be conducted in the presence of detergent (e.g., 0.05% SDS) to eliminate uncrosslinked DNA. Many antibodies, such as those against H3K4me3 and H2A.Z, are commercially available, providing a level of standardization and quality control of antibody specificity. However, one limitation of any antibody targeted against a modification is its potential to cross-react with the same or a similar modification located at other sites. Alternatively, an antibody may not recognize its epitope if a nearby amino acid is also modified, and such interfering modification might be present in only a subpopulation of the nucleosomes. Synthetic peptides harboring the modification or potentially confounding secondary modifications can be used to verify antibody specificity.

Detection

Historically, genome-wide detection of chromatin began with the use of low-resolution DNA microarrays in yeast. PCR probes of each intergenic and genic region were arrayed onto glass slides upon which fluorescently labeled ChIP material was hybridized (reviewed in Jiang and Pugh, 2009). Higher resolution was achieved with microarrays containing overlapping 50-nucleotide probes tiled every 20 base pairs across a small region of the yeast genome. Next high-density microarrays that spanned entire genomes were developed. These arrays, which remain in use today probably for a limited time, can generate maps of individual nucleosomes but with lower resolution compared to deep sequencing. Deep sequencing has the additional advantages of less background, better coverage, and a larger dynamic range compared to microarrays. That said, the fuzziness of nucleosome positions over a population precludes full realization of deep sequencing's intrinsic high resolution.

Regardless of the fragmentation method or whether ChIP is used, the resulting DNA should be gel purified in the $120 - 170$ base-pair range to remove nonspecific, subnucleosomal, and polynucleosomal DNA fragments (Fig. 2.2, step 5). Currently, deep sequencing of nucleosomal DNA requires library preparation, which essentially involves ligating DNA adapters to the ends of gel-purified mononucleosomal DNA (Fig. 2.2, step 6). This allows

for PCR amplification of the sample and creates a template by which sequencing initiates. By this stage, users typically have given their samples to a sequencing facility, which will construct the libraries for sequencing using kits provided by the manufactures of the sequencing instrument (Fig. 2.2, step 7). Research laboratories that produce large numbers of libraries may develop their own library preparation protocols, which enhance cost efficiency. Adaptor sequences are available at company websites, and their ligation involves standard molecular biology manipulations. In this case, greater DNA yields may be obtained by gel purifying after library preparation and PCR amplification.

Currently, the Illumina Genome Analyzer and the Applied Biosystems SOLiD sequencers are the most widely used deep sequencers for this type of work. Although a variety of deep sequencers will likely be available in the near future, the key instrument parameter for nucleosome mapping (and for ChIP-Seq in general) is not the read length but rather the tag count, which is the number of different DNA molecules that can be sequenced and mapped to the reference genome. In general, a technology platform should meet these minimum specifications : minimal steps for library construction, a sequencing read or tag length of ~ 35 nucleotides, read accuracy of $> 99\%$, turnaround time of less than a few days.

In principle, biases during ligation, post-construction PCR amplification, gel purification, and sequencing can result in biased tag production, which may influence the apparent occupancy level and position of nucleosomes (Stein et al., 2010). Nevertheless, such biases may be compensated computationally because they manifest as anomalously high tag counts at specific genomic coordinates. For example, setting an upper limit on the normalized tag counts at a particular coordinate may correct such statistical outliers (Kaplan et al., 2009). In practice, sequencing bias may be rather innocuous because data are often aggregated in a way that eliminates outliers.

Sequencing Tags

Sample processing that includes MNase digestion, immunoprecipitation, and gel purification of mononucleosomes eliminates nonspecific background contamination of genomic DNA, which would otherwise degrade the quality of the maps. As such, each sequencing tag represents a measured nucleosome position, generally without the need for background correction.

A general rule of thumb is that the number of sequencing tags needed to uniquely identify $> 90\%$ of all nucleosomes is minimally ten times the number of estimated nucleosomes. An estimated number of total nucleosomes is the genome size divided by 200 (i.e., the average base-pair distance covered by a nucleosome core particle plus linker). Thus, complete yeast nucleosome maps require at least 600000 tags, whereas human nucleosome maps require at least 150 million tags.

However, more or fewer tags may be needed depending upon the goal of the experiment. If the goal is to measure occupancy levels, then $\sim 3-5$ times more tags may be required to provide robust quantitative numbers of tags per nucleosome position or per genomic coordinate. If data are to be aggregated, for example by averaging the distribution of tags around a collection of genes, then substantially fewer tags may be sufficient. Indeed, not every nucleosome would need to be detected. Such minimal coverage is cost efficient when many experiments are conducted simultaneously, such as screening samples or titrating conditions. Because only a small portion of the library is sequenced, more coverage can be achieved by sequencing more of the library as needed. Similarly, histone modification states typically occur at only a fraction of all nucleosomes and, thus, in principle, require fewer tags.

The number of needed sequencing tags for each sample must dovetail with the minimal sequencing "bandwidth". Each channel of the sequencer flow cell (the current Illumina sequencer has 8 channels and the current SOLiD sequencer has 1-8 channels) represents the minimal bandwidth of the sequencer. If a sequencer delivers, for example, 40 million mappable tags as its minimal bandwidth per channel, and the user requires ~ 10 million tags per sample, then 4 multiplexed samples can be placed into each channel. Sample multiplexing, which is also called indexing or barcoding, is achieved by using commercially designed adapters. These adapters contain a unique predefined $5-10$ nucleotide DNA sequences used to identify the sample."

2.1.2 Données expérimentales

Dans ce qui suit nous traiterons essentiellement de distribution de nucléosomes le long des génomes obtenues à partir des différentes méthodes que l'on vient de présenter. Les figures 2.5, 2.6, 2.7, 2.8, 2.9, 2.10, 2.11 et 2.12 reportent les différentes cartographies des nucléosomes obtenues le long des génomes de :

(i) ***S. cerevisiae*** **:** cette levure, dite levure du boulanger, comporte 16 chromosomes, de taille allant de $230000\ pb$ (chromosome 1) à $1400000\ pb$ (chromosome 4) pour au total un génome de $12\ Mpb$. Parmi les premières expériences haute résolution et à une échelle chromosomique retenons celles de Yuan *et al.* (Yuan et al., 2005) sur le chromosome 3 de la levure 2.6 ; c'est une expérience de MNase-chip, avec un pavage d'oligos de 50 pb toutes les 25 pb, avec deux canaux d'hybridation, un d'ADN chromatinien mononucleosomal, l'autre d'ADN nu extrait par digestion MNase également. Vinrent ensuite les données MNase-chip de Lee et al. (Lee et al., 2007a), sur l'ensemble du génome de la levure avec un pavage d'oligo de $25\ bp$ toutes les $8\ pb$ (Fig. 2.5) sur chaque brin

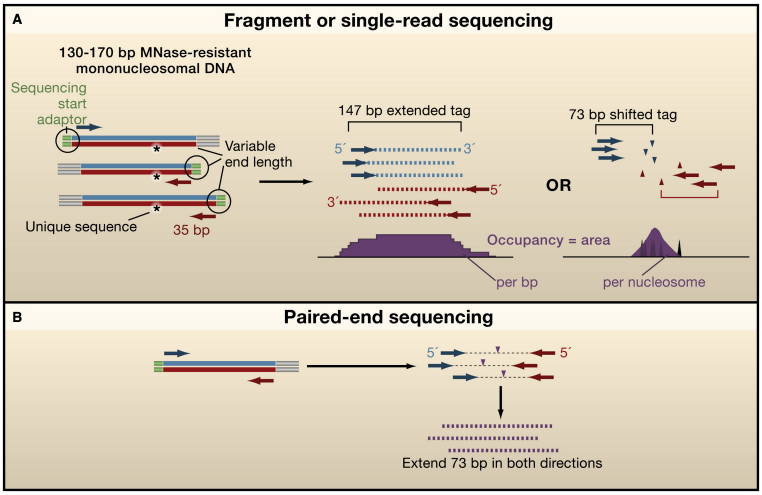


FIGURE 2.3 : *Turning Tags into Nucleosomes : Solid red and blue lines represent nucleosomal DNA ready for sequencing. The nucleosomal DNA is produced as MNase-resistant DNA fragments. In a population of molecules, the DNA fragments will have heterogeneous ends due to biases in digestion efficiency at different sequences, as well as the nucleosome not residing at a single position (i.e., "fuzzy" positioning). The asterisk, representing a unique sequence, provides a frame of reference. (A) In "fragment" or "single-read sequencing", the DNA library is sequenced from only one of the adapters (green) (except in reading the barcode) and in the direction indicated by the blue or red arrows. Consequently, each nucleosome border is measured independently as a population. Tags can then be extended to 147 nucleotides or their 5′ ends shifted by 73 nucleotides, as indicated to the right. Either way, the resulting frequency distributions, although looking different, have exactly the same uncertainty. (B) Paired-end sequencing allows both ends of the same DNA molecule to be sequenced. The midpoint of the pair defines the consensus nucleosome midpoint, which can be extended 73 nucleotides in both directions (right side).*

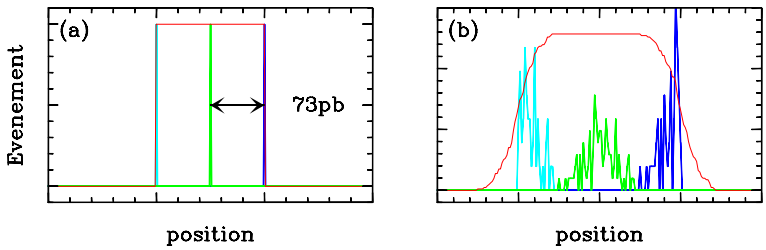


FIGURE 2.4 : *Comment établir la probabilité d'occupation à partir de tags de séquençage. (a) Cas idéal : pour chaque nucléosome positionné, on relève un tag à l'entrée du nucléosome (bleu clair), sur le brin (+) et un tag en sortie (en bleu foncé) sur le brin (-). On en déduit la position de la dyade (et donc la densité) en rajoutant (ou en enlevant) 73 pb (en vert). La probabilité d'occupation (en rouge) est obtenue en convoluant la densité par une fenêtre rectangulaire d'épaisseur 146pb. (b) Simulation d'évènements réels de séquençage. Même si l'on considère un nucléosome extrêmement bien positionné, les tags peuvent ne pas commencer strictement au début ou à la fin (du fait par exemple du rognage de la MNase), on obtient une distribution de tags autour de l'entrée et de la sortie. Le positionnement de la dyade (en vert) que l'on déduit est d'autant délocalisé, et la probabilité de positionnement (en rouge) est étalée.*

(avce un décalage de 4 pb entre brins). Les données nucléosomales sont normalisées par une expérience contrôle correspondant à la digestion d'ADN génomique. Notons aussi dans la même veine que les étude de Yuan, celles menées par Whitehouse et al. (Whitehouse et al., 2007), toujours par MNase-Chip sur l'ensemble du génome et dans deux types de souches de levure, la sauvage et une mutante qui exprime une protéine Iswi2 non catalytique. Les premières données de MNase-Seq sur l'ensemble de la levure furent celles de Albert *et al.* (Albert et al., 2007), qui se concentrèrent sur la distribution du variant H2A.Z (Htz1). La couverture est assez réduite avec au maximum 25 reads par nucléosome H2AZ. Depuis, les données se succèdent, avec Iyer et al. (Shivaswamy et al., 2008) qui obtiennent une carte des nucléosomes avant et après application d'un choc thermique pour étudier dans quelle mesure la dynamique transcriptionnelle induite (activation ou répression) est couplée à la dynamique du profil nucléosomal. Notons ensuite celles de Mavrich et al. (Mavrich et al., 2008a) avec une couverture mediane de 14 "reads" par nucléosome ; celles de Kaplan *et al.* (Kaplan et al., 2009a), pour différentes conditions de croissance (milieu riche avec glucose, milieu avec ethanol et avec galactose) avec une couverture d'environ 100 "évènements" ("reads") par nucléosomes ; et celles de Weiner *et al.* (Weiner et al., 2010) qui ont obtenu des cartes à différents niveaux de digestion MNase de la chromatine ainsi que dans une souche où l'inactivation de la polII peut être commandé par une hausse de température ; ces derniers ont ainsi pu étudier l'effet de la polymerase polII sur la structuration du chapelet nucléosomal.

(ii) ***Levures Hemiascomycota*** : Une cartographie exhaustive des nucléosomes par MNase-seq a été récemment obtenue par le groupe de O. Rando sur un grand nombre de levure du groupe Hemiascomycotes (Tsankov et al., 2010). L'objectif était d'étudier dans quelle mesure les divergences d'expression (ou de stratégies de régulation) de gènes orthologues sont associées à des différences d'organisation du chapelet (notamment au niveau des promoteurs) elle-mes induites soient par des changement dans la séquence génomique (en "cis") soient par des modifications du système de régulation moléculaire (en "trans").

(iii) ***S. Pombe*** : La première carte génomique chez *S. Pombe* a été obtenue par Lantermann *et al.* (Lantermann et al., 2010) par digestion MNase puis hybridation sur puces avec une résolution de $25\ pb$. La cartographie a été faite sur une souche sauvage et sur une souche mutante n'exprimant pas le facteur de remodelage MIT1 (remodelleur de type Snf2).

(iv) ***C. Elegans*** : Valouev et al. (Valouev et al., 2008), ont cartographié les nucléosomes le long du génome de *C. Elegans* par MNase-seq, avec une couverture de l'ordre de 70 "évènements" (reads) par nucléosome. L'expérience contrôle correspond à la digestion légère de l'ADN génomique résultant en un jeu de séquences de taille $\sim 400 - 800pb$. Les données d'occupation (ie convolution des données de tags par une fenêtere de taille $146\ pb$) présentées en Fig. 2.12 (vert) ont été normalisées par ces données de contôle.

(v) ***D. Melanogaster*** : La mouche est un autre organisme modèle très important ; dans notre étude Miele et al. (Miele et al., 2008), nous avons étudié les données MNase-chip de Mito *et al.* (Mito et al., 2005, 2007) d'une résolution de $100\ pb$. Deux cartographies plus récente des nucléosomes (le tout venant) par MNase-chip avec une résolution de $36\ pb$ et des nucléosomes contenant H2A.Z par MNase-Chip-seq cette fois-ci ont été extraites par Mavrich *et al.* Mavrich et al. (2008a).

(vi) ***Homme*** : Les données génomiques auxquelles nous nous interesserons dans ce manuscrit sont celles obtenues par Schones et al. (Schones et al., 2008) par MNase-seq avec une couvertaure d'à peu près 10 reads par nucléosomes et ce dans deux types de cellules T CD4+ : "non activées" ou "activées".

Ce que révèle l'ensemble de ces données, c'est que le positionnement est globalement hétérogène le long des génomes. Nous avons la confirmation que l'arrangement nucléosomal n'est pas une assemblée régulière de nucléosomes. Il s'agit plutôt d'une alternance de zones vides de nucléosomes, de taille $\sim 100 - 200\ pb$, bordées par des régions bien organisées, avec un positionnement systématique des nucléosomes les uns à côté des autres, et de zones remplies de nucléosomes mais mal organisées. Lorsqu'une zone est occupée par des nucléosomes, mais sans que cette occupation soit systématiquement la même d'une cellule à l'autre, alors le profil de positionnement mesuré par ces techniques (positionnement "moyen") sera flou. On remarque par ailleurs que les différents profils d'occupation obtenues chez *S. Cerevisiae* par les différentes équipes ne coincident pas exactement : la différence apparaît d'ailleurs essentiellement au niveau de l'occupation et moins au niveau des positions qui semblent finalement assez bien conservées (Figs. 2.6 et 2.7). Ce qu'on remarque également c'est que le positionnement à savoir

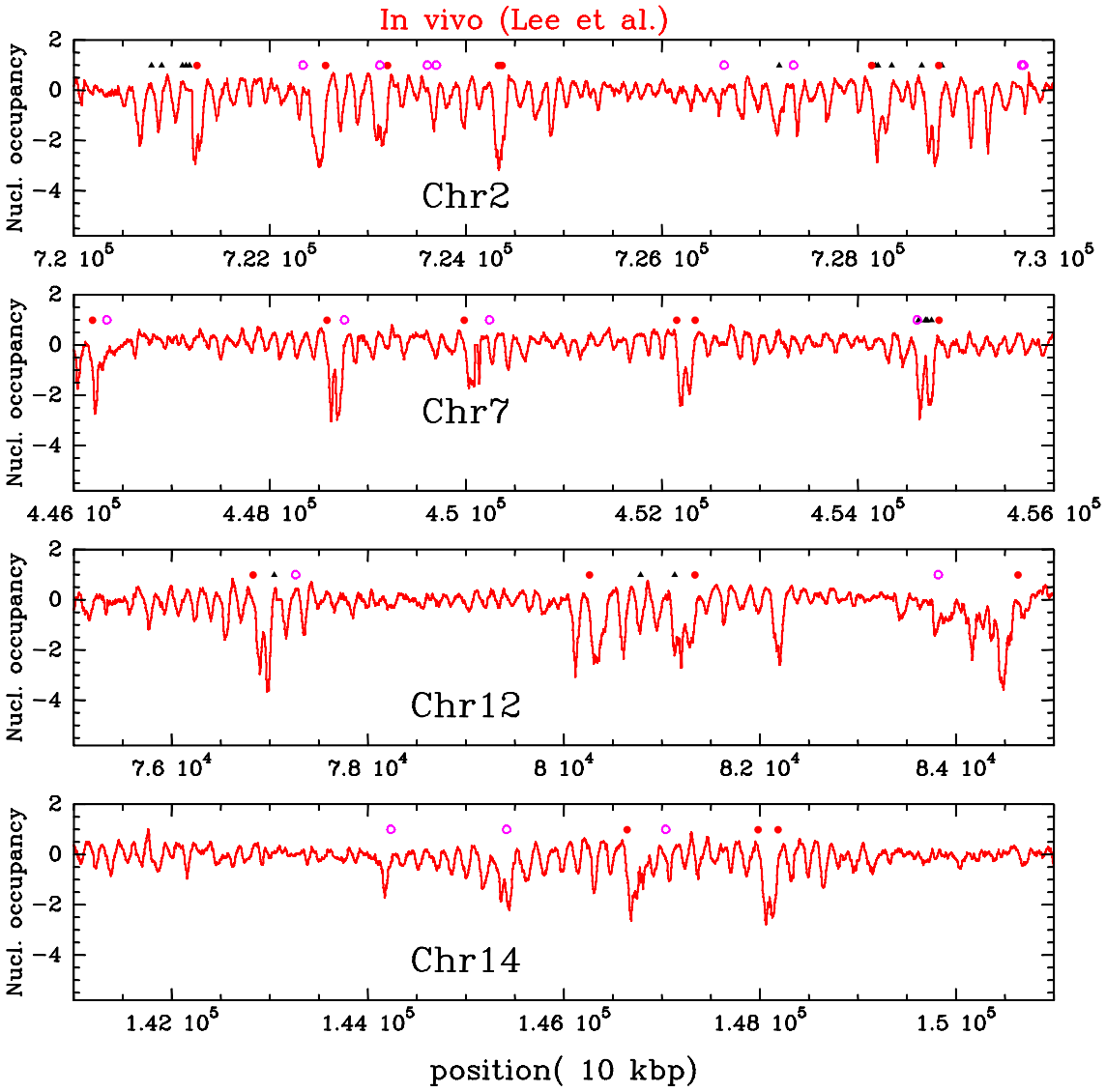


FIGURE 2.5 : *Cartes nucléosomales expérimentales obtenues par MNase-chip chez* S. Cerevisiae *par Lee* et al. *(Lee* et al.*, 2007*a*). log_2 du signal d'hybridation le long de $10kpb$ des chromosomes (a) 2, (b) 7, (c) 12 et (d) 14. Les ronds rouges désignent les TSS, les cercles roses, les TTS et les triangles, des sites de facteurs de transcription*

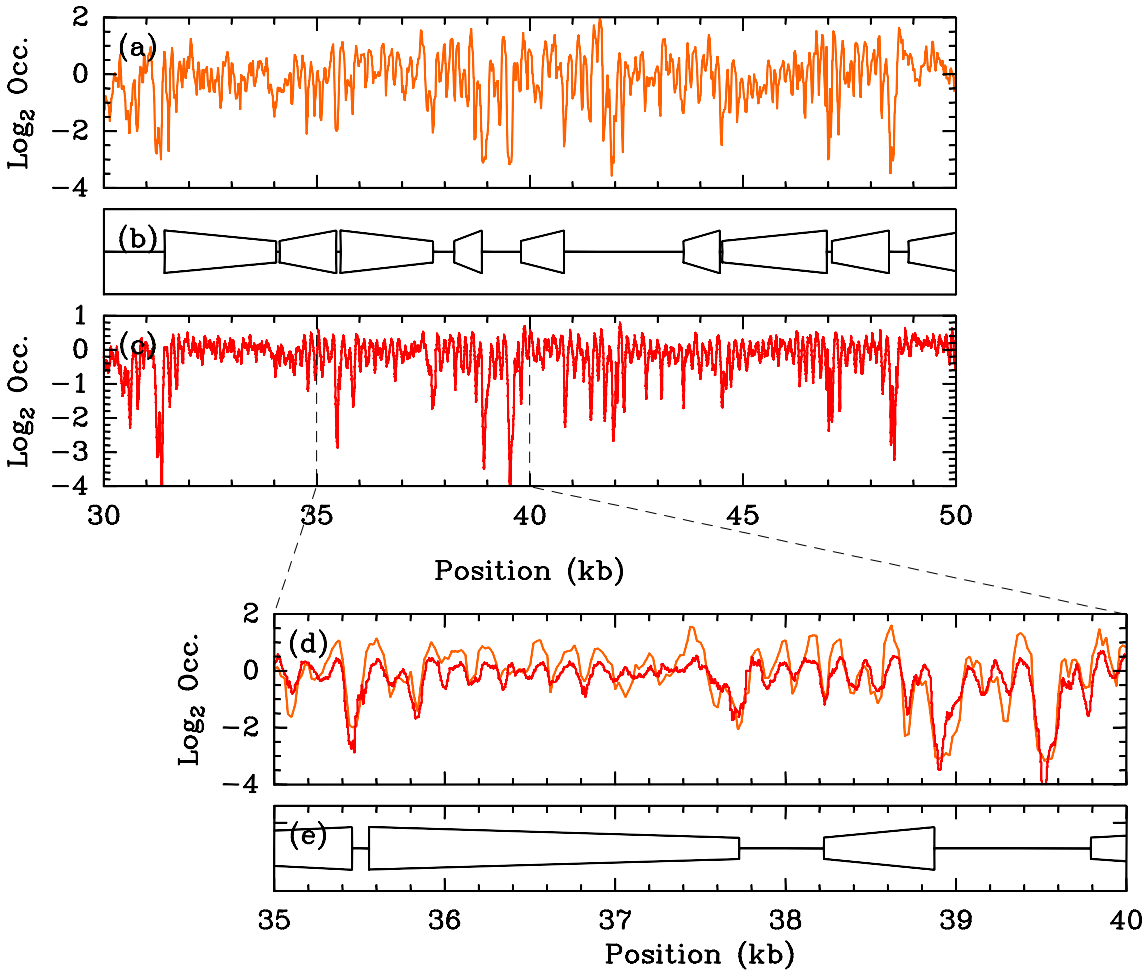


FIGURE 2.6 : *Comparaison entre cartes nucléosomales expérimentales obtenues par MNase-chip chez* S. Cerevisiae *par Yuan* et al. *(Yuan* et al.*, 2005) (a) et Lee* et al. *(Lee* et al.*, 2007*a*) (c) (log_2 du signal d'hybridation le long de $10kpb$ du chromosome 3). (b) Représentation schématique des gènes le long de cette même région), le TSS étant indiqué par le côté le plus large. (d) Superposition des deux jeux de données sur 5000 pb (orange, Yuan* et al.*, rouge Lee* et al.*). (e) Même représentation qu'en (b)*

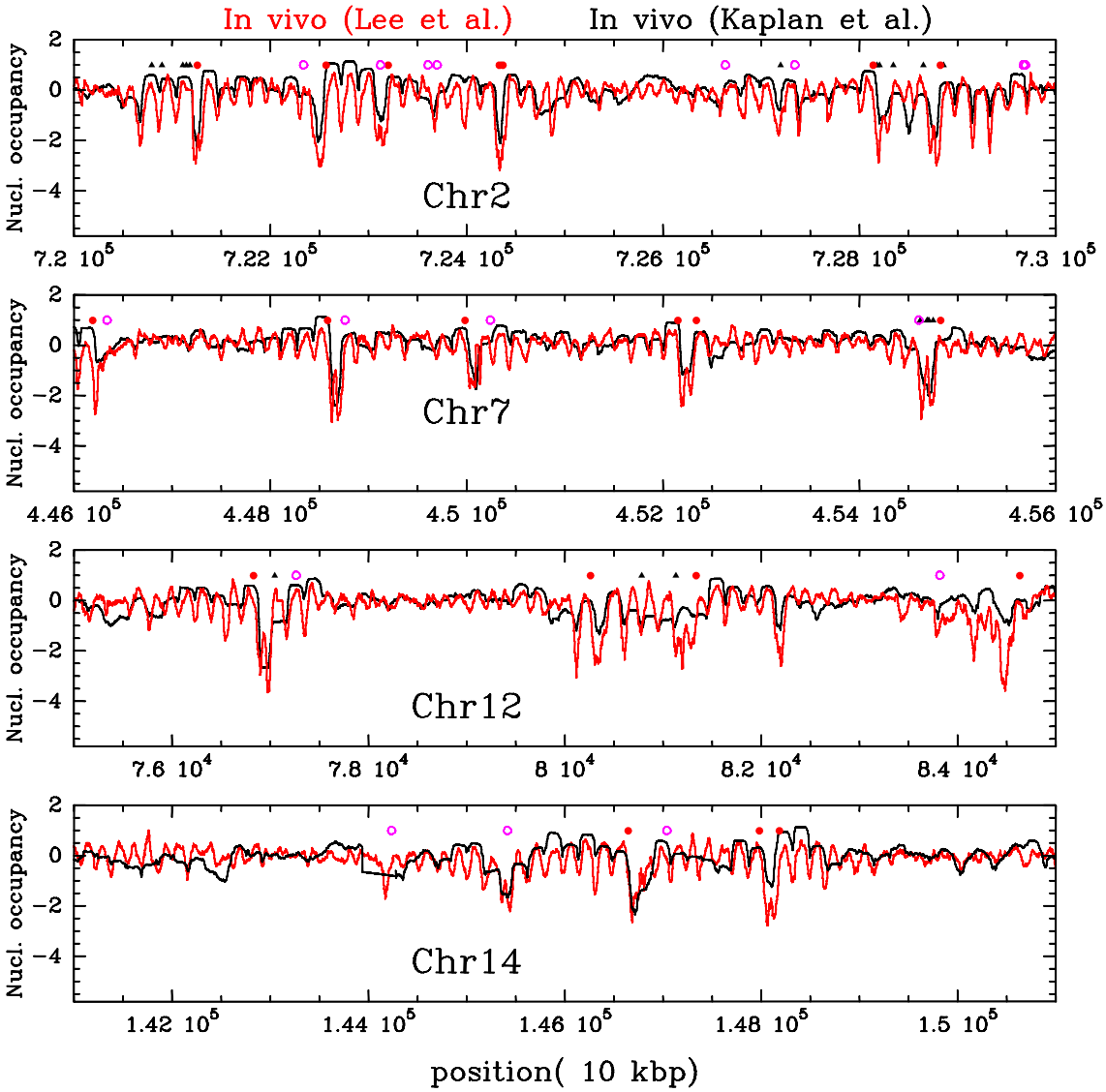


FIGURE 2.7 : *Idem qu'en Figure 2.5 avec superposées aux données de Lee* et al.*, les données experimentales Mnase-Seq de Kaplan* et al. *(Kaplan* et al.*, 2009a) (gris) (log_2 du signal d'occupation construits comme indiqué en Fig. 2.4).*

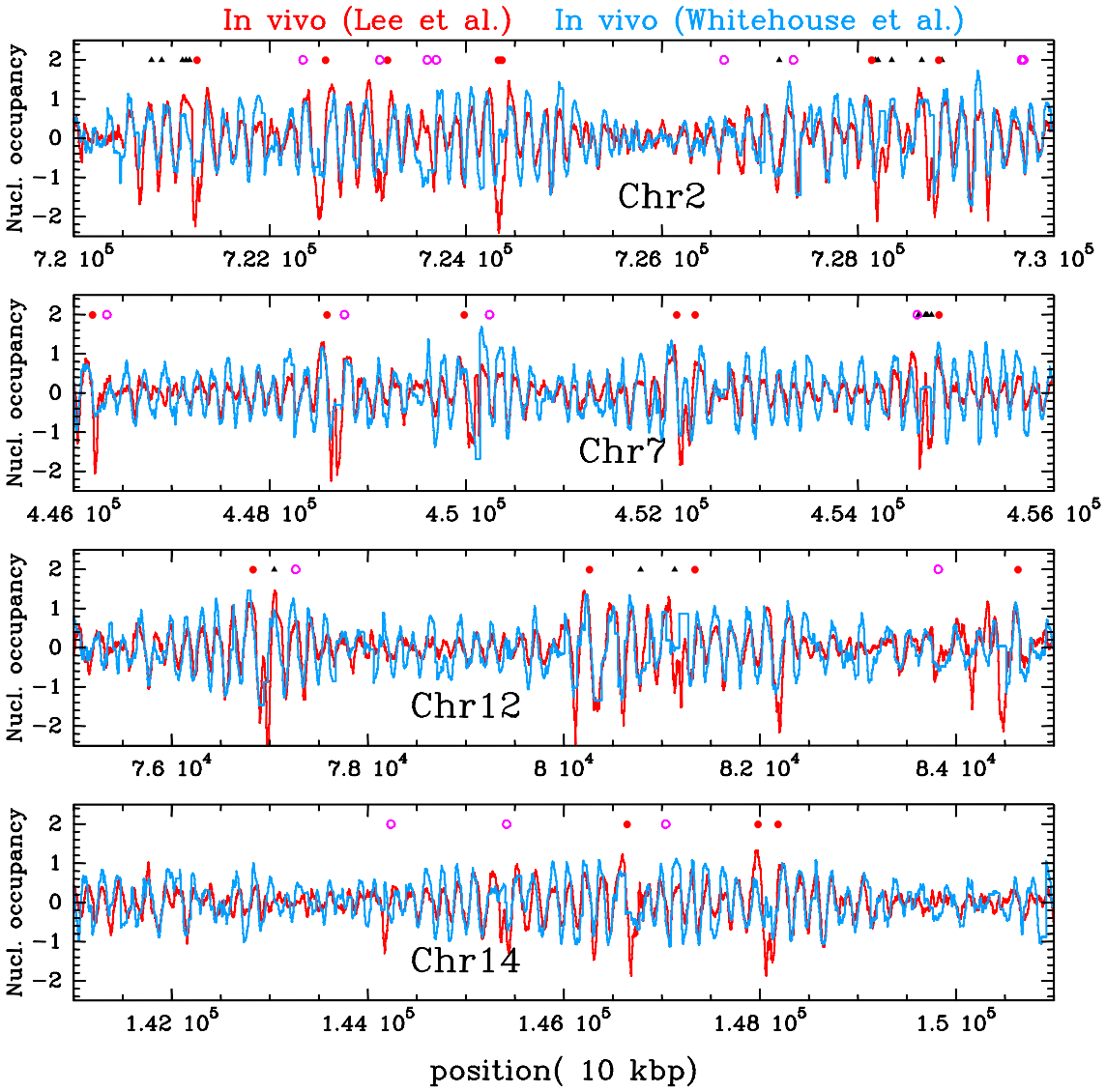


FIGURE 2.8 : *Données de MNase-Chip sur la levure obtenues par Whitehouse* et al. *(Whitehouse* et al.*, 2007) (bleu clair) ; le profil correspond aux données normalisées selon une procédure assez originale : ayant constaté que lors de l'hybridation des brins d'ADN nucléosomaux entiers (∼ 150pb), les extrémités étaient favorisées par rapport aux centres (gène stérique ?) Whitehouse* et col. *ont considéré que prendre le log-ratio des données d'hybridation issues des brins d'ADN nucléosomaux réduits (digestion par DNase résultant en une population de taille moyenne 50 pb) sur celles issues des brins entiers leur donnait une très bonne caractérisation du positionnement des nucléosomes ; le résultat de cette normalisation est notamment de filtrer les variations d'amplitude "basses fréquence" ; à titre de comparaison, la composante filtrée du profil de Lee (sur une fenêtre de 330 pb) a été superposée (rouge).*

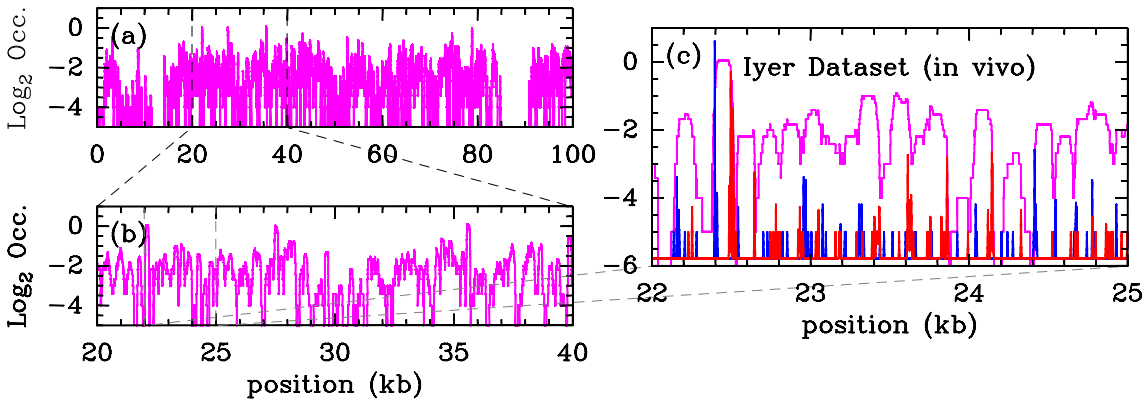


FIGURE 2.9 : *Données de MNase-Seq chez* S. Cerevisiae *obtenues par Shivaswamy et al. (Shivaswamy* et al.*, 2008). (a,b,c) Profils d'occupation obtenu à partir des reads des extrémités 5' (bleu) ou 3' (rouge) des ADN nucléosomaux représentés en (c). (b) Focus sur une région de* 20 *kpb (c) Focus sur une région de* 2 *kpb.*

le rapport d'enrichissment d'un site par rapport à un site voisin (le contratse) n'est jamais plus grand que $15 - 20$; donc assez faible si par exemple on compare avec l'enrichissement de 1000 mesuré *in vitro* sur la séquence artificielle 601 par rapport à une séquence aléatoire mais probablement suffisamment important pour cibler un facteur (activateur ou represseur) à un site nucléique.

2.1.3 Un positionnement hétérogène

À l'aide des distributions expérimentales, on caractérise l'hétérogénéité du positionnement sur les différentes données expérimentales.

The experimental density distributions yield information on the heterogeneity of positioning.

D'un point de vue plus général, les distributions de valeurs éclairent les différences entre les différentes données. Représentés en $log - log$, les histogrammes des valeurs d'occupation n'ont pas la même forme (figure 2.13) pour chacun des jeux de données. Il est important de noter que nous représentons ici les histogrammes des données brutes. Aucun traitement autre que l'application du logarithme sur les données de séquençage n'est effectué. Alors que les données de Lee et de Lanterman sont relativement peu dispersées autour de leur valeur moyenne (figure 2.13 (a) et (c)), les données de Kaplan ainsi que celle de Valouev ont une distribution beaucoup plus large, on peut imaginer que la méthode utilisée pour obtenir les données n'est pas sans incidence.

Sous échantillonnage

Lorsque les données sont sous échantillonnées, le problème apparaît clairement dans la distribution qui possède une queue anormalement haute (Iyer et Schones sur la figure 2.13 (d) et (d)). Lorsque l'essentiel du signal est à 0, comme sur les données de Schones sur l'homme, il devient très difficile d'interpréter les variations de positionnement qui sont observées.

Asymétrie

La distribution des données de Valouev est très différente (figure 2.13 (d)). Elle est enrichie en valeurs extrêmes, c'est-à-dire qu'il y a beaucoup d'évènements de très haute occupation, ce qui semble peu compatible avec la forme typique de données de densité convoluée. Cette surreprésentation des hautes valeurs est peut-être due aux séquences répétées, qui sont difficiles à positionner sur le génome. Un tag de positionnement attribué à une séquence qui est répétée ailleurs pourrait augmenter artificiellement le nombre de tag d'une position donnée. On trouve également beaucoup d'évènements de faible occupation. On retrouve aussi une queue de faible occupation dans les données de Lee, avec une surreprésentation des faibles occupations par rapport à une gaussienne par exemple. Cette queue peut avoir plusieurs origines : un sous échantillonnage (mais les données d'hybridation n'ont pas ce problème normalement), un bruit de fond dû à la fluorescence, ou bien plus probablement, la présence d'éléments déstabilisants le long du génome. Ils peuvent être directement liés à la séquence (barrières

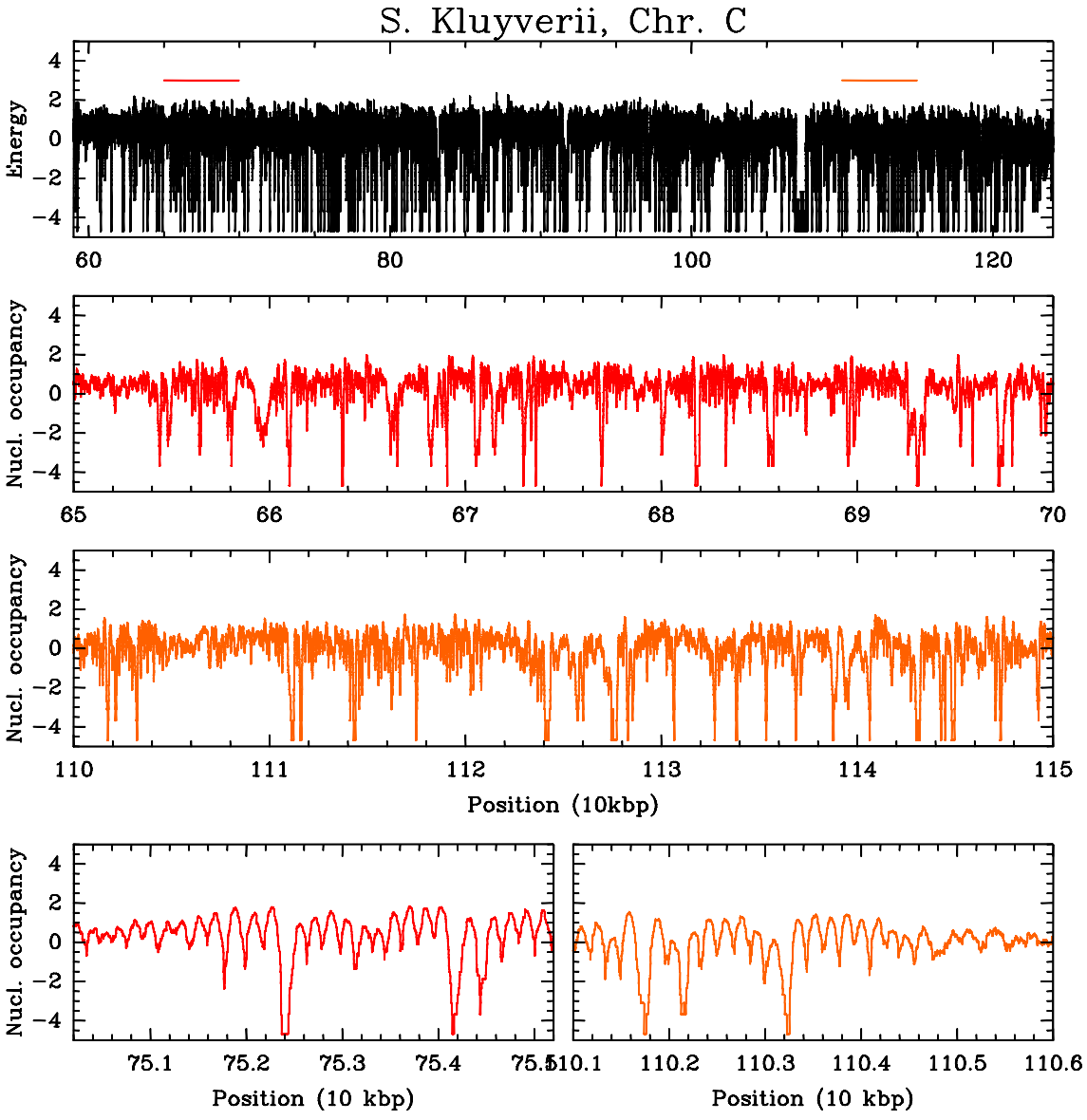


FIGURE 2.10 : *Données de MNase-Seq chez* S. kluyverii *obtenues par Tsankov* et al. *(Tsankov* et al.*, 2010) le long du chromosome C. Ce chromosome a la particularité de présenter une partie gauche de* $\sim 1Mpb$ *à* 53% *en G+C, donc riche par rapport à la moyenne* 40%*. (a) Tout le génome. (b,d) Profil d'une région de* $50kpb$ *"riche" en G+C indiquée en (a) par la bande rouge ; (d) extrait correspondant de 5 kpb. (c,e) Profil dans la région à G+C moyen indiquée en (a) par la bande bleue ; (e) extrait correspondant de 5 kpb*

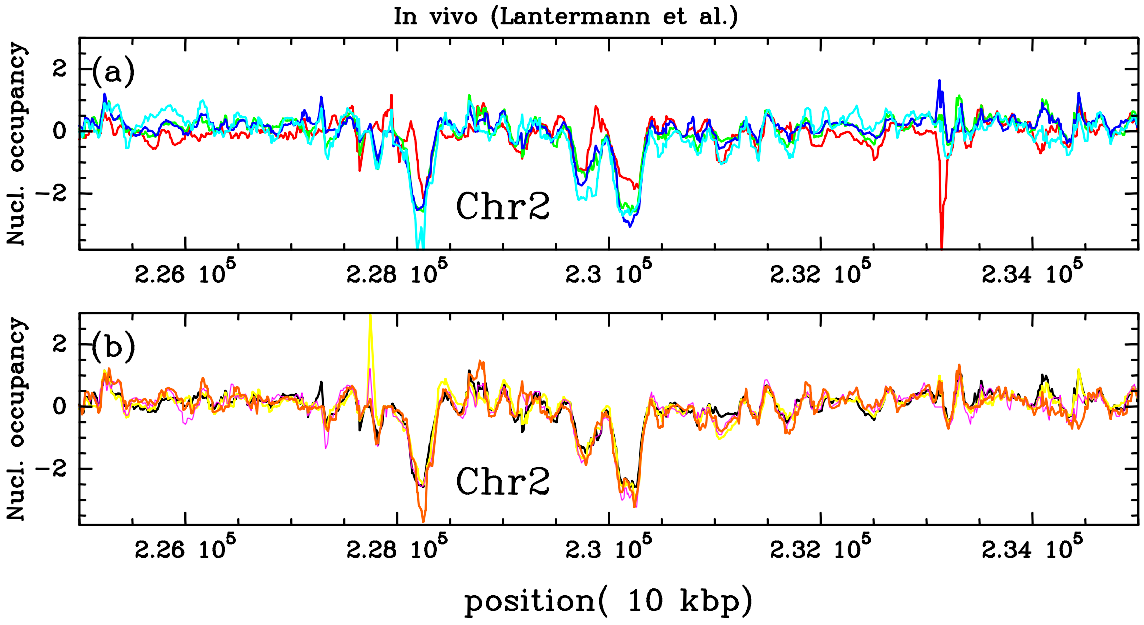


FIGURE 2.11 : *Données de MNase-chip sur la levure* S. Pombe *obtenues par Lanterman* et al. *(a) Souche sauvage (les quatres profils correspondent à quatre expériences d'extraction et d'hybridisation différentes, "réplicas") On voit ainsi que le réplica 1 est nettement différent des trois autres ... (b) Souche mutante pour le facteur de remodelage MIT1 (les 3 profils colorés correspondent à trois réplicas, le profil noir est un de ceux de la souche sauvage (cf (a)))*

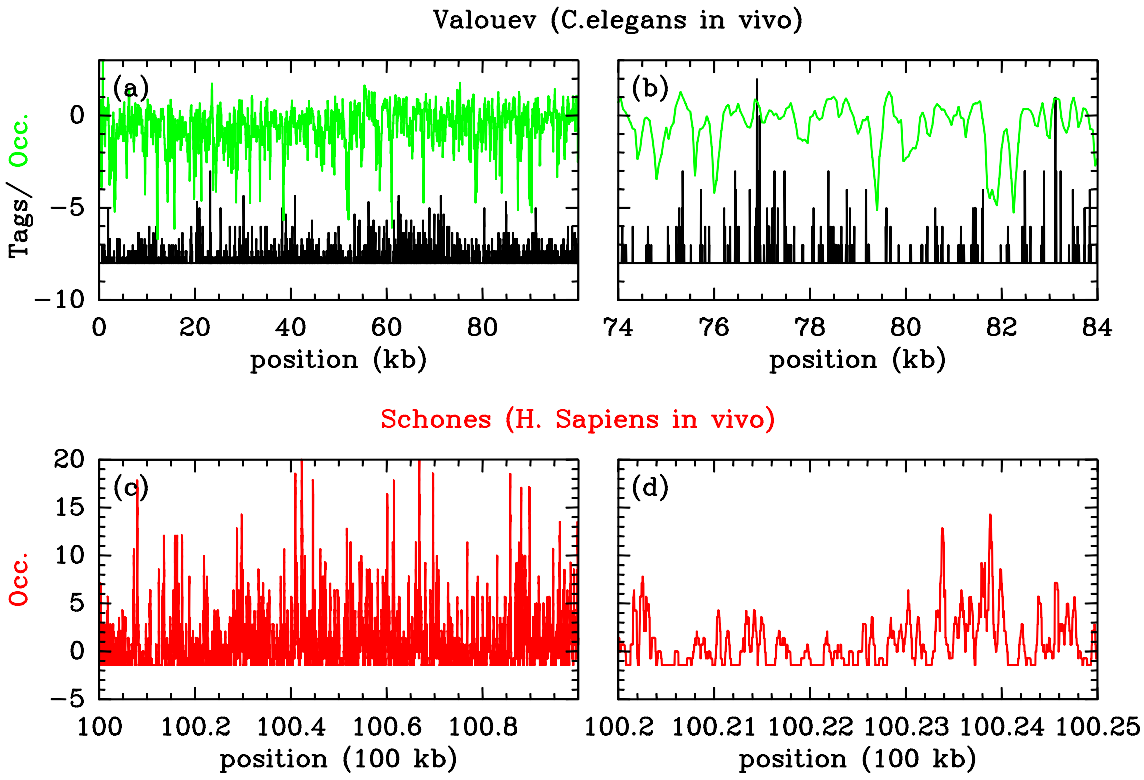


FIGURE 2.12 : *Cartes génomiques des nucléosomes obtenue par MNase-Seq : (a,b) chez C. Elegans, distribution des dyades (noire) 2.4 et profil d'occupation correspondant (vert) (Valouev* et al.*, 2008) ; (c,d) chez l'homme dans des cellules $CD4^+$ "au repos" (phase G_0), profil d'occupation (rouge) (Schones* et al.*, 2008).*

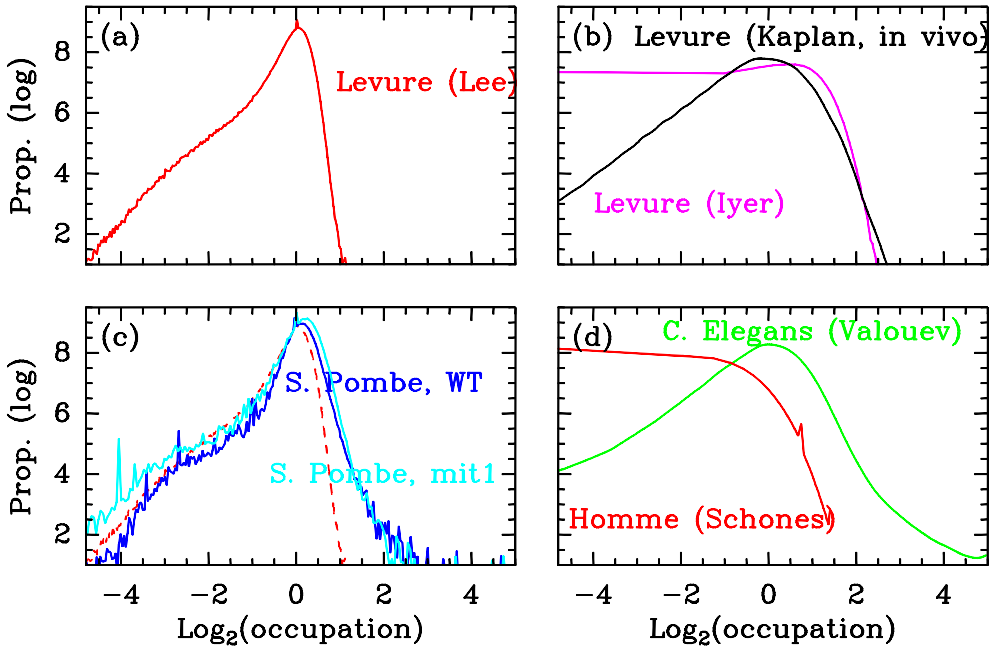


FIGURE 2.13 : *Distributions statistiques des valeurs expérimentales pour différents jeux de données d'occupation. Représentation en $log - log_2$ (a) Données de Lee (Lee* et al.*, 2007*a*) sur S. cerevisiae, données MNase-Chip. (b) Données in vivo de Kaplan (noir) (Kaplan* et al.*, 2009*a*) et de Iyer (rose) (Shivaswamy* et al.*, 2008) sur la levure (MNase-seq). (c) Données de MNase-chip chez S. Pombe, souche sauvage (bleu foncé), souche mutante* mit1 *(cyan) ; données de Lee (cf (a)) superposées (tirets rouges). (d) Données (normalisées) de Valouev (Valouev* et al.*, 2008) sur C. elegans. (MNase-seq), et données de Schones (Schones* et al.*, 2008) sur l'homme en rouge.*

énergétiques issues de rigidité de la séquence), Ils peuvent aussi venir de phénomènes extérieurs à la séquence, lorsqu'un objet biologique empêche le positionnement des nucléosomes.

2.1.4 L'organisation linéaire du chapelet nucléosomal

Pour caractériser globalement l'organisation du chapelet nucléosomal, on peut calculer la fonction d'autocorrelation des profils, qui nous renseigne à la fois sur le caractère désordonné mais aussi régulier de la distribution des nucléosomes. Les fonctions d'autocorrélation de chacun des jeux de données sont présentés sur les figures 2.14 et 2.15. Comme indiqué précédemment, les données de Lee présentent des régions oscillantes ; cela se traduit au niveau de l'autocorrélation par une modulation périodique de 168 pb (Fig. 2.14 (a)). L'autocorrélation présente également une décroissance globale qui s'étale sur une échelle de l'ordre de quelques centaines de paires de bases avant de tendre lentement vers 0. Comme on le verra plus tard, cette décroissance est en loi de puissance et reflète la présence de corrélations à longue portée dans les séquences génomiques (Vaillant et al., 2007). L'autocorrélation sur les données de MNase-seq (données de distributions de tags, non convoluées) de Iyer présente également une périodicité de 168 *pb* (figure 2.14 (a)). Les données d'occupation correspondantes ainsi que celles de Kaplan ne présentent pas d'oscillations si prononcées (figure 2.14 (b)) : l'effet de la convolution est de réduire l'amplitude de cette modulation périodique mais aussi d'en modifier légèrement la période. Sur le nématode *C. elegans*, comme l'indique la fonction d'autocorrélation 2.15(a), les données d'occupation ne révèlent tout simplement pas d'oscillations. Cependant, Valouev fournit également des données de tags de séquençage brutes, non normalisés par les effets de séquences (en noir sur la figure 2.12), avec lesquels on peut retrouver un profil homogène à une densité. Certes, les données ne sont plus normalisées par les variations d'affinité de la MNase avec la séquence mais elles ne sont plus des données d'occupation, mais des données de densité brute qui n'ont pas été convoluées par la taille du nucléosome. Elles sont donc plus résolues, et elle présentent effectivement une oscillation caractéristique située à $\approx$ 172 pb (figure 2.15 (a) en noir). Un aspect intéressant est le fait que l'on trouve des oscillations rapides dans l'autocorrélation du signal de densité de tags qui correspond à la quantification à 10 pb du positionnement nucléosomal. Ces oscillations sont d'ailleurs discutées en détail dans l'article original

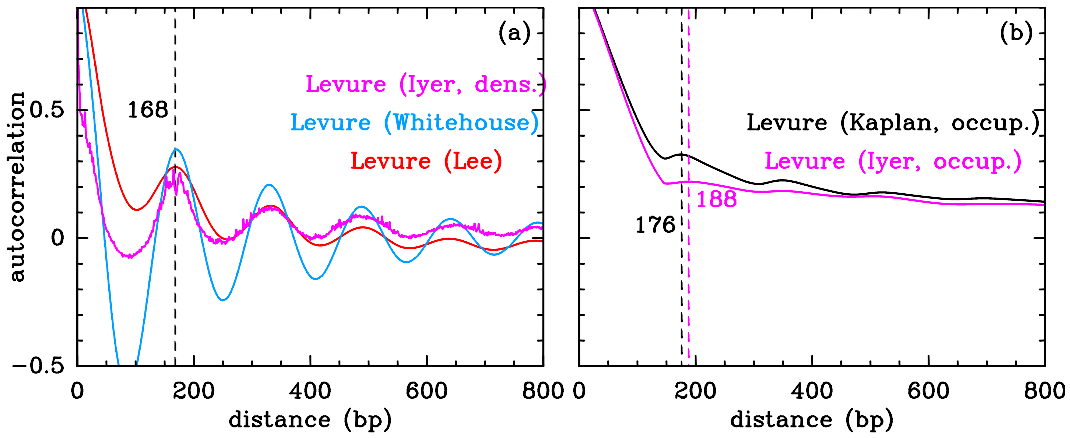


FIGURE 2.14 : *Profils d'autocorrélation des données d'occupation nucléosomale sur la levure. (a) Données de Lee (en rouge) et Iyer in vivo (densité en noir, occupation en rose) Pour plus de visibilité, l'autocorrélation des données de Iyer (densité en noir) a été multipliée par 10. (b) Kaplan in vitro, in vivo, et Zhang in vitro. Le premier maximum de l'autocorrélation est utilisé pour définir le pseudo-NRL.*

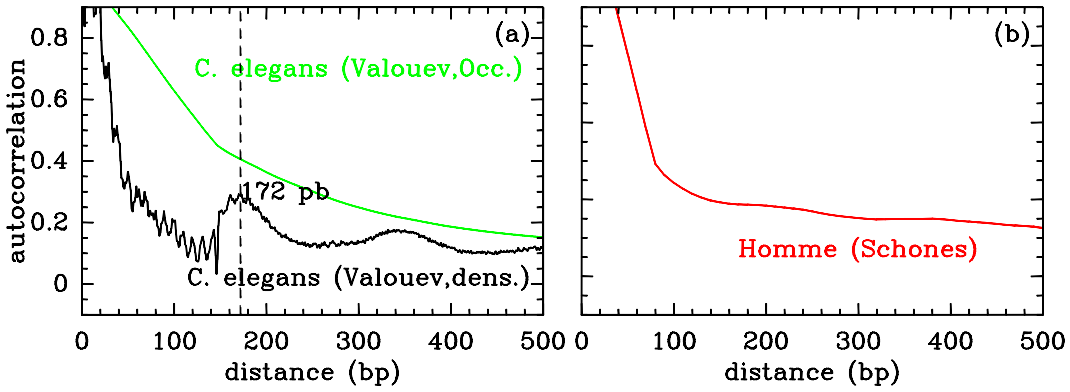


FIGURE 2.15 : *Profils d'autocorrélation des données d'occupation nucléosomale sur différents organismes. C. elegans, densité en noir (a), Occupation normalisée en vert. (b) Homme en rouge (données d'occupation de Schones)*

de Valouev (Valouev et al., 2008). Au final, ces données suggèrent que la compaction dans la levure est légèrement plus élevée que sur *C. elegans*, ce qui pourrait d'ailleurs être dû à l'absence de l'histone *linker* H1 dans la levure. Il est possible de relier l'autocorrélation (via le pseudo-NRL) à la distance entre nucléosomes (cf Chapitre 5 Section 2.3). On peut se demander dans quelle mesure cette distance varie le long des génomes. Si les nucléosomes sont plus confinés, on s'attend à ce qu'ils se rapprochent en moyenne par exemple. Nous nous intéressons donc aux variations du pseudo-NRL, en se concentrant uniquement sur les jeux de données avec lesquels il est possible de définir clairement un pseudo-NRL, c'est-à-dire les données de Lee, et les données brutes de tags de Valouev. Ces jeux de données présentent des maximum d'autocorrélation bien marqués. La figure 2.16 montre l'évolution de ce pseudo-NRL calculé dans des fenêtre de 10 kb le long du génome entier de la levure (tous les chromosomes sont mis bout à bout). Ce pseudo-NRL s'étale entre 140 et 210 paires de base avec un pic bien marqué autour de 168 pb (encart (b)). Si l'on compare ce résultat à un autre organisme (*C. elegans*, figure 2.17) on s'aperçoit que le pseudo-NRL y est en moyenne plus grand. On remarque sur cette même figure que la distribution de NRL est très similaire entre les deux organismes, et que la distribution théorique de pseudo-NRL est d'ailleurs globalement identique (figure 2.17). Certes la distribution de la levure est enrichie en valeurs faibles de NRL et la distribution de *C. elegans* est enrichie en valeurs grandes de NRL, le pic des trois distribution est situé au même endroit, c'est à dire entre 168 et 175 pb. La variabilité est toutefois tout à fait comparable, ce qui suggère que le mécanisme d'organisation des nucléosomes est similaire dans les deux organismes, mais que l'interaction effective entre nucléosomes est différente. Le fait qu'il n'y ait pas d'histones *linker* dans la levure, mais dans *C. elegans* si, n'est peut-être pas pour rien dans cette ob-

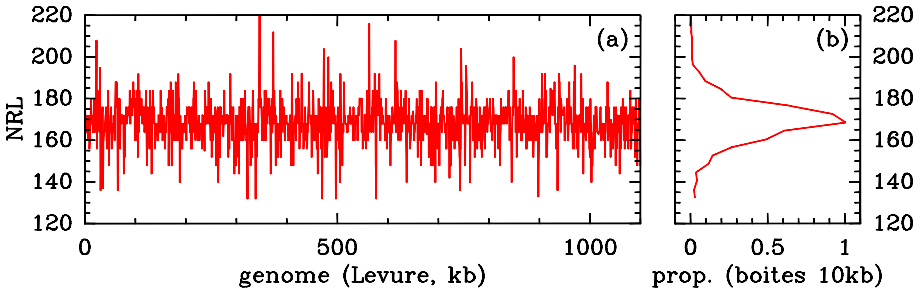


FIGURE 2.16 : *(a) Évolution du pseudo-NRL, calculé sur les données d'occupation de Lee, le long de l'intégralité du génome de la levure (chromosomes mis bout à bout), dans des boîtes de 10kb. (b) distribution des valeurs de pseudo-NRL.*

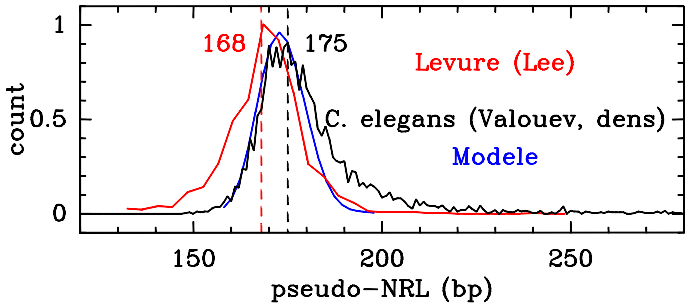


FIGURE 2.17 : *Comparaison des distributions de pseudo-NRL dans deux organismes différents. En rouge, distribution du pseudo-NRL calculé dans des fenêtres de 20 kb sur les données de Lee (levure), En noir, distribution de NRL calculé dans des fenêtres de 20kb calculée sur les données de densité (non normalisées) de Valouev sur C. elegans. En bleu, la distribution de pseudo-NRL du modèle Vaillant, calculée sur la levure (cette distribution est identique sur C. elegans dans le modèle).*

servation. Comme l'a remarqué Woodcock (Woodcock et al., 2006), il existe en effet une relation linéaire entre la valeur du NRL mesuré dans un organisme ou un type cellulaire et la stoechiométrie histones de liaison H1 vs. nucléosomes (Fig. 2.18). Notons enfin que chez S. Pombe, comme l'indique Lantermann et al. (Lantermann et al., 2010) la période nucléosomale est de $154\ pb$; l'évaluation de cette période par la fonction d'autocorrélation ne permet pas de la mesurer (on mesure plutôt la double période à $\sim\ 310\ pb$, non montré ici). En fait comme, on le verra au niveau des gènes 2.21 l'ordonnancemnt périodique chez *S. Pombe* est relativement faible, mais lorsqu'il est présent il se fait de façon plus nettement plus compacte que dans la plupart des organismes.

2.2 CHAPELET ET MÉTABOLISMES NUCLÉAIRES

2.2.1 Transcription

La régulation de la transcription chez les eucaryotes opère à tous les niveaux du cycle transcriptionnel, (i) activation/initiation durant laquelle le complexe de pré-initiation contenant la polymerase (I, II ou III) se forme au promoteur, (ii) l'élongation, correspondnat à l'échappement du promoteur puis à la progression de la polymerase et (iii) la terminaison.

Chez la levure *S. Cerevisiae*, le motif nucléosomal canonique d'un gène est constitué d'une zone déplétée (NFR) au niveau du promoteur, d'une taille située entre 100 et 200 paires de bases, et positionnée en amont du TSS (Fig. 2.19 et 2.20(a)). Le TSS est en bordure 5' du nucléosome +1 qui est en effet, comme l'atteste l'histogramme en Fig. 2.20(a) très phasé vis-à-vis du TSS (distance d'$\sim\ 73bp$). Une deuxième NFR est observable dans une moindre mesure au niveau du TTS, sur l'extrémité 3' du gène. (Yuan et al., 2005; Albert et al., 2007; Lee et al., 2007a; Mavrich et al., 2008a; Shivaswamy et al., 2008) (Fig. 2.19 et 2.20(b)) ; par contre dans ce cas, pas de positionnement périodique et un phasage faible. Cette archi-

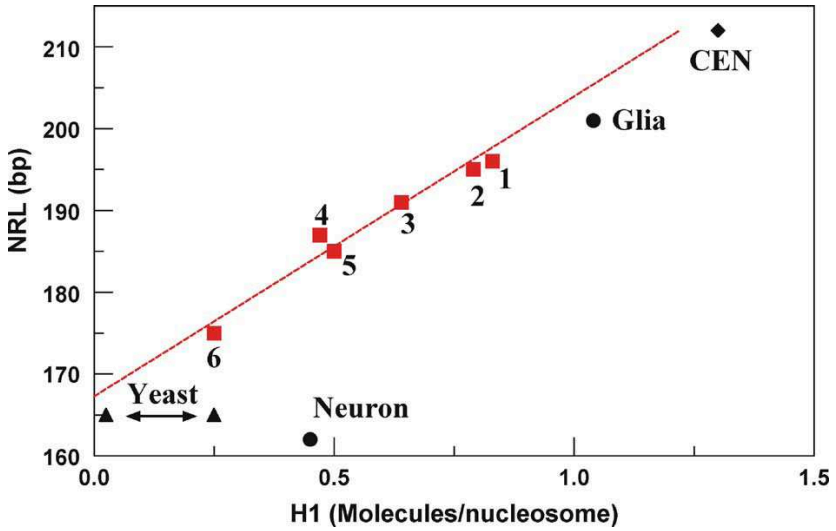


FIGURE 2.18 : *Evolution du NRL (période nucléosomale) en fonction du ratio de H1/nucléosme. Les carrés correspondent à différents types cellulaires de souris, sauvages et déplétée en H1 (Fan et al. 2003) et à des cellules ES (Fan et al. 2005) : 1. thymus, souche sauvage ; 2. foie, souche sauvage ; 3. foie, déplété en H1 ; 4. thymus, déplété en H1 ; 5. ES, sauvage ; 6. ES, déplétée en H1. Les tirest correspondnet à la régression linéaire des données de souris. Cercles : donnée de cellulles neuronales et gliales (Pearson et al. (1984)). Losanges : Chromatine d' erythrocyte de poulet (CEN). Triangles : données de* S. cerevisiae *(Freidkin & Katcoff (2001), Downs* et al. *(2003))*

tecture se retrouve dans toutes les levures *Hemyascomycta* étudiées par Tsankov *et al* (Tsankov et al., 2010). Chez S. Pombe, on retrouve au TSS la même organisation que chez ces levures avec un positionnement périodique au niveau de la partie 5′ de l'ORF moins marqué (Fig. 1 de (Lantermann et al., 2010) et Fig 2.21). Une organisation similaire au TSS et TTS est observée chez la drosophile (Mavrich et al., 2008b). Chez C. elegans on observe également, "en moyenne" une déplétion au niveau du TSS (Valouev et al., 2008) mais peu de positionnement périodique de part et d'autre (seul le nucléosome $+1$ est très bien positionné...) ; de même chez l'homme, pour les gènes riches en îlots CpG, on observe une déplétion au niveau du TSS avec un léger ordonnancement périodique dans le partie 5′ de l'ORF Schones et al. (2008) (Fig. 2.28).

Architecture du promoteur

Rappelons ici brièvement les mécanismes moléculaires associés à l'activation de la transcription au niveau des régions promotrices. Une fois des activateurs fixés à leurs séquences spécifiques ("UAS", Fig. 2.23) au niveau du promoteur (et enhancer) ceux-ci induisent une cascade de recrutements d'autres facteurs moléculaires co-activateurs conduisant à l'assemblage du complexe de pré-initiation comportant notamment la polymérase au niveau du TSS (Fig. 2.23). La compaction de l'ADN au sein des nucléosomes est un obstacle à la fixation des facteurs de transcription et en général à toute autre transaction nécessitant une fixation avec la double hélice ; parmi donc les co-activateurs on trouve des facteurs de remodellages, des enzymes de modifications des histones et des chaperones, qui souvent agissent de concert pour favoriser l'accessibilité non seulement des activateurs mais aussi des facteurs de transcription dits "généraux" associés à la formation du complexe de pré-initiation. Dans la levure, on peut classer les gènes selon leur organisation nucléosomale au niveau du promoteur. Lee *et al.* (Lee et al., 2007a) ont ainsi pu extraire quatre classes de gènes 2.24 : (i) Deux classes ont un profil très proche du profil "canonique', à savoir une NFR prononcée avec des nucléosomes bien ordonnés de part et d'autre (60% des gènes) (ii) Une classe avec une NFR moins prononcée et plus large au niveau du promoteur (15%) et (iii) Une classe sans NFR (25%). De manière générale, le niveau d'expression des gènes est anti-corrélé avec le niveau d'occupation global du promoteur (Fig. 2.25). Chez la levure (Lee et al., 2007a) on voit par exemple à partir des distributions statistiques du niveau d'expression des gènes des quatre

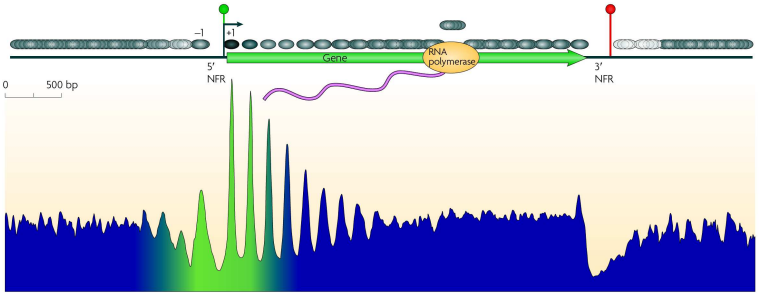


FIGURE 2.19 : *Positionnement nucléosomal canonique à proximité des gènes de la levure. La distribution consensus des nucléosomes (ovoïdes gris) aux alentours des gènes est représentée alignée sur le début et la fin de chaque gène. Les deux graphiques sont fusionnés sur le milieu du gène. La flèche située à proximité de la zone vide (NFR) en 5′ représente le TSS. La transition verte-bleue au sein du graphique représente la transition observée dans la composition et le phasage des nucléosomes : la zone verte correspond à un fort taux de remplacement de H2A par H2A.Z, ou bien à une forte acétylation ou encore à une forte méthylation, la zone bleue correspond à une zone où ces modifications sont faibles. Le cercle rouge indique la fin de la zone de terminaison de la transcription à proximité du NFR situé en 3′. Figure originale publiée par Mavrich et al (Mavrich* et al.*, 2008a)*

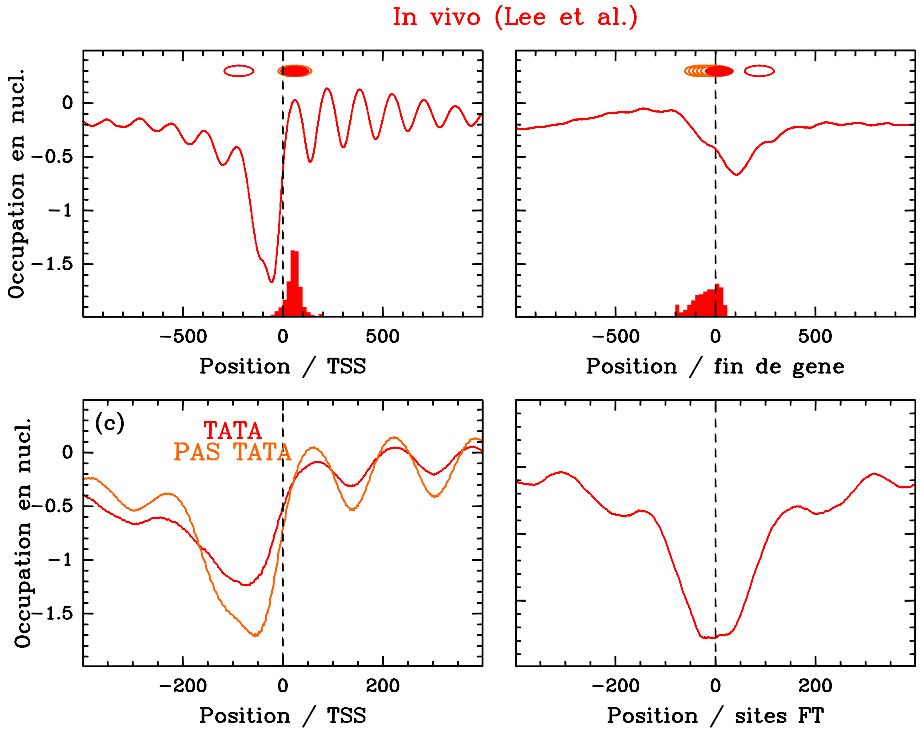


FIGURE 2.20 : *Profils moyens des données de Lee* et al. *(Lee* et al.*, 2007a) autour : du TSS (a,c) et du TTS (b) de 4554 genes (Vaillant* et al.*, 2010) et autour des sites de fixations de facteur de transcription (d). En (c), comparaison entre la moyenne pour les gènes possédant une boite TATA (rouge) et ceux sans (orange).*

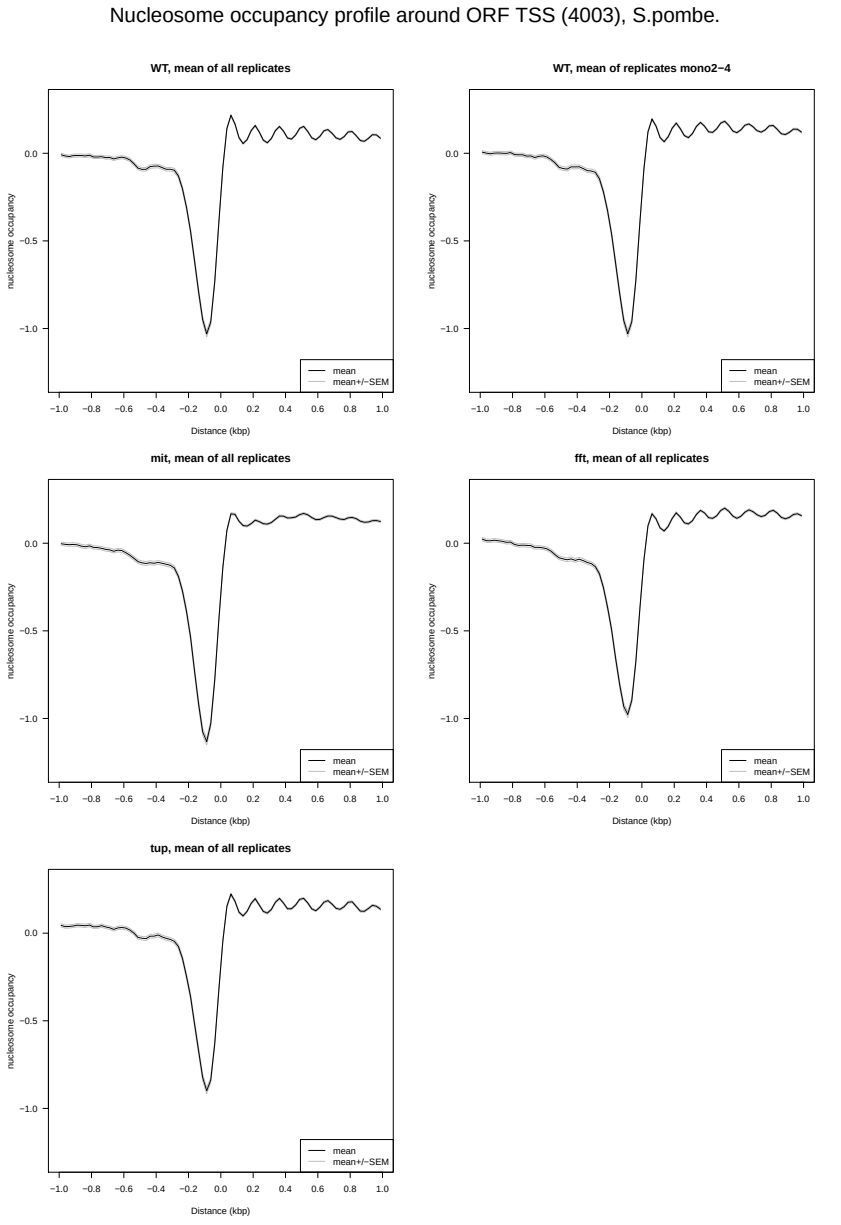


FIGURE 2.21 : *Profils moyens des données de Lantermane*t al. *(Lantermann* et al.*, 2010) autour du TSS pour (haut, gauche) souche sauvage en considérant toutes les expériences indépendantes (les "réplicas") (haut, droite) en retirant les données du premier des réplicas (Fig. 2.11) (milieu, gauche) souche mutante* mit1 *pour le facteur de remodelage Mit1, (milieu droite) pour le facteur de remodelage Fft3, et (bas) double mutant pour le co-répresseur Tup11-Tupp12-Ssn6 (Lantermann* et al.*, 2010).*

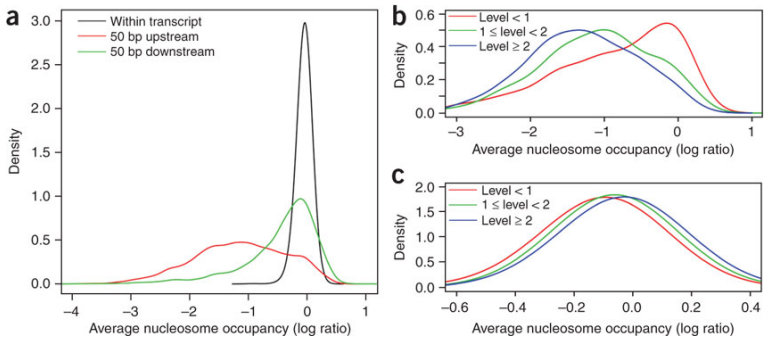


FIGURE 2.22 : *(a) Distribution statistique de l'occupation en nucléosomes mesurée dans différentes régions voisines du TSS pour* 5015 *gènes : Distribution de l'occupation moyenne pour la région de 50-bp en aval (rouge), 50-bp en amont (vert), et au sein du gène (noir). (b) Relation entre l'occupation moyenne dans la région promotrice de 50 pb juste en aval du TSS et le niveau d'expression. Distribution statistique de cette occupation moyenne pour les gènes de niveau d'expression* $< 1 (n = 759)$ *(rouge), entre* 1 *et* $2 (n = 1.859)$ *(vert), et pour les plus exprimés* $> 2 (n = 2.397)$ *(bleu). (c) Même chose, mais avec les distribution d'occupation moyenne au sein du gène.*

classes de Lee (Fig. 2.24(c)) qu'en moyenne les gènes de la classe 1, donc sans NFR sont plus faiblement exprimés que les gènes des classes 3 et 4 qui ont des NFR bien marquées mais aussi que les gènes de la classe 2 ; remarquons que malgrés une déplétion apparente faible les gènes de cette classe présentent des expressions du même ordre voire plus forte que celles des classes 3 et 4, avec en particulier une bimodalité vesr les fortes expressions. Des gènes peuvent avoir un promoteur qui en moyenne (car le profil de Lee résulte d'une moyenne sur des millions de cellules) apparait occupé mais avoir une activité transcriptionnelle forte. C'est par exemple le cas des gènes liés à la glycolise et la glucogénèse pour l'ensemble des levures *Hemiascomycota* comme l'indique l'étude de Tsankov *et al.* (Tsankov et al., 2010), qui par ailleurs confirme cette correlation négative entre occupation et expression dans l'ensemble de ces levures, principalement dans les gènes de croissance.

Chez la drosophile (Fig. 2.26 B) tout comme chez l'homme (Fig.2.28) le TSS est localisé au mileu de la zone de déplétion (NFR) contrairement à la levure S. Cerevisiae déplétion au niveau du promoteur est d'autant plus grande que le niveau d'expression est grand (Miele et al., 2008) (Fig. 2.26(b)). Des tendances similaires sont observés chez *S. Pombe* (Lantermann et al., 2010) ou encore l'homme (Schones et al., 2008).

Dans leur étude (Tirosh and Barkai, 2008a) chez la levure, Tirosh et Barkai ont classé les gènes selon leur occupation en nucléosome dans la zone promotrice proche ($[-100\ 0] pb$ du TSS) relativement à celle dans la zone plus éloignée ($[-450\ -150]\ pb$) ; ils ont ainsi pu révéler deux classes de gènes bien distinctes : (i) les gènes DPN (pour "Depleted Proximal Nucleosome") présentant une déplétion (NFR) bien localisée bordée par des nucléosomes périodiques (Fig. 2.27 et (ii) les gènes OPN (Occupied Proximal Nucleosome) dont le promoteur est occupé par un nucléosome, c'est-à-dire qui ne présente pas de NFR bien marquée (Fig. 2.27). L'étude des propriétés transcriptionnelles montre que ces deux types de gènes se distingue plus par leur "logique" de régulation que par leur activité transcriptionnelle : (i) les gènes DPN sont des gènes constitutivement exprimés, avec des sites de facteurs de transcriptions bien localisés au niveau de la NFR ; ils présentent une faible plasticité transcriptionnelle et une faible sensibilité à la régulation chromatinienne ; un bruit et une divergence transcriptionnelle ainsi qu'une dynamique d'échange d'histones H3 également légèrement moindre que la moyenne. Ils sont plutôt sans boite TATA. Par contre ils montrent un enrichissement marqué en variant H2A.Z au nucléosome +1. (ii) A l'inverse les gènes OPN sont des gènes très plastiques et fortement régulés au niveau chromatinien, bruités et divergents transcriptionnellement ; leurs sites de fixation des facteurs de transcriptions sont distribués sur l'ensemble du promoteur et ils sont enrichis en boite TATA ; enfin leur dynamique d'échange d'histone est forte et le nucléosome +1 est plutôt non enrichi en H2A.Z. Des comportements similaires sont observés chez l'homme avec par ailleurs, les gènes OPN plutôt pauvres en ilots CpG (Fig. 4 de (Tirosh and Barkai, 2008a)). On le voit en effet sur la figure 2.28, les gènes humains enrichis en ilots CpG révèlent plutôt une architecture "constitutive" avce une NFR bien marquée et un positionnement périodique (donc du type DPN), tandis que les gènes plutôt pauvres en îlots CpG ont effectivement

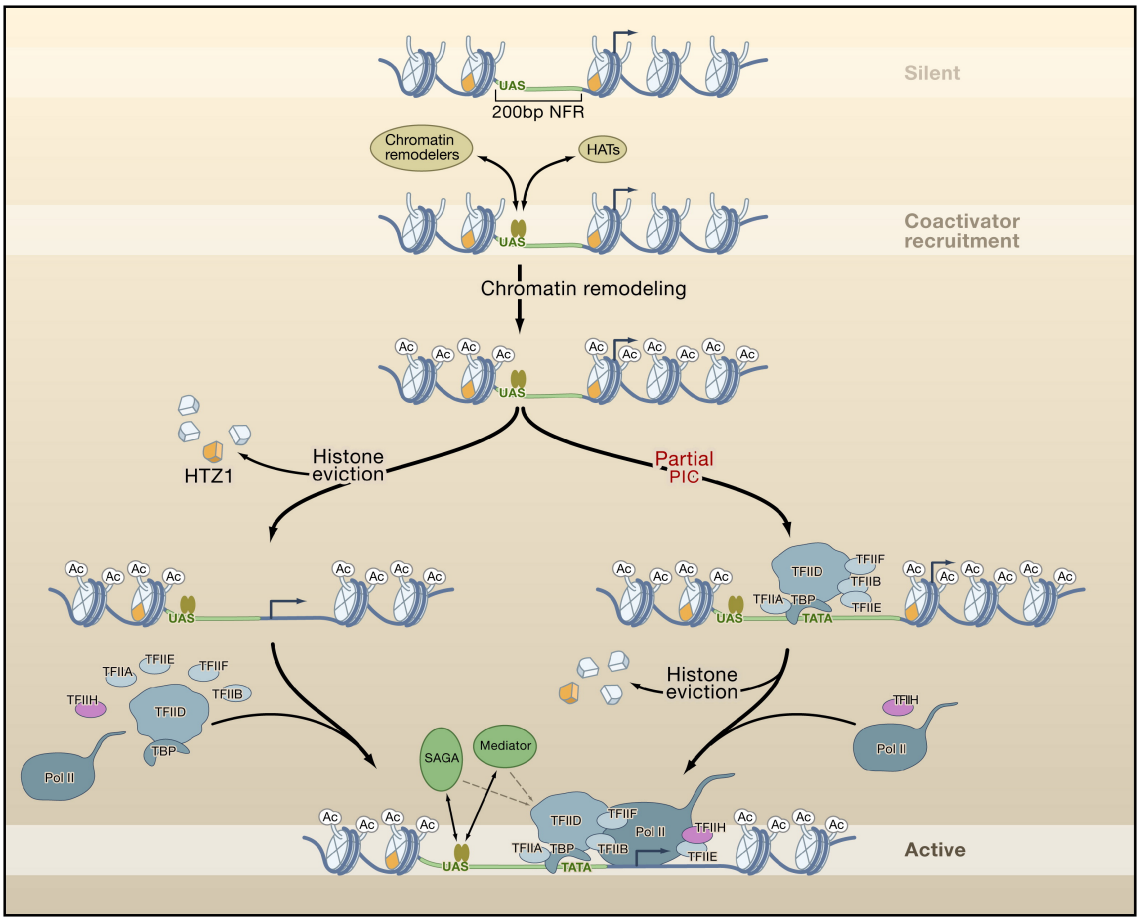
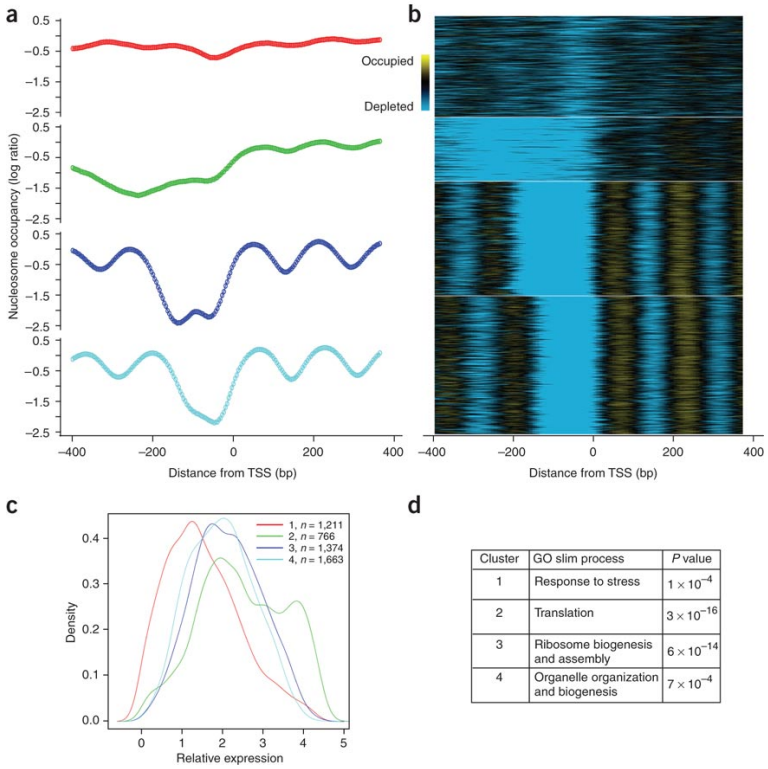


FIGURE 2.23 : *Modèles de régulation de la chromatine à l'initiation de la transcription. Au niveau des promoteurs non activés on retrouve généralment de part et d'autre d'une NFR de* $200\ pb$ *des nucléosomes enrichis en variants Htz1. Après fixation à leurs sites spécifiques (UAS, "Upstream Activator Sequence") les activateurs recrutent d'autres coactivateurs (tels que Swi/Snf ou SAGA). Ce recrutement va renforcer la fixation des activateurs, en particuliers ceux associés à des séquences protégées par des nucléosomes. Les histones proches du TSS sont acetylatés et deviennent plus mobiles. Dans un modèle (gauche), une combinaison d'acetylation et de remodelage de chromatine induit l'éviction des nucléosomes* -1 *et* $+1$*, exposant ainsi l'ensemble du promoteur au Facteurs de Transcription Généraux et de Pol II. SAGA et mediator facilite ensuite la formation du PIC par interactions directes. Dans un autre modèle (droite), qui représente l'état "remodelé", le PIC pourrait être partellement assemblé au promoteur sans éviction des nucléosomes avec histones variants Htz1. C'est la fixation de Pol II et des TFIIH qui induisent ensuite l'éviction de ces nucléosomes et la formation complète du PIC.*



Cluster	GO slim process	*P* value
1	Response to stress	1×10^{-4}
2	Translation	3×10^{-16}
3	Ribosome biogenesis and assembly	6×10^{-14}
4	Organelle organization and biogenesis	7×10^{-4}

FIGURE 2.24 : *(a) Occupation moyenne en nucléosome pour chaque classe. La classification a été obtenue par la méthode du k-mean. Les quatres classes contiennent* 1211, 766, 1374 *and* 1663 *gènes. (b) "clustergram" k-mean pour un jeu d' environ* 5000 *transcrits. Le bleu représente les régions déplétées en nucléosomes et les régions jaunes sont davantage occupées. (c) Distribution statistique des niveaux d'expression des transcrits dans chaque classe. (d) L'ontologie des gènes (GO) avec la "P-value" associée. Extrait de (*Lee et al.*, 2007*a*).*

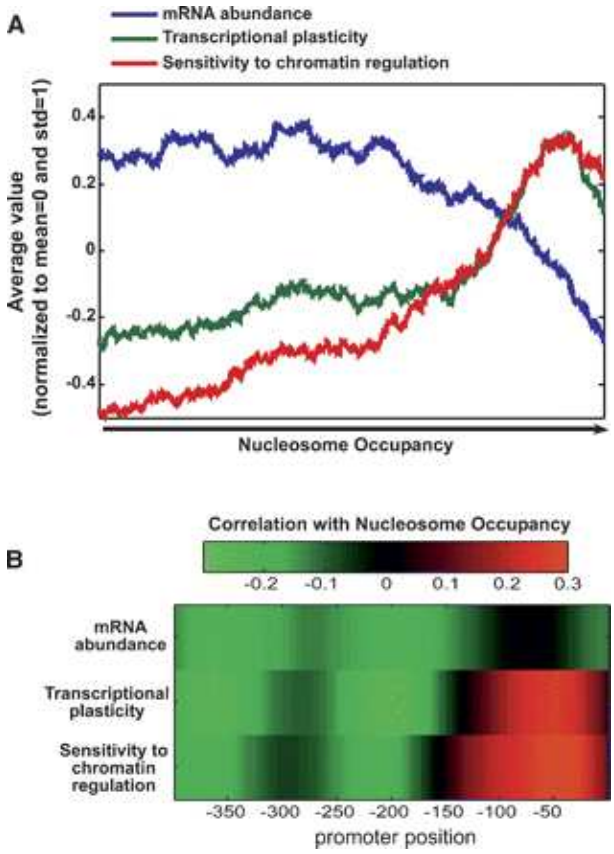


FIGURE 2.25 : *Lien entre l'expression et l'occupation nucléosomale au promoteur : (a) proportion d'ARN messager (en bleu) en fonction de l'occupation. Sensibilité de la chromatine à la régulation (en rouge), plasticité de la transcription (en vert). (b) Corrélation des trois grandeurs de la figure (a) avec l'occupation nucléosomale au niveau du promoteur. Figure adaptée de Tirosh et al. (Tirosh and Barkai, 2008a).*

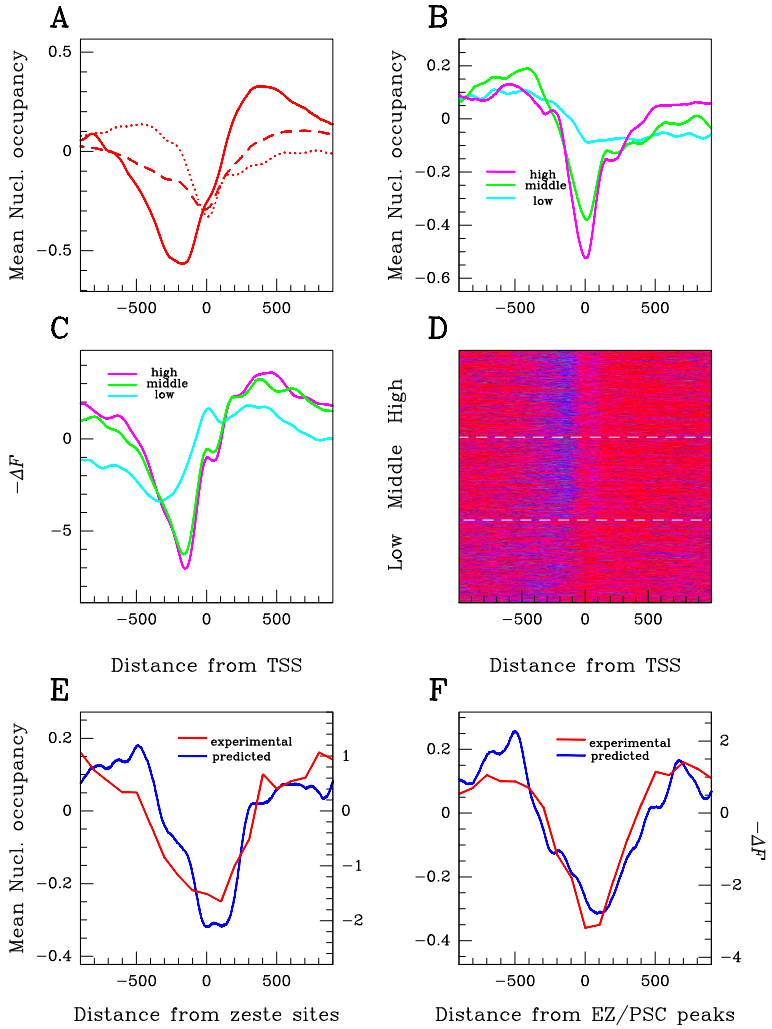


FIGURE 2.26 : *Comparaison des profils énergétiques prédits par le modèle de Miele* et al. *(Miele* et al.*, 2008) avec l'occupation en nucléosome chez* D. melanogaster. *(A) Moyennes autour du TSS de* 1610 *gènes (chromosoem 2L) des données expérimentales de Mito* et al. *(Mito* et al.*, 2007) brutes (log des données de digestion de l'ADN nucléosomal, courbe continue) et corrigées des biais de MNase (log des données nucleosomales sur les données de digestion et d'hybridation d'ADN nu génomique, pointillés) ou corrigés par les biais d'hybridation seuls (log des données nucléosomales sur les données de sonication et d'hybridation d'ADN nu, tirets). (B-D) Relation entre occupation en nucléosomes et activité transcriptionnelle. Les* 1610 *promoteurs ont été ordonnés selon leur occupation en pol II et divisée en trois classes (Mito* et al.*, 2005) : forte (bleu), moyenne (vert) et faible (rouge) densité en pol II. (B) Valeurs expérimentales corrigés par les biais de MNase et d'hybridation (log des données nucleosomales sur les données de digestion et d'hybridation d'ADN nu génomique) comme en (A). (C) Prédiction de l'occupation en nucléosomes via l'affinité $-\Delta F$ calculée à partir du modèle élastique avec les paramètres "Anselmi" (cf Chapitre 4). (D) Représentation 2D des profils $-\Delta F$ au niveau des* 1610 *promoteurs de gènes avec comme codage couleur, du rouge au bleu, des fortes énergies (faible affinité) aux faibles énergies (fortes affinité). (E,F) Comparaison entre les profils expérimentaux (comme en (B), courbe rouge) et les profils $-\Delta F$ (courbe bleue) moyennés autour de* 390 *et* 198 *sites de fixation de régulateurs du groupe Trithorax (Zeste) et Polycomb (EZ+PSC) respectivement.*

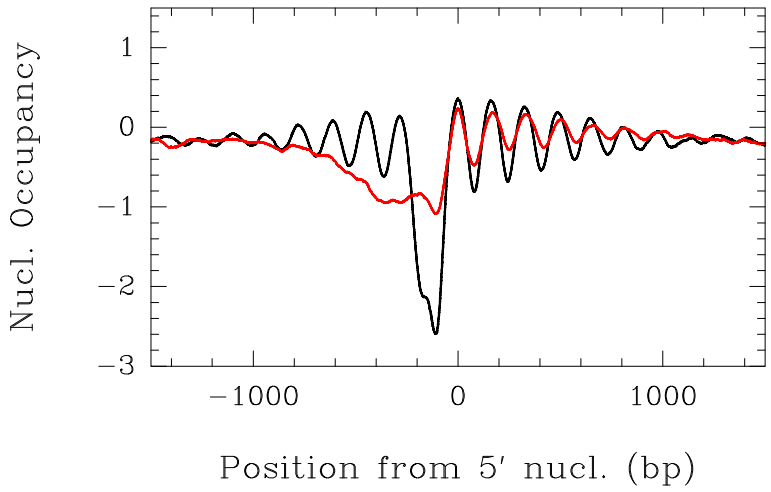


FIGURE 2.27 : *Le profil d'occupation nucléosomale des deux classes de gènes définies par Tirosh (Tirosh and Barkai, 2008a). En rouge les profils des gènes OPN (pour "Occupied Proximal Nucleosome"). En noir, les profils des gènes DPN (pour "Depleted Proximal Nucleosome").*

une occupation en nucléosome plus importante au niveau du promoteur. Chez la levure il y a principalement deux voies moléculaires associées à l'activation/initiation : la voie dite "SAGA", minoritaire, impliquant $10-15\%$ des gènes et la voie classique "TFIID" (Venters et al., 2011)(Fig. 2.23). La voie "SAGA"regrouperait plutôt des genes très fortement régulés du types de ceux de la classe 2 de Lee et des gènes OPN de Tirosh. Ce qui émérge de toutes ces études et d'autres (Tirosh and Barkai, 2008b; Lam et al., 2008; Boeger et al., 2008; Morse, 2007b), c'est que l'architecture nucléosomale du promoteur, plutôt que de jouer sur le niveau d'expression contôle la dynamique d'activation du gène ; c'est ainsi la distribution relative des sites de fixations des régulateurs vis-à-vis des positions des nucléosomes qui semble déterminante Lam et al. (2008) : un gène qui a ses sites dans une région constitutivement dépléte en nculéosome est "activable" plus facilement (rapidement) qu'un autre dont le site serait inclus dans un nucléosome.

Transcription (2) Elongation

Quel est le lien entre organisation du chapelet et les mécanismes d'élongation ? Comme pour l'activation, la progression de la polymérase est fortement régulée par un ensemble de co-facteurs 4.7. Peu de chose sont finalement connus quant à l'influence de l'organisation du chapelet sur la progression du complexe d'élongation. A ma connaissance il n'est pas encore clair si les nucléosomes sont totalement dissociés ou en partie (H2A-H2B) au passage de la polymérase. Ce qui apparait plus clairement c'est qu'assez rapidement après le passage du complexe, les nucléosomes se reforment.

Genes de levure

On souhaite observer de façon globale l'organisation nucléosomale sur l'ensemble des gènes de la levure sans faire de moyenne, à cause de la mauvaise annotation des fins de gènes. Nous proposons donc de les ranger par ordre de taille (ℓ) qui sépare le début (TSS) de la fin (TTS) de chaque gène. Chacun des profils de positionnement expérimentaux (données de Lee) est ensuite représenté et rangé dans un graphique (figure 2.30) où chaque point représente un minimum local du profil de positionnement (voir inserts dans la figure 2.31). La taille ℓ varie du plus petit (en haut) au plus grand (en bas). Cette représentation permet de bien figurer l'arrangement des nucléosomes à l'intérieur des gènes. La structuration apparaît très forte à proximité du TSS, et l'influence se ressent sur une grande distance (avec une portée de l'ordre de 5 à 7 nucléosomes, ce qui est cohérent avec les observations et les prédictions de la partie 5.3. Le TTS est moins marqué, mais on distingue toutefois une ligne incertaine qui marque la fin des gènes et qui manifeste la présence de trous de nucléosomes.

Nous envisageons alors l'utilisation d'un autre ordre pour ranger les gènes. Nous introduisons la distance L qui sépare le premier nucléosome en aval du TSS (le "$+1$") du dernier nucléosome en amont du TTS (le "-1"). Les nucléosomes $+1$ et -1 sont déterminés à partir de l'occupation nucléosomale de Lee, lissée par une gaussienne d'épaisseur 25 pb. Les maxima de ce signal sont détectés, et le $+1$ est recherché dans une fenêtre située dans l'intervalle $[-50\ bp, +200\ bp]$ autour du TSS. Le -1 est détecté dans

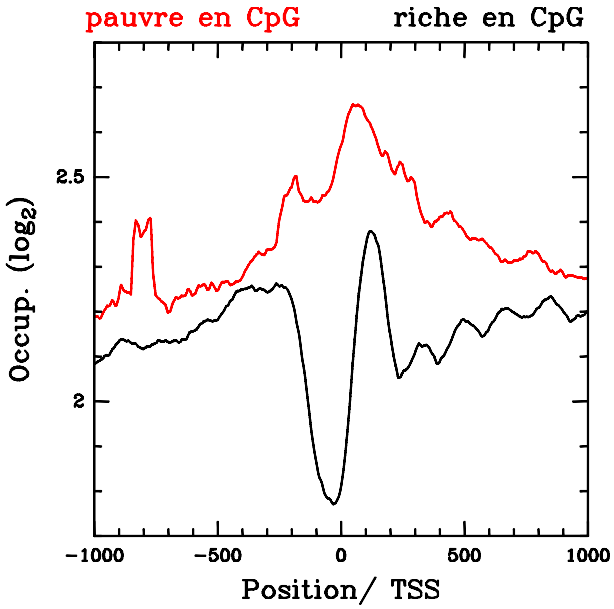


FIGURE 2.28 : *Profils moyens des données de Schones* et al. *(Schones* et al.*, 2008) autour du TSS des gènes riches en îlots CpG (noir) et pauvres en îlots CpG (rouge).*

l'intervalle $[-200\ bp, 50\ bp]$ autour du TTS. Les 50 paires de bases d'erreur que l'on s'accorde reflètent l'incertitude liée à la mauvaise annotation possible des TSS ou TTS. La figure que l'on construit avec la taille ℓ (figure 2.30 (A) vs (B)) présente les mêmes grandes caractéristiques que celle que l'on construit avec la distance L. Ce choix de l'ordre L a le désavantage de systématiquement produire du signal : si en guise de contrôle, on applique l'algorithme d'ordonnancement sur un set de "faux gènes" c'est-à-dire sur un jeu de données de TSS-TTS similaire au jeu réel, mais positionnés aléatoirement sur le génome de la levure, une partie du signal observé expérimentalement est reproduite (figure 2.30 (D)). Il faut donc bien prendre conscience que la méthode d'alignement et le choix de l'ordre ne sont pas anodins et auront tendance à produire un signal beaucoup très net naturellement, même sur des séquences de contrôle aléatoires. Le signal est pourtant plus contrasté dans la figure 2.30 (C) que dans le contrôle (figure 2.30 (D)), et c'est la différence entre ces figures qui explicite les particularités des gènes. De ce point de vue, un gène est signalé par une NFR prononcée au TSS et dans une moindre mesure au TTS. Une structuration dissymétrique très nette à l'intérieur du gène atteste d'une forte interaction défavorable au niveau du TSS.

La carte nucléosomale des gènes reportée donc à la figure 2.31 illustre donc bien la présence de déplétion de part et d'autres des gènes, avec une déplétion plus marquée et plus large au niveau du TSS (en fait ici au niveau du $+1$) qu'au TTS (-1) (voir encarts, figure 2.31). Pour les petites gènes ($L = distance(+1, -1) < 1500pb$) on observe un arrangement très périodique sur tout le long du gène. Pour les plus grand gènes, cette periodicté reste ensuite confinée au voisinage des deux zones de déplétion, et l'intérieur semble plutôt révéler un profil non stucturé (flou). Dans la région des petits gènes on voit bien la succession des gènes à $2, 3, 4, 5...7$ nucléosomes ordonnés périodiquement à mesure que la taille augmente. Cette "cristalisation"pour ces gènes et illustrée pour les cas $n = 5$ et $n = 6$ par le profil moyen (encarts). On remarque par ailleurs qu'entre ces domaines (domaines de distance L) de gènes cristallins il y a quelques gènes qui ont un profil flou comme l'atteste aussi la moyenne pour ceux qui sont entre les gènes cristallins $n = 5$ et $n = 6$. L'interprétation de cette organisation remarquable du chapelet intre-génique fera l'objet du chapitre 10.

La question générale que posent ces observations expérimentales est de savoir quelles sont les mécanismes moléculaires qui contrôlent le chapelet nucleosomal le long des génomes, à savoir l'occupation en nucléosome à tout échelle : qu'est ce qui détermine les zones de déplétion, les zones de positionnement périodique et les zones "floues" ? Qu'est ce qui contôle la caractère cristallins ou flous des gènes chez le levure ? Dans quelle mesure ces différentes organisations du chapelet sont elles codées dans la séquence ?

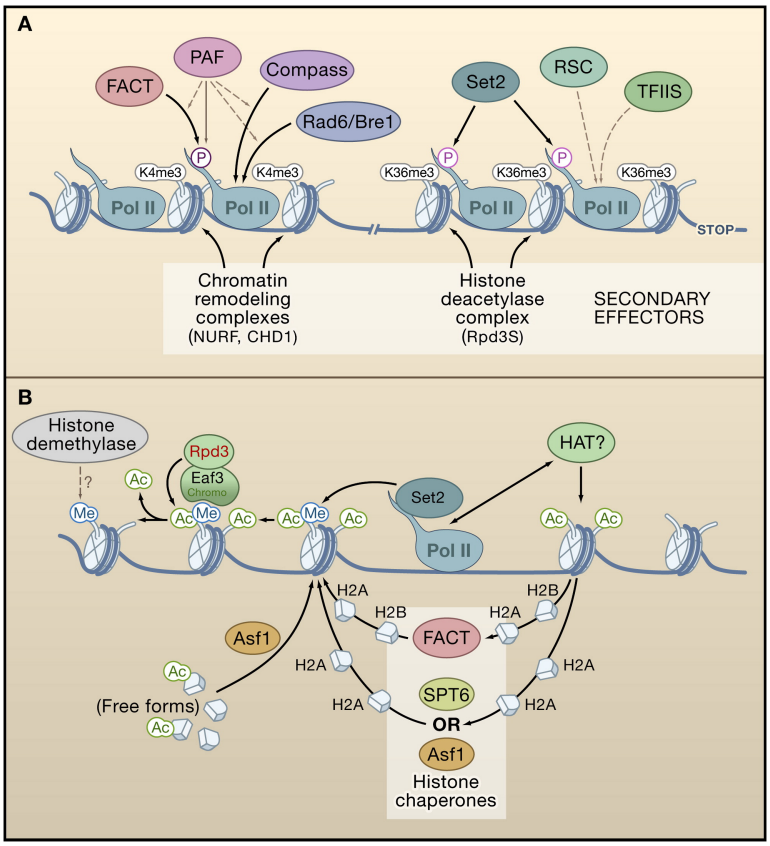


FIGURE 2.29 : *Régulation de la chromatin durant l'élongation. (A) Le paysage chromatinien au cours de l'élongation est déterminé par des fcatreurs associés à différents états de pol II. PAF facilite la fixation de FACT, COMPASS, et Rad6/Bre1 au domaine CTD Ser5-phosphorylé, induisant l'ubiquitination d'H2B et l'accumulation de H3K4me3 à la fin 5′ de l'ORF. Set2 qui intéragit directement avec le CTD Ser2-phosphorylé, induit la methylation H3K36 à la fin 3′. (B) Maintien de la stabilité nucléosomale durant la transcription. Lorsque Pol II s'éloigne du promoteurs, lorsque l'influence des HATs recrutées par les activateurs diminuent, Pol II utilise d'autres HATs pour acetyler les nucléosomes au devant de la machines d'élongation. Le passage de Pol II provoque l'éviction des nucléoosmes qui sont redéposés ensuite derrière la Pol II via les actions concertées de chaperones d'histones. Un pool d'histones libres dans le noyau est également disponible pour la redéposition. Ces histones déposés de novo sont en général hyperacétylés et sont de suite methylés par set2. La methylation H3K36 est reconnue par le chromodomaine de Eaf3, qui recrute ensuite le complexe de deacétylation Rpd3S qui supprime le groupe acetyl stabilisant le nucléosome. La méthylation de H3K36 est ensuite éliminée par une histone demethylase lorsque que le gène est desactivé.*

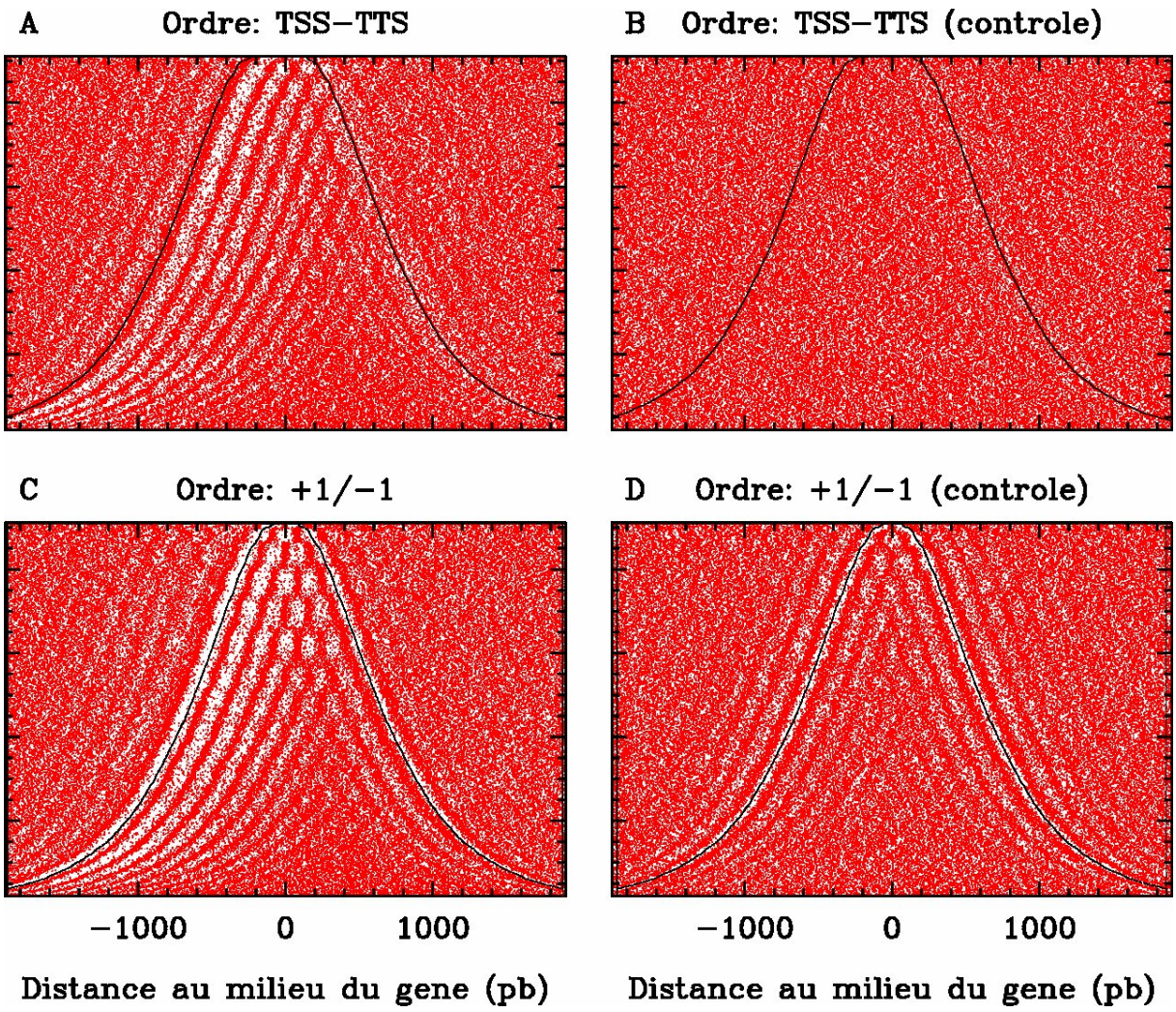


FIGURE 2.30 : *Carte nucléosomale 2D le long des gènes de la levure. (A)* 4554 *gènes sont organisés verticalement en fonction de la distance TSS-TTS. Une ligne horizontale correspond donc à l'occupation au sein d'un gène. En rouge : les minima locaux de l'occupation des données de Lee (Lee* et al.*, 2007a) sont représentés. Les nucléosomes occupent les zones blanches. (B) La même carte, mais dessinée sur un jeu de gènes dit de contrôle, dont les positions sont choisies aléatoirement sur le génome. (C) Les gènes sont ordonnés en fonction de la longueur* L*, qui sépare le nucléosome* $+1$ *du* -1*. (D) Même figure que C, mais réalisée sur le jeu de gènes de contrôle.*

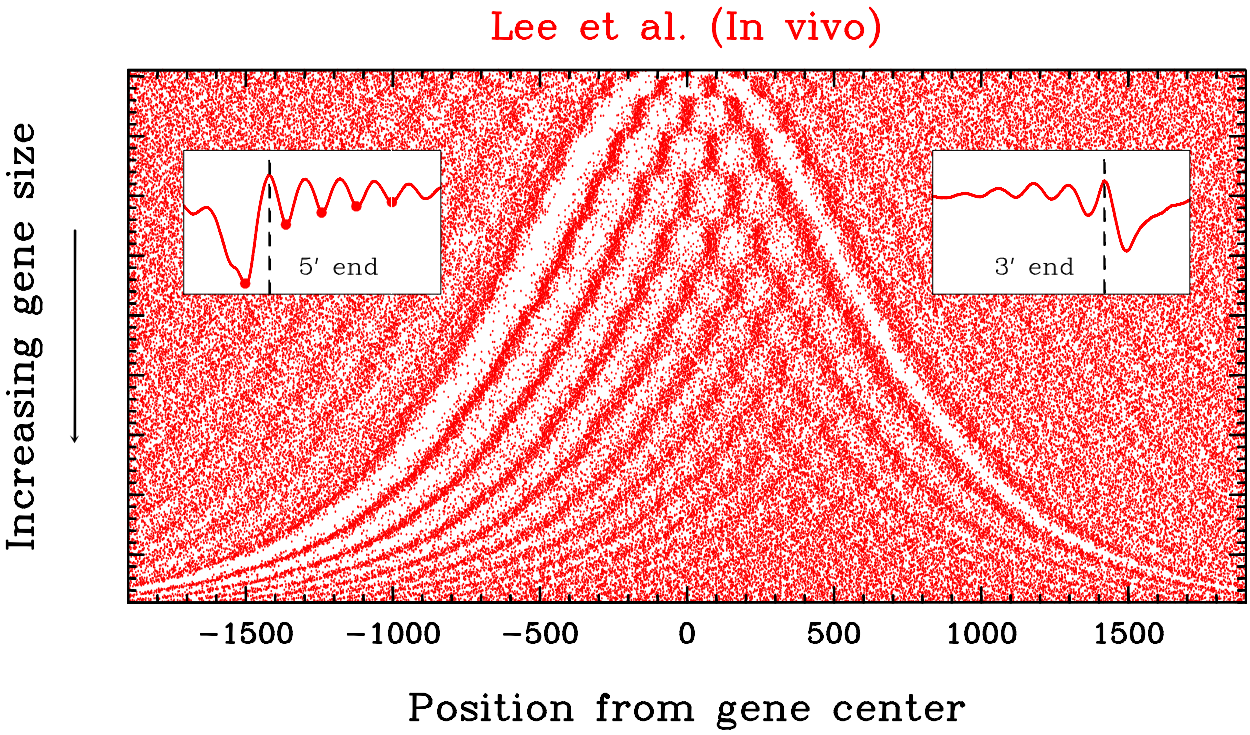


FIGURE 2.31 : *Carte nucléosomale 2D le long des gènes de la levure. (a)* 4554 *gènes sont organisés verticalement en fonction de la distance ℓ qui sépare le premier nucléosome du dernier nucléosome à l'intérieur du gène (définit par la position du TSS au TTS). Une ligne horizontale correspond donc à l'occupation au sein d'un gène. En rouge : les minima locaux de l'occupation des données de Lee (Lee* et al.*, 2007*a*) sont représentés. Les nucléosomes occupent les zones blanches. Dans les inserts sont reportées les profils moyens centrés sur le nucléosome +1 en 5′ (gauche) et -1 en 3′ (droite).*

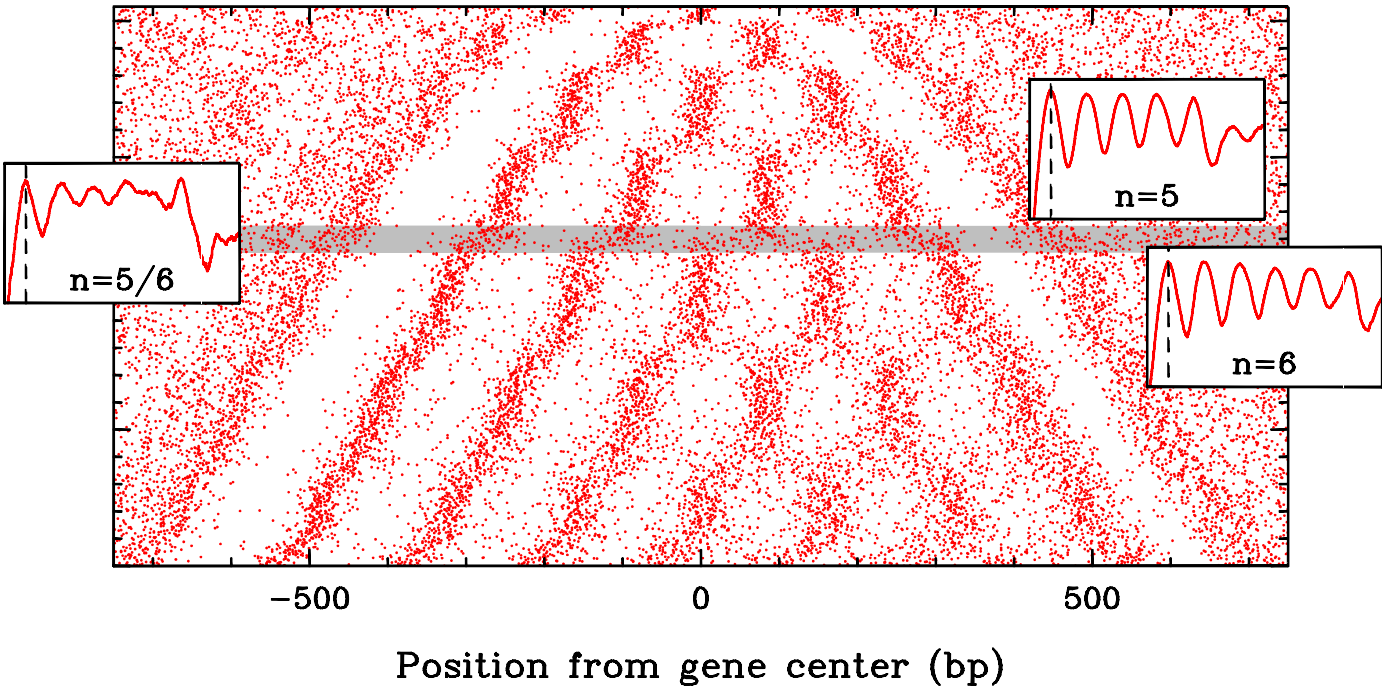


FIGURE 2.32 : *Carte nucléosomale 2D le long des gènes de la levure. (a)* 4554 *gènes sont organisés verticalement en fonction de la distance ℓ qui sépare le premier nucléosome du dernier nucléosome à l'intérieur du gène (défini par la position du TSS au TTS). Une ligne horizontale correspond donc à l'occupation au sein d'un gène. En rouge : les minima locaux de l'occupation des données de Lee (Lee* et al.*, 2007a) sont représentés. Les nucléosomes occupent les zones blanches. Dans les inserts sont reportées les profils moyens centrés sur le nucléosome +1 en 5′ (gauche) et -1 en 3′ (droite).*

3 CHAPELET NUCLÉOSOMAL : MODÈLE

L'objectif ici a donc été de modéliser les profils de positionnement observes dans les différents organismes. Pour cela il faut d'abord définir un modèle de déposition d'une assemblée dense de nucléosomes.

3.1 MODÈLE THERMODYNAMIQUE

3.1.1 Liquide de "sphères" dures

Si l'on oublie les interactions entre nucléosomes éloignés, et que l'on ne tient compte que des interactions entre proches voisins, il est possible de décrire l'assemblage de nombreux nucléosomes sur l'ADN comme un fluide unidimensionnel de tiges rigides. La description de ces fluides a déjà été effectuée en détail par les physiciens et il est possible d'accéder théoriquement à des paramètres importants tels la densité, les fonctions de paires, etc.

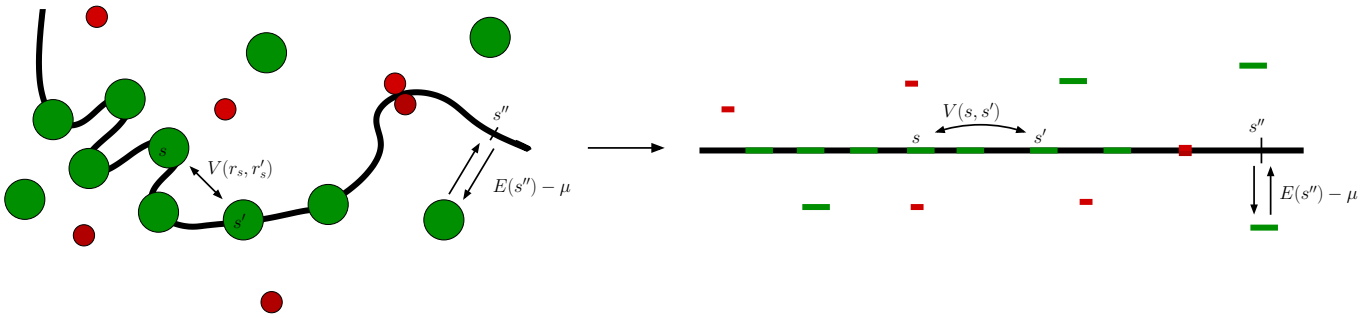


FIGURE 3.1 : *Le modèle d'assemblege des nucléosomes : un réservoir d'histones (en vert), adsorbe ou desorbe (flèches) l'ADN (en noir). Des obstacles, tels les facteurs de transcription ou tout autre molécule qui se fixe sur l'ADN (en rouge) peuvent également se fixer et empêcher les nucléosomes de se fixer. Le modèle complet considère donc une ligne unidimensionnelle sur laquelle se fixent des tiges impénétrables en équilibre avec un réservoir d'histones.*

3.1.2 Fluides de Tonks-Takahashi

Il existe un modèle détaillé de fluides en interactions de plus proches voisins qui permet d'accéder notamment à la fonction de partition et à la fonction de paire.

Les fluides de Tonks-Takahashi sont très généraux, car ils peuvent prendre en compte une interaction un peu moins contrainte que les coeurs durs :

$$u(x_{ij}) = \infty,\ |x_{ij}| < l \tag{3.1}$$
$$= \psi(x_{ij} - l),\ l < |x_{ij}| < l_2 \tag{3.2}$$
$$= 0,\ |x_{ij}| > l_2 \tag{3.3}$$

où l correspond à une distance d'interaction de coeur dur, et l_2 est la distance maximale d'interaction de deux particules. La description de ce type de système a d'abord été faite par Lord Rayleigh (Rayleigh, 1891) qui a donné la première formulation de l'équation d'état du système. Ses résultats ont été oubliés et redécouverts par Tonks et Takahashi qui y ont donné leur nom. Les fonctions de distribution pour les systèmes homogènes dans la limite thermodynamique ont été calculées par Salsburg et *al.* (Salsburg et al., 1953) et enfin, Robledo *et al.* (Robledo and Rowlinson, 1986) ont étudié exhaustivement les effets de taille finie sur les fonctions de partition. Davis a généralisé ce résultat pour des systèmes non-homogènes (Davis, 1990), et Percus a obtenu une équation guidant la densité de sphères dures dans un potentiel 1D hétérogène (Percus, 1976). Enfin Vanderlick a montré qu'il existait une solution analytique à l'équation de Percus (Vanderlick et al., 1986).

Si l'on souhaite déterminer les fonctions d'état de ce genre de système, on commence généralement par exprimer la fonction de partition de N particules plongées dans un profil énergétique $E(x)$ quelconque de taille L :

$$\begin{aligned} Z_N(L) &= N! \int_0^L dx_N \int_0^{x_N} dx_{n-1} \dots \int_0^{x_2} dx_1 \\ &\quad \times e^{-\beta(\sum_{i>j}^N u(x_i - x_j) + \sum_i^N E(x_i))} \\ &\quad \times e^{-\beta(u(x_1) + u(L - x_N))} \end{aligned} \tag{3.4}$$

où les x_i correspondent aux positions des N particules rangées dans l'ordre d'indexation, et $\beta = 1/kT$. Les conditions aux limites $u(x_1)$ et $u(L - x_N)$ considèrent qu'une particule est maintenue fixe en 0 et une autre en L. La fonction de partition grand-canonique (lorsque le système est en contact avec un réservoir de particules, N peut varier, et le potentiel chimique μ détermine l'énergie gagnée par le système lors de l'absorption d'une particule) peut être exprimée sous la forme :

$$\Xi(L) = \sum_{N=0}^{\infty} \frac{e^{\beta N \tilde{\mu}}}{N! \Lambda^N} Z_N \tag{3.5}$$

où Λ est la longueur d'onde de de Broglie que l'on prendra égale à 1 dès qu'il sera nécessaire d'implémenter les résultats, cela correspond à changer le potentiel chimique en $\mu \to \tilde{\mu} = \mu - \frac{1}{\beta} \ln \Lambda$. Enfin la

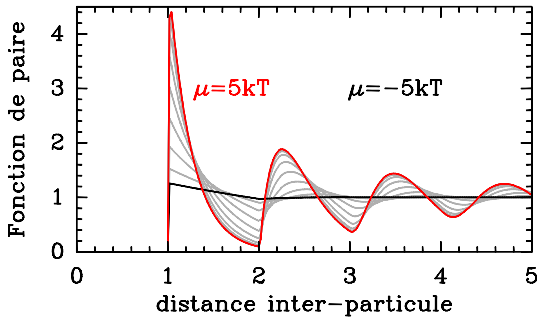


FIGURE 3.2 : *Évolution de la fonction de paire (équation 3.1.2) avec le potentiel chimique μ. Les courbes noires et rouges correspondent respectivement à un potentiel chimique de $\mu = -5$ et $+5$ kT. Les courbes grises sont des intermédiaires.*

densité $\rho(x)$ est déterminée par :

$$\rho(x) = -kT\frac{\delta \ln \Xi(L)}{\delta E(x)} \tag{3.6}$$

On peut obtenir une forme explicite pour la densité en fonction des fonctions de partitions (Robledo and Rowlinson, 1986; Davis, 1990) ainsi que pour les fonctions de paires, et d'ordre supérieurs (fonctions de distribution à k corps) :

$$\rho(x) = e^{\beta(\tilde{\mu}-E(x))}\frac{\Xi(x)\Xi(L-x)}{\Xi(L)}, \tag{3.7}$$

$$\rho^{(k)}(x_1,\ldots,x_k) = e^{\beta k\tilde{\mu}}\frac{e^{-\beta\sum_{i=1}^{k}E(x_k)}}{\Xi(L)} \times\Xi(x_1)\prod_{j=2}^{k}\Xi(x_j-x_{j-1})\Xi(L-x_k) \tag{3.8}$$

$\rho(x)$ correspond à la probabilité de trouver une particule en x. $\rho^{(2)}(x_i,x_j)$ correspond à la probabilité de trouver la particule en x_j sachant qu'il y en a une en x_i. C'est un cas particulier important parce qu'il s'agit de la fonction de paire qui permet de préciser la distance typique entre deux particules, et qui renseigne expérimentalement sur l'interaction entre particules. La fonction de paire fait également écho à la période nucléosomale mesurée classiquement sur la chromatine (voir paragraphe 5.2.2). La figure 3.2 présente la forme de la fonction de paire dans un système homogène, avec des interactions de type volume exclu. Elle est nulle sur une distance égale à la taille des particules. De plus la fonction de paire forme des oscillations dont l'amplitude dépend du potentiel chimique, autrement dit de la pression à l'intérieur du système. Lorsque la pression est forte, les particules sont collées les unes aux autres, ce qui implique qu'il est quasi-certain de trouver une particule à proximité immédiate de la particule fixée puis, puisqu'il y a certainement une particule située à une distance $1l$ de la première particule, l'espace situé entre $1l$ et $2l$ est très certainement inaccessible. Une nouvelle particule peut se placer, et le phénomène se répète à nouveau. Lorsque le potentiel chimique est faible, la fonction de paire est plate, car il est équiprobable de trouver une particule quelque soit la distance entre deux particules (sauf évidemment sur l'espace $[0,l]$). La pression dans ce type de système peut être calculée directement (Davis, 1990; Lieb and Mattis, 1966) :

$$P = kT\frac{\partial \ln \Xi(L)}{\partial L} \tag{3.9}$$

Ces formules ont beau donner une formulation explicite de grandeurs thermodynamique d'intérêt, elles ne nécessite pas moins de calculer les fonctions de partition totales du système. Dans certaines circonstances, il est possible de réaliser des approximations permettant de simplifier les expressions obtenues. Si l'on s'intéresse par exemple à un milieu infini et homogène, à savoir $v = 0$ et $L \to \infty$, alors, moyennant l'introduction de la transformée de Laplace $K(s) = \int_0^\infty e^{-sy-\beta\psi(y)}dy$ de l'interaction entre particules, on peut montrer (Lieb and Mattis, 1966; Salsburg et al., 1953) que :

$$\beta\tilde{\mu} = \beta Pl - \ln K(\beta P) \tag{3.10}$$

qui dans le cas des tiges rigides ($K(s) = 1/s$) donne les équations d'état déterminées par Rayleigh en 1891 (Rayleigh, 1891), à savoir :

$$\beta\tilde{\mu} = \beta Pl + \ln(\beta Pl) + \ln(1/l) \tag{3.11}$$

$$\beta P = \frac{\rho}{1-\rho l} \tag{3.12}$$

résultat qui vaut pour n'importe quel fluide homogène de tiges rigides.

En particulier si la formation des nucléosomes était indépendante de sa position sur le génome, si la séquence génomique ne jouait pas sur l'affinité du nucléosome avec l'ADN, et si les nucléosomes étaient strictement non inter-pénétrables, alors c'est cette densité nucléosomale que l'on obtiendrait.

Il n'existe pas de solution analytique pour l'expression des fonctions de distribution avec des interactions quelconques. Percus a proposé des solutions particulières dans le cadre d'interactions de plus proches voisins différentes du simple cas des sphères dures (Percus, 1982). Enfin, le cas particulier d'un champ extérieur uniforme est analysé et une solution exacte est possible (Ibsen et al., 1997).

3.1.3 Équation de Percus

L'équation de Percus précise la description de Tonks-Takahashi dans le cas de particules plongées dans un profil inhomogène. La densité peut être déterminée via une équation reliant le potentiel énergétique et le potentiel chimique.

The Percus equation gives the density of a Tonks-Takahashi fluid in an inhomogeneous energetic field as a function of the chemical potential and the temperature.

À partir des équations d'état, il est possible de construire une formulation qui ne fait intervenir que la densité, le potentiel chimique et la taille des tiges rigides. En remplaçant P dans l'équation 3.11, par son expression dans l'équation 3.12, on obtient :

$$\beta\mu = \frac{\rho l}{1-\rho l} + \ln\frac{\rho l}{1-\rho l} + \ln(1/l) \tag{3.13}$$

cette équation n'est autre que le cas particulier $E = 0$ de l'équation de Percus (Percus, 1976) dont on se servira dans le reste de l'étude. L'équation de Percus 3.14 a valeur relativement générale, puisqu'elle prend en compte la non homogénéité du profil énergétique dans lequel sont plongées des tiges rigides.

$$\beta\mu = \beta E(s,l) + \ln\,\rho(s) - \ln\left(1 - \int_s^{s+l}\rho(s')ds'\right) + \int_{s-l}^{s}\frac{\rho(s')}{1-\int_{s'}^{s'+l}\rho(s'')ds''}ds' \tag{3.14}$$

ÉQUATION DE PERCUS

où

- s correspond à la position le long du potentiel, en l'occurrence il s'agit de la paire de base sur laquelle est centré le nucléosome, aussi appelée "dyade".
- l correspond à la taille de la tige rigide, ici le nucléosome, donc environ 146 paires de bases.
- ρ est la densité en tiges rigides (les nucléosomes) : c'est ce que l'on cherche à établir, et c'est ce à quoi on accède indirectement de façon expérimentale. (Bien souvent, c'est la probabilité d'occupation qui est mesurée, qui correspond à la probabilité qu'un site donné soit recouvert par un nucléosome. D'un point de vue mathématique, il s'agit simplement de la convolution de la probabilité par une fenêtre rectangulaire de l'épaisseur d'un nucléosome.)
- μ représente le potentiel chimique, il détermine l'énergie gagnée par le système lorsque l'on rajoute une tige rigide dans le système.
- β est l'inverse de l'énergie de bain thermique à savoir $(kT)^{-1}$. Puisque l'on ne connait pas la température effective qui affecte notre système, on prendra simplement $\beta = 1$. Les énergies que nous obtenons sont donc en échelles de kT.
- $E(s,l)$ est l'énergie libre de formation du nucléosome le long de la séquence. C'est en particulier le potentiel à 1D qui dépend de la séquence dont il est question au chapitre 4.

3.1.4 Bilan des méthodes utilisables

Pour déterminer le positionnement de tiges rigides dans un potentiel inhomogène, deux écoles s'affrontent. La première méthode consiste à calculer directement la fonction de partition, soit par une méthode de chaine de Markov (Segal et al., 2006) soit par la méthode des matrices de transfert (Teif and Rippe, 2009). Il est possible de calculer la densité autrement, en résolvant l'équation de Percus directement (Vanderlick et al., 1986). Dans ce dernier cas, les calculs sont beaucoup moins lourds mais le calcul direct de la fonction de partition permet plus de liberté sur le type d'interactions entre particules, et permet aussi le calcul facile de fonctions de distributions autres que la densité (fonction de paire, etc).

3.1.5 Résolution de Percus rapide

Il existe une implémentation simple mais approximative qui permet de résoudre l'équation de Percus.
There is an approximate but very fast way to solve the Percus equation.

Notes sur le domaine d'étude

Pour peu que l'on s'intéresse à des systèmes physiques réels non infinis, il est toujours possible de restreindre l'étude au domaine $s \in [0; L]$ où L désigne la longueur totale du potentiel 1D en supposant que toutes les fonctions d'intérêt sont nulles en dehors de ce domaine (à l'exception de E qui serait infini en dehors du domaine, de façon à ce que les équations restent valides). Dans la suite on ne se préoccupera donc pas des dimensions finies de nos systèmes. Par ailleurs, il peut paraitre curieux d'utiliser un modèle continu discrétisé quand bien même le brin d'ADN apparaît fondamentalement discontinu puisque l'on considère une succession de paires de bases comme structure de notre potentiel. Mais cela n'est pas du tout injustifié : d'abord, considérer que l'ADN est quantifié est erroné, puisque *a priori* rien n'empêche un nucléosome de se fixer entre deux bases, même si la structure cristallographique, qui ne prend en compte aucune dynamique, suggère que non. Définir le début et la fin d'un nucléosome précisément est d'ailleurs hasardeux. Expérimentalement, on détermine les positions de nucléosomes au mieux à la paire de base près. Ensuite, la description continue n'est que mathématique, dès lors que l'on implémentera la solution de façon numérique, il faudra bien quantifier l'espace. Le pas de quantification peut être pris à 1 pb, mais il peut très bien être pris plus petit, si l'on souhaite examiner des phénomènes de très petite échelle (cela n'a d'ailleurs pas encore d'intérêt puisqu'aucune étude expérimentale ne permet d'établir une résolution supérieure à la paire de base pour le moment) ou plus grand, si l'on veut faire abstraction de phénomènes de petite échelle.

Percus donc…

Il est intéressant de noter que dans l'équation de Percus apparaît assez naturellement l'intégrale $\int_s^{s+l} \rho(u)du$ qui correspond à la probabilité d'occupation dans l'intervalle s et $s+l$; *i.e.* si un nucléosome est systématiquement situé à la position s_0 alors cette intégrale vaut 1 et le terme en $1 - \int_s^{s+l} \rho(u)du$ qui correspond à l'espace libre à proximité de s s'annule et la densité vaut nécessairement 0. Notons également qu'il est possible de réécrire l'équation sous une forme moins explicite :

$$f(s) = exp\left(\beta\mu - \beta E(s,l) - \int_{s-l}^{s} f(s')ds'\right) \tag{3.15}$$

$$\frac{\rho(s)}{1 - \int_s^{s+l} \rho(s')ds'} = exp\left(\beta\mu - \beta E(s,l) - \int_{s-l}^{s} \frac{\rho(s')}{1 - \int_{s'}^{s'+l} \rho(s'')ds''}ds'\right) \tag{3.16}$$

$$f(s) \equiv \frac{\rho(s)}{1 - \int_s^{s+l} \rho(s')ds'} \tag{3.17}$$

Comme on peut le voir ici, la fonction f –pour *forward*– que nous avons introduite est une fonction qui ne dépend que du *passé* (en effet, si l'on connait f sur l'intervalle $]s-l; s[$ on peut déterminer $f(s)$). Si l'on néglige le passage à la limite discrète, il est possible de résoudre l'équation 3.17 itérativement, en partant d'une condition initiale simple sur f : $f(s) = 0,\ s \in [-l; 0]$. Lorsque l'on essaye d'implémenter numériquement cette solution, il se pose le problème du calcul de l'intégrale de $s-l$ à l qui, s'il est possible dans une description continue, ne l'est plus dans le cadre d'une discrétisation. Étant donné que l'on ne connait pas la valeur de f en s, on commet nécessairement une erreur dans le calcul de la somme

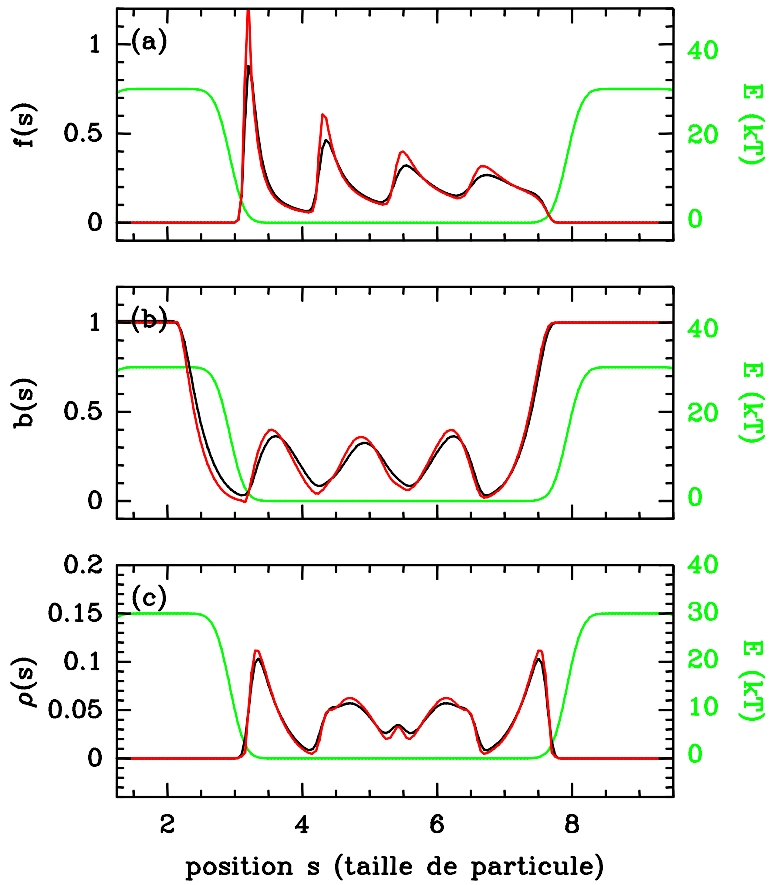


FIGURE 3.3 : *Performances de la méthode de résolution rapide de Percus comparée avec la solution exacte de Vanderlick (Vanderlick* et al.*, 1986). Profil énergétique utilisé pour le calcul en vert. Calculs avec Percus rapide en rouge, calculs avec Vanderlick en noir, corrélation de Pearson entre les profils :* 0.9956*. (a)* $f(s)$ *fonction "forward" (b)* $b(s)$ *fonction "backward" (c)* $\rho(s)$ *densité résultante. Paramètres : Amplitude du potentiel* $+30$ *kT,* $\mu = +3$ *kT*

discrète. Si l'on écarte un instant ce problème, il faut encore déterminer la densité ρ à partir de f. Pour cela, on introduit une nouvelle fonction, b telle que :

$$f(s) = \frac{\rho(s)}{b(s)} \tag{3.18}$$

Des équations 3.17 et 3.18 on déduit facilement :

$$b(s) = 1 - \int_s^{s+l} \rho(s')ds' = 1 - \int_s^{s+l} f(s') \cdot b(s')ds' \tag{3.19}$$

Cette fois, la fonction b –pour *backward*– ne dépend que du *futur*, il est de nouveau possible de la résoudre itérativement en imposant la condition initiale $b(s) = 1,\ s \in [L; L + l]$. Une fois qu'on a calculé f puis b, on peut retrouver la densité $\rho(s) = f(s) \cdot b(s)$. On obtient des résultats très similaires à la solution exacte du problème (Vanderlick complet, voir ci-après), on obtient par exemple une corrélation de Pearson de 0.9956 entre les deux profils de la figure 3.3 (c). Cependant cette méthode très simple, ne permet pas de résoudre correctement le problème car dès que les intégrales deviennent grandes, les erreurs commises lors de la discrétisation des intégrales sous forme de sommes discrètes deviennent non négligeables.

3.1.6 Résolution de Percus par Vanderlick

Une astuce calculatoire permet de résoudre exactement l'équation de Percus et permet donc de déterminer la densité exactement à partir d'un profil énergétique inhomogène.

A technical trick is proposed by Vanderlick in order to solve Percus exactly, so that the density of hard rods in an inhomogeneous field can be tracted.

Comme on l'a vu, tout le problème tient à ce que, pour déterminer une fonction en s, on doive calculer une intégrale de $s-l$ jusqu'à s inclus. Il est possible de remédier à ce problème grâce à la solution proposée par Vanderlick, Scriven et Davis en 1986 (Vanderlick et al., 1986). Partant de l'équation (3.17) on déduit :

$$\frac{df}{ds}(s) = f(s) \cdot \left[-\beta \frac{\partial E}{\partial s}(s,l) + f(s-l) - f(s) \right] \tag{3.20}$$

de solution générale :

$$f(s) = \frac{u(s)}{\frac{exp(-\beta E(s_0,l)))}{f(s_0)} + \int_{s_0}^{s} u(s')ds'} \tag{3.21}$$

où

$$u(s) \equiv exp\left(-\beta E(s,l) + \int_{s_0}^{s-l} f(s')ds' \right)$$

Il est possible de déterminer $f(s)$ itérativement, comme précédemment sauf que cette fois, $f(s)$ ne dépend que des valeurs de f pour des abscisses inférieurs à $s-l$. Cette solution permet une implémentation exacte, sans approximation liée aux intégrales, autre que la discrétisation. Pour déterminer la densité $\rho(s)$ on passe à nouveau par la fonction $b(s)$ qui cette fois obéit à la relation (dérivée de l'équation (3.19)) :

$$b'(s) = f(s)b(s) - f(s+l)b(s+l) \tag{3.22}$$

de solution générale :

$$b(s) = b(s_0)exp(\int_{s_0}^{s} f(s')ds') + \int_{s}^{s_0} \left[e^{-\int_{s}^{s''} f(s')ds'} b(s''+l)f(s''+l) \right] ds'' \tag{3.23}$$

où là encore la résolution des intégrales portant sur b a été déplacée de l. Les conditions aux limites sont les mêmes que précédemment, *i.e.* $f(s<0)=0$ et $b(s>L)=1$. Il est à noter que l'article original de Vanderlick propose de découper la ligne 1D en segments de taille l, ce qui n'est absolument pas nécessaire à la résolution de l'équation autrement que pour réduire les valeurs numériques des intégrales qui sont calculées.

3.1.7 Solution de Segal

La résolution de l'équation de Percus n'est pas la seule méthode pour déterminer la densité de tiges rigides dans un potentiel 1D. Dans un article publié en 2006 (Segal et al., 2006), Eran Segal propose une autre solution, où les fonctions de partitions sont calculées explicitement. La méthode employée est elle aussi exacte mais donne lieu au calcul de sommes dont les termes peuvent rapidement prendre des valeurs très importantes. D'après Percus (Percus, 1982), on a :

$$\rho(s) = w(s)\frac{\xi(s)\hat{\xi}}{\xi_T} \text{ avec } w(s) = e^{\beta(\mu - E(s))}$$

où

$$\begin{aligned}
\xi(s) &= \sum_{N=0}^{\infty} e^{N\beta\mu} Q_N(s) \\
\hat{\xi}(s) &= \sum_{N=0}^{\infty} e^{N\beta\mu} \hat{Q}_N(s) \\
Q_N(s) &= Z_N^{gauche}(0 \leftrightarrow s) \\
\hat{Q}_N(s) &= Z_N^{droite}(s \leftrightarrow L)
\end{aligned}$$

Toute la difficulté réside dans la détermination de Z_N^{gauche} et de Z_N^{droit} qui sont respectivement le nombre de configurations accommodant N particules dans l'espace situé à gauche de s ($0 \leftrightarrow s$) et le nombre de configurations accommodant N particules dans l'espace situé à droite de s ($s \leftrightarrow L$). Il y a bien une formulation explicite (Percus, 1982) :

$$Z_N = \int \cdots \int \prod_{i=2}^{N} e^{-\beta\phi(s_i - s_{i-1})} \prod_{i=1}^{N} e^{-\beta E(s_i)} \prod_{i=1}^{N} ds_i \tag{3.24}$$

où $\phi(s_i - s_{i-1})$ est le potentiel d'interaction entre deux particules, nul si $s_i < s_{i-1}$. Malheureusement, cette formulation n'est pas facile à calculer. Cependant, il est possible d'effectuer une récurrence sur $Q_N(s)$ dans le cas particulier d'une interaction de tiges rigides :

$$Q_N(s+1) = Q_N(s) + Q_{N-1}(s-l)e^{-\beta E(s+1-l)}$$

où $E(s)$ est l'énergie d'un nucléosome qui débute en s. Alors si l'on note N_0 le nombre maximum de particules que l'on peut mettre dans s,

$$\begin{aligned}
\xi_\mu(s+1) &= \sum_{N=0}^{\infty} e^{N\beta\mu} Q_N(s+1) \\
&= \sum_{N=0}^{N_0} e^{N\beta\mu} Q_N(s) + \sum_{N=0}^{N_0} e^{N\beta\mu} Q_{N-1}(s-l) e^{-\beta E(s+1-l)} \\
&= \xi_\mu(s) + \sum_{N=0}^{N_0-1} e^{N\beta\mu} e^{+\beta\mu} Q_N(s-l) e^{-\beta E(s+1-l)} \\
&= \xi_\mu(s) + \xi_\mu(s-l) e^{\beta(\mu - E(s+1-l))}
\end{aligned}$$

NB : dans $s - l$ il est possible de mettre au maximum $N_0 - 1$ particules, donc $\xi_\mu^{N_0-1} = \xi_\mu^{N_0}$. On opère strictement de la même façon pour déterminer $\hat{\xi}$ et la densité ρ peut être déterminée (équation 3.1.7).

3.1.8 Solution de Teif

Récemment, Vladimir B. Teif (Teif and Rippe, 2009) a proposé une méthode fondée sur les matrices de transfert permettant de calculer la fonction de partition totale d'un système constitué d'un bain de particules pouvant se lier à un brin d'ADN. L'idée est d'associer à chaque position dans la séquence une matrice de $R \times R$ où R définit le nombre d'états accessibles à chacune des paires de bases. Dans le cas qui nous intéresse, il y a 146 états, correspondant à chacune des positions possibles dans le nucléosome. De plus, il faut rajouter deux états pour définir les conditions aux limites : fin de la séquence à gauche et fin de la séquence à droite. Les éléments de la matrice $Q_n(i,j)$ correspondent physiquement à la probabilité pour la paire de base n d'être dans l'état i sachant que la paire de base $n+1$ est dans l'état j. Ces éléments sont en fait non nuls si et seulement si la paire de base $n+1$ est dans l'état $i+1$. Pour rendre compte de l'influence de la séquence, l'élément $Q_n(1,2)$, qui définit la probabilité que la paire de base n soit occupé par le début d'un nucléosome, est assigné le poids statistique $w = e^{\beta(\mu - E(n))}$. Enfin on peut déterminer la fonction de partition totale du système en sommant sur tous les états accessibles, *i.e.* :

$$Z = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \times \prod_{i=1}^{N} Q_n \times \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix} \tag{3.25}$$

Notons qu'il est possible de calculer cette fonction de partition par récurrence :

$$Z = A_N \times \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix},\ A_i = A_{i-1} \times Q_n,\ A_0 = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \tag{3.26}$$

Et pour déterminer la densité de nucléosomes en n, il suffit de dériver la fonction de partition par rapport à $K_n = e^{-\beta E(n)}$. Encore une fois, il est possible de procéder par récurrence :

$$\frac{\partial Z}{\partial K_n} = \frac{\partial A_N}{\partial K_n} \times \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix}, \frac{\partial A_n}{\partial K_n} = \frac{\partial A_{n-1}}{\partial K_n} \times Q_n + A_{n-1} \times \frac{\partial Q_n}{K_n} \tag{3.27}$$

L'avantage de cette méthode est qu'elle permet d'introduire des états supplémentaires correspondant à la liaison avec d'autres protéines, ou bien à l'interaction avec des nucléosomes voisins (Solis et al., 2004, 2007). Le désavantage est qu'elle oblige à calculer la fonction de partition totalement, imposant ainsi une lourdeur numérique assez conséquente.

Maintenant que nous avons déterminé théoriquement la densité de particules dans un profil inhomogène, on peut s'intéresser à quelques cas particuliers, et préciser les effets rencontrés fréquemment dans les systèmes de tiges rigides.

3.2 EXEMPLES SIMPLES

Dans les cas simples, la forme de la densité peut être interprétée facilement. Lorsque le potentiel est homogène, à proximité d'un ou deux murs infinis, les situations sont certes différentes, mais elles restent faciles à interpréter.

In simple situation, the shape of the density profiles can easily be interpreted : when the energetic field is homogeneous, near an infinite wall, or constrained between two walls.

3.2.1 Barrière verticale

Dans le cas simple d'une barrière infinie (figure 3.4 (a)), juxtaposée à un potentiel strictement plat, il apparaît des oscillations caractéristiques d'autant plus prononcées que le potentiel chimique est élevé. Ces oscillations émergent naturellement et reflètent simplement le fait que la pression exercée par les nombreux nucléosomes situés sur la ligne impose à tout nucléosome situé à proximité de la barrière de rester "collé" contre cette barrière. Le deuxième nucléosome voit une situation sensiblement similaire dans le sens où le premier nucléosome constitue pour lui un mur infranchissable, à ceci près que la position de ce mur est très légèrement variable (en fait cela dépend de la pression), par conséquent, sa position est elle aussi imposée avec une contrainte légèrement moins forte que pour le premier nucléosome. Le phénomène continue pour les nucléosomes suivant, créant les oscillations caractéristiques dans le profil de positionnement.

3.2.2 Puits

Dans le cas très légèrement plus raffiné du puits de potentiel (figure 3.4 (b)), c'est-à-dire de deux barrières infinies juxtaposées, on observe un phénomène de cristallisation, à mesure que le potentiel chimique augmente. La situation est tout à fait similaire au rangement de billes dans un tube ou de balles de tennis dans un étui. Lorsqu'une seule balle est présente dans le tube, elle est libre de se déplacer, la probabilité de la trouver est strictement la même partout. Plus on rajoute de balles, plus le choix devient restreint pour les positions prises par une balle en particulier, jusqu'à ce que le tube soit plein, et que les seules positions acceptables que peut occuper une balle données sont distribuées très finement autour de $i \cdot L/N$ (i le numéro de la balle, N le nombre de balles et L la taille du tube). Rajouter des balles correspond ici à augmenter le potentiel chimique.

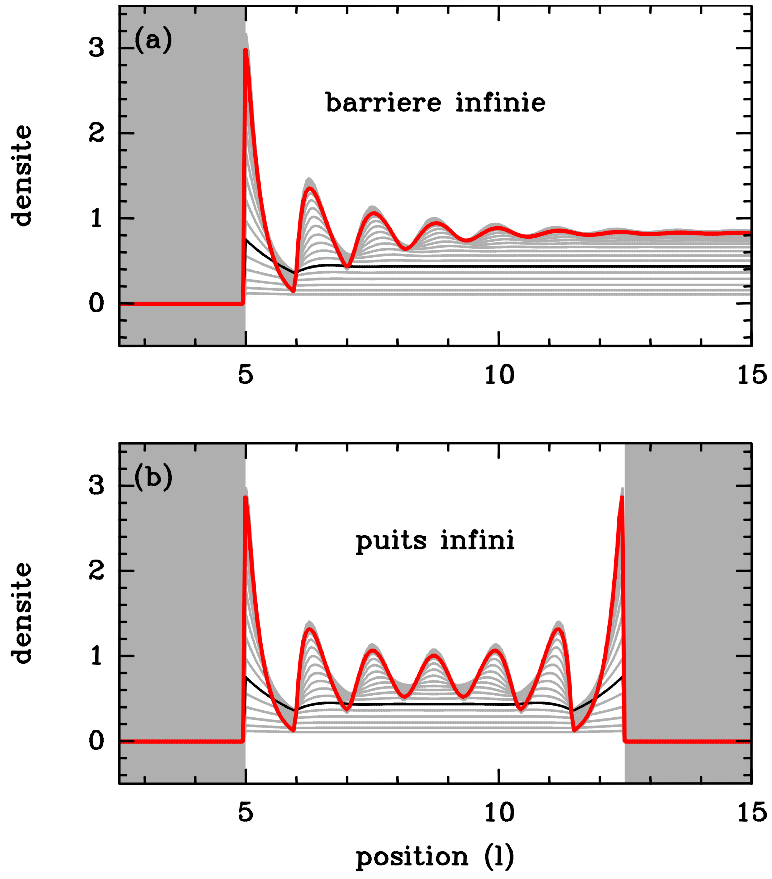


FIGURE 3.4 : *Cas simples de résolution de la densité de tiges rigides au voisinage d'une barrière infinie (a) et dans un puits infini (b) pour des potentiels chimiques faible (noir, $\mu = -5$ kT) ou élevé (rouge, $\mu = +5$ kT). La densité représentée sur l'axe des ordonnées est normalisée par la taille de la particule,* i.e. $\rho_{\text{norm}} = \rho \cdot l$

Clone	$\Delta\Delta G_D^o$ (kcal mol^{-1})		$\Delta\Delta G_E^o$ (kcal mol^{-1})	
607	-2.89 ± 0.44	(n=5)	-1.26 ± 0.37	(n=4)
611	-2.83 ± 0.22	(n=6)	-1.37 ± 0.23	(n=4)
623	-2.75 ± 0.28	(n=4)	-1.30 ± 0.28	(n=4)
601	-2.74 ± 0.33	(n=6)	-1.29 ± 0.29	(n=4)
612	-2.70 ± 0.46	(n=6)	-1.41 ± 0.12	(n=4)
626	-2.60 ± 0.25	(n=6)	-1.19 ± 0.19	(n=4)
603	-2.46 ± 0.37	(n=4)	-1.22 ± 0.39	(n=4)
618	-2.15 ± 0.26	(n=6)	-1.22 ± 0.28	(n=4)
TATA	-1.82 ± 0.29	(n=6)	-0.78 ± 0.46	(n=4)
CAG	-0.78 ± 0.15	(n=6)	-0.36 ± 0.46	(n=4)
NoSecs	-0.37 ± 0.11	(n=6)	0.27 ± 0.47	(n=4)
L. variegatus 5 S RNA gene	(0.00)	(n=6)	(0.00)	(n=4)
BadSecs	0.29 ± 0.11	(n=6)	0.07 ± 0.41	(n=4)
CA	0.31 ± 0.19	(n=6)	0.32 ± 0.26	(n=4)
Mouse Minor satellite	0.35 ± 0.03	(n=3)	(n.d.)	
TGA	1.15 ± 0.28	(n=6)	0.62 ± 0.51	(n=4)
TGGA	1.22 ± 0.24	(n=6)	0.97 ± 0.36	(n=4)

FIGURE 4.1 : *Energies libres d'association par rapport à la séquence 5S qui est incluse dans toutes les expériences et a donc une différence d'énergie libre de* 0*. ce choix de référence n'affecte pas l'ordre ni la différences entre les autres séquences.* $\Delta\Delta G_D^o$ *est l'énergie libre mesurée par la méthode de dialyse ;* $\Delta\Delta G_E^o$*", par la méthode "d' échange"(H.R. Widlund, H. Cao, S. Simonsson, E. Magnusson, T. Simonsson, P.E. Nielsen, J.D. Kahn, D.M. Crothers and M. Kubista, Identification and characterization of genomic nucleosome-positioning sequences.* J. Mol. Biol. **267** *(1997), pp. 807-817.). Les valeurs correspondent à la moyenne* $\pm$ *une déviation standard (*n*, nombre d'expériences ; n.d., non déterminée). La séquence TATA est la séquence naturelle la plus favorable mais son affinité est beacoup plus faible que la plupart des séquences non naturelles. Extrait de (Thaström* et al.*, 1999)*

4 POSITIONNEMENT "INTRINSÈQUE" : EFFETS DE SÉQUENCE

4.1 SPÉCIFICITÉ DE SÉQUENCE : EXPÉRIENCES IN VITRO

Pour quantifier la spécificité de séquence de l'interaction ADN-histones il faut recourir à des reconstitutions (assemblages) *in vitro* de nucléosomes. L'idée ensuite est de mesurer une spécificité relative, à savoir dans quelle mesure une séquence est elle plus affine qu'une autre ? Ou une séquence est elle plus réfractaire que d'autre à son enroulement autour de l'octamère ? La méthode couramment utilisée est celle mise en place par Shrader and Crothers (T.E. Shrader and D.M. Crothers, Artificial nucleosome positioning sequences.*Proc. Natl. Acad. Sci. USA* **86** 19 (1989), pp. 7418-7422, Effects of DNA sequence and histone-histone interactions on nucleosome placement. *J. Mol. Biol.* **216** 1 (1990), pp. 69-84.) qui mesure des affinités relatives de l'interaction octamère-ADN pour différents fragments d'ADN, et non l'affinité absolue.

A partir de ce type d'expériences il a été possible d'établir des spécifités relatives pour un certains nombre de séquences naturelles et artificielles comme celles reportées dans le tableau 4.1 où la référence est la séquence $5S$. Ces résultats indiquent en particulier que la séquence 601 (dans ces conditions de reconstitutions) a une affinité mille fois supérieure qu'une séquence aléatoire. Cependant, comme le montrent Lowary et Widom par ce même type d'expériences, $> 95\%$ de l'ADN génomique a une énergie libre qui ne diffère que marginalement ($0 \pm 0.3 kcal.mol^{-1}$, soit $0.6\ kT$) d'une séquence ADN synthétique aléatoire (Lowary and Widom, 1997).

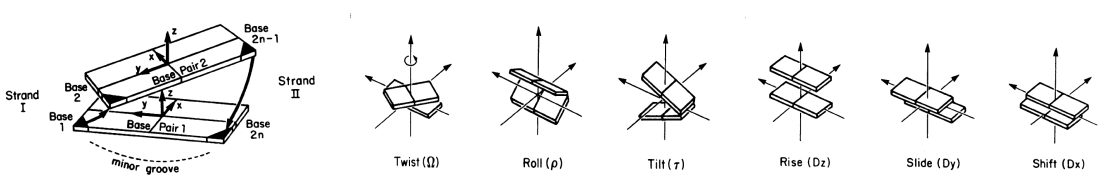


FIGURE 4.2 : *Paramètres hélicoidaux décrivant le configuration locale des paires de bases successives, définis par le groupe de travail "DNA Curvature and Bending" organisé par l'EMBO, Cambridge, 1998*

4.2 INFLUENCE DE LA SÉQUENCE SUR LES PROPRIÉTÉS ÉLASTIQUES DES CHAINES ADN

L'analyse structurale de la double hélice autour de l'octamère d'histone suggère que la formation du nuclésome induit des déformations assez importantes de la double hélice. L'ADN étant un hétéropolymère une des contribution à la spécificité de séquence (variabilité vis à vis de la séquence) de l'interaction ADN-histone est celle, "indirecte", provenant des propriétés élastiques de la séquence qui défavoriserait plus ou moins cette déformation.

Elasticité de chaines ADN :

Le modèle standard est celui décrivant la double hélice comme un empilement hélicoidal de paires de bases, prises comme objet "rigides". La conformation de la double hélice est donnée localement par les valeurs des paramètres hélicoidaux definissant les orientations et positions relatives des paires de bases successives. Si on considère le modèle simplifié où la paire de bases est considérée rigide (il n'y a pas de déformation interne) alors les paramètres hélicoidaux sont au nombre de 6 (Figure 4.2) :

Trois dégrés de libertés de courbure :

1. Le "roll" ρ définit la rotation relative des paires de bases selon la direction (et sens) sillon majeur $\rightarrow$ sillon mineur (Un roll positif correspond a une compression du sillon majeur).
2. Le "tilt" τ définit quant à lui la rotation relative des paires de base selon la direction tranverse à la chaine sucre-phosphate.
3. Le "twist" Ω definit la rotation autour de l'axe longitudinal.

Trois degrés de liberté de translation :

1. Le "Rise" Dz définit la distance séparant les paires de bases selon l'axe de l'hélice.
2. Le "Shift" Dx définit le décalage entre les paires de bases selon l'axe sillon majeur $\rightarrow$ sillon mineur.
3. Le "Slide"Dy définit le décalage entre les paires de bases selon l'axe transverse à la chaine sucre-phosphate.

Dans un modèle élastique linéaire, le coût énergétique associé à une déformation par rapport à une conformation au repos est quadratique en cette déformation à savoir localement :

$$\begin{aligned}
\delta E_{el} &= \frac{1}{2}\left(\Gamma-\Gamma_0\right)^T \mathbf{A}\left(\Gamma-\Gamma_0\right) \\
&= \frac{A_1}{2}(\rho-\rho_o)^2+\frac{A_2}{2}(\tau-\tau_o)^2+\frac{A_3}{2}(\Omega-\Omega_o)^2 \\
&+ D_1(Dx-Dx_o)^2+D_2(Dy-Dy_o)^2+D_3(Dz-Dz_o)^2+...
\end{aligned}$$

où $\Gamma=(\rho,\tau,\Omega,Dz,Dx,Dy)$ et $\Gamma_o=(\rho_o,\tau_o,\Omega_o,Dz_o,Dx_o,Dy_o)$ décrivent les configurations locales (du dinucléotide) dans l'état déformé et au repos respectivement, et $\mathbf{A}$ est la matrice des coefficients d'élasticité. Notons ici que plus une constante élastique est grande plus le coût elastique associée à la déformation associée est important, donc plus il est difficile de déformer, donc plus l'objet est (pour cette déformation) rigide. Si on considère la forme canonique de l'ADN B (Fig. 4.3) telle qu'elle est définie

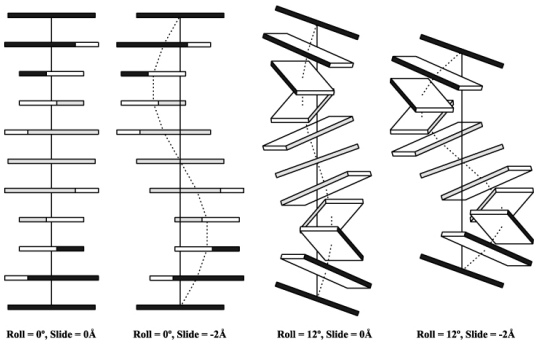


FIGURE 4.3 : *Exemples de configurations "régulières" de la doucle hélice vue comme un empilement hélicoïdal de paires de bases rigides (a) Cas ou seul le twist et le rise sont non nuls :* $\Omega \sim 34.3\,\mathrm{deg}$, $Dz = 0.34nm$ *correspondant à la forme canonique de l'ADN-B. (b) Cas correspondant à un slide constant négatif* $Dy = -2$ *(c) Désormais la double hélice présente une courbure locale : le roll est non nul,* $\rho = -12\,\mathrm{deg}$ *(d) le slide et le roll sont non nuls*

classiquement au repos, on a $\Omega_o = 34.3\,\mathrm{deg}/pb$ qui représente l'hélicité naturelle bien connue de la double hélice, et $Dz_o = 0.34\ nm$ qui représente l'écart entre paires de bases successives, et en général, un roll ρ_o et tilt τ_o nuls. En fait il est désormais admis que les propriétés élastiques locales, donc **A** et Γ_o dépendent de la séquence et que cette dépendance joue un rôle important dans les processus de reconnaissance et d'interaction entre complexes protéiques et l'ADN (Crothers, 1998) : les chaines ADN présentent, en particulier, des courbure intrinsèques (ρ_o, τ_o) non nulles et des flexibilités associées (A_1, A_2) qui dépendent de la séquence. Depuis trente ans, un grand nombre d'études expérimentales ont été mises en oeuvre pour rendre compte et quantifier cet effet de séquence, notamment dans le cadre du modèle introduit plus haut. L'électrophorèse sur gel polyacrilamide fut la première technique utilisées pour étudier la courbure et flexibilite de fragments d'ADN. Certains fragments révélaient en effet des migrations anormales qu'on imputait au fait que ces sequences etait naturellement courbe (Marini *et al.*, 1982 ; Wu and Crothers, 1984 ; Koo *et al.*, 1986 ; Hagerman, 1985, 1986 ; Diekmann, 1987 ; Chastain II *et al.*, 1995 ; Chastain and Sinden, 1998). Plusieurs modèles ont été proposés pour relier ce retard a la courbure intrinsèque et la flexibilité (Trifonov, 1980 ; Lumpkin *et al.*, 1985 ; De santis *et al.*, 1986, 1988 ; Trifonov *et al.*, 1987 ; Koo and Crothers, 1988 ; Levene and Zimm, 1989 ; Bolshoy *et al.*, 1991 ; Olson and Zhurkin, 1996).

Dans le cas où on néglige les couplages et les déformations des paramètres de translation ($D_{z,x,y}$), le coût élastique est donnée par la forme plus simple :

$$\frac{\delta E}{k_b T} = \frac{A_1}{2}(\rho - \rho_o)^2 + \frac{A_2}{2}(\tau - \tau_o)^2 + \frac{A_3}{2}(\Omega - \Omega_o)^2 \tag{4.1}$$

où A_1, A_2, A_3 sont les constantes d'élasticité associées à chacun des angles de déformation. Sous l'effet de l'agitation thermique l'ADN fluctue et ses fluctuations sont caractérisées par des distributions statistiques, en l'occurence ici gaussiennes des paramètres (ou "degrés de libertés") hélicoidaux :

$$\begin{aligned} P(\rho) &= \frac{A_1}{\sqrt{2\pi}} \exp\left(-\frac{A_1}{2}(\rho - \rho_o)^2\right) \\ P(\tau) &= \frac{A_2}{\sqrt{2\pi}} \exp\left(-\frac{A_2}{2}(\tau - \tau_o)^2\right) \\ P(\Omega) &= \frac{A_2}{\sqrt{2\pi}} \exp\left(-\frac{A_3}{2}(\Omega - \Omega_o)^2\right). \end{aligned}$$

Les courbures intrinsèques ρ_o, τ_o et Ω_o correspondent ainsi aux valeurs thermodynamiques "moyennes" ($\rho_o =< \rho >= \int P(\rho) d\rho$) où $< . >$ indique la moyenne thermodynamique) de la chaine libre et les fluctuations autour des ces valeurs d'équilibres sont d'amplitudes $< (\tau - \tau_o)^2 >= 1/A_1$, $< (\rho - \rho_o)^2 >= 1/A_2$ et $< (\Omega - \Omega_o)^2 >= 1/A_3$: donc plus la rigidité A_i est grande plus les fluctuations sont faibles. C'est dans le cadre de ce modèle gaussien, en evaluant à la fois les valeurs moyennes et les fluctuations qu'il est en effet possible d'extraire les paramètres élastiques en fonction de la séquence, expérimentalement (Scipioni et al., 2002b; Olson et al., 1998) et in silico (Lankas (Lankas et al., 2003)).

Periodicité , Corrélations à longue portée dans les séquences : interprétations structurales

L'analyse des séquences induisant cette courbure a révélé une périodicité à 10-11 bp (citations ?). Dans le cadre d'un modèle géométrique, on voit en effet qu'une telle periodicité dans la distribution du roll induit naturellement une courbure "mesosocopique" plane :

Par exemple une sequence $i = 1, ..146$ telle que :

$$\begin{aligned} \rho_0(i) &= \kappa_0 \cos(\omega i) \\ \tau_0(i) &= \kappa_0 \sin(\omega i) \\ \Omega_0(i) &= \Omega_0 \end{aligned}$$

forme un solénoide régulier (cf Fig. 4.20) de courbure géométrique κ_0 de torsion géométrique $t = \omega - \Omega_0$ qui est donc nulle pour $\omega = \Omega_0 \sim 2\pi/10.5bp$.

La présence de periodicités à $10\ pb$ dans la séquence est donc susceptible de favoriser le positionnement d'un nucléosome en induisant une courbure naturelle ou une flexibilité anisotrope (cf plus bas). On peut se demander quelle est la part de cette périodicité à l'échelle génomique ?

Considérons ainsi le profil de courbure "intrinsèque" locale (roll $\rho_0(i)$) obtenu à partir de la table "PNuc"(Vaillant, 2001) le long de différents génomes, par exemple celui de la levure *S. cerevisiae* (Fig. 4.4(a)) et celui de *E. Coli* (Fig. 4.4(b)). L'apparence des ces profils est plutôt bruitée et déjà à l'oeil on peut discerner des différences entre les deux expèces. Pour s'en convaincre on peut procéder à leur étude spectrale, par exemple calculer leuron spectre de puissance $(S(f)^2$ où $S(f)$ est la transformée de fourier du signal). Dans un signal périodique, de période T_o, donc de fréquence $f_o = 1/T_o$, le spectre de puisance se concentre en un seul pic à $f = f_o$. Dans un signal bruité dit "blanc", à savoir un signal aléatoire sans corrélation entre les valeurs successives, le spectre est plat indiquant qu'un tel signal ne présente aucune échelle caractéristique. Il y a de manière plus générale, les signaux bruités dit invariants d'échelle et monofractals (Audit, 1999), les bruits en "$1/f$" qui ont un spectre de puissance en loi de puissance $S(f)^2 = (1/f)^\nu$; $\nu = 2H - 1$, avec H le coefficient de Hurst caractérisant les propriété d'invariance d'échelle du signal. Dans ce cas la fonction d'autocorrélation du signal n'est plus "nulle" à grande distance mais est finie et décroît très lentement en loi de puissance réflétant une persistence dans la distribution des valeurs du signal. Lorsque $H = 1/2$,on est dans le cas non corrélé, bruit "blanc" ; lorsque $1/2 < H < 1$ on est dans le cas corrélé à longue portée, persistent. Les séquences ADN et les profils structuraux tels que le roll "Pnuc" présentent des corrélations à longues portée. L'analyse de ces corrélations via une méthode d'analyse en ondelette fut le sujet principal de la thèse de Benjamin Audit(Audit, 1999) et d'une partie de la mienne (Vaillant, 2001). A partir de l' analyse d'un grand nombre de génomes d'espèces (mon début de thèse a coincidé avec l'obtention de la séquence de la levure *S. Cerevisiae*) des différents règnes (eubactéries, eucaryotes et virus) nous avons montré que les eucaryotes et certains virus (retrovirus, ceux qui s'insère dans le génome "hôte") se distinguaient des eubactéries par la présence de corrélation à longue portée "à petite échelle" (entre 20 et $150 - 200$ pb) avec $H = 0.54 - 0.65$ (cela veut dire que le spectre de puissance est en loi de puissance $S(f)^2 = (1/f)^\nu$ ($\nu = 2H - 1$) dans la gamme de fréquence $[1/200 - 1/20]$). Toutes les génomes (ou presque) présentaient par contre des corrélations à longue portée au delà de cette échelle caractéristique avec $H = 0.8$. Nous avions fait alors la conjecture que ces corrélations à longue portée avaient une origine structurale en lien avec l'organisation des génomes : avec le nucléosome donc pour les eucaryotes à petite échelle (l'échelle caractéristique étant proche de celle d'un nucléosome) ; avec des organisations supérieurs (fibres de 30 nm chez les eucaryotes ou domaines de surenroulement chez les eubactéries) à grande échelle. La seconde partie de ma tèse a donc été consacrée a essayer de voir quel pouvait être l'effet de telles corrélations à longue portée sur l'élasticité de longues chaîne ADN. La persistence par exemple dans la distribution du roll intrinsèque pourrait elle favoriser (statistiquement) la compaction de la chaîne ? (On serait dans même registre structural que la périodicité qui correspondrait à une "persistence" stricte, déterministe). L'étude spectrale de la distribution de la courbure intrinsèque telle que mesurée par le codage "Pnuc" pour différents organismes est reportée à la figure 4.5. On vérifie bien la présence de corrélations à longue portée à petite échelle (hautes fréquences $[1/200 - 1/20]$) pour les eucaryotes, son absence chez E. Coli et la présence de corrélations plus persistentes ($H = 0.8$) à plus grande échelle (basse fréquence). A titre compartatif, le spectre d'une séquence aléatoire ($H = 0.5$) à toute échelle est reportée en Figure 4.5(a).

On remarque aussi que la contribution de péridocité à $10.2pb$ et qui se manifeste par un pic à la fréquence $1/10.2$ est, il est vraie significative, (sauf pour S. Pombe, ou elle est totalement absente...) mais très faible ($< 5\%$ de la contribution totale au spectre) mis à part chez *C. Elegans*. Chez *E. Coli* la périodicité

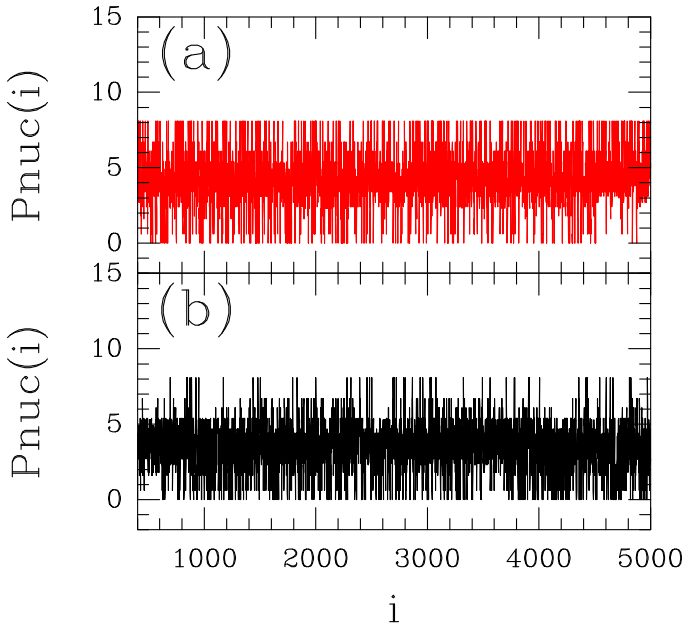


FIGURE 4.4 : *Profils de courbure intrinsèque "Pnuc" (roll ρ_o, mesuré en degré) (Satchwell* et al.*, 1986; Goodsell and Dickerson, 1994; Vaillant, 2001) le long de contigs de* $5000\ pb$ *(a) de la levure* S. Cerevisiae *et (b) de* E. Coli.

contribue également très peu à l'organisation du génome et se situe plutôt à $11.2pb$ (Fig. 4.5).

N.B. : Dans tous les spectres mis à part tout à fait logiquement dans le cas purement aléatoire (Fig. 4.5(a)), le pic à $3pb$ reflète l'organisation en codons des séquences codantes.

De cette petite étude on e déduit que la périodicité à $10.2pb$ dans les séquences génomique reste marginale et ne peut donc intervenir dans le poitionnement des nucléosomes que de façon marginale.

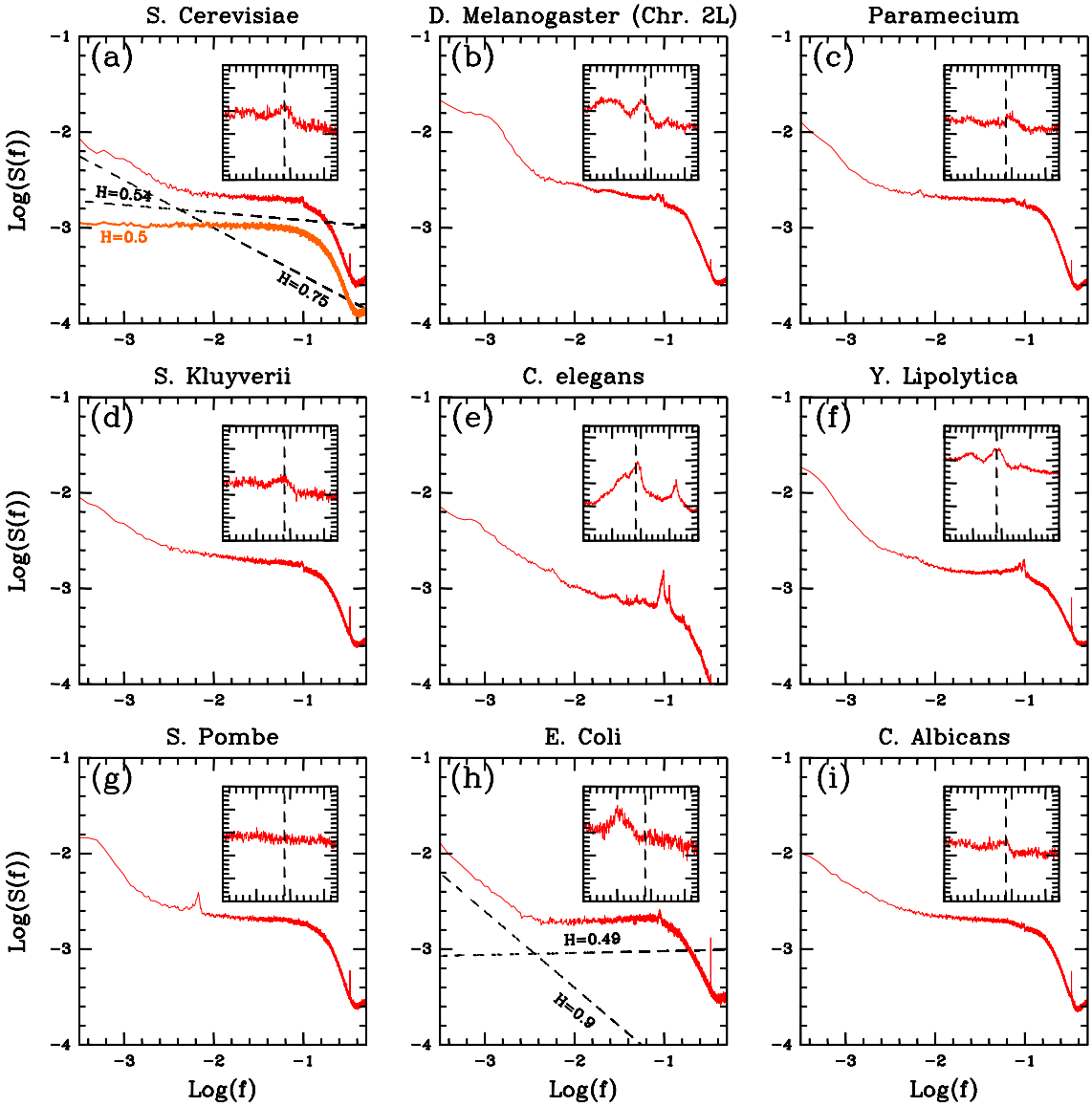


FIGURE 4.5 : *Analyse spectrale des profils de courbure intrinsèque "Pnuc" pour différents organismes. Spectre de puissance $S^2(f)$ en représentation log-log. En (a) et (h) les tirets correspondnet aux comportements linéaires associés aux petites et grandes échelles avec les valeurs de H coorespondantes (la pente est donnée par $\nu = 2H - 1$). Dans (a-i), en inserts est reportée un zoom sur la contribution spectrale voisine de la période* $10.2\ pb$.

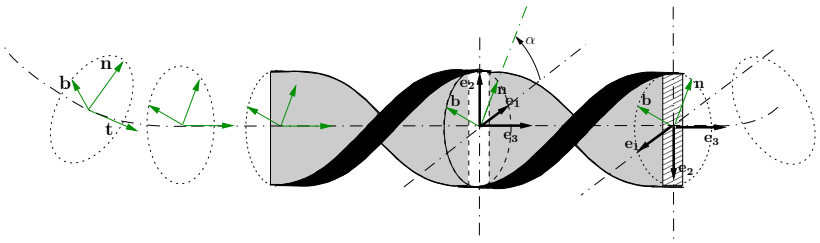


FIGURE 4.6 : *Représentation schématique de la double hélice vue comme un ruban naturellement torsadé (Vaillant, 2001).*

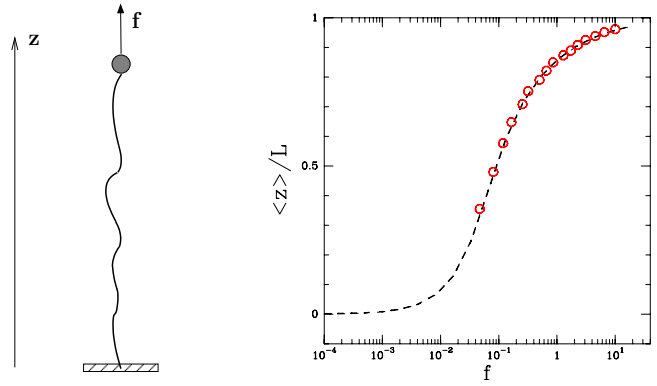


FIGURE 4.7 : *Expérience d'élongation de chaînes ADN par pincette magnétique. (gauche) Principe : la chaine est accorché à la surface et à une bill paramagnétique à son autre extrémité. Un champ magnétique permet de manipuler la bille, à savoir exercer une force et/ou imposer une torsion. (droite) Courbe d'extension* vs *force exercée, obtenue sur l'ADN du phage λ ; données expérimentales (cercles rouges), prédiction théorique à partir du modèle du ver (Eq. 4.2) avec $A = 51\ nm$ (courbe en tirets).*

L'objet ici est d'étudier dans quelle mesure une distribution désordonnée de courbure ou/et flexibilité intrinsèque influence les propriétés de conformation d'une chaine ADN. Et en particulier dans quelle mesure un désordre "corrélé" à longue portée affecte ces propriétés structurales. On s'intéresse à une influence "statistique", donc influnece à l'échelle d'un génome, tout au moins d'un grand nombre de séquence ou d'une très grande séquence dont les propriétés d'organisation (composition en motifs, distribution de ces motifs) (mises à part les cas des séquences répétées périodiquement sur de très longues distances) sont définies "statistiquement" (par des lois probabilistes).

1. **Longues chaines faiblement contraintes :**
   Pour commencer à comprendre ce lien entre chromatine et correlations à longue portée, nous avons, dans un premier temps, étudié comment la séquence via ses propriétés structurales particulières influence les propriétés de conformation des chaines ADN à grande échelle. Dans les années 90, plusieurs expériences de micromanipulation de chaines ADN ont permis la mesure du comportement de telles chaines soumises soit à une élongation simple (Fig. 4.7) soit à une élongation couplée à une torsion (Smith et al., 1992; Perkins et al., 1994; Cluzel et al., 1996; Smith et al., 1996; Strick et al., 1996; Wang et al., 1997) and twisting (Strick et al., 1996; Allemand et al., 1998; Léger et al., 1999).
   Il se trouve que dans une certaine gamme de forces et de surenroulement, les résultats obtenus peuvent être déduits, avec excellente précision, à partir d'un modèle minimal de corde élastique homogène et isotrope défini par(Kratky and Porod, 1949; Schellman, 1974; Bustamante et al., 1994; Vologoskii, 1994; Marko and Siggia, 1995a; Bouchiat et al., 1999; Grossberg and Khoklov, 1994) :

$$\frac{\delta E}{k_b T} = \frac{A_{eff}}{2}(\rho^2 + \tau^2) + \frac{A_{3eff}}{2}(\Omega)^2 \tag{4.2}$$

   Par rapport au modèle plus général de corde élastque, ce modèle correspond donc au cas où $\rho_o = \tau_o = 0$ et $A_1 = A_2 = A_{eff}$. La composante courbure peut s'écrire simplement à partir de la

courbure de l'axe de la corde définie par $\kappa = ||\frac{d\vec{t}}{ds}|| = \sqrt{\rho^2 + \tau^2}$ où $\vec{t}$ designe l'orientation de l'axe (Fig. 4.6) :

$$\frac{E_{courbure}}{k_b T} = A_{eff} \left(\frac{d\vec{t}}{ds} \right)^2 ds \tag{4.3}$$

C'est le modèle le plus simple dit modèle du ver qui caractérise les propriétés de conformation de polymère dit "semi-flexibles" (les polymères "flexibles" ne présentent pas d'élasticité enthalpique et seule la contribution entropique intervient).

Les fluctuations de l'axe de la chaine sont ainsi contrôlées par la longueur de persistence de cette chaine ; cette longueur de persistence corerspond à la longueur l_p d'ADN au bout de laquelle l'orientation de l'axe se décorrèle (du fait de l'agitation thermique) :

$$\langle \vec{t}(s).\vec{t}(s') \rangle = \exp -|s' - s|/l_p \tag{4.4}$$

Généralement il est difficile d'avoir accès à cette fonction de corrélation et on mesure plus facilement le rayon de gyration de la chaine qui mesure le degre de compaction d'une chaine :

$$\langle R^2 \rangle = \int_0^L \int_0^L \langle \vec{t}(s).\vec{t}(s') \rangle ds ds' = 2l_p^2(\exp(-L/l_p) - 1 + L/l_p), \tag{4.5}$$

Pour de longues chaines on extrait la longueur de persistence par la relation asymptotique :

$$\ell_p = \lim_{L \to \infty} \langle R^2 \rangle / 2L. \tag{4.6}$$

Dans le cadre du modèle du ver 4.3 on montre qu'on a tout simplement : $l_p = A$. Plus la chaine est rigide plus la longueur de persistence est importante, et moins la chaine est compacte ($\langle R^2 \rangle$ grand). Dans les conditions physiologiques, la plupart des ecperiences donne une estimation comparable se $l_p = A_{eff} \simeq 50$ nm (~ 170 bp). Dans le cas ou il y a une contrainte de surenroulment, il faut prendre en compte la contribution de torsion dont C_{eff} contôle les fluctuation ; de la même manire que pour la courbure de l'axe, C_{eff} détermine la longueur de persistence associée à ces fluctuations le long de la chaines. Les estimations faites notamment à partir des courbes d'extension en fonction du surenroulement donne une valeur comprise entre $C_{eff} = 75\ nm$ et $110\ nm$ (Marko and Siggia, 1995b; Moroz and Nelson, 1997; Vologodskii and Marko, 1997; Bouchiat and Mezard, 1998, 2000).

Mais si on tient compte de la séquence génomique, on doit considérer (au minimum) le modèle semi-flexible inhomogène 4.1 avec donc des distributions de courbures et torsions intrinsèques ρ_o, τ_o et Ω_o. Ces distributions obtenues à partir de la sequence genomique via des tables structurales telles que celle "PNuc" (Satchwell et al., 1986; Goodsell and Dickerson, 1994), ou "Anselmi" (Scipioni et al., 2002b) sont notamment caractérisées par leur loi statistique (en général gaussienne) de variance σ_o^2 et leur propriétés spectrales (periodicités, corrélations).

On montre facilement (Vaillant, 2001; Vaillant et al., 2003) que la longueur de persistence est désormais fonction de ces courbures intrinsèques et dans le cas d'une distribution "désordonnée" qui nous intéresse on a la relation introduite par Trifonov (Trifonov et al., 1987) et en considérant par ailleurs le modèle isotrope $A_1 = A_2 = A$:

$$l_p = A \left(\frac{1}{1 + A\sigma_o^2} \right) = A_{eff} \tag{4.7}$$

Et donc que la chaine se comporte comme une chaine "idéale" sans courbure intrinsèque mais avec une rigidité "effective" A_{eff} plus petite que celle de la chaine "idéale" et de même "vraie" rigidité de courbure de l'axe A. Plus le désordre intrinsèque est fort (σ_o^2 grand), plus la longueur de persistence est faible, plus la chaine apparait compacte : à la compaction induite par les fluctuations thermiques s'ajoute la compaction "intrinsèque".

La modélisation des expériences d'extension de l'ADN du phage lambda en prenant désormais en compte le désordre de courbure intrinsèque indique que pour observer une rigidité apparente de $A_{eff} \sim 50\ nm$, la chaine doit présenter une rigidité "vraie" de l'ordre de $A = 60 - 80\ nm$. En fait tout dépend de la valeur de σ_o^2, qui selon les tables structurales varie entre 0.01 et 0.05 (Vaillant,

2001; Vaillant et al., 2003)... Ces valeurs sont en accord avec celles extraites d'expériences de molécules uniques pour des chaines artificielles intrinsèquement droites (donc avec $\sigma_o = 0$) (Song and Schurr, 1990; Bednar et al., 1995; Furrer et al., 1997). Nos travaux (théoriques) indiquent aussi que la conformation à 3D de chaines ADN faiblement contraintes depend essentiellemnt de la "force" du désordre intrinsèque et non des propriétes de corrélations à longue portée (Vaillant, 2001; Vaillant et al., 2003). Cela dit on est loin ici des contraintes mécaniques induites par l'enroulement autour du coeur d'histones...

2. **Petites chaines contraintes : adsorption sur une surface**

   Parmi les méthodes expérimentales de caractérisation de molécules uniques, parce qu'elle fournit une visualisation des conformations 2D des molécules d'ADN, l'AFM apparait comme une technique très performante, non seulement pour estimer des longueurs de persistence via la mesure de la distance bout-à-bouts (R), (Rivetti et al., 1998; Anselmi et al., 2005; Cognet et al., 1999; Lyubchenko et al., 1993; Hansma et al., 1996; Rivetti et al., 1996) mais aussi et surtout parce qu'elle permet d'avoir accès, après moyennation thermodynamique sur une population de ces conformations, aux courbures intrinsèques et flexibilités de la doubles hélice et ce avec une résolution de quelques pas d'hélices (Cognet et al., 1999; Zuccheri et al., 2001; Scipioni et al., 2002a,c; Marilley et al., 2005). L'analyse des données AFM d'ADN a été principalement faite dans le cadre du modèle du ver "idéal", donc décrivant le comportement de chaines semi-flexibles intrinsèquement droites (Rivetti et al., 1998; Anselmi et al., 2005; Cognet et al., 1999). Notre objectif a été d'utiliser la microscopie AFM dans l'air et en liquide pour mettre en évidence un effet des corrélations à longue portée (H) sur les conformations de molécules d'ADN nues déposés sur une surface de mica dans des conditions où elles sont contraintes à s'équilibrer à 2D (Moukhtar et al., 2007, 2010).

   Nous avons étudié les propriétés de conformation de chaines d'ADN (i) artificielle de $800\ pb$ intrinsèquement droites (construites à partir de la ligation de quatre chaines de $200\ pb$ intrinsèquement droite), (ii) de $2200\ pb$ du virus ARN de l'hépatite C, dont la séquence est décorrélée ($H = 0.5$), (iii) du génome humain, de tailles comprises entre 1000 et $3000\ pb$, dont les séquences présentent des corrélations à longue portée ($H = 0.73$) et (i) du génome humain de taille $2200pb$ correpondant à différente concentration en G+C (31%, 38%, 41% et 46%),

   A partir des images de molécules déposées sur la surface il est possible d'avoir accès à la longueur de persistence de la chaine, par exemple en mesurant le rayon de gyration R^2 pour chaque molecule isolée et en moyennant toutes les mesures. Dans le cadre du modèle du ver sans courbure intrinsèque et avec une rigidité de courbure A (qui correspond à la longueur de persistence à 3D), sa longueur de persistence lorsqu'elle est confinée à 2D est $l_p^{2D} = 2A$. Si on introduit une courbure intrinsèque, désordonnée et de variance σ_o^2, on obtient une relation de "renormalisation" qui dépend désormais explicitement des corrélations à longue portée (via H)(Moukhtar et al., 2009; Moukhtar, 2008) :

$$l_p^{2D} = 2A_{eff} = 2A\left(\frac{1}{1 + \frac{(2A)^{2H}\sigma_o^2}{2}\Gamma(2H+1)}\right) \tag{4.8}$$

   Pour une distribution de courbure intrinsèque non corrélée, $H = 1/2$ et on retrouve la formule de Trifonov avec $A_{eff} = A(\frac{1}{1+A\sigma_o^2})$: la chaine se comporte comme une chaine idéale mais avec une flexibilité apparente réduite ; si on introduit désormais des corrélations à longue portée dans la distribution des courbures intrinsèques, la chaine se comporte, de même, comme une chaine idéale avec une rigidité apparente qui est d'autant plus faible que H augmente. A 2D, comme illustré aux figures 4.8 et 4.9, la présence de corrélation à longue portée induit une compaction intrinsèque plus importante : en effet la persistence dans la valeur et le signe de la courbure intrinsèque fait que, à même force du désordre σ_o^2 (Fig. 4.8), la probabilité pour faire des boucles est plus importante dans le cas corrélé que dans le cas non corrélé (pas de persistence) (Fig. 4.8). Une chaine avec des corrélatiosn à longue portée est donc intrinsèquement plus compacte que chaine non corrélé et l'agitation thermique ne fait qu'accentuer cette différence (Fig. 4.9).

   Ces effets de séquences et de corrélations à longue portée sur la compaction des chaines à 2D sont en effet confirmées par les observations expérimentales. Les mesures indiquent que la chaine intrinsèquement droite présente une rigidité de $A = 70\ nm$ en accord avec les mesures faites par Bednar et al. (Bednar et al., 1995; Furrer et al., 1997). Pour la séquence non corrélée du VHC, on mesure une longueur de persistence de $50\ nm$ correspondant à une rigidité $A = 70\ nm$ et $\sigma_o = 0.02$. Pour les séquences humaines corrélées à longue portée, les résultats sont égalements

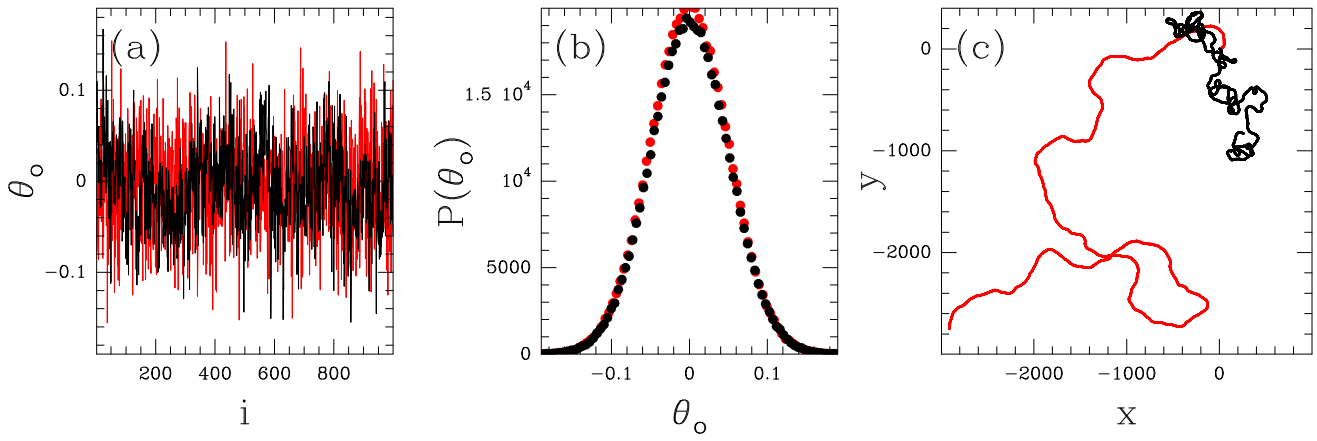


FIGURE 4.8 : *(a) Profils des courbures intrinsèques (2D, θ_o) d'une chaîne désordonnée non corrélée ($H = 1/2$, rouge) et corrélée à longue portée ($H = 0.8$, noir). (b) Histogramme des valeurs correspondant à une distribution gaussienen de même variance $\sigma_o = 0.05$ (fort désordre) (c) Trajectoires 2D intrinsèques (d'équilibre) construites à partir des profils de courbure représentés en (a).*

consistants avec le modèle du ver généralisé, pour $A = 70\ nm$, $\sigma_o = 0.02$ et $H = 0.6$ $l < 200\ pb$ et $H = 0.73$ $l > 200\ pb$. Les figures 4.10, 4.11 et 4.12 rassemblent quelques images représentatives des conformations pour chaque type de chaines ADN et illustrent ainsi comment le désordre et les corrélations peuvent induire des conformation plus compactes et courbes.

3. **Petites chaines contraintes : boucles à 2D**

   Nous venons de voir qu'à 2D, les corrélations à longue portée se manifestent par des courbures macroscopiques intrinsèques pouvant ainsi favoriser la formation de boucles. Pour modéliser justement la formation de boucles de taille ℓ au sein d'un chaine de plus grande de taille L et étudier les effets de correlations nous avons considéré des chaines sous les contraintes géométriques suivantes (Vaillant et al., 2005, 2006) :

   Contrainte d'enroulement : Cette contrainte (Fig. 4.13(a)) correspond à fixer la variation angulaire de l'axe de la chaine sur une distance (curviligne) ℓ :

   $$\int_s^{s+\ell} \dot{\theta}(u)du = \alpha. \tag{4.9}$$

   Contrainte de cyclisation : La contrainte de cyclisation (Fig. 4.13(b)) consiste à imposer que les deux extrémités de la chaîne se rejoignent au bout de la distance ℓ :

   $$\int_s^{s+\ell} \cos(\theta(u))du = \int_s^{s+\ell} \sin(\theta(u))du = 0. \tag{4.10}$$

   À la Fig.4.14(a) sont représentés les profils énergétiques pour $A = \tilde{A}/k_BT = 200\ bp$, $\ell = 200\ bp$, $\sigma_o = 0.01$ et $\alpha = 2\pi$, pour une chaine non corrélée ou avec CLP ; les fluctuations du profil pour cette dernière sont plus importantes. Comme l'indique la Fig. 4.14(c), à la limite des désordres faibles, à savoir quand $\tilde{A}^2\alpha^2\sigma_o^2\ell^{2H-2} \ll 1$ (voir Eq. (4.12)), les statistiques du profil énergétique sont gaussiennes et on obtient l'expression suivante pour la moyenne :

   $$\overline{E}(\ell) = \frac{\tilde{A}}{2}[\frac{\alpha^2}{\ell} + \sigma_o^2\ell^{2H-1}], \tag{4.11}$$

   et pour la variance :

   $$\Lambda(\ell) = \overline{(E(\ell) - \overline{E}(\ell))^2} = \tilde{A}^2\alpha^2\sigma_o^2\ell^{2H-2}. \tag{4.12}$$

   Comme l'indiquent les Figs.4.14(b) et 4.14(d), lorsque la force du désordre augmente ($\sigma_o = 0.05$), les distributions statistiques de l'energie commencent à présenter une queue exponentielle aux

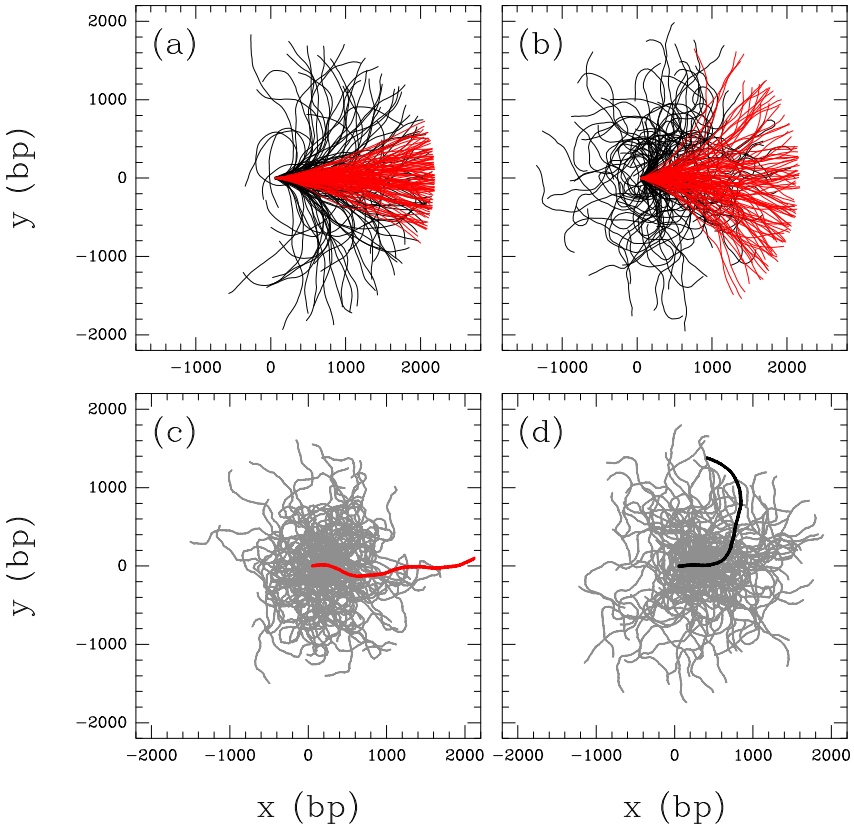


FIGURE 4.9 : *(a) $N = 100$ trajectoires "intrinsèques" de chaînes de taille $L = 2200$ bp construites numériquement avec une distribution de courbure intrinsèque non corrélée ($H = 1/2$, noir) et corrélée à longue portée ($H = 0.73$, rouge) d'amplitude ("force") $\sigma_o = 0.007$; (b) même chose qu'en (a) avec $\sigma_o = 0.022$. (c) simulations de $N = 100$ conformations 2D à l'équilibre (trajectoires grises) pour une chaine de taille $L = 2200$ pb, de flexibilité $2A = 490$ bp et avce une distribution de courbure intrinsèque non corrélée ($H = 1/2$) et d'amplitude $\sigma_o = 0.022$ (trajectoire intrinsèque, noire) ; (d) même chose qu'en (c) mais pour une chaîne avec une distribution de courbure intrinsèque corrélée ($H = 0.73$) et d'amplitude $\sigma_o = 0.007$ (trajectoire intrinsèque, rouge) Moukhtar* et al. *(2010).*

fortes valeurs ; plus H est grand, moins les statistiques sont gaussiennes et plus les barrières énergétiques sont importantes Vaillant et al. (2005, 2006).

A température finie, il faut considérer les effets des fluctuations thermiques et donc calculer l'energie libre associée à la formation d'une boucle Vaillant et al. (2005, 2006) :

$$\beta f(s,\ell) = \beta E(s,\ell) - \Delta S(s,\ell), \tag{4.13}$$

où $\beta = 1/k_B T$. Sous l'approximation d'harmonicité des petites fluctuations autour de la configuration d'équilibre mécanique (i.e. fluctuations gaussienes), le coût entropique peut être calculé analytiquement pour la contrainte d'enroulement :

$$\Delta S_w(s,\ell) = b_w - \frac{1}{2}\ln \ell, \tag{4.14}$$

où b_w est une constante indépendante de ℓ. Pour la contrainte de cyclisation, nous avons besoin d'avoir recours au calcul numérique ; on obtient un comportement logarithmique similaire mais avec un préfacteur différent :

$$\Delta S_c(s,\ell) = b_c - \frac{7}{2}\ln \ell, \tag{4.15}$$

où b_c est de nouveau indépendante de ℓ. Dans les deux cas, puisque $\Delta S(s,\ell)$ ne dépend pas de la courbure intrinsèque $\dot{\theta}_o(s)$, elle est indépendante de la position s le long de la chaine (ce qui ne serait plus la cas si on considérait une flexibilité $\tilde{A}(s)$ dépendant de la position).

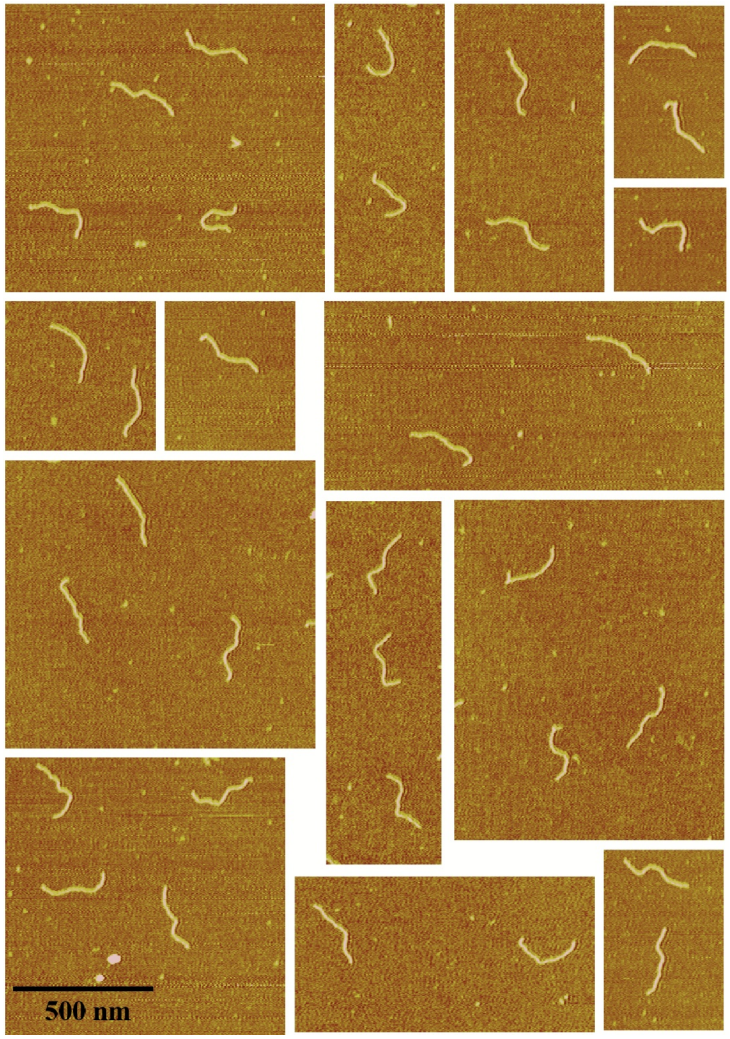


FIGURE 4.10 : *Images AFM (dans l'air) de molécules d'ADN artificielles de 800 pb, intrinsèquement droites (Arneodo* et al.*, 2008).*

Les propriétés statistiques de l'énergie libre se déduisent directement de celle de l'énergie mécanique. À partir des Eqs. (4.11), (4.12), (4.14) et (4.15), on obtient les expressions suivantes pour l'énergie libre moyenne dans la limite des faibles désordres Vaillant et al. (2005, 2006) :

$$\beta\overline{f}(\ell) = \frac{A}{2}[\frac{\alpha^2}{\ell} + \sigma_o^2\ell^{2H-1}] + c\ln\ell - b, \tag{4.16}$$

où $c = c_w = 1/2$ (resp. $c_c = 7/2$) pour la contrainte d'enroulement (resp. de cyclisation). Enfin, à un facteur multiplicatif près β^2, la variance des fluctuations de l'energie libre est donnée par la variance de celles de l'énergie (Eq. (4.12)) :

$$\Lambda(\ell) = \beta^2\overline{(f(\ell) - \overline{f}(\ell))^2} = A^2\alpha^2\sigma_o^2\ell^{2H-2}. \tag{4.17}$$

Les propriétés thermodynamiques d'une boucle 2D de taille ℓ se formant au sein d'une chaine désordonnée de longueur L sont décrites par la fonction de partition Vaillant et al. (2005, 2006) :

$$Z(\ell, L) = \int_0^{L-\ell} \exp\left[-\beta f(s,\ell)\right]ds, \tag{4.18}$$

qui tient compte de toutes les positions possibles de la boucle le long de la chaine. L'énergie libre du système (relativement à celle $\beta F_o = \frac{L}{2}\ln(A/2\pi)$ de la chaine sans boucle) est donnée par :

$$\beta F(\ell, L) = -\ln(Z(\ell, L)). \tag{4.19}$$

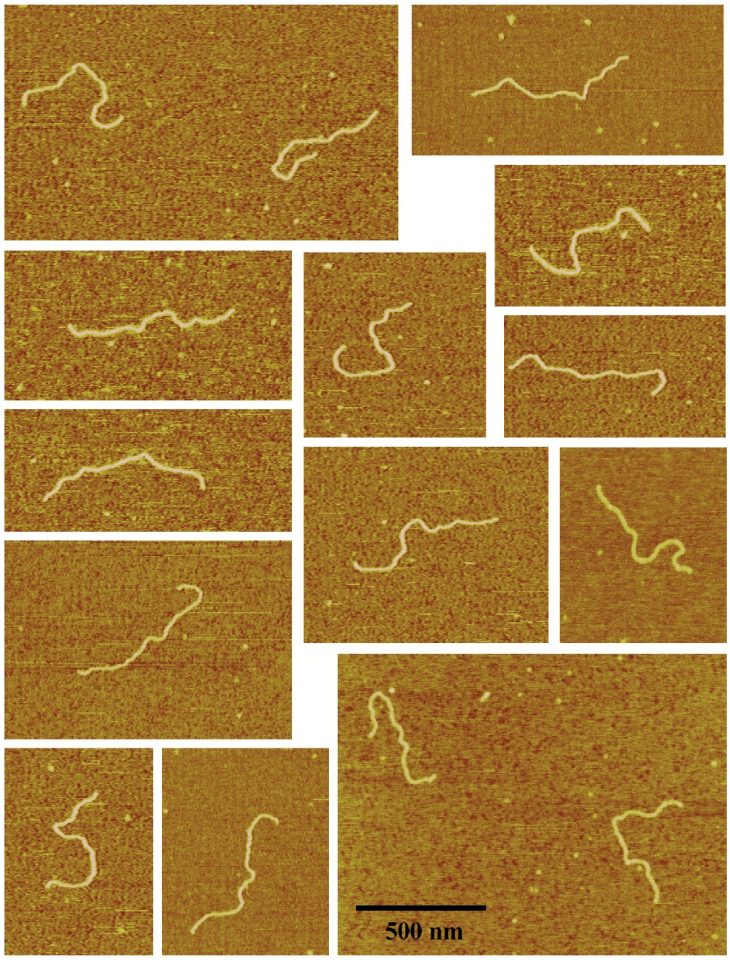


FIGURE 4.11 : *Images AFM (dans l'air) de chaines ADN de $L = 2200$ bp extraite du VHC, et de séquence non corrélée ($H = 0.5$) (*Arneodo et al.*, 2008).*

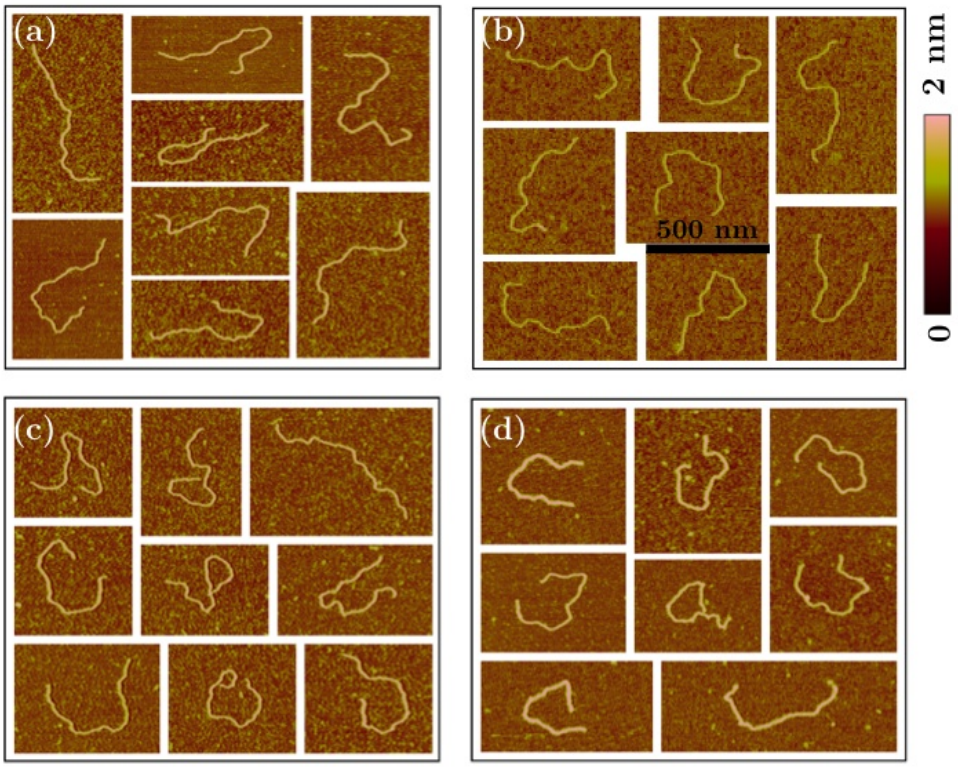


FIGURE 4.12 : *Images AFM de fragments ADN du génome humain à différents contenus en GC (Moukhtar* et al.*, 2010). (a) chr 21 :* $L = 2200$ *bp, G+C=31% ; (b) chr21 :* $L = 2006$ *bp, G+C=38% ; (c) chr 21 :* $L = 2206$ *bp, G+C=43% ; (d) chr 8 :* $L = 2189$ *bp, G+C=46%.*

Pour un désordre faible, cette énergie libre vaut en moyenne (Vaillant et al., 2006) :

$$\begin{aligned}\beta\overline{F}(\ell, L) =& \frac{A\alpha^2}{2\ell} + \frac{A\sigma_o^2}{2}\ell^{2H-1} - \frac{A^2\alpha^2}{2}\sigma_o^2\ell^{2H-2} \\ &+ c\ln\ell - b - \ln(L-\ell).\end{aligned} \tag{4.20}$$

et donc décroît lorsque H augmente.

- En absence de désordre ($\sigma_o = 0$), le système "idéal" a une énergie libre qui présente un minimum pour une taille $\ell^* = \alpha^2 A/2c$. Cette taille optimale (~ 1100 *bp* pour les paramètres utilisés à la figure Fig. 4.15, sépare le domaine enthalpique à petite échelle ℓ caractérisé par une décroissance rapide de l'énergie libre et le domaine entropique à plus grande échelle caracterisé par une croissance faible, logarithmique.
- Lorsqu'on ajoute du désordre intrinsèque non corrélé ($H = 1/2$), la dépendance en ℓ de l'énergie libre se réduit à celle du cas homogène avec une flexibilté de courbure effective renormalisée $A_{eff} = A(1 - A\sigma_o^2)$, en accord avec la version "faible désordre" de la relation de Trifonov *et al.* (Trifonov et al., 1987; Nelson, 1998). Ainsi, il n'y pas de différence qualitative entre une chaîne désordonné non corrélée et une chaine idéale, mais introduire un désordre diminue l'énergie libre du système et favorise la formation de boucles de petites tailles $\ell^*(H = 1/2) = \alpha^2 A_{eff}/2c$.
- Lorsqu'on introduit un désordre corrélé à longue porté (CLP) le système ne se comporte pas comme un système homogène et de façon importante, pour les boucles de petites tailles ($\lesssim 1000$ *pb*), l'énergie libre decroît quand H augmente. Colle l'indique la Fig. 4.15 l'énergie libre du système présente un minimum bien défini dont la valeur est fonction de la taille ℓ, et de la présence ou non de corrélation à longue portée.

4. **Minicercle : effets de séquences**

Dans les complexes nucléo-protéïques, la contribution de la séquence dans l'affinité d'association est souvent indirecte, par opposition à une reconnaissance directe où les bases interagissent

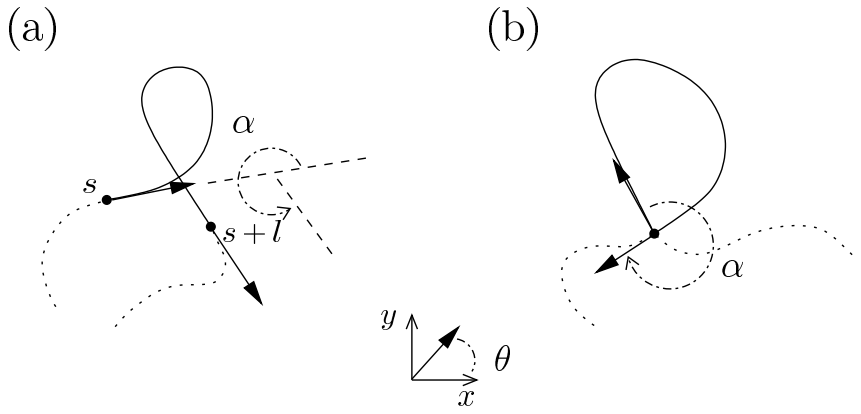


FIGURE 4.13 : *Contraintes géométriques : (a) contrainte d'"enroulement" ; (b) contrainte de "cyclisation" (Vaillant* et al.*, 2006).*

avec les acides aminés via des liaisons hydrogènes. Par exemple, la protéïne E2 du papillomavirus s'associe à deux sites distants de 4 pb, à savoir "ACCGNNNNCGGT",où N_4 est une séquence variable n'interagissant pas avec la protéïne mais qui influence la stabilité du complexe nucléoprotéïque (Zhang et al., 2004). Le rôle de cette séquence "intermédiaire" a été également mis en avant dans le cas de la protéïne CRP ("Cyclic AMP Receptor Protein", également connue sous le nom CAP, pour "Catabolic Activator Protéïne") (Ivanov et al., 1995; Gartenberg and Crothers, 1988). Les expériences avec des ADN de 147 à 163 pb contenant des boites TATA ont révélé une forte propension à la cyclisation (Shore et al., 1981; Kahn and Crothers, 1992) suggérant une forte flexibilité et/ou courbure intrinsèque. Il a été par ailleurs montré qu'une contrainte de courbure pouvait fortement favoriser l'association de la TPB ("TATA Binding Protein") avec sa boite TATA. Vu qu'il y a peu de contacts directs par liaison hydrogène (Juo et al., 1996; Kim et al., 1993), il a été proposé que les propriétés mécaniques de la boite TATA sont certainement fortement impliqués dans les mécanismes d'association avec la TBP (Davis et al., 1999). Cette hypothèse est renforcée par de récents calculs "tout atome" qui prédisent une reconnaissance principalement indirecte (Paillard and Lavery, 2004). Cependant, la nature exacte de cette flexibilité n'est pas connue. Sous contrainte, la séquence pourrait induire un "kink" comme ceux observés dans la structure de l'ADN nucléosomal (Fig. 4.19). On s'attend à ce q'un tel "kink" puisse se former dans des mini-cercles d'ADN de $158pb$ ($\sim$ longueur de persistence de l'ADN), là où la courbure moyenne imposée est forte et les fluctuations thermiques à priori donc de moindre amplitude, et se visualiser par cryo-microscopie électronique (Fig. 4.17) (Un cliché de cryo-EM d'un mini-cercle, correspond à une trajectoire thermique "figée"). Nous avons donc observé et comparé les configurations 3D de minicercles d'ADN de longueur 158 pb avec, inséré, un fragment de 18 pb contenant soit la boite TATA soit le site de CAP (CRP). Bien que les minicercles "TATA" cyclisent avec deux ordres de grandeurs plus efficacement que les minicercles "CRP", aucune différence significative n'est observée dans les reconstructions 3D des deux types de minicercles. Nous avons conclus que les fluctuations thermiques sont encore suffisamment importantes pour gommer les différences entre TATA et CAP.

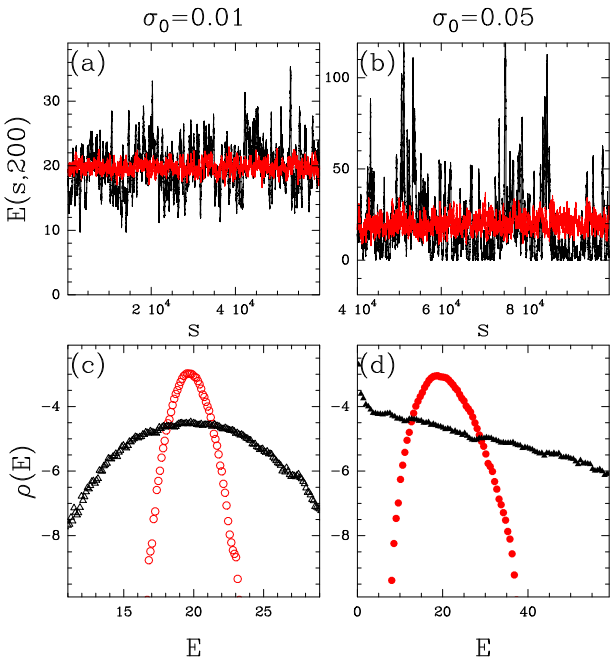


FIGURE 4.14 : *Statistiques de l'énergie d'une boucle de taille $\ell = 200$ bp se formant le long d'une chaine de taille $L = 2\ 10^5$ bp, en considérant la contrainte "d'enroulement" et pour $A(= \tilde{A}/k_BT) = 200$ bp et $\alpha = 2\pi$. (a) $E(s, 200)$ le long d'une chaine non corrélée ($H = 0.5$, noir) et corrélée ($H = 0.8$, rouge) dans le cas du désordre faible , $\sigma_o = 0.01$. (b) Profils energétiques pour un désordre plus fort, $\sigma_o = 0.05$. (c),(d) Histogramme des profils énergétiques reportés en (a), (b). (Issu de Ref. Vaillant* et al. *(2006)).*

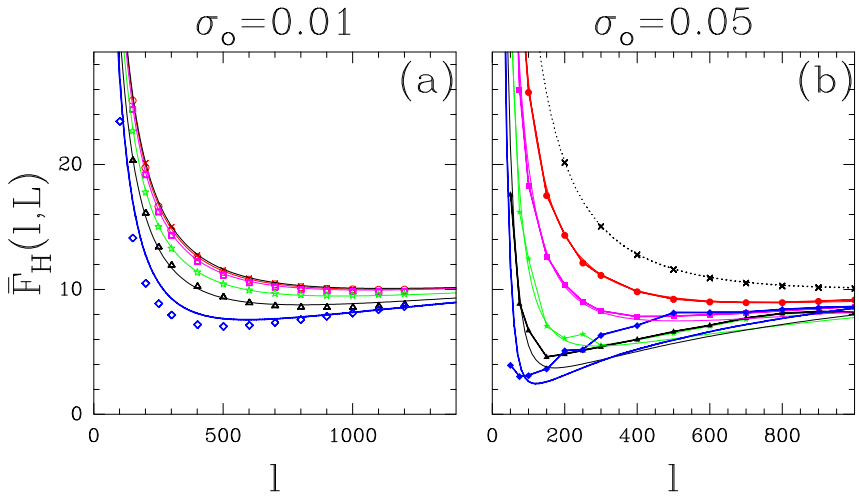


FIGURE 4.15 : *Energie libre du système "simple boucle" $F_H(\ell, L)$* vs *ℓ en considérént la contrainte de "cyclization" et pour $L = 15000$ bp, $A = 200$ bp, $\alpha = 2\pi$ and $\sigma_o = 0.01$ (a) et 0.05 (b). Les symboles correspondent à l'énérgie libre d'une seule chaine $F_H(\ell, L)$ obtenue à partir d'un calcul numérique pour $H = 0.5$ (○, ●), 0.6 (□, ■), 0.7 (losanges), 0.8 (△, ▲) and 0.9 (◊, 117); les (×) correspondent au cas "pur", sans désordre. Les lignes continues correspondent aux expressions "approchées" de la valeur moyenne $\overline{F}_H(\ell, L)$. Les pointillés correspondent à l'expression exacte pour le cas idéal (Vaillant* et al.*, 2006).*

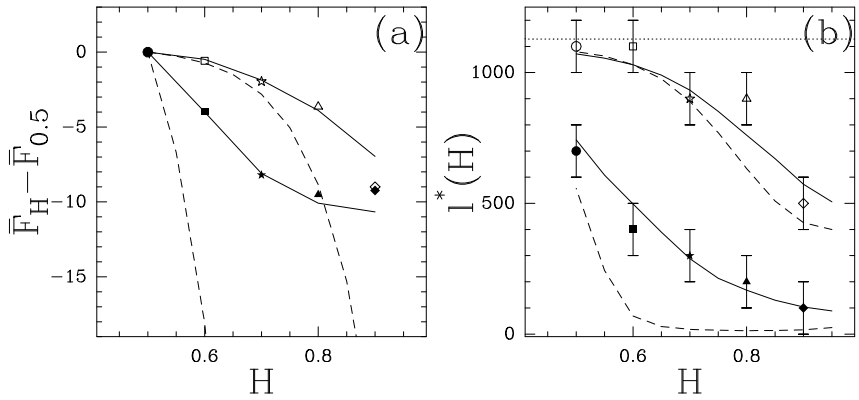


FIGURE 4.16 : *Energie libre et taille optimal de boucle* vs H *sous contrainte de "cyclisation" pour* $L = 15000$ *pb,* $A = 200$ *pb,* $\alpha = 2\pi$ *et* $\sigma_o = 0.01$ *(symboles ouverts) et* 0.05 *(symboles noirs). Les symboles et les courbes continues ont la même signification qu'à la Fig. 4.15. (a)* $F_H(\ell, L) - F_{1/2}(\ell, L)$ vs H *poue des chaine CLP et non corrélées, pour une boucle de taille* $\ell = 200$ *pb ; les courbes en tirets correspondent à l'approximation perturbative. (b)* $\ell^*(H)$ vs H *; les courbes en tirets correspondent à la solution de l'expression perturbative et les lignes horizontales en pointillés indiquent la taille de boucle optimale pour le système "idéal"* $\ell^* = 1128$ *bp.*

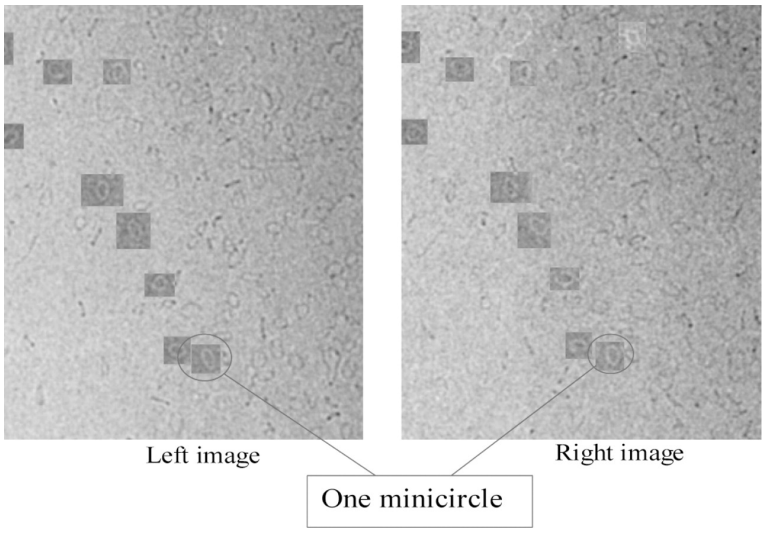


FIGURE 4.17 : *Paires stéréographiques de minicercles d'ADN obtenues par Cryo-EM et utilisées pour la reconstruction (Fig. 4.18)*

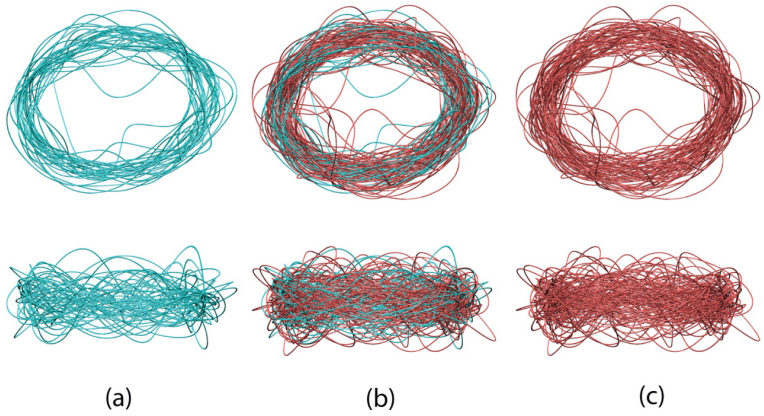


FIGURE 4.18 : *Toutes les trajectoires reconstruites sont alignées par superposition de leurs axes principaux d'inertie. (a) Tous les 31 minicercles CAP (bleu), (b) tous les 95 minicercles et (c) tous les 64 minicercles TATA (rouge).*

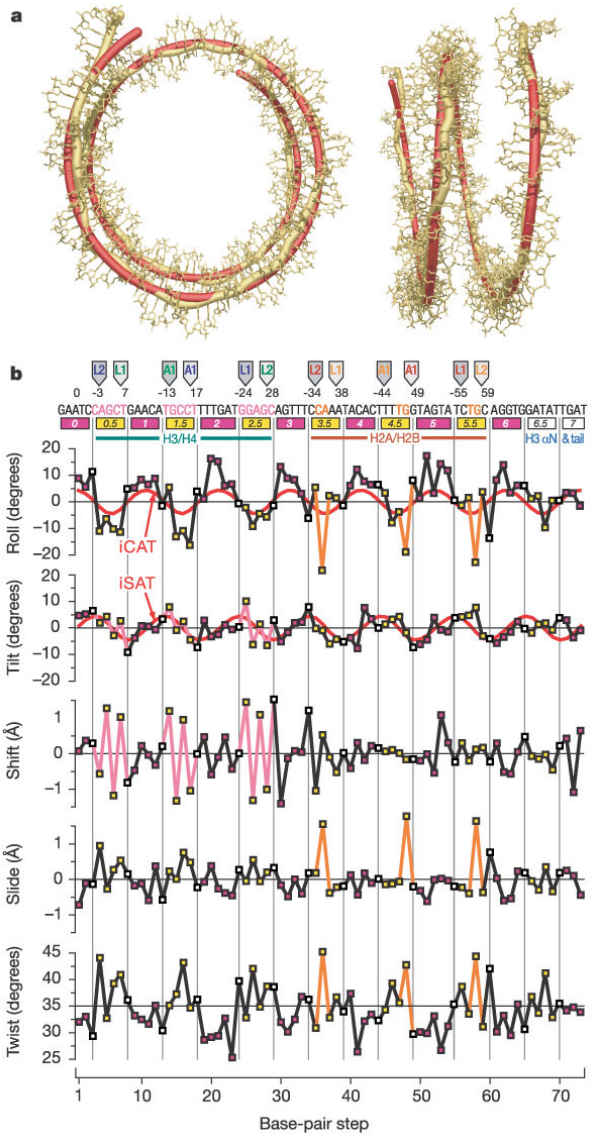


FIGURE 4.19 : *Structure du nucléosome par cristallographie aux rayons X à la résolution de 1.9 Angstrom (Richmond and Davey, 2003b)*

4.3 NUCLÉOSOMES : MODÈLES PHYSIQUES

On construit des modèles directement à partir des propriétés physiques de l'interaction nucléosome-ADN. Ces modèles exploitent essentiellement le coût en énergie de la déformation de l'ADN et l'interaction électrostatique entre l'ADN et le nucléosome. Ils sont généralement moins performants que les modèles d'apprentissage.

Some other models are solely based on the enrgetics of the nucleosome formation. They are generally less performant than the trained-based algorithms.

Pour déterminer le lien entre positionnement du nucléosome et séquence ADN, il est raisonnable d'exploiter les propriétés élastiques et mécaniques de l'ADN. Le but est cette fois de partir de la physique pour construire un modèle théorique de formation du nucléosome. Il s'agit alors d'essayer d'évaluer le coup énergétique lié à la déformation du brin d'ADN et les gains énergétiques liés à la liaison chimique entre le nucléosome et l'ADN. La base est de se référer aux observations obtenues dans les études menées sur la structure cristallographique du nucléosome (Luger and Richmond, 1998). L'ADN y est décrit par un empilement de paires de bases considérées comme des plaques indéformables (voir figure 4.20) dont la trajectoire générale enroule le nucléosome sous une forme solénoïdale. À partir de

là, les modèles énergétiques usuels exploitent les six degrés de libertés (figure 4.20 (b)) qui relient deux paires de bases consécutives pour construire une énergie de déformation de l'ADN (Olson et al., 1998). On construit alors une énergie élastique de déformation E_{el} en prenant en compte la déviation des paramètres par rapport à leur valeur au repos, pour chacune des successions de paires de bases. Cela permet donc d'expliciter une énergie qui dépend de la position le long de la séquence (figure 4.21).

A titre d'exemple, on peut construire E_{el} comme (Tolstorukov et al., 2007) :

$$E_{el} = \frac{1}{2} \sum_{i=-N/2}^{N/2} \left[\Gamma_{nuc}^{n(i)} - \Gamma_0^{n(i)}\right]^T \mathbf{A}^{n(i)} \left[\Gamma_{nuc}^{n(i)} - \Gamma_0^{n(i)}\right] \tag{4.21}$$

où Γ est le vecteur à six composantes correspondant aux degrés de libertés des dinucléotides évoqués plus haut et illustrés de nouveau à la figure 4.20 B. $N = 146$, la somme est effectuée sur l'intégralité des paires de bases affectées par le nucléosome. $\Gamma_0^{n(i)}$ correspond aux valeurs d'équilibre dans le cas libre prises par le dinucléotide de type n se trouvant à la position i au sein dun nucléosome ($i = 0$ à la dyade). $\Gamma_{nuc}^{n(i)}$ correspond aux valeurs prises par le dinucléotide de type n à la position i au sein dun nucléosome ($i = 0$ à la dyade). Enfin, $\mathbf{A}^{n(i)}$ est la matrice de coefficients d'élasticité calculée de telle sorte que $\mathbf{A}^n = (\mathbf{C}^n)^{-1}$, où : $C_{jk}^n = \langle(\Gamma_j^n - \Gamma_{0j}^n)(\Gamma_k^n - \Gamma_{0k}^n)\rangle$ correspondnat à la co-variance des fluctuations des dégré de liberté Γ_j et Γ_j dans le cas libre (sous l'effet de l'agitation thermique) pour le dinucléotide de type n.

Cette formule très générale est employée par Tolstorukov (Tolstorukov et al., 2007), Morozov (Morozov et al., 2009) et Becher et al. (Becker and Everaers, 2009). Il s'agit de calculer le coût élastique associé à toutes les déformations que ce soit angulaires (roll, tilt et twist) que de translation (slide, rise, shift) et de prendre en compte aussi tous les couplages possibles entre ces déformations. Les déformations sont mesurées par rapport à l'état de référence correespondnat à l'état d'équilibre de la double hélice "libre". Les valeurs des degrés de liberté au sein du nucléosome sont (i) soient "figées" (comme dans nos modèles présentés ci-après et dans le modèle de Tolstorukov (Tolstorukov et al., 2007)) aux valeurs correspondant à une conformation régulière de superhélice idéale (Fig. 4.20 A avec comme distributions de roll et tilt celles dénommées iCAT et iSAT à la Figure 4.19) ou à celles correspondant directement aux données cristallogarphiques (Fig. 4.20 A et 4.19) (ii) soient soumises à une relaxation à partir d'un état proche de ces configurations nucléosomales "figées" (Morozov et al., 2009; Becker and Everaers, 2009). Dans le cas du modèle de Tolstorukov (Tolstorukov et al., 2007), les paramètres élastiques, fléxibilité et valeurs d'équilibre (Γ_0^n et $\mathbf{A}^n$), qui dépendent de la séquence à l'échell du dinucléotide, sont tirés de données cristallographiques complexes ADN-prtoéines (Olson et al., 1998).

Si l'on considère que les autres sources de gain et de coût énergétiques sont indépendants de la séquence (l'interaction attractive entre le nucléosome et l'ADN nait ainsi de l'écrantage électrostatique du brin d'ADN qui, absorbé sur le nucléosome chargé positivement, n'est plus contraint par la forte charge négative qui est distribuée le long du brin) ce modèle permet d'évaluer l'influence de la séquence seule sur l'énergie de formation du nucléosome.

Il est aussi possible d'utiliser des formulations plus simples en ne prenant en compte que certaines parties des coefficients. On peut par exemple négliger les effets de translation entre deux paires de bases consécutives pour ne s'intéresser qu'aux variations angulaires. L'une de ces approches consiste à considérer qu'au sein de l'élasticité de l'ADN, seuls les paramètres angulaires importent, où plutôt les déviations des angles imposés par le nucléosome par rapport à la situation au repos (Miele et al., 2008; Anselmi et al., 2000).

Par exemple, le modèle développé par Miele *et al.* (Miele et al., 2008) considère que l'ADN forme une super-hélice de rayon $R = 4.19$ nm et de pas $P = 2.52$ nm. Les trois angles (tilt –τ–, roll –ρ–, twist –Ω–) des plaques de paire de base sont alors imposés :

$$\rho(j) = \kappa\, cos(\omega(j-i) + \phi_i) \tag{4.22}$$

$$\tau(j) = \kappa\, sin(\omega(j-i) + \phi_i) \tag{4.23}$$

$$\Omega(j) = 2\pi/10.3\, pb^{-1} \tag{4.24}$$

avec $j = i, \ldots, i + L - 1$,$\kappa = (2\pi)^2 R/(P^2 + (2\pi R)^2)$, $\omega = \Omega_3(j) - 2\pi P/(P^2 + (2\pi R)^2)$. Si l'on considère que l'ADN est inextensible et élastique, on peut alors déterminer l'énergie élastique requise pour former cette hélice :

$$\frac{\Delta E(i,L)}{k_b T} = \sum_{j=i}^{i+L-1} \left(\frac{A_1}{2}(\tau - \tau_o)^2 + \frac{A_2}{2}(\rho - \rho_o)^2 + \frac{A_3}{2}(\Omega - \Omega_o)^2 \right) \tag{4.25}$$

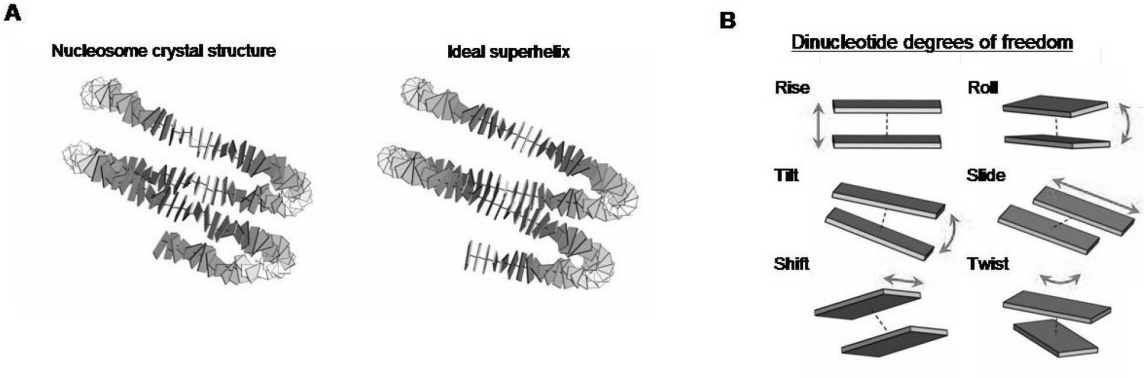


FIGURE 4.20 : *(a) Illustration de l'empilement de paires de bases imposé par l'enroulement de l'ADN autour du nucléosome, dérivé de la structure cristallographique du nucléosome (Richmond and Davey, 2003a) et juxtaposée avec une structure de type "super-hélice idéale". (b) Les différentes relations entre nucléotides consécutifs. Il existe six degrés de libertés, dont trois de déplacement des paires de bases (rise, shift, slide) et trois de rotations (twist, roll et tilt). Figure tirée de Tolkunov et Morozov (Tolkunov and Morozov, 2009)*

où A_1, A_2 et A_3 sont les flexibilité de tilt, roll et twist, contrôlant les fluctuations autour des valeurs d'équilibre τ_o, ρ_o et Ω_o, respectivement. Ces paramètres dépendnet de la séquence aux positions $j = i, ..., i + L - 1$ le long de l'ADN nucléosomal. Les valeurs d'équilibre et les flexibilités sont déterminés comme en (Scipioni et al., 2002b). Nous négligeons l'anisotropie de coubure locale et retenons donc que la composante isotrope controllée par la flexibilité de courbure "isotrope" $A = (A_1+A_2)/2$. Nous considérons $A(j) = A^* t_m(j)$, où $A^* = 50$ nm est la flexibilté "standard" d'une séquence (aléatoire) et t_m est un facteur de modulation (proche de 1) dépendnat de la séquence. De la même façon, $A_3(j) = A_3^* t_m(j)$ avec $A_3 = 75$ nm (Neukirch, 2004). Comme décrit dans (Scipioni et al., 2002b), $t_m(j)$ est definie comme le rapport de la température locale de fusion sur sa valeur moyenne pour un ADN standard : $t_m(j) = T_m(j)/T_m^*$. C'est une grandeur qu'on peut relier à l'énergie d'empilement et à la longueur de persistence (flexibilié de courbure) de la double hélice (Scipioni et al., 2002b). Nous prenons aussi en compte le coût entropique correspondnat à la transition de l'état libre à la configuration contrainte (figée) du nucléosome :

$$\Delta S(i, L) = 3/2 L ln(t_m(i, L)), \tag{4.26}$$

où

$$t_m(i, L) = (1/L) \sum_{j=i}^{j=i+L-1} t_m(j). \tag{4.27}$$

Au final on considère comme profil énergétique "nucléosomal" l'énergie libre donnée par :

$$\frac{\Delta F(i)}{k_B T} = \frac{\Delta E(i)}{k_B T} - \Delta S(i). \tag{4.28}$$

La figure 4.21 montre la forme du potentiel énergétique de Miele *et al.* sur une partie du chromosome I de la levure. On voit bien la composante oscillante à $10pb$ qui dépend essentiellement du cartère anisotrope de la distraibution de courbure intrinsèque. On ne retiendra pour notre part que l'enveloppe inférieure correspondant aux états "locaux" (à $10\ pb$ de paires près) de plus faibles énergie libre. On pourrait penser qu'oublier cette haute fréquence est critique pour déterminer le positionnement nucléosomal, pourtant, tant que l'on ne travaille pas à une résolution inférieure à 10 pb, ce n'est pas un problème, puisque le résultat du positionnement une fois filtré à l'échelle de quelques paires de bases ne change pas fondamentalement avec ou sans les hautes fréquences (cf. chapitre 4 de la thèse de Guillaume (Chevereau, 2010) revient plus en détail sur le rôle joué par une haute fréquence sur le positionnement et montre qu'effectivement, une haute fréquence ne change pas le positionnement des nucléosomes à une résolution de l'ordre la dizaine de paires de bases). Ce modèle élastique obtient des résultats très satisfaisants par rapport aux autres modèles physiques (tableau 6.1). De fait, il ne tient compte que de la déviation de la trajectoire de l'ADN par rapport à une trajectoire gelée qui dépend de la séquence (les angles Ω_o). Le choix des constantes d'élasticité est en fait primordial et selon les tables de coefficients

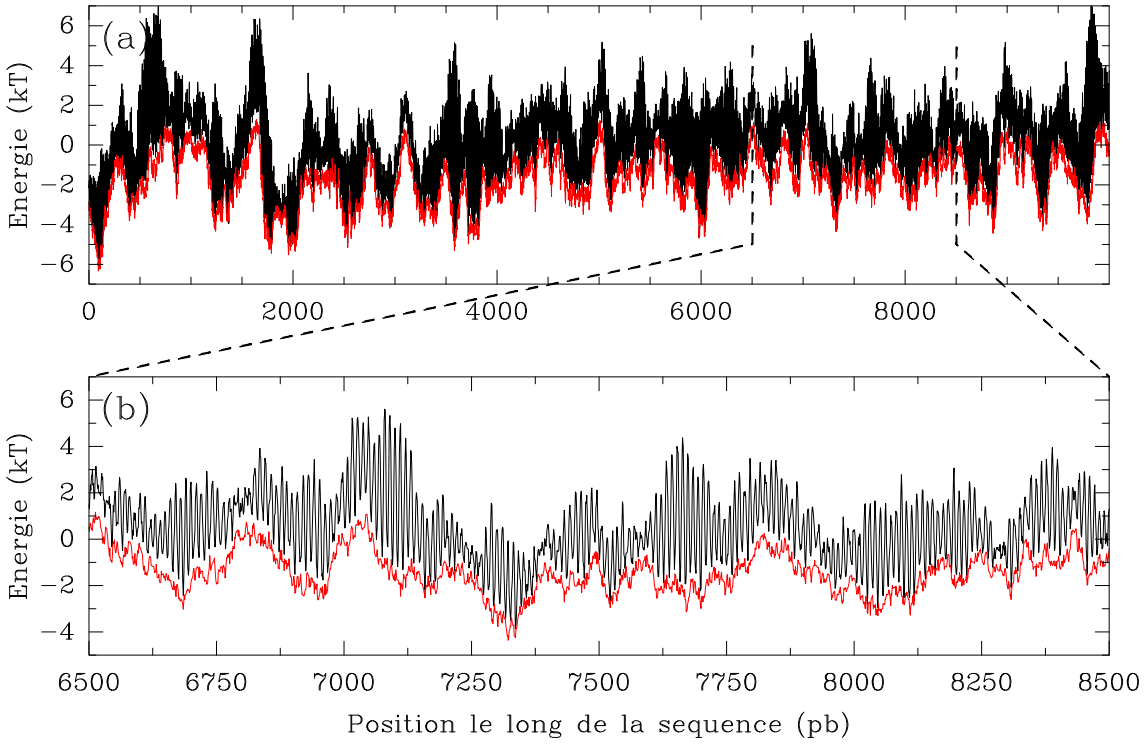


FIGURE 4.21 : *Forme générale d'un profil nucléosomal sur le chromosome I de la levure à grande échelle (a) et à moyenne échelle (b) : En noir le modèle de Miele et al. (Miele* et al.*, 2008). En rouge, l'enveloppe inférieure du profil de Miele et al.*

utilisées, les résultats des modèles physiques peuvent changer du tout au tout. Nous utilisons également un modèle (Vaillant) en tout point similaire (isotrope, sans degré de liberté de translation, avec des coefficients élastques constants, une trajectoire nucléosomale de type solénoïde régulier) mais qui utilise une autre table déterminée à partir de valeurs d'équilibre de roll établie par Goodsell et Dickerson (Goodsell and Dickerson, 1994; Satchwell et al., 1986). Le calcul du profil énergétique est le même que pour le modèle Miele, simplement on considère le tilt intrinsèque nul ($\tau_o \equiv 0$) et le twist intrinsèque constant $\Omega_o = 34.3$. Contrairement à la table d'Anselmi cette table n'est pas d'origine physique, elle est établie justement à partir de données de positionnement de nucléosomes (Goodsell and Dickerson, 1994; Satchwell et al., 1986).

4.4 MODÈLES PHÉNOMÉNOLOGIQUES ET PROBABILISTES : PÉRIODICITÉ À 10.2 PB ET "BIAIS" DE COMPOSITION

L'approche la plus directe dans le sens où elle ne requiert pas de calculs physiques, a d'abord été d'inférer le positionnement en fonction de la séquence et ce, à partir de données expérimentales. L'idée est, à partir d'un jeu de séquences issu d'expériences d'extraction d'ADN nucleosomal (in vivo et in vitro) ou de competition, d'identifier certaines propriétés/organisation génomiques particulières et qui seraient donc susceptible de favoriser l'interaction ADN-histones :

(i) positions préférentielles de certains di-,tri-,tetra-... nucléotides le long de la trajectoire de l'ADN autour de l'octamere (position qu'on mesure généralement à partir de la dyade), et qui par exemple indique dans quelle mesure un certain motif préfère avoir son grand sillon oriente vers le coeur plutôt que vers l'extérieur, ou être davantage localisé vers la dyade qu'aux extremites... L'hypothese que, quelle

que soit la séquence, l'organisation de la double hélice autour du coeur d'histones est proche de celle mesuree sur le cristal (en particuliers les distributions de roll etc...) fait qu'effectivement les positions le long de l'ADN nucleosomal ne sont pas équivalentes. L'octamères d'histones n'offre pas une surface d'interaction uniformes avec notamment les points de contacts preferentiels revenant tous les $10\ bp$: une séquence qui de par ses proprietes physico-chimique particulières s'adapterait plus a cette surface d'interaction serait donc plus favorable. A partir d'un jeu de séquences identifiées comme étant préférentiellement "nucléosomales", il est possible d'établir à l'échelle du di- tri- ou tetra-nucléotide une table résumant pour chaque motif (i.e. AA, AAA, AAAA ...) dans quelle mesure celui-ci est distribué périodiquement à $\sim 10pb$ (en appliquant la transformée de fourier sur le profil de répartition du-dit motif le long de la séquence nucléosomale) et avec quelle phase (petit ou grand sillon orienté vers la surface de l'octamère ?). C'est d'ailleurs l'origine exacte de la table de roll "Pnuc" (Satchwell et al., 1986). Au delà de la distribution de certains motifs on peut aussi s'interesser à la périodcité dans les distributions de certaines propriétés structurales dépendantes de la séquence comme les courbures et flexibilités intrinsèques ou la taille du grand sillon etc...

(ii) composition moyenne en certains mono-, di ... nucléotides : si l'organisation (périodicté, corrélations) de certains motifs le long d'une séquence peut effectivement avoir un effet favorable ou défavorable, on peut imaginer aussi que la composition moyenne (peu importe l'ordre) compte aussi ; par exemple, une séquence très flexible (longueur de persistence faible) induite par une composition moyenne en TA (ou G+C) importante peut être plutôt plus favorable à son enroulement simplement parce que le coût élastique est plus faible (il y a cependant un coût entropique plus important...) (Virstedt et al., 2004). Inversement une composition forte en poly(dA :dT) indiquera plutôt une séquence anti-nucléosomale. De même que pour les périodicté, on peut tout aussi bien s'interesser à l'effet d'une propriété structurale "moyenne" (temperature de fusion, energie d'empilement, taille du grand sillon, flexibilité...).

De nombreux types de code ont été établis pour prédire l'affinité du nucléosome avec la séquence : soit simplement par recherche de périodicités et par corrélation avec le profil de contenu nucléotidique expérimental (Ioshikhes et al., 2006; Segal et al., 2006), soit en rajoutant des termes liés aux k-mères de nucléotides (figure 4.26 (b)), grâce à l'utilisation d'algorithmes d'apprentissage (Peckham et al., 2007; Field et al., 2009).

4.4.1 Périodicité simple

Périodicité à 10 pb :

Dans un premier temps, la recherche de périodicités dans le contenu nucléotidique a constitué la piste principale des études portant sur la liaison nucléosome-ADN (Satchwell et al., 1986; Trifonov and Sussman, 1980; Cohanim et al., 2005, 2006). Pourquoi la périodicité ? Essentiellement pour des raisons structurales. L'ADN est très rigide et est fortement contraint lorsqu'il est enroulé autour du nucléosome. Si certains mots nucléotidiques peuvent plus facilement relaxer cette contrainte que d'autresde part notamment des propriétés elastiques particulières, ils favoriseront l'adsorpton autour de l'octamère. C'est cette idée qu'ont introduite Trifonov et Sussman dans leurs travaux précurseurs en révélant la présence d'une périodicité préférentielle à 10 pb dans des chaines ADN nucleosomales (Trifonov and Sussman, 1980). Ceux-ci ont suggéré que cette periodicité induirait une déformabilité anisotrope de la chaine facilitant ainsi son enroulement au sein du nucléosome. Il avait été observé que les motifs riches en A/T and G/C se courbent plus facilement vers le petit et grand sillon respectivement ; lorsque ces séquences sont arrangées avec une période de $10\ pb$, le nucléosome se positionne de telle sorte que les A+T riches ont leur petit sillon orienté vers la surface de l'octamère et les G+C riches le petit sillon orienté vers l'extérieur (Drew and Travers, 1985; Travers and Klug, 1987; Satchwell et al., 1986; Shrader and Crothers, 1989; Travers, 1991) (donc toujours à une position où "leur" courbure préférentielle est en accord avec celle imposée par l'interaction avec la surface). Le séquencage d'un grand nombre de fragments d'ADN nucléosomaux d'erythrocytes de poulet (Satchwell et al., 1986), SV40 (Bina, 1994) et de levure (Segal et al., 2006) ont révélé que les dinucléotides AA/TT présentent une période de 10 pb avec un dephasage de 5 pb avec GC (Fig. 4.23).

Modèles probabilistes :

De nombreux modèles basés sur cette périodicité dans la ditribution de mots dinucleotidiques ont été mis en place :

1. Modèle de **Ioshikhes** (Ioshikhes et al., 2006) :

Calcul de la corrélation entre les motifs AA/TT périodiques d'une séquence donnée avec ceux issus d'un jeu de 204 séquences nucléosomales eucaryotes et virales extraites par des expériences *in vivo* et *in vitro*. Les séquences sont converties en signal de distribution en AA/TT/AT et ces signaux sont moyennes de tell sorte à obtenir un profil "moyen" (consensus) qu'on correle ensuite au signal AA/TT/AT de la séquence voulue. Au final en toute position on obtient un score nucleosomal (Fig. 4.25(a)).

2. Modèle de **Segal** (Segal et al., 2006)
   Le modèle de Segal de 2006 diffère de celui de Field ou Kaplan en cela qu'il ne tient compte que du caractère positionnant (périodique) de la séquence. C'est un modèle donc probabiliste entrainé sur la levure (données de Yuan et al. (Yuan et al., 2005)) à partir d'une matrice de score positionnement déduite d'un jeu de séquences nucléosomales issues d'expériences de sélection *in vitro*.
3. **Modèle "Pnuc" :**
   Le modèle que nous choisissons d'utiliser (Modèle "Pnuc") est hybride : Les coefficients sont établis à partir de données de positionnement de nucléosomes (Satchwell et al., 1986) : l'analyse spectrale de 177 séquences nucléosomales chez le poulet a permis de quantifier à l' échelle du "trinucléotide" le caractère périodique ($10.2pb$) des motifs ainsi que leur phase préférentielle (petit ou grand sillon vers la surface de l'octamère) ; donc de mesurer une préférence nucléosomale et localisation préférenteille pour chaque trinucléotides. Ces coefficient ont été ensuite traduits en terme de paramètres élastique et notamment d'angles de roll intrinsèques ρ_o (Goodsell and Dickerson, 1994). Le modèle élastique utilisant les paramètres "Pnuc" n'est donc pas stricto-sensus un modèle physique.

Periodicité à 10 pb : Interprétation physique :

De manière générale, toute séquence dont les propriétés élastiques favoriseraient une distribution de roll, tilt etc... voisines de celles observées dans le cristal (Fig. 4.19) est susceptible de favoriser son enroulement. D'un point de vue énergétique de telles séquences anisotropes induisent localement de fortes variations dans le profil énergétique, de période $10pb$: si en effet on considère que la configuration de la double hélice autour du nucléosome est fortement contrainte (notamment par les points de contacts qui "préfèreraient" les petits sillons), partant d'un configuration favorable, déplacer la chaine ADN de $5\ pb$ est effectivement très coûteux. Dans le cadre des modèles élastiques présentés plus haut (Miele et al., 2008) on peut effectivement illuster quantitativement l'effet d'une telle périodicité (Fig. 4.22). On voit effectivement que pour la séquence construite par mes soins les profils oscillent beaucoup plus qu'au niveau des séquences aléatoires avec donc, à la fois des positions plus favorables et des positions défavorables. Le profil énergétique correspondant aux états optimaux (enveloppe inférieure du profil initial oscillant) présente ainsi au niveau de cette séquence un minimum, plus marqué d'ailleurs pour les modèles "Anselmi"(Fig. 4.22 (c)) et "Pnuc" (Fig. 4.22 (a)) que pour le modèle "Tolstorukov"(Fig. 4.22 (e)). La séquence super-positionante ne l'est par contre pas pour nos différents modèle, sauf légèrement pour le modèle "Pnuc" (Fig. 4.22)(b)).

4.4.2 Autres "règles" génomiques

Rajouter une pénalité pour certains motifs nucléotidiques (mono,di,...,pentamères de nucléotides particuliers) permet d'améliorer les prédictions des modèles fondés sur l'analyse de séquences.

Introducing a penalty to mono, di, ..., and pentameres of particular nucleotides will produce a very good predicting tool, if used in conjunction with a training algorithm.

L'évolution naturelle de ce type d'approche a été de rechercher des motifs différents (plus longs par exemple), et d'y associer une pertinence vis-à-vis du nucléosome. Un certain nombre de règles de positionnement ont été établies en extrayant des données de positionnement les configurations de nucléotides influençant la formation des nucléosomes. Les modèles récents introduisent une pénalité pour certains k-mères de nucléotides que l'on a noté comme étant défavorables ou favorables aux nucléosomes. On peut alors établir une table, qui pour chaque séquence nucléotidique, définit un score nucléosomal. Pour plus de performance, on entraine un modèle (un ensemble de règles permettant d'établir un score parfois appelé code de positionnement sur un jeu de données connues puis on confronte le modèle à un autre jeu de données pour déterminer sa pertinence. Les règles peuvent être modifiées jusqu'à ce qu'il y ait une bonne concordance entre modèles et expériences. Cette approche a notamment été initiée par

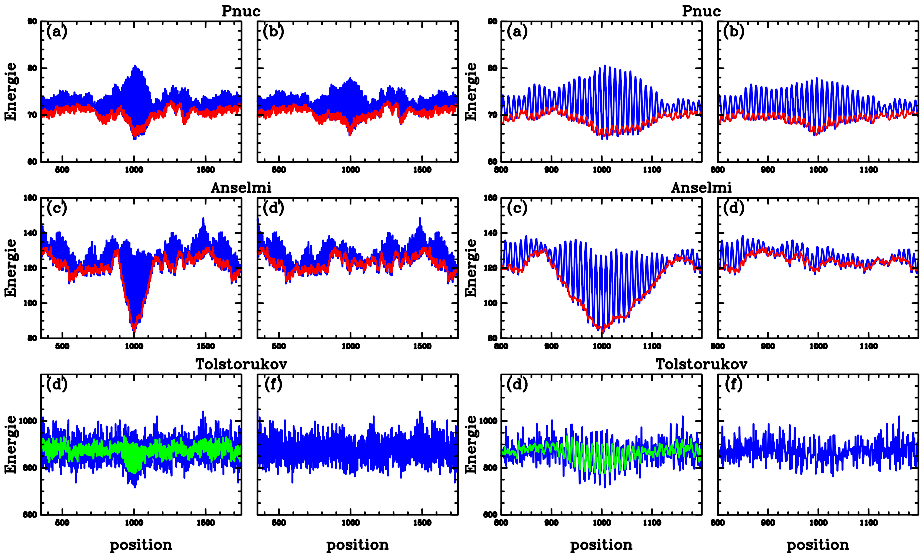


FIGURE 4.22 : *Profils énergétiques (la figure de gauche correspond à un zoom de cell de droite) calculé à partir des modèles élastiques "simplifiés" (Miele* et al.*, 2008) avec les paramètres "Pnuc" (Satchwell* et al.*, 1986; Goodsell and Dickerson, 1994; Vaillant* et al.*, 2007) (a,b), "Anselmi" (Scipioni* et al.*, 2002*b*) (c,d) et le modèle plus complet de Tolstorukov (Tolstorukov* et al.*, 2007)avec les paramètres "Olson" (Olson* et al.*, 1998) pour deux séquences : Une séquence aléatoire avec incluse au mileu au choix, deux séquences "périodiques" de* $\sim 130pb$*, l'une construite par mes soins (a,c,e) et une autre (b,d,f), dite "super-positionnante", construite par J. Widom et E. Segal pour l'étude de Wang et al. (Wang* et al.*, 2011). Les courbes rouges en (a,b,c,d) correspondent aux profils énergétiques "optimaux".*

Yuan *et al.* et Peckham *et al.* (Yuan and Liu, 2008a; Peckham et al., 2007) dont un exemple est représenté sur la figure 4.25(b).

Le principe général de ce type de modèle est détaillé en figure 4.26. La périodicité est utilisée comme base pour le modèle, un raffinement est apporté en favorisant (resp. pénalisant) les pentamères qui sont effectivement surreprésentés dans les nucléosomes (resp. *linkers*). Il est difficile d'établir une signification physique associée à ces codes, puisque ce n'est plus seulement la périodicité qui les guide, mais la présence ou non de certains k-mères favorables ou pénalisant.

1. Modèle de **Yuan et Liu** (Yuan and Liu, 2008b)
   Calcul de l'occupation en nucléosome à partir d'un signal dinucléotidique périodique extrait de séquences d'ADN nucléosomal et de linker extraits d'expriences *in vivo* et *in vitro* menées chez la levure.
2. Modèle de **Peckham** (Peckham et al., 2007)
   Classification (par "Support Vector Machine") de l'affinité de séquence par l'étude de la surreprésentation des k-mers (k=1-6) dans un jeu "d'entrainement" de séquences occupées ou déplétées en nucléosomes extraites des données *in vivo* de la levure.
3. Modèles de **Field** (Field et al., 2008a) & **Kaplan** (Kaplan et al., 2009a)
   Modèles probabilistes basés sur la préférence des 5-mers mesurée *in vivo* (resp. *in vitro*) chez la levure et sur la périodicité à 10.2 pb des dinucléotides.
4. Modèles de **Lasso** Lee et al. (2007a); Tillo and Hughes (2009) :
   Modèles de régression linéaire entrainés sur des données *in vivo* d'occupation en nucléosomes. Comem entrée, ces modèles utilisent aussi des paramètres structuraux, des séquences anti-positionnantes et des sites de facteurs de transcription.

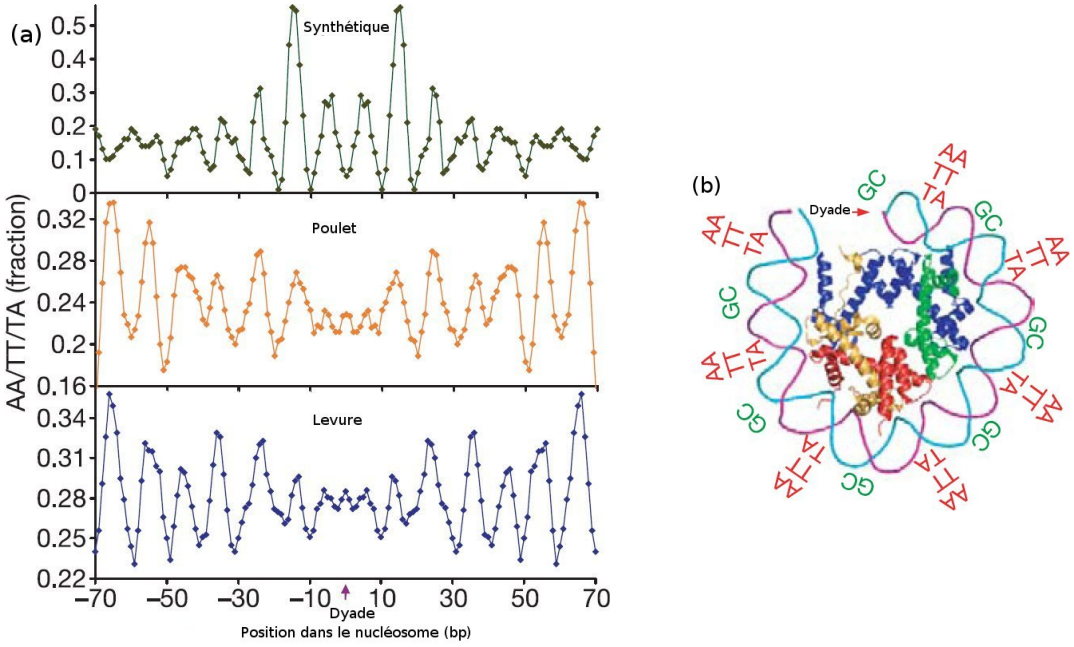


FIGURE 4.23 : *(a) Fraction en dinucléotides AA/TT/TA (moyenne sur* 3 *paires de bases) alignée sur des séquences extraites de la levure, du poulet et de séquences artificielles où se fixent expérimentalement des nucléosomes. (b) Position préférentielle des dinucléotides essentiels tirés des alignements de séquences par rapport à la structure du nucléosome. Figure adaptée de l'article de Eran Segal (Segal* et al.*, 2006).*

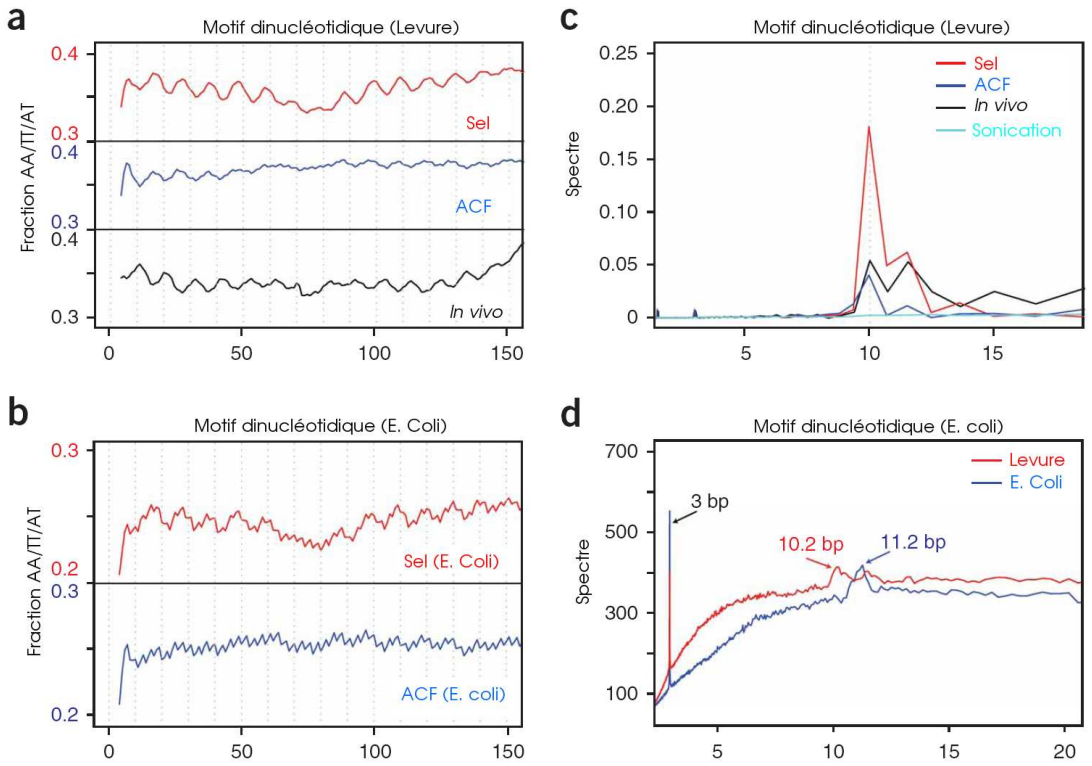


FIGURE 4.24 : *(a) Les motifs préférés des nucléosomes selon les conditions (Sel = reconstitution in vitro par dialyses, ACF = reconstitution en présence d'un facteur de remodelage et d'une chaperonne (NAP)). (b) La même chose chez la bactérie E. coli. (c) Spectre de Fourier des données de (a). (d) Spectre du signal issu des séquences de la levure et de E. coli associé aux motifs AA/TT/AT dans les deux organismes. (figure traduite de Zhang et Struhl (Zhang* et al.*, 2009))*

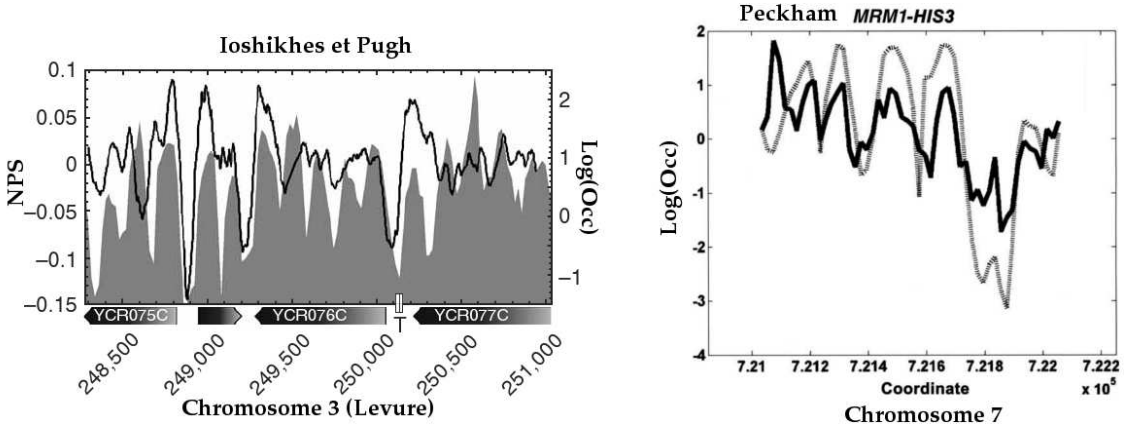


FIGURE 4.25 : *Modèles in silico. À gauche : modèle de périodicité pur de Ioshikhes (Ioshikhes* et al.*, 2006) (NPS) sur un extrait du chromosome III de la levure (données expérimentales (Yuan* et al.*, 2005) de positionnement en fond). À droite, le modèle de Peckham (Peckham* et al.*, 2007) (noir) comparé aux données expérimentales (Yuan* et al.*, 2005) (apprentissage en gris) sur un extrait du chromosome VII.*

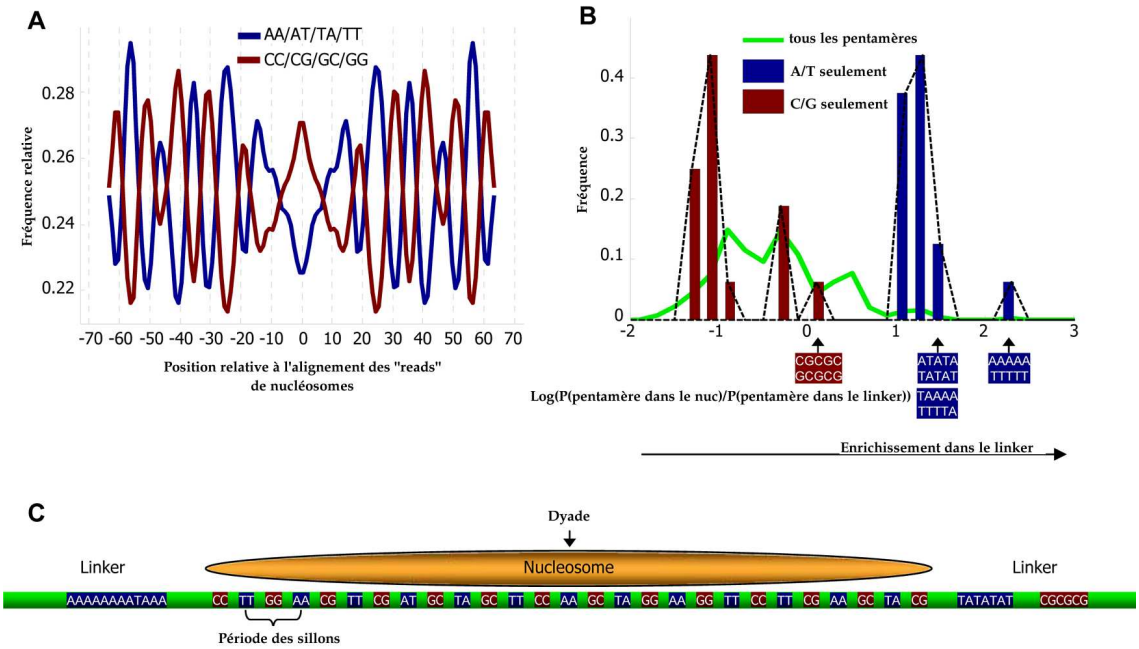


FIGURE 4.26 : *(A) Fraction de dinucléotides AA/AT/TT (en bleu) et de CC/CG/GG (en rouge) sur les positions relatives au centre des séquences liées aux nucléosomes. (B) Un certain nombre de pentamères sont plus fréquents dans les linkers ou bien dans les régions nucléosomales : ratios en* $\log_2$ *des pentamères dans les linker sur les pentamères dans les nucléosomes. En vert l'ensemble des pentamères, en rouge les pentamères constitués de C ou de G seulement. En bleu les pentamères contenant des A ou des T seulement. (C) illustration des caractéristiques principales du modèle d'apprentissage. La dyade désigne le milieu de la séquence sur laquelle se fixe un nucléosome.*

5 POSITIONNEMENT STATISTIQUE

Les obstacles dans un profil énergétique provoquent un positionnement qui peut s'étendre sur des distances allant jusqu'à 5 ou 6 fois la taille d'une particule. On parle alors de positionnement statistique, par opposition au positionnement local, induit par le profil énergétique, et particulièrement par des puits énergétiques (séquences favorables).

Here we investigate the statistical positioning as opposed to the intrinsic positioning (sequence induced), that arise next to boundary conditions.

5.1 MESURES PERTINENTES

Quels cas particuliers sont pertinents en termes de positionnement nucléosomal, quels sont les paramètres qui peuvent jouer dans ce cas ?

What parameters are relevent to this type of study ?

5.1.1 Position du problème

D'un point de vue théorique, plusieurs paramètres peuvent affecter l'équation de Percus, quels est le rôle respectif de chacun d'eux ?

A few parameters can affect the Percus equation. What are their respective role ?

On dispose à présent d'un modèle énergétique pour décrire la formation des nucléosomes sur l'ADN, et d'un modèle d'interaction lors de leur assemblage collectif. L'étude détaillée de la construction du profil énergétique a été évoqué au chapitre précédent. Nous nous intéressons à la deuxième étape, à savoir l'assemblage collectif des nucléosomes. On considère le potentiel de formation des nucléosomes comme une donnée, une entrée dans le problème. Avant tout il convient d'identifier avec soin le rôle joué par chacun des paramètres de l'équation de Percus, sans quoi il sera difficile d'estimer les éléments déterminants du positionnement nucléosomal. Le rôle de ce chapitre est donc d'analyser comment la forme du potentiel utilisé, l'épaisseur du nucléosome, l'amplitude des variations du profil énergétique, le potentiel chimique et la température, bref tous les paramètres libres, peuvent influencer la structure globale des résultats. Nous mettrons particulièrement l'accent sur le rôle du potentiel chimique et celui de l'amplitude de variation du potentiel, car ce sont les deux paramètres que l'on peut changer après avoir calculé le profil énergétique. Par souci de clarté et de simplicité, on ne s'intéresse d'abord qu'à des situations relativement simples, à savoir des profils homogènes, éventuellement ponctués de barrières énergétiques. L'objectif est de fixer les idées sur des situations intuitives. Par ailleurs, biologiquement, l'effet d'un certain nombre de protéines peut être modélisé, au premier ordre, par l'ajout d'une barrière dans le potentiel créé par la séquence d'ADN. Lorsqu'une protéine se fixe sur l'ADN, elle interdit généralement la formation d'un nucléosome à cet endroit, on peut donc interpréter la présence de cette protéine en terme de barrière énergétique à partir du moment où la protéine a un temps de résidence important. De plus, il a été montré que les potentiels générés à partir de séquence réelles présentaient un nombre élevé de séquences défavorables, localisées précisément à des endroits cruciaux du génome (Sekinger et al., 2005; Yuan et al., 2005; Peckham et al., 2007). Ces zones peuvent encore une fois être assimilées à des barrières énergétiques, induites par la séquence cette fois. L'étude détaillée des liens entre les barrières énergétiques et la biologie sera détaillée au chapitre 9, mais les aspects physiques du problème forment le coeur de ce chapitre. Le processus mis en évidence ici est essentiellement un effet de positionnement statistique, résultant de l'interaction des particules avec des éléments positionnant du profil énergétique. En terme de nucléosome, on peut donner l'exemple d'une zone favorable, de quelques paires de bases seulement, bordée par des zones défavorables de quelques dizaines de pb : un nucléosome s'y fixera quasi-systématiquement, et de ce fait, constituera à son tour une barrière énergétique pour les autres nucléosomes, provoquant ainsi des oscillations de positionnement à proximité immédiate de cette nouvelle barrière.

5.1.2 Quels sont les paramètres pertinents ?

On distingue trois paramètres essentiels : la forme de l'énergie E (en particulier son amplitude δ), le potentiel chimique μ et la taille des nucléosomes l.

Three parameters are essential : The shape of the energetic profile E, its amplitude δ, the chemical potential μ and of course the size of the nucleosomes l.

Les paramètres qui sont susceptibles de modifier les résultats de l'équation de Percus

$$\beta\mu = \beta E(s,l) + ln\,\rho(s) - ln\left(1 - \int_s^{s+l} \rho(s')ds'\right) + \int_{s-l}^{s} \frac{\rho(s')}{1 - \int_{s'}^{s'+l} \rho(s'')ds''} ds' \tag{5.1}$$

sont :

- la forme du potentiel E (le profil énergétique généré par la séquence, sur lequel on peut jouer en modifiant l'affinité du nucléosome pour la séquence localement.)
- le potentiel chimique μ (L'affinité moyenne du nucléosome pour la séquence, sur lequel on peut jouer en modifiant les conditions globales d'adsorption du nucléosome : concentration saline, taux moyen de G+C, méthylation de l'ADN)
- la taille des tiges rigides l (est déterminée par l'interaction entre particules. Ce paramètre peut vraisemblablement être modifié par la presence ou non d'histones *linkers*, la respiration du nucléosome, donc son état général : acétylé ou non, présence de variants d'histones, etc.)
- et la température $\beta^{-1} = kT$.

Souvent, les paramètres seront mêlés de façon à ce qu'un minimum de paramètres sans dimension soit nécessaire à une complète description. Par exemple, il sera toujours possible de renormaliser E et μ par la température, on utilisera donc indifféremment E pour βE et μ pour $\beta\mu$. L'énergie de formation des particules E (ou potentiel) n'est pas un paramètre au sens stricte, dans le sens où il s'agit d'un profil particulier à une séquence, et qu'à chaque séquence différente correspondra un comportement différent.

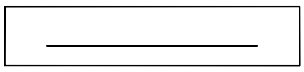
5.1.3 Quelles formes de E étudier ?

Plusieurs cas particuliers de formes de E sont intéressants : présentons ici quelques situations typiques avec des exemples tirés de situations réelles.

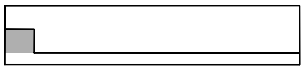
A few real life exemples of peculiar energetic field shapes.

On peut considérer que chaque forme générale du potentiel constitue une instance donnée du paramètre "Énergie". On distinguera donc plusieurs cas typiques :

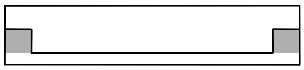
– $E = 0$ ou potentiel plat.



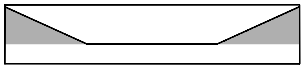
– Une barrière infinie, bordé par un potentiel plat ailleurs.



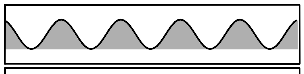
– Un puits bordé par des barrière infinies. La "boîte"

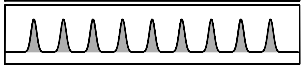


– Un puits bordé par des rampes (confinement par une force).

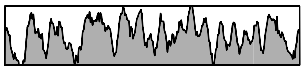


– Potentiel E périodique (sinusoïdal, puis peigne convolué).





– E calculé à partir d'une séquence (aléatoire, puis génomique).



Dans chacune de ces conditions (mises à part le potentiel périodique, cf. thèse de G. Chevereau), nous étudierons l'influence des paramètres pertinents sur la densité de particules. Ces situations, qui pour-

raient apparaître comme idéales, sont loin d'être absentes des cas réels. La figure 5.1 présente plusieurs profils énergétiques, calculés sur le chromosome I de *C. elegans*, qui correspondent plus ou moins à ces situations. Par exemple, les 10 kb situés aux alentours de la $9820000^{\text{ème}}$ paire de base de ce chromosome présentent des profils énergétiques très variés qui couvrent la quasi-totalité des cas d'école suscités. On y trouve :

- des profils oscillants très rapidement qui résultent au final en un profil relativement plat à l'échelle du nucléosome (figure 5.1 (a) B et D, et figure 5.15). Mais aussi des puits de potentiel rectangulaires de taille variable (figure 5.1 (a) C,C' et C''). Notons que systématiquement, ces profils énergétiques très particuliers sont issus de séquences répétées. Les données de positionnement nucléosomal utilisées dans cette figure sont issues de séquençages et par conséquent, ont beaucoup de difficultés à échantillonner ces zones répétées. La difficulté liée au séquençage des régions comportant des répétitions pourrait ainsi expliquer la qualité peu satisfaisante des données expérimentales de Valouev (Valouev et al., 2008) dans ces zones. Ainsi, dans la figure 5.1 (a) la zone A correspond à 15 répétitions d'une séquence de 200 paires de bases la zone C à 35 copies de 15 paires de bases et la zone C' à 8 copies de 200 paires de bases (Wormbase [a]).
- Par ailleurs, on trouve aussi des profils extrêmement plats, encore une fois issu de séquences en tandem, comme celui de la zone E (figure 5.1 (b)), constituée de 76 copies de 45 paires de bases (zone non codante mais qui correspond à un intron au sein d'un gène).
- Aussi certaines zones peuvent présenter des profils quasi-sinusoïdaux (figure 5.1 (c) et (d), zone F, au sein du gène *ZC247.1*).
- Enfin, certaines régions abritent même des séquences répétées avec une barrière énergétique prononcée qui pourraient correspondre au peigne convolué décrit plus haut (figure 5.1 (e), zone G). Cette dernière zone est d'autant plus intéressante qu'elle correspond à un gène prédit par Mark Borodovskpar [b] des méthodes HMM, où chacune des barrières énergétiques correspondrait à un intron.

Mais l'ensemble des profils présentés sur la figure 5.1 ne sont généralement pas la norme. Par exemple, chez la levure, on trouve en effet très peu de zones répétées. Dans d'autres organismes, leur occurrence peut ne pas être anecdotique (Huda et al., 2009).

a. `http://www.wormbase.org/`
b. cf. Wormbase

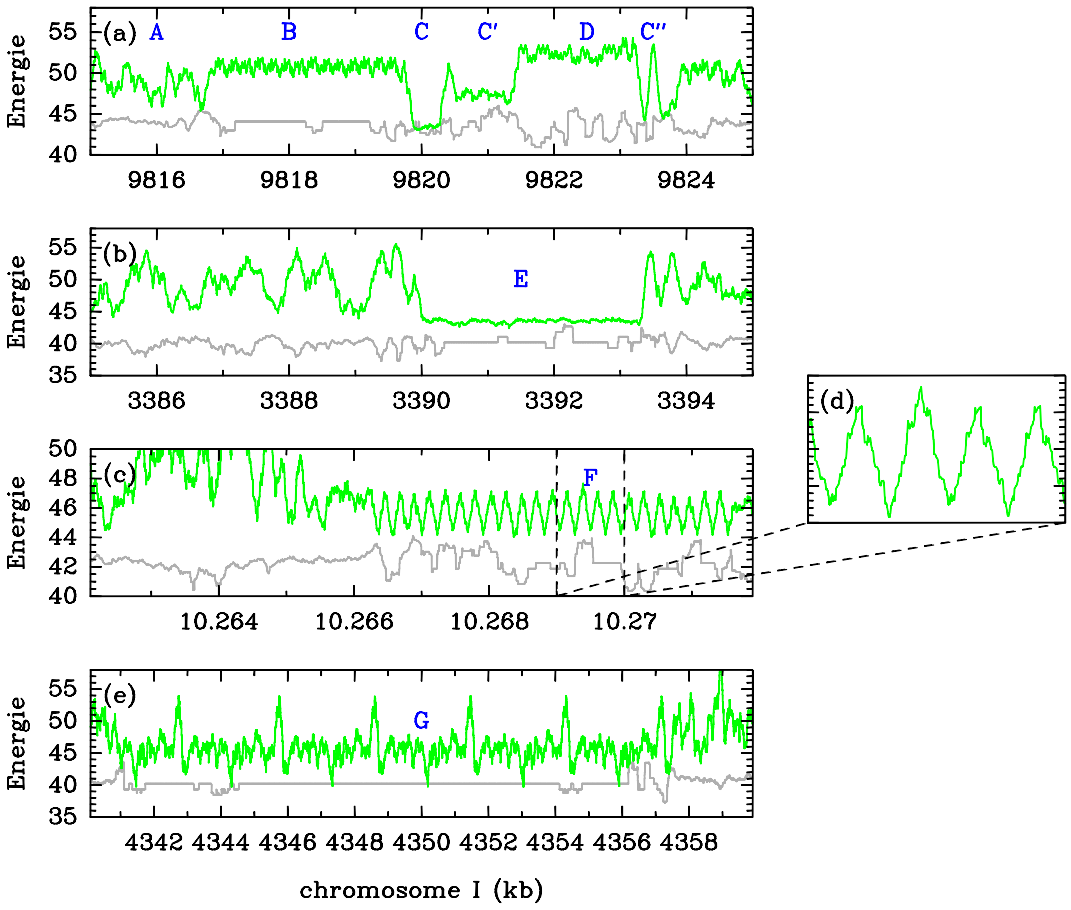


FIGURE 5.1 : *Extraits du chromosome I de C. elegans. En vert, le profil énergétique calculé avec le modèle Vaillant (Vaillant* et al.*, 2007). En gris, les données expérimentales de Valouev (Valouev* et al.*, 2008) (données normalisées, représentées en* log*). Les zones répétées, difficiles à échantillonner sont plates, car vides de données.*

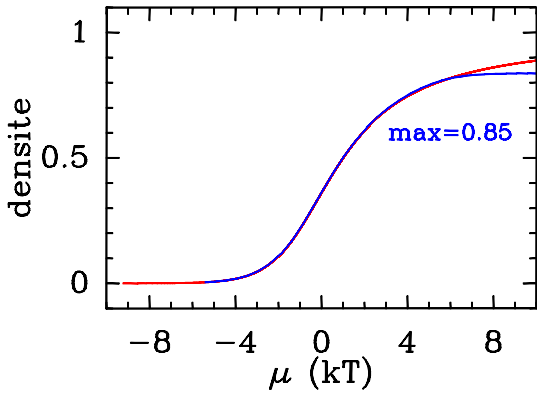


FIGURE 5.2 : *La densité de tiges rigides dans un profil homogène en fonction du potentiel chimique μ. Courbe théorique en rouge (tirée de l'équation 3.12), courbe numérique (Méthode de Vanderlick (Vanderlick* et al.*, 1986)) en bleu.*

5.2 $E = 0$, PROFIL HOMOGÈNE.

Dans un profil de potentiel homogène, tout dépend seulement du potentiel chimique.
In an homogeneous field, the chemical potential controls everything.

5.2.1 Densité dans un profil homogène

La densité dans un profil homogène augmente avec μ en passant par une transition brusque autour de $\mu = 0$.
The density in an homogeneous field increases with μ and goes through a transition around $\mu = 0$.

Dans le cas idéal du potentiel nul –qui trouve parfois un écho biologique relatif, voir figure 5.1 zone (E), où la séquence produit un potentiel quasiment plat sur 400 pb avec notre modèle énergétique–, la situation est simple, seul le potentiel chimique est un paramètre. l ne peut pas jouer puisque le résultat doit être invariant par translation. μ contrôle directement la densité moyenne, de façon monotone, mais pas linéaire : la variation de la densité en fonction du potentiel chimique dans un profil homogène est présentée en rouge sur la figure 5.2. Elle est tout simplement calculée à partir de l'équation d'état établie au chapitre 3 (équation 3.12).

Notons que l'implémentation numérique de l'algorithme de Vanderlick ne permet pas d'atteindre des densités moyennes très élevées, ce qui est traduit par le plateau correspondant à une densité maximale de l'ordre de 0.85 observée en bleu sur cette même figure. Cette saturation n'apparait que pour des densités *moyennes* élevées et, dans le cas de profils énergétiques non homogènes, il est tout à fait possible que localement, la densité atteigne des valeurs très élevées. Preuve en est que l'occupation peut valoir 1 à proximité d'un mur par exemple, comme on peut le voir sur la figure 5.7 (b) en vert.

En réalité, augmenter le potentiel chimique dans un profil homogène résulte en une cristallisation du système, c'est-à-dire que les particules sont toutes collées les unes contre les autres et la densité pour une configuration donnée ressemble à un peigne de Dirac. Comme il n'y a pas de conditions aux bords, la phase du peigne de Dirac n'est pas définie et il existe une infinité de ces configurations cristallines décalées les unes par rapport aux autres. Au final, ceci produit un profil de densité plat, alors qu'il s'agit bien de configurations cristallines prises indépendamment ; on parle alors de *Mode de Goldstone*. À faible potentiel chimique, le profil de densité est plat parce que peu de particules y sont présentes, et que celles-ci peuvent se fixer de manière équiprobable le long du profil énergétique plat. Lorsque le potentiel chimique augmente, le profil de densité reste plat parce qu'il existe un continuum de configurations cristallines décalées les unes par rapports aux autres.

Au final, la densité part de zéro aux faibles potentiels chimiques pour tendre petit à petit vers 1 en passant par une zone de transition située effectivement entre -4 et $+4\ kT$ (figure 5.2). Point important : la probabilité pour qu'une particule se fixe est guidée par $e^{\mu - E}$ (voir par exemple l'équation de Percus 5.1), c'est donc la différence entre le potentiel chimique et l'énergie qui est importante. De manière générale, on parlera de potentiel chimique sans préciser qu'il s'agit du potentiel chimique relatif à la valeur moyenne de l'énergie auquel on s'intéresse. C'est lorsque la densité est proche de 0.5 que la va-

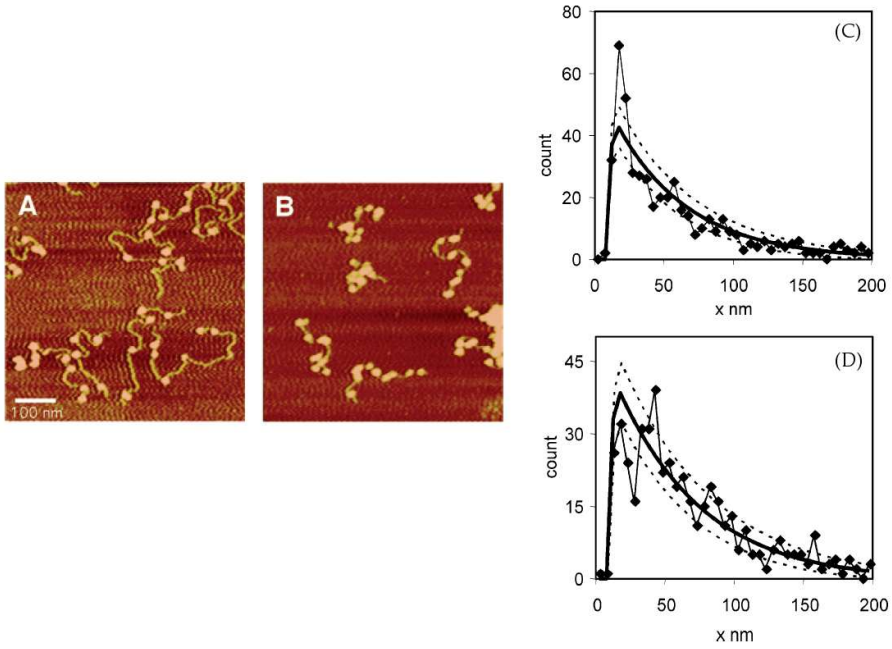


FIGURE 5.3 : *Expérience de mesure de la distance inter-nucléosomale. A et B visualisation par microscopie AFM de nucléosomes fixés sur des séquences MMTV (Mouse Mammary Tumor Virus) à différentes densités nucléosomales. (C) et (D) distribution des distances mesurées (génomique) entre nucléosomes (points), distribution aléatoire (Solis* et al.*, 2007) théorique en ligne continue noir. (C) l'ADN est peu acétylé, les nucléosomes sont plus souvent proches les uns des autres que dans le cas acétylé (D) ce qui suggère une coopérativité de l'accrochage nucléosomal. Figure adaptée de Solis et al (Solis* et al.*, 2007).*

riation de densité en fonction du potentiel chimique est la plus forte. On parle alors de susceptibilité pour traduire la sensibilité de la densité avec μ. Cette susceptibilité fait écho à la notion de *robustesse* présentée notamment sur la figure 5.26, au paragraphe 5.6.5 qui traduit la sensibilité des nucléosomes aux variations de potentiel chimique dans un profil inhomogène. On l'a vu, la forme de la transition peut être déterminée explicitement via l'équation d'état d'un fluide de Tonks-Takahashi (paragraphe 3.1.2), il ressort alors que la transition a lieu à un potentiel chimique de l'ordre de $\mu = 0 - ln(l)$. Dans le cas des nucléosomes de taille 147 paires de base la transition a donc lieu à des potentiels chimiques de l'ordre de $-5kT$. Ceci explique pourquoi nombre d'études en fonction de μ sur le nucléosome seront centrées sur $-ln(147) = -5\ kT$ (voir par exemple la figure 5.18 (a)).

5.2.2 Distance inter-nucléosomale dans un profil homogène

La distance entre deux nucléosomes adjacents dépend de l'interaction entre les deux particules.
The internucleosomal distance is a function of the interaction between two particles.

On peut parfois vouloir accéder à des paramètres statistiques plus sophistiqués que la densité et en tout premier lieu la distance entre deux particules. Cette distance dépend de l'interaction qui existe entre les particules. Biologiquement, cette distance correspond à la taille du *linker* qui sépare deux nucléosomes. Mesurer expérimentalement la taille du *linker* peut s'avérer difficile. Solis (Solis et al., 2007) propose de la mesurer par une observation par microscopie par force atomique, ou AFM (voir chapitre 5, paragraphe 7.2.3) et ses résultats son présentés sur la figure 5.3

Les expériences d'AFM permettent difficilement d'établir la véritable valeur de cette distance internucléosomale. De plus, la séquence influence localement la taille du *linker*, nous proposons donc une étude prenant en compte l'effet de la séquence en calculant directement la distribution de *linkers* associée à la séquence observée sous AFM (cette fois en phase liquide, expérience menée conjointement avec Pascale Milani, figure 5.4).

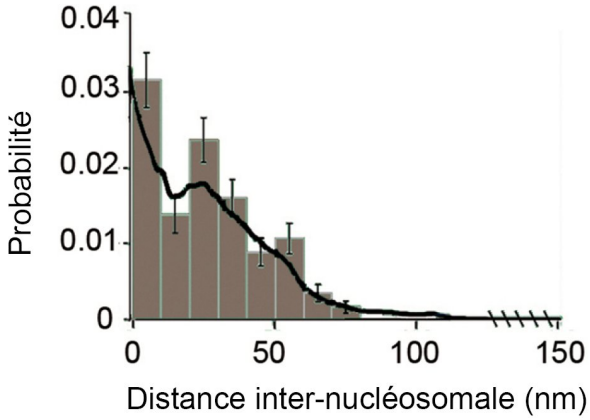


FIGURE 5.4 : *Distribution de la taille des linkers mesurés par AFM en phase liquide sur la séquence du gène YGR105W de la levure (Reconstitution in vitro de dinucléosomes (Milani* et al.*, 2009)). La courbe noire est une simulation calculée à partir du profil énergétique de la séquence sous-jacente (voir paragraphe 7.2.3).*

La distance typique qui sépare deux particules (la distance la plus probable qui sépare deux particules) est définie théoriquement comme la position du premier maximum de la fonction de paire. Puisque celle-ci correspond à la probabilité de trouver une particule à une distance donnée d'une autre particule, la fonction de paire permet bien d'établir la distance typique qui sépare deux particules adjacentes. Le problème avec cette définition est que pour des tiges rigides le premier maximum de la fonction de paire est systématiquement situé à la distance l (figure 3.2). Ceci signifie que le *linker* le plus probable est de taille nulle. Ceci est ennuyeux car au sein de la chromatine, les nucléosomes sont espacés régulièrement par un *linker* que l'on suppose être non nul (figure 5.3). La distance qui est mesurée dans les expériences classiques de biochimie ne correspond en fait pas à la position du premier pic de la fonction de paire (voir paragraphe 5.2.3). Toutefois dans l'expérience de Solis on mesure effectivement la probabilité de trouver la prochaine particule à une distance donnée. L'équation 7.1 propose une façon théorique de calculer cette probabilité dans un profil inhomogène avec un exemple en figure 7.6 (F). De plus la figure 5.3 (C et D) montre le premier mode de la fonction de paire dans un profil homogène. Il s'agit simplement de la fonction de paire, dans laquelle on a tronqué les oscillations, et gardé le premier pic. On sait toutefois que si l'on rajoute une répulsion sur une petite échelle entre particules, alors la fonction de paire change, et la distance la plus probable n'est plus strictement l, sans que les profils de densité associés ne changent qualitativement (voir figure 5.20). Le *linker* typique n'est plus de taille nulle, et varie selon la taille de la répulsion.

Pertinence biologique du NRL

La distance inter-nucléosome expérimentale est appelée "NRL", elle est évaluée par des expériences de biologie moléculaire (digestions enzymatiques), sans que cette mesure ne renvoie exactement la distance typique entre particules. Elle dépend beaucoup des organismes et des conditions expérimentales.

Experimentally, the internucleosomal distance is called the NRL, Nucleosome Repeat Length. It is measured through an enzymatic digestion procedure, that does not exactly yield the typical distance between nucleosomes. It depends on the organism it is measured in, and on the experimental conditions.

Lorsque la taille du *linker* est évaluée de façon globale par des méthodes biochimiques, on parle de NRL (Nucleosome Repeat Length). Tout comme pour la densité, il peut être intéressant de connaître le NRL, non pas sa valeur globale, mais localement en toute position génomique. Il faut alors exploiter des données de dinucléosomes. Cette donnée est d'autant plus importante que le NRL semble jouer un rôle non négligeable lors de la formation de la structure supérieure de la chromatine (Bednar et al., 1998; Mergell et al., 2004; Kepper et al., 2008).

Accéder expérimentalement à la distance entre nucléosomes peut se faire soit par visualisation directe comme dans la figure 5.3 (on mesure alors la Distance Typique entre Nucléosomes, DTN), soit indirectement par des expériences de digestion enzymatique (Woodcock et al., 2006; Blank and Becker, 1995; Noll and Kornberg, 1977) (mesure du NRL). Les méthodes biochimiques utilisent une faible quantité d'enzyme (MNase) qui découpe la chromatine préférentiellement au niveau des *linkers*. Les fragments

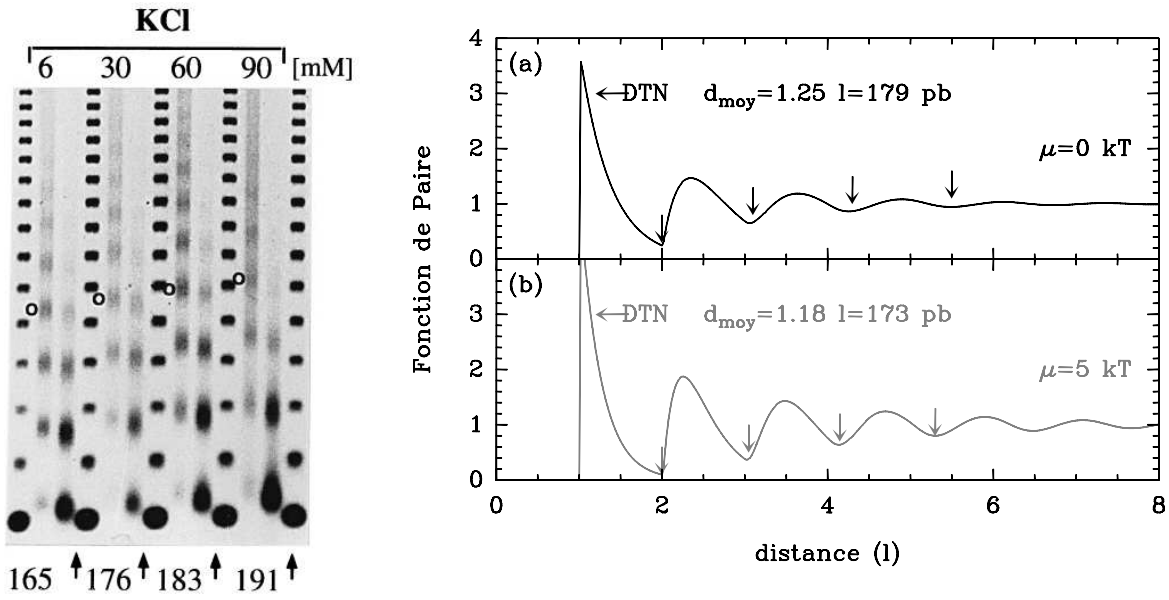


FIGURE 5.5 : *À gauche : Mesure du Nucleosome Repeat Length (NRL) par digestion MNase. Digestions dans différentes concentrations salines (KCL), Pour chaque concentration, un échelon (signalé par les flèches) à* 123 *paires de bases sert de référence et permet d'évaluer le NRL (en bas), deux migrations à des temps de digestion différents sont proposées (1 minute au milieu et 5 minutes à droite de chaque groupe). Figure adaptée de Blank et Becker (Blank and Becker, 1995). A droite : Ce qui est mesuré en réalité lors de la digestion enzymatique : la position des minima de la fonction de paire (donc la position des linkers, signalés par les flèches). Deux représentations à μ différents en (a) et (b) montrent deux mesures de –NRL– différents, alors que la distance typique (la position du premier pic, DTN) ne change pas.*

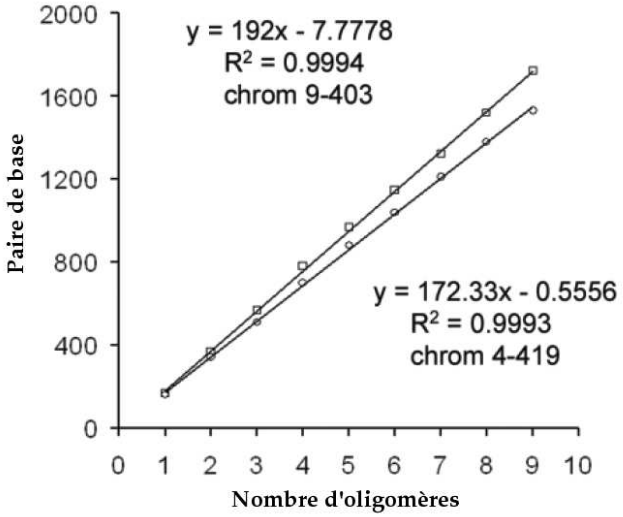


FIGURE 5.6 : *Détermination expérimentale du Nucleosome Repeat Length (NRL) à partir d'un gel. Les maxima d'intensité du gel de digestion (similaire à fig 5.5) sont représentés en fonction de la taille de l'ADN correspondant. Le premier point correspond à un nucléosome seulement. La valeur de la pente de la régression linéaire définit le NRL. Les résultats de deux gels différents sont représentés. Figure adaptée de Cioffi et al. (Cioffi* et al.*, 2006)*

d'ADN obtenus sont ensuite séparés selon leur taille par migration sur un gel (par électrophorèse), et comparés avec un échelon de référence (figure 5.5). Le résultat de la digestion peut donner des mononucléosomes (si la nucléase coupe deux *linkers* consécutifs), des dinucléosomes (un *linker* n'est pas coupé), des trinucléosomes, etc. Ces différents fragments migreront avec un vitesse différente selon la longueur des fragments. La distance qui sépare chacune des bandes sur le gel est interprétée comme une distance *linker-linker* moyenne. Ceci est tout à fait discutable : en effet, lorsque l'on récupère un fragment de taille L, cela signifie simplement que l'enzyme a coupé l'ADN à deux endroits espacés d'une distance L. Si l'on suppose que l'enzyme ne peut couper que dans les *linkers* alors ce que l'on mesure effectivement ce sont des fragments dont la longueur suit la loi de probabilité de la fonction de paire des *linkers*. On mesure de fait la probabilité qu'un *linker* se trouve à la distance L d'un autre *linker*. Il est naturel de penser que la fonction de paire des *linkers*, et la fonction de paire des particules est la même. Par conséquent, lorsqu'on "mesure" le NRL en évaluant la pente de la courbe obtenue à partir de la position des maxima d'intensité du gel (figure 5.5) en fonction de la distance, ce que l'on évalue réellement n'est pas la distance moyenne entre nucléosome, mais c'est bien la distance moyenne qui sépare les pics de la fonction de paire. Ce n'est pas la distance typique entre deux particules (DTN). Le NRL est nécessairement plus grand que la taille des particules.

Notons que cette distance varie manifestement en fonction de nombreux paramètres : acétylation de la séquence, organisme considéré, domaine dans le chromosome, mais aussi en fonction de la concentration saline, comme le montre justement la figure 5.5. Ceci suggère effectivement une interaction électrostatique, soit entre les histones et l'ADN, soit entre les nucléosomes entre eux.

5.2.3 Clarification sur les distances inter-nucléosomale d'intérêt

Selon les définitions, on peut distinguer trois distances intéressantes.

Depending on the various definitions of the distance between nucleosomes, we distinguish three different relevent distances : the NRL, the Pseudo-NRL and the Typical Distance between Nucleosomes.

A ce stade de l'étude, il convient de revenir sur les différentes définitions de distances inter-nucléosomales auxquelles on fera référence par la suite. On peut distinguer trois cas :

- **Distance Typique entre Nucléosomes (DTN, figure 5.5 (a)) :** Comme son nom l'indique, il s'agit de la distance la plus probable entre deux particules. Dans un modèle de sphères dures, c'est systématiquement la taille de la particule (147 pb). Il s'agit plus généralement de la position du premier pic de la fonction de paire. Cette distance est la plus significative d'un point de vue physique, puisqu'elle détermine l'environnement le plus probable d'un nucléosome.
- **Le "Nucleosome Repeat Length" (NRL) :** Il s'agit d'une mesure expérimentale, qui fait référence à la pente de la régression entre les maxima d'intensité d'un gel de digestion d'ADN par la MNase (qui préfère les *linkers*). Si la MNase préfère effectivement les *linkers*, les maxima du gel de digestion devraient correspondre avec les minima de la fonction de paire. Le premier maximum d'intensité doit correspondre au DTN (effectivement, sur la figure 5.6 les premiers points expérimentaux de chaque droite coïncident, ce qui confirme bien que le DTN est relativement constant). Cette mesure n'est pas la mesure d'une distance typique entre particule ! En tout état de cause, ce qui est mesuré ici est très similaire à la pseudo période décrite dans la figure 5.7 (c) en rouge. La pseudo période décrit effectivement la distance moyenne entre les pics de densité (mais aussi entre les *linkers*) à proximité d'un mur, or la densité à proximité d'un mur est exactement la fonction de paire. Pseudo période et NRL mesuré sont donc deux grandeurs similaires. On peut voir que cette pseudo période dépend fortement du potentiel chimique, jusqu'à ce que ce dernier atteigne une valeur suffisamment élevée. La pseudo période tend alors lentement vers son minimum. Le NRL n'est autre qu'une mesure expérimentale de cette pseudo période.
- **Le pseudo-NRL :** On peut enfin évaluer la distance qui sépare deux particules en moyenne en calculant l'autocorrélation du signal de densité (cette méthode ne marche que si les profils sont irréguliers). La position du premier pic de l'autocorrélation évolue de façon conjointe avec le DTN (voir 5.6.3). Cette distance est donc un intermédiaire entre le DTN et le NRL.

5.3 EFFET D'UNE BARRIÈRE : OSCILLATIONS LIÉES AU CONFINEMENT

Une barrière provoque des oscillations dans la densité qui s'étendent d'autant plus loin que le potentiel chimique est élevé.

A vertical wall produces nearby oscillation in the density that ranges according to the chemical potential.

5.3.1 Paramètres pertinents

Seul μ est véritablement important.

μ is the governing parameter.

Prenons maintenant le cas d'une barrière verticale infinie bordée par un potentiel plat sur un demi espace infini (figure 5.7). Ce problème a été abordé d'un point de vue théorique par Robledo et Rowlinson (Robledo and Rowlinson, 1986) (l'expression de la densité à proximité d'un mur est donnée par une série de $(x - k \cdot l)^k$, où x est la position, l est la taille des particules, et k le nombre de particules). Voyons comment cela se traduit d'un point de vue qualitatif. Une nouvelle fois, seul le potentiel chimique est véritablement un paramètre, puisque pour obtenir les résultats pour différentes valeurs de l il suffit de redimensionner les résultats par rapport au l de référence. La température β ne joue pas d'autre rôle que celui, implicite, de définir l'échelle de variation du potentiel chimique (μ est exprimé en kT, c'est donc $\beta\mu$ que l'on considère ici).

5.3.2 Oscillations

La présence d'une barrière fait apparaître des oscillations dans le profil de densité.

How a vertical barrier will produce oscillations.

L'un des effets les plus marquants issus de l'interaction de tiges rigides est le phénomène de positionnement statistique. On parle de positionnement statistique lorsque la position de plusieurs particules est imposée par la présence d'un élément fixe dans le potentiel, par exemple une barrière énergétique, ou bien une particule fixée par un puits local profond. La pression exercée par les particules voisines induit un confinement qui à son tour provoque le positionnement statistique. Comme on peut le voir sur la figure 3.4, l'effet de cette barrière est de provoquer des oscillations qui décroissent à mesure que l'on s'éloigne de la barrière. Ces oscillations sont d'autant plus prononcées que le potentiel chimique est élevé. La période de ces oscillations est directement reliée à la taille des particules : les oscillations sont légèrement plus grandes que l car la position de deux particules successives vaut au minimum l et la contrainte entropique veut qu'un espace minimum sépare deux particules en moyenne. D'un point de vue biologique, dans le cadre de l'étude de nucléosomes, cela signifiera que tout élément fixe dans l'ADN qui empêchera la formation d'un nucléosome induira automatiquement un effet de confinement à sa proximité. Sous la pression exercée par l'ensemble des nucléosomes fixés sur l'ADN, les particules ont tendance à se coller contre les éléments fixes. Bien évidemment des particules sous pression mais sans élément fixe se collent également contre les autres nucléosomes, mais comme leur position n'est pas systématiquement la même, le profil de positionnement ne présente pas d'oscillations.

5.3.3 Description phénoménologique

On peut décrire l'évolution de la portée de l'interaction simplement.

On peut essayer de décrire phénoménologiquement la forme de ces oscillations en introduisant une pseudo période des oscillations l_0 (l_0 correspond également au NRL expérimental) ainsi qu'une distance typique de décroissance λ. La densité elle-même présente des discontinuités très prononcées, et il sera difficile de l'approximer par des fonctions usuelles. Mais si l'on prend en compte la probabilité d'occupation (la densité convoluée par une fenêtre carrée de taille l) ou mieux encore la densité lissée par une fenêtre gaussienne, alors il devient très facile de réaliser un "fit" des signaux par des sinusoïdes amorties du type $e^{-\frac{s}{\lambda}} cos(\frac{2\pi s}{l_0})$ où λ est le coefficient d'atténuation, l_0 la période de la sinusoïde, figure 5.7 (b). Bien qu'il soit possible de déterminer une solution explicite pour la forme des oscillations de confinement à proximité d'une barrière infinie (Robledo and Rowlinson, 1986), ces solutions ne donnent pas d'expression transparente pour la période des oscillations ni sur la décroissance de ces oscillations. L'idée est

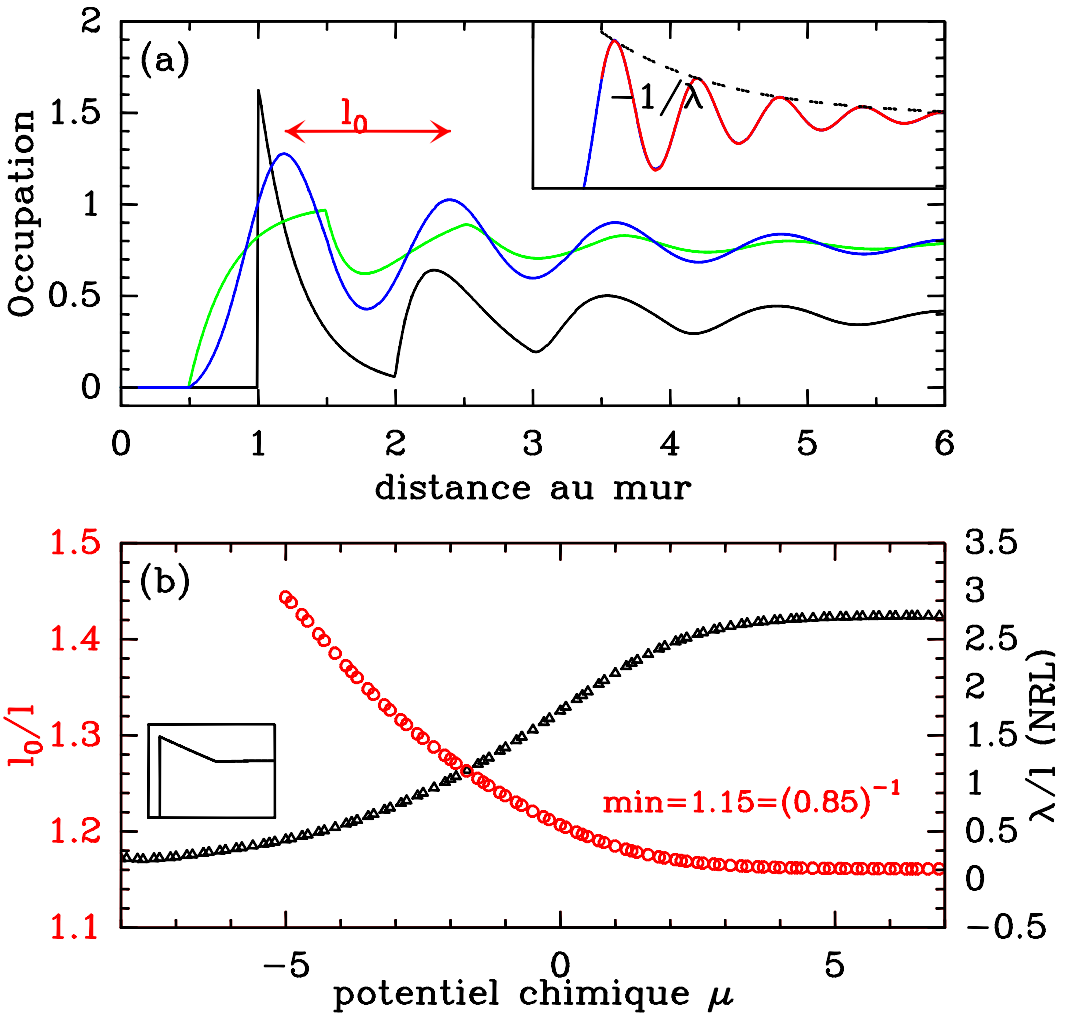


FIGURE 5.7 : *(a) Profils à proximité d'un mur infini : en noir la densité, en vert l'occupation, en bleu la convolution de la densité par une fenêtre gaussienne. En insert : comparaison entre la convolution (en bleu) et le fit sinusoïdal (de longueur d'onde l_0 atténué (en rouge), décroissance exponentielle λ en pointillé). (b) Évolution de l'atténuation de l'exponentielle (en noir) et de la longueur d'onde l_0 du fit en fonction du potentiel chimique. toutes les longueurs sont exprimées en unités naturelles (l). La zone de potentiel chimique inférieur à $-5kT$ correspond à la zone ou le fit devient difficile puisque la densité n'oscille plus ou très peu (profil de densité dans cette zone présenté en insert).*

donc ici de gommer les variations anguleuses de la densité à proximité du mur en filtrant avec une gaussienne, afin d'obtenir un signal dont la forme est très proche du sinus amorti, de manière à obtenir une expression phénoménologique de la période des oscillations. Il devient donc possible d'établir numériquement la dépendance de la période des oscillations ainsi que celle de l'amortissement avec le potentiel chimique (figure 5.7). Conformément à l'intuition, on observe qu'une augmentation du potentiel chimique induit une diminution de la distance entre deux particules mesurée par l_0 jusqu'à tendre vers une valeur proche de l. L'espace restant ne peut être comblé sans contraindre l'entropie du système. Cela correspond bel et bien au fait que μ contrôle la densité, et donc le nombre moyen de particules implantées dans le système. Ce nombre croissant induit une augmentation de la pression et rapproche les particules les unes des autres, tout en exacerbant l'effet positionnant sur la première des particules, et par propagation de l'effet positionnant, des nombreuses particules suivantes. L'amortissement de son côté, se fait sentir sur une distance de plus en plus grande lorsque l'intensité du confinement augmente. Cela correspond au fait que la propagation de proche en proche de l'effet positionnant est d'autant plus efficace que les particules sont proches.

5.3.4 Conclusion, prédictions

La portée maximale de l'effet d'une barrière est de l'ordre de $5-7$ distances particulaires. L'ajout d'un bruit détériore le signal.

The boundary condition will produce an effect that has a range of at most $5-7$ particles in typical nucleosomal conditions. Adding noise, that is sequence contribution, will rapidly blur the positioning signal.

Le résultat essentiel de cette étude est que l'effet positionnant d'un obstacle possède une portée finie dont la valeur dépend fortement du potentiel chimique μ. Si l'on définit la portée maximale comme, à un potentiel chimique très élevé ($> +10kT$), la distance à laquelle les oscillations provoquées par la présence de la barrière ont été diminuées d'un facteur 10, alors cette portée vaut $\lambda_{\max} ln(10) \approx 6.3\ l$ pour les valeurs maximales de μ qui nous intéressent (les valeurs de μ attendues *in vivo* et *in vitro* pour les nucléosomes sont situées entre -6 et $+2\ kT$, voir chapitre 6). Si l'on est moins strict sur le terme de décroissance, amplitude divisée par 100, alors la portée maximale est d'environ $12\,l$. Dans le meilleur des cas (pas d'autres évènements dans le potentiel que la barrière en question, *i.e.* le potentiel est plat ailleurs que sur la barrière), l'influence d'un objet ponctuel ne pourra donc se faire sentir qu'à une distance de l'ordre de la dizaine de particules. La saturation de la portée de l'interaction observée sur la figure 5.7 (b) est à mettre en doute. En effet, comme on l'a vu sur la figure 5.2, il y a une certaine forme de saturation à très haut potentiel chimique ($> +8kT$) ou bien lorsque la densité devient plus grande que 0.85. Le palier observé vient d'un problème numérique. Il faut toutefois se rendre compte que les valeurs de pression nécessaire pour augmenter la densité deviennent drastiques dès qu'on dépasse des densités de l'ordre de 0.8 justement, cela se traduit naturellement par des valeurs très élevées pour les intégrales calculées dans l'algorithme de Vanderlick. Il faut donc s'attendre à ce que la portée maximum réelle ne suive pas la courbe numérique (figure 5.7 (b), noir), et de même, il faut s'attendre à ce que la période minimum tende vers 1 et non 1.15 dans la figure 5.7 (b) en rouge. Ce résultat reste important pour nous, car il définit la portée effective de l'influence d'une protéine ou d'un objet positionnant quelconque dans la fibre d'ADN.

5.3.5 Pertinence biologique

On peut observer l'effet positionnant en moyennant autour d'une position que l'on sait inaccessible.

The positioning effect of a boundary element can be accessed through the averaging of density profiles next to them.

Biologiquement, cette portée d'interaction est étudiée en détail dans le cas particulier de la protéine régulatrice humaine CTCF dans un article de Fu *et al* (Fu et al., 2008). CTCF est une protéine (un insulateur) qui se lie très fortement à l'ADN induisant un très fort positionnement des nucléosomes voisins (figure 5.8). Même si l'environnement du site de fixation de CTCF n'est pas homogène, une moyenne sur l'ensemble des sites de fixations de CTCF permet d'évaluer la portée expérimentale d'un objet fixe dans le génome. La figure 5.8 présente les résultats de digestions MNase effectuées sur l'ensemble de ces sites de fixation. L'origine est donnée par la séquence consensus de fixation de CTCF, et on a représenté séparément les "Tags" de séquences sens en bleu, et les "Tags" de séquence antisens en orange. Ces Tags permettent respectivement de situer la partie gauche et la partie droite de chaque nucléosome (en

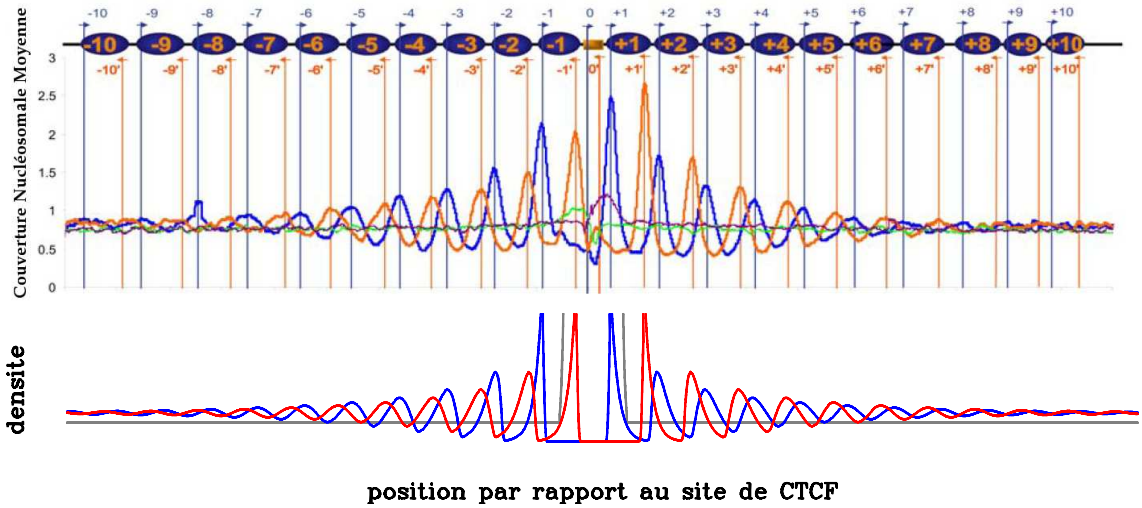


FIGURE 5.8 : *En haut : Signaux génomiques moyens autour des sites de fixation de CTCF. La position du site de fixation est indiquée par la boîte orange. Les résultats de la MNase-seq sont présenté pour le brin sens (en Bleu) et antisens (en orange). Le contrôle en l'absence de CTCF est représenté en Violet pour le brin anti-sens, et en vert pour le brin sens. Figure adaptée de Fu et al (*Fu et al.*, 2008). En bas : modélisations (rouge et bleu) avec un profil énergétique plat, obstrué par un obstacle positionné en zéro (en gris), laissant une épaisseur* 240 *pb inaccessibles.*

moyenne) (Fu et al., 2008). Tout comme dans notre modélisation, les nucléosomes voisins de la barrière énergétique créée par CTCF sont positionnés d'autant mieux qu'ils sont proches de la barrière. Ils sont espacés en moyenne de 185 paires de bases, ce qui est compatible avec les périodes prédites pour des potentiels chimiques de l'ordre de $\mu = -2kT$ (1 nucléosome tous les 200 pb environ). À comparer à $146 \cdot 1.3 \approx 190$ paires de bases, valeurs expérimentale du repeat. Le positionnement des nucléosomes présente également une décroissance exponentielle et la portée de l'effet positionnant s'étend également sur une distance de l'ordre de quelques nucléosomes, ce qui est une nouvelle fois compatible avec notre modélisation. En guise de contrôle négatif, les tags sens (en violet) et antisens (en vert) de sites inoccupés par CTCF sont représentés également.

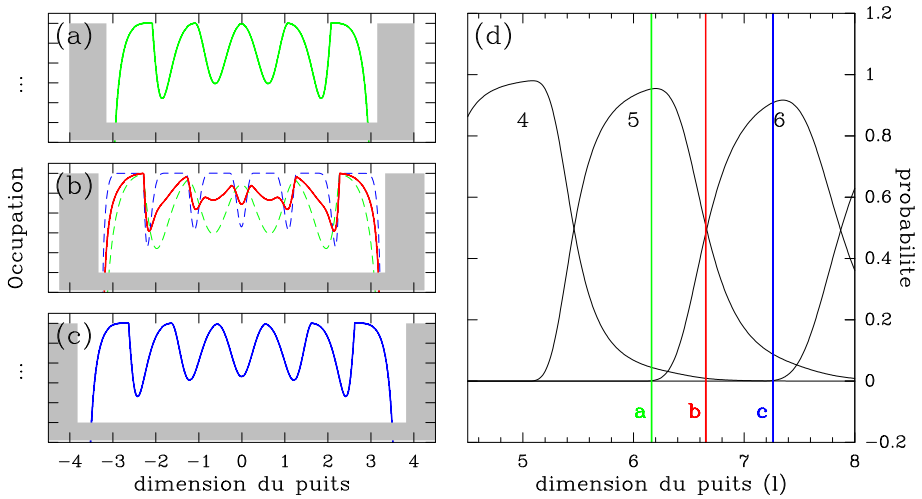


FIGURE 5.9 : *Confinement dans une boîte (potentiel plat bordé par des murs infinis). (a) Occupation dans une boîte suffisamment grande pour accueillir* 5 *particules. (b) Compétition entre configurations à* 5 *et à* 6 *particules (pointillés) qui résulte en un profil non structuré (en rouge) en grand-canonique. (c) La boîte s'agrandit encore, la configuration à* 6 *particules domine. (d) probabilités associées aux configurations à N fixé pour un potentiel chimique donné ($\mu = +4kT$). Les barres verte, rouge et bleue correspondent aux tailles de boîte utilisées dans les encadrés (a), (b) et (c). La situation rouge correspond bien à une situation équiprobable entre* 5 *et* 6 *particules.*

5.4 PARTICULES COINCÉES DANS UNE BOÎTE

Les profils de positionnement dans une boîte sont quantifiés.
Particles in a well will produce quantized configurations.

5.4.1 Profils de densités

Confinés dans une boîte, les particules n'ont qu'un petit nombre de configurations autorisées.
Confined, the particles will only adopt a few configurations.

Le problème devient légèrement plus complexe lorsque deux barrières deviennent proches : le phénomène de positionnement statistique qui apparaissait à proximité d'une barrière seule se révèle d'autant plus, puisqu'il y a des conditions aux limites imposées des deux côtés du système. En termes de paramétrisation, cette fois il n'est plus possible d'ignorer la taille l des particules : le rapport L/l où L est la distance qui sépare les deux barrières devient la grandeur pertinente. Comme précédemment, on peut oublier le paramètre β puisque les barrières sont de taille infinie. Les paramètres d'importance sont donc μ/β et L/l, que l'on notera simplement μ et L par simplicité dans le reste du manuscrit. Considérons d'abord une situation où N, le nombre de particule, est fixé : augmenter L permet de libérer de plus en plus d'espace, et le nombre de configurations augmente rapidement avec la taille accessible. D'un autre côté, si l'on rajoute une particule, le système gagne une énergie μ, encore faut il qu'il y ait suffisamment de place pour accueillir la particule supplémentaire. En grand-canonique, il existe donc une compétition entre configurations à N différents. Cette compétition tourne à l'avantage des N les plus grands quand L augmente. (figure 5.9). Lorsque deux configurations *canoniques* sont équiprobables (figure 5.9 (d) en $l = b$), le profil *grand-canonique* résultant peut ne plus être structuré (c'est-à-dire ne plus présenter des oscillations bien prononcées, figure 5.9 (b)) quand bien même les configurations dont il résulte présentent de grandes oscillations. Mais le déphasage entre chacune des configurations joue de telle sorte que la somme des profils est non structurée. Cela ne se produit évidemment que si au moins deux configurations ont une probabilité non négligeable, ce qui n'arrivera que sur des toutes petites fenêtres de longueur L pour un haut potentiel chimique. On parle alors de fenêtre de bistabilité.

5.4.2 Évolution des probabilités de chacune des configurations "canoniques" (N fixé) en fonction de μ et de L

Selon la taille accessible et dans une moindre mesure, selon μ il sera probable que N, $N+1$, etc particules se fixent dans la boîte.

According to the size available, and according to μ, N or $N+1$ particles may enter the well.

La coexistence de différentes configurations dépendra des deux paramètres μ et L, mais dans le cas où le potentiel chimique est élevé, les dimensions pour lesquelles la coexistence est possible sont restreintes. Sur la figure 5.10 (a) (b) et (c) il apparaît que :

- (i) Si l'énergie apportée par une nouvelle particule est grande, c'est la configuration avec l'occupation maximale pour une longueur donnée qui domine, les transitions sont rapides et au maximum deux configurations coexistent.
- (ii) Si l'ajout d'une particule est neutre en terme énergétiques, c'est le nombre de configurations (la fonction de partition à N fixé) seul qui avantage les configurations à grand N. Les transitions sont plus lentes, mais il faut aller très loin en longueur de boîte (ou puits) pour que plus que deux configurations puisse coexister.
- (iii) si l'ajout de particule est pénalisé, les transitions sont très lentes. Un grand nombre de configurations coexistent quelque soit la longueur du puits considéré.

Dans la suite, on fera référence à ces zones en tant que zones *bistables* ou alors *multistables* pour signifier l'existence de plusieurs configurations acceptables. Ces zones bistables sont espacées régulièrement quand L augmente. Le potentiel chimique gouverne l'épaisseur de ces transitions et le nombre de configurations qui peuvent coexister. Afin de représenter l'évolution des profils de densité à l'intérieur de la boîte quand L augmente à un μ donné, on peut tracer une carte 2D tractant l'évolution du positionnement avec L (figure 5.10, les formules théoriques permettant d'établir ces résultats sont présentés dans le chapitre 3). Puisque la taille du puits accessible augmente de haut en bas, le nombre de particules qui remplissent le puits (en blanc, plus le blanc est intense, plus la particule est localisée) passe successivement de 1 à 4 particules lorsque le potentiel chimique est élevé (figure 5.10 en haut à droite). Mais lorsque le potentiel est plus faible, on retrouve des transitions plus floues, et surtout plus larges.

5.4.3 Effet sur la distance entre deux particules.

Agrandir un puits permet l'ajout discontinu de particules ce qui conduit à des variations brutales de la taille des linkers dans le puits.

Increasing the size available will strongly affect the internucleosomal length.

Un effet secondaire attenant à l'agrandissement de l'espace accessible entre deux barrières est la variation de la distance qui sépare deux particules dans le puits. Prenons un exemple simple : si une boîte est suffisamment grande pour accueillir deux particules tout juste, la distance qui sépare ces deux particules est nulle. Si l'on agrandit cette boîte, la distance qui les sépare augmente jusqu'à ce qu'il y ait suffisamment de place pour accueillir une nouvelle particule. Cela arrivera quand l'espace qui sépare les deux premières particules est égale à l. Mais à mesure que l'on agrandit l'espace accessible, il est possible de rajouter une particule à chaque fois que L augmente de l, cet espace l libre juste avant l'apparition d'une nouvelle particule, est distribué sur les particules existantes. Par conséquent, plus le nombre de particule déjà présent est grand, moins l'agrandissement n'aura d'influence sur la distance qui sépare deux particules. Sur la figure 5.11, cela correspond au fait que l'amplitude de variation de la taille du *linker* diminue à mesure que la taille du puits augmente. L'espace qui sépare deux particules dans un tout petit puits est susceptible de changer quand le puits s'agrandit, tandis que pour un puits déjà long, l'ajout d'une nouvel objet n'a quasi pas d'incidence sur la distance typique entre deux particules au sein du puits. Cette observation somme toute relativement triviale trouvera un écho inattendu lors de l'étude de la structure nucléosomale au sein d'un gène (chapitre 9, figure 9.10). En effet, l'espace qui sépare deux nucléosomes, le *linker*, joue un rôle important dans le niveau de transcription des gènes. Nous reparlerons plus en détail de cet effet dans le chapitre 9.

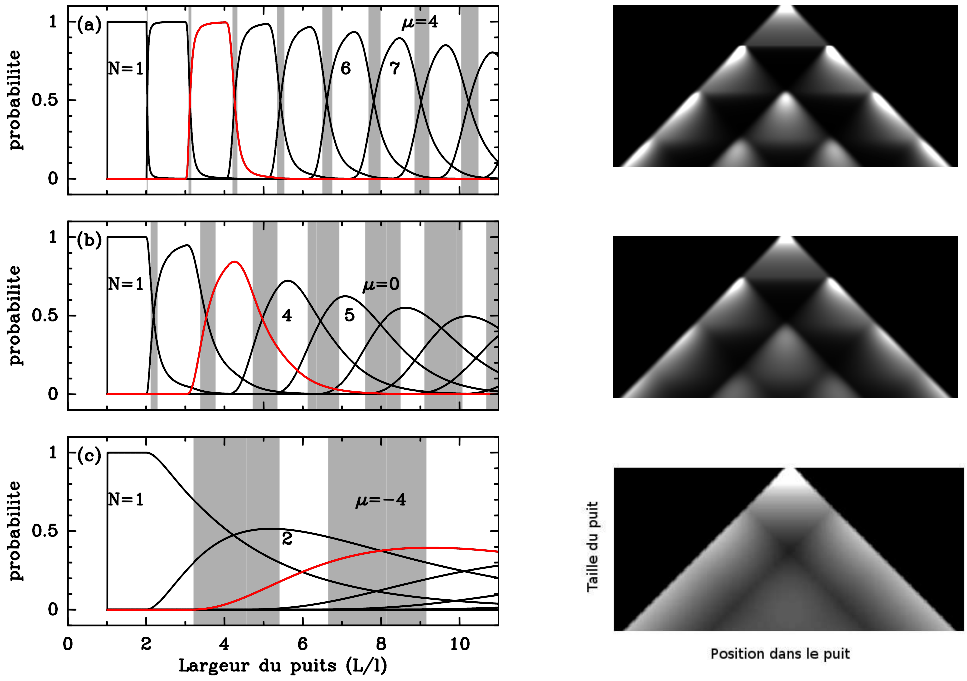


FIGURE 5.10 : *Évolution des probabilités de configurations canoniques (N fixé) en fonction du potentiel chimique. À gauche : (a) $\mu = +4kT$, (b) $\mu = 0kT$, (c) $\mu = -4kT$. Les zones grisées correspondent à des tailles de puits telles que au moins deux configurations ont une probabilité supérieure à 0.3. Les courbes rouges correspondent à la probabilité de trouver $N = 3$ particules dans le puits pour une longueur L/l donnée. A droite : évolution de la densité (codée en niveau de gris relatifs, le blanc correspond à une forte densité, le noir à une densité nulle ou très faible) en fonction de la taille de la boîte pour chacun des potentiel chimiques $(4, 0, -4)$. La taille augmente de haut en bas, et passe de 0 à $4.5\ L/l$.*

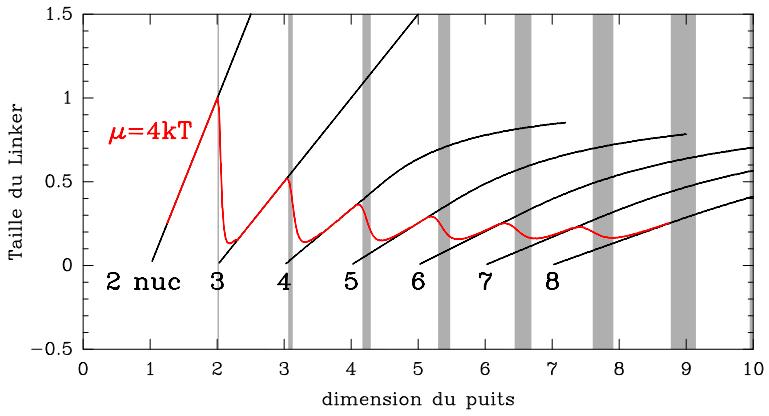


FIGURE 5.11 : *Évolution du linker, ici défini comme la distance qui sépare la deuxième de la première particule dans le puits, en fonction de la taille du puits. En noir sont représentées les distances associées à chaque configuration canonique (N fixé), en rouge le linker moyen calculé en pondérant le linker de chacune des configurations canoniques (N fixé) par le poids associé à la probabilité grand-canonique ($\mu = +4kT$) (figure 5.10 (a)).*

5.5 CONFINEMENT PAR UNE FORCE

L'interaction avec un facteur de transcription est mieux décrit par une force répulsive que par un blocage purement stérique (chapitre 6). L'effet de confinement est moins violent, mais tout à fait similaire aux barrières verticales.

Since a boundary element will often correspond to some kind of protein, such as transcription factors, the confinement effect is better described by a force applied on the edge rather than a steric hindrance.

Imposer une barrière verticale correspond à limiter l'espace accessible aux particules. Il arrive plus souvent que ce soit une force qui soit appliquée aux bord plus qu'une véritable contrainte sur la position. Même si une particule impose un encombrement stérique, le fait que le nucléosome respire conduit à une situation où la dyade peut s'approcher à une distance inférieure à 73 pb de l'objet stérique. On modélise très bien ce phénomène par une force de répulsion. On peut donc se poser la question de l'influence d'une barrière non verticale sur les bords : pour modéliser une force, il suffit de placer des barrières en forme de rampe, la force appliquée étant la pente de la rampe en question. Biologiquement cela pourrait correspondre à la présence d'une protéine dont l'interaction avec les nucléosomes est plus que la simple contrainte stérique. Il peut aussi s'agir tout simplement d'une assemblée de particules mobiles, qui ensemble, créent une pression. Ainsi, les nucléosomes eux mêmes créent une pression sur leur voisins, qui se modélise à 1D comme une force. A ces échelles, l'amplitude typique des forces mise en jeu par des moteurs moléculaires par exemple est de l'ordre du picoNewton (Mihardja et al., 2006; Ladoux et al., 2000; Bennink et al., 2001; Claudet et al., 2005; Kruithof et al., 2009). Puisque deux paires de bases sont espacés d'environ $0.36nm$, cela signifie que l'ordre de grandeur des pentes que l'on peut introduire doit être autour de $0.1kT.nm^{-1}$ soit environ $5\ kT$ par nucléosome. On peut généraliser le problème en introduisant des formes de barrière quelconques, mais par soucis de simplicité, on ne s'intéresse d'abord qu'aux rampes. Du coup, un nouveau paramètre intervient : l'intensité de la force appliquée, ici désignée par $f = \frac{\Delta E}{l}$ où ΔE désigne la variation d'énergie sur l'échelle l. Globalement, la situation est inchangée, dans le sens où μ permet toujours de peupler plus avant le puits de potentiel, la longueur du puits détermine toujours le nombre de particules qui peuvent y accéder, et la taille du premier *linker* est toujours une fonction décroissante et légèrement oscillante de la taille du puits. Toutefois, le fait qu'il devienne possible de *mordre* sur la barrière, ou, en termes énergétiques, le fait que la première particule ait la possibilité de se fixer sur une partie défavorable de la barrière pour permettre une relaxation des particules voisines, induit des transitions plus douces, et ce d'autant plus que la force appliquée est faible. Certes la particule en question perd de l'énergie, mais du point du vue du système complet, le gain entropique compense cette perte.

5.5.1 Évolution de la densité avec le potentiel chimique μ

L'ajout de particules à proximité de la force répulsive induit un positionnement statistique relaxé par rapport à une barrière verticale.

The effect of a force or pressure instead of steric hindrance is smoother.

Lorsque le potentiel chimique augmente, encore une fois, le paramètre qui influe de façon majeure sur la forme du profil est le potentiel chimique. Cela est dû au fait que le potentiel chimique détermine la zone du potentiel qui est accessible puisque le poids statistique associé à une position s donnée dépend de $e^{\beta(\mu - E(s))}$. En conséquence, augmenter le potentiel chimique dans un puits de potentiel dont les côtés sont des pentes, fixées par l'intensité de la force appliquée, revient à agrandir la zone accessible, tout en défavorisant le positionnement sur les côtés. Lorsque la pente est faible (figure 5.12), on retrouve effectivement le même type d'évolution que dans la figure 5.9 (a),(b),(c) ou la figure 5.10 (droite), où c'est la taille du puits qui augmente. Lorsque le confinement est important –si la force est importante– (figure 5.12 (b)), il faut monter à des potentiels chimiques plus élevés pour pouvoir peupler ne serait-ce que les abords immédiats du fond plat. L'évolution avec μ est ici beaucoup plus proche de ce qui se passait dans le cas des barrières infinies (et c'est normal, à la limite $f \to \infty$ la barrière est verticale) avec une simple apparition d'oscillations dont le nombre N dépend uniquement de L. Sur les bords apparaissent des oscillations supplémentaires, qui correspondent à des configurations à $N+1$ voire $N+2$ particules fortement défavorisées par la nécessité de *mordre* haut en énergie sur la barrière pour les particules des bords.

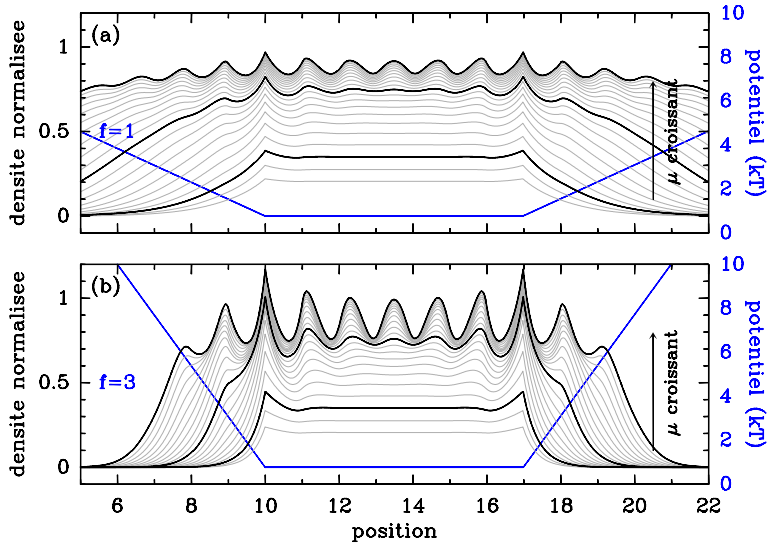


FIGURE 5.12 : *Évolution de la densité de particules confinées par des forces d'amplitude fixées : (a) $f = 1\,(kT).l^{-1}$ la force est de faible intensité, le potentiel correspondant est représenté en bleu. En noir sont représentés les courbes de densité pour un potentiel chimique de valeur $-4, 0$ et $5\,kT$, courbes intermédiaires en gris. (b) Similaire à (a) sauf que la force de confinement est plus intense ($f = 3\,kT.l^{-1}$)*

5.5.2 Évolution de la densité avec la largeur du confinement/taille des particules L/l

On observe des zones de transitions $N \to N + 1$.

L'évolution avec la taille du confinement est très similaire à ce qui se passe dans la cas de barrières infinies (figures 5.9 et 5.10). De légères différences apparaissent toutefois : d'abord les transitions sont plus douces, dans le sens où le passage d'une configuration de N particules à une configuration de $N+1$ particules intervient sur un domaine de variation de L plus large (figure 5.13 (d)). Les transitions apparaissent pour les mêmes longueurs de confinement, et l'épaisseur de chacun des profils de probabilité est sensiblement la même que pour les murs infinis. L'effet de parking est lui aussi adouci : la proximité d'une force répulsive provoque également l'apparition d'un positionnement de la première particule adjacente, mais ce positionnement est moins violent, la forme de la première oscillation de probabilité de positionnement est moins anguleuse (figure 5.13 (a),(b),(c)).

5.5.3 Évolution de la densité avec l'intensité du confinement f

Plus la force augmente, plus on se rapproche de la configuration "stérique".
As the pressure applied increases, the situation tends toward the steric hindrance problem.

Comme cité plus haut, l'évolution de la densité avec l'intensité du confinement revient à diminuer l'espace accessible pour les particules. On construit le potentiel en fixant l'espace où le profil énergétique est plat, on rajoute des barrières en forme de pente sur chaque côté dont le pied abouti toujours au même endroit (en $+2$ et -2 de la figure 5.14). Augmenter la pente correspond bien à diminuer l'espace sur lequel la grandeur $\mu - E(s)$ est inférieure à une constante par exemple. De fait, il est donc naturel de retrouver le même type d'évolution que ce qui a été vu pour l'évolution de la densité avec L dans le cas de barrières infinies (figure 5.10). Augmenter f diminue l'espace accessible, et provoque des transitions dans le nombre des oscillations du profil. À mesure que la force augmente, les particules qui résidaient sur les pentes sont éjectées. Les oscillations situées sur les bords du profil bleu de la figure 5.14 correspond à cette situation (force relativement faible, de l'ordre de $1kT.l^{-1}$). Lorsque la force augmente, on éjecte cette particule, et on finit par obtenir le profil rouge, sans oscillations sur les bords, car tout est confiné à l'intérieur strict du puits. Notons que lorsque la force est de faible intensité, l'amplitude des oscillations est très faible ce qui suggère très peu de positionnement dans ce cas, alors que lorsque la force dépasse les quelques $kT.l^{-1}$, l'amplitude des oscillations coïncide avec les amplitudes générées

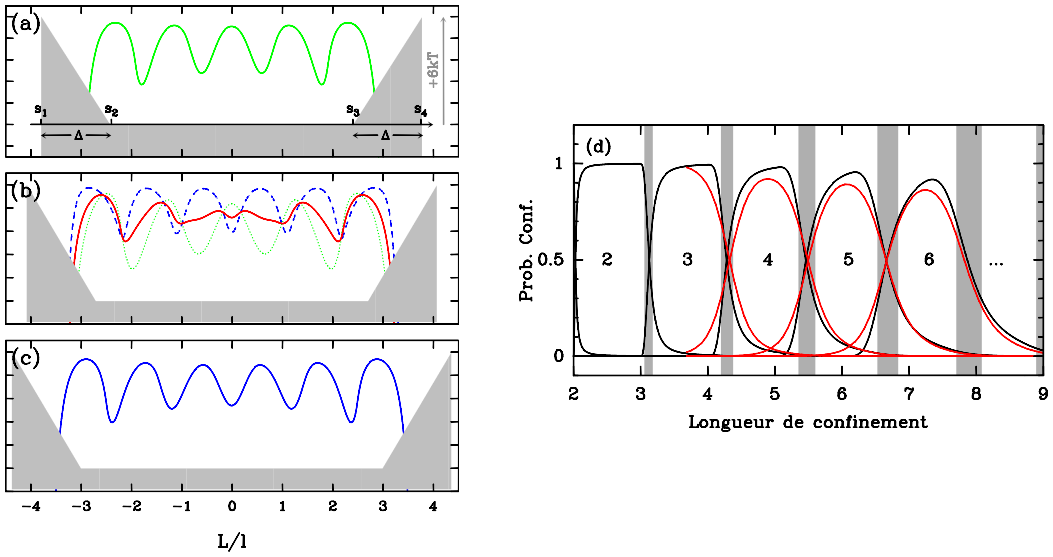


FIGURE 5.13 : *Influence de la longueur de l'espace confiné sur les profils de densité (a),(b) et (c). (d) probabilité de chacune des configurations à N particules fixés dans le cas de force appliquées aux bords (en rouge) vs barrières infinies (en noir). Le potentiel chimique est fixé à $+4kT$. La longueur de confinement est facile à définir dans le cas des barrières infinies, c'est la distance entre les murs. Dans le cas des forces, on définit la longueur de confinement comme la distance entre les pieds des pentes énergétiques à laquelle on rajoute la distance nécessaire pour que l'énergie de la barrière soit égale à $\mu = +4kT$. La force de confinement vaut $+6kT.l^{-1}$*

par des barrières verticales.

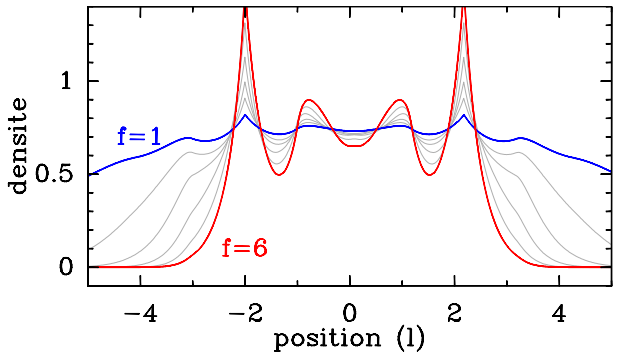


FIGURE 5.14 : *Influence de l'intensité de la force sur le positionnement : en bleu, densité dans le cas d'une force faible ($f = 1\,(kT).l^{-1}$), en rouge, densité dans le cas d'une force forte($f = 6\,(kT).l^{-1}$). En gris, densité pour des forces intermédiaires ($f = 2, 3, 4, 5\,(kT).l^{-1}$) L'intensité du confinement détermine à la fois l'espace accessible ainsi que l'amplitude des oscillations dues au confinement.*

5.5.4 Et en réalité ?

Le confinement entre deux barrières a donc pour principale conséquence de quantifier les états accessibles pour le nombre de particules. Seuls quelques modes sont autorisés, et il en résulte un profil de densité oscillant. Si l'on reprend les profils réels que l'on présentait sur la figure 5.1 sur *C. elegans* et si l'on calcule directement la probabilité d'occupation (avec un choix de paramètres similaires à ce qui sera utilisé dans la levure (voir Chapitre 6), on retrouve les grandes lignes de ce qui a été présenté durant ce chapitre.

- Lorsque le profil est plat (figure 5.15 E), la densité résultante est plate, et le léger bruit perturbe la densité et donne des oscillations proportionnelles au bruit de la séquence.
- À proximité d'une barrière (figure 5.15 E_1 et E_2 toujours), des oscillations nettes et décroissantes avec la distance sont observables.
- Confinés dans un puits (figure 5.15 A, C, C',C''), les nucléosomes adoptent des profils bien oscillant (C'') ou "fuzzy" (C) selon que la taille accessible permet à plusieurs configurations de coexister ou non.

5.6 PROFILS ÉNERGÉTIQUES ALÉATOIRES

Dans un potentiel aléatoire, le densité moyenne est déterminée par un couple de valeurs (δ, μ). L'échelle de variation du potentiel joue également. Le pseudo-NRL est déterminé par la pression exercée par les particules, et par la pression intrinsèque imposée par la séquence (δ).

In random profiles, the mean density is defined by a set (δ, μ). The scale of varaition bares definitely some importance. The pseudo-NRL is defined by the pressure exerted by the particles as well as by the pressure naturally induced by the sequence.

Précédemment nous avons illustré le positionnement statistique, à savoir l'émergence d'une périodicité dans l'organisation linéaire du chapelet nucléosomal induite par la présences de barrières energétiques dont l'effet est de confiner les nucléosomes. Comme on le verra plus tard ces barrières sont très souvent *in vivo* le fait de facteurs extrinsèques qui se lient à l'ADN comme dans le cas CTCF illustré plus haut. Comme discuté précédemment pour C. elegans (Fig. 5.15), et comme on le verra plus tard au niveau des gènes de la levure, les profil énergétiques "intrinsèques", calculés à partir des modèles présentés au chapitre 4, peuvent effectivement présenter des barrières énergétiques contribuant ainsi à un positionnement statistique. Cependant, comme l'indiquent les profils énergétiques intrinsèques calculés le long du génome de la levure et C. elegans (Fig. 5.16) par exemple à partir du modèle "Pnuc", la topographie de ces profils est plutôt désordonnée, ressemble donc à un bruit. La distribution statistique des valeurs de cette énergie intrinsèque est en première approximation gaussienne caractérisée par une variance δ^2 (Fig. 5.16(b)). Ce désordre résulte directement du désordre "génomique". L'objectif ci-après est de caractériser l'impact d'un tel désordre génomique sur l'organisation du chapelet. On voit bien-sûr ici qu'un paramètre important est la variabilté (mesurée par δ) de ces profils énergétiques qui va effectivement déterminer dans quelle mesure la séquence contribue au positionnement des nucléosomes : plus la séquence induit des barrières et/ou des puits importants plus l'effet de confinement (local et non local) sera grand. On s'interessera donc à l'influence de δ ("force" du bruit) dans le cas du génome de la levure (bon exemple de genome ou le profil apparait essentiellement desordonné) ; pour étudier l'influence des corrélations à longue portée on s'interessera également à l'étude de séquences artificielles non corréles ($H = 0.5$) et corrélées à longue portée ($H = 0.8$).

5.6.1 Statistiques : amplitude δ de variation

Une question : qu'est ce qui contrôle la variabilité δ des profils énergétiques intrinsèques ? D'après le modèle énergétique décrit au chapitre 4, notamment celui utilisant la table "PNuc", on peut en première approximation considérer que les fluctuations de cette énergie vis-à-vis de la séquence sont données par une relation du type de celle obtenue pour les boucles : $\delta/kT \sim \frac{A_2}{2}\rho_{nuc}\sigma_o l^H$ où (i) A_2 (la flexibilité de roll), ρ_{nuc} (courbure globale de la double helice autour du nucleosome) ne dépendent pas de la séquence, (ii) σ_o caractérise les fluctuations de la distribution de roll intrinsèque, (iii) H caractérise les corrélations à longue portée à petite échelle ($< 150pb$) dans les profils de roll intrinsèque et (iv) l est la taille de la sonde nucléosomale ($l = 125$ dans notre cas). Si on garde l fixée on voit donc que, pour

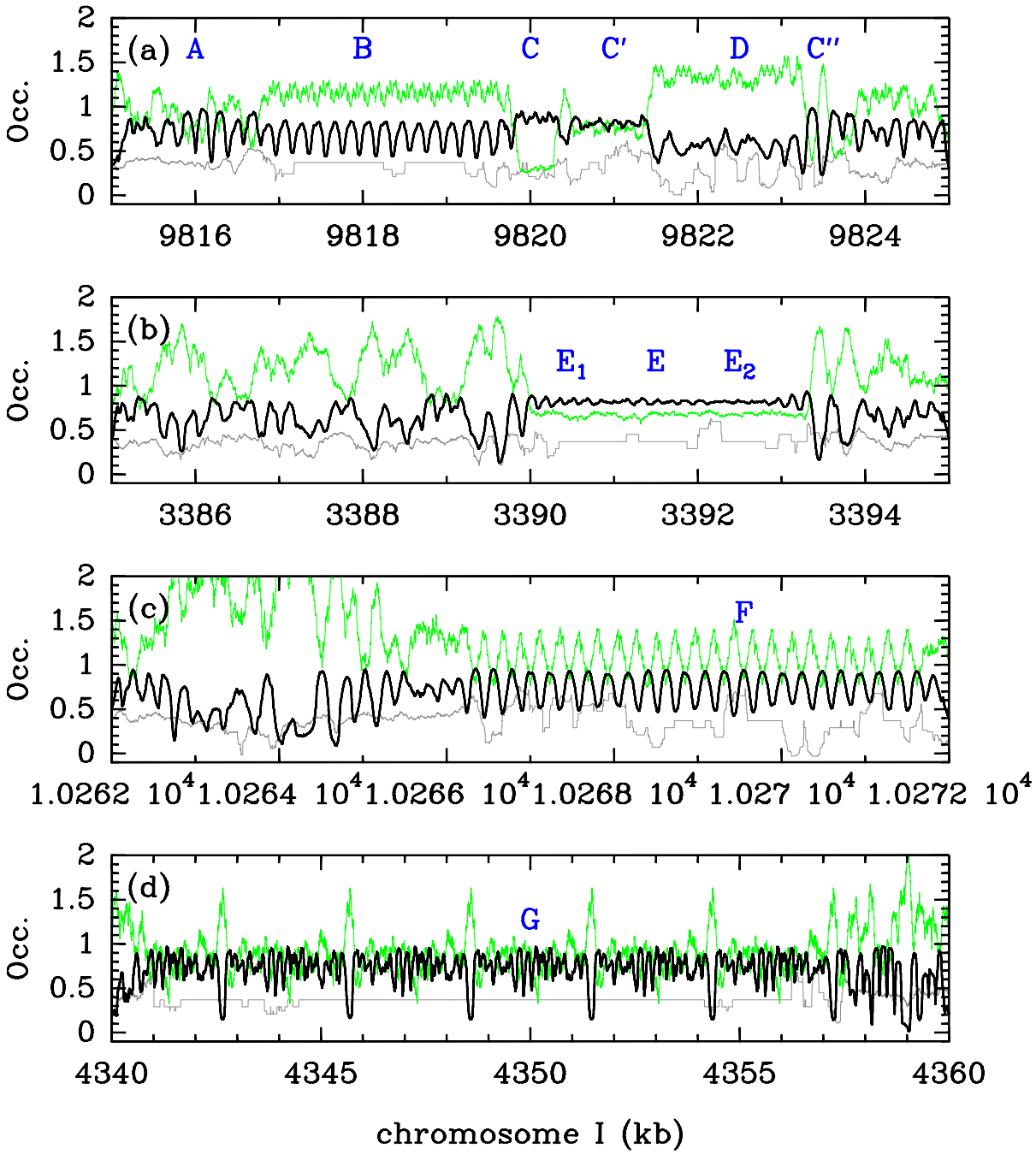


FIGURE 5.15 : *Extraits du chromosome I de C. elegans. En vert, le profil énergétique que l'on calcule avec le modèle Vaillant. En gris, les données expérimentales de Valouev (données normalisées, représentées en* log*). En noir l'occupation déterminée avec l'algorithme de Vanderlick* ($(\mu, \delta) = (-1.3, 2)$ kT*.*

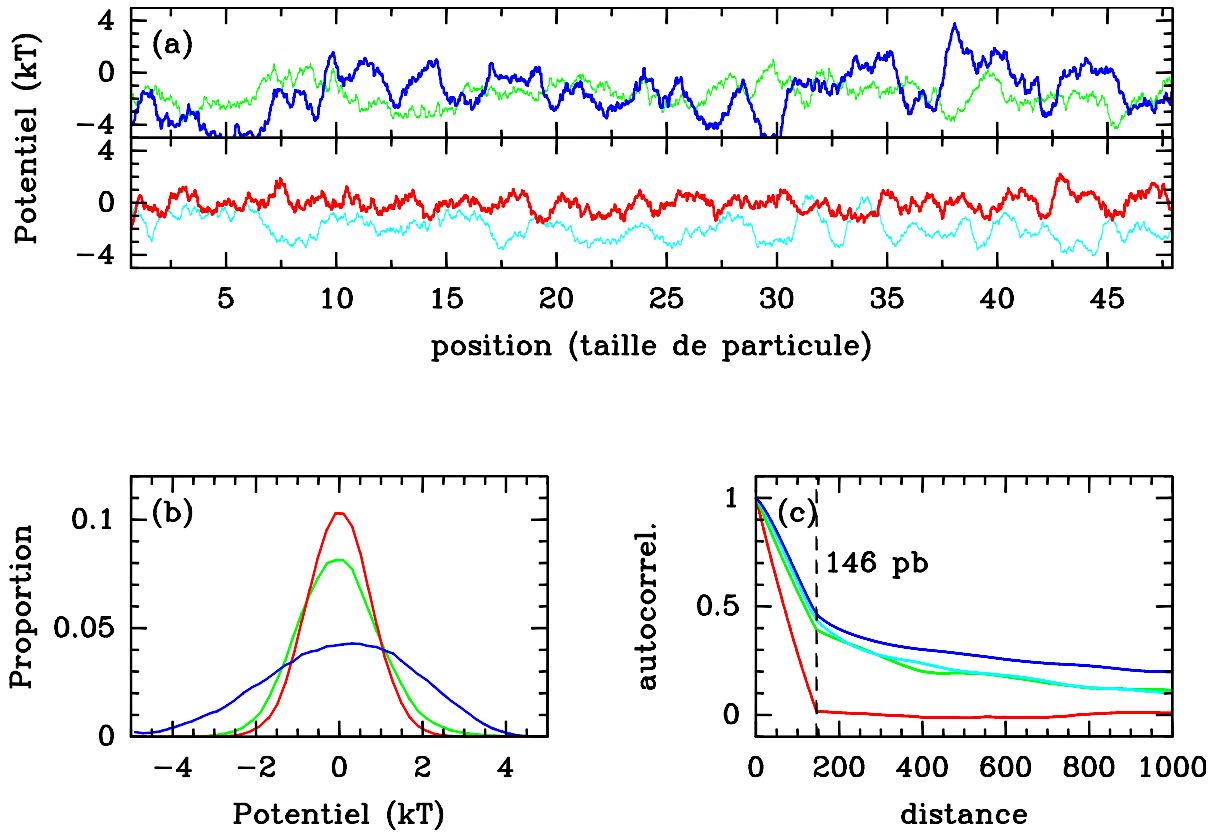


FIGURE 5.16 : *(a) Différence entre une énergie de formation issue d'une séquence générée avec un exposant de Hurst de* 0.5 *en rouge, par rapport à un profil généré par une séquence d'exposant* 0.8 *(en bleu). En cyan, un extrait du chromosome I de C. elegans et en vert un extrait du chromosome III de la levure, donnés à titre de comparaison. (b) Distribution de l'énergie en fonction du Hurst. (c) Fonction d'autocorrélation des profils énergétiques : en vert la levure, en cyan C. elegans, en bleu aléatoire* $H = 0.8$*, en rouge* $H = 0.5$*.*

une séquence donnée on peut moduler l'amplitude du profil énergétique soit en contrôlant la flexibilité de la double hélice, soit en modifiant σ_o ou H. σ_o dépend essentiellement de la composition moyenne en G+C (Vaillant et al., 2003) ; à même composition moyenne en G+C, même flexibilité (moyenne) A_2, une manière d'augmenter l'amplitude des fluctuations de l'énergie est de distribuer la séquence avec des corrélation à longue portée (augmenter H). Dans la levure $H = 0.54$, donc les rapport d'amplitude entre la levure et des séquences (de même compositions) non corrélées ($H = 0.5$) et CLP ($H = 0.8$) sont $\delta_{lev}/\delta_{0.5} \sim 125^{0.04} = 1.2$ et $\delta_{lev}/\delta_{0.8} \sim 125^{-0.26} = 0.3$.

5.6.2 Dépendance de la densité avec les paramètres classiques μ et δ

La densité moyenne augmente avec μ *et diminue avec* δ*. La densité maximale est limitée par* δ*.*
The mean density increases with μ *and decreases with* δ*. The maximum density is actually limited by* δ*.*

Les figures 5.17 et 5.18 montre la dépendance de la densité avec les deux paramètres classiques. La densité est maximale lorsque μ est élevé, et lorsque δ est faible. La dépendance avec le potentiel chimique est triviale (Fig. 5.18 (a)), l'augmenter correspond à améliorer l'affinité des particules avec le système et donc à peupler d'avantage le milieu. Mais à mesure que l'on augmente les amplitudes de variations du profil énergétique (Fig. 5.18(b)), le système se dépeuple car une partie du système a une énergie trop élevée pour être peuplée. En effet si les variations du potentiel augmentent, alors une partie de plus en plus importante de la séquence devient inaccessible, et la densité moyenne diminue. Cela est vrai tant que μ est assez fort, puisque comme le montrent les figures (Figs. 5.18 (a) et (b)) pour de très faibles valeurs de μ ($\mu < -5$) c'est l'effet inverse qui est observé : la densité augmente lorsque δ augmente jusqu'à une valuer caractéristique δ_m au delà de laquelle le dépeuplement intervient. La transition entre séquence faiblement occupée et séquence fortement peuplée a lieu aux alentours de

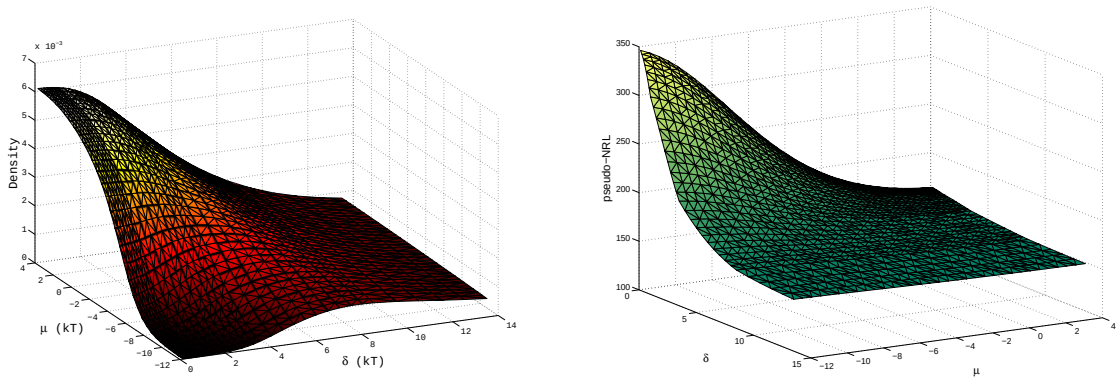


FIGURE 5.17 : *À gauche, nappe de densité calculée sur un profil énergétique généré par une séquence réelle (l'intégralité du génome de la levure), pour des nucléosomes de taille $l = 146$ pb. A droite, nappe de pseudo-NRL calculée sur un profil énergétique généré par une séquence réelle (l'intégralité du génome de la levure, $(\mu, \delta) = (-1.3, 2)$).*

$\mu = -5kT$ (figure 5.18 (b)) quand les amplitudes de variations du potentiel sont faibles, ce qui concorde avec la théorie dans le cas homogène. Lorsque l'amplitude grandit, la transition a lieu plus tôt, c'est-à-dire à potentiel chimique inférieur à $-5kT$.

Si on compare la levure avec les cas $H = 0.5$ et 0.8 (Fig. 5.19), la différence est faible tant que l'amplitude (celle du profit énergétique de la levure) est faible ($< 2-3kT$). Les differences observées concordent avec le fait que les amplitudes sont respectivement moins fortes dans le cas $H = 0.5$ et plus fortes dans le cas $H = 0.8$.

5.6.3 Pseudo-NRL dans les profils inhomogènes

Le pseudo-NRL diminue avec δ et avec μ.
The pseudo-NRL decreases with both μ and δ.

Lien entre fonction de paire et autocorrélation

Pour accéder à la distance typique entre deux particules, le mieux serait de pouvoir calculer la fonction de paire, malheureusement, ce n'est pas possible à partir de la simple donnée de la densité, encore moins à partir des données d'occupation. Même si l'autocorrélation et la fonction de paire ne correspondent pas à la même mesure, leur forme est relativement similaire, et la position du premier pic est quasiment la même dans les deux courbes.

Even if the autocorrelation and the pair function are not the same, their shape is relativly similar in noisy energetic landscapes, and the position of the first peak is similar in both curves.

Lorsque l'on est dans un cas simple comme le potentiel homogène, il est facile de calculer théoriquement la fonction de paire, mais si le profil n'est plus homogène le problème devient plus compliqué. Si l'on ne dispose que du profil de densité expérimental, accéder à ce paramètre se révèle impossible. C'est pourquoi, lorsque nous essaierons de définir la distance qui sépare deux particules, nous utiliserons une définition alternative : la position du premier maximum de l'autocorrélation du signal de densité. L'autocorrélation correspond à la quantification des propriétés d'invariance d'un signal, si un signal se ressemble à lui-même lorsque l'on le translate, alors la valeur de l'autocorrélation est forte pour la translation donnée. Si deux particules sont souvent juxtaposées, le signal de densité présentera deux oscillations espacées de la distance typique entre deux particules et l'autocorrélation sera forte pour une translation de cette distance. Dans le cas du potentiel homogène, l'autocorrélation est complètement plate et par conséquent fonction de paire et autocorrélation sont complètement différentes. Mais quand le potentiel est non homogène, il est possible de relier les deux plus clairement : la figure 5.20 montre comment l'autocorrélation s'approche de la fonction de paire sur ce genre de signaux, à partir de deux modélisations différentes (avec une interaction de tiges rigides, ou bien avec une interaction de coeur dur doublée d'une répulsion sur une échelle de l'ordre de 60 paires de bases). Que ce soit avec les coeurs durs purs, ou bien avec les coeurs durs et une répulsion, la fonction de paire et l'autocorrélation coïncident à partir du premier pic justement (figure 5.20 (b)).

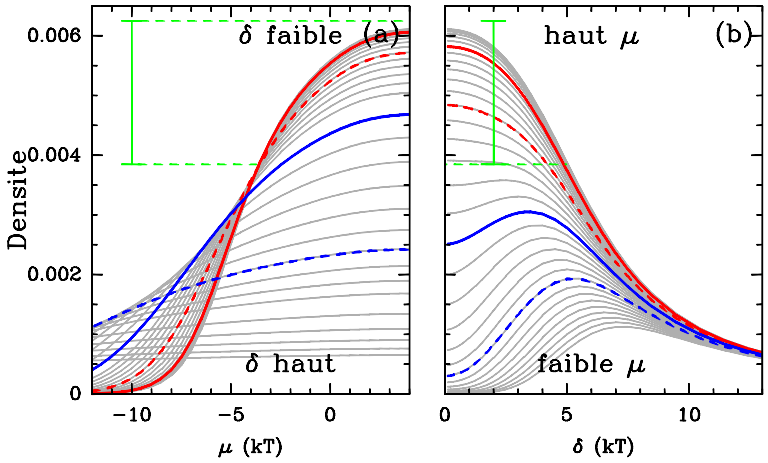


FIGURE 5.18 : *(a) Évolution de la densité avec le potentiel chimique pour différentes valeurs de δ. Calculs effectués sur l'intégralité du génome de la levure ; pointillé bleu : $\delta = 7$, en bleu 4, en pointillé rouge 2.25, en rouge $0.8kT$. Le domaine raisonnable de densité (situé entre 1 nucléosome tous les $160bp$ et 1 nucléosome tous les 250 paires de bases) est indiqué en vert. (b) Évolution de la densité avec l'écart type du potentiel, pour des μ variables (pointillé bleu : -9, bleu -6, pointillé rouge -3, rouge $0kT$).*

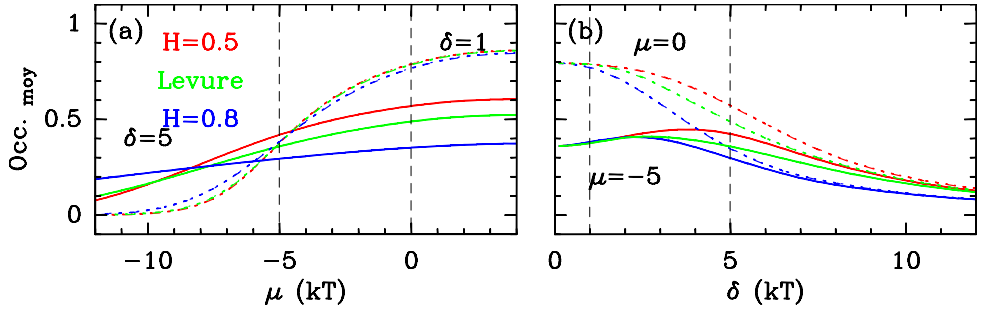


FIGURE 5.19 : *L'occupation moyenne dans un profil énergétique selon la séquence qui le génère : (a) évolution en fonction de μ pour deux valeurs de δ différentes ($\delta = 1$, en pointillé et $\delta = 5$, en trait plein), et pour la séquence de la levure (en vert), une séquence aléatoire de $H = 0.5$ (en rouge) et une séquence aléatoire de $H = 0.8$ (en bleu). δ se rapporte à l'amplitude de variation du profil de la séquence de la levure, rappelons qu'avec les mêmes paramètres du modèle l'amplitude de variation d'une séquence de $H = 0.5$ sera plus petite, et $H = 0.8$ plus grande (figures 5.16 (a) et* **??** *(b)). (b) évolution en fonction de δ (l'amplitude rapportée à la levure), pour deux valeurs du potentiel chimique $\mu = -5$ (en trait plein) et $\mu = 0$ (en trait pointillé).*

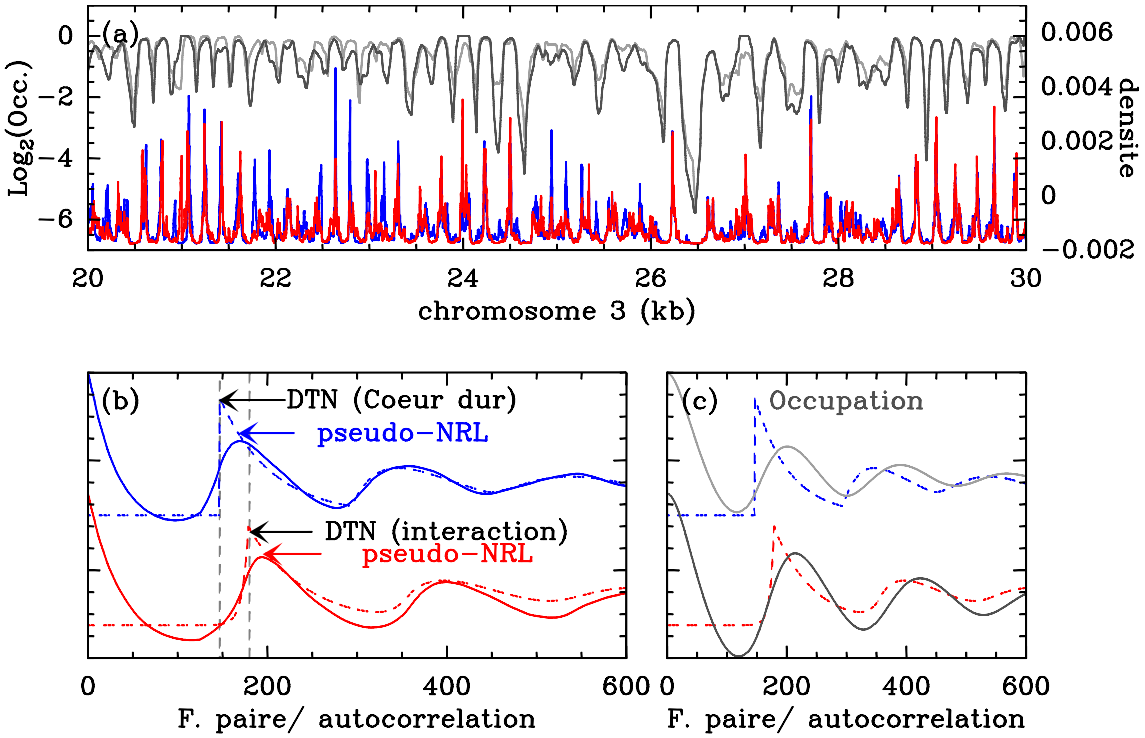


FIGURE 5.20 : *Lien entre la fonction de paire et l'autocorrélation. (a) Densité (resp. occupation) calculée pour des tiges rigides (bleu/gris clair), et pour une interaction répulsive plus lisse (repulsion répartie sur 60 pb), sur un morceau du chromosome 3 de la levure. (b) Comparaison entre la fonction de paire (en pointillé) et l'autocorrélation des signaux de densité de l'encart (a) (même code couleur). (c) Comparaison entre la fonction de paire et l'autocorrélation de l'occupation. Occupation coeur dur pur en gris clair, et occupation coeur dur et interaction en gris foncé*

Parfois, il arrive que seule l'occupation soit expérimentalement accessible. L'autocorrélation du signal d'occupation est plus éloignée de la fonction de paire dans ce cas là (figure 5.20 (c)), surtout pour les tiges rigides pures et dans une moindre mesure pour des particules avec une interaction répulsive. Du fait du caractère très doux de la fonction d'autocorrélation par rapport à la fonction de paire, prendre le premier maximum de l'autocorrélation revient à sur-estimer de quelques paires bases la DTN, comme on peut le voir sur l'encart (b). Le premier pic de l'autocorrélation est donc désigné sous l'appellation pseudo-NRL. Le phénomène important ici, est que la fonction de paire et l'autocorrélation dépendent de la même façon de l'interaction entre particules.

La figure 5.21 présente différentes formes de profil d'autocorrélation pour des valeurs de potentiel chimiques croissantes. Le pseudo-NRL extrait de ces courbes, est aussi fonction des différents paramètres δ et μ (figure 5.17 et 5.23). Les dépendances correspondent à ce que l'on a décrit dans le cas des barrières : c'est l'intensité du confinement qui détermine le pseudo-NRL. Le pseudo-NRL diminue de façon monotone lorsque la densité augmente (confinement créé par une forte population) ou bien lorsque l'amplitude de variation du potentiel augmente (confinement créé par le potentiel). La valeur du pseudo-NRL varie autour de 1 nucléosome tous les 180 paires de bases, et peut descendre minimum aux alentours de 1 nucléosome tous les 160 pb quand l'amplitude δ est très élevée. La dépendance de l'autocorrélation avec l'exposant de Hurst est plus subtile. Comme le montre la figure 5.22(a), changer le H de la séquence n'affecte pas particulièrement la forme de l'autocorrélation. On observe simplement le renforcement des oscillations, qui découle de l'augmentation de l'amplitude naturelle de variation du profil énergétique avec H. Le confinement est plus fort, les oscillations de la fonction d'autocorrélation sont plus fortes également. D'un point de vue théorique, on peut accéder au NRL directement, en calculant la pseudo-période de la fonction de paire comme dans la figure 5.7. Le NRL déterminé de cette façon varie peu avec le Hurst (figure 5.22 (b)), et varie de façon similaire au pseudo-NRL.

La densité et le pseudo-NRL sont deux paramètres reliés, mais qui ne sont pas strictement équivalents. Les deux dépendent indirectement de l'amplitude du potentiel ainsi que du potentiel chimique. La figure 5.23 résume ces diverses dépendances. On voit que pseudo-NRL et distance inter-nucléosomale

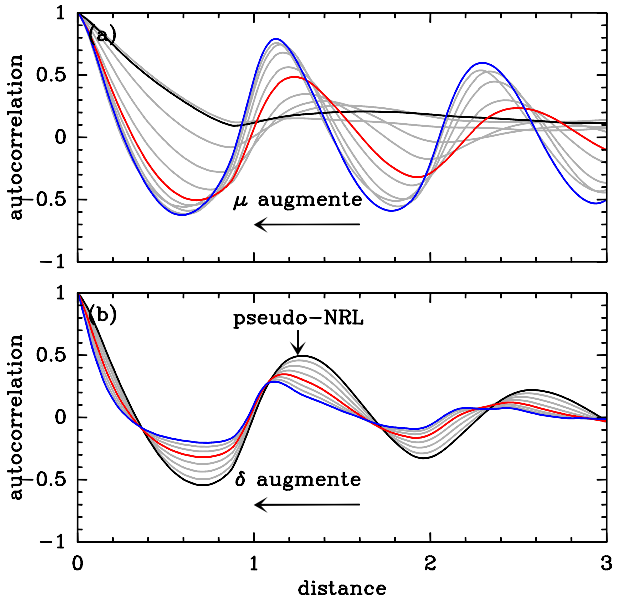


FIGURE 5.21 : *Profil d'autocorrélation d'un signal de densité typique (généré à partir d'une séquence aléatoire de Hurst* 0.8*). (a)* μ *variable,* $\delta = 1$*,* $\mu = -7kT$ *en noir,* $-2kT$ *en rouge, et* $+3kT$ *en bleu, intermédiaires en gris. (b)* δ *variable,* $\mu = -2kT$*,* $\delta = 0.5kT$ *en noir,* $\delta = 1kT$ *en rouge,* $\delta = 3kT$ *en bleu*

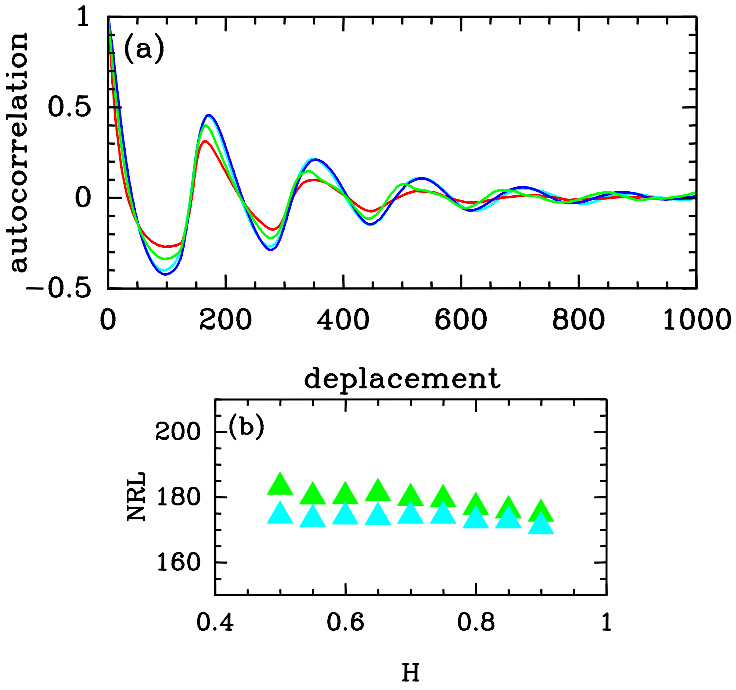


FIGURE 5.22 : *(a) Les fonctions d'autocorrélation selon l'exposant de Hurst, calculées sur des profils de densité déterminés avec* $\mu = -1.3$ *et* $\delta = 2$*. En rouge,* $H = 0.5$*, en bleu* $H = 0.8$*, en vert l'autocorrélation de la densité calculée sur le profil énergétique de la levure, en cyan, celle sur C. elegans. (b) La variation du pseudo-NRL (cyan) et du NRL (en vert, calculé comme la pseudo-période de la fonction de paire, voir figure 5.7) en fonction de* H $(\mu, \delta) = (-1.3, 2)$.

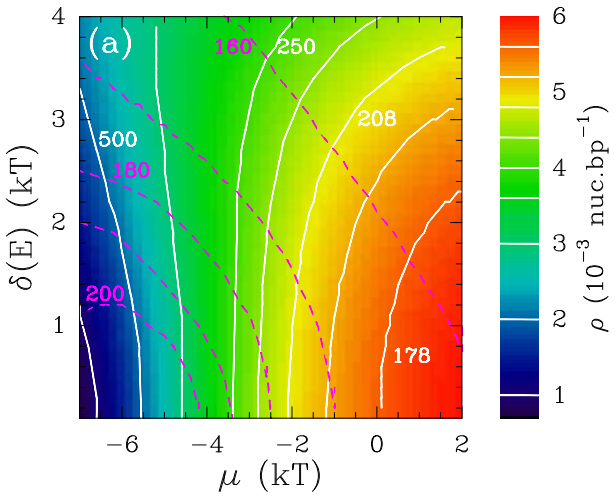


FIGURE 5.23 : *Superposition de la nappe de densité avec une représentation des différentes iso, calculée sur la levure. En blanc : iso-densités, en rose, iso-pseudo-NRL.*

moyenne coïncident presque lorsque $\delta \to 0$. En effet lorsque le profil est homogène, le repeat moyen est directement reliée à la densité moyenne par NRL $= 1/\rho_m$. Ailleurs, à chaque couple (μ, δ) correspond une relation (ρ_m, pseudo-NRL) moins triviale.

Pour une séquence désordonnée, on retiendra donc que si la variabilité δ augmente, la densité diminue sauf à bas potentiel chimique où elle commence par augmenter légèrement avant de diminuer (figure 5.19 (b)). Par contre, plus cette variabilité augmente et plus le NRL (et pseudo-NRL) diminue du fait du plus fort confinement. L'influence des corrélations à longue portée (celles observées à petite échelle) est ici relativement faible tant que l'amplitude des variations du profil énergétique n'est pas trop grande $\delta < 2 - 3kT$. Cette influence est essentiellement d'augmenter le confinement. On verra par ailleurs au chapitre suivant que les corrélations à longue portée observées à grande échelle cette fois-ci ($H = 0.8$) se réflète dans l'organisation du chapelet en augmentant la portée du confinement.

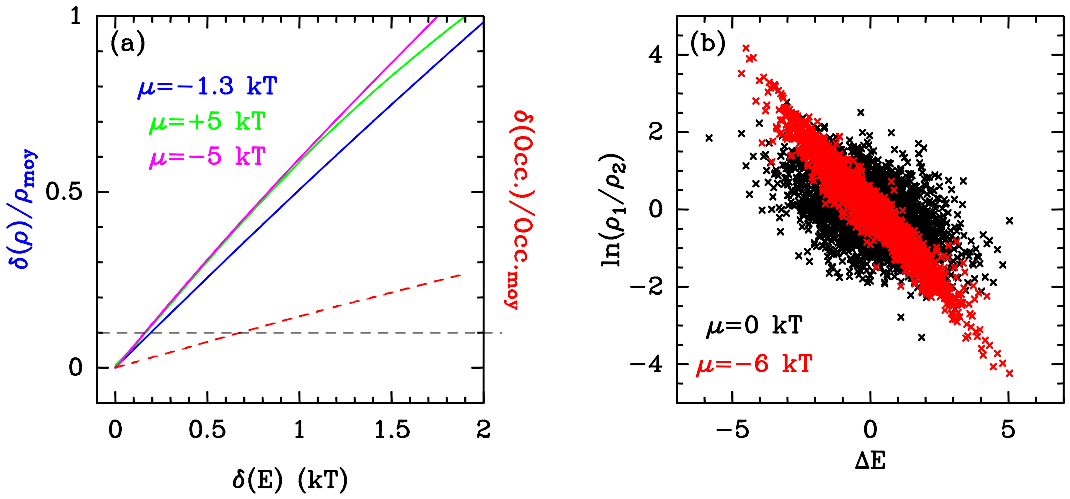


FIGURE 5.24 : *(a) Évolution du contraste (écart type de la densité rapportée à la densité moyenne, en bleu) en fonction de l'écart type du profil énergétique aléatoire (δ) pour $\mu = -1.3$ kT. Le contraste de l'occupation est montré en pointillé rouge également. La limite à 10 % de l'amplitude moyenne est représentée en pointillé noir. La dépendance du contraste avec le bruit dépend très peu du potentiel chimique (courbes de densité à $\mu = -5$ en rose et $\mu = +5$ en vert). (b) Logarithme du rapport de densité entre deux points séparés de 1 kb ($\rho(s_1)$ et $\rho(s_2)$) du profil (ici calculé sur le génome de la levure) en fonction de la différence d'énergie ($E(s_1) - E(s_2)$) de ces deux points : en rouge dans un milieu dilué ($\mu = -6kT$), en noir dans un milieu dense ($\mu = 0kT$).*

5.6.4 Degré de positionnement

La figure 5.24 montre comment le contraste dans la densité évolue en fonction du bruit (l'écart type du profil énergétique, ici calculé en prenant la séquence de la levure). On s'intéresse à ce contraste car c'est une façon de quantifier l'intensité du positionnement, de faibles variations dans le profil de positionnement signifient que certaines positions sont en moyenne légèrement plus souvent occupées que d'autres, mais globalement le positionnement est homogène. Si le contraste est fort, alors on peut parler de nucléosomes "bien positionnés" et de zone déplétées. Le contraste de densité est proportionnel au bruit qui est mis dans l'énergie, et devient non négligeable (écart type supérieur à 10% de la valeur moyenne) pour une amplitude de l'ordre de $0.2 - 0.6$ kT selon que l'on s'intéresse au contraste de la densité ou au contraste de l'occupation respectivement.

Les rapports de densités entre un point s_1 et un point s_2 sont guidés par des termes d'ordre de grandeur $e^{-\beta(E(s_1)-E(s_2))}$ si le milieu est suffisamment dilué (points rouges, pour $\delta = 1$ kT, figure 5.24 (b)). Si le milieu n'est pas dilué, alors la différence d'énergie entre deux points ne suffit plus à spécifier la différence de densité entre ces deux points, du fait justement des interactions entre particules. Comme l'indique la plus grande dispersion dans la relation, à haute densité, une même énergie de formation peut mener à des densités très différentes du fait d'un environnement énergétique différent. On ne considèrera pas des variations de potentiel plus rapides que quelques kT par paire de bases qui provoqueront des positionnements de type tout ou rien (figure 5.29 (b)). Ce problème nous intéressera plus loin, car l'affinité du nucléosome avec la séquence d'ADN possède naturellement une composante oscillante de très grande amplitude, correspondant au phasage avec les sillons majeurs et mineurs de l'ADN (figure 4.21). Les détails de ce problème seront abordés plus exhaustivement au paragraphe 5.7, souvenons nous juste que cette composante haute fréquence aura beau affecter sérieusement le positionnement résolu à la paire de base près, le positionnement à gros grain –c'est-à-dire résolu uniquement à l'échelle de quelques dizaines de paires de base– restera quant à lui relativement insensible à ces variations de haute fréquence.

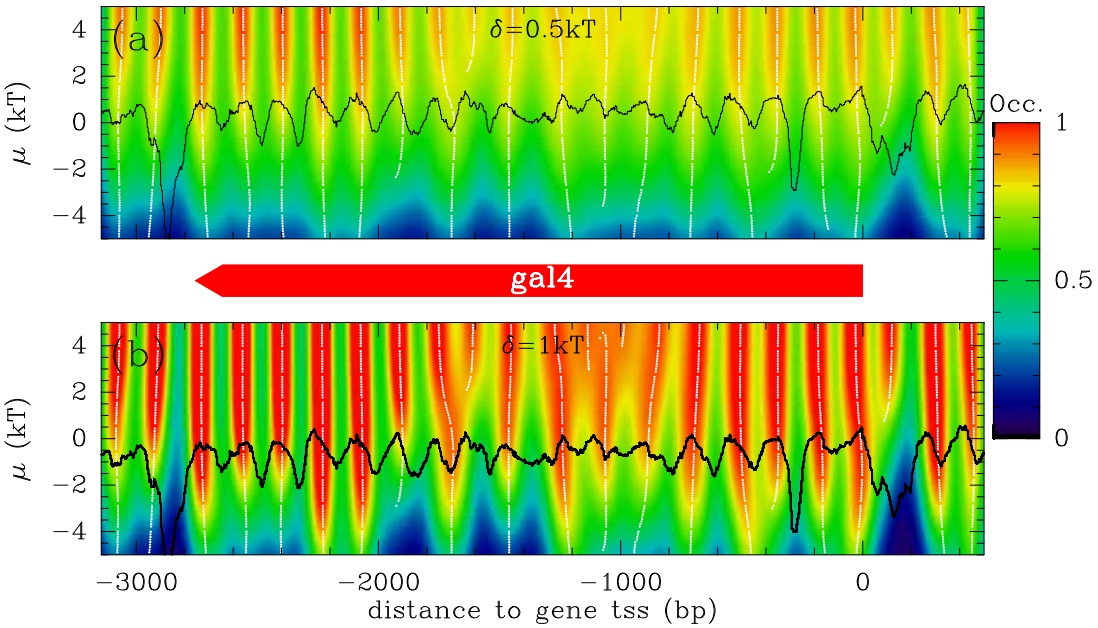


FIGURE 5.25 : *Évolution du positionnement prédit le long du gène gal4 de la levure. Le profil expérimental (Lee* et al.*, 2007a) est représenté en noir. Le potentiel chimique varie verticalement, le niveau d'occupation est traduit par un code de couleurs : rouge (très fortement occupé) vers bleu (très faiblement occupé). Les positions des maxima locaux d'occupation sont matérialisés par des lignes blanches. (a) L'amplitude du bruit généré par la séquence est faible (écart type $\delta = 0.5$ kT), le contraste en occupation est donc faible. (b) L'amplitude du bruit généré par la séquence est plus élevé (écart type $\delta = 1$ kT).*

5.6.5 La robustesse des nucléosomes

Lorsque la séquence n'est plus tout à fait homogène, et qu'une faible perturbation est apportée, les positions relatives des particules se phasent. Ce phasage est susceptible de changer lorsque le potentiel chimique augmente.

The noise in the energetic landscape phases the nucleosomal array and some nucleosome are better fixed than other. We can evaluate the robustness of nucleosome.

Influence du peuplement sur la position locale des nucléosomes

La robustesse des nucléosomes est déterminée à la fois par le positionnement intrinsèque et par l'environnement. Un nucléosome robuste est un maximum de la densité qui ne change pas de position lorsque le potentiel chimique varie.

The robustness of nucleosomes is determined both by the intrinsic contribution of the sequence and by the neighbouring nucleosomal environement. A robust nucleosome is defined by a peak in the density whose position does not change when the chemical potential varies.

Lorsque l'on peuple progressivement une séquence avec des nucléosomes, il est inévitable que certaines positions occupées lorsque la densité est faible deviennent inoccupées du fait de la contrainte stérique imposée par le nombre de plus en plus grand de voisins. La figure 5.25 présente l'évolution de l'occupation dans un profil énergétique classique, en l'occurrence le profil énergétique généré par la séquence du gène gal4 de la levure. Lorsque le potentiel chimique est faible, le profil présente des pics relativement espacés, et la valeur de la densité est faible. Mais à mesure que le potentiel chimique augmente, les pics de densité se rapprochent (Le NRL et le pseudo-NRL diminuent), la densité moyenne augmente, et la position de chacun des maxima de densité peut légèrement changer. En effet lorsque, du fait de la pression, il devient plus avantageux de mettre deux nucléosomes sur des parties de l'énergie défavorables que de laisser un seul nucléosome, alors il se produit une transition –ou une bifurcation– dans le positionnement du maximum de densité (cf le tracé des positions des maxima en blanc de la figure 5.25). À partir de ces tracés de positions, il devient possible de définir la robustesse d'un nucléosome. La figure 5.25, en représentant l'occupation à différents potentiels chimiques, permet bien de visualiser les nucléosomes présents quel que soit μ et ceux qui n'apparaissent que tardivement. Un maximum local auquel on associe un nucléosome donné, sera d'autant moins robuste que sa position à

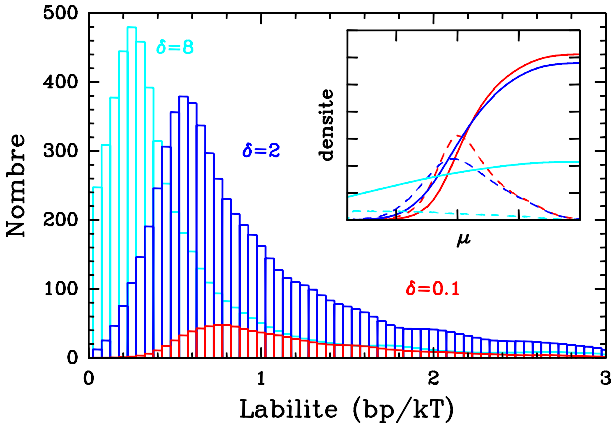


FIGURE 5.26 : *Variation de la robustesse avec l'amplitude de variation du potentiel énergétique ($\mu = -1.3$) : histogramme de la labilité des nucléosomes (maxima de positionnement prédits en blanc sur la figure 5.25) calculé sur le chromosome III entier de la levure en fonction de l'amplitude du bruit généré par la séquence δ. En insert, un rappel de la forme de la transition dans la densité en fonction de μ pour des bruits de diverses amplitudes. En pointillés, la dérivée des courbes de transitions (la susceptibilité nucléosomale).*

faible et à fort potentiel chimique sera différente. La dérivée de la position d'un nucléosome en fonction de μ permet d'établir de combien se déplace un nucléosome avec les variations du potentiel chimique. C'est à partir de là que l'on construit la robustesse R du nucléosome N comme :

$$R_{\delta,N}(\mu_0) = \left(\frac{\partial X_N}{\partial \mu}\right)^{-1}_{\mu_0} \tag{5.2}$$

où X_N correspond à la position du nucléosome N au potentiel chimique μ_0. L'inverse de la robustesse correspondrait à ce qu'on pourrait appeler la labilité d'un nucléosome, c'est-à-dire sa propension à changer de position avec des variations du potentiel chimique. Avec cette définition, un nucléosome sera d'autant plus robuste qu'il faudra une grande variation de potentiel chimique pour le bouger. Des nucléosomes peuvent très bien être robustement positionnés sans que l'énergie effective qu'il faille fournir pour les mouvoir soit grande s'ils étaient seuls dans la séquence. Non, ce qui précise la robustesse d'un nucléosome c'est à la fois l'environnement énergétique (la position du nucléosome est elle placée au fond d'un puits d'énergie local ?) et l'environnement stérique (Est-ce que bouger ce nucléosome permettrait de rajouter une particule dans le système ?). Les configurations les plus robustes sont donc celles où des puits d'énergie sont répétés dans l'espace à une distance idéale pour que les interactions stériques ne soient pas trop fortes.

Évolution de la robustesse avec l'amplitude de variation du potentiel

Sous la pression grandissante de leurs voisins, les nucléosomes changent peu de position, un petit nombre de nucléosomes se déplace suffisamment pour accueillir de nouvelles particules.

As the pressure increases in the hard rods line, few particles move. Only a subset of particles will move sufficently so that new particles may settle in the potential. Collective remodeling is hard.

La robustesse des nucléosomes dépend évidemment du profil énergétique et si les amplitudes de variations du potentiel sont grandes alors, comme on l'a vu, le confinement est plus fort. La robustesse des nucléosomes sera en conséquence plus grande également. La figure 5.26 permet de confirmer ce résultat : elle présente les histogrammes de labilité. Exprimée en paires de bases par kT, la labilité est l'inverse de la robustesse et quantifie le changement de position subi par les nucléosomes (maxima de position, figure 5.25) lorsque le potentiel chimique augmente. Effectivement, plus l'amplitude de variation du potentiel est grande, plus la distribution de la labilité se rapproche de zéro. Il est normal que les histogrammes ne soient pas de même somme : lorsque l'amplitude de variation de l'énergie est faible il y a moins de maxima locaux que lorsque l'amplitude est plus élevée. Pour des séquences réelles, on s'attend à ce que l'amplitude liée à la séquence soit située dans un intervalle situé entre $\delta = 0.5$ et $\delta = 3$ kT (voir chapitres 5 et 6). Il faudra donc s'attendre à ce que les nucléosomes bougent en moyenne de quelques paires de bases tout au plus, lorsque le potentiel chimique augmente. Attention, il s'agit d'une

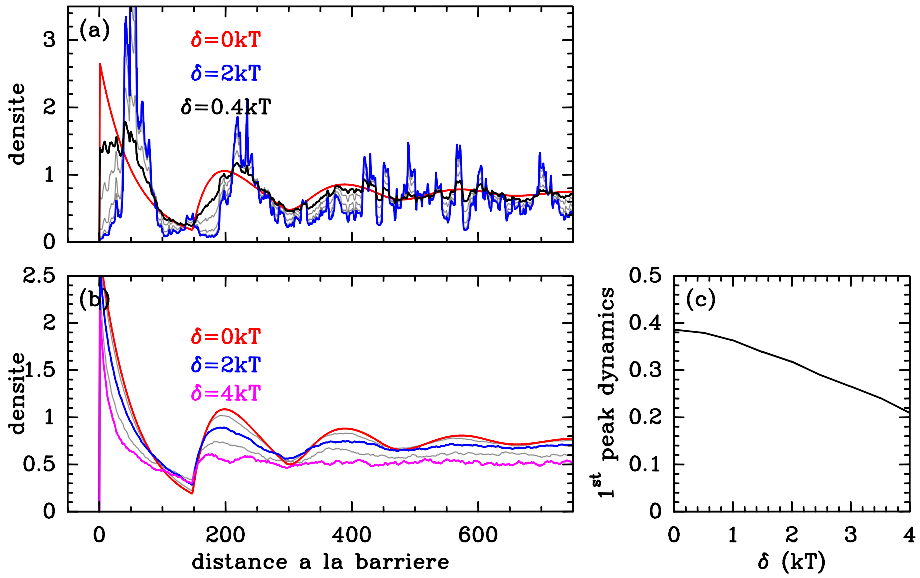


FIGURE 5.27 : *La portée dépend aussi du bruit. (a) Un exemple simple : à proximité d'une barrière verticale infinie, la densité oscille d'autant plus que l'influence de la séquence est faible. (b) Profils de positionnement moyens à proximité d'une barrière infinie, jouxtée par un bruit d'amplitude* 0 *kT en rouge,* 2 *kT en bleu et* 4 *kT en rose (potentiel chimique* $\mu = -1.3kT$*). (c) Dynamique,* i.e. *l'amplitude du deuxième pic sur celle du premier pic de la densité à proximité d'une barrière, en fonction du bruit.*

moyenne, et l'histogramme montre toujours une queue d'évènements de forte labilité. Ceci signifie que la plupart des nucléosomes changent peu de position et qu'un petit nombre est véritablement déplacé du fait de la pression et du positionnement statistique.

5.6.6 Effet du "bruit" de séquence sur le positionenment statistique au voisinage des barrières

Comme on l'a vu notamment avec CTCF, une barrière induite par la présence de protéines ou complexes protéiques, induit un ordonnancement périodique. L'étude de ce positionnement statistique a été fait dans le cas "idéal" où, mise à part au niveau de la barrière, le profil énergétique était uniforme. Si on considère désormais un modèle énergétique mixte avec une barrières (zones d'exclusions forte) se "rajoutant" au profil énergétique intrinsèque.

Comme le montre la figure 5.27 (a), dès que l'on rajoute de l'inhomogénéité il y a compétition entre le positionnement local induit par la barrière et celui induit par les inhomogénéités. Plus l'amplitude de varaiation du profil intrinsèque augmente, plus le positionnement local induit par la séquence perturbe le positionnement statistique lié à la barrière. Donc l'effet non local de positionnement d'une barrière est maximal lorsque le profil énergétique environnant est uniforme. En moyenne, cela se traduit par des profils de positionnements qui sont d'autant moins oscillants que l'amplitude des variations du profil énergétique δ est forte. La dynamique du premier pic, qui correspond à l'amplitude du deuxième pic de positionnement rapporté à l'amplitude du premier pic, passe de 0.4 à $[0.1 - 0.2]$ lorsque le bruit est de l'ordre de $4\ kT$. L'effet de séquence devient sensible à partir de $\delta = 1kT$.

Une dépendance similaire est observée biologiquement par Tirosh et Barkai (Tirosh et al., 2010) qui montrent que la suppression du remodeleur ISW1 dans la levure, diminue l'amplitude du positionnement périodique à proximité des promoteurs (Fig. 5.28). Or effectivement, ISW1 facilite le mouvement des nucléosomes sur la séquence, et peut être considéré comme un "fluidificateur" en diminuant la rugosité du profil énergétique intrinsèque.

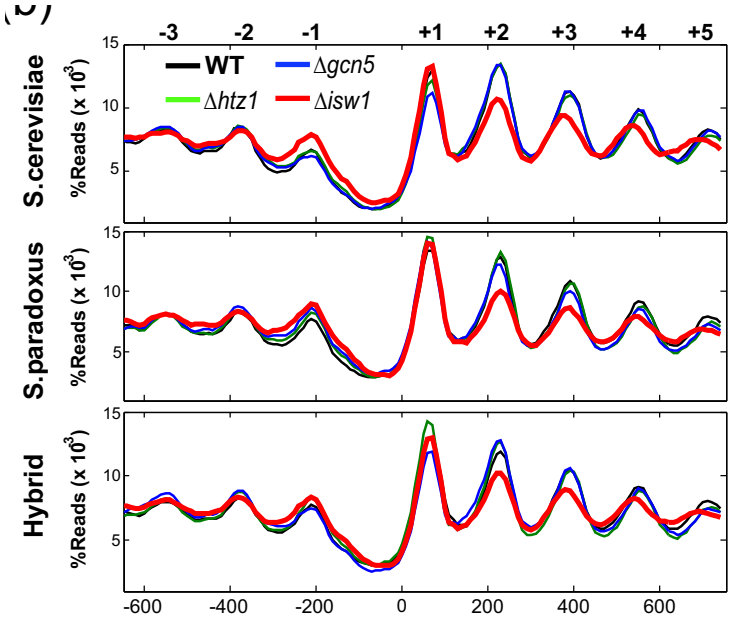


FIGURE 5.28 : *Profil nucleosomal moyen pour tous les gènes de la levure dans la souche sauvage (WT) et trois souches mutantes ; en toute position les valeurs correspondent au pourcentage de reads mesurés relativement à la valeur au TSSs. La même analyse a été menée pour* S. cerevisiae *(haut),* S. paradoxus *(milieu) et leur hybride (bas)*

5.7 PÉRIODICITÉ RAPIDE DANS L'ADN

Revenons aux périodicités rapides dans le potentiel liées au phasage avec les sillons mineurs de l'ADN : quelle est leur influence ? Elles ne changent pas le positionnement à gros grain, mais affectent la valeur moyenne du positionnement, en excluant une partie de la séquence.

What with the high frequency component linked to the necessary phasing with the minor groove of the DNA. Will it affect coarse-grained positioning ?

Nous avons vu au chapitre 2 que l'énergie présente une composante haute fréquence (à environ 10 pb) correspondant au fait que l'enroulement autour du nucléosome est anisotrope. Les points d'adhésions nécessitent une interaction entre le petit sillon et l'octamère. Notre modèle en fait abstraction puisqu'il prend l'enveloppe inférieure du potentiel de formation d'un nucléosome, et de ce fait, oublie cette périodicité. Nous proposons ici d'analyser la situation lors de l'ajout d'une haute fréquence dans le potentiel. On compare donc les profils de densité issus du modèle avec et sans la haute fréquence. Prendre l'enveloppe inférieure correspond bien à retirer la haute fréquence liée au phasage avec les sillons. Ce qu'il faut retenir, c'est qu'au vu de la figure 5.29 (a) et (b), la présence d'une haute fréquence ne changera pas fondamentalement la forme des résultats peu résolus. En effet, à haute résolution, les profils générés avec la haute fréquence ressemblent à une discrétisation du profil de densité généré sans la HF ; une fois filtrés par une fenêtre de quelques paires de bases (11 paires de base dans la figure), les résultats redeviennent très similaires. Imposer des variations à haute fréquence dans le potentiel énergétique a des conséquences sur la densité moyenne. Puisqu'une partie de la séquence devient effectivement beaucoup plus difficile à atteindre, il n'est pas étonnant que la densité moyenne diminue par rapport à une énergie non bruitée, il faut un potentiel chimique plus élevée pour atteindre les mêmes densités moyennes. On voit en outre que le filtrage des données n'est pas sans effet, puisque le filtrage par une fenêtre rectangle de la taille de la périodicité les a fait disparaître.

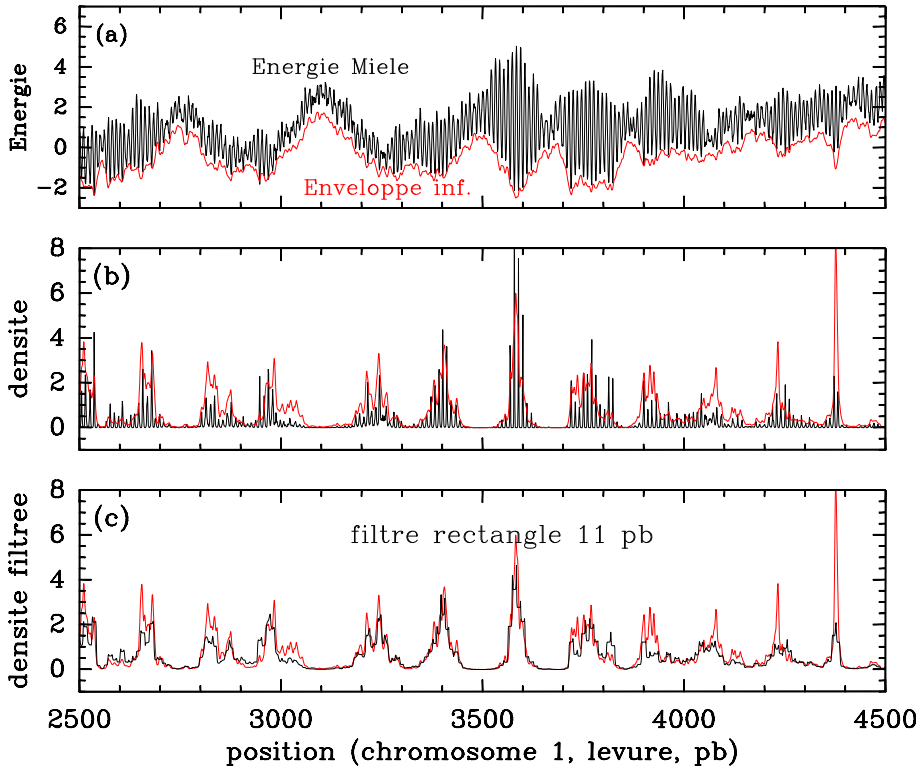


FIGURE 5.29 : *(a) Le profil énergétique avec (en noir) et sans la haute fréquence (en rouge) liée au phasage avec les petits sillons de l'ADN, calculée selon le modèle de Miele et al.* Miele et al. *(2008). (b) Influence d'une haute fréquence (en rouge) ($T = 10/146$, $A = 5kT$) sur les profils de densité. Comparaison avec les profils de densité sans haute fréquence (en noir). (c) La densité haute fréquence est filtrée par une fenêtre carrée de taille* $11/146$.

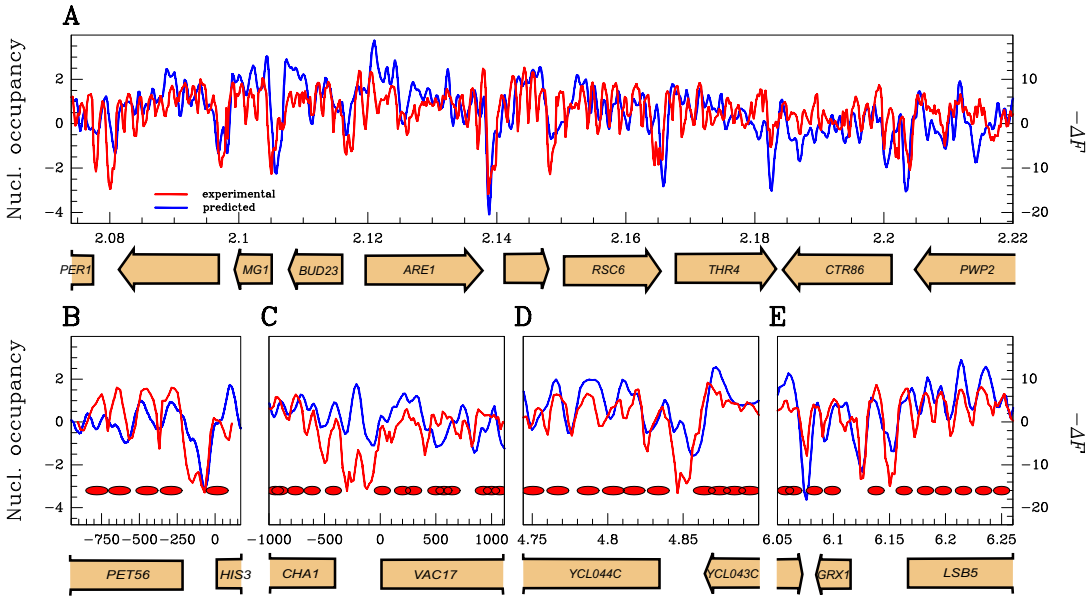


FIGURE 6.1 : *Comparaison entre le profil d'affinité $-\Delta F$ (courbe bleue, unité en kT) avec les profil expérimentaux d'occupation chez* S. Cerevisiae *(courbes rouges : log2ratio des données d'hybridation obtenues par Yuan* et al. *(Yuan* et al.*, 2005). Les ovales rouges représentent les positions de nucléosomes prédites par Yuan* et al. *par méthode HMM. (A) Portion représentative du chromosome III (B-D) Comparaisons à certaines régions particulières ; (B) promoteur HIS3 ; (C) promoteur CHA1 (le profil théorique prédut un nucléosome séparant les promoteurs des deux gènes divergents. Ce nucléosome est bel et bien détecté expérimentalement mais à un niveau plus faible, certainement du fait de la présence des machineries de transcription) ; (D,E) Vues locales de certaines régions du chromosome III montrant à la fois des nucléosomes localisés et des nucléosomes délocalisés. En abscisse, les unités sont $100\ kb$ dans (A), $1\ pb$ dans (B,C), $10\ kpb$ dans (D,E). La coorelation de pearson entre les données expérimentales et $-\Delta F$ calculée sur tout le chromosme III est de $r = 0.45,\ P < 10^{-15}$*

6 IN VIVO : POSITIONNEMENT INTRINSÈQUE ?

A partir des modèles énergétiques présentés précédemment nous pouvons donc calculer un profil de positionnement de nucléosomes le long des génomes.

6.1 LEVURES IN VIVO

– ***S. Cerevisiae*** : *Données de Yuan et al. (2005)* : Dans nos premiers travaux (Miele et al., 2008) nous avons directement comparé les données de MNase-chip de Yuan et al. (Yuan et al., 2005) avec l'énergie libre $\Delta F(s)$ issue du modèle élastique présenté au chapitre 5 avec les paramètres "Anselmi" ; le modèle thermodynamique du chapelet nucléosomal, pris comme un fluide non uniforme de nucléosomes avec volume exclu le long des génomes n'avait pas été mis en place. Nous avons calculé le profil énérgétique $\Delta F(s, l)$ pour différentes tailles d'enroulement l, et la meilleure corrélation a été obtenue pour une taille $l = 73pb$ correspondnat à un demi-nucléosome ou plutôt à l'enroulement autour du tetramère. Comme l'indique la (Fig. S1 de (Miele et al., 2008)) figure 6.1, ce modèle reproduit de façon très satisfaisante les profils expérimentaux, avce une corrélation de Pearson sur l'ensemble du chromosome III de $r = 0.45$. Le profil énergétique reproduit en particu-

liers très bien les zones de déplétion. Ces prédictions sont établies avec une sonde nucléosomale de type super-hélice idéale et dans le cas isotrope ; on a montré qu'introduire une sonde nucléosomale de type de celle extraite du cristal ou/et introduire de l'anisotropie ne changeait rien quant à la performance du modèle (Fig. 6.2 C et D). De même les perfomramces équivalentes ($r = 0.35$) sont obtenues par le modèle plus complet introduit par Tolstorukov *et al.* (Tolstorukov et al., 2007) qui prend en compte les déformations associées aux six degrés de liberttés hélicoïdaux (Chapitre 5) et leurs couplages, et qui utilise les paramètres élastique de Olson *et al.* (Olson et al., 1998) (Fig. 6.2 A et B).

– ***S. Cerevisiae*** : *Données de Lee et al. (2007)* :
Ayant ensuite développé notre modèle thermodynamique de positionnement de nucléosomes, d'abord par le biais de simulation de type Monte Carlo (Vaillant et al., 2007) puis en utilisant la relation de Percus et l'algorithme de Vanderlick (Chevereau et al., 2009; Vaillant et al., 2010; Milani et al., 2009) (chapitre 5), nous avons essayé de modéliser la structure primaire du chapelet nucléosomal à savoir le profil d'occupation en nucléoosme obtenu par Yuan et al. (Yuan et al., 2005) puis Lee et al. (Lee et al., 2007a) notamment chez la levure. Il s'agissait desormais d'inclure les interactions entre nucléosomes et de comparer les données non pas à un profil énergétique (comme montré précédemment dans notre première étude comparative (Miele et al., 2008)) mais à un profil d'occupation. Le profil énérgétique $\Delta F(s)/kT$, dans le cadre d'un tel modèle thermodynamique, rend finalement bien compte de l'occupation tant que la densité moyenne est faible puisqu'on a localement $\ln \rho(s) \propto -\Delta F(s)$ (cf. relation de Percus) ; cependant à forte densité l'effet des interaction n'est plus négligeable et cette relation de proportionnalité n'est plus valable : c'est ce qu'illustre la figure pédagogique 6.3, où on voit bien que le profil d'occupation à faible densité (faible μ) est finalement bien décrit (sa topographie) par le profil d'énergie (on a une corrélation (ici négative) très forte entre les deux profils) ; tel n'est plus le cas à haute densité du fait des oscillations caractéristique du positionnement statistique induit par les barrières. Que ce soit pour le profil énergétique caculé avec les paramètres "Anselmi" ou "Pnuc", leur fonction d'autocorrélation ne présente pas de modulation périodique à $168\ pb$ (Fig. 6.8(d), vert), indiquant qu'effectivement cette régularité n'est pas encodée dans la séquence mais résulte d'un positionnement statistique (Vaillant et al., 2007).
Nous avons, depuis les études de Miele *et al.* (Miele et al., 2008), préféré utiliser le modèle élastique avec les paramètres "Pnuc", simplement parce que les profils d'occupation correspondant présentent une corrélation plus forte avce les données expérimentales que le modèle avec les paramètres "anselmi". En gardant évidemment à l'esprit qu'il s'agit plus de paramètres phénoménologiques qu'élastiques.
Comme paramètres nous avons alors choisi :

1. La taille effective d'enroulement $l = 125$ pb pour calculer les profil énergétiques ; contrairement à l'étude de Miele et al.(Miele et al., 2008) où la taille optimale de la "sonde" nucléosomale était plutôt $73pb$ ($\sim$ le tétramère) ici, la plus forte corrélation est atteinte pour cette valeur $l = 125$. La différence vient du fait qu'ici on compare les expériences à un profil d'occupation non à une énergie. Cela dit, on remarque que la contribution du tetramère est ici aussi très importante (corrélation $r = 0.71$ avec $l = 80$) mais n'est pas suffisante. En d'autre terme la spécificité de séquence de l'octamère ne se réduit pas exclusivement (mais en grande partie tout de même) à celle du tetramère.
2. L'amplitude du profil énergétique $\delta = 2kT$; comme l'indique la figure 6.6(d) la corrélation évolue en fait très peu avec δ, avec un léger optimum pour $\delta = 1.3\ kT$. Nous avons choisi $\delta = 2$ car c'est pour cette valeur que les histogrammes des profils d'occupations (Fig. 6.5(a)) ainsi que les fonctions d'autocorrélation (Fig. 6.5(a)) sont le plus proches de leur pendants expérimentaux.
3. Le potentiel chimique $\mu = -1.3$ est choisi pour obtenir une densité moyenne de $1nuc./190 - 200pb$.

Donc, à partir des séquence nous calculons le profil énergétique (Fig. 6.4) à partir duquel on peut calucler la densité puis l'occupation en nucléosome grâce à l'algorithme de Vanderlick (Chapitre 3). La comparaison des prédictions de notre modèle avce les données de Lee sur quelques contigs des chromosme de *S. cerevisiae* sont reportées à la figure 6.4. On constate à l'oeil un bon accord général avec tout de même une majorité de région mal prédites mais aussi avec des régions très bien décrites par ce modèle. Sur toute la levure, on obtient en moyenne une corrélation de Pearson

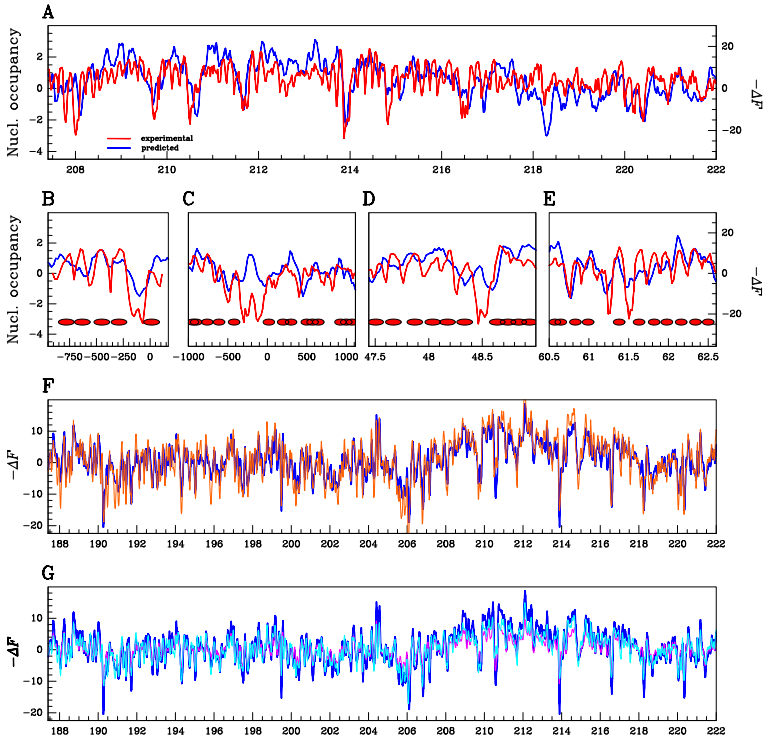


FIGURE 6.2 : *Comparison of energy profile $-\Delta E$ (blue line, ordinate units are in kT) as obtained by the Tolstorukov Tolstorukov* et al. *(2007) method with experimentally determined nucleosome occupancy (red line : log2ratio of hybridization data retrieved from Yuan* et al. *(2005)). The regions that were analyzed in Fig. 1 (A-E) are presented in panels A-E. The correlation between experimental data and $-\Delta E$ values over the whole chromosome III region analyzed is $r = 0.35,\ p < 10^{-15}$. (F, G) Comparison of the predicted $-\Delta F$ profile obtained with our model (dark blue) with refined anisotropic models : (F) the "crystal" nucleosomal model with no bending anisotropy ($A_2 = A_1 = 50t_m$, orange) for which the "ideal" superhelix trajectory of the nucleosomal DNA has been replaced by the trajectory derived from the crystal structure without changing the roll/tilt/twist of the Anselmi (Scipioni* et al.*, 2002b) parametrization ; (G) with the bend anisotropic model for which we have introduced a roll/tilt bending anisotropy ($A_1 = 2A_2 = 66t_m$) with either the "ideal" superhelix nucleosomal DNA model (magenta) or the "crystal" nucleosomal DNA model (cyan). In all cases, these anisotropic models, as the original one, consider only the roll/tilt/twist deformations. Thus, the shape of the coarse-grained energy landscapes are not significantly affected by using additional refinements of our simple model that either consider the "crystal" structure of Richmond (as the model of Tolstorukov (Tolstorukov* et al.*, 2007) does) or take into account an anisotropic roll/tilt flexibility. This does not imply that these parameters were properly described in the original model of Anselmi (Scipioni* et al.*, 2002b) but simply that they are not playing a determinant role at this scale.*

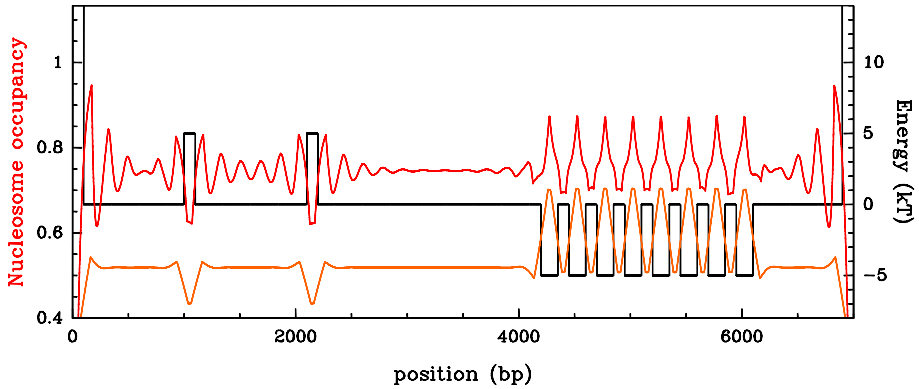


FIGURE 6.3 : *Illustration de l'effet de densité sur le profil d'occupation dans un paysage énergétique non uniforme : Le profil énergétique (noir) est constitué de deux barrières infinies aux bords, de deux barrières finies et d'une série de puits énergétiques. A partir de la relation de Percus et de l'algoritme de Vanderlick (Chapitre 4) on calcule la densité puis l'occupation en nucléosomes à faible (orange,* $1nuc./500pb$*) puis à forte densité (rouge,* $1nuc./200pb$*.) en considérant un volume exclus de* $146\ pb$

de $r \sim 0.3$, ce qui veut dire qu'en moyenne 10% du profil nucléosomal observé est en accord avec notre modèle.

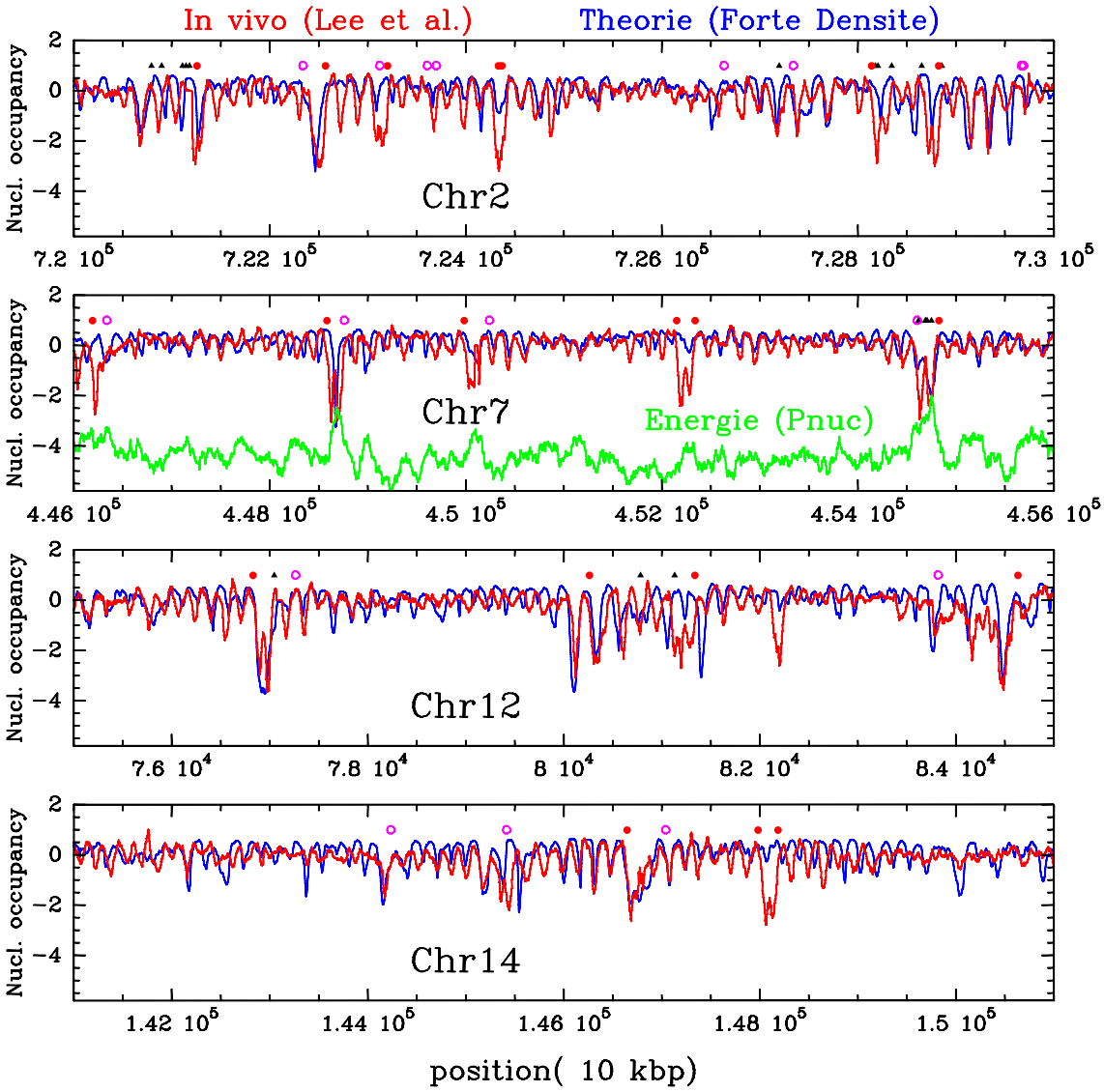


FIGURE 6.4 : *Comparaison entre notre modèle théorique haute densité (bleu, paramètres "Pnuc", $\mu = -1.3,\ \delta = 2$) et les données de Lee* et al. *(rouge) (Lee* et al.*, 2007a).En (b) est reporté le profil énergétique calculé avec les paramètres "Pnuc" (vert). Symboles : même signification qu'en Fig. 2.5*

– Levure : *S. Kluyverii*
  Notre modèle sur les données *in vivo* chez S. Kluyverii mais globalment chez l'ensemble des Hemiascomycota est aussi (ou aussi peu, selon l'humeur) "prédictif". On mesure ainsi en moyenne, une corrélation de Pearson de $r \sim 0.32$.
– Levure : *S. Pombe*.
  La comparaison avec les données de nucléosome de *S. Pombe* révèle ici (beaucoup) moins de corrélation. C'est d'ailleurs un des enseignements essentiels qu'en tirent Lanterman *et al.* dans leur papier (Lantermann et al., 2010). Heureusement que O. Rando n'a pas choisi cette levure pour ses premières études (Yuan et al., 2005)... Cependant, cela ouvre des perspectives d'étude et d'approdondissement de notre modèle de positionnement très intéressantes.

Au vu de ces analyses, on pourrait remettre en question la valeur de notre modèle énergétique, i.e. de sa dépendance vis-à-vis de la séquence génomique mais aussi de notre modèle d'assemblage à savoir de notre modèle d'interaction entre nucléosome (hypothèse sphère dure et d'interaction entre plus proches voisins).

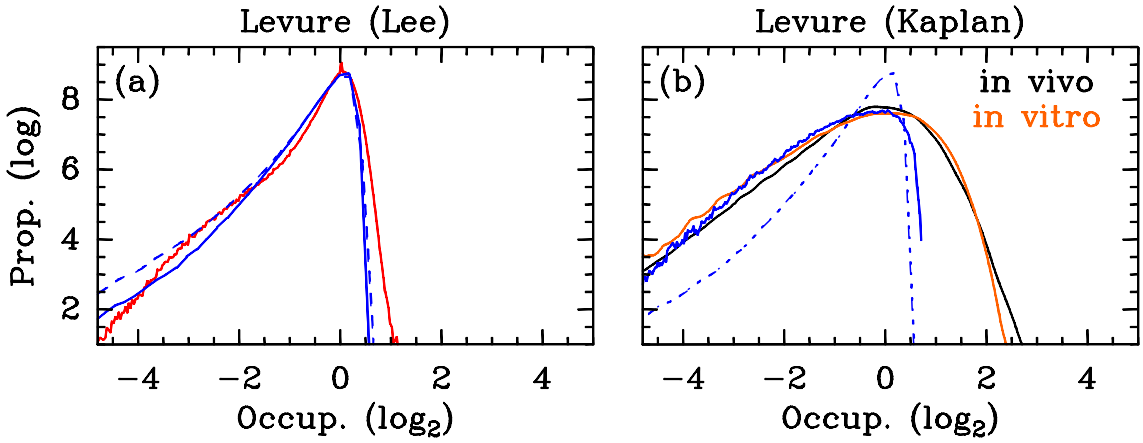


FIGURE 6.5 : *Comparaison entre les distributions statistiques des données expérimentales et les distributions des profils théorique (a) Données* in vivo *de Lee* et al. *(Lee* et al.*, 2007*a*) (rouge), prédictions de notre modèle à haute densité (Fig. 6.4) (bleu) et d'un modèle équivalent mais en incluant l'effet de barrières induites par la présence de facteurs de transcriptions (tirets bleus). (b) Données* in vivo *(noir) et* in vitro *(orange) de Kaplan* et al. *(Kaplan* et al.*, 2009*a*), prédiction à faible densité (Fig. 6.7) (bleu) et à forte densité (tirets-pointillés bleus)*

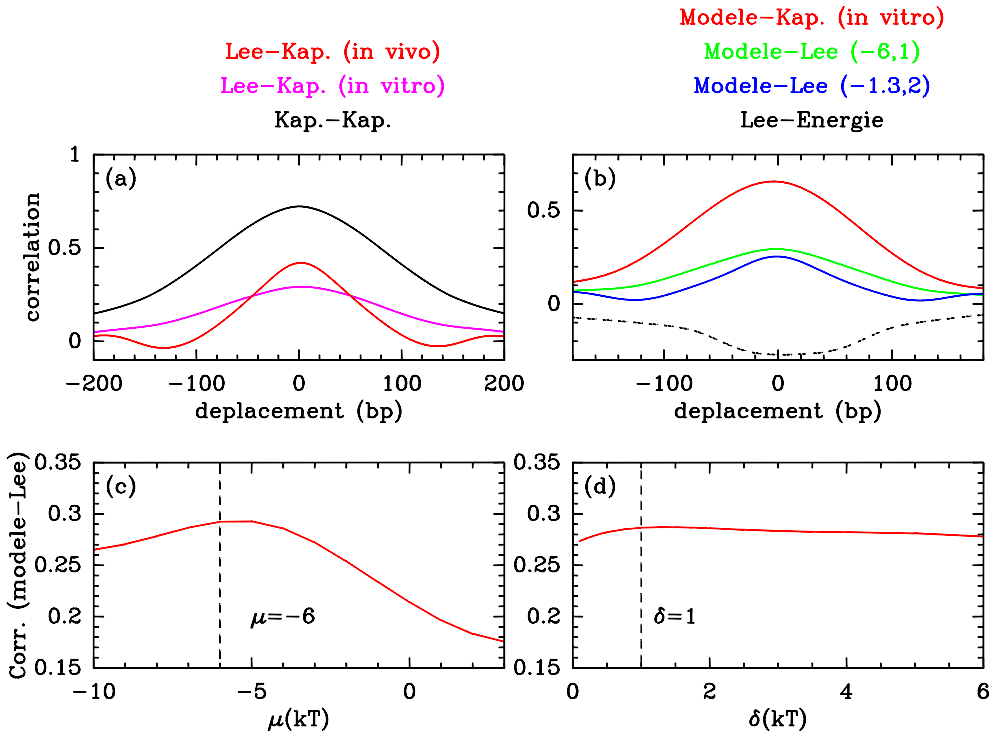


FIGURE 6.6 : *(a) corrélation entre les différentes données expérimentales : en noir corrélation entre les données in vivo et in vitro de Kaplan en fonction du déphasage (déplacement) entre les données. En rouge, corrélation entre Lee et Kaplan in vivo, en rose corrélation entre Lee et Kaplan in vitro. (b) En vert corrélation entre le modèle avec les paramètres* $(-6, 1)$ *kT et Lee ; en Bleu, corrélation entre le modèle avec les paramètres* $(-1.3, 2)$ *kT et Lee. En pointillé noir, la corrélation entre le profil énergétique et les données de Lee. (c) Évolution de la corrélation entre le modèle et les données de Lee, en fonction de* μ *à* δ *fixé à* 1 *kT. (d) Évolution de la corrélation entre les modèles et les données de Lee, en fonction de* δ *et à* μ *fixé (*-6 *kT).*

6.2 LEVURE IN VITRO

Un moyen d'évaluer la spécificité de séquence de l'octamère d'histone est de mesurer le profil d'occupation sur une chromatine reconstituée *in vitro* à faible densité où les effets de volume exclus (et plus généralement d'interaction nucleosome-nucleosome) sont faibles. A ce sujet la figure 5.24 indique comment la spécificité de séquence est modifiée (de facon non triviale) à mesure que la densité augmente. Les rapports de densités entre un point s_1 et un point s_2 sont guidés par des termes d'ordre de grandeur $e^{-\beta(E(s_1)-E(s_2))}$ si le milieu est suffisamment dilué (points rouges, pour $\delta = 1$ kT, figure 5.24 (b)). Si le milieu n'est pas dilué, alors la différence d'énergie entre deux points ne suffit plus à spécifier la différence de densité entre ces deux points, du fait justement des interactions entre particules. Comme l'indique la plus grande dispersion dans la relation, à haute densité, une même énergie de formation peut mener à des densités très différentes du fait d'un environnement énergétique différent. Ainsi toute méthode basée sur un apprentissage sur un jeu de données nucléosomales extrait d'une chromatine dense sera entachée d'une incertitude d'autant plus importante.

Kaplan *et col.* ont ainsi extrait les cartes genomique des nucléosomes sur une chromatine de levure reconstituée *in vitro* (Kaplan et al., 2009a) ; la reconstitution est effectuée par gradient de dialyses sur des fragments d'ADN génomique de la levure d'environ $10\ kpb$. L'assemblage par de telle méthode ne permet pas d'obtenir de très forte densités ; en l'occurence ici la densité moyenne obtenue est de $1nucl/500pb$. L'extraction des ADN nucleosomaux se fait de la même facon par digestion MNase puis par séquençage. Les données pour quelques régions du génome de la levure sont reportées en Figure 6.7. Comme en atteste le profil de corrélation de la figure 6.6 (b), modélisation et expériences *in vitro* sont très corrélés. Nous avons choisi les paramètres de la modélisation en concordance avec les données *in vivo* nous choisissons la même amplitude de variation liée à la séquence, à savoir $\delta = 2$ et nous diminuons simplement le potentiel chimique à $\mu = -6$ kT pour rendre compte de la diminution de densité moyenne sur les données de reconstitution. On obtient alors une corrélation de 0.74. Si on regarde directement les profils (figure 6.7), l'accord est alors plus que satisfaisant.Non seulement on capture les basses fréquences, mais l'intégralité du signal est reproduit. Nous avons donc toute confiance dans ce modèle pour reproduire les résultats d'une chromatine uniquement construite à partir de la séquence. Donc notre modèle énergétique capte de façon plus que satisfaisante la spécificité de séquence de l'octamère. Les différences observées entre notre prédiction à forte densité et les données *in vivo* sont donc le fruit d'actions extrinsèques. Il y a peut etre aussi une défaillance du modèle d'interaction et nous sommes actuellement en train de prendre en compte dans notre modèle thermodynamique l'effet d'une telle interaction qui porrait etre modulée épigénétiquement via les modifications des queues d'histones. L'idée étant ensuite de voir dans quelle mesure une interaction "non sphere dure" permettrait elle d'améliorer nos prédictions.

Remarque : Il y a une certaine bizarrerie dans les données de Kaplan, puique les formes des distributions *in vitro* et *in vivo* de Kaplan sont similaires (Fig. 6.5), bien qu'elles ne soit censées ne pas représenter les mêmes niveaux d'occupation ($\sim 1/200$ *in vivo* et $1/500$ *in vitro*). Il y a eu une normalisation des données pour qu'elles aient la même valeur moyenne, mais il est étonnant qu'elles aient la même dynamique, c'est-à-dire la même forme autour de la valeur moyenne (figure 6.5 (b)) : comme montré en annexe de la thèse de Guillaume, la distribution de la densité et de l'occupation change avec μ et donc avec la densité moyenne en nucléosomes. Or la chromatine *in vitro* est deux à trois fois moins dense que la chromatine *in vivo*. Comme cela est montré à la figure 6.5(b), la distribution du profil d'occupation *in vitro* est effectivement proche de celle associée au profil théorique à faible densité (bleu), contrairement donc au profil *in vivo* qui est significativement différent de ce à quoi on s'attendrait pour une chromatine à haute densité (rouges, tirets-pointillés).

6.3 CORRÉLATIONS À LONGUE PORTÉE : CONFIRMATION EXPÉRIMENTALE

Comme nous l'avions tout d'abord montré dans notre étude (Vaillant et al., 2007) pour les cartes nucléosomales de Yuan et al. (Yuan et al., 2005), les profils d'occupation en nucléosomes de Lee et al. (Lee et al., 2007a) et celles *in vivo* et *in vitro* de Kaplan *et al.* présentent des corrélations à longue portée comme

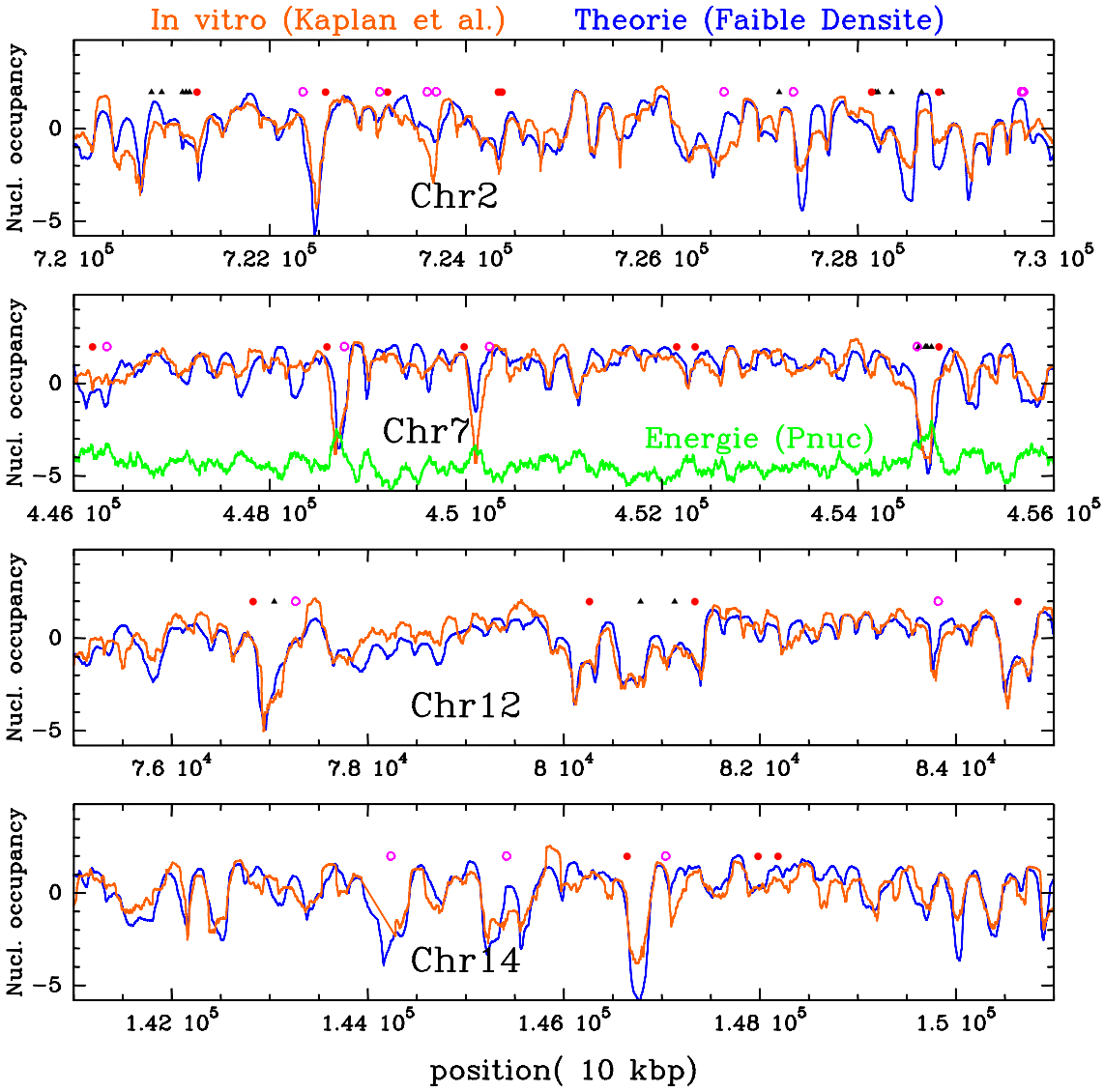


FIGURE 6.7 : *Comparaison entre notre prédiction à faible densité (bleu, paramètres "Pnuc", $\mu = -6$, $\delta = 2$) et les données* in vitro *de Kaplan* et al. *(orange) (Kaplan* et al.*, 2009*a*).En (b) est reporté le profil énergétique calculé avec les paramètres "Pnuc" (vert). Symboles : même signification qu'en Fig. 2.5*

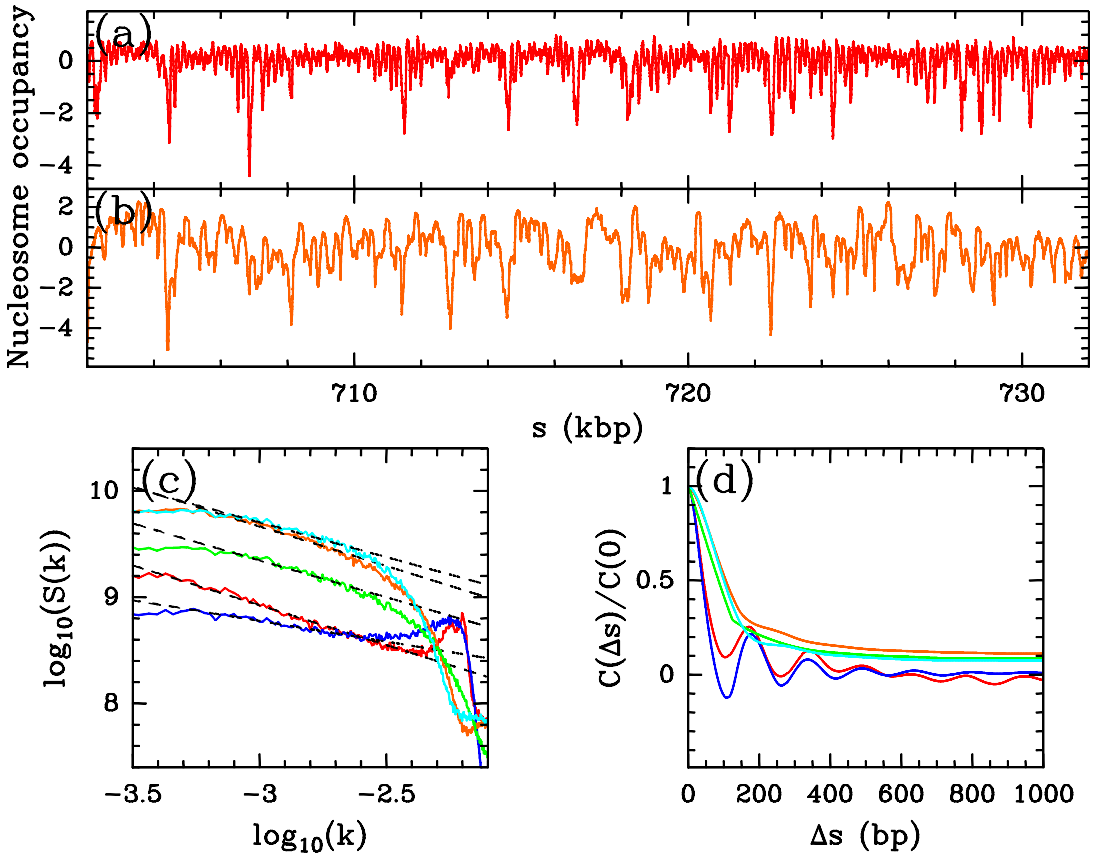


FIGURE 6.8 : *Nucleosome occupancy profile ($\delta Y(s) = Y(s) - \bar{Y}$, where $Y(s) = \log_2(P(s))$) along the yeast chromosome II : (a)* in vivo *data Lee* et al. *(2007b) (red) ; (b)* in vitro *data Kaplan* et al. *(2009b) (orange) ; (c) corresponding power spectra ; (d) autocorrelation function ($C(\Delta s) = \langle \delta Y(s) \delta Y(s + \Delta s) \rangle$). In (c) and (d) are also represented the results of our theoretical modeling (Sect.* **??***) : energy landscape (green) for $\delta E = \langle (E - \bar{E})^2 \rangle^{1/2} = 2$ kT ; nucleosome occupancy profile for $\delta E = 2$ kT and $\tilde{\mu} = \mu - \bar{E} = -1.3$ kT (dark blue) and -6 kT (cyan) so that the mean* in vivo *and* in vitro *nucleosome densities are correctly reproduced respectively. In (c) and (d) the data correspond to the results obtained when averaging over the* 16 *yeast chromosomes. In (c), the dashed lines correspond to the power-law scaling exponents $\nu = 0.65$, 0.74, 0.68, 0.74 and 0.46 from top to bottom corresponding to $H = 0.82$, 0.87, 0.84, 0.87 and 0.77 LRC properties respectively.*

l'atteste le comportement spectral en loi de puissance (linéaire en log-log, Fig. 6.8 (c)), dont l'origine peut s'expliquer via notre modèle thermodynamique : ces corrélations à longue portée sont simplement le reflet de celles présentes dans le profil énergétique elles-même induites par celles caractérisant les profils de courbure "Pnuc" le long des génomes. Comme l'indique la figure 6.3(a-b) et en acord avec ce qu'on avait démontré dans le cas des boucles 2D dans une chaine avec courbure intrinsèque désordonnées (chapitre 4, Fig. 4.14), les corrélations à longue portée induisent des profils énergétique avec une amplitude plus importante, des régions de plus grande affinité et des régions de plus faible affinité. On remarque également que ces corrélations induisent naturellement (cf. toujours boucles 2D, Fig. 4.14) un enrichissement vers les valeurs à plus haute énergie, donc vers les séquences défavorables ; dans le cas corrélé, la distribution est symétrique indiquant que barrières et puits (mesurés par rapport à la moyenne) sont équiprobables. La topograghie des profils est également différentes avec notamment une persistence plus importante (donc un profil plus "lisse") dans le cas du chromosome 7 de la levure (corrélé à longue portée) que dans le cas non corrélé. Cette persistence ($H = 0.8$) qui opère au delà de l'échelle du nucléosome induirait ainsi un confinement à plus longue distance que dans le cas non corrélé et renforcerait ainsi le phénomène de positionnement statistique. L'étude théorique plus précise de cet effet de corrélation sur le positionnement statistique est en cours. Notez qu' au chapitre 5, nous avons étudié et interprété l'effet des corrélations à longue portée présentes à petite échelle, $H = 0.54$.

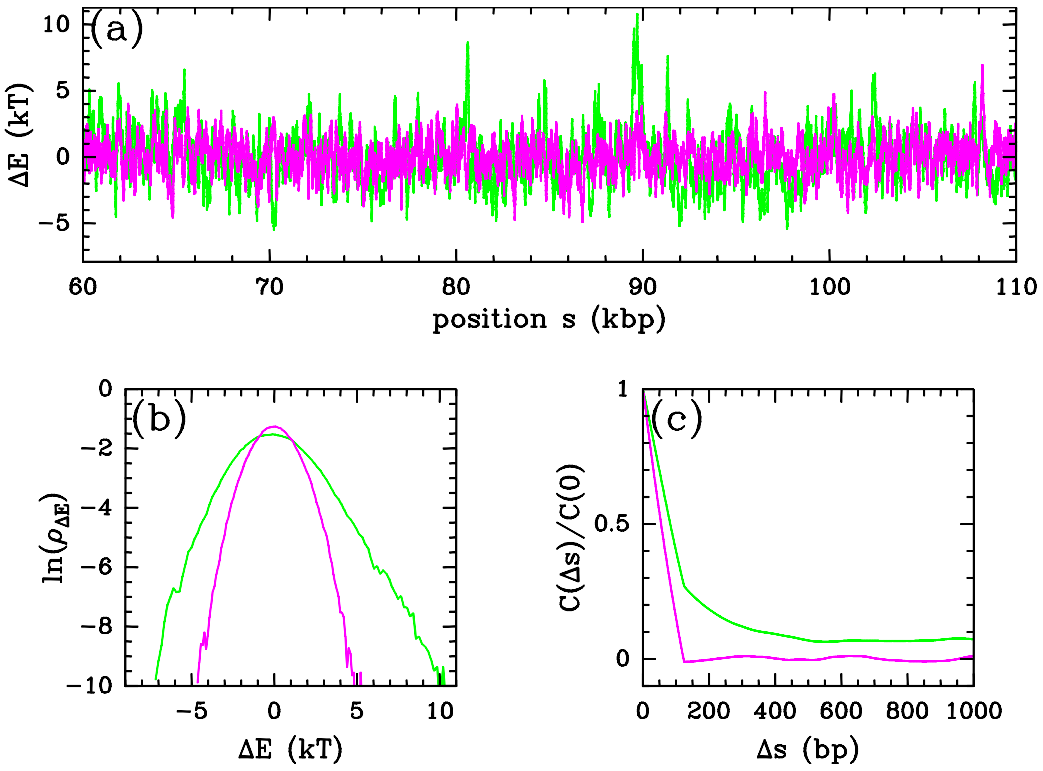


FIGURE 6.9 : *(a) Profils énergétiques, modèle "Pnuc", le long du chromsome 7 de* S. Cerevisiae *(vert) et d'une séquence aléatoire (de même composition moyenne en G+C) (magenta) (b) Distribution satistiques correspondantes (densite de proabbilité en ln-ln) (c) Fonctions d'autocorrélations correspondantes*

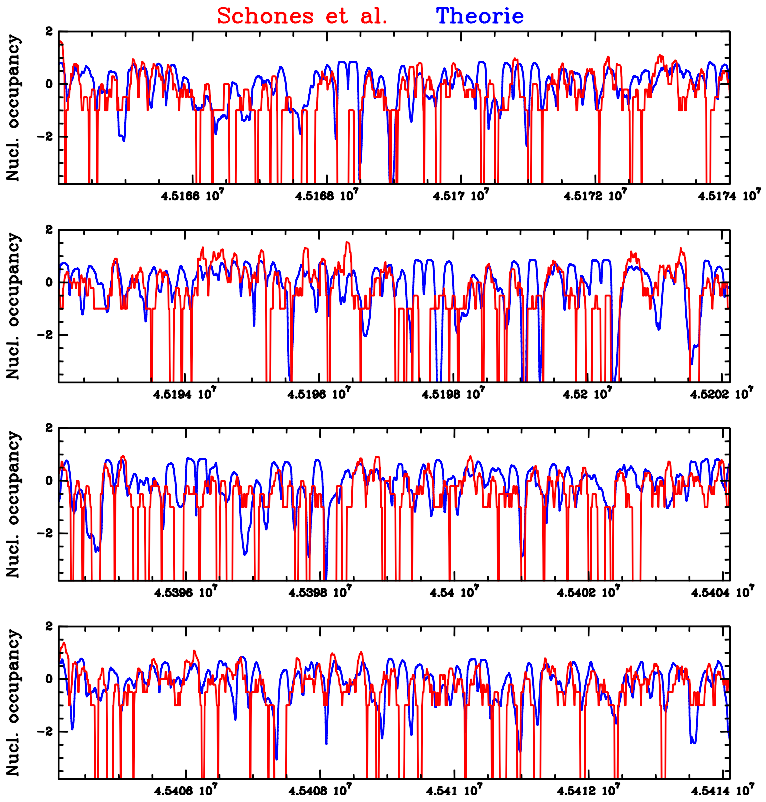


FIGURE 6.10 : *Comparaison entre les données de Schones (Schones* et al.*, 2008) pour les cellules T $CD4^+$ non activées le long du chromosome 6 (rouge) et les prédiction du modèle thermodynamique avec les mêmes paramètres que pour la levure :$\mu = -1.3,\ \delta = 2$ (bleu).*

6.4 C. ELEGANS

Comme l'indique le tableau 6.1, notre modèle d'occupation prédit toujours de façon assez modeste les données *in vivo* mais cela dit, un peu mieux que dans le cas de la levure : $r = 0.43 - 0.47$.

6.5 HOMME

Au vu les données de Schones (Fig. 2.12(c,d)) il parait difficile de faire étude comparative quantitative. Comme le montre cependant la figure 6.5, il semble cependant que l'accord est dans certain cas très correct. Si on compare la théorie et l'expérience, par exemple en moyenne au niveau des TSS l'accord est effectivement bon essentiellement dans le cas des gènes pauvres en CpG (Fig. 2.28, là où on n'observe pas de dépletion..., données non reportées ici). Une étude plus détaillée est en cours avec notamment les nouvelles données à la fois *in vivo* et *in vitro* de Valouev *et al.* (Valouev et al., 2011).

6.6 PERFORMANCES DES DIFFÉRENTS MODÈLES

On peut désormais s'intéresser aux "performances" relatives des différents modèles de positionnement "intrinsèque". Comme le confirme le Tableau 6.1 extrait de l'article de Tillo (Tillo and Hughes, 2009), la meilleure prédiction est globalement obtenue par le modèle publié par Kaplan en 2009 (Kaplan et al., 2009a), puisqu'il obtient une corrélation de Pearson de 0.89 avec les données de nucléosomes re-

Modèle								
	oligo-nuc. sur puce (contrôle)	oligo-nuc. (séquen-çage)	levure *in vitro*	levure *in vivo*	*C. elegans*	*C. elegans* (normalisé)	Homme (Schones)	$\%G+C$
Kaplan 2009	0.51	0.45	0.89	0.34	0.47	0.61	0.28	0.87
Lasso 2009	0.44	0.41	0.86	0.38	0.49	0.66	ø	0.85
Field 2008	0.47	0.45	0.74	0.39	0.46	0.61	ø	0.64
% G+C	0.53	0.49	0.78	0.25	0.42	0.47	ø	1
Peckham 2007	0.43	0.39	0.48	0.22	0.29	0.33	ø	0.57
Miele 2008	0.32	0.26	0.38	0.22	0.21	0.25		0.49
Tolstorukov 2007	0.01	0.004	0	-0.001	-0.001	-0.001		-0.0003
Cette étude	ø	ø	0.74	0.33	0.43	0.47	ø	0.86

TABLE 6.1 : *Performances comparées de différents modèles (apprentissage et énergétiques). Les modèles sont comparés à différents jeux de données : "Oligo-nuc." correspond à la préference du nucléosome pour différents oligo quantifiés sur différentes séquences, préférence mesurée par hybridation ou séquençage (pour s'affranchir de l'effet de la MNase). Les données de la levure sont obtenues in vitro par séquençage Kaplan. Les données in vivo sont obtenues par hybridation sur puce par Lee (Lee* et al.*, 2007a). Les données de C. elegans sont tirées de l'article de Valouev (Valouev* et al.*, 2008). Le modèle Lasso, tiré de l'article de Tillo et Hughes (Tillo and Hughes, 2009) est lui aussi un modèle d'apprentissage. Field est un modèle similaire à Kaplan (Field* et al.*, 2008a). Peckham est un autre modèle construit par apprentissage, plus ancien (Peckham* et al.*, 2007). Miele est un modèle énergétique simplifié exploitant la déformation du brin d'ADN seule (Miele* et al.*, 2008). Tolstorukov est un modèle physique très complet (Tolstorukov* et al.*, 2007). Enfin cette étude exploite un modèle similaire à Miele 2008, mais avec une table de coefficients tirée de Goodsell et Dickerson (Goodsell and Dickerson, 1994; Satchwell* et al.*, 1986). Tableau adapté de (Tillo and Hughes, 2009)*

constitués sur la levure (levure *in vitro*) et 0.34 avec les données *in vivo* de Lee *et al.*. Notons toutefois que le modèle est "entrainé" justement à partir des données expérimentales. Les performances des autres modèles d'apprentissage sont d'ailleurs très similaires (Field, Tillo, Peckham). Notre modèle (Vaillant) obtient également des performances très proches de celles de Kaplan. Notre autre modèle énergétique utilisant la table d'Anselmi (Miele) obtient des performances légèrement moins bonnes (corrélation de Pearson de 0.38 sur la levure *in vitro*), mais il seste satisfaisant. Enfin, le modèle construit physiquement par Tolstorukov *et al.* (Tolstorukov et al., 2007) n'obtient pas de bon résultats (corrélation de -0.001). Dans tous les cas, il est difficile d'obtenir des prédictions concordantes avec les données expérimentales (Morozov et al., 2009), à partir des modèles purement physiques construits *ab initio*. En conclusion, on voit donc qu'au mieux, tout modèle confondu, la séquence compte en moyenne pour 16% dans les profils d'occupation observé *in vivo* par Lee et Kaplan pour la levure *S. Cerevisaie* et de $25-44\%$ pour C. elegans.

Il faut cependant se méfier des valeurs moyennes. Nous avons ainsi sur la levure S. Cerevisiaie calculé les corrélations de Pearson entre notre modèle d'occupation et les données de Lee sur des fenêtres glissantes de $1000\ pb$ sur tout le long du génome. L'histogramme de ces valeurs est reporté en figure 6.11(a) (points bleus) et est comparé à un contrôle aléatoire (points noirs). On remarque bien que si en moyenne la corrélation de Pearson est de $r \sim 0.3$ il y a une part significative de régions avec des corrélations plus fortes (mais aussi des régions beaucoup moins corrélées et anti-corrélées). En effet comme l'indique la figure 5.25 pour le locus Gal4 mais aussi la figure 6.4 notamment autour de la position $7.7\ 10^4\ pb$ du chromosome 12, nos prédictions théoriques se superposent très bien avec les données expérimentales et ce sur des régions assez étendues (au moins $1000pb$). Il y a donc plus d'effets de séquence qu'on ne le pense. Un même effet est observé chez C. Elegans (Fig. 6.12).

6.7 CONCLUSION SUR LA PERTINENCE DU MODÈLE

Notre modèle énergétique (comme ceux de Kaplan, Lasso, Yuan) donne une excellente prédiction pour les expériences de reconstitution *in vitro*, moins bonne en moyenne *in vivo* mais avec des régions tout de même très bien prédites, comem par exemple Les différences que l'on observe sur les données

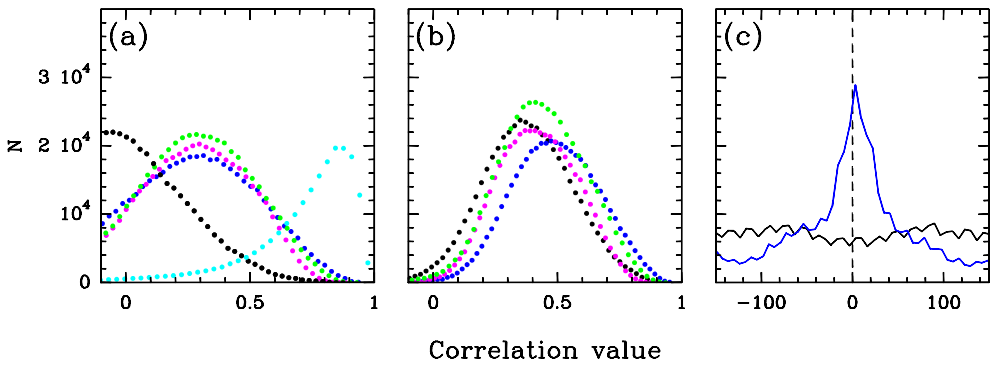


FIGURE 6.11 : *Histogramme des valeurs de corrélation de Pearson entre les données de Lee et al. (Lee* et al.*, 2007*a*) et nos prédictions à haute densité ((a,b) points bleus), l'énergie (en fait $-E(s,l)$) ((a,b), points verts), le N-score de Yuan (Yuan and Liu, 2008*a*) ((a,b), rose) et un contrôle aléatoire ((a,b), noir) mesurées dans une fenêtre glissante de taille* $1000\ pb$ *: (a) pour un décalage* $d = 0$ *; (b) pour le décalage* $-200 < d < 200$ *qui maximise la corrélation ; (c) histogramme de ces décalages "optimaux"* d*, pour nos prédiction (bleu) et pour le contrôle aléatoire (noir). En (a), histogramme des valeurs de corrélation de Pearson entre les données* in vitro *de Kaplan* et al. *(Lee* et al.*, 2007*a*) et nos prédictions à faible densité (points bleus clairs).*

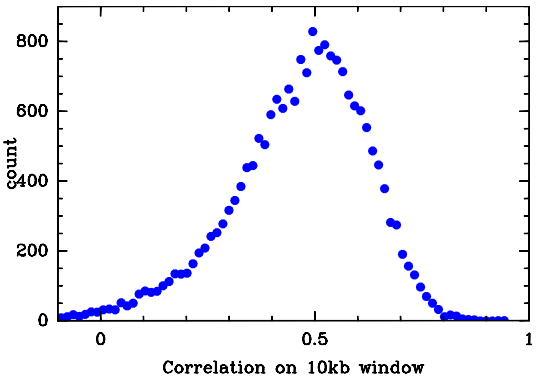


FIGURE 6.12 : *Histogramme des valeurs de corrélation de Pearson entre les données de Valoeuv et al. (Valouev* et al.*, 2008) et nos prédictions à haute densité (mes paramètres que la levure) mesurées dans une fenêtre glissante de taille* $10000\ pb$*.*

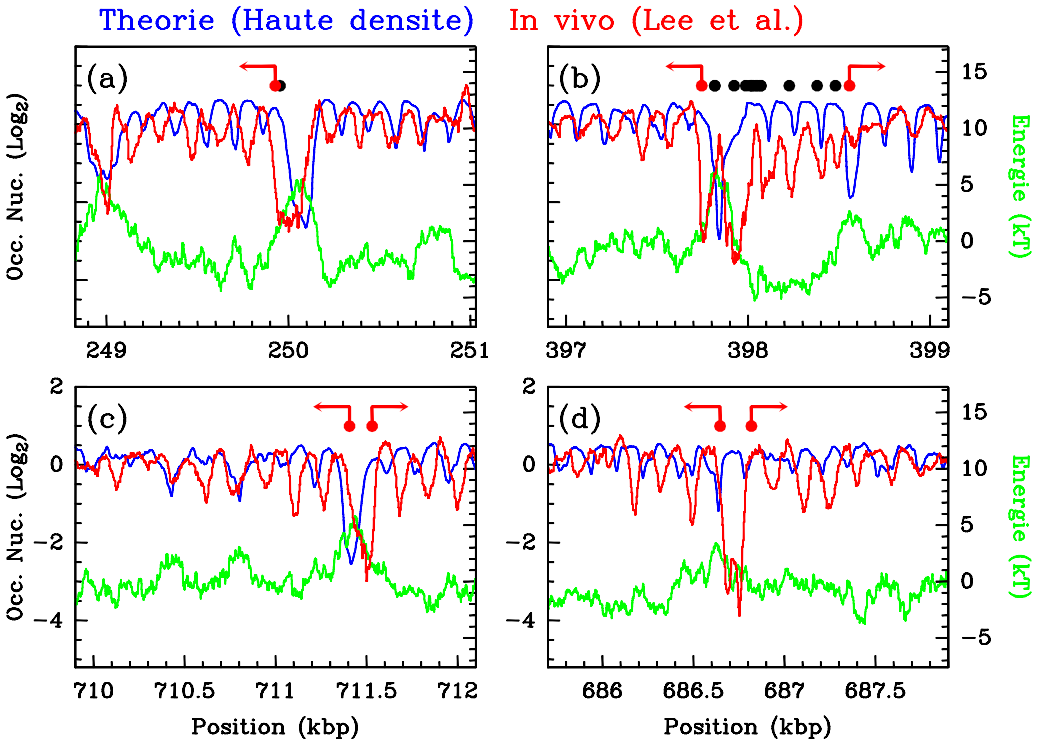


FIGURE 6.13 : *Des exemples de profils d'occupation théoriques (en bleu) similaires aux profils expérimentaux (Lee* et al.*, 2007*a*) (en rouge) à un remodelage globale près. Les gènes sont indiqués par les points (TSS) et flèches rouges. Les sites de fixation des facteurs de transcription sont marqués par les points noirs.*

in vivo ne peuvent donc être interprétées que par des phénomènes extérieurs à la séquence. On pourrait imaginer que le modèle soit bon à basse densité et mauvais à haute densité, comme c'est le cas *in vivo*. Toutefois, ceci serait peu probable, puisque que comme le montre la figure 5.25, les positions changent finalement assez peu avec le potentiel chimique. Il est cependnat possible qu'introduire des intéractions de type attractives ou d'autoriser une longueur d'enroulement variable autour de l'octamère puisse un peu améliorer nos prédictions. On sent bien surtout que l'essentiel des différences entre notre modèle actuel *in vitro* et les données *in vivo* est le fruit de l'action extrinsèque de facteurs. C'est particulièrement clair sur les profils reportés à la figure 6.4, par exemple pour le chromosome 2 en position $7.245\ 10^5\ pb$, le chromosome 7 en position $4.52\ 10^5\ pb$ et le chromsome 14 en position $1.48\ 10^5$ pb où le modèle peine à prédire les déplétions observées *in vivo* qui sont a priori induits par les complexes transcriptionels (dans les deux cas ce sont des promoteurs de gènes divergents). Il y a effectivement quel que soit l'organisme une anticorrélation importante entre les données de position de polII et les données de nucléosomes (Schones et al., 2008).

Comparer ainsi nos modélisations à haute densité et les donnée *in vivo* peut nous renseigner sur les mécanismes extrinsèques qui agissent et modèlent *in fine* le chapelet nucléosomal. Si les profils expérimentaux *in vivo* ne concordent pas tout le temps avec la prédiction, c'est que d'autres facteurs que la séquence interviennent. Toutefois, il est très fréquent que les profils nucléosomaux générés par la séquence soient similaires au profil expérimental à un léger décalage près. La figure 6.13 présente quelques exemples, situés à proximité des gènes de la levure. Il suffit parfois de décaler légèrement ces profils pour obtenir un "fit" correct des données (figure 6.13 (a) et (c)). Dans d'autre cas, il faut parfois invoquer la suppression d'un nucléosome et un remodelage plus complexe (figure 6.13 (b) et (d)). Enfin, certains cas révèlent très bien la compétition entre les facteurs de transcription et les nuclésomes comme l'illustre la figure 6.13(b).

Pour rendre compte statistiquement de ces "remodelages" du chapelet, on a calculé la corrélation des profils expérimentaux et théoriques dans des fenêtres glissantes de 1 kb en autorisant désormais un décalage des signaux (figure 6.11 (b)). La corrélation moyenne augmente évidemment, et ce autant que dans le contrôle aléatoire. Le contrôle est ici réalisé en corrélant les données expérimentales avec des

profils générés par des séquences choisies ailleurs sur le chromosome. Mais le décalage nécessaire pour obtenir ces corrélations est réparti aléatoirement pour le contrôle, alors qu'il est très fortement piqué autour de quelques paires de bases seulement pour notre modélisation (décalage typique de l'ordre de 60 pb). L'interprétation que l'on peut donner ici, est que la séquence via la spécificité de séquence des histones produit naturellement une chromatine très proche de ce qui est nécessaire pour la cellule.

D'ailleurs si on regarde la corrélation de nos prédictions avec les données expérimentales mais à des échelles plus grandes on s'aperçoit (Fig. 6.14) que la corrélation augmente atteignant 0.6 à l'échelle de $1000\ pb$, échelle où les profils sont moins sensibles à des décalage locaux de quelques dizaines paires da base. La séquence interviendrait donc plus globalement en controlant la densité à plus grande échelle. Nous avons enfin essayé d'améliorer nos prédiction en incluant l'effet local de déplétion induit par les facteurs de transcription. La modélisation très basique qui assume une barrière énergétique fixe aux sites de fixations des facteurs de transcription en plus du profil énergétique induit par la séquence permet tout de même d'améliorer les performances des prédictions avec des corrélation de $r = 0.4$ et $r = 0.65$ à une échelle de $1000\ pb$ (Fig. 6.14).

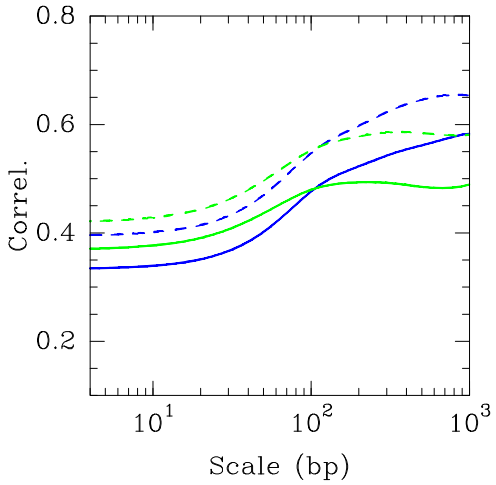


FIGURE 6.14 : *Etude multi-échelle des corrélationc de Pearson entre notre prédiction à aute densité (courbe bleue), celle incluant l'effet des facteurs de transcription (via une barrière énergétique) (courbe bleue en tirets), le profil énergétique (en fait le gain $-E(s,l)$, vert), le profil énergétique avec les barrières au niveau des sites de facteurs de transcription (tirets vert) et ce, calculé à différentes échelles (par transformée en ondelette, gaussienne).*

6.8 GC : ARTEFACT OU RÉALITÉ ?

Dès les données génomiques de Yuan *et col.* en 2005 (Yuan et al., 2005), il nous était apparu que le profil d'occupation en nucléosomes était très corrélé aux variations du contenu en G+C (Miele et al., 2008). Comme l'indique le tableau 6.1, quel que soit le jeu de données, le G+C est un excellent "prédicteur" de l'occupation (intrinsèque) en nucléosomes, suggérant ainsi que l'affinité augmente avec la teneur en GC. C'est d'autant plus frappant avec les données *in vitro* de Kaplan *et al.* pour lesquelles on mesure une corrélation de 0.78 ! *In vivo* cette corrélation reste également très forte dans le cas des données de Kaplan ($r = 0.4$) et dans une moindre mesure avec les données de Lee ($r = 0.25$). De tous les motifs nucléotidiques le simple contenu en G+C est celui qui contribue le plus au positionnement "intrinsèque" du nucléosome, à la fois *in vitro* (pour 50%) et *in vivo* (10%). Du coup tout modèle conduisant à une affinité fortement corrélée au G+C (comme nos modèles "Anselmi" et "Pnuc") est un modèle qui décrit bien la spécificité de séquence (ou "intrinsèque") observée expérimentalement. Comme le confirme les récentes études de Field (Field et al., 2008a) et Tillo (Tillo and Hughes, 2009), un autre facteur important est la composition en poly(dA :dT), connus pour être plutôt défavorables à la formation du nucléosome car plus rigides et qui sont effectivement surepresentés dans les régions de forte déplétion, en l'occurrence

les promoteurs chez la levure (Yuan et al., 2005). La mesure de l'affinité sur des oligo synthétiques de $150\ pb$ (quantification sur puces ou par séquençage) confirme ces effets de composition : entre 20 et 60% de G+C, l'affinité intrinsèque augmente à peu près d'un facteur 5, ce que revient ici à un ecart moyen de $1.6kT$ entre les séquence à bas GC, plutôt défavorables au nucléosome et les séquences à fort G+C, plutot favorables (la variation maximale est de $\sim 3kT$) (Fig. 6.15 A). Quant aux poly-A, ils induisent une déplétion d'autant plus importante que leur taille et leur degré homopolymérique est importante ; entre une taille de 5 et 15 pb, ce qui correspond à la gamme de taille des poly-A dans les séquences naturelles, on observe un niveau de dépletion qui va d'un facteur 2 ($-0.7kT$) à $5-6$ ($-1.7kT$) par rapport à une séquence aléatoire (Fig.6.15 B). Pour un poly-A de $25\ pb$ le niveau de déplétion est d'un facteur 30 soit $-3.4kT$ (Fig. 6.16). On voit donc à partir de ces données in vitro (et MNase-indépendante) qu'on est, en accord par ailleurs avec les expériences antérieures de compétition in vitro (cf Chapitre 4) dans une gamme d'affinité ou de variabilité énergétique pour les séquences génomiques assez faible (comparativement au 601, par exemple) mais non négligeables.

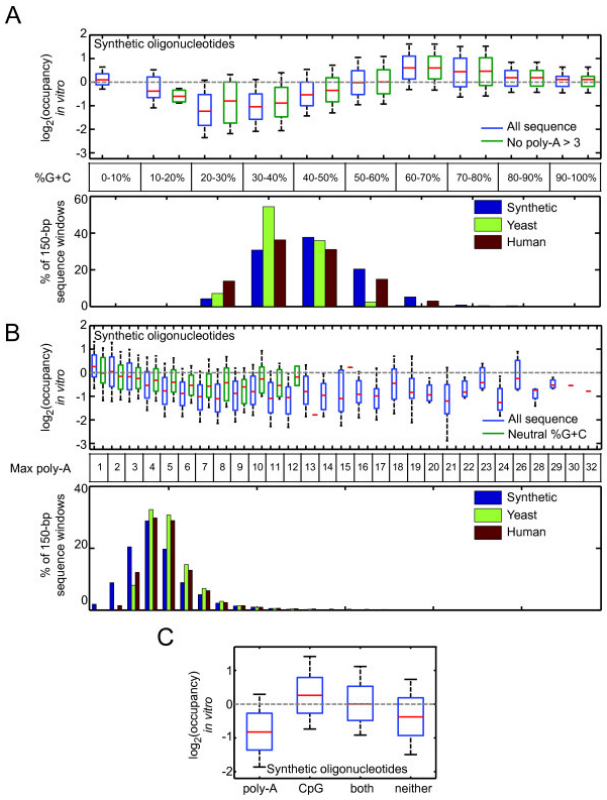


FIGURE 6.15 : *Relative nucleosome preference of different subsets of synthetic 150-mers. (A) and (B) Dependence of relative nucleosome preference (as log2(occupancy ratio)) on G+C content (A) and maximum poly-A length (B). Oligonucleotides categorized as "Neutral %G+C" in (B) are those with $45-55\%$G+C. Graph below shows the frequency of the selected attribute in the oligonucleotides analyzed, and also the human and yeast genomes. (C) Dependence of relative occupancy on poly-A content and CpG status. Poly-A containing oligonucleotides are defined as containing at least four consecutive adenine bases. CpG oligonucleotides are defined as having a G+C content $> 50\%$, with an observed/expected CpG ratio > 0.6 (Obs/Exp CpG = Number of CpG * N/(num G * num C), where N = length of sequence (Gardiner-Garden J. Mol. Biol.* **196***, 261-282 (1987))). The sequencing readout (rather than array readout) data from the Kaplan paper was used in this analysis. On all box plots, whiskers indicate 10th and 90th percentiles.*

Ce lien avec le GC, bien qu'évoqué très tôt déjà dans nos travaux n'a pas été vraiment compris. Plusieurs pistes sont possibles :

1. C'est un artefact experimental : il est en effet connu que la MNase presente une spécificité de séquence assez forte coupant préférentiellement les dinucleotides AT. Comme l'a récemment montré Chung *et col.* (Chung et al., 2010) et Fan *et col.* (Fan et al., 2010) très récemment, les "profils" issus de la digestion MNase sur de l'ADN nu sont effectivement très corrélés aux fluctuations du

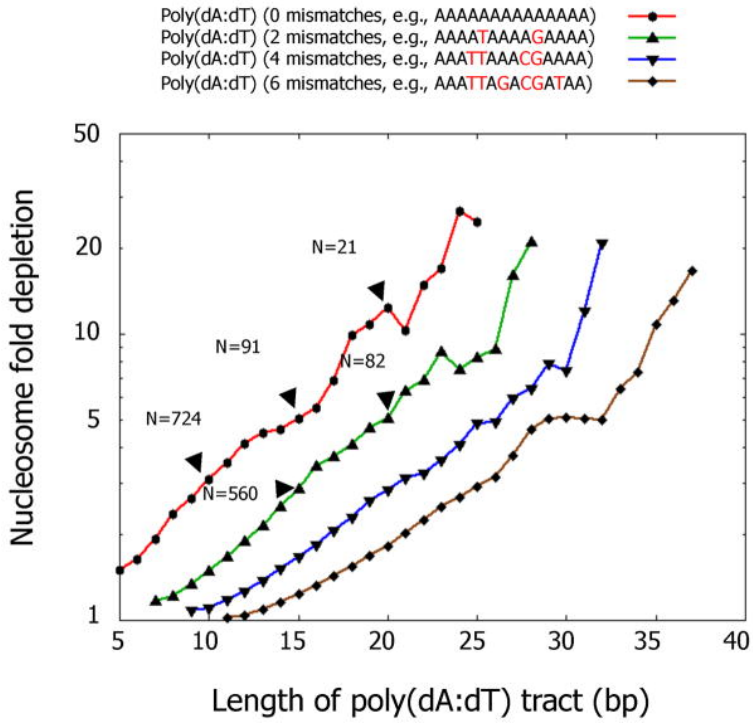


FIGURE 6.16 : *Nucleosome are relatively depleted over poly(dA :dT) tracts in vivo. Shown is the combined nucleosome fold depletion over all poly(dA :dT) tracts of length k, for $k = 5, 6, 7, ...$, and for tracts with exactly 0, 2, 4, or 6 base substitutions. Each graph is trimmed at a length K at which there are less than 10 such tracts in the S. cerevisiae genome, and the fold depletion at this final point is computed over all elements whose length is at least K. The number of underlying elements at various points in the graph is indicated (N).*

G+C (contenu en G+C sur fenêtre de taille 147 pb) et aux profils de digestion des chromatines *in vitro* (quel que soit la méthode de reconstitution) et *in vivo*. Comme nous l'avons montré également (Miele et al., 2008), dans la levure que les profils de digestion par la MNase de l'ADN génomique au niveau des promoteurs sont en effet fortement corrélés aux données nucléosomales. Une manière de s'affranchir de ce "biais" est donc de normaliser les données notamment avec l'ADN nu digéré. On peut aussi ne s'interesser qu'au positionnement et non au profil d'occupation complet. A priori, s'il y a biais il est probable que les données in vitro a faible densite soient le plus affectées.

2. On peut imaginer aussi que le taux de G+C détermine en partie les propriétés mécaniques essentielles au positionnement du nucléosome. Cette corrélation n'est pas relevée dans la plupart des modèles énergétiques qui prennent en compte l'intégralité des interactions nucléosome-ADN (Lavery, Morozov). Seuls ceux qui ne s'intéressent qu'à la déformation du brin d'ADN (Miele, Vaillant, Travers) révèlent une corrélation avec le taux de G+C et donc avec les données expérimentales. Dans le modèle Miele, avec paramètre Anselmi, c'est la contribution venant du roll intrinsèque qui domine et qui est effectivement positivement corrélé au G+C...

6.9 PÉRIODICITÉ

Les modèles présentés ici qui se basent uniquement sur la périodicité à 10 pb (Ioshikhes, Segal-2006) des dinucléotides AA/TT/AT et GC rendent en fait assez mal compte du positionnement observé *in vivo*.

En outre, les séquences synthétiques sélectionnées artificiellement pour leur forte affinité avec le nucléosome présentent une très forte variation dans le profil du contenu en nucléotide (variation de

la composition de 0.1 à 0.5). Ces séquences, au nombre desquelles on compte le 601 (Lowary and Widom, 1998; Thåström et al., 1999) – séquence issue d'une sélection drastique fondée sur l'énergie de formation du complexe nucléosome-ADN (Lowary and Widom, 1998)– sont très particulières, de part leur affinité hors norme avec les nucléosomes. À titre d'indication, la différence d'énergie de formation du complexe ADN-nucléosome est estimée à -4.9 ± 0.55 kT (Lowary and Widom, 1998) par rapport à une séquence de référence (la séquence "5S"). Toutefois, il n'existe pas d'organisme connu contenant une telle séquence, sans doute parce que l'affinité avec du nucléosome avec celle-ci serait trop forte et qu'une fois positionné, il serait trop difficile d'altérer ce nucléosome, ce qui est parfois nécessaire lors de la réplication et de la transcription. Les séquences qui accrochent les nucléosomes sur des organismes réels ont certes le même type de variations dans le contenu nucléotidique, mais avec des amplitudes très faibles (amplitude de variation du contenu nucléotidique de l'ordre de $5 - 10\%$ chez le poulet, et $3 - 12\%$ chez la levure (fig. 4.23 (a) et fig. 4.24 (a))). Ce que traduit ce profil, c'est qu'effectivement, les nucléosomes se fixent naturellement mieux – c'est-à-dire plus souvent – sur une séquence dont le profil en contenu "AA/TT/AT" est oscillant avec une période 10 pb, et où les "AA/TT/AT" sont positionnés dans les sillons mineurs.

Toutefois, dans une étude *in vitro* récente menée à la fois sur la levure et sur la bactérie *E. coli* qui n'héberge pas de nucléosomes puisqu'il s'agit d'un organisme procaryote – sans noyau –, Zhang et Struhl (Zhang et al., 2009) montrent que les périodicités associées à la formation du nucléosome n'existent que dans une très faible proportion au sein du génome des organismes vivants. Certes la séquence moyenne sur laquelle se fixent les nucléosomes présente une périodicité (figure 4.24 (a), (b)). Les nucléosomes préfèrent d'ailleurs une séquence dont le profil en AA/TT/AT présente des périodicités sauf si l'on rajoute le facteur de remodelage –ACF– et une chaperonne -NAP-. Le rôle de la chaperone est de faciliter l'absorption des nucléosomes sur n'importe quelle séquence tandis que le facteur de remodelage ACF équirépartit équitablement les nucléosomes (Shundrovsky et al., 2006; Nakagawa et al., 2001). Le positionnement devient quasi-équiprobable, ce qui explique l'aplatissement du profil. Mais cette périodicité ne semble pas particulièrement présente dans les séquences réelles (figure 4.24 (d)). Au vu des spectres de Fourier, il existe bel et bien une légère périodicité à 10 paires de bases dans le génome de la levure (figure 4.24 (d)), mais elle existe aussi (plus vers 11 pb) chez la bactérie *E. coli* qui ne possède pas de nucléosomes ! Si périodicité il y a, elle est sans commune mesure avec la périodicité induite par le code génétique (pic visible à 3 paires de bases). Il existe donc une légère différence entre le spectre de *E. Coli* et celui de la levure, mais il semble difficile d'imputer cette différence à la présence des nucléosomes. La différence entre les deux spectres est beaucoup plus prononcée dans l'intervalle de périodes $[4 - 10]$ pb par exemple. L'interprétation de la différence entre levure et la bactérie est sujette à caution tant les deux pics à 10.2 et 11.2 émergent faiblement du bruit de fond du spectre. Le nucléosome a vraisemblablement une légère préférence pour un motif périodique en dinucléotides "AA/TT/AT" puisque clairement les séquences sur lesquelles il se fixe facilement ont cette périodicité. Cependant, les génomes n'exploitent pas ou peu cette préférence pour positionner les nucléosomes puisque cette périodicité n'est pas visible dans les spectres.

Toute l'analyse comparative précédente a été faite en considérant le profil d'occupation ; on peut également comparer les positions préférentieles des nucléosomes extraites de ces profils d'occupations :

6.10 NUCLEOSOMES BIEN POSITIONNÉS

On peut ainsi étudier dans quelle mesure les nucléosomes bien positionnés *in vivo* sont prédits par notre modèle. Par soucis de comparaison avec les prédictions d'autres groupes Yuan and Liu (2008a); Albert et al. (2007); Segal et al. (2006); Ioshikhes et al. (2006), nous utilisons une approche similaire qui consiste à appliquer un algorithme de prédiction par chaine de Markov cachée (HMM, "Hidden Markov Chain") sur nos simulations pour obtenir un jeu de nucleosomes bien positionnés. La comparaison avec le jeu obtenu par Lee *et al.* Lee et al. (2007a) sur leurs données expérimentales est reportée à la Figure 6.17(a). Notre modèle prédit 48.7% de "vrais" positifs à une distance de 35 *bp* près qu'on peut comparer aux 42% attendus par "chance". Ce résultat pour les nucléosomes positionnés, ne révèle rien d'autre que la "distance de remodelage" qui a été quantifié précédemment pour le profil d'occu-

pation (Fig. 6.11). Dans ce cas, cette distance caractéristique est environ de $35\ pb$. Cette comparaison indique qu'en moyenne notre modèle, tout comme pour le profil d'occupation, ne prédit que modestement la positions des nucléosomes bien positionnés en accord avec les conclusions de Yuan *et al.* (Yuan and Liu, 2008a) et Peckham Peckham et al. (2007). Si par contre on calcule le performance (à $35pb$ près) dans une fenêtre glissante de taille $5000\ bp$, on observe un comportement bien différent du cas "aléatoire" comme l'indique la Figure 6.17(b) : si la performance est homogènement distribué autour d'une valeur de 0.4 pour le contrôle aléatoire, il y a une plus grande hétérogénéité pour nos prédictions avec un grand nombre de régions génomique où la prédiction est nettement meilleure que dans le cas aléatoire et d'autres pire. C'est, encore une fois, le reflet de ce qui a été observé pour les profils d'occupation.

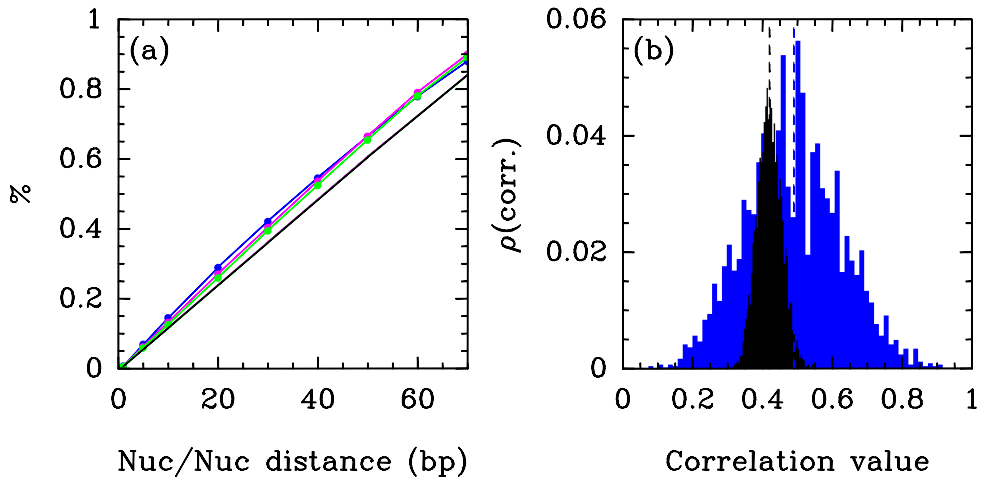


FIGURE 6.17 : *Performance of our theoretical model and of Yuans N-score model in terms of well positioned nucleosome as obtained by HMM methods. The comparison is made against the set of well positioned nucleosomes obtained by Lee* et al. *on their genomic wide experimental data using a similar HMM algorithm (Lee* et al.*, 2007*a*). Performance is measured by the proportion of true positive* i.e. *well predicted positioned nucleosomes at a given overlapping distance of an experimental nucleosome. (a) Mean performance value* vs. *the overlapping distance for the model (blue) the energie model (green) and the Yuan N-score model (magenta). (b) Statistics of the performance values (at* $35\ bp$ *accuracy) computed on a sliding window of size* $5000\ bp$ *along the entire genome for the prédictions (blue) and for the random control (black). The vertical dashed line (black and blue) indicates the corresponding expectation values.*

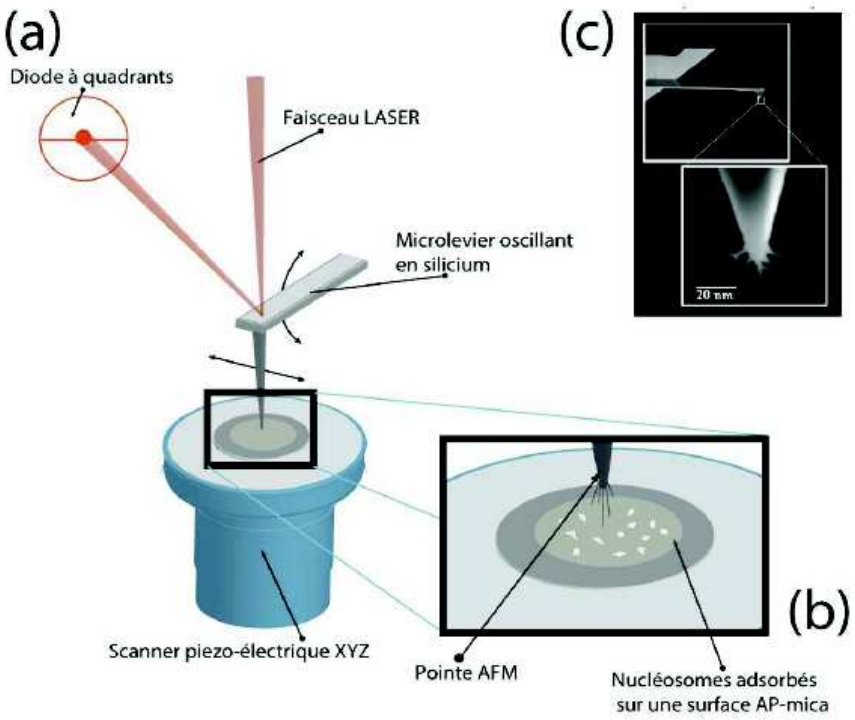


FIGURE 7.1 : *Principe de la visualisation par microscopie AFM : (a) Schéma général : un faisceau LASER permet d'évaluer la déviation de la pointe de silicium (c) induite par les forces de Van de Waals au contact avec les nucléosomes absorbés sur une surface très plane (mica, (b)) (Montel, 2008).*

7 VISUALISATION DIRECTE DE NUCLÉOSOMES SUR DES SÉQUENCES GÉNOMIQUES.

Observés directement par microscopie à force atomique en milieu liquide, les nucléosomes se fixent conformément aux prédictions du modèle énergétique sur des petits fragments d'ADN. Les résultats sont cohérents avec les méthodes biochimiques menées sur des séquences similaires.

Under direct AFM visualization in liquid, nucleosomes reveal a positioning compatible with the predictions of our model.

L'objectif ici, est notamment de s'affranchir des biais de MNase mais aussi pour travailler à l'echelle de la molécule unique.

7.1 L'AFM ET LA VISUALISATION DE NUCLÉOSOMES.

L'AFM est un outil tres performant pour mesurer le positionnement des nucleosomes le long d'un frament d'ADN. Apres reconstitution en volume par bains de dyalises successives (ou autre methode) les mono- di-...le nucléosomes sont déposés à surface et topographiés par la pointe AFM. La mesure des hauteurs permet de localiser le nucleosome le long du brin (Milani et al., 2009). Nous proposons ici une étude relativement originale, puisque nous étudions des séquences génomiques (extraites de la levure et de l'homme) en milieu aqueux.

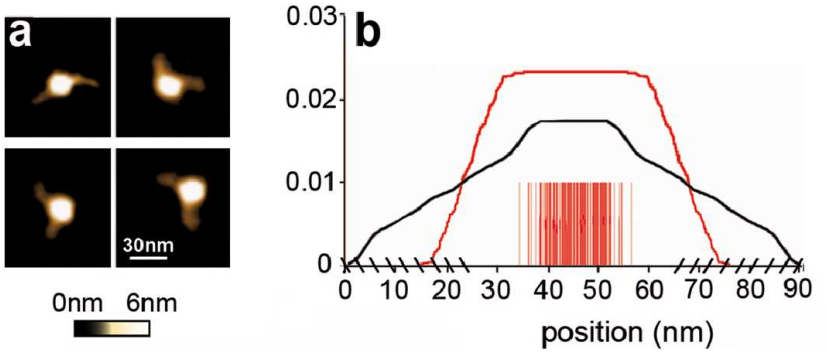


FIGURE 7.2 : *(a) Images AFM en milieu liquide de fragments 601 (255 pb). (b) Distribution de 105 évènements de positionnement (barres verticales). En rouge, l'occupation expérimentale déduite des barres verticales. En noir, la prédiction d'occupation $P = \int \rho(s)ds$ du modèle énergétique (Milani* et al.*, 2009).*

À travers ces expériences, l'AFM apparaît comme un outil intéressant pour observer le positionnement de nucléosomes sur des séquences ADN, avec une limitation dans la résolution de l'ordre de la trentaine de paires de bases. Les résultats sont cohérents avec les expériences biochimiques (notamment celles menées sur le 601).

7.1.1 Choix des séquences pour illustrer le profil énergétique

Trois fragments d'ADN génomique de la levure sont choisis pour leur profil énergétique simple et relativement symétrique et didactique.

We investigate three genomic fragment with peculiar energetic profiles.

Nous avons choisi des fragments génomiques (extraits de la levure) pour plusieurs raisons : l'essentiel des études AFM classiques sont menées sur des séquences artificielles et puisque nous nous intéressons à l'influence de la séquence sur le positionnement, il est naturel de s'intéresser à des séquences réelles. C'est ensuite la forme du profil énergétique que ces séquences génèrent qui guide notre choix. Nous voulons illustrer le positionnement par exclusion, c'est-à-dire le positionnement induit à proximité d'une barrière énergétique. Nous choisissons donc des séquences qui, selon notre modèle, présentent des zones très fortement défavorables et qui jouent le rôle de ces barrières énergétiques.

Dans une première expérience, on reconstruit des mononucléosomes sur des petits fragments d'ADN (figures 7.3 et 7.5) de 394, 386 et 387 pb de long, respectivement nommés A, B et C. Ces trois fragments ont été sélectionnés au sein du chromosome III de la levure *S. cerevisiae*. Le contexte génomique de ces différentes séquences est présenté sur la figure 7.3. Deux de ces profils présentent des zones fortement défavorables par rapport au reste du fragment. Ces *barrières énergétiques* liées à la séquence correspondent d'ailleurs à des zones vides de nucléosomes (NFR) sur les mesures biochimiques *in vivo* (figure 7.3 (a) et (b)).

- Pour le premier fragment, labellisé A, une barrière énergétique observée à la fois sur le profil énergétique que nous calculons et sur le score nucléosomal du modèle de Field, est située au centre du fragment (figure 7.3 (a))
- Pour le fragment B il y a une barrière énergétique sur chaque bord de la séquence (figure 7.3 (b)),
- Enfin le fragment C ne présente pas de zone particulièrement défavorable et sert de référence (figure 7.3 (c)).

Remarquons que les trois fragments présentent des profils énergétiques relativement symétriques, ce qui est nécessaire puisque les expériences ne peuvent distinguer le sens des brins d'ADN. Si, pour orienter le brin, ils avaient été labellisés par une streptavidine par exemple, nous aurions altéré l'affinité de la séquence avec le nucléosome.

On parle ici de barrière parce que le potentiel énergétique s'élève localement sur une hauteur de l'ordre de la dizaine de kT, sur une distance de l'ordre de la taille du nucléosome, ce qui correspond à une force de l'ordre de $0.007\ kT.pb^{-1}$. À titre de comparaison, l'écart type d'un profil énergétique moyen est de l'ordre de deux ou trois kT. Avec les paramètres que nous utilisons pour la prédiction sur la levure, $\delta = 2\,kT$ est l'écart type du profil énergétique sur l'ensemble du génome. Rapporté à la taille

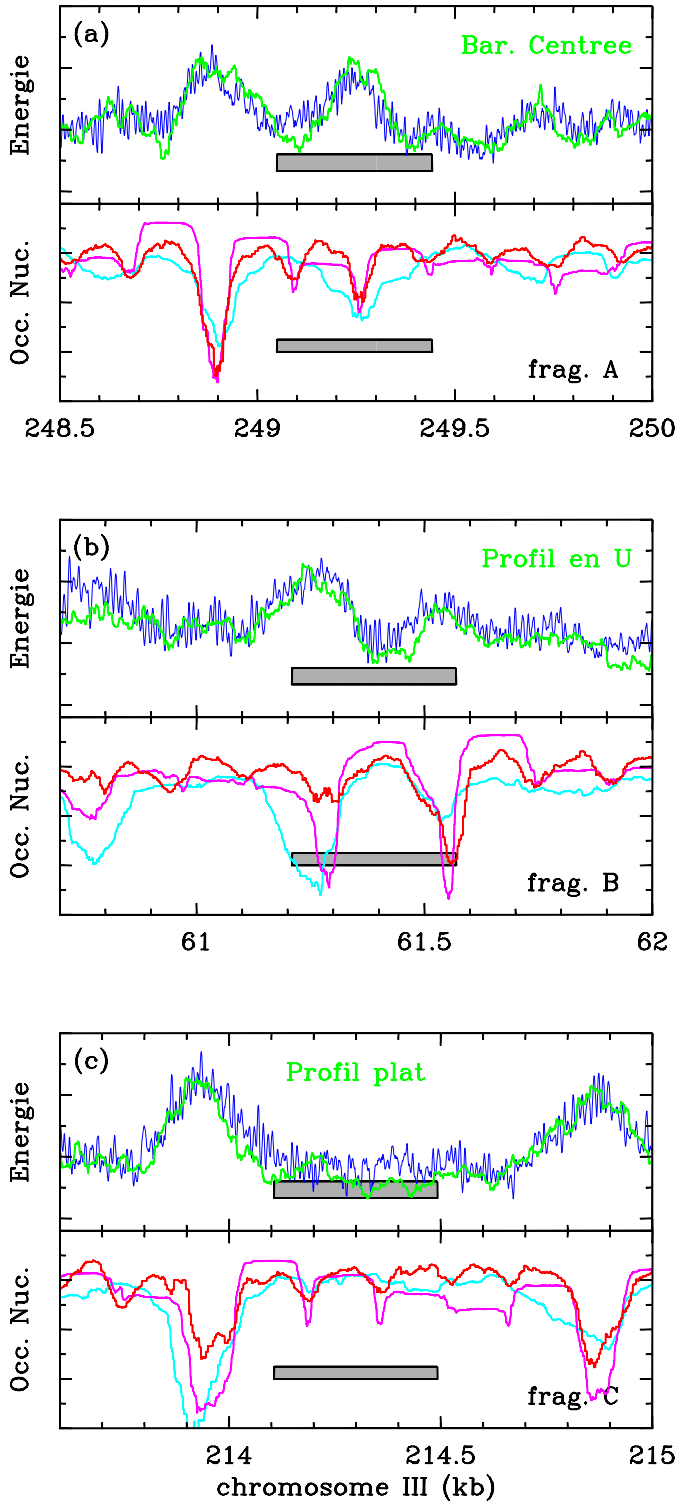


FIGURE 7.3 : *Les séquences choisies et leur contexte génomique. (a) position du fragment A (barrière centrée) le long du chromosome III. En bleu clair, les données de reconstitution in vitro de Kaplan (Kaplan* et al.*, 2009a), en rose, les données in vivo de Kaplan, en rouge, données de positionnement de nucléosomes de Lee (Lee* et al.*, 2007a), en vert le profil énergétique que l'on calcule ($\delta = 2\,kT$) et en bleu le score nucléosomal de Field (Field* et al.*, 2009). (b) idem pour le fragment au profil énergétique en forme de U (fragment B). (c) idem pour le profil énergétique relativement plat (fragment C).*

du nucléosome, cela signifie que les forces typiques liées à la séquences sont de l'ordre de $0.0015\ kT.pb^{-1}$. L'écart énergétique moyen entre deux zones du génomes espacées de plus d'une taille de nucléosome est de l'ordre de 2 kT (cf Fig. 4.9 (Chevereau, 2010)), ce qui signifie bien que les forces mises en jeu par la séquence sont de l'ordre de $0.0015\ kT.pb^{-1}$. On constate qui plus est que le profil énergétique que nous calculons est très similaire au score nucléosomal de Field (Field et al., 2009) (figure 7.3 (a) (b) et (c)), à l'exception évidemment des variations rapides qui ont été éliminées dans notre modèle. On ne tient pas compte de la valeur absolue du profil énergétique, seules les variations locales sur l'échelle du fragment sont prises en compte pour le choix de ces séquences, parce que cette valeur absolue n'influence que le rapport ADN nu sur ADN complexé. Elle n'influence pas la distribution relative des nucléosomes sachant que seuls les ADN complexés sont pris en compte.

Le positionnement *in vivo* des nucléosomes est en partie dû aux particularités énergétiques de ces séquences. On observe *in vivo* (données de Lee et Kaplan (Lee et al., 2007a; Kaplan et al., 2009a)) et *in vitro* (sur les données de Kaplan (Kaplan et al., 2009a))

- une dépletion en nucléosome sur les bords de A (figure 7.3 (a)), qui va de paire avec un enrichissement au milieu.
- B est appauvri en nucléosome en son centre (figure 7.3 (b)) et
- C présente une forme de positionnement faible (figure 7.3 (c)), que l'on soupçonne être issu du positionnement statistique créé par les fortes barrières énergétiques situées respectivement en 214 kb et 214.8 kb.

Évidemment ces données sont issues d'un contexte nucléosomal imposé par l'ensemble de la séquence et par d'autres facteurs dans le cas *in vivo*. La question fondamentale ici est de déterminer dans quelle mesure la séquence seule détermine le positionnement observé *in vivo* des nucléosomes.

7.2 RÉSULTATS EXPÉRIMENTAUX

Le positionnement observé par AFM sur trois petits fragments révèle l'influence effective de la séquence. Les données sur le gène YGR105W confirment que les zones anti-positionnantes peuvent jouer le rôle de barrières énergétiques et créer un effet de positionnement statistique par confinement.

Along with the three genomic fragment, the positioning on the gene YGR105W reveal that the sequence may position nucleosome through excluding mechanisms.

7.2.1 Petits fragments

Les images AFM de nucléosomes positionnés sur les trois fragments sont compatibles avec la description énergétique que l'on donne de ces séquences.

AFM imaging of nucleosomes on those sequences is fully compatible with our model.

Les nucléosomes sont ensuite reconstitués sur ces fragments en solution par dialyses successives jusqu'à ce que l'on observe des nucléosomes isolés sur les fragments par AFM en milieu liquide. Quelques exemples d'images AFM caractéristiques de chacun des fragments sont présentés sur la figure 7.4. A est visible en figure 7.4 (a), B en figure 7.4 (b) et C en figure 7.4 (c).

Au regard des figures 7.4 (a) et (d), les nucléosomes se forment principalement sur les bords du fragment A, ce qui est *a priori* en accord avec la prédiction du modèle énergétique. Une analyse statistique effectuée sur $N = 107$ molécules permet d'évaluer la probabilité de positionnement de la dyade le long du fragment A et confirme effectivement un défaut de positionnement au centre de la séquence, là où une barrière énergétique est prédite.

De façon similaire, 102 molécules ont été analysées sur le fragment (B) (figure 7.4 (b) et (e)) et ici quasiment aucun nucléosome ne s'est formé sur les bords du fait du coût énergétique trop élevé induit par la séquence sur les bords, comme notre modèle énergétique le prévoit.

Pour finir, le fragment C censé ne présenter aucune caractéristique énergétique particulière est observé dans $N = 105$ images (figure 7.4 (c) et (f)) et ne présente effectivement que très peu de variation dans son profil expérimental d'occupation.

Nous avons également présenté sur la figure 7.2 (b) la prédiction obtenue sur la séquence 601, où l'on voit que notre profil énergétique ne montre pas de particularité. L'effet positionnant qui existe au sein de la séquence 601 n'est pas bien capturé par notre modèle énergétique. La séquence 601 positionne le nucléosome essentiellement par l'effet très fort des périodicités, effet absent des séquences génomiques que nous étudions.

Enfin ces séquences ne présentaient pas de périodicité à 10 paires de bases particulière (les variations du score nucléosomal du modèle de Field de la figure 7.3 sont du même ordre de grandeur que partout ailleurs sur la séquence) et c'est donc plus l'effet anti-positionnant des barrières qui détermine la position des nucléosomes que les effets des périodicités.

7.2.2 Prédictions sur les fragments

Les prédictions sont relativement robustes avec le choix des paramètres, et les incertitudes expérimentales sont plus grandes que les différences faibles qui émergent en changeant les quelques paramètres du modèle.

Predictions are robust with changes in parameters. The experimental inaccuracy is greater than the small diferences that arise from changes in μ or δ.

Si l'on ne s'intéresse qu'aux fragments isolés de leur contexte génomique, alors on peut utiliser notre modèle pour prévoir le positionnement des nucléosomes en isolant le profil énergétique sur la zone d'intérêt, et en imposant des conditions aux limites (figure 7.5 (a) (b) et (c)). Pour établir le profil énergétique, nous utilisons le modèle Vaillant. L'amplitude de variation de l'énergie ($\delta = 2$) est choisi conformément au choix du chapitre 6, le potentiel chimique μ est choisi de telle sorte qu'un seul nucléosome est fixé sur la séquence en moyenne. Notons que les prédictions que nous obtenons sont très similaires à celles tirées du modèle récent de Field (Kaplan et al., 2009a), comme on peut le voir sur la figure 7.5. Les deux modèles sont obtenus indépendamment puisque le premier est issu d'une description physique du problème tandis que le modèle de Field est purement issu d'apprentissage sur les données de la levure, mais ils donnent des résultats quasi-identiques. On observe bien des différences entre les deux profils de prédiction, particulièrement sur le fragment C (figure 7.5 (c)), mais il n'est pas possible de

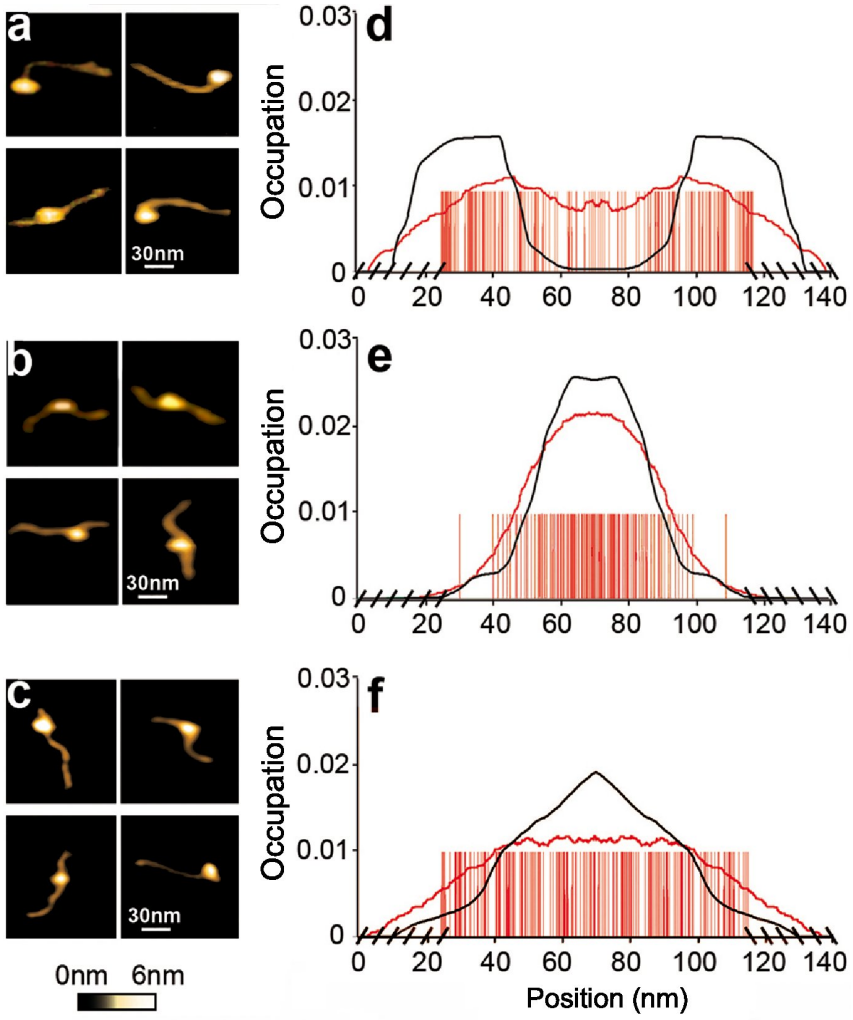


FIGURE 7.4 : *Images AFM en liquide des mononucléosomes reconstruits sur les différents fragments. (a et d) : fragment A extrait de la levure, $L = 394$ pb, $N = 107$ évènements de positionnement observés. (b et e) : fragment B, $L = 386$ pb, $N = 105$ évènements. (c et f) : fragment C, $L = 387$ pb, $N = 105$ évènements. (a) à (c) montrent quatre exemples d'images sur chacun des fragments. (d) à (f) présentent les densités symétrisées déterminées expérimentalement à partir de ces images. Une barre verticale correspond à un évènement de positionnement de dyade. La moyenne du positionnement est présentée en rouge, elle est évaluée par la convolution des données de positionnement par une fenêtre de taille 43.2 nm soit 120 pb. Les résultats théoriques sont présentés en noir ($\delta = 2\ kT$ et μ tel que la densité totale sur le fragment est 1).*

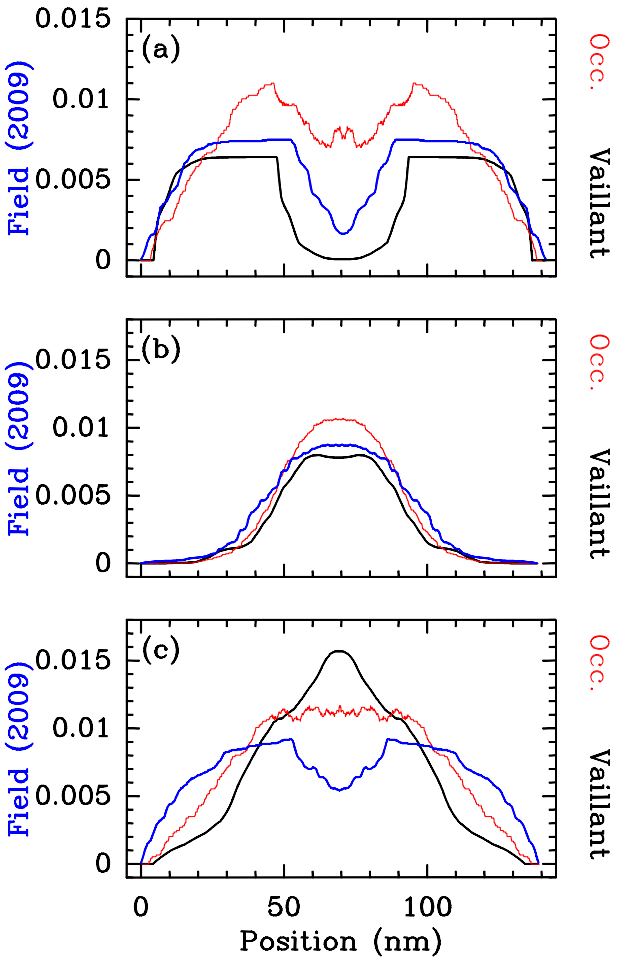


FIGURE 7.5 : *Les prédictions symétrisées du modèle (Vaillant, en noir) et du modèle récent de Field (Kaplan* et al.*, 2009*a*) (*`http://www.genie.weizmann.ac.il`*, en bleu) sur les trois fragments (A,a) (B,b) et (C,c). Profils d'occupation expérimental en rouge.*

déterminer quel modèle est le plus fidèle à l'expérience étant donné l'incertitude expérimentale sur le positionnement.

7.2.3 Positionnement intrinsèque ou positionnement statistique

Les nucléosomes fortement positionnés que l'on observe à l'intérieur du gène YGR105W de la levure ne sont pas positionnés par la séquence mais par l'effet confinant des barrières énergétiques induites par la séquence aux extrémités de ce gène.

Strongly bound nucleosomes observed inside the gene YGR105W are not positioned by the sequence, but by the excluding barriers at the edge of the sequence.

Le positionnement de mono- et de dinucléosomes sur le gène YGR105W

Les mononucléosomes se fixent de façon relativement homogène sur le brin d'ADN, à l'exception des bords. Les dinucléosomes sont répartis de part et d'autre du milieu du fragment, les bords restent désertés.

Mononucleosome will bind the YGR105W sequence evenly, outside of the edges. Dinucleosomes will not bind the edges either, and are thus constrained in the middle.

Dans une deuxième série d'expériences AFM (figure 7.6), on a cette fois reconstruit des mono- et des dinucléosomes sur un fragment sensiblement plus long ($L = 595bp$) issu du chromosome VII de la levure. Ce fragment contient un gène entier (YGR105W) qui code pour une protéine membranaire. *In vivo*, les données de positionnement de Lee (Lee et al., 2007a) (points noirs de la figure 7.6 (g)) suggèrent une organisation très régulière. Deux nucléosomes y sont systématiquement juxtaposés et sont bordés par deux zones vides (NFR : lorsqu'une zone du génome est vide de nucléosome) situées respectivement sur

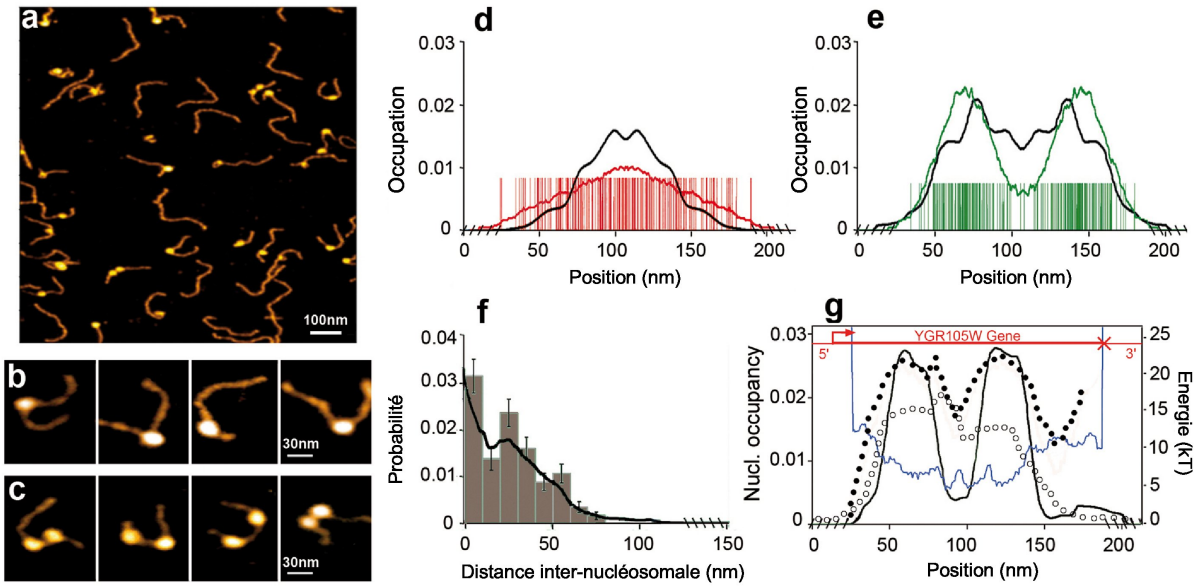


FIGURE 7.6 : *Imagerie AFM de mononucléosomes (a et b) et dinucléosomes (a et c), effectuée sur le gène YGR105W (595 pb) du chromosome VII de la levure. (d) Densité symétrisée de mononucléosomes (113 évènements). Les barres verticales correspondent aux positions des dyades relevées expérimentalement, ainsi que leur symétrique. La courbe d'occupation expérimentale en rouge est comparée à la courbe prédite en noir. (e) Densité symétrisée de dinucléosomes (62 évènements). Pour chaque dinucléosome observé, quatre barres verticales vertes sont tracées, aux positions des dyades, et sur leur symétrique. L'occupation expérimentale en verte est comparée à la prédiction en noir. (f) Distribution statistique de la taille des linkers mesurée sur les dinucléosomes. La courbe théorique (voir paragraphe 5.2.2) est tracée en noir. (g) Occupation expérimentale in vivo (données chip de Lee (Lee* et al.*, 2007a), points noirs), Occupation en reconstitution in vitro (Kaplan* et al.*, 2009a) (cercles noirs). La position du gène YGR105W est indiquée en rouge. La flèche indique le TSS (Transcription Start Site), et la croix le TTS (Transcription Terminaison Site). Profil énergétique associé à la séquence (en bleu). Occupation prédite par notre modèle en noir.*

le "Transcription Start Site" ou TSS (L'accès au promoteur est facilité) et sur le "Transcription Terminaison Site" ou TTS (l'utilité de cette NFR est encore discutée, il est possible qu'elle soit impliquée dans la terminaison de la transcription, le recyclage de la polymérase ou bien dans l'initiation antisens (Mavrich et al., 2008a).).

Les images AFM effectuées sur ce gène (figure 7.4 (d)) montrent qu'effectivement, peu de mononucléosomes se mettent sur les bords, ce qui suggère la présence de barrières énergétiques sur les côtés du fragment, tout comme le fragment B de la première expérience (figure 7.4 (e)). Lorsqu'un nucléosome seul se forme sur ce fragment, il ne peut pas se former sur les bords, mais il peut se mettre à peu près où il veut sur le milieu du fragment comme en atteste l'occupation relativement homogène au milieu du fragment sur les $N = 117$ molécules observées (figure 7.4 (d)). Cette homogénéité de positionnement au centre du gène est confirmée par des données *in vitro* obtenues sur l'ensemble du génome de la levure par une reconstitution similaire (Kaplan et al., 2009a) (figure 7.4 (g), cercles noirs). Si cette fois on oblige deux nucléosomes à se former sur ce brin, les choses changent drastiquement en terme de positionnement. Les 62 dinucléosomes observés par AFM se placent systématiquement de part et d'autre du centre du gène. Le résultat important n'est pas ici de trouver que les deux nucléosomes ne se mettent jamais au milieu (ce serait impossible, car si un nucléosome se place au milieu, alors, il n'y a pas de place pour former un deuxième nucléosome sur la séquence (figure 7.4 (e))). Non, le résultat important est l'absence de positionnement particulier (figure 7.4 (d)) des mononucléosomes sur les deux parties occupées par les dinucléosomes. Ceci implique donc que ce n'est pas un caractère positionnant de la séquence qui induit la formation régulière observée *in vivo*, mais c'est plutôt le caractère anti-positionnant de la séquence qui par effet de confinement induit un ordonnancement régulier des nucléosomes au sein du gène. On est ici dans la situation décrite au chapitre 4, correspondant à deux barrières bordant un profil plat, faiblement bruité par la séquence. Les oscillations de positionnement sont dues au confinement et non un code déterminant la position de chacun des nucléosomes.

Distance inter-nucléosome

La distance inter-nucléosome dépend de la séquence et est compatible avec un modèle de sphère dures.
The internucleosomal length of dinucleosome definitely argues in favor of a strong sequence influence in YGR105W.

Puisqu'on observe des dinucléosomes, il est possible ici d'évaluer systématiquement la distance qui sépare deux nucléosomes (donc la "DTN" et le "NRL", cf Chapitre 5.2.3) afin notamment de sonder l'interaction (i.e. déterminer la portée effective de répulsion entre les particules). Théoriquement, on peut aussi évaluer la distribution de cette distance à partir du profil énergétique : il suffit de sommer les poids de Boltzmann associés à chacune des configurations pour une distance donnée

$$P(x = l + x_l) = \frac{1}{\Gamma} \int_0^{L-l} e^{\beta(2\mu - (E(s)+E(s+l)))} ds \tag{7.1}$$

où L est la taille du fragment, l est la taille d'un nucléosome, x_l est la taille du *linker*. L'intérêt est d'ici de sonder l'interaction qui affecte les nucléosomes entre eux (paragraphe 5.2.2). Lorsque l'on compare les résultats expérimentaux obtenus pour la distance inter-nucléosomale par rapport à la distribution théorique pour des sphères dures (figure 7.6 (f) et figure 7.7), on observe un bon accord. La distribution de distance inter-nucléosomale dans un modèle où la séquence ne jouerait pas (profil homogène) est présenté sur la figure 7.7 en rouge : clairement, la distribution expérimentale ne correspond absolument pas à cette description. La distance maximale autorisée est matérialisée par les traits obliques sur la figure 7.6 (f), et n'est jamais atteinte expérimentalement. Si le positionnement était strictement homogène sur le fragment, il devrait y avoir quelques évènements où un nucléosome est à chaque extrémité. Puisque la séquence n'est pas favorable aux nucléosomes sur les bords, ces évènements sont très sous-représentés dans la courbe théorique, et aucun évènement de ce type n'est observé expérimentalement. La décroissance suit globalement les mêmes lois, et on prédit même le sursaut de distance inter-nucléosomale situé aux alentours de 30 nm. Il faut toutefois rappeler que les incertitudes expérimentales sont relativement grandes puisque seuls 62 dinucléosomes sont observés et l'incertitude sur les distances entre nucléosomes est de l'ordre de 10 nm. Cette expérience confirme simplement que le modèle tige rigide est très raisonnable pour donner une description de ce système et que le profil énergétique que nous calculons est tout à fait valide pour décrire ce fragment.

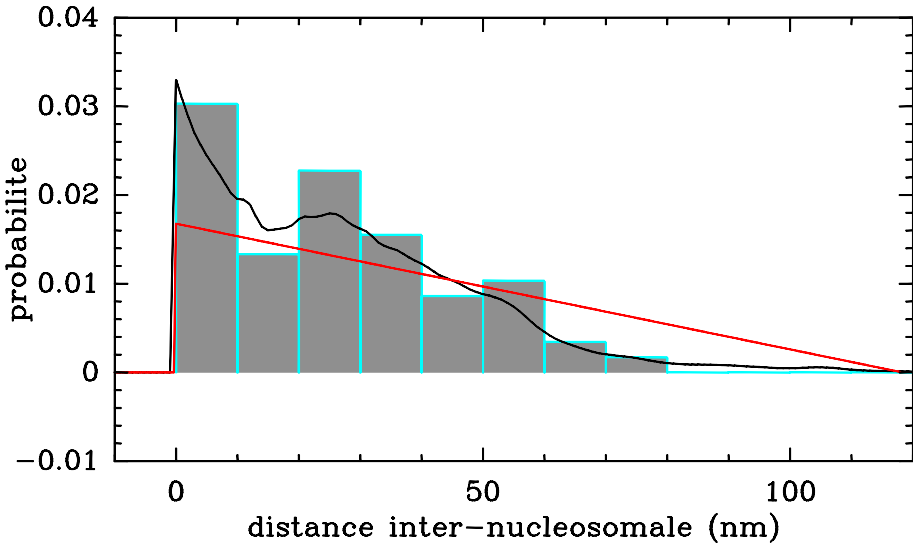


FIGURE 7.7 : *Distribution de la taille de linker sur le fragment qui contient le gène YGR105W. Expérimentale (histogramme), théorique en prenant un profil homogène (en rouge), et théorique en prenant en compte l'influence de la séquence (en noir).*

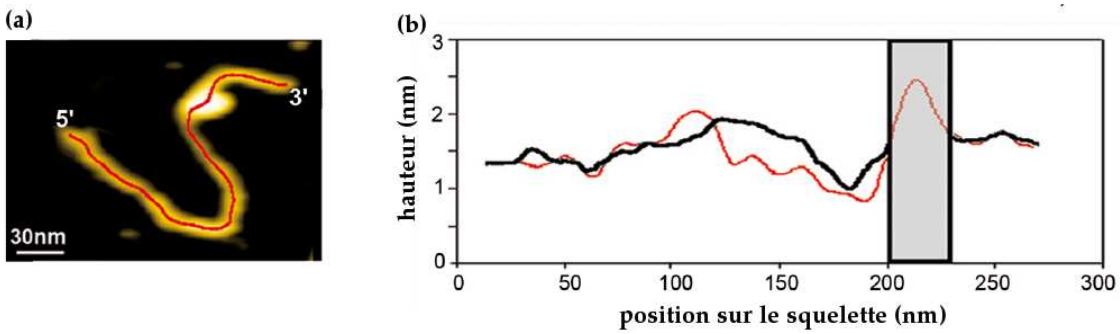


FIGURE 7.8 : *Illustration de la méthode d'orientation des brins d'ADN grâce à la topologie des images AFM. (a) Image AFM du fragment IL2RA (898 pb de long) (b) Le profil de hauteur le long du fragment (en rouge) est comparé au calcul de variation de pitch (en noir) obtenu avec la table de Bolsoy et al. (Bolshoy* et al.*, 1991).*

7.2.4 Régulation de l'induction de la transcription du gène IL2RA

Les barrières d'énergie peuvent influencer le positionnement près des sites de régulation de façon à réprimer la transcription par la présence d'un nucléosome inhibiteur.

We investigate the human gene IL2RA, where the positioning of nucleosomes near regulation sites will influence transcrption.

Nous avons également étudié un fragment suffisamment long pour qu'il puisse être orienté grâce à la topologie de l'image AFM (figure 7.8). Il s'agit d'un extrait de 898 pb de long du génome humain, qui contient le promoteur du gène IL2RA qui joue un rôle majeur dans le contrôle de la réponse immunitaire. Cet extrait contient en outre divers éléments fonctionnels : (i) le TSS du gène, (ii) le promoteur et la TATA-box, (iii) deux sites de régulation (Positive Regulatory Region, PRR) PRRI et PRRII situés aux positions -289/-216 et -137/-64 en amont du TSS (figure 7.9 (b)). Ces sites de régulations sont importants pour la régulation de l'induction de la transcription du gène IL2RA (Kim et al., 2006). De précédentes études *in vivo* menées dans des cellules T non stimulées ont révélé qu'un nucléosome inhibiteur était positionné sur la TATA-box et le TSS avant la transcription (Reeves et al., 2000). Le complexe de préinitiation — PIC, Pre Initiation Complex en anglais— de la transcription ne pouvant être recruté, la transcription est inhibée. De surcroît, ce nucléosome pourrait également prévenir la dénaturation non contrôlée de la région du TSS (Milani, 2007). Ces propriétés font de ce promoteur un modèle exemplaire pour l'analyse du rôle réel joué par la séquence dans un environnement chromatinien singulier.

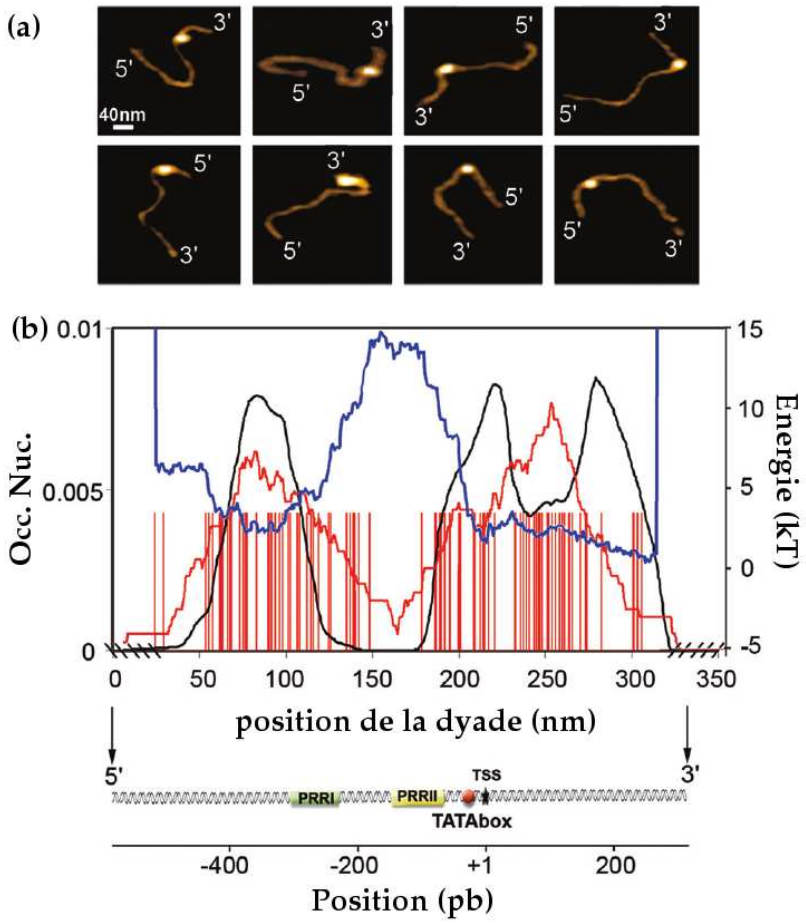


FIGURE 7.9 : *(a) images AFM orientées de brin de longueur* 898 *pb contenant le promoteur du gène IL2RA. (b) analyse statistique du positionnement de la dyade de* $N = 100$ *molécules (barres rouges) relativement à la position des deux régions régulatrices PRRI et PRRII, la TATA-box, et le TSS. L'occupation expérimentale que l'on déduit est présentée sur la courbe rouge. Le profil prédit (pour un potentiel chimique qui conduit à une occupation moyenne de 1 nucléosome) est présenté en noir. Le profil énergétique associé à la séquence du brin est présenté en bleu.*

Les $N = 100$ mononucléosomes positionnés sur les séquences orientées que nous avons observés (figure 7.9 (b)) confirment le positionnement naturel d'un mononucléosome sur la TATA-box et le TSS, comme observé *in vivo*. Nous observons également une région favorable environ 350 pb en amont du TSS qui couvre la zone régulatrice PRRI (figure 7.9 (b)). L'occupation nucléosomale observée *in vitro* est en accord remarquable avec le profil théorique que nous calculons. Selon notre modèle, il existe une séquence fortement défavorable 150 paires de bases en amont du TSS qui couvre la zone de PRRII, ce qui explique l'absence de positionnement à cet endroit (-200 jusqu'à -100 pb). Cette zone d'exclusion induit naturellement un positionnement fort des nucléosomes sur ses bords : à droite il s'agit du nucléosome réprimant le TSS et la TATA-box, et à gauche il s'agit du nucléosome recouvrant en partie la zone de PRRI. Ces résultats démontrent le rôle fondamental joué par la séquence en produisant un profil nucléosomal singulier qui contribue à la répression de ce gène. De plus, en inhibant la formation de nucléosome sur la zone de PRRII, la séquence favorise l'accès pour les protéines HMGA qui sont connues pour induire un stress tortionnel négatif. L'accumulation de ce stress conduit à un superenroulement qui favorise sans doute le remodelage du nucléosome ou même son éjection (Milani, 2007; Nissen and Reeves, 1995). Au final, ces observations expérimentales mettent en lumière l'importance de l'organisation nucléosomale induite par la séquence pour la régulation de l'induction de la transcription du gène IL2RA. Cette étude a d'ailleurs été approfondi par uyne étude structurale plus complète de ce locus, résumée à la figure 7.10 (Milani et al., 2011).

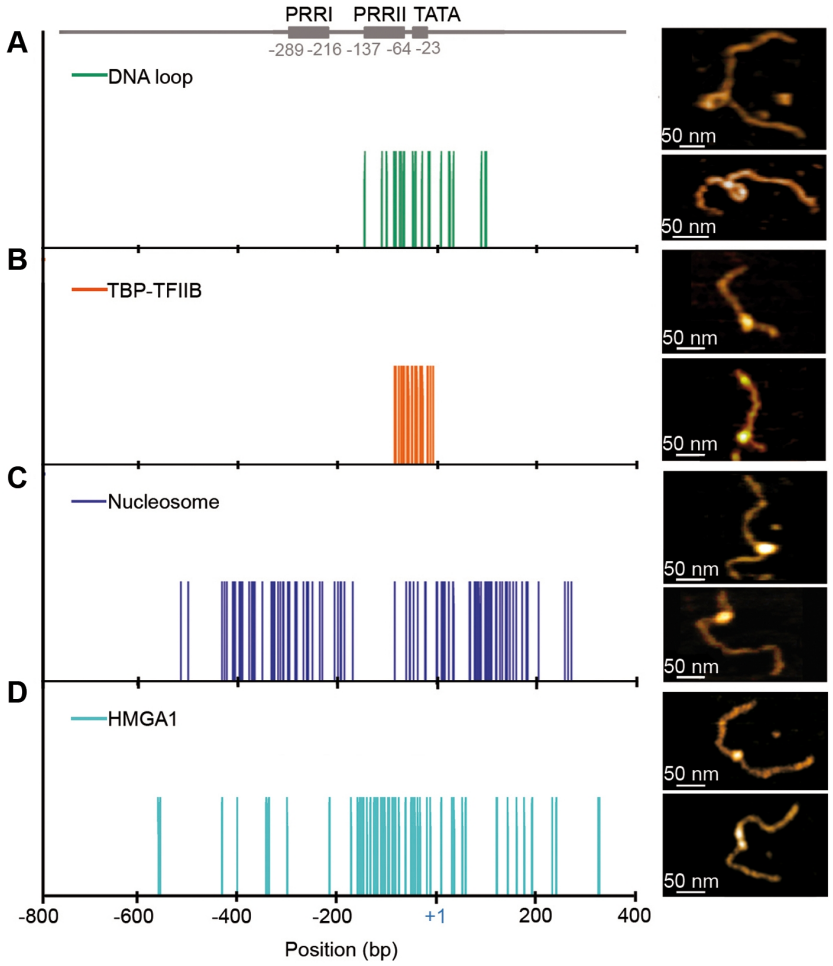


FIGURE 7.10 : *AFM imaging of DNA loop, GTF, nucleosome and HMGA1 positioning on the IL2RA gene promoter. (A) AFM imaging in liquid of DNA loop positioning along the 1290 bp IL2RA fragment (upper right panel) and the 898 bp IL2RA fragment (lower right panel) and statistical analysis (left panel) of their positioning given by the loop middle point (green vertical bars) from N = 21 AFM images. (B) AFM imaging of TBP-TFIIB positioning along the 563 bp IL2RA fragment (right panels) and statistical analysis (left panel) of their positioning (orange vertical bars) from N = 51 AFM images. (C) AFM imaging of mononucleosome positioning along the 898 bp IL2RA fragment (right panels) and statistical analysis (left panel) of dyad positioning (blue vertical bars) from N = 100 AFM images. (D) AFM imaging of HMGA1 positioning along the 898 bp IL2RA fragment (right panels) and statistical analysis (left panel) of dyad positioning (blue vertical bars) from N = 73 molecules.*

8 LES "TROUS" DE NUCLÉOSOMES

Nous présentons maintenant le positionnement nucléosomal et ses liens avec les fonctions du génome. Nous analysons en particulier la présence d'évènements marquants comme les NFR ("Nucleosome Free Region").

We will now focus on the positioning and its links with genomic regulation sites. We will particularily emphasize the role played by Nucleosome Free Regions.

À petite échelle, le positionnement n'est pas sans conséquence sur les différents aspects de la régulation génétique. La régulation de la transcription par exemple est généralement une conséquence de la modulation de l'accessibilité des sites de régulation par leur facteur de transcription associé. (Kornberg and Lorch, 1999; Li et al., 2007; Morse, 2007a; Rando and Ahmad, 2007; Segal and Widom, 2009) (Cf. Cahpitre 2).

8.1 LA POSITION DES BARRIÈRES ÉNERGÉTIQUES ET DES NFRS À PROXIMITÉS DES ZONES FONCTIONNELLES

On trouve des NFRs quasi-systématiquement au promoteur des gènes de la levure, et très fréquemment au niveau de la fin des gènes. Les sites de fixation de facteurs de transcription correspondent également à des NFRs expérimentales mais pas liées à la séquence.

NFR are almost systematically found near the promoteur region of genes. There are also NFR at the end of yeast genes. Transcription factor binding site also correspond to NFR definitely not induced by the sequence.

Comme on l'a vu au chapitre précédent, la séquence seule permet de prédire quasi- complètement le positionnement *in vitro*.

Les données *in vivo* révèlent des zones qui devraient être occupées par des nucléosomes, et qui ne le sont pas en réalité. Il faut donc invoquer la présence de facteurs extérieurs à la séquence pour expliquer ces zones déplétées. Nous nous concentrons maintenant sur ces NFRs, en analysant le rôle d'exclusion qu'elles jouent, le rôle régulateur qu'elles peuvent produire et leur positionnement le long du génome. La modélisation générale que nous utilisons confirme le rôle essentiel joué par les zones d'exclusion, en induisant un positionnement statistique à proximité. Il est donc naturel de s'interroger sur les aspects fonctionnels de l'organisation nucléosomale induite par ces NFRs. Nous nous penchons donc sur la position de ces zones vis-à-vis de régions fonctionnelles telles que le TSS, le TTS, et les sites de fixation de facteurs de transcription.

8.1.1 Définition des NFRs

Pour définir une NFR dans les données in vivo de Lee, nous faisons un simple seuillage. Pour définir une NFR théorique, c'est-à-dire liée à la séquence, nous seuillons et nous vérifions qu'il existe une barrière énergétique locale.

In order to define a NFR in vivo in Lee's data, we use a simple threshold. In order to define a theoretical NFR, that is a sequence induced NFR, we threshold and we check for an energetic barrier in the potential.

– Définition par seuillage : À partir des données de positionnement expérimentales, on peut par exemple effectuer un simple seuillage en dessous duquel une zone est dite déplétée. Toutefois cette méthode peut définir comme NFR un *linker* situé entre deux nucléosomes particulièrement bien positionnés. Certes la zone est effectivement peu occupée, mais ce n'est pas du fait d'une déplétion à proprement parlé, il s'agit plutôt d'une conséquence du positionnement statistique. Lors de la recherche de NFR nous souhaitons exclure les *linkers* ; pour cela il faut définir la largeur d'une zone vide de nucléosome et exclure les zones trop petites. La solution que nous adoptons consiste d'abord à définir les bords des potentielles NFRs, que l'on appellera une IP (pour Inflexion Point, le point d'inflexion de la remontée de probabilité d'occupation qui borde une NFR). Si la zone déplétée est bordée par des pentes situées à moins de quelques dizaines de paires de bases (50), c'est que l'on vient de détecter un *linker*. La détection des pentes est effectuée en corrélant le signal d'oc-

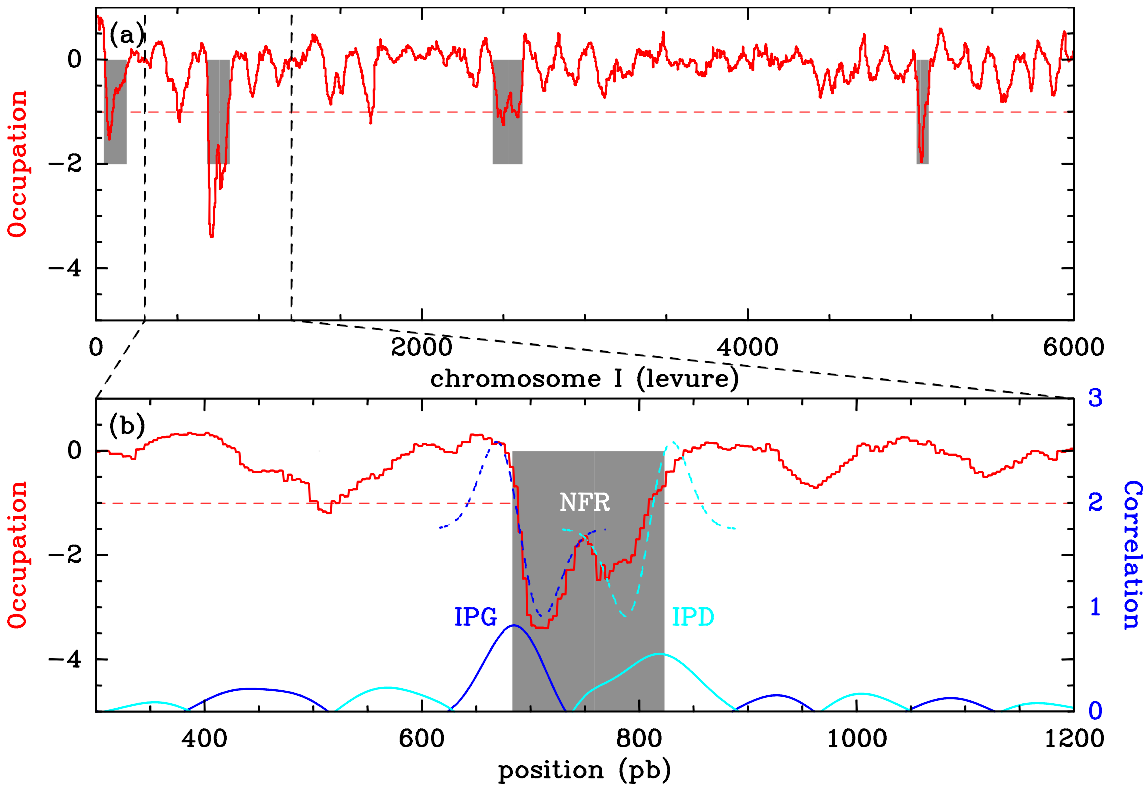


FIGURE 8.1 : *Détection des NFRs dans les données d'occupation en nucléosome de Lee (Lee* et al.*, 2007*a*). (a) Illustration de la détection sur un extrait du chromosome I de la levure. En gris : les zones définies comme NFR. En pointillé rouge : le seuil en dessous duquel l'occupation est considérée comme anormalement basse et qui permet de définir les NFRs (b) Détail sur une NFR. L'occupation est corrélée avec les deux ondelettes (bleue et cyan) et le résultat de la corrélation (courbe continue en bas) permet de déterminer les bords de la NFR.*

cupation avec la dérivée d'une gaussienne de largeur 30 pb (figure 8.1) (la dérivée de la gaussienne permet de détecter la pente montante, permet de détecter les IP Droite. La dérivée de l'opposée de la gaussienne permet de déterminer une pente descendante, ou IP Gauche). En corrélant le signal d'occupation avec une fenêtre glissante, on compare localement le signal à cette ondelette. En seuillant cette corrélation ($corr > 0.4$), on construit une banque de donnée de positionnement d'IP (gauche et droite) qui détermine l'ensemble des positions où le signal de positionnement présente une forte pente, ce qui expérimentalement, correspond généralement à un bord d'un nucléosome bien positionné. Les NFRs sont ensuite validées par un seuillage sur la valeur de l'occupation entre deux pentes ($Occupation < -1$).

– Lorsque l'on essaie de déterminer une NFR sur les données modélisées, il est possible de détailler de façon plus précise et de distinguer les zones vides de nucléosomes du fait de la séquence de celles qui n'y sont pas liées et qui sont de simples zones vides du fait du positionnement statistique. Puisque le modèle est uniquement défini par la séquence, la seule source possible de déplétion nucléosomale est *a priori* une séquence fortement désavantageuse pour la formation du nucléosome. Il arrive évidemment que par effet de positionnement statistique, un *linker* se trouve fortement peu peuplé, mais tout comme avec les données expérimentales, on ne souhaite pas définir ces zones comme des NFRs. Une première idée serait donc de prendre le profil énergétique de formation des nucléosomes, et seuiller en énergie. Cette fois ci, toute zone dont l'énergie de formation serait trop élevée correspondrait à une NFR. Malheureusement, cette méthode est problématique lorsqu'une large zone (plusieurs kb) est défavorable : certes le niveau d'occupation moyen sera faible, mais il y aura au sein de cette zone des oscillations de positionnement. L'idée est donc de rechercher les NFRs dans le profil de positionnement modélisé, et de conditionner la validation de cette zone en tant que NFR au fait de trouver une surélévation locale du potentiel énergétique. De cette façon, on s'assure que les NFRs que l'on définit sont bien des zones déplétées du fait de la séquence, et

non du positionnement statistique. Pour définir la position des NFRs théoriquement, nous avons d'abord seuillé sur l'occupation prédite à partir d'une séquence ($P < 0.35$) et capturé les minima locaux du profil résultant. Nous avons ensuite vérifié qu'il existait une barrière dans $E(s)$ après avoir éliminé les variations lentes du potentiel. C'est-à-dire que nous avons vérifié que le minimum d'occupation correspondait bien à une surélévation ($> +3\ kT$) locale du profil énergétique. Si une NFR théorique (liée à la séquence) est détectée de cette façon, nous lui attribuons un score correspondant au nombre moyen de nucléosome dans une fenêtre de 125 pb autour du minimum d'occupation.

8.1.2 NFR *in vitro*

Une petite proportion des TSS et des TTS possèdent une NFR encodée dans la séquence.

À partir d'un simple seuillage ($\log_2(Occ.) < -3$), nous avons extrait des données *in vitro* de Kaplan Kaplan et al. (2009a) environ 3500 NFRs qui sont associées de manière prédominante avec les TSS (24% d'entre eux en ont une) et les TTS (55% d'entre eux ont une). Une très grande majorité ($\approx 63\%$) co-localise (à 70 pb près) avec les NFRs prédites par notre modèle à faible densité. De fait, 63% des NFRs situées aux TSS dans les données *in vitro* correspondent à des NFRs prédites. De façon tout à fait similaire, 70% des TTS qui présentent une NFR *in vitro* présentent une NFR prédite. Puisque seulement 24% des TSS ont une NFR prédite par notre modèle, cela signifie que seulement 15% des TSS de la levure sont déplétés constitutivement. De même, seulement 30% des TTS sont déplétés constitutivement. Il est probable que ces déplétions soient issues de l'évolution de la séquence dans ce sens (Washietl et al., 2008).

Sur la figure 8.2 (a), on a représenté la moyenne de l'occupation *in vitro* obtenue autour des 4554 TSS de la levure. La déplétion observée est reproduite de façon remarquable par notre modèle (il en va de même du modèle de Field). On retrouve bien l'effet de la présence de barrières anti-positionnantes liées à la séquence (voir le modèle énergétique en vert) au niveau du TSS. On observe également une déplétion au niveau du TTS (en 5', figure 8.2 (b)). En fait, le profil moyenné au niveau du 3' et celui du 5' sont très similaires. L'organisation *in vitro* est donc très symétrique pour un gène. Les deux profils ne présentent guère de positionnement bien défini pour les nucléosomes à l'intérieur du gène.Toutefois, lorsque l'on moyenne sur la position du nucléosome en $+1$ du TSS et du nucléosome en -1 du TTS, il apparaît que l'énergie présente bel et bien une barrière, mais aussi que le $+1$ et le -1 sont légèrement positionnés. Ceci correspond vraisemblablement aux positions des nucléosomes que l'on sait être positionnés par des séquences périodiques AA/TT/AT (Mavrich et al., 2008a; Shivaswamy et al., 2008; Ioshikhes et al., 2006). Lorsqu'on analyse en effet la periodicité de ces motifs tout comme à la figure 4.5 via le spectre de fourier, mais cette fois-ci soit dans la region promotrice $[-200 - 0]\ pb, 0 = TSS$ ((Fig. 8.3, rouge) et dans la région génique $[0 - 200]\ pb$ (Fig. 8.3, vert), on remarque que le (faible) signal nucléosomal de periodicité $10.2pb$ observé génomiquement (Fig. 4.5(a)) est en fait principalemnt localisé dans la région génique, au niveau donc du nucléosome $+1$.

8.1.3 NFR *in vivo*

In vivo, on trouve très fréquemment une NFR au niveau du TSS, beaucoup plus souvent que ce que la séquence seule peut induire.

In vivo NFR at the TSS cannot be accounted for solely by the sequence.

Lorsque l'on étend cette étude à la situation *in vivo* (seuillage à $\log_2(Occ.) < -2$), on trouve significativement plus de NFR (4800), et cette fois 70% des TSS en possèdent une, et 33% des TTS. Clairement, comparativement à la situation *in vitro*, un grand nombre de ces NFRs *in vivo* ne peuvent être interprétées par l'effet défavorable de la séquence. Il y a là la signature d'agents extérieurs qui déstabilisent le nucléosome au TSS notamment : on peut penser aux facteurs de transcription (TF), aux remodeleurs Hartley and Madhani (2009) et aux chaperonnes. Voilà qui permet d'expliquer pourquoi notre modèle, même à haute densité ($\mu = -1.3$ kT), prédit un plus petit nombre de NFRs (29%) au niveau du TSS, mais continue d'être performant au niveau du TTS qui sont sans doute moins occupés par des TFs. En effet, 46% des TTS ont une NFR théoriquement, et 40% des NFRs expérimentales sont correctement prédites — dont 67% sont déjà présentes *in vitro*, à basse densité —. Ainsi, la plupart des NFRs observées *in vitro* correspondent à des barrières énergétiques imposées par la séquence, et un certain nombre des NFRs observées *in vivo* ne sont pas explicables par la séquence seule.

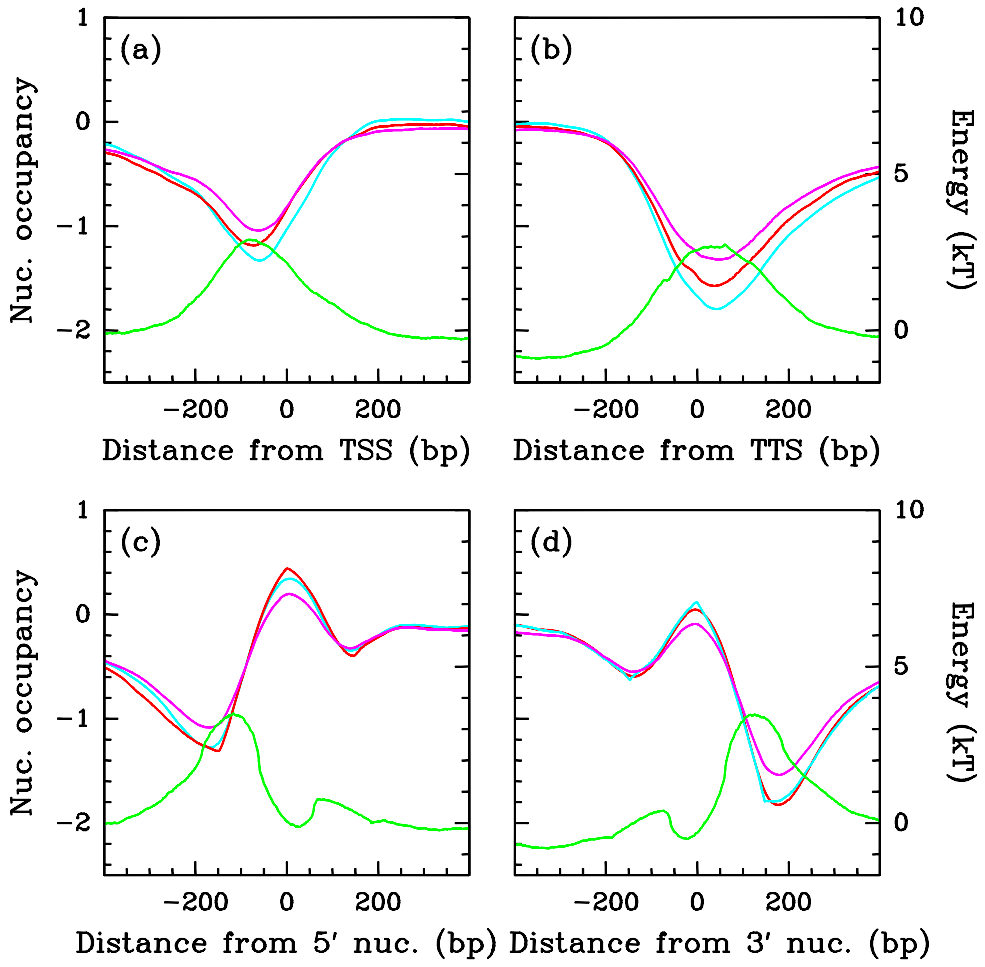


FIGURE 8.2 : *Moyenne (sur* 4554 *gènes) du* $\log_2$ *de l'occupation in vitro (orange) et théorique Vaillant (Vaillant* et al.*, 2010) (cyan), théorique Field (Field* et al.*, 2009) (rose), ainsi que du profil énergétique en vert autour du (a) TSS, (b) TTS, (c) nucléosome en* $+1$ *du TSS, (d) nucléosome en* -1 *du TTS. Paramètres du modèle :* $(\mu, \delta) = (-6, 2)$.

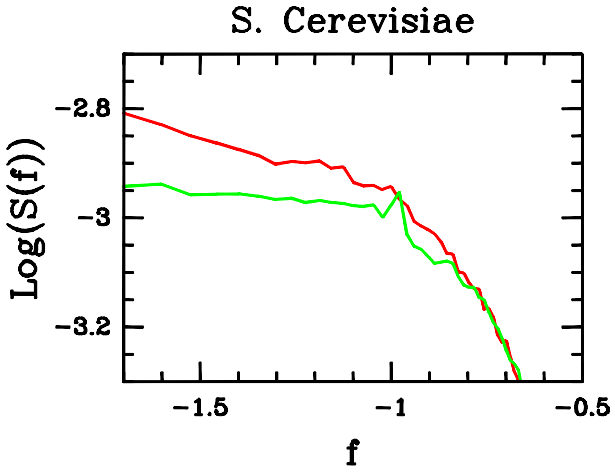


FIGURE 8.3 : *Spectre de puissance du signal de distribution de courbure intrinsèque (*ρ_o*) donnée par la table trinucléotide "PNuc" dans la région promotrice (*$[-200; 0]$ *pb, rouge) et 5′ (*$[0; 200]$ *pb, vert) ; moyenne sur les gènes de la levures.*

Sur la figure 8.4, on représente la moyenne *in vivo* de l'occupation sur les 4554 gènes de la levure, et on observe à nouveau une forte déplétion. Toutefois, cette déplétion est flanquée d'une série d'oscillations à l'intérieur du gène, de périodicité 167 pb, correspondant à la succession de nucléosomes bien positionnés. Si notre modèle reproduit correctement la déplétion, il ne rend pas compte des oscillations. Nous interprétons cette différence par une différence de phasage : alors que la NFR est systématiquement située juste en amont du TSS expérimentalement, la déplétion théorique n'est pas précisément positionnée. Ceci mène par moyennage à l'élimination des oscillations et à un agrandissement de la zone déplétée (figure 8.4 (a)). L'histogramme de la figure 8.4 (a) montre bien que la distance entre le TSS et le nucléosome $+1$ expérimental est piquée autour de 70 pb.

Le nucléosome $+1$ prédit est quand à lui vaguement distribué autour de ces 70 pb. Si l'on moyenne sur le nucléosome $+1$ cette fois (figure 8.4 (e)) ou même sur la NFR directement (figure 8.4 (c)), on retrouve complètement les oscillations. La situation au TTS est relativement similaire (figure 8.4 (b), (c) et (f)).

Afin d'élucider le mécanisme sous-jacent à la présence de ces NFRs non liées à la séquence, nous avons moyenné le profil d'occupation autour des sites de fixation des facteurs de transcription (TFS). Comme le montre la figure 8.5 (b), la plupart des TFS résident *in vivo* dans des régions déplétées en nucléosome expérimentales (Yuan et al., 2005; Albert et al., 2007; Lee et al., 2007a), mais quasiment pas *in vitro* (figure 8.5 (a)). Voilà qui suggère que les facteurs de transcription ont gagné la compétition pour l'accès à l'ADN contre le nucléosome (Segal and Widom, 2009).Effectivement, notre modèle énergétique ne prévoit qu'une très faible déplétion induite par la séquence ((figure 8.5 (a), notez la forme particulière du profil énergétique (vert) et le léger enrichissement dans notre modèle faible densité (bleu clair) au niveu du site), ce qui se traduit également à haute densité par une légère déplétion, et un très léger phasage des nucléosomes (figure 8.5 (b)). Afin de prendre en compte la présence des facteurs de transcription sur l'ADN, nous avons rajouté dans le profil énergétique $E(s)$ des barrières artificielles de forme trapézoïdale pour imiter l'effet des facteurs de transcription. Comme le montrent la figures 8.5(b), cette astuce permet de retrouver en grande partie le profil *in vivo*.

La figure 9.1 résume la situation au TSS et au TTS : si on se contente de regarder les profils moyennés, on n'observe pas d'oscillation dans la modélisation du fait du mauvais phasage avec le TSS ou le TTS (figure 9.1 (a) et (b)). Si on s'affranchit du problème de phasage en centrant les profils moyennés sur le nucléosome, on retrouve des oscillations avec le modèle. L'amplitude de ces oscillations est trop faible au TSS et on ne peut les retrouver que si l'on rajoute une barrière énergétique en plus de la séquence au TSS (figure 9.1 (c) et (d)) ; on peut supposer que cette barrière, introduite donc *ad hoc*, correspond *in vivo* à la présence du complexe de pre-initiation. Si, comme déjà discuté plus haut pour la déplétion au niveau des sites de facteurs de transcription, on ne rajoute une barrière qu'au niveau de ces sites (et non pas systématiquement au niveau des TSS), on améliore aussi, mais dans une moindre mesure, nos prédictions des déplétions observées *in vivo* au TSS (Fig. 8.4). Si on s'intéresse aux données *in vitro*, données expérimentales et théoriques coïncident parfaitement (figure 9.1 (e) et (f)).

8.2 RÉGULATION DE L'ACTIVATION : CONTRIBUTION DE LA SÉQUENCE ET ÉVOLUTION

On a vu donc qu'au niveau du promoteur, la séquence peut rendre compte pour au plus 30% des déplétion observé (à un décalage près). La question qu'on se pose ici est de savoir dans quelle mesure l'architecture des promoteurs, au delà du NFR est-elle prédite par la séquence et ainsi dans quelle mesure on retrouve les classes structurelles et fonctionnelles extraites par Lee (Lee et al., 2007a) (Fig. 2.24) et Tirosh (Tirosh and Barkai, 2008a) (Fig. 2.27) discutées au chapitre 2. Au regard des figures 8.6 et 8.7, on peut dire que : (i) l'architecture du chapelet à la fin des gènes est la même quelle que soit les classes (ce qui n'avait pas été noté au chapitre 2), (ii) que cette architecture est donc tout comme pour la moyenne globale très bien décrite par notre modèle "simple" d'occupation à haute densité (séquence + volume exclus), (iii) que par contre, ce modèle simple ne semble pas vraiment discriminer les architectures au niveau du TSS observées *in vivo* mais (iv) que rajouter au profil énergétique des barrières au niveau des sites de facteurs de transcription permet désormais de mieux rendre compte des différentes classes (Simplement parce que dans la classe rouge, il n'y a que peu de sites, dans la classe verte il y a plus de sites

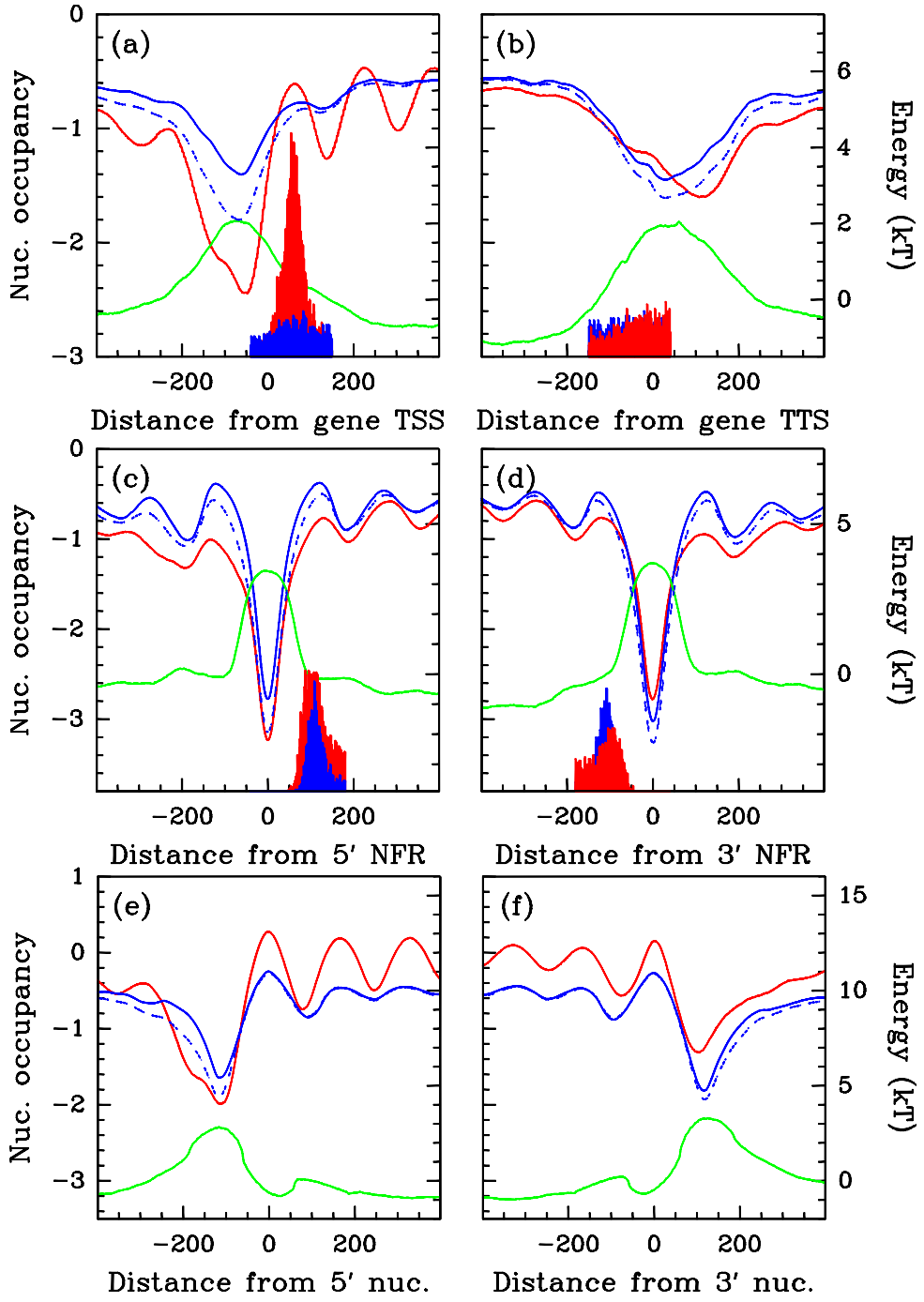


FIGURE 8.4 : *Moyenne du* $\log_2$ *de l'occupation in vivo (en rouge), théorique (en bleu) et de l'énergie (en vert) centré sur (a) le TSS, (b) le TTS, (c) la NFR en 5′, (d) la NFR en 3′, (e) le nucléosome en* $+1$ *du TSS et (f) le nucléosome en* -1 *du TTS. Les paramètres du modèle sont* $(\mu, \delta) = (-1.3, 2)$*. La courbe en pointillé bleu correspond au profil théorique obtenu en rajoutant une barrière énergétique au niveau des sites de fixation des facteurs de transcription.*

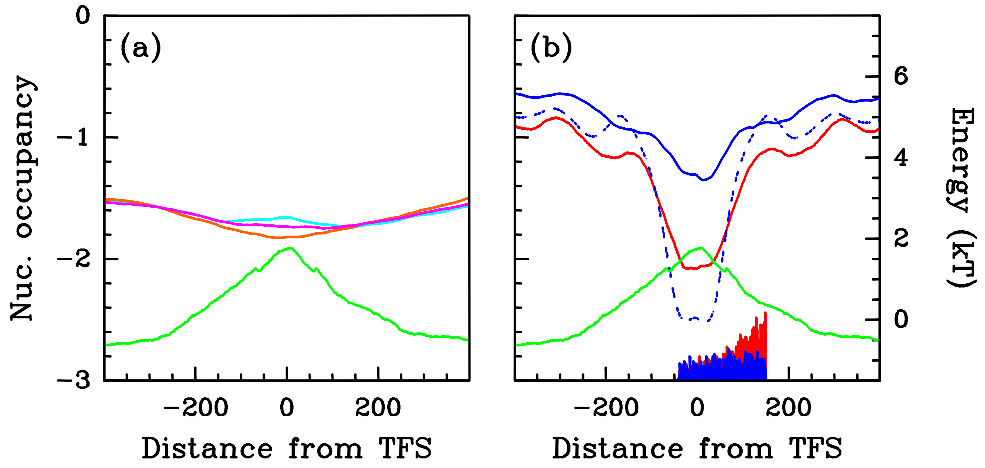


FIGURE 8.5 : *Moyenne du $\log_2$ de l'occupation autour des sites de fixation des facteurs de transcription (TFS). (a) In vitro (orange), théorique Vaillant (bleu clair), théorique Field (rose) et profil énergétique (en vert). (b) In vivo (rouge), théorique (bleu) et modèle avec des barrières énergétiques supplémentaires sur les TFS (en pointillé bleu). L'histogramme rouge (resp. bleu) correspond à la distance expérimentale (resp. théorique) entre la NFR et le TFS.*

délocalisés au niveau du promoteur et dans les classes bleues, il y a des sites également mais concnetrés au voisinage du TSS). En analysant maintenant les deux classes "OPN" et "DPN" de Tirosh & Barkai (Fig. 8.8) on remarque tout de même que notre modèle et ce en accord avec les données *in vitro*(Fig. 8.8 (a,b), orange) révèle une barrière énergétique plus forte (Fig. 8.8 (a,b), vert) et donc une déplétion plus marquée à faible densité (Fig. 8.8 (a,b), cyan) dans le cas des gènes "DPN" (Fig. 8.8 (b)) que des gènes "OPN"(Fig. 8.8 (a)). Cette distinction "OPN" *vs.* "DPN" s'observe donc également dans nos prédictions à haute densité (Fig. 8.8 (c,d)) ; mais bien sur comme dans le cas général, notre modèle ne peut rendre compte qu'en partie du profil *in vivo* quelle que soit la classe OPN ou DPN (contrairement aux cas in vitro qui coincident parfaitement, figure 8.8 (a,b)) : les déplétions prédites dans les deux cas sont trois fois moins importantes qu' *in vivo* (cf. légende de la figure. 8.8), ce qui correspond bien à environ $30\ \%$ de bonne prédiction.

Dans une expérience pionnière, Yuan *et al.* (Yuan et al., 2005) a étudié la conservation des séquences de 7 espèces différentes de *Saccharomyces*.Quand on aligne les promoteurs sur la NFR, et en moyennant les scores de conservation, il apparait clairement que la zone qui contient la NFR est conservée alors que les zones intergéniques le sont beaucoup moins. Puisque ces régions conservées contiennent des zones bien au delà des sites bien connus de fixation des facteurs de transcription (Kellis et al., 2003), ils ont interprété ces résultats par la présence de multiples séries de poly($dA : dT$) qui sont connus pour leur faible affinité avec le nucléosome (Iyer and Struhl, 1995; Suter et al., 2000; Field et al., 2008a; Mavrich et al., 2008a; Kaplan et al., 2009a; Bao et al., 2006; Whitehouse and Tsukiyama, 2006).Dans une étude plus récente, Washietl *et al.* Washietl et al. (2008) a confirmé qu'en moyenne, on observe un taux plus faible de substitutions dans les *linkers* que là où se positionnent les nucléosomes, et ce, que ce soit dans les régions intergéniques ou intragéniques.En fait, dans les régions intergéniques, les NFRs observées au TSS et au TTS semblent relativement plus conservées. Une partie de cette conservation est probablement due aux sites de fixation des facteurs de transcription. Une autre peut être interprétée par la présence des séries de poly($dA : dT$).

Ces études ont été récemment complétées par Tirosh et Barkai (Tirosh et al., 2010) qui ont étudié comment l'évolution de la structure de la chromatine et la divergence d'expression sont-elles liées et dans quelle mesure ce lien peut il s'expliquer par des mutations dans la séquence (effets en *cis*). Ils montrent à travers la comparaison entre *S. Cerevisiae* et *S. Paradoxus* que la majorité ($\sim 70\%$) des différences inter-espèces de structure du chapelet sont dues à des différences de séquence locale (effets "intrinsèques", en *cis*) laissant donc une part significative (30%) significatives à des différences d'action de facteurs "extrinsèques" (effets en *trans*). En accord avec l'étude plus exhaustive de Tsankov et al. (Tsankov et al., 2010), que les effets en *cis* sont principalement dus à des mutations au niveau des séquences riches en AT (les poly($dA : dT$) et ne sont pas associés à des divergences de séquences positionnantes.

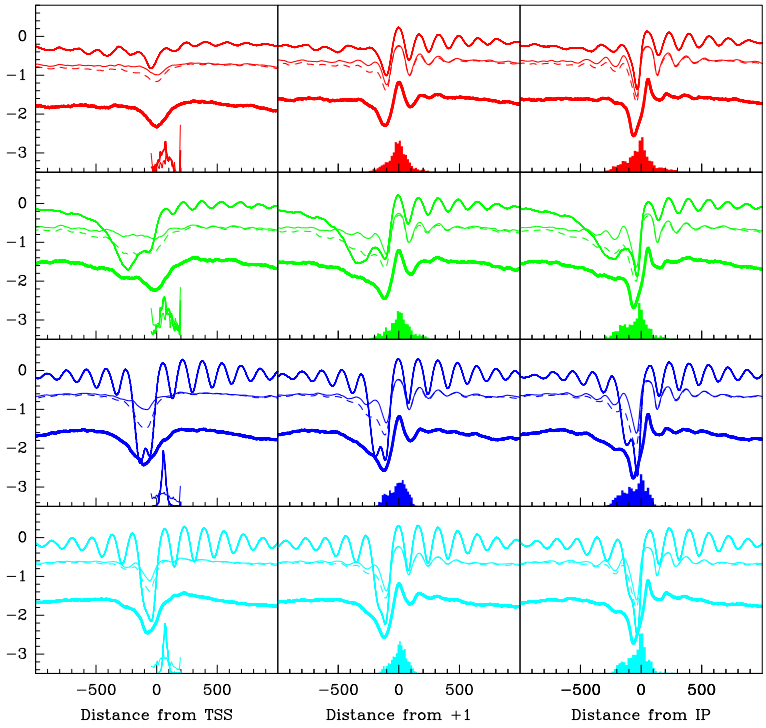


FIGURE 8.6 : *Comparaison entre nos prédictions théoriques et les données expérimentales de Lee au niveau du promoteur. Moyennes sur les differentes classes extraites par Lee (même codage couleur que dans la figure 2.24. A gauche, l'alignement est sur le TSS ; au milieu, réalignement au niveau des nucléosomes +1 des profils individuels ; à droite, réalignement au niveau de la remontée 3′ des NFR des profils individuels. Pour plus de lisibilité (...) les courbes ont été décalées. Moyenne de l'énergie (en fait $-E(s,l)$) (courbe continue, trait épais, décalée vers le bas). Moyennes des données expérimentales de Lee* et al. *(Lee* et al.*, 2007a) (courbe continue, trait assez épais, décalée vers le haut). Moyennes de nos prédictions d'occupation à haute densité dans le modèle sans barrières aux sites des FT (courbes continue fine, milieu) et avec barrières aux sites des FT (tirets, milieu).*

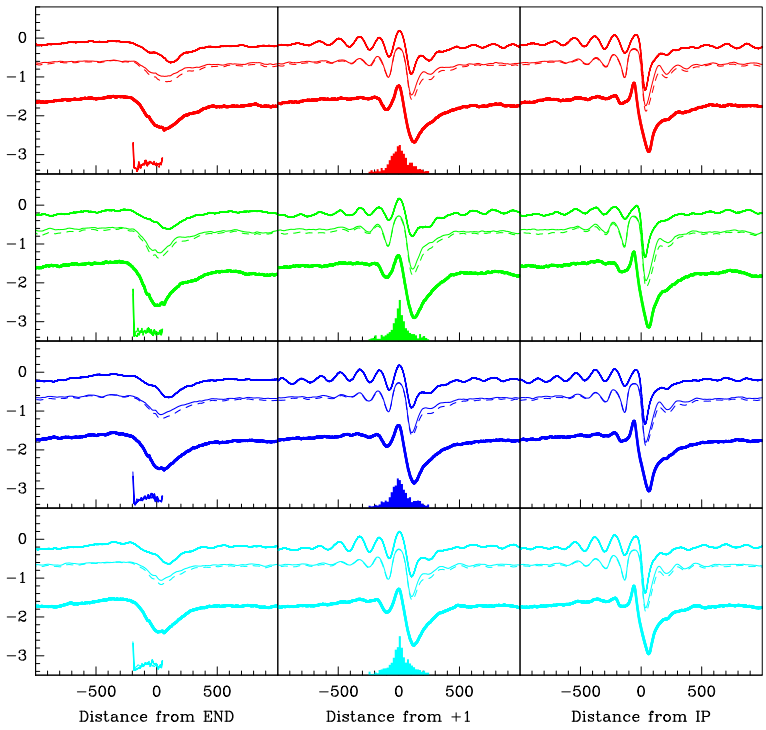


FIGURE 8.7 : *Idem qu'à la figure 8.6 mais au niveau du TTS*

9 L'ORGANISATION NUCLÉOSOMALE DES GÈNES DE LA LEVURE & RÉGULATION DE LA TRANSCRIPTION

Certes, une faible proportion des nucléosomes $+1$ qui jouxtent le promoteur sont situés sur des séquences qui présentent des périodicités AA/TT (Ioshikhes et al., 2006; Mavrich et al., 2008a; Shivaswamy et al., 2008) (Fig. 9.2(D)), mais il est peu probable que les successions de nucléosomes bien positionnés qui sont observées expérimentalement sur l'ensemble du génome et en particulier à l'intérieur des gènes soient induites par une périodicté dans la séquence(Fig. 9.2(E)). La plupart des nucléosomes observés expérimentalement ne sont pas associés à des séquences fortement positionnantes ((Peckham et al., 2007; Yuan and Liu, 2008a)). L'arrangement nucléosomal observé expérimentalement n'est pas une conséquences de séquences qui positionneraient chacun des nucléosomes, mais plus probablement du positionnement statistique induit par la présence des NFRs aux extrémités des gènes. C'est ce que démontrent la comparaison entre les profils *in vitro* et in vivo au niveau des extrémité 5′ et 3′ des gènes de la levure, reportés à la figure 9.1 : dans le cas *in vivo* on voit un ordonnancement périodique émanant des NFR tandis que *in vitro* seuls les nucleosomes $+1$ et -1 semblent être bien positionnés. D'ailleurs ce positionnement observé dans cette chromatine in vitro faiblement dense résulte très certainement de l'effet de la présence d'une barrière énergétique au niveau du NFR (Fig. 9.1 (e,f)) qui induit un confinement favorisant le positionnement du nucléosome adjacent, positionnement renforcé par la présence de périodicité (voir plus haut) et comme le révèle notre profil énergétique par la présence d'une petite "barrière" locale (lègère séquence antipositionnante) sur l'extrémité 3′ (resp 5′) des nucléosomes +1 (resp. -1).

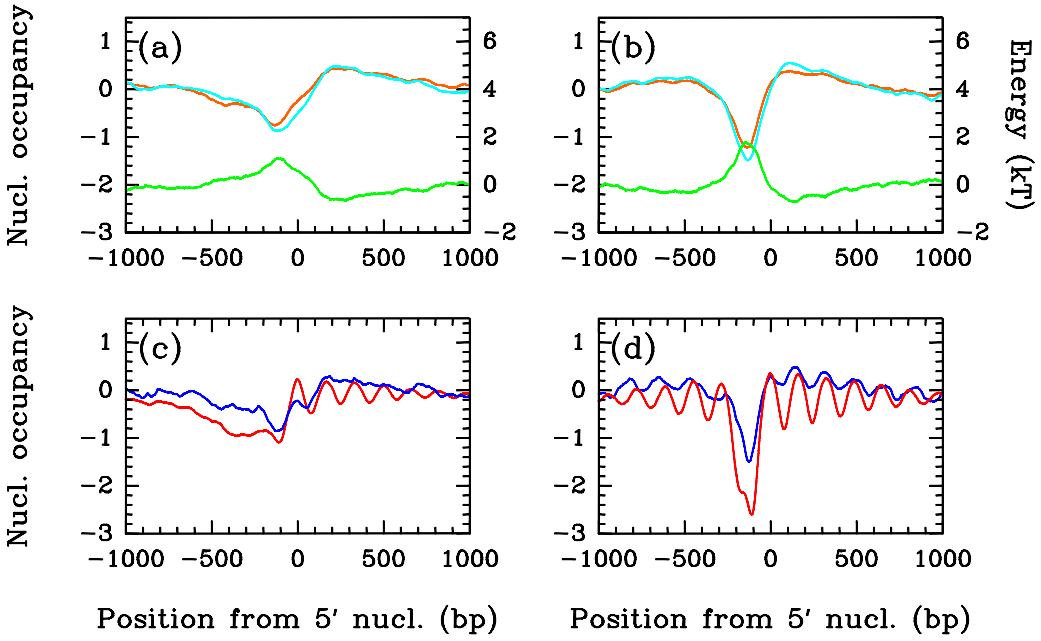


FIGURE 8.8 : *Average nucleosome occupancy profiles plotted around the 5′ nucleosome of* ($N = 503$) *OPN (a,c) and* ($N = 458$) *DPN (b,d) yeast genes (Tirosh and Barkai, 2008*b*).* In vitro *data (orange),* in vivo *data (red), physical model* $\delta E = 2$ *kT,* $\tilde{\mu} = -6$ *kT (light blue) and* -1.3 *kT (dark blue). In (a,b) are shown for comparison the corresponding theoretical energy profiles. Note that in (c,d), the mean theoretical nucleosome occupancy profile has been multiplied by a factor* 3 *indicating that the* in vivo *data result from the combination of "intrinsic" and "extrinsic" (TFs, chromatin regulators, Pols) sequence specific force fields.*

Structure interne du gène

Revenons à la carte nucléosomale 2D des gènes de la levure présentée en section ... L'organisation périodique au sein des gènes reflète ce qui a été observé aux sections 5.4 et 5.5 lorsqu'on confine des nucléosomes entre deux barrières, à savoir une alternance de zones bien ordonnées et de zones floues. Par exemple si l'on observe en détail les petits gènes (d'une taille L inférieure à 1.5 kb), on s'aperçoit (figure 9.3 (C)) que des zones avec n nucléosomes bien positionnés alternent avec des zones où il est difficile de déterminer le nombre et la position des nucléosomes (zones grisées, dites bistables, voir définition dans la partie 9.1). À mesure que L augmente, les gènes se séparent en groupes différents selon que l'on compte $n = 2$ jusqu'à $n = 10$ nucléosomes ($L < 1.5$ kb). Lorsque la taille des gènes est plus grande, l'arrangement périodique des nucléosomes reste visible, mais apparaît de plus en plus flou puisque le confinement induit par les bords est trop éloigné. La dissymétrie de la carte (l'influence de la NFR située en $3'$ s'étend sur 5 à 7 nucléosomes, alors que l'influence de la NFR située en $5'$ ne s'étend que sur 3 à 4 nucléosomes), provient à la fois de la difficulté de définir clairement la fin du gène (figure 2.30 (A)) mais aussi probablement d'un effet supplémentaire qui a tendance à positionner fortement les nucléosomes à proximité du TSS. Le profil moyen centré sur le "$+1$" ou le "-1" s'affranchit du problème d'annotation et conforte cette observation : l'amplitude d'oscillation est plus faible à proximité du TTS que près du début du gène, ce qui suggère effectivement un plus fort confinement au début du gène.

On observe donc un arrangement nucléosomal à l'intérieur des gènes conforme à l'arrangement que prendrait des sphères dures confinées dans un puits de potentiel, dont les bords seraient constitués par les bords de chaque gène. La périodicité du profil de positionnement, l'intensité et la portée du confinement dépendent essentiellement de l'intensité de l'effet d'exclusion imposé sur les bords (en fait la force qui est appliquée sur les bords des gènes) et de la densité moyenne. Les zones intergéniques ont en outre une structure différente (figure 9.4). La structuration est faible à l'intérieur de la zone intergène par rapport à la zone extérieure (*i.e.* la zone intragène). Il est probable que cette forte différence tienne à la dissymétrie de la NFR, qui remonte très rapidement dans la direction du gène, mais dont l'épaisseur et l'extension vers l'extérieur du gène est très variable. Par conséquent, la portée de l'influence de la NFR est plus faible dans la zone intergénique que dans la zone intragénique.

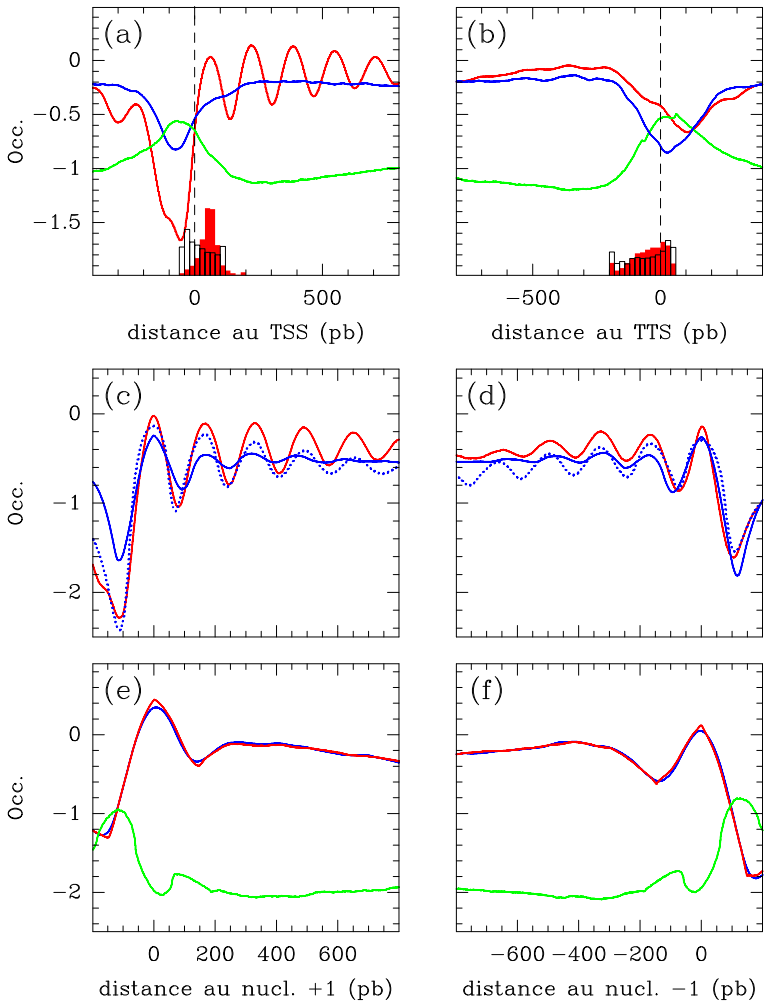


FIGURE 9.1 : *Moyennes in vivo (rouge) et théorique (bleu) de l'occupation nucléosomale et du profil énergétique (en vert) autour (a) du TSS et (b) du TTS. (c) et (d) La même chose, mais où les profils sont alignés sur le nucléosome qui borde le TSS ou TTS dans la direction du gène (+1/-1). La ligne pointillée bleu représente le résultat obtenu lorsque l'on rajoute des barrières énergétiques en plus de l'effet de séquence* $(\mu, \delta) = (-1.3, 0.8)$. *(e) et (f) La même chose que (c) et (d) pour les profils in vitro. En (a) et en (b) l'histogramme rouge (resp. blanc) correspond à la distribution expérimentale (resp. théorique) du TSS (a) ou du TTS (b).*

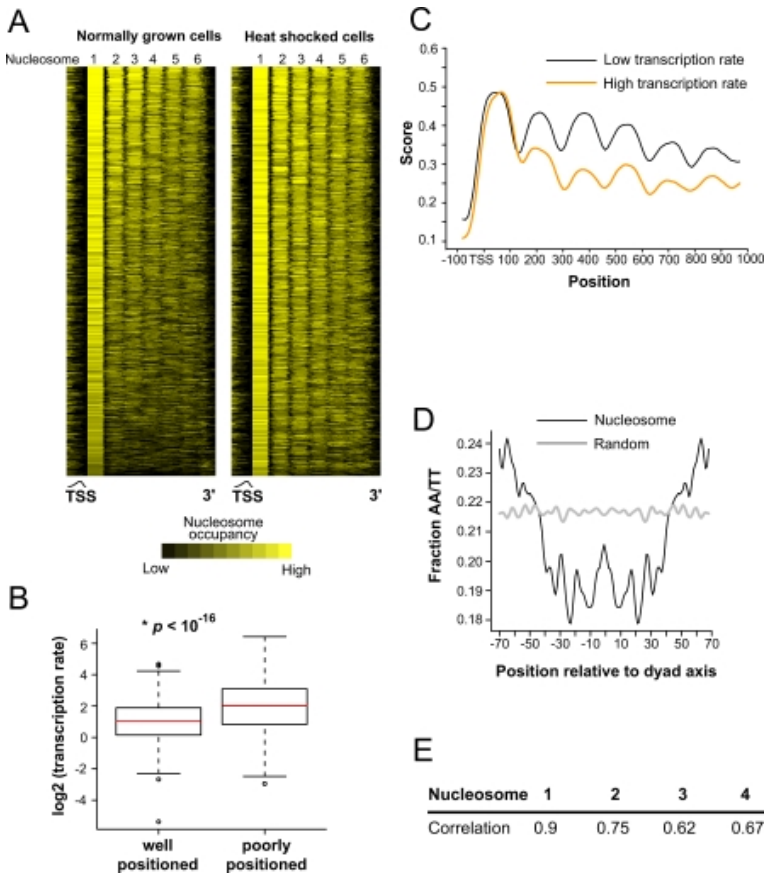


FIGURE 9.2 : *Nucleosome Positioning over Coding Regions Depends on Transcription Rate and Sequence Characteristics (A) Genes were aligned to the first nucleosome downstream of the TSS and sorted by their nucleosome positioning periodicity (NPP) score (see Materials and Methods). Genes were sorted by their NPP scores in normally growing cells, and the data from heat-shocked cells are shown in the same order. The unaligned TSS is indicated by the approximate curve. (B) The transcription rate of genes with high NPP scores (well-positioned nucleosomes) is significantly lower than that of genes with low NPP scores (poorly positioned). In these box plots, the red line indicates the median, the upper and lower bounds of the box indicate the interquartile range, the horizontal lines that are connected to the box by a dashed line indicate the upper and lower bounds of nonoutlier values, and the open circles indicate outliers. (C) Genes were sorted in descending order according to their transcription rates, and the average nucleosome profiles over the coding regions for top 500 genes (orange) and the bottom 500 genes (black) are plotted. (D) Frequency of AA/TT dinucleotide at each position in the DNA sequence associated with the most strongly positioned first nucleosomes. The frequency profiles for the dinucleotides AA and TT for the first nucleosome shown in (A) were summed and smoothed using a 3-bp moving average. The same analysis was also performed for a comparable set of randomly chosen DNA sequences from the yeast genome. (E) Correlation coefficients of the AA/TT profiles for the DNA sequences underlying each of the indicated coding nucleosome positions from (A), with the positioning profile derived earlier. Each of the correlation values was significantly higher than background.*

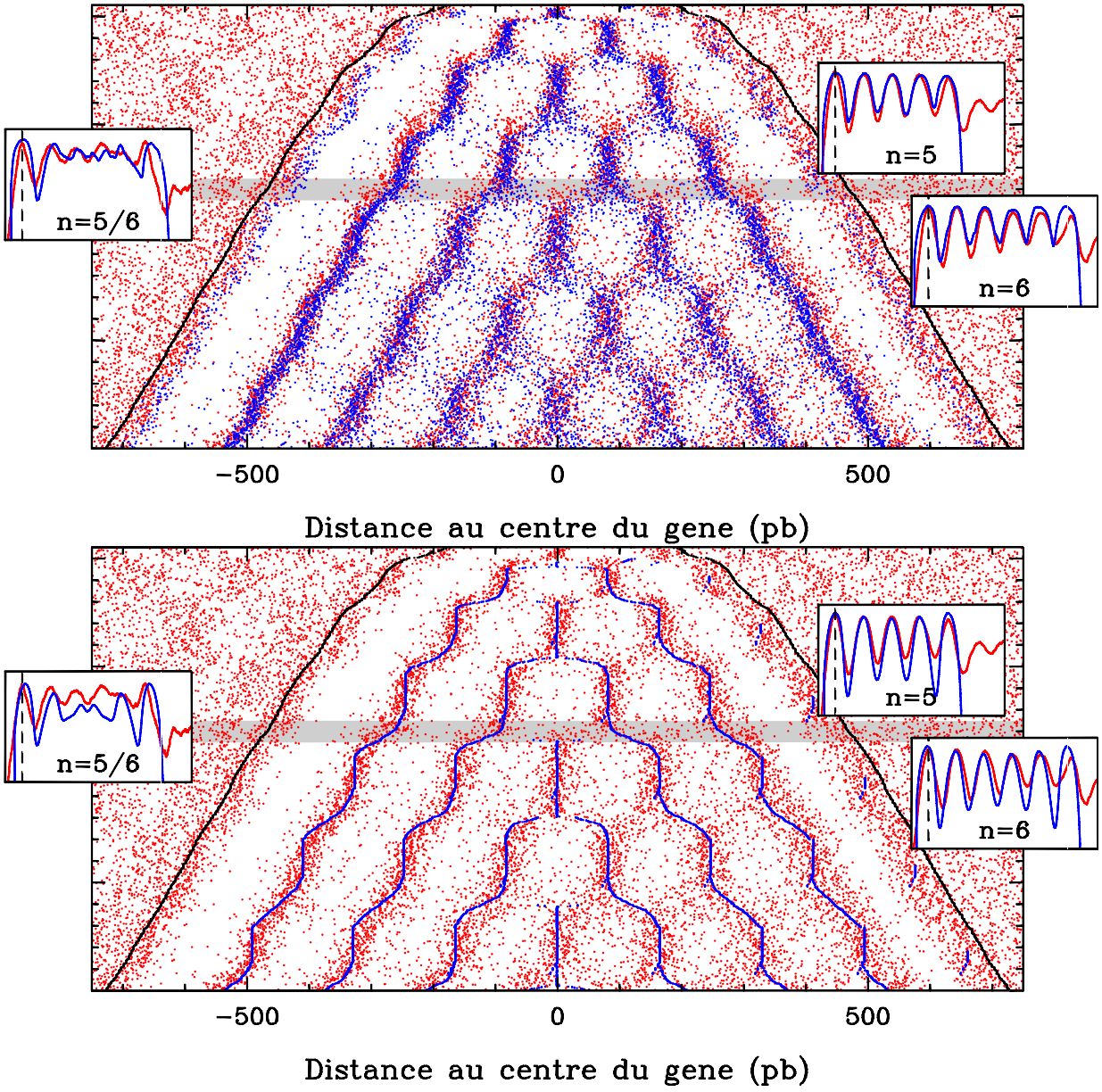


FIGURE 9.3 : *Carte nucléosomale 2D le long des gènes de la levure, pour les gènes de taille* $< 1500pb$*, cf. Figure 2.31 : (haut) : On a rajouté en bleu les résultats du modèle considérant un gène comme un profil énergétique faiblement influencé par la séquence, bordé par des barrières énergétiques (voir figure 9.7) ; (bas) : On représente cette fois les résultats du modèle considérant un gène comme un profil énergétique plat bordé par des barrières énergétiques (équation 9.1).*

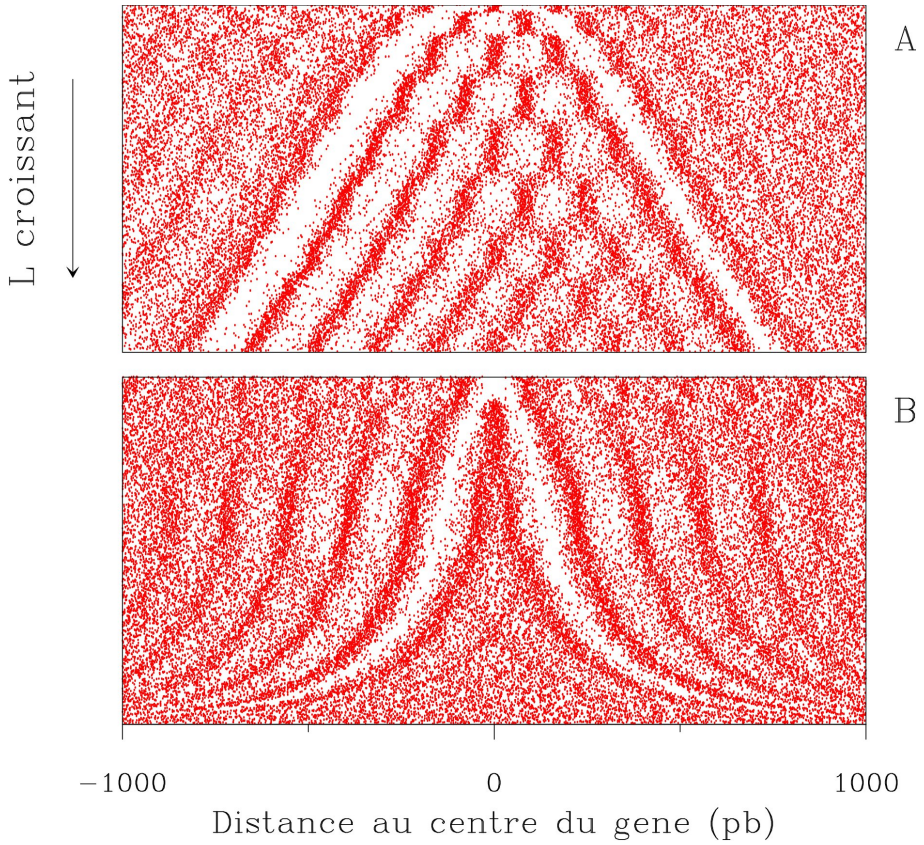


FIGURE 9.4 : *Comparaison entre l'occupation à l'intérieur et à l'extérieur des* 2500 *gènes les plus courts. (a) Occupation intragénique (figure 2.30) (b) Occupation intergénique :* 3300 *zones intergéniques sont considérées. Les profils sont rangés verticalement en fonction de la distance entre les nucléosomes situés en* $+1$ *du TTS et* -1 *du TSS.*

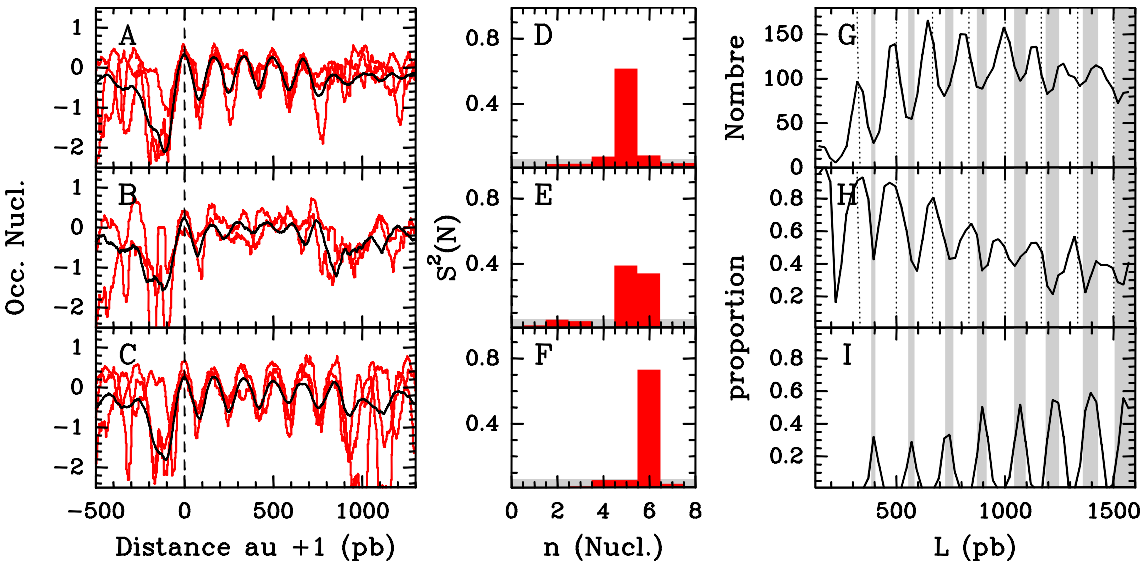


FIGURE 9.5 : *Première colonne : occupation moyenne dans les zones cristallines et bistables (en noir) superposée avec des exemples de gènes particuliers (en rouge). (A) Gènes cristallins avec* $n = 5$ *nucléosomes. (B) Gènes bistables à la transition* $n = 5$ *et* $n = 6$*. (C) gènes cristallins avec* $n = 6$*. Deuxième colonne : spectres de puissance de l'occupation de Whitehouse (Whitehouse* et al.*, 2007). (D) Spectre de puissance d'un gène particulier avec un pic dominant à* $n = 5$*. (E) Spectre d'un gène bistable. (F) Spectre d'un gène à* $n = 6$*. Troisième colonne : (G) nombre de gène en fonction de la taille* L*. (H) Proportion des gènes cristallins. (I) Proportion de gènes bistables* vs L*.*

9.1 DEUX ORGANISATIONS DISTINCTES CRISTALLINES ET BISTABLES

Selon la distance qui sépare les barrières inhibitrices de début et de fin de gène, deux cas sont envisageables : bistable ou cristallin. Ces deux situations correspondent à des configurations d'expression différentes.

Bistable and cristalline genes correspond to two distinct modes of expression.

Afin de distinguer les différentes populations de gènes selon leur structure nucléosomale, nous avons mesuré le spectre de Fourier associé à chacun des profils d'occupation. Nous les avons assigné aux catégories suivantes : cristalline (1940/4554), bistable (946/4554) et indistincte (1668/4554). Les gènes avec un seul pic bien distinct dans leur spectre de Fourier sont considérés comme cristallins (figure 9.5 A,C,D,F). Les gènes avec deux pics sont considérés comme bistables et enfin, les gènes dont le spectre de Fourier est trop confus, ou sans contribution prépondérante sont dits indistincts. Notons que parmi les grands gènes, certains spectres de Fourier présentaient plus que deux pics bien distincts, suggérant une "multi-stabilité" de la chromatine de ces gènes. Pour chacun des 4554 gènes, nous avons calculé le spectre de puissance $S^2(k), k = 0, ..., L/2$ et le spectre de puissance normalisé $\tilde{S}^2(k) = S^2(k)/\sum_{i=1} S^2(i)$ de l'occupation située entre les nucléosomes $+1$ et le -1. Les principaux maxima de ce spectre ont été déterminés, et seules les périodes situées dans l'intervalle $[125\ bp, 210\ bp]$ sont considérées comme informatives. Un gène est cristallin si son spectre possède un maximum unique d'amplitude supérieure à 0.006. Un gène est bistable si les deux maxima sont d'amplitudes supérieures à 0.006 et si la moyenne des périodes est située dans l'intervalle $[160\ bp, 170\ bp]$.

Distribution des gènes selon L

Il existe une certaine forme de quantification de la distribution de taille L dans la population totale des gènes (figure 9.5 G,H,I), ce qui est naturel étant donné que la transition entre une configuration à n nucléosomes et une configuration à $n + 1$ nucléosomes se fait très rapidement à mesure que la taille accessible augmente (les zones de transition sont très peu épaisses en terme de L). Par conséquent, Les gènes dans une configuration bistable sont sous représentés par rapport au reste de la population. En outre, le critère d'appartenance à un statut bistable que l'on choisit est très strict, et ce pour ne pas contaminer la catégorie bistable avec des gènes cristallins ou plus surement indistincts. La proportion de gènes bistables présente une périodicité à 167 pb (figure 9.5 H), c'est-à-dire qu'à chaque fois que l'on augmente L de 167 on retrouve une nouvelle zone bistable, ce que l'on pouvait observer déjà sur la figure 9.3 (haut et bas). Cette périodicité apparait clairement jusqu'à ce que L atteigne les 1500 paires de bases environ. Les variations de proportion de gènes bistables diminuent à mesure que L augmente, conformément à l'idée que les transitions entre deux régimes sont de plus en plus élargies lorsque l'on augmente l'espace accessible pour des sphères dures par exemple (voir les figures 5.13 et 5.11 où les zones de transitions dans un modèle simple sont évaluées analytiquement). En conclusion, il est possible de définir un régime de cristallisation associé à la longueur L. De larges zones cristallines alternent avec de courtes zones bistables. Aux transitions entre les domaines n et $n + 1$ (figure 9.5 I), on peut établir une fenêtre dans laquelle l'occupation est floue (zone grisée sur les figures 9.3 et 9.5 G,H,I). Comme le suggèrent les spectres de puissance mesurés, il est possible que le caractère brouillé des profils émane d'une superposition de configurations à n et $n + 1$ nucléosomes plus qu'à un caractère "liquide" de la chromatine à cet endroit.

Répercutions fonctionnelles

Une question intéressante est de savoir si cette structuration a des répercutions biologiques, et nous avons notamment recherché si les différentes classes (cristallines et bistables) présentaient des spécificités fonctionnelles génétiques. Un faible enrichissement en gènes responsables de l'organisation du cytosquelette est observé sur les gènes bistables (56%,$P = 10^{-2}$). De même, un enrichissement en gènes impliqués dans la translation est observé parmi les gènes cristallins (74%,$P = 10^{-2}$). Ces observations sont sujettes à caution étant données les valeurs des P-values, bien trop élevées pour qu'une conclusion soit tirée.

9.1.1 Une chromatine intragénique construite en accord avec l'équilibre statistique

L'arrangement nucléosomal général à l'intérieur des gènes est expliqué simplement par un équilibre thermodynamique entre deux NFRs.

The nucleosomal array inside genes is simply explained by a thermodynamic equilibrium between two excluding energetic barrier.

Au vu des observations, on souhaite naturellement décrire la structure intragénique de la chromatine comme un fluide de nucléosomes en équilibre thermodynamique. Si l'on applique purement et simplement notre modèle énergétique de formation des nucléosomes, couplé au modèle d'interaction, il n'est pas possible de retrouver cette structure. En effet, même si le modèle énergétique prévoit effectivement une dépletion nucléosomale au début et à la fin d'un grand nombre de gènes, cette déplétion est souvent insuffisante pour rendre compte des résultats expérimentaux. De façon assez claire, la séquence contribue effectivement à dépléter les zones comme les sites de facteurs de transcription, les promoteurs et les fins de gènes, de sorte qu'elles sont généralement disponibles pour l'accès de protéines qui se fixent sur l'ADN (figure 9.1). Toutefois, il faut nécessairement que ces protéines se fixent et c'est leur interaction avec les nucléosomes qui détermine la structure de la chromatine à cet endroit.

Séquence seule

Seule, la séquence ne peut pas prétendre décrire correctement l'arrangement nucléosomal à l'intérieur des gènes (figure 9.1 (a),(b), (c) et (d)). La déplétion induite par la séquence au promoteur et à la fin des gènes n'est ni suffisamment profonde, ni suffisamment phasée pour permettre un arrangement correct intragène (carte 2D non reportée ici).

Barrières seules

Rajouter des barrières énergétiques de forme trapézoïdales (c'est à dire appliquer une force de répulsion sur les bords des gènes) permet de retrouver une structure similaire, sans dispersion (figure 9.3 (bas)). La seule véritable contrainte qui semble ici s'appliquer est une forte déplétion au bord des gènes. Étant donné les ordres de grandeurs d'énergie mis en jeu, et si l'on ne s'intéresse qu'à la structure globale d'un gène et non comment localement les nucléosomes se positionnent, on peut totalement oublier la séquence et modéliser un gène par une simple boîte aux bords de laquelle serait appliquée une force. La situation est donc strictement identique à ce qui se passait dans la partie 5.5. Comme le rappelle la figure 9.6, le potentiel $E(s)$ associé à un gène complet est déterminé par :

$$\left|\begin{array}{lll} E(s) = E_M & & s < s_0 = s_{TSS} + \delta_{TSS} - \Delta \\ E(s) = E_m + (E_M - E_m)(1 - (s - s_0)/\Delta) & & s_0 < s < s_1 = s_{TSS} + \delta_{TSS} \\ E(s) = E_m & \text{for} & s_1 < s < s_2 = s_{TTS} - \delta_{TTS} \\ E(s) = E_m + (E_M - E_m)((s - s_2)/\Delta) & & s_2 < s < s_3 = s_{TTS} - \delta_{TTS} + \Delta \\ E(s) = E_M & & s > s_3 \end{array}\right| \tag{9.1}$$

avec $E_M = 6$kT et $\Delta = 80$pb. Le choix d'une force plutôt qu'une barrière verticale permet d'obtenir une figure 9.3 (bas) cohérente avec l'expérience. Si on choisit une barrière verticale, les zones de transitions n'ont ni la bonne épaisseur, ni le bon espacement. Il faut donc ajuster la force de répulsion pour que les zones de transition soient d'épaisseur et espacées de façon similaire à l'expérience.

Dans cette situation, l'augmentation de la taille accessible (*i.e.* la taille du gène) permet de passer successivement de configurations n à des configuration $n+1$ et ce avec des zones de transitions espacées régulièrement (figure 9.6 à droite). Lors d'une transition, le profil de positionnement est flou et rappelle une situation liquide, alors qu'il est issu de la superposition de configurations déphasées. Il résulte de cette description une carte de positionnement des sphères dures tout à fait similaire à ce que l'on obtient expérimentalement (figure 9.3 (bas)), où les zones de transitions expérimentales présentent des profils conformes à la modélisation (encarts de la figure 9.3 (bas)) Ceci renforce l'idée qu'il est possible de donner une description thermodynamique relativement valide de la structure d'un gène, sans tenir compte de la séquence, simplement en imposant des forces sur les bords du gène.

Contribution de la séquence

Pour rendre compte de l'effet complet de la séquence, des barrières énergétiques et des remodeleurs il suffit de construire un modèle prenant en compte ces trois effets (figure 9.3 (haut)). Si on garde les

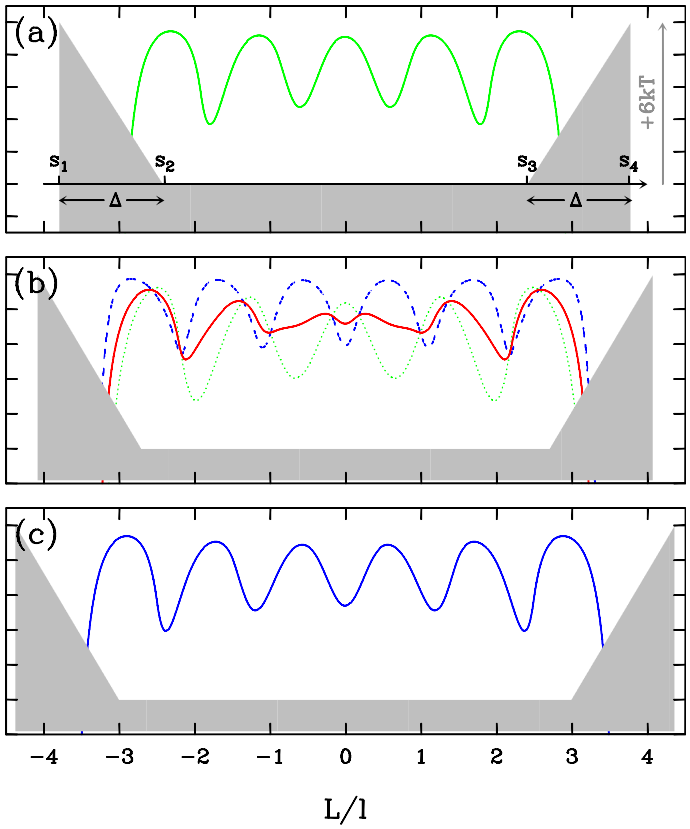


FIGURE 9.6 : *Rappel de l'effet de l'augmentation de la taille accessible L/l sur les profils d'occupation. Transition d'un état à $n = 5$ (a) vers un état à $n = 6$ (c) en passant par une zone bistable n=5/6 (b)*

paramètres que l'on a utilisé sur la levure, à savoir $\delta = 2\ kT$, le fait de rajouter des barrières énergétiques conduit à des profils moyennés incorrects ainsi qu'à des profils trop oscillant de manière générale (figure 9.7 (c)). Pour obtenir une description valide du positionnement, il faut diminuer l'influence de la séquence, à savoir diminuer δ à une valeur $\delta = 0.8$ (figure 9.7 (c) et figure 9.1 (c) et (d) pour la moyenne au TSS et TTS). Encore une fois, il s'avère que la description des données *in vivo* est améliorée si on fait appel à l'effet des remodeleurs. En augmentant la température effective des nucléosomes, ils diminuent de fait l'effet de la séquence. Ils n'ont pas ou peu d'effet sur les facteurs de transcription ce qui explique l'amplitude relative très grande des barrières énergétiques par rapport à la séquence (figure 9.7 (a)).

Lorsque l'on compare les cartes 2D d'occupation des nucléosomes des gènes de levure obtenues avec, d'une part, les données *in vitro* (Fig. 9.8(a)) et, d'autre part, celles prédites par le modèle à faible densité (Fig. 9.8(b)), on observe comme prévu un positionnement des nucléosomes 3' et 5' aux extrémités du gène. Cependant, dans les régions intragéniques, aucune des cartes 2D ne fait apparaître la structure régulière cristalline et bistable des motifs nucléosomaux observée *in vivo*. Comme dans de précédentes études, ces résultats démontrent que les séquences intragéniques sont incapables de générer *in vitro* ces motifs fortement réguliers (à faible densité de nucléosome et en l'absence de facteurs externes comme des remodeleurs, des facteurs de transcription, etc.). Elles confirment que ces motifs cristallins et bistables résultent d'un agencement thermodynamique obtenu à haute densité et en présence de barrières d'énergie aux extrémités des gènes. Cette analyse de l'organisation des nucléosomes intragéniques, en terme d'agencement statistique, est corroborée par la situation dans la région centrale (1kb) des long gènes de la levure, notamment ceux avec $L > 3000$ pb. Dans ces régions éloignées des extrémités et de l'influence des barrières, nous avons trouvé 1270 nucléosomes bien ancrés (Un nucléosome est dit bien ancré si la valeur de l'occupation dépasse localement les 0.2 (sur les données de Lee) ou bien 1.5 (sur celles de Kaplan). Pour le profil théorique, un nucléosome est dit bien ancré si la probabilité d'occupation dépasse localement 0.718). Cela représente une couverture d'environ 39% de la séquence ; c'est moins que la couverture de 80% observée aux extrémités du gène à cause du confinement lié aux effets de bord, mais significativement plus que zéro, ce qui indique que la séquence joue tout de même un rôle dans le positionnement des nucléosomes centraux.

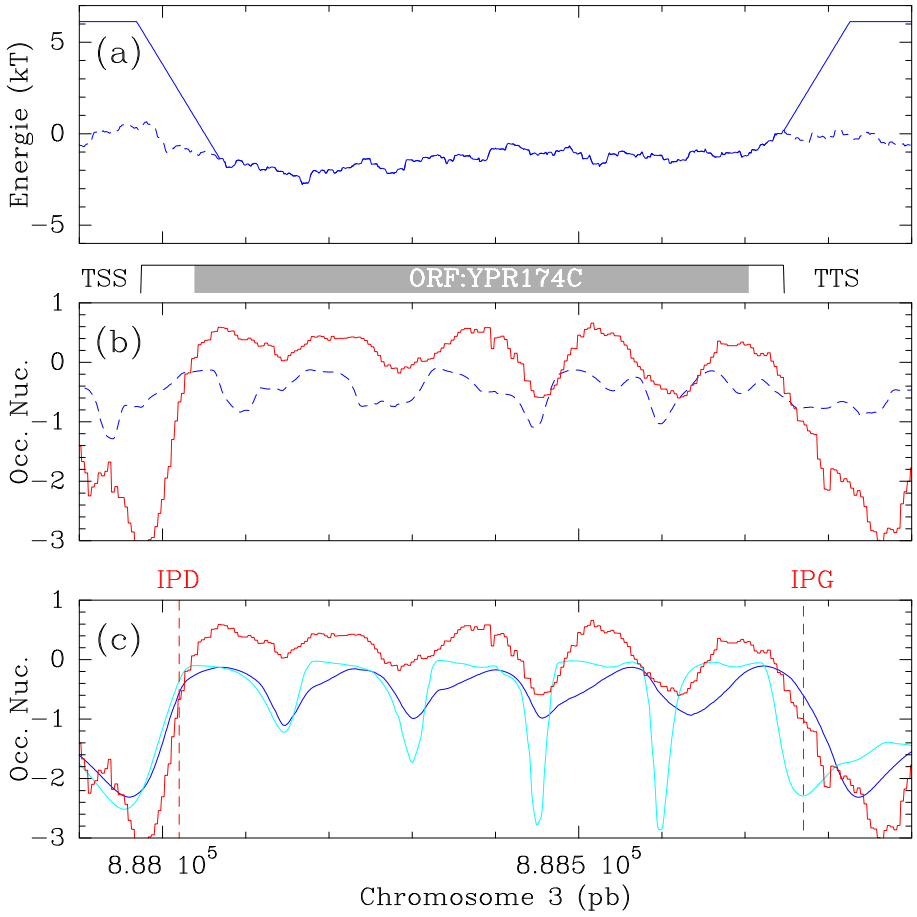


FIGURE 9.7 : *Une modélisation raisonnable d'un gène : une force est appliquée sur les bords, et la séquence contribue au positionnement au milieu ($\delta = 1$). Profil énergétique correspondant en (a). (b) Comparaison entre le profil prédit (avec un modèle sans force appliquée aux bords) et les données de Lee. (c) Comparaison entre la prédiction effectuée avec une force appliquée aux bords en plus de la séquence et les données de Lee. En bleu foncé, $\delta = 0.8$, cyan, $\delta = 2$.*

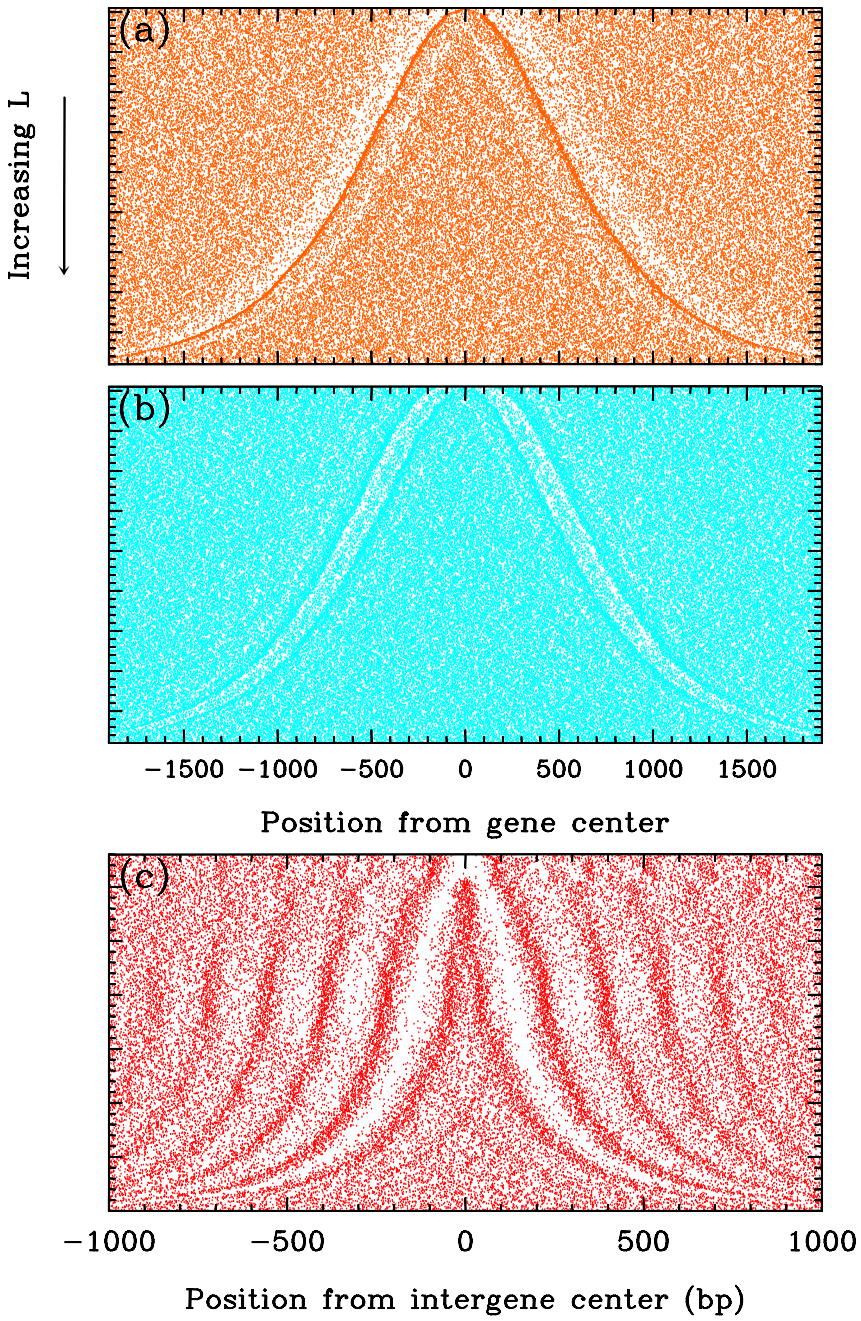


FIGURE 9.8 : *Carte 2D de l'ocupation en nucléosomes le long des gènes de la levure : (a) données* in vitro *de Kaplan* et al. *(Kaplan* et al.*, 2009*b*) ; (b) prédictions de notre modèle avec* $\delta E = 2$ *kT et* $\tilde{\mu} = -6$ *kT, donc à faible densité en nucleosome. (c) Carte 2D pour les régions intergénique (même figure qu'en 9.4)*

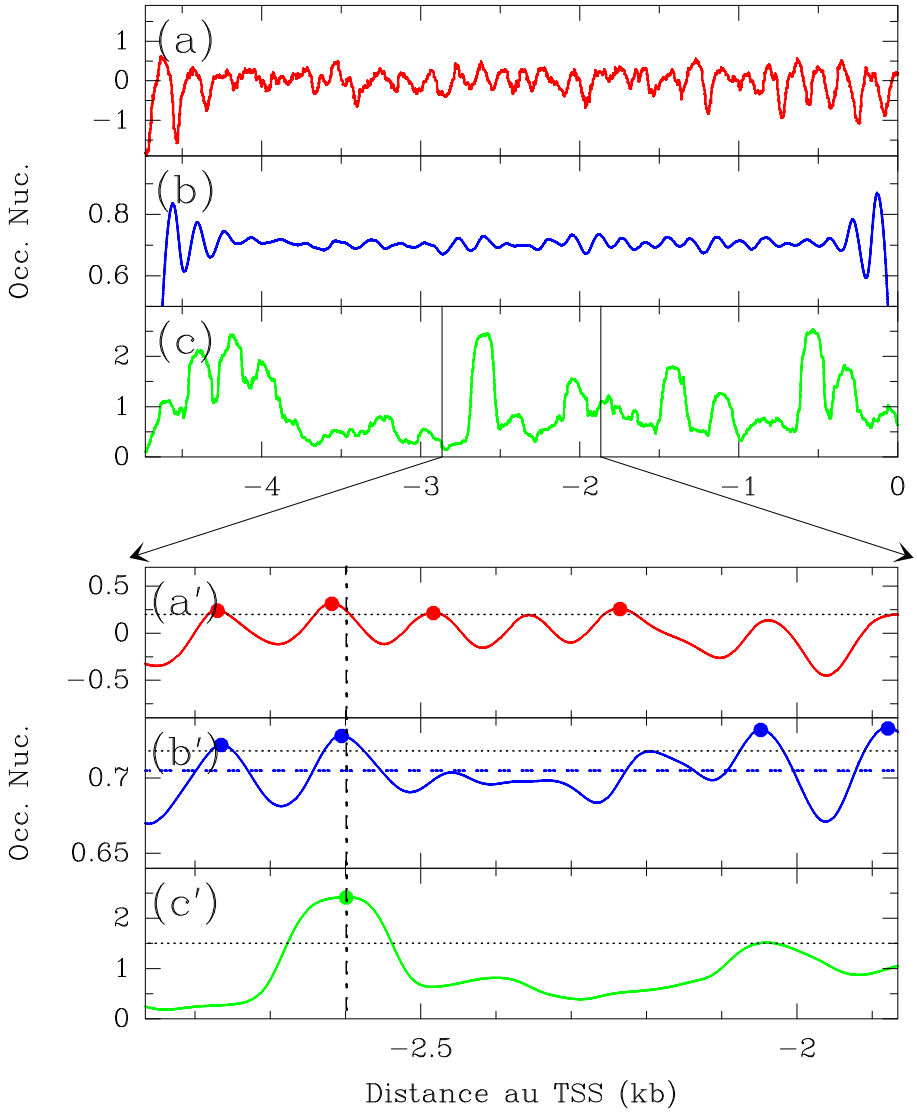


FIGURE 9.9 : *Un exemple de profil d'occupation dans la région centrale d'un long gène. Profil du gène KOG1 ($TSS = 480731$, $TTS = 475999$) du chromosome VIII : (a) données* in vivo *de Lee (Lee* et al.*, 2007*a*) ; (b) modèle théorique avec des barrières énergétiques au bords, et le profil généré par la séquence au milieu (figure 9.7) ; (c) données* in vitro *de Kaplan (Kaplan* et al.*, 2009*a*). (a'-c') Sont des zooms sur la zone centrale ; la ligne pointillée représente le seuil de détermination d'un nucléosome bien ancré (points) ; La ligne verticale pointillée dénote la position d'un nucléosome détecté* in vitro *et* in vivo *et dans le modèle.*

Il est important de noter que parmi ces 1270 nucléosomes observés *in vivo*, seulement 191 correspondent (à une précision de 35bp près) à des nucléosomes bien ancrés également observés dans les données *in vitro* de Kaplan. Puisque l'interaction intrinsèque entre ADN et histones semble influer *in vitro* (donc à faible densité) sur l'ancrage des nucléosomes, cela signifie que moins de 15% $(191/1270)$ des nucléosomes bien ancrés *in vivo* dans les régions intragéniques centrales peuvent être attribués à des séquences ADN fortement positionnantes. Cette fois pour modéliser correctement ce qui se passe loin de barrières énergétiques, on est obligé d'introduire un effet de séquence. Mais cette effet de séquence agit plus par effet de positionnement statistique que par positionnement direct de chacun des nuclésosomes. Les fluctuations dans le profil énergétique phasent la chromatine et la figent dans une situation où chacun des nucléosomes est indépendamment peu positionné (ce qui est déterminé par les données *in vitro*) mais où l'ensemble est relativement robuste à forte densité car changer un nucléosome de place impliquerait un mouvement collectif de beaucoup de nucléosomes voisins. Quand on introduit cet effet de séquence entre les barrières énergétiques et que l'on détermine les nucléosomes bien ancrés, on trouve 1309 nucléosomes dans cette situation. On retrouve notamment 28% des nucléosomes bien positionnés *in vivo* $(360/1270$ ce qui correspond tout à fait au fait que la corrélation entre les données *in vivo* et le modèle tiré de la séquence seule est de $0.3)$. On retrouve également 53% $(100/191)$ des nucléosomes bien ancrés *in vitro*. Si au lieu de rajouter un effet lié à la séquence sous-jacente, on rajoute un effet lié à une séquence aléatoire, on trouve un nombre similaire de nucléosomes bien ancrés (1140).Puisque la séquence est aléatoire, il y a peu de chance qu'une séquence soit particulièrement favorable, et les nucléosomes qui sont bien ancrés le sont essentiellement par positionnement statistique. Toutefois, si l'on applique une séquence aléatoire, on perd une majorité des nucléosome bien ancrés *in vitro* $(74/100)$.

En conclusion, une majorité du positionnement observé *in vivo* des nucléosomes tient au positionnement statistique, c'est-à-dire par effet d'interaction stérique entre les nucléosomes eux-mêmes mais aussi avec tout autre objet qui se trouve sur l'ADN. Une petite partie du positionnement peut être attribué à la séquence seule qui impose notamment la fixation d'un petit nombre de nucléosomes. Puisqu'on ne connait pas la position et l'empiètement exacts de tous les agents qui se fixent sur l'ADN, il est naturel qu'il soit difficile de prévoir le positionnement observé *in vivo*. En revanche, la prédiction est possible pour les expériences *in vitro* parce que ces agents ne sont pas présents, et que seule la séquence intervient.

9.1.2 La densité nucléosomale intragène corrèle avec le taux de transcription

Tout comme l'occupation au promoteur influence l'expression des gènes, la structure interne du gène joue aussi un rôle.

As much as the occupation at the promoteur influence gene expression, the nucleosomal structure inside the gene is very related to the gene transcritpion regulation.

Ensuite, il est possible d'évaluer les implications fonctionnelles des différents motifs chromatiniens que l'on observe au sein des gènes. Commençons par le taux de transcription, que l'on évalue par l'intermédiaire de la densité en ARN-polymérase II (figure 9.10 (A) (Steinmetz et al., 2006)). Nous recherchons comment ce taux de transcription moyen des gènes change en fonction de leur taille L. Le taux de transcription moyen des gènes dont la taille est située dans une boîte glissante de 50 pb est évalué, puis normalisé par le nombre de gènes dans cette même boîte de 50 pb. Il apparaît clairement que dans chacun des domaines à n fixé, le taux de transcription décroit avec L, de façon plus chiffrée, il apparaît notamment une anti-corrélation entre le taux de transcription et les domaines cristallins : $n = 3$, $r = 0.2,\ P = 3.6\ 10^{-2}$; $n = 4$, $r = -0.24,\ P = 7\ 10^{-4}$; $n = 5$, $r = -0.21,\ P = 3.5\ 10^{-3}$; $n = 6$, $r = -0.3,\ P = 1.3\ 10^{-3}$ et $n = 7$, $r = -0.23,\ P = 5\ 10^{-3}$. Plus l'assemblage nucléosomal à l'intérieur du gène est compact, plus le *linker* est petit, et plus la transcription est élevée. Á chaque fois que l'on passe d'un domaine à n vers un domaine à $n + 1$ la compaction augmente subitement, et du coup le taux de transcription moyen augmente subitement lui aussi.

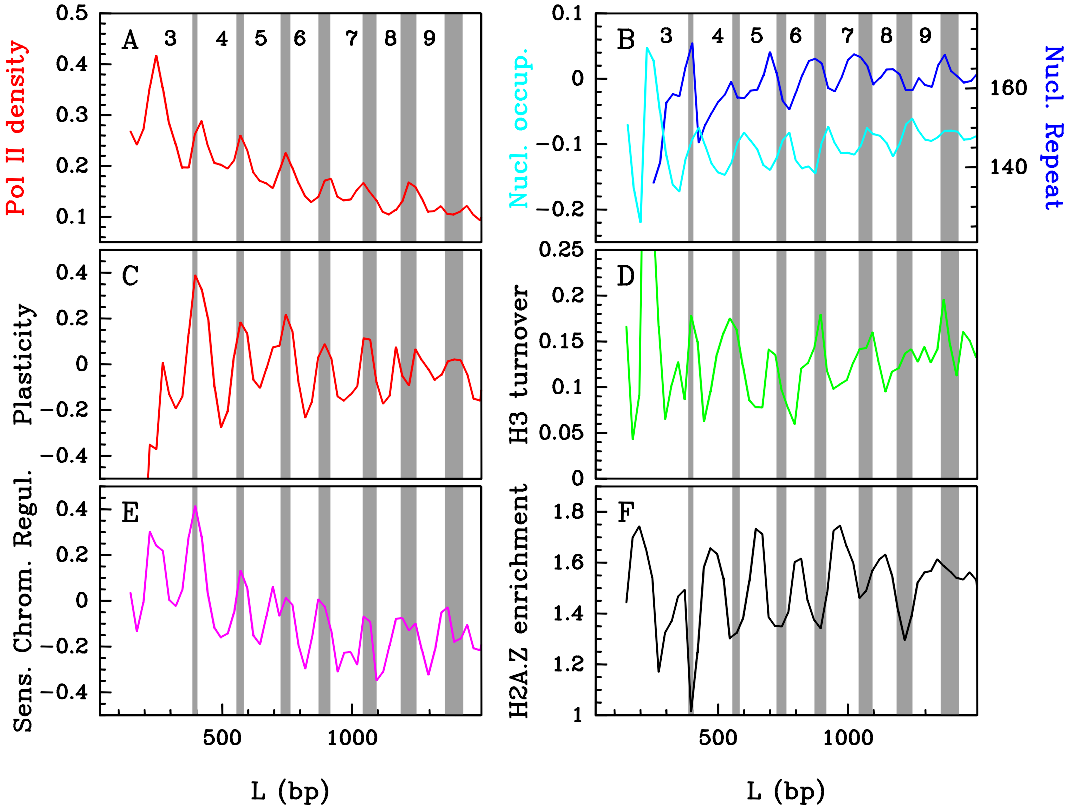


FIGURE 9.10 : *La bistabilité des gènes contrôle l'expression des gènes. (A) Analyse glissante (fenêtre de 50 pb) en fonction de L sur : (A) la moyenne du taux de transcription, estimé à partir de la densité de Pol II (Steinmetz* et al.*, 2006). (B) pseudo-NRL (bleu foncé) et occupation moyenne intragénique (bleu clair). (C) Plasticité de la transcription (Tirosh and Barkai, 2008*a*). (D) Taux de remplacement de H3 (Dion* et al.*, 2007). (E) Sensibilité de la chromatine (Steinfeld* et al.*, 2007). (F) Taux d'occupation en H2A.Z (Tirosh and Barkai, 2008*a*). Les zones grisées correspondent aux domaines bi-stables.*

9.1.3 Les gènes bistables ont une expression régulée

Les gènes bistables ont deux configurations chromatinienne distinctes. Ils ont donc une expression particulière.
Bistable genes show two distinct chromatin features. They in turn show a peculiar expression.

Si l'on observe également la distribution de la plasticité, qui quantifie les variations de niveau d'expression d'un gène en fonction de variations environnementales (Tirosh and Barkai, 2008a), il apparait également des fortes variations corrélées avec les différentes zones nucléosomales. Les gènes bistables présentent généralement une forte plasticité (figure 9.10 C). Or les gènes bistables présentent deux états possibles de compaction : l'un relativement dilué avec un NRL de l'ordre de L/n et l'autre plus compact avec un NRL de l'ordre de $L/(n+1)$. Or si l'on croit les résultats selon lesquels le niveau de transcription dépend de la compaction, il est naturel que les zones bistables présentent deux niveaux de transcription différents. Il n'est donc pas étonnant de trouver que leur plasticité, c'est-à-dire la variance du niveau d'expression, est elle aussi élevée. La corrélation entre les zones bistables et la plasticité est donc très grande ($r = 0.42, P = 1.1 \times 10^{-3}$).

Une étude récente (Steinfeld et al., 2007) vient confirmer que le niveau d'expression des gènes peut être bouleversé par un changement de forme de la chromatine. Si l'on affecte les régulateurs de la chromatine tels que HAT (Histone Acetylase Transferase), HDAC (Histone DeACetylase), HMTs (Histone Methyl Transferase) ou tout autre remodeleur de la chromatine fonctionnant à l'ATP, l'expression des gènes concernés est changée. Ces régulateurs de la chromatines ne sont pas distribués de façon homogène sur le génome de la levure, et il se trouve que les gènes bistables sont significativement plus sensibles aux régulateurs que les autres gènes, notamment les cristallins (figure 9.10 E). Il est possible d'interpréter cette observation par le fait que les gènes bistables n'ont pas de configuration nucléosomale fixe. La configuration nucléosomale structurale des gènes cristallins n'existe pas pour les gènes bistables.

Au final, les gènes bistables sont des gènes dont l'architecture nucléosomale est plus plastique et plus dynamique sans doute que les gènes cristallins. Puisque l'expression d'un gène est certainement en partie régulée par cette architecture sous-jacente, les gènes bistables sont des éléments privilégiés pour une régulation fine de l'expression.

9.1.4 Spéculation

Il est très vraisemblable que la structure nucléosomale à l'intérieur des gènes affecte directement le niveau d'expression d'un gène. Comment et pourquoi la distance inter-nucléosomale semble directement affecter cette expression reste à déterminer. Il est intéressant de noter que la relation directe entre expression et chromatine que l'on observe à l'intérieur des gènes est similaire à ce qui a déjà été obtenu sur le promoteur par Tirosh et Barkai (Tirosh and Barkai, 2008a). Au vu de cette proximité fonctionnelle, nous avons recherché un lien entre ces gènes OPN et DPN avec nos gènes bistables et cristallins, mais nous n'en n'avons pas trouvé. Rappelons que la définition DPN/OPN fait référence à ce qui se passe à l'intérieur du promoteur et non à l'intérieur du gène. Ceci argumente en faveur de deux mécanismes différents de régulations, reliés à la structure de la chromatine sur le promoteur et à l'intérieur du gène. Dans tous les cas, ces observations posent la question du mécanisme à travers lequel la forme locale de la chromatine affecte l'expression. Il a été proposé par Morse (Morse, 2007a) qu'une chromatine compacte restreint les possibilités d'attachement de facteurs de transcription et d'autres facteurs. Il pourrait donc y avoir une perturbation du bon déroulement de l'élongation, de la progression de PolII ou encore un déclenchement de transcriptions cryptiques indésirables. Par ailleurs, selon la modélisation géométrique tridimensionnelle obtenue par Lesne et Victor (Lesne and Victor, 2006), une structure nucléosomale régulière avec une NFR courte telle que celle produite dans les gènes cristallins, favoriserait la progression séquentielle de PolII. Par exemple, les réversomes (des nucléosomes de chiralité inversée par rapport au nucléosome classique) ont récemment fait l'objet d'études (Lavelle and Prunell, 2007) et semblent se former sous l'action du superenroulement de l'ADN à proximité de l'ARN-polymerase. Ce réversome faciliterait l'élongation en permettant à la polymérase de faire levier sur les dimères H2A-H2B bien ancrés sur l'ADN et qui perturbent la transcription en l'absence de tout autre facteur (Bancaud et al., 2007). Un arrangement nucléosomal bien régulier favoriserait la formation d'une onde réversome qui progresserait très vite à travers la chromatine et permettrait une transcription rapide des gènes.

Un gène peut vraisemblablement être décrit phénoménologiquement par : (i) une NFR imposée en partie par la séquence au TSS et en partie par la présence de facteurs extérieurs. (ii) une NFR en fin de gène, essentiellement encodée dans la séquence. (iii) une séquence intragénique peu influente, ou bien dont l'influence a été diminuée par les facteurs de remodelage.

10 CHAPELET ET INSERTION VIRALE

Dans une récente collaboration avec Marc Lavigne (Laboratoire Joliot-Curie) et Vincent Parissi (Université de Bordeaux 2) (Lesbats et al., 2011) nous avons étudié dans quelle mesure le chapelet nucléosomal *in vitro* pouvait influencer l'activité de l'intégrase du virus HIV. Nous avons montré, expériences *in vitro* et modélisation à l'appui, que l'intégration est favorisée dans des chapelets plus labiles, avec des nucléosomes plus faciles à déplacer (ceci n'excluant pas le fait que l'intégration puisse *in fine* être favorisée au niveau d'un nucléosome). Comme le montre la figure 10.1, nos prédictions du chapelet nucléosomal indique que les sites d'intégration de HIV dans des cellules humaines Wang et al. (2007) sont effectivement plutôt localisées dans un chapelet nucléosomal moins dense qu'en moyenne. (La comparaison avec le données *in vivo* de Schones (Schones et al., 2008) des cellule CD4+ ne donnent par contre pas d'effet apparent).

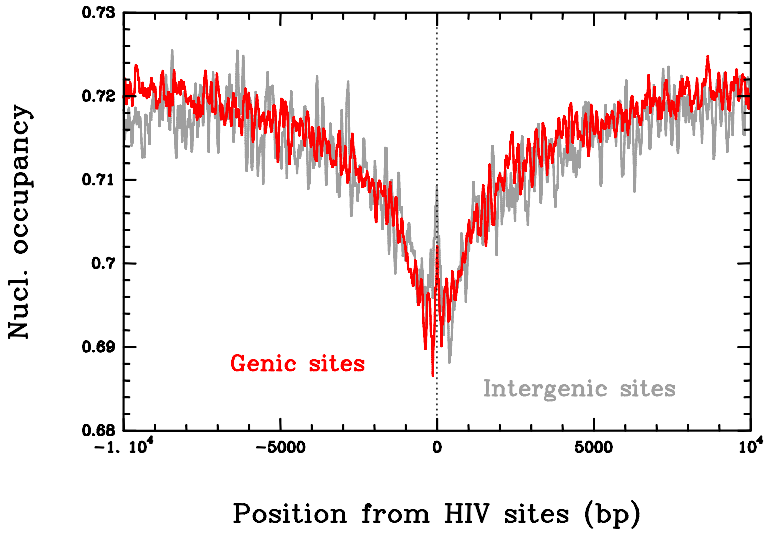


FIGURE 10.1 : *Profils théoriques "moyens" d'occupation en nucléosome autour de sites d'intégration de HIV dans des cellules humaines (type Hela) ; sites intergéniques (gris), sites géniques (rouge).*

11 CONCLUSIONS ET PERSPECTIVES

Le fil directeur des mes projets de recherche depuis la thèse est l'étude du couplage à toute échelle entre la chromatine et les différents métabolismes nucléaire. L'objectif fixé à mon entrée au CNRS était d'identifier et modéliser des mécanismes "chromatiniens" de régulation de la transcription et/ou réplication ; il s'agissait par ailleurs de concentrer sur la contribution de la séquence génomique.

11.1 EFFETS DE SÉQUENCES : QUEL AVENIR ?

Dans ce manuscrit, j'ai donc essayé, au fil de mes travaux réalisés depuis la thèse, de présenter un aperçu de nos contributions et celles d'autres groupes dans l'étude de l'influence de la séquence génomique sur l'organisation du chapelet nucléosomal.Il y a encore évidemment beaucoup de points à

éclaircir quant au rôle précis de la séquence sur la structuration du chapelet nucléosomal. Il n'y a déjà toujours pas de vision clair concernant la spécificité de séquence de l'octamère, et notamment la relation avec le GC ne me parait pas "entendue". Les études récentes sur la cartographie des nucléosomes chez l'homme dans des types cellulaires différents (Valouev *et al.*, Determinants of nucleosome organization in primary human cells. *Nature* (2011)) semblent cependant valider l'effet central de la composition moyenne en G+C. Certains groupes ont ainsi établis des modèles ou émis des conjectures d'évolution de la séquence et notamment de la composition en G+C en rapport avec les données chromatiniennes (Kenigsberg et al., 2010). Donc, au jour d'aujourd'hui on peut dire que les principaux "déterminants" génomiques qui structurent le chapelet in vivo et in vitro via la spécificité de séquence "intrinsèque" de l'interaction histones-ADN sont les $poly(dA : dT)$ et la composition en $G + C$. Pour des génomes où la séquence est faiblement hétérogène comme pour les levures, les effets sont assez faibles on l'a vu, avec une variabilité typique de $1 - 2kT$ mais suffisant pour conditioner le positionnement en cis et/ou le recrutement en trans de régulateurs chromatiniens et/ou fonctionnels. On trouve par ailleurs plutôt un enrichissment vers les sequences défavorables suggérant une règle génomique d"'antipositonnement" plutôt qu'une règles de "positionenment" liée à la périodicité à $10.2pb$. La question est de savoir si chez des organismes comme l'homme, les fortes variations de compositions de G+C (isochores) se reflètent au niveau du chapelet comme on s'y attendrait ; et si, tel n'est pas le cas, quels sont les mécanismes "épigénétiques " de compensation. On voit ici qu'un projet tout à fait intéressant sera, à partir de notre modèle simple, d'étudier le phénomène de co-évolution compensatoire dans les mécanimes de régulation du chapelet nucléosomal et ce à toute échelle. Bien entendu un tel projet benéficiera fortement de l'amélioration (voir plus bas) de notre modèle, pour pouvoir prendre en compte d'autres facteurs régulateurs dont l'activité est liée à l'organisation génomique *via* leurs sites de fixations comme les facteurs de transcriptions et insulateurs (et qui probablement participent à son évolution).

11.2 Dynamique "spatio-temporelle" de l'hétérochromatine

Les chromosomes eucaryotes sont en général composés de deux types de domaines structurels & fonctionnels chromatiniens : l'*euchromatine*, ouverte et généralement accessible, où l'on retrouve la plupart des gènes actifs et *l'hétérochromatine*, fortement condensée, riche en éléments transposables et en séquences répétées. D'un point de vue fonctionnel, l'hétérochromatine contrôle plusieurs aspects fondamentaux du fonctionnement nucléaire : (i) assemblage du kinetochore (ii) cohésion des chromatides soeurs assurant ainsi la bonne ségrégation durant la division cellulaire (iii) recombinaison : inhibition de toute recombinaison inopinée au niveau des séquences répétées garantissant ainsi une stabilité génomique (iv) expression des gènes : répression de la transcription des séquences sous-jacentes et voisines (v) organisation de la fibre de chromatine à grande échelle : activation/repression de certaines interactions à longue distance. L'hétérochromatine est ainsi associée à la différentiation cellulaire, même dans les organismes unicellulaires ou elle contrôle le type cellulaire et la reproduction sexuée. Dans les organismes multicellulaires l'hétérochromatine est impliquée dans la maintenance de l'identité cellulaire au cours du développement. Malgré son importance, les mécanismes de formation et de maintenance (heritabilité), et la manière dont ce type de chromatine exécute ces différentes fonctions restent encore mal connues. D'un point de vue moléculaire, l'hétérochromatine se caractérise par des signatures biochimiques et structurelles particulières : l'hypo-acetylation des histones, la méthylation spécifique H3K9me/H3K27me, l'association avec d es protéines structurales de la famille HP1/Polycomb et par une distribution des nucléosomes très périodique. Il a été montré *in vitro* que l'hypo-acétylation et la régularité de positionnement des nucléosomes facilitent la compaction de la fibre de $30\ nm$. Ainsi le scenario actuellement proposé pour l'assembalge heterochromatinienne repose sur un mécanisme de recrutement et d'action combinée des enzymes de méthylation, des protéines HP1/polycomb, des enzymes de déacetylation et des facteurs de remodelage conduisant à une structure compacte et stable de la fibre. Comme l'indique les récentes études chez la drosophile plusieurs type de domaines hétérochromatiniens et euchromatinienes, se différenciant par des structures, signatures biochimiques particulières et partitionnent le génome. En plus des hétérochaomatines "classiques", "HP1" et "Polycomb" cette étude à notamment identifié un nouveau type d'hétérochromatine, extrêmement peu active transcrip-

tionnellement, la chromatine "noire" (Filion et al., 2010) qui se distingue notamment par son association forte avec les lamines. Il y a par ailleurs dans les processus d'hétérochromatinisation un couplage avec l'organisation spatiale au sein du noyau avec en particulier une tendance forte de localisation de l'hétérochromatine à la périphérie et au voisinage du nucléole (Meister et al., 2010).

Mon objectif au sein du LJC dans les prochaines années est de mettre en place une modélisation biophysique des processus de nucléation, de propagation et de maintenance de l'hétérochromatine le long des génomes qui permettent à (long) terme de modéliser l'orchestration des différents types chromatiniens le long du génome et de leur localisation spatiale au sein du noyau.

Nous nous appuierons sur nos récents travaux qui ont essentiellement porté sur l'étude du chapelet nucléosomal. L'objet de mes futurs travaux sera d'affiner ce travail de modélisation en appui/relation avec des données expérimentales obtenues au laboratoire Joliot-Curie ou récoltées par ailleurs dans des organismes unicellaires tels que la levure *S. Cerevisiae*, *S. Pombe* ou la paramécie, et multi-cellulaires comme le ver *C. Elegans*, la mouche *Drosophila Melanogaster*, Arabidopsis Thaliana et l'Homme. Il faut en effet profiter du "flot" accru de donnée génomiques de plus en plus résolues concernant la structure à toute échelle et le "contenu" biochimique (donc "signalétique") de la chromatine (Filion et al., 2010; Venters et al., 2011; Roudier et al., 2011) : stabilité, positionnement et taille des nucléosomes, marques épigénétiques (modification des queues des histones, variants d'histones, methylation de l'ADN), localisations de régulateurs chromatiniens (remodelleurs, facteurs de transcription, isolateurs, cohésines, lamines), interactions à distance de la fibre ; ces cartes génomiques de la chromatine commencent par ailleurs à être (et le seront à terme de plus en plus fréquemment) extraites de différents types cellulaires à différents stades du développement, à différentes phases du cycle cellulaire. Ces cartes peuvent être par ailleurs de plus en plus fréquemment confrontées aux données transcriptomiques et aux données de timing de réplication.

On peut donc commencer à envisager d'entreprendre (initier) l'étude et la modélisation de la dynamique spatio-temporelle de la chromatine à petite (à l'echelle d'un gène) et grande échelle (à l'échelle d'un domaine) et étudier son couplage avec les dynamiques spatio-temporelles de la transcription et réplication.

Pour mieux appréhender les mécanismes de structuration du chapelet intervenant *in vivo* notamment dans les processus d'hétérochromatinisation, nous proposons donc :

(1) D'affiner notre modèle intrinsèque en autorisant notamment une enroulement (ou protection) de l'ADN fluctuante et à introduire des intéraction effectives entre nuclésomes (plus proches voisins), des intéractions spatiales entre nucléosomes induites par des contraintes de compaction du chapelet en fibre de $30nm$ (par exemple par la présence de H1 et/ou d'hypo-acétylation H4,H3).

(2) D'intégrer des facteurs extrinsèques comme :

- la fixation de facteurs de transcription, d'insulateurs (qui peuvent présenter une spécificité de séquence plus ou moins forte) ou autres protéines architecturales comme des histones de liaison (H1) ou des protéines de type HP1 ciblés par le présence de certaines modificatiosn des queues d'histones (et qui peuvent "ponter" deux nuclésomes successifs).
- la présence de certains variants d'histones (modulant la stabilité du nucléosome), de certaines modifications biochimiques des queues des histones (modulant l'interaction nucléosome-nucléosome, ou recrutant des protéines hétérochromatinienne)
- l'action de facteurs de remodelages qui peuvent repositionner ou éjecter des nucléosomes (et ce de facon plus ou moins dépendante de la séquence génomique) et qui peuvent être recrutés par les facteurs de transcriptions et les modifications des queues d'histones.

(3) Mettre en place un modèle de fibre de chromatine (niveau de compaction, structure) ainsi qu'un modèle d'organisation de cette fibre au sein des noyaux. L'idée étant dans un premier temps d'étudier les mécanisme de ségrégation spatiales des domaines génomùiques eu- et hetero-chromatiniens.

(4) D'introduire l'effet de la réplication ; à partir des connaissances sur les bases moléculaires associées à la restructuration de la chromatine après le passage de la fourche, on s'attachera à étudier l"effet de la réplication sur la stabilité des domaines héterochromatiniens et donc son influence sur les transitions épigénétiques à grande échelle des domaines chromatiniens.

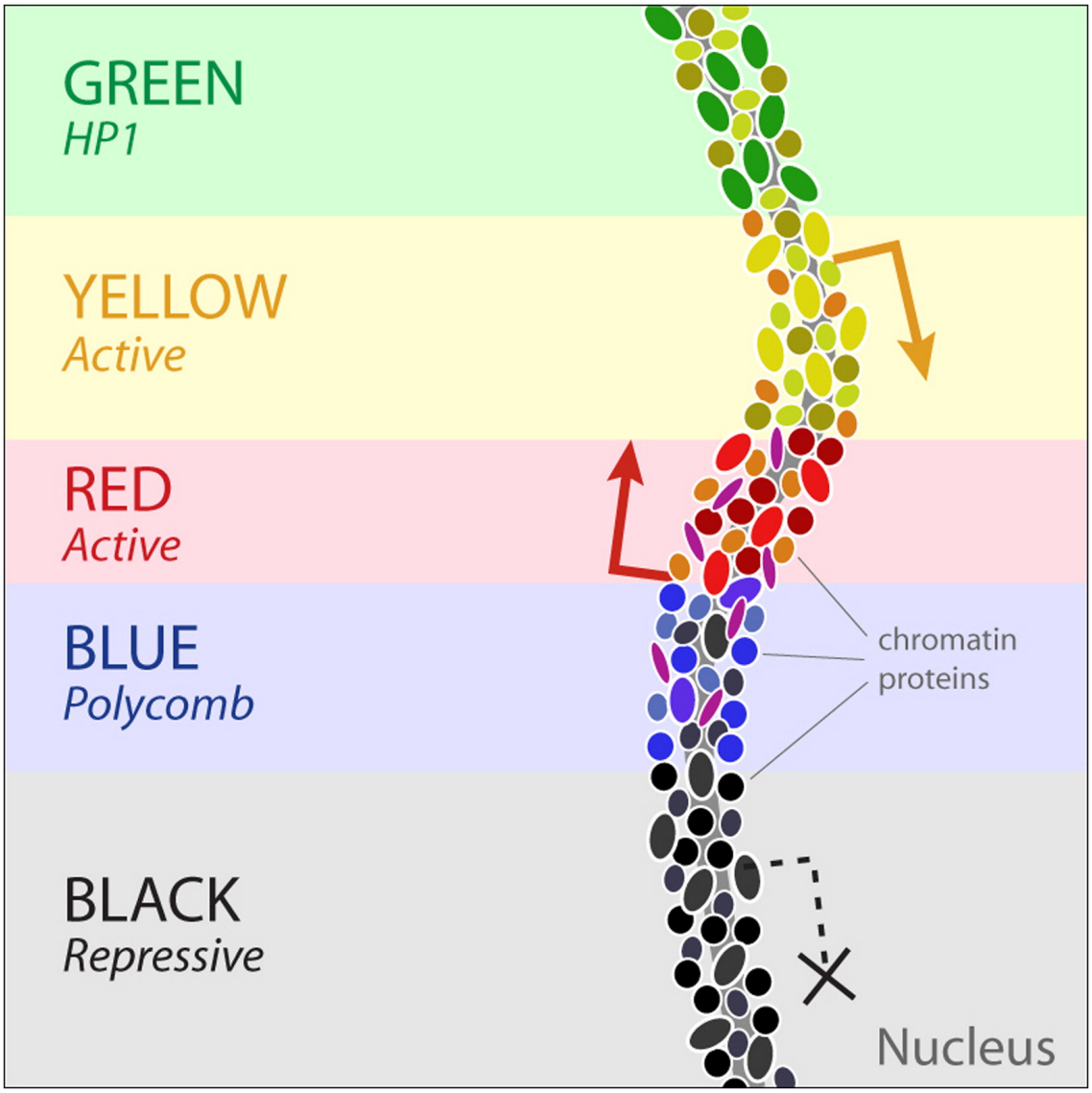


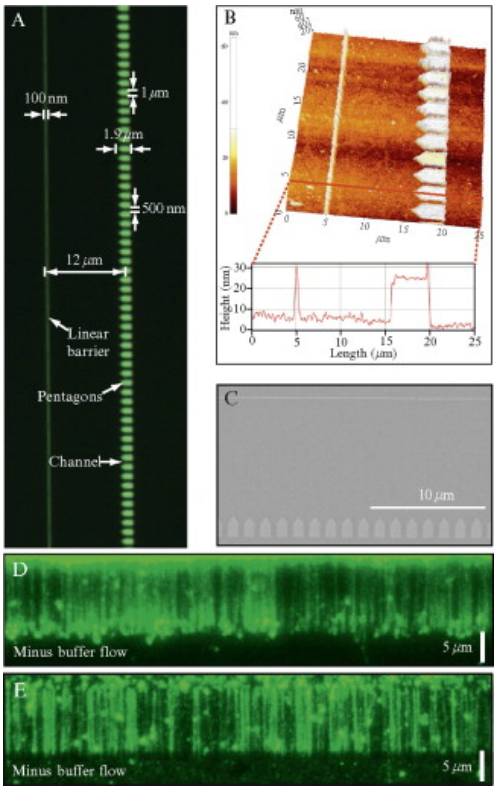
FIGURE 11.1 : *Les 5 types chromatiniens chez la drosophile (Filion* et al.*, 2010)*

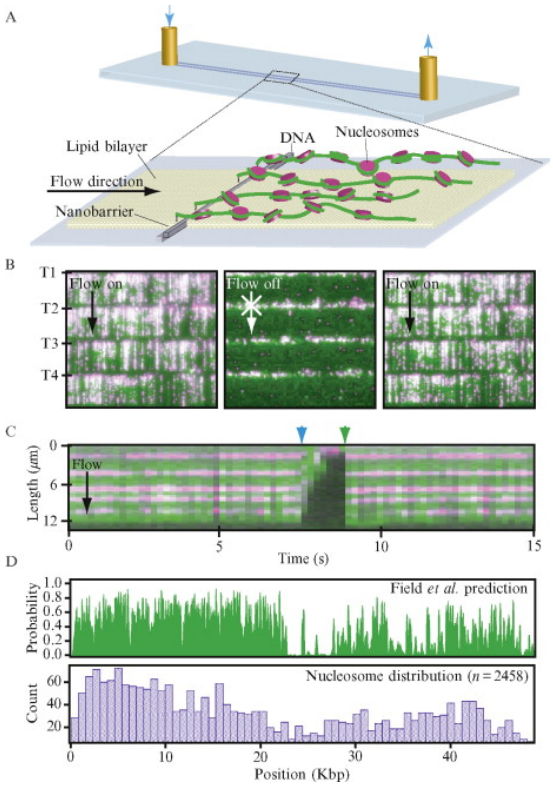
11.3 VOIR L'HÉTÉROCHROMATINE

Les méthodes biochimiques ont l'avantage des statistiques et de la résolution. Elles ont le désavantage des artefacts pouvant être associés à l'extraction (par exempel MNase) et à la cartographie (par ex., PCR, hybridation, séquencage). Par ailleurs , ces cartographies correspondent à une moyenne sur un ensemble de chromatines extraites de cellules souvent non synchronisées et de types cellulaires différents, bien que désormais les cartes sont extraites dans des types bien définis notamment chez l'homme (Valouev et al., 2011) ou à des stades bien définis du développement (Zhang et al., 2011). Même dans ce type d'expérience sur un jeu de cellules "homogènes" on a simplement accès à des distribution "moyennes" ; or on aimerait pouvoir tester nos hypothèses et nos modèles à des comportements "individuels", à savoir à des configurations de chromatines individuelles qui seules, si on est capable d'en mesurer suffisamment, peuvent nous renseigner à la fois sur la variabilité d'un trait chromatinien (periode nucleosomale, marque épigénétique...) au sein d'une population synchronisée ou non , de même type cellulaire ou non mais aussi sur son degré de coopérativité le long des fibres (persistence dans la distribution d'une proteine ou tout marque chromatinienne). En somme caractériser la dynamique de la chromatine par des expériences de molécules uniques, méthodes optiques de préférence. C'est d'ailleurs le travail que nous avons déjà mis en place concernant l'étude du positionnement *in vitro* des nucléosomes par imagerie AFM (Chapitre 7).

11.3.1 Le rideau de chromatine

Récemment le groupe de Greene a mis au point un dispositif experimental permettant la mesure de localisation de protéines marquées le long de molécules étirées : nucléosomes reconstitués sur l'ADN du phage λ, protéines diffusant le long de la chaîne... La résolution est encore assez faible mais nous projetons, en collaboration avec Aurelien Bancaud du LAAS (Toulouse) d'améliorer le dispositif et introduire des systèmes de super-résolution optique. Nous comptons également couplé le dispositif avec la technique AFM, seule méthode actuelle permettant d'avoir une mesure assez résolue des nucléosomes (Chapitre 7). C'est un projet que je compte personnellement développer au laboratoire en collaboration avec les expérimentauers bilogiste et physiciens. Une ANR (Projet Blanc, début 2011, non acceptée) dont j'étais le coordinateur et un projet "soutien à la prise de risque" (Interface Chimie-Physique-Biologie) a été déposé et des fonds de recherche de l'ENS ont été attribué pour débuter le projet. Dans le projet actuel on a l'intention de mieux comprendre les processus d'inactivation des copies rDNA, et ce en relation avec le projet géneral sur la dynamique d'heterochromatinisation (voir plus haut) ; il s'agit d'identifier les contributions génomiques et épigénétiques par retours successifs entre observations et modélisation. Expérimentalement il s'agira d'extraire des fibres de chromatine native, de les peigner et d'extraire les cartes chromtinienens de fibres individuelles. La tâche est ambitieuse et requierera au moins un postdoctorant à plein temps sur ce projet. On peut cependant changer de système (j'avais pensé aussi au site de "mating" chez *S. Pombe*) et s'associer avec un laboratoire plus aguérri à ce sujet (extraction de fibres natives...). Une première tâche pourrait être d'extraire les carte de méthylation de l'ADN sur les chaines ADN natives mais nues (si on considère un système mammifère).





Left :

FIGURE 11.2 : Nanofabricated rack patterns for making "double-tethered" DNA curtains. *(A) shows an optical image of a single barrier set and relevant pattern dimensions are indicated. An AFM image rack pattern is shown in (B) highlighting the height of the linear barriers and the pentagons as well as the distance between the barrier elements. An SEM image of a rack pattern is shown in (C). Examples of a "double-tethered" DNA curtain stained with YOYO1 are shown in (D) and (E). The "double-tethered" curtain shown in (D) was made by ebeam lithography, and the curtain shown in (E) was made by nanoimprint lithography. Adapted from Gorman* et al., Langmuir **26**, *1372-79 (2010).*

Right :

FIGURE 11.3 : Visualizing fluorescently tagged nucleosomes. *(A) depicts the experimental design. YOYO1-stained DNA curtains (green) bound by nucleosomes (magenta) that were tagged with anti-FLAG QDs are shown in (B). The tethered end of each curtain is indicated as T1-T4, and arrows indicate the direction of flow. A kymogram illustrating five nucleosomes on one DNA molecule is shown in (C). The nucleosomes disappear when flow is temporarily interrupted (blue arrowheads) and reappear when flow is resumed (green arrowheads), verifying that they are bound to the DNA and do not interact with the lipid bilayer. The top panel in (D) shows the theoretical distribution of nucleosomes on λ-DNA as predicted by Field* et al. *Field* et al. *(2008b). The lower panel shows the observed distribution of nucleosomes obtained from DNA curtain experiments. Adapted from Visnapuu and Greene,* Nat. Struct. Mol. Biol. **16**, *1056-1062 (2009).*

BIBLIOGRAPHIE

Albert, I., Mavrich, T. N., Tomsho, L. P., Qi, J., Zanton, S. J., Schuster, S. C. and Pugh, B. F. (2007). Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature *446*, 572–576.

Allemand, J. F., Bensimon, D., Lavery, R. and Croquette, V. (1998). Stretched and overwound DNA forms a Pauling-like structure with exposed bases. Proc. Natl. Acad. Sci. USA *95*, 14152–14157.

Anselmi, C., Bocchinfuso, G., Santis, P. D., Savino, M. and Scipioni, A. (2000). A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. Biophys J *79*, 601–613.

Anselmi, C., DeSantis, P. and Scipioni, A. (2005). Nanoscale mechanical and dynamical properties of DNA single molecules. Biophys. Chem. *113*, 209–221.

Arneodo, A., Audit, B., Faivre-Moskalenko, C., Moukhtar, J., Vaillant, C., Argoul, F., d'Aubenton-Carafa, Y. and Thermes, C. (2008). From DNA sequence to chromatin organization : the fundamental role of genomic long-range correlations. In Mémoire de la Classe des Sciences, Collection en -8^{o}, 3^{e} série, Tome XXVIII, n^{o} 2049. Académie Royale de Belgique Bruxelles.

Audit, B. (1999). Analyse statistique des séquences d'ADN par l'intermédiaire de la transformée en ondelettes. PhD thesis, Université de Paris VI Pierre et Marie Curie.

Bancaud, A., Wagner, G., Silva, N. C. E., Lavelle, C., Wong, H., Mozziconacci, J., Barbi, M., Sivolob, A., Cam, E. L., Mouawad, L., Viovy, J.-L., Victor, J.-M. and Prunell, A. (2007). Nucleosome chiral transition under positive torsional stress in single chromatin fibers. Mol Cell *27*, 135–147.

Bao, Y., White, C. L. and Luger, K. (2006). Nucleosome core particles containing a poly(dA.dT) sequence element exhibit a locally distorted DNA structure. J Mol Biol *361*, 617–624.

Becker, N. B. and Everaers, R. (2009). DNA nanomechanics in the nucleosome. Structure *17*, 579–589.

Bednar, J., Furrer, P., Katritch, V., Stasiak, A. Z., Dubochet, J. and Stasiak, A. (1995). Determination of DNA persistence length by cryo-electron microscopy : Separation of the static and dynamic contributions to the apparent persistence length of DNA. J. Mol. Biol. *254*, 579–594.

Bednar, J., Horowitz, R. A., Grigoryev, S. A., Carruthers, L. M., Hansen, J. C., Koster, A. J. and Woodcock, C. L. (1998). Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. Proc Natl Acad Sci U S A *95*, 14173–14178.

Bennink, M. L., Leuba, S. H., Leno, G. H., Zlatanova, J., de Grooth, B. G. and Greve, J. (2001). Unfolding individual nucleosomes by stretching single chromatin fibers with optical tweezers. Nat Struct Biol *8*, 606–610.

Bina, M. (1994). Periodicity of dinucleotides in nucleosomes derived from simian virus 40 chromatin. J. Mol. Biol. *235*, 198–208.

Blank, T. A. and Becker, P. B. (1995). Electrostatic mechanism of nucleosome spacing. J Mol Biol *252*, 305–313.

Boeger, H., Griesenbeck, J. and Kornberg, R. D. (2008). Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. Cell *133*, 716–726.

Bolshoy, A., McNamara, P., Harrington, R. E. and Trifonov, E. N. (1991). Curved DNA without A-A : experimental estimation of all 16 DNA wedge angles. Proc Natl Acad Sci U S A *88*, 2312–2316.

Bouchiat, C. and Mezard, M. (1998). Elasticity model of a supercoiled DNA molecule. Phys. Rev. Lett. *80*, 1556–1559.

Bouchiat, C. and Mezard, M. (2000). Elastic rod model of a supercoiled DNA molecule. Eur. Phys. J. E *2*, 377–402.

Bouchiat, C., Wang, M. D., Allemand, J., Strick, T., Block, S. M. and Croquette, V. (1999). Estimating the persistence length of a worm-like chain molecule from force-extension measurements. Biophys. J. *76*, 409–413.

Bustamante, C., Marko, J. F., Siggia, E. D. and Smith, S. (1994). Entropic elasticity of lambda-phage DNA. Science *265*, 1599–1600.

Chevereau, G. (2010). Thermodynamique du positionnement des nucléosomes. PhD thesis, Université de Lyon-Ecole Normale Supérieure.

Chevereau, G., Palmeira, L., Thermes, C., Arneodo, A. and Vaillant, C. (2009). Thermodynamics of intragenic nucleosome ordering. Phys Rev Lett *103*, 188103.

Chung, H.-R., Dunkel, I., Heise, F., Linke, C., Krobitsch, S., Ehrenhofer-Murray, A. E., Sperling, S. R. and Vingron, M. (2010). The effect of micrococcal nuclease digestion on nucleosome positioning data. PLoS One *5*, e15754.

Cioffi, A., Fleury, T. J. and Stein, A. (2006). Aspects of large-scale chromatin structures in mouse liver nuclei can be predicted from the DNA sequence. Nucleic Acids Res *34*, 1974–1981.

Claudet, C., Angelov, D., Bouvet, P., Dimitrov, S. and Bednar, J. (2005). Histone octamer instability under single molecule experiment conditions. J Biol Chem *280*, 19958–19965.

Cluzel, P., Lebrun, A., Heller, C., Lavery, R., Viovy, J. L., Chatenay, D. and Caron, F. (1996). DNA : an extensible molecule. Science *271*, 792–794.

Cognet, J. A., Pakleza, C., Cherny, D., Delain, E. and Le Cam, E. (1999). Static curvature and flexibility measurements of DNA with microscopy. A simple renormalization method, its assessment by experiment and simulation. J. Mol. Biol. *285*, 997–1009.

Cohanim, A. B., Kashi, Y. and Trifonov, E. N. (2005). Yeast nucleosome DNA pattern : deconvolution from genome sequences of S. cerevisiae. J Biomol Struct Dyn *22*, 687–694.

Cohanim, A. B., Kashi, Y. and Trifonov, E. N. (2006). Three sequence rules for chromatin. J Biomol Struct Dyn *23*, 559–566.

Crothers, D. M. (1998). DNA curvature and deformation in protein-DNA complexes : a step in the right direction. Proc Natl Acad Sci U S A *95*, 15163–15165.

Davis, H. T. (1990). Density distribution functions of confined Tonks-Takahashi fluids. Journal of Chemical Physics *93*, 4339–4344.

Davis, N. A., Majee, S. S. and Kahn, J. D. (1999). TATA box DNA deformation with and without the TATA box-binding protein. J Mol Biol *291*, 249–265.

Dion, M. F., Kaplan, T., Kim, M., Buratowski, S., Friedman, N. and Rando, O. J. (2007). Dynamics of replication-independent histone turnover in budding yeast. Science *315*, 1405–1408.

Drew, H. R. and Travers, A. A. (1985). DNA bending and its relation to nucleosome positioning. J Mol Biol *186*, 773–790.

Fan, X., Moqtaderi, Z., Jin, Y., Zhang, Y., Liu, X. S. and Struhl, K. (2010). Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. Proc Natl Acad Sci U S A *107*, 17945–17950.

Field, Y., Fondufe-Mittendorf, Y., Moore, I. K., Mieczkowski, P., Kaplan, N., Lubling, Y., Lieb, J. D., Widom, J. and Segal, E. (2009). Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. Nat Genet *41*, 438–445.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., Widom, J. and Segal, E. (2008a). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol *4*, e1000216.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., Widom, J. and Segal, E. (2008b). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput. Biol. *4*, e1000216.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J. and van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell *143*, 212–224.

Fu, Y., Sinha, M., Peterson, C. L. and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet *4*, e1000138.

Furrer, P., Bednar, J., Stasiak, A. Z., Katritch, V., Michoud, D., Stasiak, A. and Dubochet, J. (1997). Opposite effect of counterions on the persistence length of nicked and non-nicked DNA. J. Mol. Biol. *266*, 711–721.

Gartenberg, M. R. and Crothers, D. M. (1988). DNA sequence determinants of CAP-induced bending and protein binding affinity. Nature *333*, 824–829.

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

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

Hansma, H. G., Revenko, I., Kim, K. and Laney, D. E. (1996). Atomic force microscopy of long and short double-stranded, single-stranded and triple-stranded nucleic acids. Nucleic Acids Res. *24*, 713–720.

Hartley, P. D. and Madhani, H. D. (2009). Mechanisms that specify promoter nucleosome location and identity. Cell *137*, 445–458.

Huda, A., Mariño-Ramírez, L., Landsman, D. and Jordan, I. K. (2009). Repetitive DNA elements, nucleosome binding and human gene expression. Gene *436*, 12–22.

Ibsen, J., Cordero, P. and Tabensky, R. (1997). Hard rods in the presence of a uniform external field. J. Chem. Phys. *107*, 5515.

Ioshikhes, I. P., Albert, I., Zanton, S. J. and Pugh, B. F. (2006). Nucleosome positions predicted through comparative genomics. Nat Genet *38*, 1210–1215.

Ivanov, V. I., Minchenkova, L. E., Chernov, B. K., McPhie, P., Ryu, S., Garges, S., Barber, A. M., Zhurkin, V. B. and Adhya, S. (1995). CRP-DNA complexes : inducing the A-like form in the binding sites with an extended central spacer. J Mol Biol *245*, 228–240.

Iyer, V. and Struhl, K. (1995). Poly(dA :dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J *14*, 2570–2579.

Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J. and Dickerson, R. E. (1996). How proteins recognize the TATA box. J Mol Biol *261*, 239–254.

Kahn, J. D. and Crothers, D. M. (1992). Protein-induced bending and DNA cyclization. Proc Natl Acad Sci U S A *89*, 6343–6347.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J. and Segal, E. (2009a). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J. and Segal, E. (2009b). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254.

Kenigsberg, E., Bar, A., Segal, E. and Tanay, A. (2010). Widespread compensatory evolution conserves DNA-encoded nucleosome organization in yeast. PLoS Comput Biol *6*, e1001039.

Kepper, N., Foethke, D., Stehr, R., Wedemann, G. and Rippe, K. (2008). Nucleosome geometry and internucleosomal interactions control the chromatin fiber conformation. Biophys J *95*, 3692–3705.

Kim, H. P., Imbert, J. and Leonard, W. J. (2006). Both integrated and differential regulation of components of the IL-2/IL-2 receptor system. Cytokine Growth Factor Rev *17*, 349–366.

Kim, Y., Geiger, J. H., Hahn, S. and Sigler, P. B. (1993). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512–520.

Kornberg, R. D. and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell *98*, 285–294.

Kratky, O. and Porod, G. (1949). X-Ray investigation of dissolved chain molecules. Recueil : J. Roy. Netherlands Chem. Soc. *68*, 1106–1123.

Kruithof, M., Chien, F.-T., Routh, A., Logie, C., Rhodes, D. and van Noort, J. (2009). Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber. Nat Struct Mol Biol *16*, 534–540.

Ladoux, B., Quivy, J. P., Doyle, P., du Roure, O., Almouzni, G. and Viovy, J. L. (2000). Fast kinetics of chromatin assembly revealed by single-molecule videomicroscopy and scanning force microscopy. Proc Natl Acad Sci U S A *97*, 14251–14256.

Lam, F. H., Steger, D. J. and O'Shea, E. K. (2008). Chromatin decouples promoter threshold from dynamic range. Nature *453*, 246–250.

Lankas, F., Sponer, J., Langowski, J. and Cheatham, T. E. (2003). DNA basepair step deformability inferred from molecular dynamics simulations. Biophys J *85*, 2872–2883.

Lantermann, A. B., Straub, T., Strålfors, A., Yuan, G.-C., Ekwall, K. and Korber, P. (2010). Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. Nat Struct Mol Biol *17*, 251–257.

Lavelle, C. and Prunell, A. (2007). Chromatin polymorphism and the nucleosome superfamily : a genealogy. Cell Cycle *6*, 2113–2119.

Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., Hughes, T. R. and Nislow, C. (2007a). A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet *39*, 1235–1244.

Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., Hughes, T. R. and Nislow, C. (2007b). A high-resolution atlas of nucleosome occupancy in yeast. Nat. Genet. *39*, 1235–1244.

Léger, J. F., Romano, G., Sarkar, A., Robert, J., Bourdieu, L., Chatenay, D. and Marko, J. F. (1999). Structural transitions of a twisted and stretched DNA molecule. Phys. Rev. Lett. *83*, 1066–1069.

Lesbats, P., Botbol, Y., Chevereau, G., Vaillant, C., Calmels, C., Arneodo, A., Andreola, M.-L., Lavigne, M. and Parissi, V. (2011). Functional coupling between HIV-1 integrase and the SWI/SNF chromatin remodeling complex for efficient in vitro integration into stable nucleosomes. PLoS Pathog *7*, e1001280.

Lesne, A. and Victor, J.-M. (2006). Chromatin fiber functional organization : some plausible models. Eur Phys J E Soft Matter *19*, 279–290.

Li, B., Carey, M. and Workman, J. L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Lieb, E. and Mattis, D. (1966). Mathematical Physics in One Dimension. Academic Press.

Lowary, P. T. and Widom, J. (1997). Nucleosome packaging and nucleosome positioning of genomic DNA. Proc Natl Acad Sci U S A *94*, 1183–1188.

Lowary, P. T. and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J Mol Biol *276*, 19–42.

Luger, K. and Richmond, T. J. (1998). DNA binding within the nucleosome core. Curr Opin Struct Biol *8*, 33–40.

Lyubchenko, Y., Shlyakhtenko, L., Harrington, R., Oden, P. and Lindsay, S. (1993). Atomic force microscopy of long DNA : imaging in air and under water. Proc. Natl. Acad. Sci. USA *90*, 2137–2140.

Marilley, M., Sanchez-Sevilla, A. and Rocca-Serra, J. (2005). Fine mapping of inherent flexibility variation along DNA molecules : validation by atomic force microscopy (AFM) in buffer. Mol. Genet. Genomics *274*, 658–670.

Marko, J. F. and Siggia, E. D. (1995a). Stretching DNA. Macromolecules *28*, 8759–8770.

Marko, J. F. and Siggia, E. D. (1995b). Statistical mechanics of supercoiled DNA. Phys. Rev. E *52*, 2912–2938.

Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., Qi, J., Schuster, S. C., Albert, I. and Pugh, B. F. (2008a). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res *18*, 1073–1083.

Mavrich, T. N., Jiang, C., Ioshikhes, I. P., Li, X., Venters, B. J., Zanton, S. J., Tomsho, L. P., Qi, J., Glaser, R. L., Schuster, S. C., Gilmour, D. S., Albert, I. and Pugh, B. F. (2008b). Nucleosome organization in the Drosophila genome. Nature *453*, 358–362.

Meister, P., Towbin, B. D., Pike, B. L., Ponti, A. and Gasser, S. M. (2010). The spatial dynamics of tissue-specific promoters during C. elegans development. Genes Dev *24*, 766–782.

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

Miele, V., Vaillant, C., d'Aubenton Carafa, Y., Thermes, C. and Grange, T. (2008). DNA physical properties determine nucleosome occupancy from yeast to fly. Nucleic Acids Res *36*, 3746–3756.

Mihardja, S., Spakowitz, A. J., Zhang, Y. and Bustamante, C. (2006). Effect of force on mononucleosomal dynamics. Proc Natl Acad Sci U S A *103*, 15871–15876.

Milani, P. (2007). Caracteristiques structurales et dynamique du promoteur du gene IL2RA. Etude par modelisation et par microscopie a force atomique. PhD thesis, Faculte de Medecine, Marseille.

Milani, P., Chevereau, G., Vaillant, C., Audit, B., Haftek-Terreau, Z., Marilley, M., Bouvet, P., Argoul, F. and Arneodo, A. (2009). Nucleosome positioning by genomic excluding-energy barriers. Proc Natl Acad Sci U S A *106*, 22257–22262.

Milani, P., Marilley, M., Sanchez-Sevilla, A., Imbert, J., Vaillant, C., Argoul, F., Egly, J.-M., Rocca-Serra, J. and Arneodo, A. (2011). Mechanics of the IL2RA gene activation revealed by modeling and atomic force microscopy. PLoS One *6*, e18811.

Mito, Y., Henikoff, J. G. and Henikoff, S. (2005). Genome-scale profiling of histone H3.3 replacement patterns. Nat Genet *37*, 1090–1097.

Mito, Y., Henikoff, J. G. and Henikoff, S. (2007). Histone replacement marks the boundaries of cis-regulatory domains. Science *315*, 1408–1411.

Montel, F. (2008). Dynamique à l'équilibre et hors-équilibre de la chromatine visualisée par Microscopie à Force Atomique : effet des variants d'histone et des facteurs de remodelage. PhD thesis, Ens-Lyon.

Moroz, J. D. and Nelson, P. (1997). Torsional directed walks, entropic elasticity, and DNA twist stiffness. Proc. Natl. Acad. Sci. USA *94*, 14418–14422.

Morozov, A. V., Fortney, K., Gaykalova, D. A., Studitsky, V. M., Widom, J. and Siggia, E. D. (2009). Using DNA mechanics to predict in vitro nucleosome positions and formation energies. Nucleic Acids Res *37*, 4707–4722.

Morse, R. H. (2007a). Transcription factor access to promoter elements. J Cell Biochem *102*, 560–570.

Morse, R. H. (2007b). Transcription factor access to promoter elements. J. Cell. Biochem. *102*, 560–570.

Moukhtar, J. (2008). Effet de séquence sur les propriétés thermodynamiques de brins d'ADN : de la théorie à l'expérience. PhD thesis, Ecole Normale Supérieure de Lyon.

Moukhtar, J., Faivre-Moskalenko, C., Milani, P., Audit, B., Vaillant, C., Fontaine, E., Mongelard, F., Lavorel, G., St-Jean, P., Bouvet, P., Argoul, F. and Arneodo, A. (2010). Effect of genomic long-range correlations on DNA persistence Length : From theory to single molecules experiments. J. Phys. Chem. B *114*, 5125–5143.

Moukhtar, J., Fontaine, E., Faivre-Moskalenko, C. and Arneodo, A. (2007). Probing persistence in DNA curvature properties with atomic force microscopy. Phys. Rev. Lett. *98*, 178101.

Moukhtar, J., Vaillant, C., Audit, B. and Arneodo, A. (2009). Generalized wormlike chain model for long-range correlated heteropolymers. Europhys. Lett. *86*, 48001.

Nakagawa, T., Bulger, M., Muramatsu, M. and Ito, T. (2001). Multistep chromatin assembly on supercoiled plasmid DNA by nucleosome assembly protein-1 and ATP-utilizing chromatin assembly and remodeling factor. J Biol Chem *276*, 27384–27391.

Nelson, P. (1998). Sequence-disorder effects on DNA entropic elasticity. Phys. Rev. Lett. *80*, 5810–5812.

Neukirch, S. (2004). Extracting DNA twist rigidity from experimental supercoiling data. Phys Rev Lett *93*, 198107.

Nissen, M. S. and Reeves, R. (1995). Changes in superhelicity are introduced into closed circular DNA by binding of high mobility group protein I/Y. J Biol Chem *270*, 4355–4360.

Noll, M. and Kornberg, R. D. (1977). Action of micrococcal nuclease on chromatin and the location of histone H1. J Mol Biol *109*, 393–404.

Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. and Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A *95*, 11163–11168.

Paillard, G. and Lavery, R. (2004). Analyzing protein-DNA recognition mechanisms. Structure *12*, 113–122.

Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., Struhl, K. and Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. Genome Res *17*, 1170–1177.

Percus, J. K. (1976). Equilibrium state of a classical fluid of hard rods in an external field. Journal of Statistical Physics *15*, 505–511.

Percus, J. K. (1982). One-dimensional classical fluid with nearest-neighbor interaction in arbitrary external field. Journal of Statistical Physics *28*, ?

Perkins, T. T., Quake, S. R., Smith, D. E. and Chu, S. (1994). Relaxation of a single DNA molecule observed by optical microscopy. Science *264*, 822–826.

Rando, O. J. and Ahmad, K. (2007). Rules and regulation in the primary structure of chromatin. Curr Opin Cell Biol *19*, 250–256.

Rayleigh, L. (1891). On the Virial of a System of Hard Colliding Bodies. Nature London *45*, 80–82.

Reeves, R., Leonard, W. J. and Nissen, M. S. (2000). Binding of HMG-I(Y) imparts architectural specificity to a positioned nucleosome on the promoter of the human interleukin-2 receptor alpha gene. Mol Cell Biol *20*, 4666–4679.

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

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

Rivetti, C., Guthold, M. and Bustamante, C. (1996). Scanning force microscopy of DNA deposited onto mica : equilibration versus kinetic trapping studied by statistical polymer chain analysis. J. Mol. Biol. *264*, 919–932.

Rivetti, C., Walker, C. and Bustamante, C. (1998). Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. J. Mol. Biol. *280*, 41–59.

Robledo, A. and Rowlinson, J. (1986). The distribution of hard rods on a line of finite length. Molecular Physics *58*, 711–721.

Roudier, F., Ahmed, I., BÃľrard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., DesprÃľs, B., Drevensek, S., Barneche, F., DÃĺrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., Martin-Magniette, M.-L., Robin, S., Caboche, M. and Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J *30*, 1928–1938.

Salsburg, Z. W., Kirkwood, J. and Zwanzig, R. (1953). Molecular Distribution Functions in a One-Dimensional Fluid. J. Chem. Phys *21*, 1098.

Satchwell, S. C., Drew, H. R. and Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. J Mol Biol *191*, 659–675.

Schellman, J. A. (1974). Flexibility of DNA. Biopolymers *13*, 217–226.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G. and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Scipioni, A., Anselmi, C., Zuccheri, G., Samori, B. and De Santis, P. (2002a). Sequence-dependent DNA curvature and flexibility from scanning force microscopy images. Biophys. J. *83*, 2408–2418.

Scipioni, A., Zuccheri, G., Anselmi, C., Bergia, A., Samor ?, B. and Santis, P. D. (2002b). Sequence-dependent DNA dynamics by scanning force microscopy time-resolved imaging. Chem Biol *9*, 1315–1321.

Scipioni, A., Zuccheri, G., Anselmi, C., Bergia, A., Samorì, B. and De Santis, P. (2002c). Sequence-dependent DNA dynamics by scanning force microscopy time-resolved imaging. Chem. Biol. *9*, 1315–1321.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z. and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Segal, E. and Widom, J. (2009). What controls nucleosome positions ? Trends Genet *25*, 335–343.

Sekinger, E. A., Moqtaderi, Z. and Struhl, K. (2005). Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. Mol Cell *18*, 735–748.

Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M. and Iyer, V. R. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol *6*, e65.

Shore, D., Langowski, J. and Baldwin, R. L. (1981). DNA flexibility studied by covalent closure of short fragments into circles. Proc Natl Acad Sci U S A *78*, 4833–4837.

Shrader, T. E. and Crothers, D. M. (1989). Artificial nucleosome positioning sequences. Proc. Natl. Acad. Sci. USA *86*, 7418–7422.

Shundrovsky, A., Smith, C. L., Lis, J. T., Peterson, C. L. and Wang, M. D. (2006). Probing SWI/SNF remodeling of the nucleosome by unzipping single DNA molecules. Nat Struct Mol Biol *13*, 549–554.

Smith, S. B., Cui, Y. and Bustamante, C. (1996). Overstretching B-DNA : the elastic response of individual double-stranded and single-stranded DNA molecules. Science *271*, 795–799.

Smith, S. B., Finzi, L. and Bustamante, C. (1992). Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. Science *258*, 1122–1126.

Solis, F. J., Bash, R., Wang, H., Yodh, J., Lindsay, S. A. and Lohr, D. (2007). Properties of nucleosomes in acetylated mouse mammary tumor virus versus 5S arrays. Biochemistry *46*, 5623–5634.

Solis, F. J., Bash, R., Yodh, J., Lindsay, S. M. and Lohr, D. (2004). A statistical thermodynamic model applied to experimental AFM population and location data is able to quantify DNA-histone binding strength and internucleosomal interaction differences between acetylated and unacetylated nucleosomal arrays. Biophys J *87*, 3372–3387.

Song, L. and Schurr, J. M. (1990). Dynamic bending rigidity of DNA. Biopolymers *30*, 229–237.

Steinfeld, I., Shamir, R. and Kupiec, M. (2007). A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription. Nat Genet *39*, 303–309.

Steinmetz, E. J., Warren, C. L., Kuehner, J. N., Panbehi, B., Ansari, A. Z. and Brow, D. A. (2006). Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. Mol Cell *24*, 735–746.

Strick, T. R., Allemand, J. F., Bensimon, D., Bensimon, A. and Croquette, V. (1996). The elasticity of a single supercoiled DNA molecules. Science *271*, 1835–1837.

Suter, B., Schnappauf, G. and Thoma, F. (2000). Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res *28*, 4083–4089.

Teif, V. B. and Rippe, K. (2009). Predicting nucleosome positions on the DNA : combining intrinsic sequence preferences and remodeler activities. Nucleic Acids Res *37*, 5641–5655.

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

Thåström, A., Lowary, P. T., Widlund, H. R., Cao, H., Kubista, M. and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. J Mol Biol *288*, 213–229.

Tillo, D. and Hughes, T. R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442.

Tirosh, I. and Barkai, N. (2008a). Two strategies for gene regulation by promoter nucleosomes. Genome Res *18*, 1084–1091.

Tirosh, I. and Barkai, N. (2008b). Two strategies for gene regulation by promoter nucleosomes. Genome Res. *18*, 1084–1091.

Tirosh, I., Sigal, N. and Barkai, N. (2010). Widespread remodeling of mid-coding sequence nucleosomes by Isw1. Genome Biol *11*, R49.

Tolkunov, D. and Morozov, A. V. (2009). Nucleosome positioning and energetics : Recent advances in genomic and computational studies. arXiv :0912.3954 *?*, ?

Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K. and Zhurkin, V. B. (2007). A novel roll-and-slide mechanism of DNA folding in chromatin : implications for nucleosome positioning. J Mol Biol *371*, 725–738.

Travers, A. A. (1991). To bend or... ? Curr Biol *1*, 171–173.

Travers, A. A. and Klug, A. (1987). The bending of DNA in nucleosomes and its wider implications. Philos Trans R Soc Lond B Biol Sci *317*, 537–561.

Trifonov, E. N. and Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci U S A *77*, 3816–3820.

Trifonov, E. N., Tan, R. K. Z. and Harvey, S. C. (1987). In DNA bending curvature, (Olson, W. K., Sarma, M. H. and Sundaralingam, M., eds), p. 243, Academic Press, Schenectady.

Tsankov, A. M., Thompson, D. A., Socha, A., Regev, A. and Rando, O. J. (2010). The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol *8*, e1000414.

Vaillant, C. (2001). Influence de la séquence sur les propriétés élastiques des chaînes ADN. PhD thesis, Université Paris VI, Pierre et Marie Curie.

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

Vaillant, C., Audit, B. and Arneodo, A. (2007). Experiments confirm the influence of genome long-range correlations on nucleosome positioning. Phys Rev Lett *99*, 218103.

Vaillant, C., Audit, B., Thermes, C. and Arneodo, A. (2003). Influence of the sequence on the elastic properties of long DNA chains. Phys. Rev. E *67*, 032901.

Vaillant, C., Audit, B., Thermes, C. and Arneodo, A. (2006). Formation and positioning of nucleosomes : effect of sequence-dependent long-range correlated structural disorder. Eur. Phys. J. E *19*, 263–277.

Vaillant, C., Audit, B., Thermes, C. and Arnéodo, A. (2003). Influence of the sequence on elastic properties of long DNA chains. Phys Rev E Stat Nonlin Soft Matter Phys *67*, 032901.

Vaillant, C., Palmeira, L., Chevereau, G., Audit, B., d'Aubenton Carafa, Y., Thermes, C. and Arneodo, A. (2010). A novel strategy of transcription regulation by intragenic nucleosome ordering. Genome Res *20*, 59–67.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K., Sidow, A., Fire, A. and Johnson, S. M. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res *18*, 1051–1063.

Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z. and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *00*.

Vanderlick, Scriven and Davis (1986). Solution of Percus's equation for the density of hard rods in an external field. Phys Rev A *34*, 5130–5131.

Venters, B. J., Wachi, S., Mavrich, T. N., Andersen, B. E., Jena, P., Sinnamon, A. J., Jain, P., Rolleri, N. S., Jiang, C., Hemeryck-Walsh, C. and Pugh, B. F. (2011). A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. Mol Cell *41*, 480–492.

Virstedt, J., Berge, T., Henderson, R. M., Waring, M. J. and Travers, A. A. (2004). The influence of DNA stiffness upon nucleosome formation. J Struct Biol *148*, 66–85.

Vologodskii, A. V. and Marko, J. F. (1997). Extension of torsionally stressed DNA by external force. Biophys. J. *73*, 123–132.

Vologoskii, A. (1994). DNA extension under the action of an external forces. Macromolecules *27*, 5623–5625.

Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. and Bushman, F. D. (2007). HIV integration site selection : analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res *17*, 1186–1194.

Wang, M. D., Yin, H., Landick, R., Gelles, J. and Block, S. M. (1997). Stretching DNA with optical tweezers. Biophys. J. *72*, 1335–1346.

Wang, X., Bryant, G. O., Floer, M., Spagna, D. and Ptashne, M. (2011). An effect of DNA sequence on nucleosome occupancy and removal. Nat Struct Mol Biol *18*, 507–509.

Washietl, S., Machné, R. and Goldman, N. (2008). Evolutionary footprints of nucleosome positions in yeast. Trends Genet *24*, 583–587.

Weiner, A., Hughes, A., Yassour, M., Rando, O. J. and Friedman, N. (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res *20*, 90–100.

Whitehouse, I., Rando, O. J., Delrow, J. and Tsukiyama, T. (2007). Chromatin remodelling at promoters suppresses antisense transcription. Nature *450*, 1031–1035.

Whitehouse, I. and Tsukiyama, T. (2006). Antagonistic forces that position nucleosomes in vivo. Nat Struct Mol Biol *13*, 633–640.

Woodcock, C. L., Skoultchi, A. I. and Fan, Y. (2006). Role of linker histone in chromatin structure and function : H1 stoichiometry and nucleosome repeat length. Chromosome Res *14*, 17–25.

Yuan, G.-C. and Liu, J. S. (2008a). Genomic sequence is highly predictive of local nucleosome depletion. PLoS Comput Biol *4*, e13.

Yuan, G.-C. and Liu, J. S. (2008b). Genomic sequence is highly predictive of local nucleosome depletion. PLoS Comput. Biol. *4*, e13.

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

Zhang, L., Ma, H. and Pugh, B. F. (2011). Stable and dynamic nucleosome states during a meiotic developmental process. Genome Res *21*, 875–884.

Zhang, Y., Moqtaderi, Z., Rattner, B. P., Euskirchen, G., Snyder, M., Kadonaga, J. T., Liu, X. S. and Struhl, K. (2009). Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol *16*, 847–852.

Zhang, Y., Xi, Z., Hegde, R. S., Shakked, Z. and Crothers, D. M. (2004). Predicting indirect readout effects in protein-DNA interactions. Proc Natl Acad Sci U S A *101*, 8337–8341.

Zhang, Z. and Pugh, B. F. (2011). High-resolution genome-wide mapping of the primary structure of chromatin. Cell *144*, 175–186.

Zuccheri, G., Scipioni, A., Cavaliere, V., Gargiulo, G., De Santis, P. and Samorì, B. (2001). Mapping the intrinsic curvature and flexibility along the DNA chain. Proc. Natl. Acad. Sci. USA *98*, 3074–3079.