Influence of the genomic sequence on the primary structure of chromatin

Guillaume Chevereau, Alain Arnéodo, Cédric Vaillant

Influence of the genomic sequence on the primary structure of chromatin

G. Chevereau[a,b], A. Arneodo[a,b] and C. Vaillant[a,b]

[a] *Université de Lyon, F − 69000 Lyon, France;* [b] *Laboratoire de Physique, CNRS, ENS-Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France*

As an important actor in the regulation of nuclear functions, the nucleosomal organization of the 10 nm chromatin fiber is the subject of increasing interest. Recent high-resolution mapping of nucleosomes along various genomes ranging from yeast to human, have revealed a patchy nucleosome landscape with alternation of depleted, well positioned and fuzzy regions. For many years, the mechanisms that control nucleosome occupancy along eukaryotic chromosomes and their coupling to transcription and replication processes have been under intense experimental and theoretical investigation. A recurrent question is to what extent the genomic sequence dictates and/or constrains nucleosome positioning and dynamics? In that context we have recently developed a simple thermodynamical model that accounts for both sequence specificity of the histone octamer and for nucleosome–nucleosome interactions. As a main issue, our modelling mimics remarkably well *in vitro* data showing that the sequence signaling that prevails are high energy barriers that locally inhibit nucleosome formation and condition the collective positioning of neighboring nucleosomes according to thermal equilibrium statistical ordering. When comparing to *in vivo* data, our physical modelling performs as well as models based on statistical learning suggesting that *in vivo* bulk chromatin is to a large extent controlled by the underlying genomic sequence although it is also subject to finite-range remodelling action of external factors including transcription factors and ATP-dependent chromatin remodellers. On the highly studied *S. cerevisiae* organism, we discuss the implications of the highlighted 'positioning via excluding' mechanism on the structure and function of yeast genes. The generalization of our physical modelling to human is likely to provide new insight on the isochore structure of mammalian genomes in relation with their primary nucleosomal structure.

Keywords: chromatin; nucleosome; DNA sequence; physical model; statistical ordering

Introduction

The relation between the DNA primary structure and its biological function is one of the outstanding problems in modern cell biology. There are many objective reasons to believe that the functional role of the DNA sequence is not only to code for proteins (which represent less than 5% of mammalian genomes), but also to contribute to controlling the spatial structure and dynamics of DNA in chromatin. Nowadays, it is well recognized that the dynamics of DNA folding and unfolding within living cells plays a major role in regulating many biological processes, such as gene expression, DNA replication, recombination and repair (van Holde 1988; Wolffe 1998; Calladine and Drew 1999; Alberts et al. 2002). In that context, a very challenging issue is to decipher to which extent the compromise between the necessity of compacting DNA in the nucleus of eukaryotic cells and the required accessibility to regulatory proteins is actually encoded in the DNA sequence. If the precise influence of the genomic sequence on the different steps of DNA compaction (Alberts et al. 2002), including the nucleosomal array, its condensation into the 30 nm chromatin fiber and the formation of chromatin loops, up to a full extent of condensation in metaphase chromosomes, remains controversial, at a local scale specific elements have been identified to interact with protein components of chromatin. For instance, as far as the basic unit of eukaryotic chromatin is concerned, some motifs that favor the formation and positioning of nucleosomes were found to be regularly spaced, e.g. the 10.5 bp periodicity exhibited in some dinucleotides like AA/TT/TA (Satchwell et al. 1986; Ioshikhes et al. 1996; Widom 2001). But these periodically distributed motifs concern only 5% of sequences that present affinities for the histone octamer significantly larger than average (Lowary and Widom 1997) and cannot account for more than ∼20% of the *in vivo* nucleosome positioning above what is expected by chance (Peckham et al. 2007; Yuan and Liu 2008). Alternatively, similar sequence motifs were shown to present long-range correlations (LRC) along the genome as the signature of the majority of nucleosomes that corresponds to 95% of bulk genomic DNA sequences having an affinity for the histone octamer similar to that of random sequences (Audit et al. 2001, 2002, 2004; Arneodo et al. 2011). Hence, in contrast to the tight histone binding obtained with an adequate

periodic ('deterministic') distribution of bending sites, LRC likely facilitate the formation of some large-scale intrinsic curvature due to a persistent distribution of DNA curvature sites that predisposes DNA to make small (random) loops favoring the positioning of the histone core throughout a major part of the genome (Vaillant et al. 2005, 2006; Moukhtar et al. 2007, 2009, 2010, 2011). Furthermore, the LRC observed at scales larger than the length of DNA wrapped around the histone octamer (>150 bp) were conjectured as contributing to the collective organization and repositioning dynamics of nucleosomes along the 10 nm chromatin fiber and possibly to its packing into higher-order chromatin structures (Audit et al. 2002; Vaillant et al. 2005, 2006). This conjecture was tested and verified as quite relevant when using power spectrum and correlation function analysis to study the nucleosome occupancy landscape obtained in the pioneering *in vivo* experiment of Yuan et al. (2005) on chromosome 3 of *S. cerevisiae*. This study (Vaillant et al. 2007) actually confirmed that the spatial organization of nucleosomes is long-range correlated with characteristics similar to the LRC imprinted in the DNA sequence. Since this pioneering experiment, the recent flowering of genome-wide experimental maps of nucleosome positions for many different organisms and cell types (Fire et al. 2006; Johnson et al. 2006; Albert et al. 2007; Lee et al. 2007; Mito et al. 2007; Ozsolak et al. 2007; Whitehouse et al. 2007; Field et al. 2008; Mavrich et al. 2008a, 2008b; Schones et al. 2008; Shivaswamy et al. 2008; Valouev et al. 2008; Kaplan et al. 2009; Zhang et al. 2009; Lantermann et al. 2010; Tsankov et al. 2010; Weiner et al. 2010; Valouev et al. 2011), has provided an unprecedented opportunity to elucidate to which extent the DNA sequence participates in the positioning of nucleosomes observed *in vivo* along eukaryotic chromosomes.

The paper is organized as follows. In the next section, we review *in vivo* and *in vitro* nucleosome occupancy data in various eukaryotic genomes with special focus on *S. cerevisiae*, *S. pombe*, *C. elegans* and human. The third section is devoted to the definition of our thermodynamical model of nucleosome assembly which is inspired from the well-known thermodynamics of the Tonks–Takahashi 1D fluid (Tonks 1936; Takahashi 1942). As described by the Percus equation (Percus 1976), the hard-rod density (the nucleosome density along genomes) in an inhomogeneous energetic field (the nucleosome potential along genomes) can be determined as a function of the chemical potential (histone octamer reservoir) and the temperature. Various numerical schemes to compute the density are discussed including the technical trick proposed by Vanderlick et al. (1986) to derive an explicit solution of the Percus equation that requires numerical integration. In the fourth section, for pedagogical purposes, we investigate toy energy landscapes involving square-like wells, square-like energy barriers, infinite energy barriers, to illustrate the 'statistical ordering' mechanism originally proposed by Kornberg and Stryer (1988). If an array of well-positioned nucleosomes can be induced by a stretch of regularly distributed potential wells (e.g. a stretch of highly positioning sequences), statistical short-range ordering can also be observed near an energy barrier due to the interplay between boundary confinement and rod–rod (nucleosome–nucleosome) excluded volume interaction. In the fifth section, we develop a realistic sequence-dependent model of nucleosome assembly that relies on the computation of the free-energy cost of bending a DNA fragment of a given nucleotide sequence from its natural curvature to the final superhelical structure around the histone core. When comparing the predictions of this grand-canonical modelling at low chemical potential to *in vitro* nucleosome occupancy data, we show that our physical model performs remarkably well confirming that it accounts for both sequence specificity of the histone octamer and for nucleosome–nucleosome interactions. When tuning the chemical potential to higher value to reproduce genome coverage by nucleosomes observed *in vivo*, we show that the collective nucleosomal organization in the *in vivo* bulk chromatin is to a large extent controlled by the underlying sequence. Interestingly, when some discrepancy is observed between the numerical predictions and the *in vivo* data, it actually provides very instructive information for future modelling of both transcription factor and chromatin remodeller driven 'extrinsic' nucleosome positioning. We conclude, in the final section, by discussing some new modelling perspectives in mammalian genomes.

In vivo and *in vitro* genome-wide primary structure of chromatin

Nucleosome organization is generally analyzed by micrococal nuclease (MNase) digestion of chromatin. To perform large-scale studies, the distribution of MNase cleavage sites is determined throughout genomic regions or in the whole genome by means of either high resolution oligonucleotide tiling microarrays (MNase-chip) (Yuan et al. 2005; Lee et al. 2007; Mito et al. 2007; Ozsolak et al. 2007; Whitehouse et al. 2007; Lantermann et al. 2010) or several different massive DNA sequence technologies (MNase-seq) (Johnson et al. 2006; Albert et al. 2007; Shivaswamy et al. 2008; Valouev et al. 2008; Kaplan et al. 2009; Tsankov et al. 2010; Weiner et al. 2010; Valouev et al. 2011). In this section, we will mainly present nucleosome occupancy profile $P(s)$ as an estimate (up to some normalization) of the probability of a base pair located at position s to be occupied by a nucleosome. As illustrated in Figures 1 to 6, a semi-logarithmic representation $\delta Y(s) = Y(s) - \overline{Y}$, where $Y(s) = \log_2(P(s))$, is generally used to report the experimental MNase-chip and MNase-seq data.

In vivo *nucleosome occupancy profiles*

In Figures 1 to 6 are shown nucleosome occupancy profiles experimentally obtained *in vivo* for several eukaryotic organisms, namely *S. cerevisiae* (Figures 1 and 2),

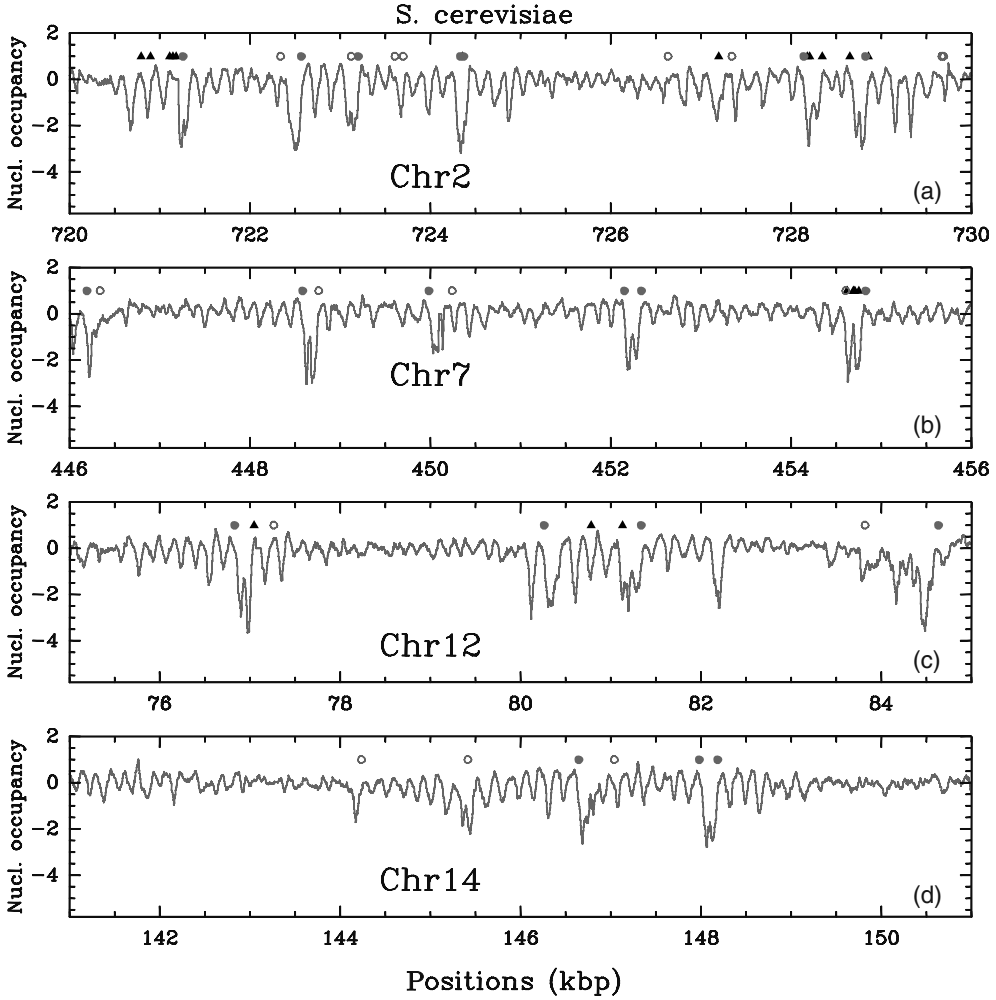

Figure 1. *In vivo* experimental nucleosome mapping along *S. cerevisiae* chromosomes obtained by Lee et al. (2007) (MNase-chip). Nucleosome occupancy profile ($\delta Y(s) = Y(s) - \overline{Y}$, where $Y(s) = \log_2(P(s))$) along a 10 kpb fragment of chromosomes 2 (a), 7 (b), 12 (c) and 14 (d). Symbols indicate regulatory sites: Transcription Start Sites (TSS, red dots), Transcription Termination Sites (TTS, pink circles) and Transcription Factors Binding Sites (TFBS, black triangles).

S. kluyveri (Figure 3), *S. pombe* (Figure 4), *C. elegans* (Figure 5) and human (Figure 6). These experimental data reveal that nucleosome occupancy is globally disordered along the genomes (at chromosomal level). This confirms that the primary structure of chromatin is not a regular array of nucleosomes. As exemplified on the highly studied unicelullar organism *S. cerevisiae* whose genome is highly compact (∼75% gene coverage), the nucleosome occupancy profile (Figure 1) reveals an alternation of (i) nucleosome depleted regions (NDR) of typical size (100–200 bp), (ii) periodic stretches of well-positioned nucleosomes (of mean period 168 bp) and (iii) occupied but unorganized ('fuzzy') regions. As reported in Figure 2, a puzzling and rather annoying observation is that the nucleosome maps obtained from different studies and different methods (MNase-chip or MNase-seq (Zhang and Pugh 2011)) do not generally coincide. Actually they mainly differ by their level of occupancy which might be due to different tiling strategies between MNase-chip approaches (Yuan et al. 2005; Lee et al. 2007) or by the sequencing depth in the MNase-seq approaches. But hopefully the positions of nucleosomes and NDRs are pretty well conserved as observed in *S. cerevisiae* (Figure 2) and *S. pombe* (Figure 4).

At a statistical level, experimental nucleosome occupancy fluctuations can be quantified by their statistical distribution – that characterizes the variability of occupancy along the genome – and their auto-correlation function – that characterizes the regularity of nucleosome occupancy (positioning) along the genome. As shown in Figure 7(a) for the budding yeast microarray data of Lee et al. (2007), the distribution of $\delta Y(s)$ is asymmetric with a large exponential tail at low occupancy values. The distributions obtained with the *S. pombe* microarray data of Lantermann et al. (2010) presents a very similar shape (Figure 7c). Interestingly, the distribution for the mutant, deficient in Mit1 activity, reveals a slight enrichment towards lower occupancy values as compared to the wild type (WT) distribution (Figure 7c) which illustrates the global action of the remodelling factor Mit1 involved in the elimination of nucleosome depletion regions (NDRs) as recently shown by Garcia et al. (2010). The distributions of MNase-seq nucleosome data for *S. cerevisiae* and *S. pombe* (Figure 7b) still present an asymmetry but are of different shapes to those of the MNase-chip data. Let us note that *S. pombe* and *S. cerevisiae* data obtained by the same method present very similar distributions. The main

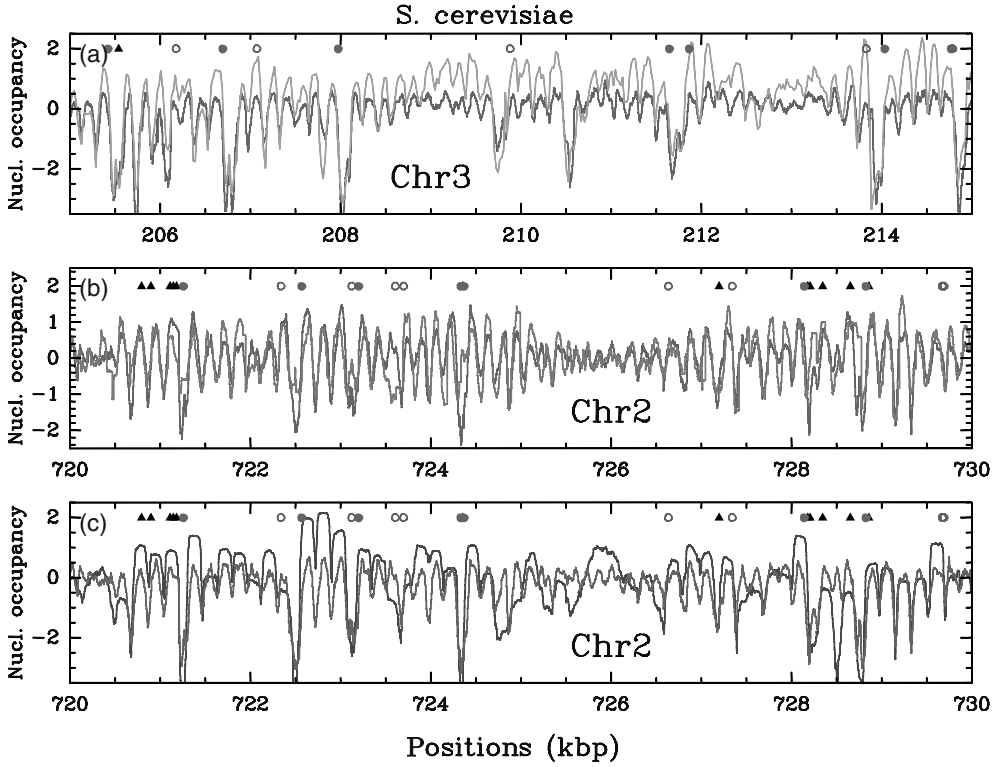

Figure 2. *In vivo* nucleosome occupancy profile $\delta Y(s)$ (see Figure 1) along a 10 kbp fragment of chromosome 3 (a), 2 (b) and 2 (c) of *S. cerevisiae*. Comparison of Lee et al. (2007) MNase-chip data (red) with (a) Yuan et al. (2005) MNase-chip data (green), (b) Whitehouse et al. (2007) MNase-chip data (blue) and (c) Kaplan et al. (2009) MNase-seq data (violet). The symbols have the same meaning as in Figure 1. In (b), the Whitehouse et al. data correspond to a detrended hybridation profile: $Y(s) = Y(s) - \int_{s-a}^{s+a} Y(s)$ with $a \sim 200$ bp. For the sake of comparison, we have applied the same detrending procedure to the Lee et al. data in (b).

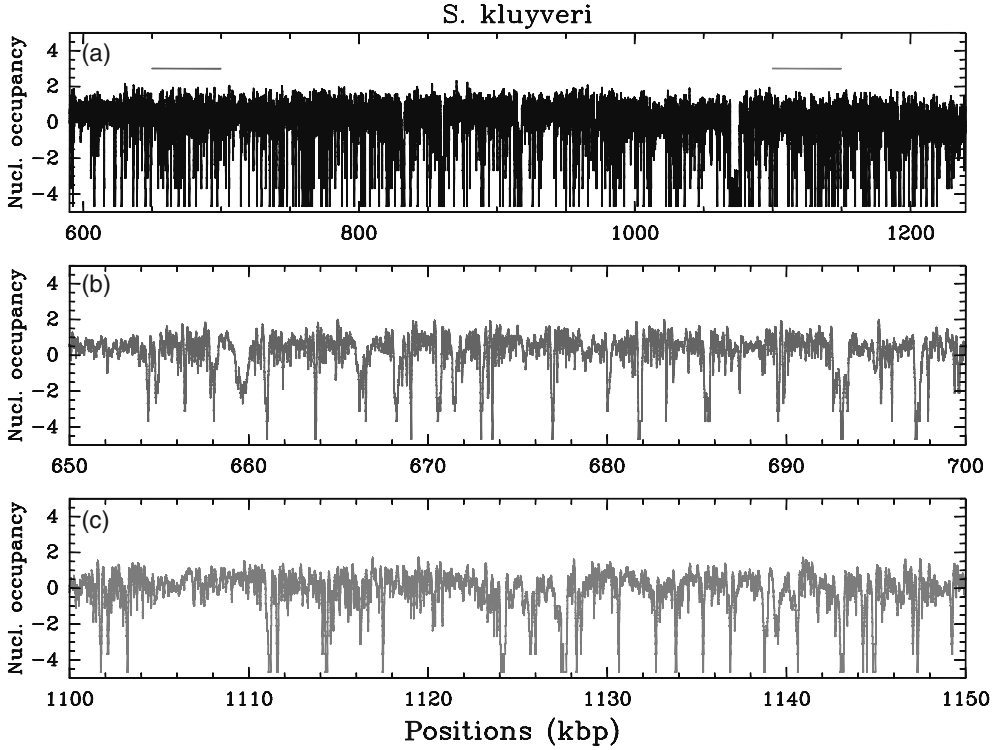

Figure 3. (a) *In vivo* nucleosome occupancy profile $\delta Y(s)$ (see Figure 1) along a 650 kbp fragment of chromosome C of *S. kluyveri*. MNase-seq data of Tsankov et al. (2010). (b) (resp. (c)), zoom on the 50 kbp region indicated in (a) by the red (resp. orange) colored segment, that corresponds to a high (resp. low) (G + C) content region, namely 52% (resp. 40%) as compared to the mean genome value (G+C) = 40%.

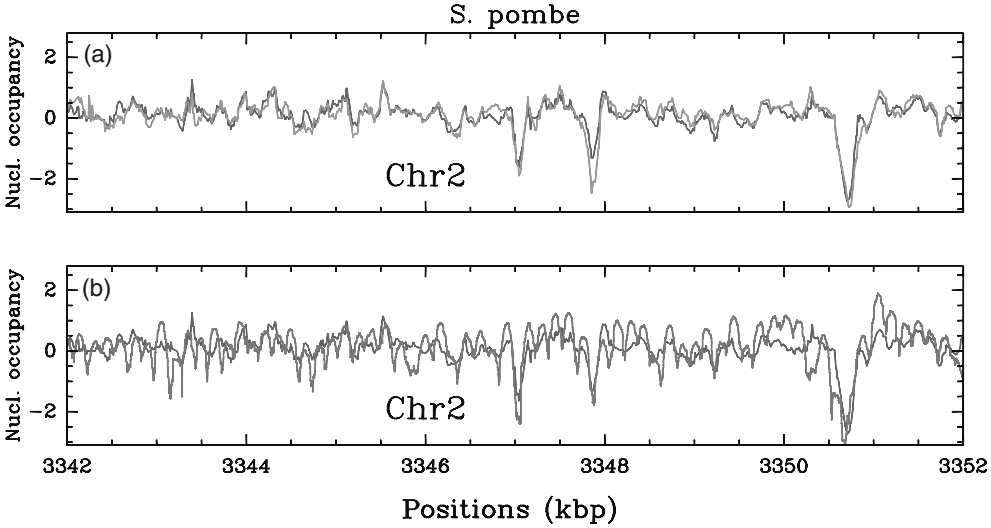

Figure 4. *In vivo* nucleosome occupancy profile $\delta Y(s)$ (see Figure 1) along a 10 kbp fragment of chromosome 2 of *S. pombe*. (a) MNase-chip data of Lantermann et al. (2010): comparison between the WT (red) and the mit1-mutant (green). (b) Comparison between WT MNase-chip data of Lantermann et al. (red, see (a)) and the MNase-seq data of Tsankov et al. (2011) (cyan).

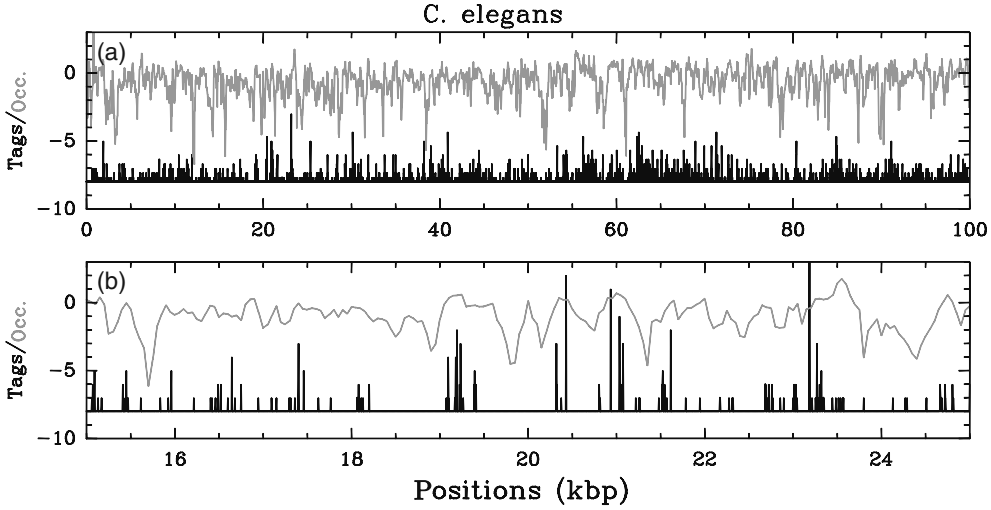

Figure 5. *In vivo* nucleosome occupancy profile $\delta Y(s)$ (see Figure 1) along chromosome 2 of *C. elegans* (Valouev et al. 2008). (a) $\delta Y(s)$ versus s (green) along a 100 kbp fragment as obtained from the 5′ and 3′ ends' tag profiles (black, see Zhang and Pugh 2011). (b) Zoom on a 10 kbp region.

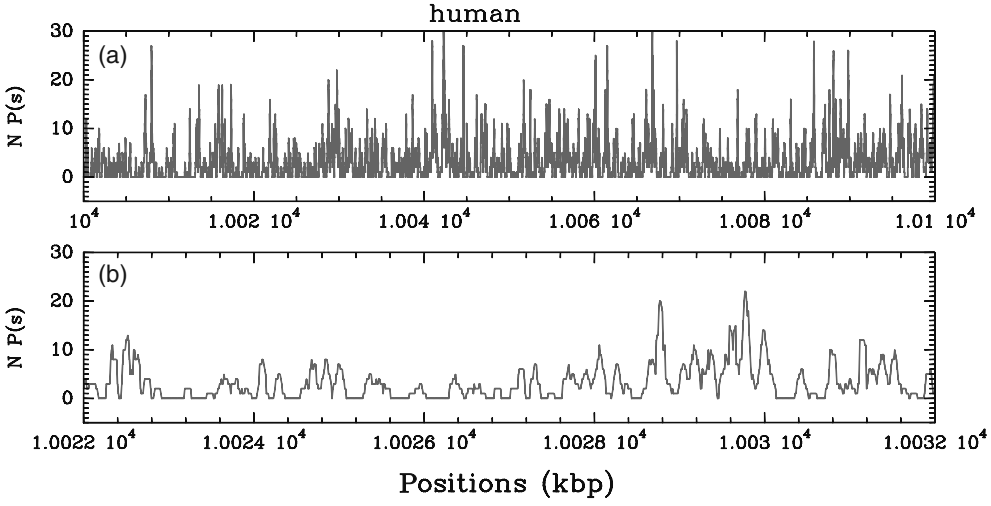

Figure 6. *In vivo* nucleosome (non-normalized) occupancy profile $P(s)$ along chromosome 2 of the human genome (Schones et al. 2008). (a) $P(s)$ versus s along a 100 kbp fragment as obtained from the 5′ and 3′ ends tag profiles (see Zhang and Pugh 2011). (b) Zoom on a 10 kbp region.

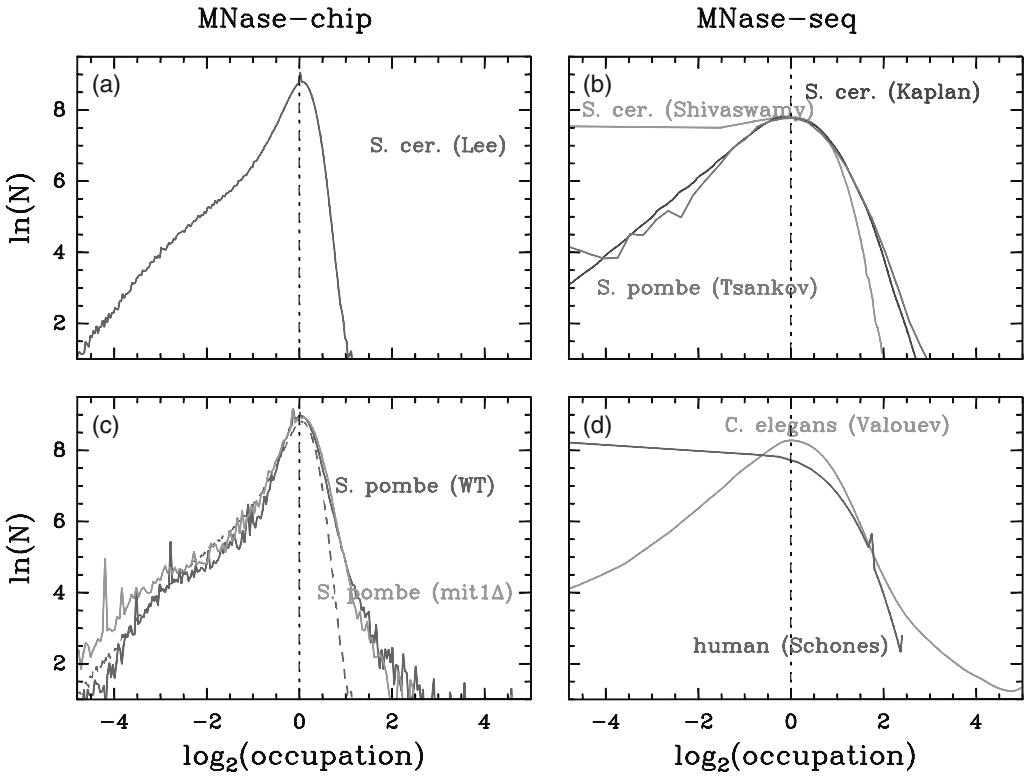

Figure 7. Histograms of nucleosome occupation $Y(s)$ values centered on their typical values (i.e. the maximum of the histogram is positioned at zero) for different sets of *in vivo* data. (a) *S. cerevisiae* MNase-chip data of Lee et al. (2007). (b) *S. cerevisiae* MNase-seq data of Shivaswamy et al. (2008) (green) and of Kaplan et al. (2009) (violet) as compared to *S. pombe* MNase-seq data of Tsankov et al. (2011) (cyan). (c) *S. pombe* MNase-chip data of Lantermann et al. (2010): WT (red) and Mit1 mutant (green); the dashed curve corresponds to the *S. cerevisiae* Lee et al. data shown in (a). (d) *C. elegans* (green) and human (red) MNase-seq data of Valouev et al. (2008) and Schones et al. (2008), respectively.

difference between the MNase-chip and MNase-seq data is at high occupancy values where the fluctuations above the typical value are significantly weaker for MNase-chip data (there is a more pronounced 'bounded-like' behavior). The origin of this discrepancy is not clear: it might be due to a lower statistical sampling in the seq experiments or to some saturation of the fluorescence in the microarray experiments. For some MNase-seq data sets such as the *S. cerevisiae* data of Shivaswamy et al. (2008) (Figure 7b) and the human data of Schones et al. (2008) (Figure 7d), the sequencing depth (Zhang and Pugh 2011) is too low to get reliable distributions. Indeed, the high enrichment observed at low occupancy values is here artificial and is a direct consequence of unmapped genomic loci. Let us point out that for MNase-seq experiments, the *C. elegans* distribution (Figure 7d) is almost similar to the *S. cerevisiae* and *S. pombe* distributions (Figure 7b) apart from the slight excess of the high occupancy values which originates from excessive read enrichment at some repeated sequences. From these distributions and despite discrepancies between experiments and/or possible mapping artefacts, we can provide a rough estimate of the variability of *in vivo* nucleosome occupancy between genomic sites: the occupancy ratio between highest affinity sites and bulk affinity sites (as defined here by the sites of typical occupancy – the most probable value) is around 2–8 fold; between the lowest affinity sites and bulk sites, the depletion ratio is around 20–30 fold. These estimates indicate that the *in vivo* variability of nucleosome occupancy is clearly weak as compared to the 1000 fold enrichment observed *in vitro* between the non-natural 601 sequence, one of the sequences with the highest affinity to the histone octamer, and a random sequence (Lowary and Widom 1998; Thåström et al. 1999). However, such a site-dependent variability may be sufficient to control the accessibility of protein complexes to their target sites and as such to participate in the regulation of nuclear function (Kornberg and Lorch 1999; Boeger et al. 2008; Lam et al. 2008; Wang et al. 2011a, 2011b).

As experienced in various studies (see for a review Arneodo et al. 2011), the statistical organization of nucleosomes along the chromatin fiber can be quantified by performing correlation and power-spectrum analysis. As shown in Figure 8(a), the auto-correlation function $C(\Delta s) = \langle \delta Y(s) \delta Y(s + \Delta s) \rangle$ of *S. cerevisiae* MNase-chip data of Lee et al. (2007) displays a slow decrease with a periodic modulation of period $l^* = 168$ bp. The presence of this small-scale periodic arrangement of nucleosomes manifests itself in the well-defined harmonic peak at $f^* = 1/l^*$ in the power spectrum of the auto-correlation function (Figure 8b). Note that the same periodic modulation is also observed for the other sets of *in vivo* data, e.g. the Whitehouse et al. (2007) MNase-chip data and the Shivaswamy et al. (2008) MNase-seq data (Figure 8). The local value of this period, hereafter called Nucleosome Repeat Length

(NRL), actually fluctuates along the chromosomes. As reported in Figure 9, the NRL distribution is rather narrowly centered on the value 168 bp with root-mean square fluctuations $\sigma = 10$–11 bp, in good agreement with the well-established 165 bp value for *S. cerevisiae* (Woodcock et al. 2006). As shown in Figure 9 (inset), the same distribution is observed for the NRL of *S. kluyveri*. Actually as already pointed out by Tsankov et al. (2010), the mean NRL of the *Hemiascomycota* fungi are relatively well conserved (around 165 bp) with the exception of *C. albicans* and *K. lactis* ($\sim$175 bp). Note that this characteristic NRL observed in the stretches of well-ordered nucleosomes (Figure 1) is significantly smaller than the mean *in vivo* NRL $\sim$210 bp estimated when assuming an homogeneous 75% coverage of the 16 yeast chromosomes by nucleosomes. This suggests the presence of some 'confining' process. The experimental two-point correlation functions in Figure 8(a) also reveal that this 'nucleosome periodicity' statistically appears as a modulation of a dominant slow decaying component which characterizes the large-scale disordered occupancy landscape fluctuations. In our concluding final section, we will discuss the fact that this decay behaves as a power-law as an experimental confirmation that the spatial organization of nucleosomes observed *in vivo* is long-range correlated with characteristics similar to the LRC imprinted in the DNA sequence (Vaillant et al. 2007).

Similar behavior of the nucleosome occupancy auto-correlation function is observed in different eukaryotic organisms. As expected from a simple visual inspection of Figure 4, for *S. pombe* the periodic patterns in the occupancy profile and consequently the periodic modulations of $C(\Delta s)$ are more pronounced and easiest to quantify when analyzing MNase-seq data (Tsankov et al. 2011) than MNase-chip data (Lantermann et al. 2010). As shown in Figure 9, the distribution of NRL values is as wide as previously obtained for *S. cerevisiae* but is significantly shifted to smaller values with a mean value $l^* = 151$ bp, which, to our knowledge, is still unexplained. On the same Figure 9 is also reported the NRL distribution obtained for *C. elegans* data (Valouev et al. 2008); fluctuations are of the same magnitude but the distribution is shifted towards larger NRL values with a larger mean value of 175 bp. Regional

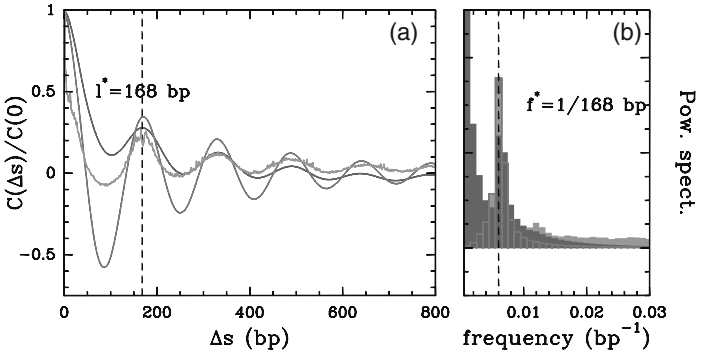

Figure 8. (a) Auto-correlation function $C(\Delta s) = \langle \delta Y(s) \delta Y(s + \Delta s) \rangle$ versus Δs. (b) Corresponding power spectrum. The different colors correspond to the following *S. cerevisiae* data sets: MNase-chip data of Lee et al. (2007) (red) and of Whitehouse et al. (2007) (blue); MNase-seq 'tag' data of Shivaswamy et al. (2008) (green).

variations of the NRL have also been measured recently in human with $l^* \simeq 205$ bp in the heterochromatin HP1 domains, a value significantly larger than in the euchromatin domains where $l^* \simeq 178$–195 bp (Valouev et al. 2011). As suggested by Woodcock et al. (2006), the control of the inter-nucleosome distance might be related *in vivo* to the stoichiometry of the linker histone H1 (per nucleosome) which can be species and cell-type specific: the more abundant the H1, the larger value the NRL (which is indeed the case for *in vitro* assembly (Blank and Becker 1995, see Figure 17 later). In *S. cerevisiae*, the H1 counterpart, Hho1 is indeed weakly present and to our knowledge no such linker histone has been found in *S. pombe*. However, the origin (and causality) of this relationship is not understood. Recent studies rather indicate that the nucleosome spacing is controlled by the action of some remodelling factors such as members of the ISWI and CHD family (ACF, Mit1, Iswi1, Chd1...) (Clapier and Cairns 2009; Garcia et al. 2010; Gkikopoulos et al. 2011) which can be selectively targeted to specific genomic and/or epi-genomic loci in a developmentally regulated manner (Moshkin et al. 2012).

In vitro *nucleosome occupancy profiles*

Only very recently, *in vitro* genome-wide nucleosome occupancy data have become available. One of the main

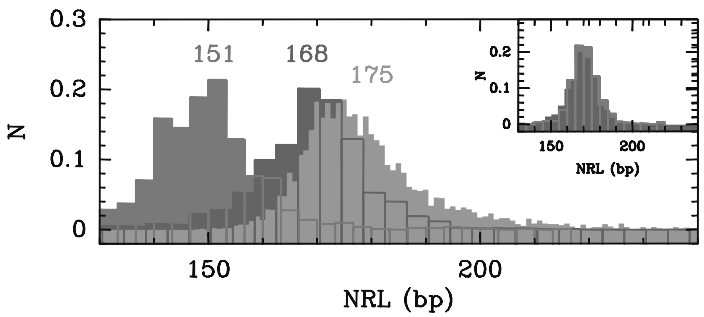

Figure 9. Histograms of local NRL values computed from different nucleosome occupancy data sets. The different colors correspond to the following data: *S. cerevisiae* MNase-chip data of Lee et al. (2007) (red); *S. pombe* MNase-seq data of Tsankov et al. (2011) (cyan); *C. elegans* MNase-seq data of Valouev et al. (2008) (green). (Inset) Comparison betwen the histograms of the *S. cerevisiae* NRLs (Lee et al. 2007) (red) and of *S. kluyveri* NRLs (Tsankov et al. 2010) (orange).

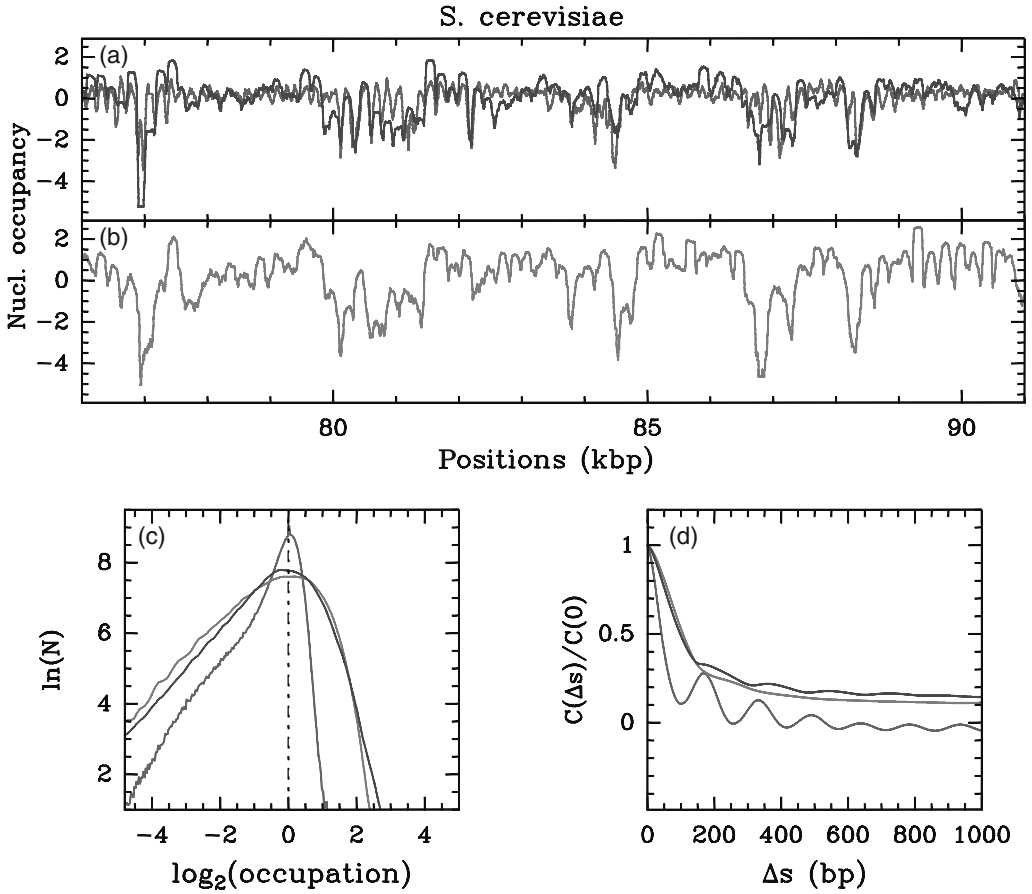

Figure 10. Nucleosome occupancy profile ($\delta Y(s) = Y(s) - \overline{Y}$, where $Y(s) = \log_2(P(s))$) along a 15 kpb long fragment of *S. cerevisiae* chromosomes 12: (a) *in vivo* MNase-chip data of Lee et al. (2007) (red) and *in vivo* MNase-seq data of Kaplan et al. (2009) (violet); (b) *in vitro* MNase-seq data of Kaplan et al. (2009) (orange); (c) corresponding histograms of $Y(s)$ values centered on their typical values; (d) auto-correlation functions $C(\Delta s) = \langle \delta Y(s) \delta Y(s + \Delta s) \rangle$.

difficulties encountered by such experiments is a dramatic limitation in the obtained genome coverage by reconstituted nucleosomes. For example, Kaplan et al. (2009) have assembled chicken erythrocyte histone octamers on purified *S. cerevisiae* genomic DNA by salt gradient dialysis. They succeeded in producing *in vitro* occupancy profiles but for a nucleosome genome coverage ∼30%, which corresponds to a much smaller nucleosome density (mean inter-nucleosome distance ∼500 bp) than previously observed *in vivo*. As shown in Figure 10, the *in vitro* nucleosome occupancy profile along a 15 kbp long fragment of *S. cerevisiae* chromosome 12 looks much more disordered as compared to the corresponding *in vivo* profile obtained either by MNase-chip (Lee et al. 2007) or MNase-seq (Kaplan et al. 2009) experiments. In particular, if we still observe rather localized nucleosome depleted regions that remarkably coincide with some of the NDRs observed *in vivo* (Figure 10a), there is no longer evidence of stretches of periodically distributed nucleosomes. Thus, if the histogram of nucleosome occupancy $\delta Y(s)$ in Figure 10(c) obtained *in vitro* almost superimposes onto the one obtained *in vivo* with the same MNase-seq technique, in contrast, the auto-correlation function $C(\Delta s)$ in Figure 10(d) does not oscillate any longer as the signature of the absence of *in vitro* well-defined NRL. Nevertheless, what is remarkable is the fact that we recover *in vitro* the same slow power-law decay of $C(\Delta s)$ as previously observed *in vivo*; this is the confirmation that the long-range order observed in the collective nucleosome organization of the *S. cerevisiae* chromatin fiber likely results from the LRC that have been imprinted in the DNA sequence (at scales >200 bp) during evolution (see final section 'Discussion') (Audit et al. 2001, 2002; Vaillant et al. 2007; Arneodo et al. 2011).

Thermodynamical model of nucleosome assembly

The mechanisms underlying the formation, the structure and the displacement of nucleosomes are still largely unknown. As a consequence, we will attempt to provide some understanding of the experimental observations reported in the previous section using a phenomenological approach based on simple physical arguments (Chevereau et al. 2009; Milani et al. 2009; Vaillant et al. 2010). When focusing on the dynamical assembly of histone octamers along the DNA chain, we first assume that chromatin can be reasonnably modelled by a fluid of rods of finite extension (the DNA wrapping length around the octamer), binding and moving in an external potential $E(s)$ (the effective nucleosome formation potential) and interacting (potential $V(s, s')$) on a 1D substrate (the DNA chain) (Figure 11). The thermodynamics of such a system has been widely investigated in the literature. In the case of monodisperse hard rods on a

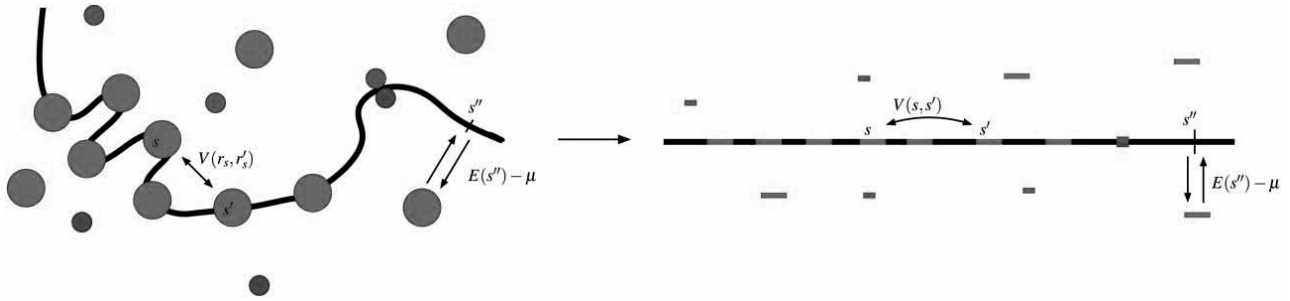

Figure 11. Grand canonical model of nucleosome assembly: bulk histones (green) may adsorb on or desorb from DNA (arrows). Barriers, such as transcription factors or other DNA binding proteins (red) can hinder nucleosome formation. The dynamics is controlled by a thermal bath (kT), the chemical potential of the histone reservoir (μ), the nucleosome–nucleosome interaction ($V(r_i, r_j)$) and the non-homogeneous adsorbing potential $E(s)$. When no tridimensional degree of freedom is considered, the system reduces to a one-dimensional Tonks–Takahashi fluid of hard-rods of hard core size the DNA wrapping length l (Tonks 1936; Takahashi 1942).

uniform external potential, this is the well-known Tonks–Takahashi gas (Tonks 1936; Takahashi 1942). In the case of a non-uniform external potential of interest here, the problem has been partly solved by Percus (1976) who derived an exact functional relationship between the residual chemical potential $\mu - E(s)$ and the hard rod density $\rho(s)$.

Tonks–Takahashi fluid

The Tonks–Takahashi fluid (Tonks 1936; Takahashi 1942) is a one-dimensional fluid of particles obeying the nearest neighbor two-body potential:

$$\begin{aligned} V(x_{ij}) &= \infty,\ |x_{ij}| < l, \\ &= \psi(x_{ij} - l),\ l < |x_{ij}| < l_M, \\ &= 0,\ |x_{ij}| > l_M \end{aligned} \tag{1}$$

where l corresponds to the hard-core length and l_M to the maximal range of ψ which is an arbitrary function of x_{ij}. Lord Rayleigh (Rayleigh 1891) first obtained the equation of state of hard-rods in an homogeneous field in the thermodynamic limit. The distribution functions in this limit were derived later by Salsburg et al. (1953). The case of hard rods in an inhomogeneous external potential was solved by Percus (1976) who derived an exact functional relationship between the local one-point density $\rho(s)$ and the residual chemical potential $\mu - E(s)$. Vanderlick et al. (1986) proposed an iterative scheme to integrate the Percus equation. Robledo and Rowlinson (1986) investigated finite-size effects on homogeneous hard-rod fluids and Davis (1990) generalized these results to homogeneous Tonks–Takahashi systems. The equilibrium properties of Tonks–Takahashi fluids can be derived from the configuration partition function of the N-rod system:

$$\begin{aligned} Z_N(L) = N! \int_0^L \mathrm{d}s_N \int_0^{x_N} \mathrm{d}s_{N-1} \cdots \int_0^{s_2} \mathrm{d}s_1 \\ \times \mathrm{e}^{-\beta(\sum_{i>j}^N V(s_i - s_j) + \sum_i^N E(s_i))} \times \mathrm{e}^{-\beta(V(s_1) + V(L - s_N))}, \end{aligned} \tag{2}$$

where s_i correspond to the positions of the N particles; L is the size of the system and $\beta = 1/kT$. Boundary conditions $V(s_1)$ and $V(L - s_N)$ assume that a particle is maintained fixed at positions 0 and L, respectively. Then, the 'Grand-canonical' partition function (the system equilibrates with a reservoir of particles of chemical potential μ) is given by

$$\Xi(L) = \sum_{N=0}^{\infty} \frac{\mathrm{e}^{\beta N \mu}}{N! \Lambda^N} Z_N(L), \tag{3}$$

where Λ is the de Broglie thermal wavelength (it comes from the integration of the kinetic part of the hamiltonian). For the sake of simplicity, in the following, we will put $\Lambda = 1$ which corresponds to a constant shift in the chemical potential $\mu \leftrightarrow \mu - \beta^{-1} \ln \Lambda$. The one-point density distribution $\rho(s)$ can be computed from the functional derivative

$$\rho(s) = -kT \frac{\delta \ln \Xi(L)}{\delta E(s)}. \tag{4}$$

The explicit forms of this density function and of the k-body density distribution function are given by (Robledo and Rowlinson 1986; Davis 1990):

$$\rho(s) = \mathrm{e}^{\beta(\mu - E(s))} \frac{\Xi(s)\Xi(L-s)}{\Xi(L)}, \tag{5}$$

$$\begin{aligned} \rho^{(k)}(s_1, \ldots, s_k) = \mathrm{e}^{\beta k \mu} \frac{\mathrm{e}^{-\beta \sum_{i=1}^k E(s_k)}}{\Xi(L)} \\ \times \Xi(s_1) \prod_{j=2}^{k} \Xi(s_j - s_{j-1}) \Xi(L - s_k). \end{aligned} \tag{6}$$

$\rho(s)$ corresponds to the probability of finding a particle at position s. $\rho^{(2)}(s_i, s_j)$ corresponds to the probability of finding a pair of particles at positions s_i and s_j. This function defines locally the equilibrium statistical distribution of distances between successive particles. The shape of this pair function directly depends on the interaction potential $V(s_i, s_j)$ (Davis 1990) (see subsection 'Homogeneous energy profile $E = E_o$'). Classical measurement of the NRL by gel analysis of chromatin digestion products is directly related to this pair function (see Figure 17 later) (Noll and Kornberg 1977; Blank and Becker 1995; Woodcock et al. 2006). The pressure of this system can be computed directly

from the partition function (Lieb and Mattis 1966; Davis 1990):

$$P = kT \frac{\partial \ln \Xi(L)}{\partial L}. \tag{7}$$

The homogeneous case $E = E_o = c^{ste}$

Evaluation of the grand canonical ensemble partition function $\Xi(L)$ is not easy in general. However, in the special case of an homogeneous potential $E = E_o$, the problem is equivalent to the $E = 0$ homogeneous case when considering a residual chemical potential $\tilde{\mu} = \mu - E_o$. To simplify the notations, we will thus treat the $(E = 0, \mu)$ case for which the canonical ensemble partition functions $Z_N(L)$ can be expressed as the N-convolution integral of the interaction Boltzmann weight $e^{-\beta V(s)}$. By denoting $K(p)$ the Laplace transform of the interaction Boltzmann weight:

$$K(p) = \int_0^\infty e^{-ps-\beta V(s)}\, ds, \tag{8}$$

the partition function then writes:

$$Z_N(L) = Z_N(L - Nl) = \frac{N!}{2\pi i} \int_{-i\infty+\tau_o}^{i\infty+\tau_o} e^{(L-Nl)p} [K(p)]^{N+1}\, dp, \tag{9}$$

from which one can express $\Xi(L)$ [Equation (3)] and the density distributions [Equations (5) and (6)] as sole functions of $K(p)$:

$$\Xi(L) = \sum_{N=0}^{\infty} e^{N\beta\mu} \int_{-i\infty+\tau_o}^{i\infty+\tau_o} \frac{e^{(L-Nl)p}}{2\pi i} [K(p)]^{N+1}\, dp. \tag{10}$$

At the thermodynamic limit, the pressure P [Equation (7)] of the bulk Tonks–Takahashi fluid at chemical potential μ obeys the equation of state (Salsburg et al. 1953; Lieb and Mattis 1966):

$$\beta\mu = \beta P l - \ln K(\beta P). \tag{11}$$

The density–pressure equation can then be derived from this chemical potential–pressure equation using the Gibbs–Duhem equation:

$$1/\rho = (\partial \beta\mu)/(\partial \beta P). \tag{12}$$

In the case of hard rods, $V(s)$ has the form of the Heaviside distribution and K takes the simple form, $K(p) = 1/p$; we thus recover the equation of state derived by Rayleigh (1891):

$$\beta\mu = \beta P l + \ln(\beta P l) + \ln(1/l), \tag{13}$$

$$\beta P = \frac{\rho}{1 - \rho l}. \tag{14}$$

From Equations (13) and (14), we deduce the following chemical potential–density relationship:

$$\beta\mu = \ln(\rho) - \ln(1 - \rho l) + \frac{\rho l}{1 - \rho l}. \tag{15}$$

Percus equation

When the energy landscape $E(s)$ is non-uniform (with the exception of energy landscapes composed of infinite energy barriers and flat potential in between) the method of Laplace transform presented above cannot be applied. In 1976, Percus derived an equation that gives the density of a Tonks–Takahashi hard-rods fluid in an inhomogeneous energetic field as a function of the chemical potential and the temperature (Percus 1976):

$$\beta\mu = \beta E(s) + \ln \rho(s) - \ln\left(1 - \int_s^{s+l} \rho(s')\, ds'\right) + \int_{s-l}^{s} \frac{\rho(s')}{1 - \int_{s'}^{s'+l} \rho(s'')\, ds''}\, ds,' \tag{16}$$

where

- s is the position along the potential; experimentally it usually corresponds to the genomic position of the nucleosome 'dyad' or of one of the nucleosome borders (5′ or/and 3′ extremity).
- l corresponds to the size of hard rod: the nucleosome wrapping length, assumed to be fixed at $l = 146$ bp.
- ρ is the density of hard rods.
- μ represents the chemical potential (i.e. the energy transferred by the bulk reservoir).
- $\beta = (kT)^{-1}$ is the reciprocal temperature.
- $E(s)$ is the free energy of nucleosome formation on a sequence at position s.

Remark: Note that for the homogeneous case $E(s) = E_o$, $\rho(s) = \rho = c^{ste}$ and the Percus equation reduces to Equation (15).

Resolution of the Percus equation: the exact solution of Vanderlick et al.

A technical trick was proposed by Vanderlick et al. (1986) in order to solve the Percus equation (16) exactly. It is indeed possible to write the Percus equation as follows:

$$f(s) = \exp\left(\beta\mu - \beta E(s) - \int_{s-l}^{s} f(s')\, ds'\right), \tag{17}$$

where

$$f(s) \equiv \frac{\rho(s)}{1 - \int_s^{s+l} \rho(s')\, ds'}. \tag{18}$$

Note that the function f – for *forward* – so introduced is a function that depends on the 'past' only. By taking the

derivative of Equation (17), we get:

$$\frac{\mathrm{d}f}{\mathrm{d}s}(s) = f(s)\cdot\left[-\beta\frac{\partial E}{\partial s}(s) + f(s-l) - f(s)\right]. \quad (19)$$

This equation has for general solution:

$$f(s) = \frac{u(s)}{\dfrac{\exp(-\beta E(s_0))}{f(s_0)} + \int_{s_0}^{s} u(s')\,\mathrm{d}s'}, \quad (20)$$

where

$$u(s) \equiv \exp\left(-\beta E(s) + \int_{s_0}^{s-l} f(s')\,\mathrm{d}s'\right). \quad (21)$$

It is thus possible to solve Equation (20) iteratively starting from a simple initial condition for f: $f(s) = 0,\ s \in [-l; 0]$. To compute the density let us introduce a new function $b(s)$ such that:

$$f(s) = \frac{\rho(s)}{b(s)}. \quad (22)$$

From Equations (18) and (22) we deduce:

$$b(s) = 1 - \int_{s}^{s+l} \rho(s')\,\mathrm{d}s' = 1 - \int_{s}^{s+l} f(s')\cdot b(s')\,\mathrm{d}s'. \quad (23)$$

The function b – for *backward* – depends on the 'future' only. When taking the derivative of Equation (23), we obtain:

$$\frac{\mathrm{d}b}{\mathrm{d}s}(s) = f(s)b(s) - f(s+l)b(s+l), \quad (24)$$

the general solution of which writes:

$$b(s) = b(s_0)\exp\left(\int_{s_0}^{s} f(s')\,\mathrm{d}s'\right) + \int_{s}^{s_0}\left[\exp\left(-\int_{s}^{s''} f(s')\,\mathrm{d}s'\right) b(s''+l) f(s''+l)\right]\mathrm{d}s''. \quad (25)$$

This equation can again be solved iteratively by considering the boundary conditions: $b(s) = 1,\ s \in [L; L+l]$. Then, once $f(s)$ and $b(s)$ are so computed, we use Equation (22) to get the density $\rho(s) = f(s)b(s)$ (Figure 12).

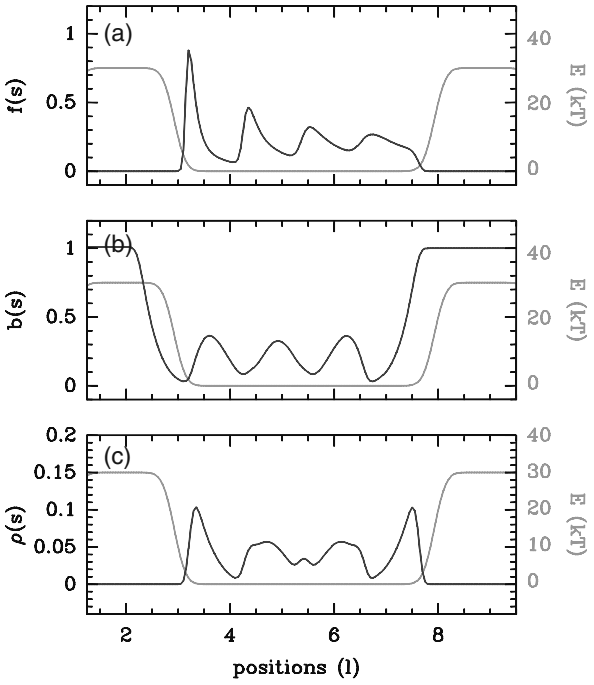

Figure 12. Illustration of the Vanderlick et al. exact solution (blue) of the Percus equation [Equation (16)]. The energy landscape ($E(s)$) used for the computation is shown in green. (a) $f(s)$ 'forward' function; (b) $b(s)$ 'backward' function; (c) the resulting density $\rho(s) = f(s)b(s)$. Model parameters: potential wall amplitude $= +30\,\mathrm{kT}$, $\mu = +3\,\mathrm{kT}$.

Segal et al. method

Recently, in the context of nucleosome positioning modelling, Segal et al. (2006) have proposed an alternative solution for the computation of the density, using a Markovian algorithm that explicitly builds the full partition function of the system. According to Percus (1982), the density can be written as:

$$\rho(s) = w(s)\frac{\xi_\mu(s)\hat{\xi}_\mu(s)}{\xi_T}, \quad (26)$$

where

$$w(s) = \mathrm{e}^{\beta(\mu - E(s))} \quad (27)$$

and

$$\xi(s) = \sum_{N=0}^{\infty} \mathrm{e}^{N\beta\mu} Q_N(s), \quad \hat{\xi}(s) = \sum_{N=0}^{\infty} \mathrm{e}^{N\beta\mu}\hat{Q}_N(s), \quad (28)$$

with

$$Q_N(s) = Z_N^{\mathrm{left}}(0 \leftrightarrow s), \quad \hat{Q}_N(s) = Z_N^{\mathrm{right}}(s \leftrightarrow L). \quad (29)$$

ξ_T is the total grand partition function. The main difficulty relies in the computation of $Z_N^{\mathrm{left}}(s)$ (resp. Z_N^{right}) corresponding to the number of configurations that accomodate N particles in the space $0 < x < s$ $(0 \leftrightarrow s)$ (resp. $s < x < L$, $(s \leftrightarrow L)$).

The explicit formulation of Z_N [Equation (2)] (Percus 1982) is:

$$Z_N = \int\cdots\int \prod_{i=2}^{N} \mathrm{e}^{-\beta V(s_i - s_{i-1})} \prod_{i=1}^{N} \mathrm{e}^{-\beta E(s_i)} \prod_{i=1}^{N} \mathrm{d}s_i, \quad (30)$$

which is unfortunately not easily tractable. However in the particular case of hard-rods, $Q_N(s)$ [Equation (29)] can be

computed recursively:

$$Q_N(s+1) = Q_N(s) + Q_{N-1}(s-l)\mathrm{e}^{-\beta E(s+1-l)}, \quad (31)$$

where $E(s)$ is the energy of a particle (nucleosome) that starts at s. Then, if we note N_M the maximum number of particles that can be put in the interval $[0, s]$, we get:

$$\begin{aligned}\xi_\mu(s+1) &= \sum_{N=0}^{\infty} \mathrm{e}^{N\beta\mu} Q_N(s+1), \\ &= \sum_{N=0}^{N_M} \mathrm{e}^{N\beta\mu} Q_N(s) + \sum_{N=1}^{N_M} \mathrm{e}^{N\beta\mu} Q_{N-1}(s-l) \\ &\quad \times \mathrm{e}^{-\beta E(s+1-l)}, \\ &= \xi_\mu(s) + \sum_{N=0}^{N_M-1} \mathrm{e}^{N\beta\mu}\mathrm{e}^{+\beta\mu} Q_N(s-l)\mathrm{e}^{-\beta E(s+1-l)}, \\ &= \xi_\mu(s) + \xi_\mu(s-l)\mathrm{e}^{\beta(\mu - E(s+1-l))}, \end{aligned} \quad (32)$$

where we have explicitly used the fact that, by definition, it is possible to put no more than $N_M - 1$ particles of size l in the interval $[0, s-l]$. By proceeding in the same way for $\hat{\xi}_\mu(s)$ [Equation (28)] we get:

$$\hat{\xi}_\mu(s-1) = \hat{\xi}_\mu(s) + \hat{\xi}_\mu(s+l)\mathrm{e}^{\beta(\mu - E(s-1+l))}. \quad (33)$$

Then we use Equation (26) to determine the density $\rho(s)$, where $\xi_\mu(s)$ and $\hat{\xi}_\mu(s)$ have been estimated recursively from the left ($s = 0$) and the right ($s = L$) ends of the system according to Equations (32) and (33) respectively. As compared to the Vanderlick et al. algorithm described in the previous subsection, the Segal et al. method (2006) has the main inconvenience of involving sums of terms that can very quickly become very important which may result in the accumulation of numerical errors.

Teif and Rippe method

Teif and Rippe (2009) have recently proposed a theoretical framework for lattice models of histone–DNA interactions. Their strategy mainly consists in using the transfer matrix method to compute the grand canonical partition function (Baxter 1982). The idea is to associate a $l \times l$ matrix to each position s along the DNA sequence, where $l = 146$ corresponds to the 146 possible positions of the nucleotide s in the nucleosome. The matrix $Q_s(i,j)$ represents the probability that the nucleotide s is in the state i knowing the fact that the nucleotide $s+1$ is in the state j. Now if $E(s)$ is the binding energy for a nucleosome starting at position s and extending up to $s + l$, then the matrix elements that are non-null write $Q_s(i, i+1) = \mathrm{e}^{\beta(\mu - E(s-i+1))}$ where $1 < i < l$. The grand canonical partition [Equation (3)] can then be expressed by summing over all the possible accessible states:

$$\Xi(L) = (1 \;\; 1 \;\; \ldots \;\; 1) \times \prod_{s=1}^{L} Q_s \times \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix}. \quad (34)$$

Note that it is possible to compute this partition function iteratively:

$$\begin{aligned}\Xi(L) &= A_L \times \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix}, \quad A_s = A_{s-1} \times Q_{s-1}, \\ A_0 &= (1 \;\; 1 \;\; \ldots \;\; 1). \end{aligned} \quad (35)$$

To determine the density distribution $\rho(s)$ [Equation (4)], we just have to derive the partition function with respect to $K_s = \mathrm{e}^{-\beta E(s)}$ and then to proceed iteratively consistently with Equation (35):

$$\frac{\partial \Xi}{\partial K_s} = \frac{\partial A_N}{\partial K_s} \times \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix}, \frac{\partial A_s}{\partial K_s} = \frac{\partial A_{s-1}}{\partial K_s} \times Q_s + A_{s-1} \times \frac{\partial Q_s}{K_s}. \quad (36)$$

The main advantage of the Teif and Rippe (2009) method is its adaptability to account for additional states, e.g. corresponding to other proteins or to interaction with neighboring nucleosomes. Among the disadvantages, the necessity to compute the total partition function is numerically costly when working at the chromosome scales.

Statistical positioning

In this section, we present and discuss the results of the numerical integration of the Percus equation (16) using the Vanderlick et al. integration scheme described in the subsection entitled 'Resolution of the Percus equation: the exact solution of Vanderlick et al.' for simple and illustrative energy landscapes made of energy barriers, traps and flat regions. The nucleosome occupancy profile $P(s)$, as defined in section '*In vivo* and *in vitro* genome-wide primary structure of chromatin', will be obtained by convolving the nucleosome density $\rho(s)$ via the rectangular function Π of width 146 bp:

$$P(s) = \rho \circ \Pi_{146}(s). \quad (37)$$

Illustrative energy landscape

For pedagogical purposes, we show in Figure 13, the theoretical occupancy profiles of nucleosomes considered as hard rods of hard-core size 146 bp obtained at different chemical potential values in a toy energy landscape bordered by two infinite walls and that displays a stretch of square-like wells on the right-side and two square-like energy barriers on the left side.

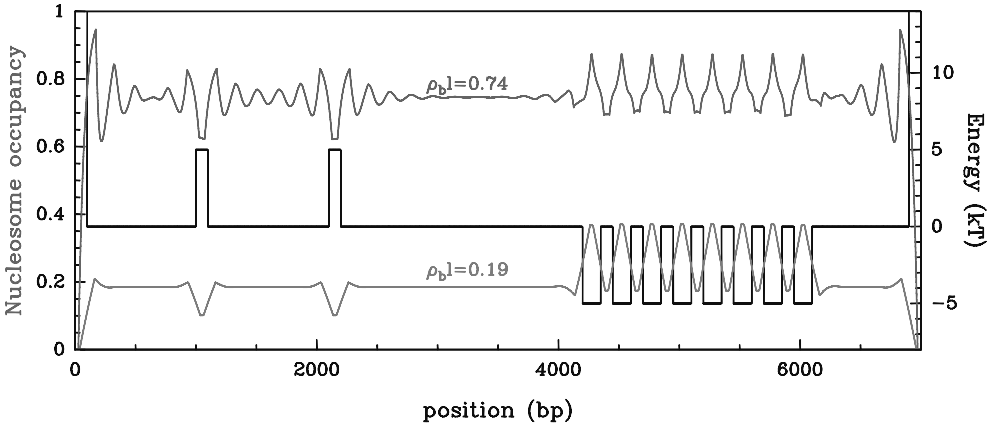

Figure 13. Hard-rod occupancy profiles $P(s)$ in a non-uniform energy landscape made of discrete energy barriers and traps and bounded by two infinite walls (black). The Percus equation (16) was solved using the Vanderlick et al. integration scheme (subsection 'Resolution of the Percus equation: the exact solution of Vanderlick et al.') at low chemical potential $\mu = -6\,\mathrm{kT}$ (orange) and high chemical potential $\mu = -1\,\mathrm{kT}$ (red). The occupancy values $\rho_b l = 0.19$ (resp. $\rho_b l = 0.74$) correspond to the bulk occupancy of the uniform system at $\mu = -6\,\mathrm{kT}$ (resp. $\mu = -1\,\mathrm{kT}$) (see Figure 14).

Low chemical potential

At low chemical potential, the occupancy profile $P(s)$ reflects the energy landscape topography: local depletion is observed at energy barriers and local enrichement at energy traps. At low chemical potential, the average density (occupancy) is weak and the Percus equation indeed reduces to its diluted (or non-interacting) approximation:

$$\beta\mu = \beta E(s) + \ln \rho(s), \tag{38}$$

where the nonlinear and nonlocal terms in ρ have been neglected. Then the density $\rho(s)$ at position s only depends on the energy $E(s)$ at this position via the simple Boltzmann relationship:

$$\rho(s) \sim \mathrm{e}^{\beta(\mu - E(s))}. \tag{39}$$

At regions where the energy landscape is uniform, density is constant and equals the bulk density [Equation (15)]: $\rho(s) = \rho_b \simeq \exp \beta(\mu - E_0)$ (see Figure 13, $\rho_b l = 0.19$).

High chemical potential

When increasing the chemical potential, the average density increases, local depletion and enrichment are still observed at barriers and traps respectively. As expected and already observed at low μ values, an array of well-positioned nucleosomes is observed as induced by the stretch of regularly distributed potential wells (e.g. a stretch of highly positioning sequences like the well-known 601 sequence (Lowary and Widom 1997)). More interesting are the oscillations that appear in $P(s)$ in the vicinity of the energy barriers and of the bordering walls. This 'periodic' positioning, which is not induced by any local periodically distributed energy traps, is purely entropic (Kornberg and Stryer 1988). It results from the confinement of the hard rods imposed by the excluding energy barriers: the pressure imposed on each particle by the surrounding particles increases with the density; so at high enough density the first particle next to (flanking) the barrier experiences a pushing force from the rest of the fluid that confines it against the barrier. Positioning (as defined by spatial localization) is thus the strongest at this location. The next particle is also confined from one side by the fluid pressure and from the other side by the (confined but) moving first particle and not by a fixed barrier. So its positioning is weaker, and so on, positioning strength decreases as the distance to the barrier increases. Far from the barrier, the density profile becomes uniform and equals the bulk value ρ_b given by Equation (15). Amplitude and range of this nonlocal-induced periodic ordering depend on the barrier height and shape (that defines the pressure exerted on the fluid) and on the chemical potential (that defines internal pressure of the fluid) as further discussed in the following subsections. Indeed this periodic ordering is an internal property of dense fluids; it is usually 'revealed' in the occupancy profile by the presence of inhomogeneities (vertical energy barriers in Figure 13) in the energy landscape.

Remark: An interesting point here is that from low density (low chemical potential) profile, we can extract the underlying energy landscape via the simple Boltzmann relationship [Equation (39)]. Actually the Percus equation (16) provides a direct computation of the energy landscape from the density profile at every chemical potential. However, since this relationship is only valid for hard-core repulsion, applying it to a Tonks–Takahashi fluid with an 'a priori' unknown interaction potential V (for example, applying it to high density *in vivo* data, e.g. the distribution of tags in MNase-seq experiments, see subsection '*In vivo* nucleosome occupancy profiles') may lead to an interaction-dependent bias in the energy landscape estimation. This 'bias' is likely to be minimized when using low density profile (for example, the low density *in vitro* data, see subsection '*In vitro* nucleosome occupancy profiles').

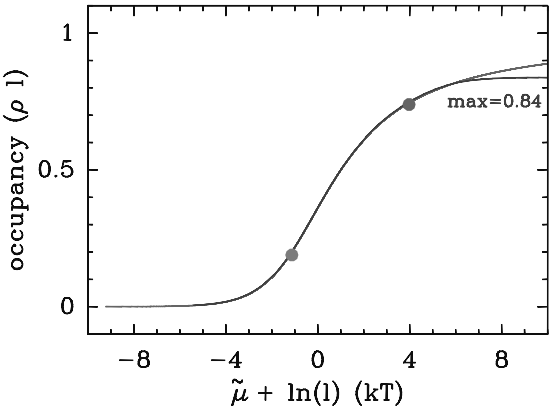

Figure 14. Occupancy (ρl) of hard rods in a homogeneous energy landscape $E(s) = E_o$, as a function of the residual chemical potential $\tilde{\mu} = \mu - E_o$. Theoretical curves obtained from Equation (14) (red, infinite size system) and from Vanderlick numerical method (blue, large but finite size system). The dots indicate the bulk occupancy values for the chemical potential values used in Figure 13: $\tilde{\mu} = -6\,\mathrm{kT}$ (orange dot) and $\tilde{\mu} = -1\,\mathrm{kT}$ (red dot).

Homogeneous energy profile $E = E_o$

Nucleosome density

In the 'ideal' case of a flat potential ($E = E_o = c^{ste}$), the density $\rho = \rho_b$ is homogeneous (the 'bulk' phase density) and is controlled by a single parameter, the residual chemical potential $\tilde{\mu} = \mu - E_o$. As reported in the section 'A sequence-dependent physical model of nucleosome occupancy' (see Figure 27b later on, the zone F_5 of 3.5 kbp), this 'ideal' situation can be realistically achieved in the frame of our sequence-dependent energy model at some specific genomic regions. We may also expect that, in some genomic regions, the action of remodelling factors can result in an effective reduction of the sequence-dependent potential fluctuations leading to an approximate local homogenization of the energy landscape ($\delta E(s) \to 0$)) (Zhang et al. 2011). In Figure 14, we show the evolution of the occupancy ($\rho_b l$) as a function of $\tilde{\mu}$ computed from Equation (15). It has a sigmoidal shape: at very low $\tilde{\mu}$ values the density is weak; then it slightly increases and goes through a transition at $\tilde{\mu}_c = -\ln(l)$, where the susceptibility $\chi = \partial\rho/\partial\mu$ is maximum. At this point, $\rho(\tilde{\mu}_c)l = 0.36$, and the occupancy rapidly increases up to $\rho l \simeq 0.75$. When further increasing $\tilde{\mu}$, it then slowly increases towards the asymptotic limit $\rho^* l = 1$. Interestingly, for very large $\tilde{\mu}$ values, finite-size systems experience a transition towards a saturated $\rho l = 0.85$ occupancy state (Figure 14, blue curve) which has been identified as a pseudo-crystalline state consisting of quasi regularly spaced particles 'self-confined' inside equipartitioned regions of length equal to the average length per particle $l_m = 1/\rho$ (Piasecki and Peliti 1993; Giaquinta 2008). In inhomogeneous fluids, this saturation only occurs for the bulk occupancy ($\rho_b l = l/L \int_0^L \rho(s)\,\mathrm{d}s$) as illustrated in Figure 13 where at the edges near the infinite energy walls, the local occupancy $\rho(s)l$ can reach a value close to one.

Inter-nucleosomal distance

As defined in the section '*In vivo* and *in vitro* genome-wide primary structure of chromatin', key distances characterizing the primary structure of chromatin are the nucleosome repeat length (NRL) $l^*(s)$ and the linker size $d(s)$, i.e. the part of unwrapped DNA that joins two successive octamers. The NRL and the linker size are directly related: $l^* = l + d$. As we shall see later, this distance can be extracted experimentally by MNase digestion and gel analysis of digestion products (Blank and Becker 1995) or directly by single molecule imaging of the 10 nm fiber (Solis et al. 2004; Milani et al. 2009). This distance differs in general from the mean inter-nucleosome distance which is simply given by $l_m = \rho_b^{-1}$. In the fluid formalism (Hansen and McDonald 2006), the statistical properties of the inter-particle distances are given by the pair density distribution $\rho^2(s_i, s_j)$ which represents the joint probability to have particles (of fixed size l) at positions s_1 and s_2. Using the Bayes formula, we have the following general decomposition:

$$\rho^2(s_1, s_2) = \rho(s_1|s_2)\rho(s_2), \tag{40}$$

where $\rho(s_1|s_2)$ is the conditional probability of having a particle at position s_1 given that a particle is fixed at position s_2, and $\rho(s_2)$ is the particle density at position s_2. In a Tonks–Takahashi fluid (see subsection 'Tonks–Takahashi fluid'), $\rho(s_1|s_2)$ corresponds exactly to the density $\rho(s_1)$ of the fluid when confined by a wall at position s_2 (imposed by the conditional fixed particle) with a wall–particle interaction corresponding to the inter-particle interaction function $V(s_2, s_1)$. The computation of the pair distribution thus only requires to compute the density $\rho_w(s)$ of the semi-confined Tonks–Takahashi fluid. In the case of an homogeneous fluid, $\rho(s_2) = \rho_b$ is the bulk density and $\rho(s1|s2) = \rho_w(r)$, where ρ_w correspond to the density of the fluid at distance $r = |s_1 - s_2|$ from the wall. As discussed in the subsection 'Tonks–Takahashi fluid', the bulk density is obtained at the thermodynamic limit by solving Equation (15). When using, Equations (3), (5) and (7), we get for $\rho_w(r)$ (Davis 1990):

$$\rho_w(r) = \exp(\beta\mu)\exp[-\beta Pr]\Xi(r - l). \tag{41}$$

For hard-rod fluids, when using Equation (10) with $K(p) = 1/p$, we get:

$$\rho_w(r) = \mathrm{e}^{\beta\mu}\exp - \left(\frac{\rho_b}{1-\rho_b l}r\right) \times \sum_{N=0}^{\infty}\frac{\mathrm{e}^{N\beta\mu}}{N!}(r-(N+1)l)^N\theta(r-(N+1)l), \tag{42}$$

where θ is the heavyside function ($\theta(r) = 0, r > 0, = 1, r < 0$). Now let us introduce the pair correlation function

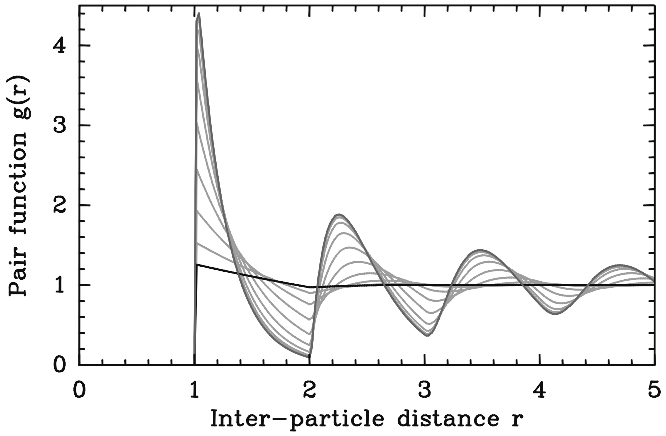

Figure 15. Evolution of the pair function $g(r)$ [Equation (44)] with the residual chemical potential $\tilde{\mu}$. Black and red curves correspond to a chemical potential $\tilde{\mu} = -5$ kT and $+5$ kT respectively. Gray curves correspond to intermediate values of $\tilde{\mu}$. The inter-particle distance $r = |s_1 - s_2|$ is expressed in l units.

$g(s_1, s_2)$ as:

$$g(s_1, s_2) = \rho^2(s_1, s_2)/\rho(s_1)\rho(s_2) = \rho(s_1|s_2)/\rho(s_1). \quad (43)$$

In the case of a uniform fluid:

$$g(r) = \rho^2(r)/\rho_b^2 = \rho_w(r)/\rho_b. \quad (44)$$

Let us note again that whatever the distribution function, the mean inter-nucleosome distance is given by: $l_m = 1/L \int_l^L r g(r)\, \mathrm{d}r = \rho_b^{-1}$. The pair function $g(r)$ can be computed from the analytical form [Equation (42)], where for short distances r only a few modes $N < r/l$ are contributing. In Figure 15, to get the complete pair function, we proceed as in Figure 13 by solving the Percus equation (16) in the semi-confinement configuration (vertical barrier: $E = \infty, x < 0; E = 0, x > 0$ (see subsection 'Statistical ordering near an energy barrier'). At very low chemical potential, i.e. at very low bulk density, the pair function is almost constant ($g = 1$), consistent with the fact that particles are almost isolated (and don't interact) leading to a constant pair distribution $\rho^2(r) = \rho_b^2$: all inter-nucleosomal distances are equiprobable. When increasing the chemical potential, particles interact resulting in oscillations in the pair function, whose amplitude (resp. period) increases (resp. decreases) with the density (Figure 15): around each particle there is preferential periodic ordering (of period l^*) of flanking particles. Sufficiently far from the reference particle, the pair function tends to the value $g = 1$ of the diluted (unorganized) fluid, as well as $\rho_w(r)$ tends to the bulk density ρ_b. The origin of the oscillations (internal ordering) is by definition the same as the one previously described in the density profiles for the semi-confined fluid near an infinite wall (Figure 13). Here, each particle acts as a confining wall in its reference frame.

It is quite interesting to note that the typical inter-nucleosome distance is actually l the size of the hard rods: the typical (most probable) configuration consists in having neighboring hard rods in contact (i.e. linker size $d(s) = 0$), and this whatever the chemical potential value (Figure 15). For hard rods, this typical distance thus differs from the NRL l^* which rather corresponds to the order parameter of the fluid. The ordered phase of the dense uniform hard-rod fluid is indeed characterized by this period l^* and the damping length λ of the pair function. As shown in Figures 16(a) and (b), the attenuation of the oscillations and the relaxation towards the asymptotic value 1 is indeed exponential with a well-defined characteristic correlation length λ; these oscillations are periodic with a well-defined period l^*. The computation of these characteristic lengths from the analytical expression of the pair function [Equations (42) and (44)] is not obvious; so we used numerical calculations to produce the results shown in Figure 16. In Figure 16(c) is reported the evolution of the characteristic lengths λ and l^*, as compared to the mean inter-nucleosome distance l_m, as a function of the chemical potential $\tilde{\mu}$. Let us first point out that periodic ordering could only be defined for $\tilde{\mu} > \mu_c = \ln l$, which corresponds to the transition point of the ρ_b versus μ plot in Figure 14. At this critical point, for hard rods of size $l = 146$ bp, $l^* = 1.45l = 212$ bp and $l_m = 0.36^{-1}l = 405$ bp. The damping length is $\lambda = 1.18l = 172$ bp. More interestingly, for a higher chemical potential $\tilde{\mu} = -1$ kT corresponding to a bulk occupancy of $\rho_b l = 0.74$, we find a period of $l^* = 1.22l = 178$ bp and $l_m = l/0.74 = 198$ bp. The damping length is $\lambda = 2.5l = 365$ bp which means that at a distance $\sim$6 from the (wall) particle, the ordering strength (i.e. the density 'excess': $\rho(6l) - \rho_b$) is 10 fold less than the one observed next to the reference particle (wall). At a distance of $12l$, ordering is lost (the density excess is 100 fold less than the one observed next to the reference particle). By further increasing $\tilde{\mu}$, we are entering the solid-like phase and the period l^* varies slowly, from $l^* = 1.2l = 175$ bp ($l_m = l/0.78 = 187$ bp) for $\tilde{\mu} = 0$ kT to $l^* = 1.16l = 169$ bp ($l_m = 1/0.84 = 174$ bp) for $\tilde{\mu} > 4$ kT. Note that the *in vivo* experimental value $l^* = 168$ bp observed in budding yeast would thus correspond to the solid-like (saturated) phase with an occupancy of at least 84%. At this global density, the *S. pombe* NRL value $l^* = 151$ bp would then correspond to a reduced hard-core length value $l = 130$ bp.

It is important to emphasize here that in this homogeneous energy profile $E = E_o$, the density profile $\rho(s) = \rho_b$ is homogeneous and doesn't present any oscillation as observed in Figure 13 far from the confining energy barriers. To observe a similar ordering pattern in the density profile one has to introduce a vertical barrier in the energy landscape. This will be discussed in the subsection 'Statistical ordering near an energy barrier'. Experimentally, the NRL is measured through an enzymatic digestion procedure followed by a gel characterization of digestion products. Restriction enzymes like MNase preferentially digest linker DNA ('naked' DNA) resulting in mono-, di-, tri- ... nucleosome DNA fragments whose length distribution is quantified by gel migration (Blank and Becker 1995). For a well-ordered primary structure, the digestion pattern

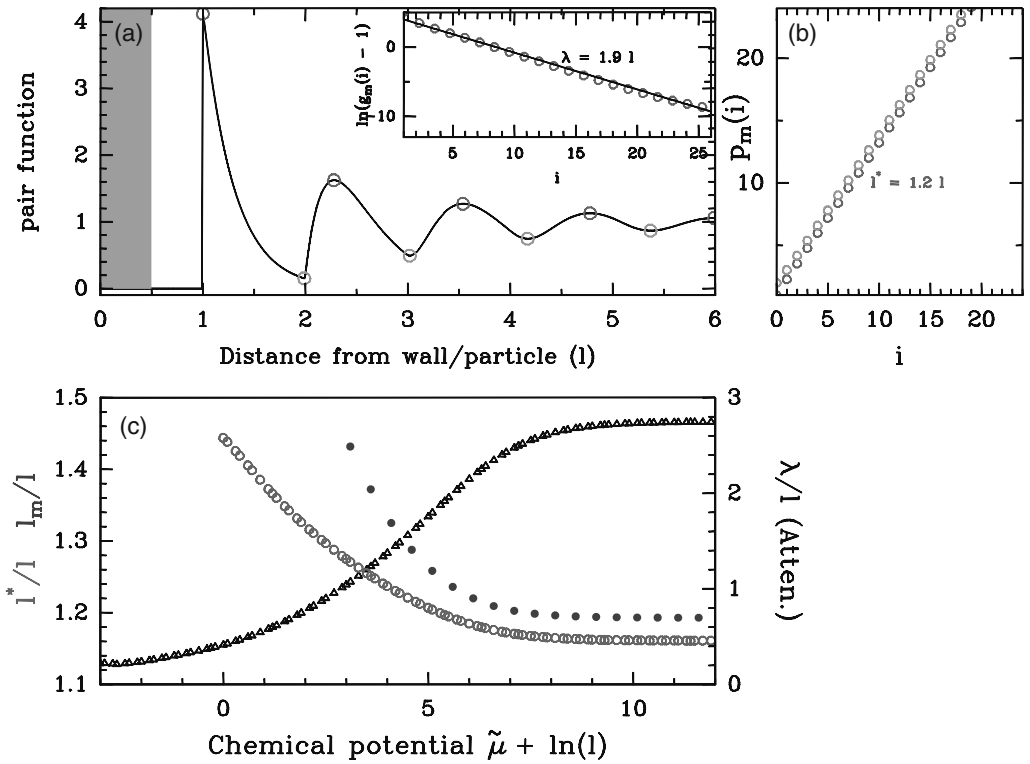

Figure 16. (a) Pair function $g(r)$ of a dense (uniform, $E(s) = E_o = c^{ste}$) hard-rod fluid. The residual chemical potential is $\tilde{\mu} = 0$ which gives a bulk density $\rho_b l = 0.78$. The statistical ordering of this dense phase is characterized by the period and the range of the oscillating pattern: (inset) ln $(g_m(i)l - 1)$ as a function of $p_m(i) - l$, where $g_m(i)$ is the value of the pair function at the ith extremum and $p_m(i)$ is the position of the ith extremum: maximum (red points) and minimum (green points) of the pair function. The linear regression gives a slope $-\lambda = -1.9l$, with λ defining the damping length. (b) Position $p_m(i)$ of the ith maxima (red) or minima (green) as a function of i. The linear regression gives the mean period of the oscillations $l^* = 1.2l$. (c) Evolution of the ordering range λ (black) and the mean period l^* (red) as a function of the chemical potential $\tilde{\mu}$. Both lengths are measured as explained in (a) and (b). In blue is reported the mean inter-nucleosome distance $l_m = \rho_b^{-1}$ as a function of $\tilde{\mu}$.

hence reveals regularly spaced gel bands (Figure 17a) corresponding to the migration of oligomers of quantized size $n \times l^*$. It is usually extracted from the intensity profile (Figure 17b) as the period of the oscillation (by a linear regression as illustrated in Figure 16). Formally this digestion profile is reminiscent of the pair distribution and the NRL coincides with the period l^* of this pair function (actually the MNase profiles are rather related to the pair function of linkers). As already mentioned, it depends on the organism and the cell type. It has been shown by *in vitro* experiments, supported by theoretical investigations, that the NRL plays a crucial role in the formation of higher-order chromatin structure (Bednar et al. 1998; Dorigo et al. 2004; Mergell et al. 2004; Lesne and Victor 2006; Robinson et al. 2006; Kepper et al. 2008; Depken and Schiessel 2009).

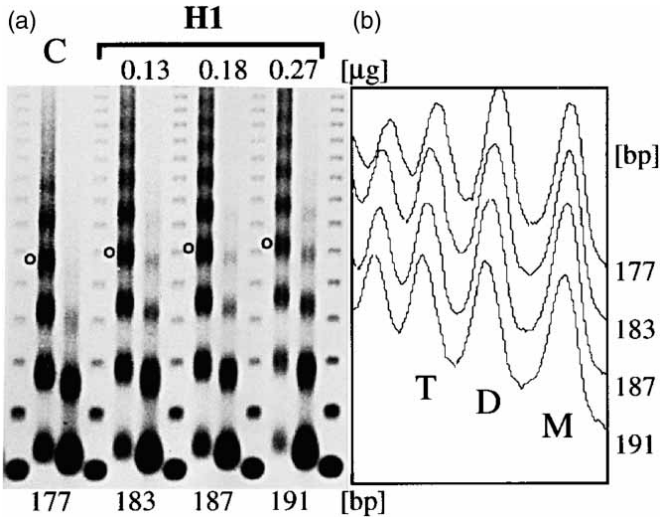

Figure 17. Gradual increase of NRL by histone H1. (a) Chromatin was assembled in histone depleted embryo extracts complemented with core histones and the indicated amounts of histone H1. (b) Plot profile of the first lane of each MNase digestion in (a). Peaks of mono- (M), di- (D) and tri-nucleosomes (T) are indicated. Adapted from Blank and Becker (1995).

Statistical ordering near an energy barrier

As already illustrated in Figure 13, a simple way to produce periodic positioning without any local 'positioning' signal (i.e. energy traps) is to introduce energy barriers, i.e. exclusion regions (Kornberg and Stryer 1988; Chevereau et al. 2009; Möbius and Gerland 2010; Vaillant et al. 2010). These inhibitory energy barriers can be encoded in the DNA sequence via unfavorable sequences that potentially resist to the structural distortions required for nucleosome formation (e.g. the presence of poly(dA:dT) (Bao et al. 2006; Field et al. 2008; Mavrich et al. 2008a; Kaplan et al.. 2009; Segal and Widom 2009b) or particular sequences that may

recruit transcription factors (Pusarla et al. 2007; Hartley and Madhani 2009; Kaplan et al. 2009) or/and other protein complexes such as chromatin regulators (Whitehouse and Tsukiyama 2006; Clapier and Cairns 2009; Teif and Rippe 2009) that may compete with the nucleosomes.

Infinite energy barriers

As shown in Figure 18(a), near an infinite energy barrier, when progressively increasing the chemical potential $\tilde{\mu}$, statistical ordering becomes more and more pronounced and manifests as periodic oscillations in the density (and occupancy) profile very similar to what we have observed in the pair function in Figure 15. Indeed, as explicitly used in subsection 'Inter-nucleosomal distance', the density $\rho(s)$ near an infinite wall [ρ_w, Equation (42)] takes the same form, up to a multiplicative factor, as the pair correlation function in an homogeneous landscape. Thus, the (positional) statistical ordering can again be characterized by the two characteristic lengths, the damping length λ and the period of the spatial modulation l^*.

Of course there is the question of the biological relevancy of such periodic ordering induced by infinite barriers. As a very convincing example of biological interest, we report in Figure 19(a), the experimental nucleosome positioning data around CTCF sites obtained by Fu et al. (2008) in human cells. The binding protein CTCF has been extensively studied for its impact on imprinting and X-inactivation (Lee 2003). It is known to bind to insulator elements to prevent heterochromatin spreading and may function as a transcriptional repressor or activator (Klenova et al. 1993; Burcin et al. 1997; Ohlsson et al. 2010). As shown in Figure 19(b), the remarkable nucleosome ordering observed on both sides of the CTCF bound proteins and that progressively vanishes for distances larger than 1 kbp is remarkably reproduced by the statistical positioning predicted by numerically solving the Percus equation [Equation (16)] in a flat energy landscape with an infinite energy barrier of width 240 bp positioned at the CTCF site. A good agreement is actually obtained for a chemical potential value $\tilde{\mu} = -2\,\mathrm{kT}$ which yields for the NRL a value $l^* = 185$ bp very close to the 190 bp estimated experimentally. Consistently, the predicted damping length accounts quite well for the exponential decay of the nucleosome ordering observed in the data.

Finite energy barrier

Inhibitory energy barriers encountered along eukaryotic chromosomes can be of variable shape. As we will discuss in the next section 'A sequence-dependent physical model of nucleosome occupancy', the ones that are encoded in the mechanical properties of the DNA double helices are typically of a few kT high. Their width is at most of the order of a nucleosome DNA wrapping length. So there is a need to quantify the effect of the energy barrier characteristics on the statistical positioning observed nearby these obstacles. In Figure 20, we summarize the numerical results obtained when solving the Percus equation with an energy barrier of variable height emerging at the center of a flat energy landscape. In agreement with the experimental observation in budding yeast (Figure 1), we fixed the barrier width to $w = 180$ bp (which allows a nuclosome to form for small barrier height as shown in Figure 20a). For the chemical potential value $\tilde{\mu} = -1, 0, 4$ and 10 kT, corresponding to high nucleosome bulk occupancy (0.74, 0.78, 0.83 and 0.84), we have computed the NRL l^* and the maximal density value ρ_w obtained near an energy barrier for different values of the barrier height ranging from 1 to 20 kT (Figure 20b). Let us first note that above 5 kT, l^* and ρ_w reach the asymptotic values of the infinite wall case. If the height of the energy barrier does not significantly affect the NRL, it has a clear influence on the maximum density ρ_w which increases between low (< 1 kT) and high (>5 kT) barrier height by 2, 2.5 and 3.5 fold for $\mu = -1, 0$ and >4 kT respectively.

Bistability induced by statistical confining in between two energy barriers

When analyzing the nucleosome occupancy profiles obtained *in vivo* in budding yeast (Figures 1 and 2), we realize that nucleosome depleted regions commonly called the Nucleosome Free Region (NFR) are distributed along the chromosomes with a mean separation distance $\sim$1–2 kbp (Chevereau et al. 2009; Vaillant et al. 2010; Arneodo et al. 2011). This observation raises the issue of the statistical confining of nucleosomes in between two inhibitory energy barriers that are separated by a distance L of a few nucleosome DNA wrapping lengths l. As illustrated in Figure 18(b) for the very simplified situation of a flat energy landscape bordered by two infinite energy barriers, when solving the Percus equation (16) for increasing values of the chemical potential $\tilde{\mu}$, we observe the establishment of a clear statistical ordering near the two bordering energy barriers, as previously shown in Figure 18(a), that progressively invades the system to transform into a clear periodic packing at high nucleosome (hard rod) density. The more nucleosomes are confined, the more they adopt a long-range and compact periodic organization with the inter-boundary distance L as a fundamental control parameter.

The problem of stacking hard rods of size $l = 146$ bp in a box of size L with infinite wall boundaries is actully easy to solve (Chevereau 2010). For a given (rather high) value of the chemical potential $\tilde{\mu}$, the theoretical weighted probability of a n-nucleosome configuration is shown in Figure 21(d) for various n-values as a function of the inter-barrier distance L. The nucleosome occupancy profile is then obtained as the weighted sum of each $n = 1, 2, \ldots$ nucleosome occupancy profiles (Figures 21a–c). Thus this theoretical

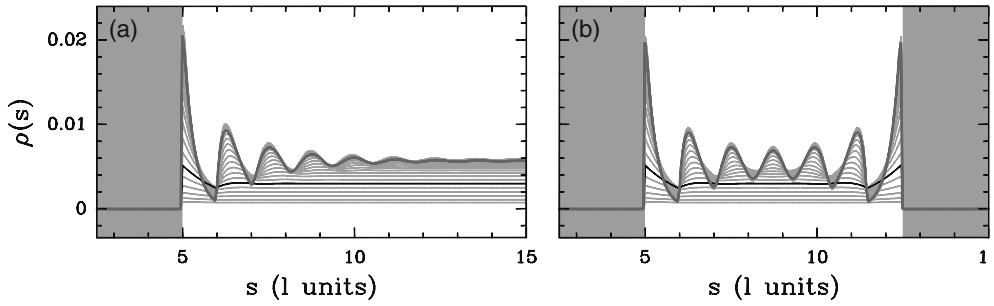

Figure 18. Evolution of the density profile $\rho(s)$ with the chemical potential $\tilde{\mu}$ from $\tilde{\mu} = -5\,\mathrm{kT}$ (black) to $+5\,\mathrm{kT}$ (red). Statistical confinement near an infinite wall (a) and in between two infinite energy barriers (b).

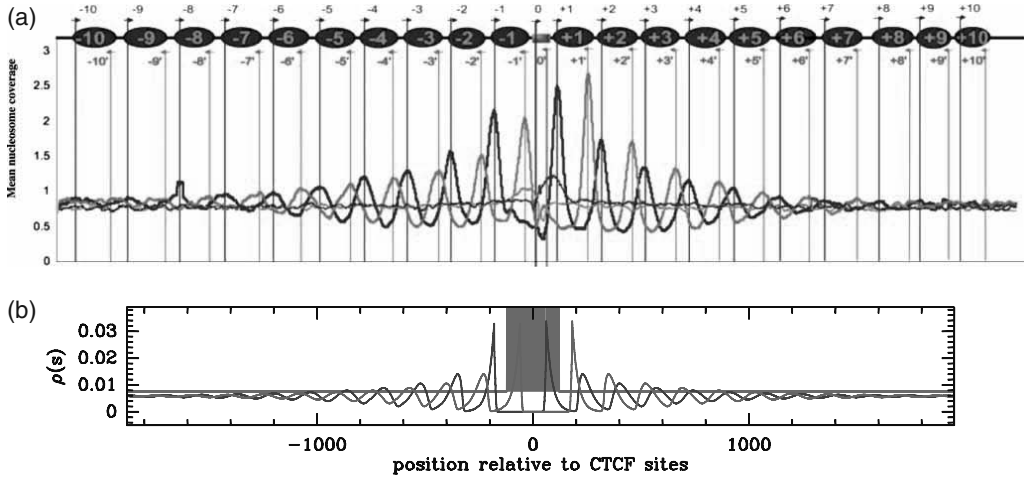

Figure 19. (a) Aggregation of nucleosome signals around CTCF sites from the experiment of Fu et al. (2008). The coordinate origin is set to the 5′ end position of the 20 bp-long CTCF sites. Schematic arrangement of nucleosomes (blue ovals) around a CTCF binding site (orange rectangle). Blue arrows indicate sequence tags on the same strand as the CTCF site (nucleosome 5′ extremity) and orange arrows indicate opposite-strand tags (nucleosome 3′ extremity). In green (resp. purple) are reported the 5′ (resp. 3′) extremity nucleosome counts in the absence of bound-CTCF. (b) Modelling of the data in (a) obtained by solving the Percus equation (16) in a flat energy landscape with an infinite energy barrier centered on the CTCF site and of width 240 bp (gray area) and a chemical potential value $\tilde{\mu} = -2\,\mathrm{kT}$.

situation predicts the existence of 'crystallization' domains that are characterized by a single dominating crystal n configuration (Figures 21a and c) with a NRL that increases with L over the range $l_{\min} < NRL \sim L/n < l_{\max}$ as shown in Figure 22. Importantly, this model also predicts that in between the crystal n (Figure 21a) and $n+1$ (Figure 21c) domains, there exists a coexistence domain where these two (or more) crystalline configurations contribute statistically to an apparently irregular occupancy profile (Figure 21b). As clearly seen on the NRL in Figure 22, as the inter-barrier distance L increases, the extent of crystal domains is expected to decrease to the benefit of the bistable (or multistable) coexistence domains. At very large L, periodic packing is lost at the center of the system where the nucleosome positioning is no longer dictated by the long-range influence of the bordering infinite walls. The NRL l^* is no more L-dependent and tends to the value of the uniform system (Figure 16c) at the chemical potential $\tilde{\mu}$.

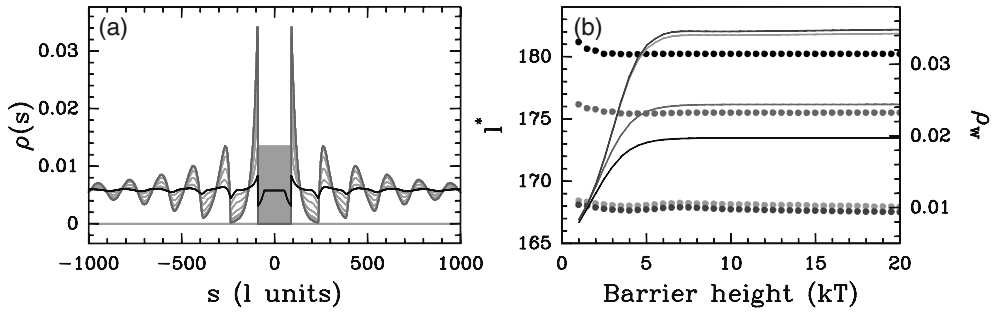

Figure 20. Statistical periodic ordering observed near an energy barrier as a fuction of the barrier height. (a) Density profiles obtained by solving the Percus equation (16) in a flat energy landscape with a finite energy barrier of width $l = 180\,\mathrm{bp}$ centered at $s = 0$ (gray area) and a chemical potential value $\tilde{\mu} = 4\,\mathrm{kT}$. The profiles correspond to barrier heights ranging from 1 kT (black) to 20 kT (red). (b) NRL l^* (dots) and wall density value $\rho(s = \pm 90) = \rho_w$ (curves) as a function of the barrier height, extracted from the density profiles (see (a)) at different values of the chemical potential : $\mu = -1\,\mathrm{kT}$ (black), $\mu = 0\,\mathrm{kT}$ (red), $\mu = 4\,\mathrm{kT}$ (green) and $\mu = 10\,\mathrm{kT}$ (blue).

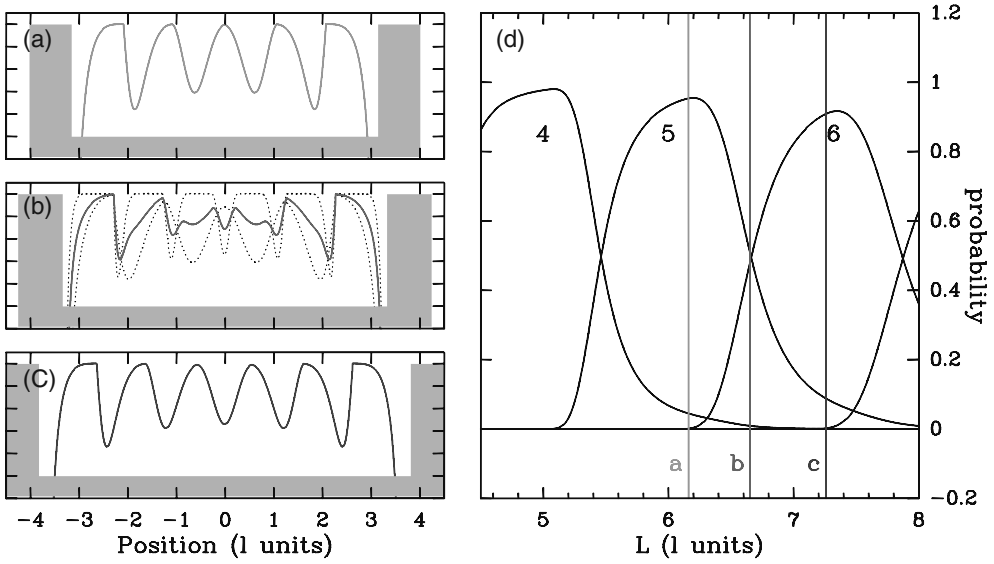

Figure 21. Theoretical probability of nucleosome occupancy at each point of a box of size L bordered by two infinite walls. (a) Box large enough to shelter five nucleosomes (green). (b) Larger box where the two dotted configurations are possible; the weighted average of the 5 and 6 nucleosome crystal-like profiles yields an irregular-looking average profile (red). (c) Larger box where six nucleosomes can be inserted without being tightly packed. (d) Probability of crystal configurations with a fixed number n of nucleosomes with respect to the box size L. Vertical colored lines correspond to the inter-barrier distances L used respectively in (a), (b) and (c). While only one configuration has clearly the highest probability for (a) and (c), two configurations are equally probable in (b), which justifies the superposition. The distances are expressed in nucleosome length units (hard core length l).

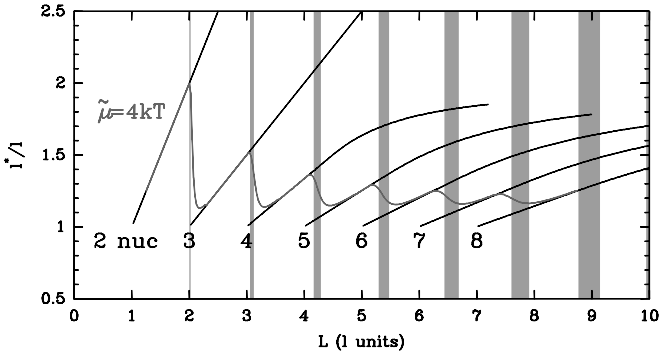

Figure 22. Theoretical NRL l^* dependency on the box size L (see Figure 21); black dotted lines correspond to a fixed number n of nucleosomes and the red lines to the NRL size at a given chemical potential $\tilde{\mu} = 4\,\mathrm{kT}$. Vertical gray shaded bands correspond to the bistable domains. The distances are expressed in nucleosome length l units.

Altogether, these theoretical results demonstrate the crucial role of the inter-barrier distance L as a fine control of chromatin structure.

As pointed out in a previous study (Chevereau et al. 2009; Vaillant et al. 2010; Arneodo et al. 2011), in *S. cerevisiae*, NFR are mainly observed *in vivo* at transcription start sites (TSS) and transcription termination sites (TTS). Whatever the origin of the underlying energy barriers, this observation suggests that the chromatin organization inside budding yeast genes is susceptible to being described by a phenomenological model as simple as the one previously studied in Figures 21 and 22. As reported in Figure 23, when ordering budding yeast genes by the distance L that separates the first ($5'$) and last ($3'$) nucleosome, we obtain a 2D map that reveals a striking organization of the nucleosomes' distribution inside the genes. (Note that we use L as a substitute of the distance $\mathcal{L} \sim L + 188$ between the $5'$ and $3'$ NFRs which is more difficult to measure accurately because of the NFR shape variability.) Small genes ($L < 1.5$ kbp) present a clear periodic packing in between the two bordering NFRs with a well-defined number n of regularly spaced nucleosomes (Figure 23c). As the interdistance L increases, these 'crystallized' genes cluster into L-domains with the gene having the same number of nucleosomes, for $n = 2$ to about 9 nucleosomes. For rather large gene sizes ($L > 1.5$ kbp), the nucleosome positioning appears periodic essentially at the two boundaries and fuzzy in the middle where the confinement induced by both boundaries is too weak to constrain the positioning of the central nucleosomes (Figure 23a). This intra-genic nucleosome organization is totally consistent with the statistical ordering mechanism induced by exclusion from the boundaries except that to quantitatively account for the *in vivo* data in Figures 23(b) and (c), we had to consider in our theoretical modelling finite-size linear energy barriers of height $E_M = 6\,\mathrm{kT}$ and width $\Delta = 80$ bp and a chemical potential value $\tilde{\mu} = 1\,\mathrm{kT}$ so that the nucleosomes cover 75% of the yeast genome as observed *in vivo* (see Vaillant et al. 2010, for more technical details). This peculiar linear shape of the bordering energy barriers actually amounts to imposing a constant force $F = E_M/\Delta = 6\,\mathrm{kT}/27.2\,\mathrm{nm} \sim 1$ pN on both sides of the intra-genic nucleosome array. Note that 1 pN is comparable and actually a little less than the few piconewton tensions generated by elongating polymerases (Wang et al. 1998; Hall et al. 2009) and helicases (Strick et al. 2003; Lionnet et al. 2006) suggesting that these enzymes can drastically affect the nucleosome ordering observed inside yeast

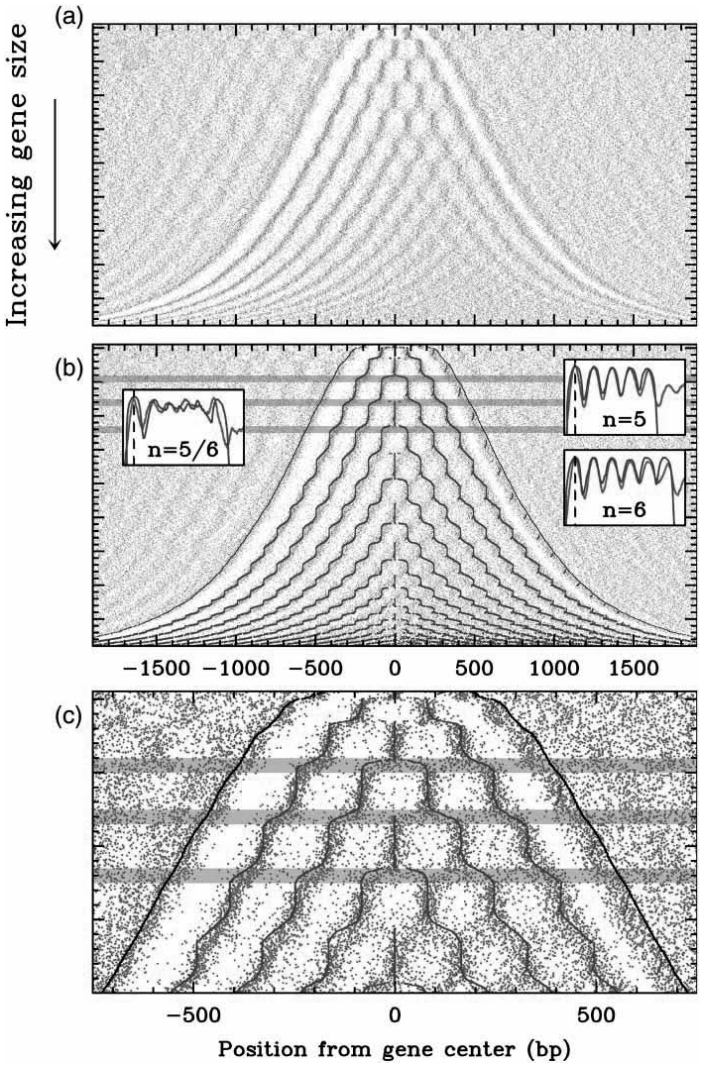

Figure 23. 2D-map of nucleosomes along budding yeast genes. (a) The 4554 genes are ordered vertically by the distance L between the first ($5'$) and last ($3'$) nucleosomes. The nucleosome occupancy profile of each gene is figured along a horizontal line: red dots correspond to the minima of nucleosome occupancy; nucleosomes occupy the white zones; *in vivo* data are retrieved from Lee et al. (2007). (b) Predictions of our theoretical modelling (blue) with fixed force boundary energy barriers (see text and Vaillant et al. (2010)) drawn on top of experimental data (red). Insets: mean experimental (red) and theoretical (blue) nucleosome occupancy profiles for crystal genes harboring 5 nucleosomes (right, top), 6 nucleosomes (right, bottom) and for bi-stable genes with 5/6 nucleosomes. (c) Zoom on the first 2000 genes in (b); gray-shaded areas correspond to some bi-stable L-domains. In (b) and (c), the black curves indicate the $5'$ and $3'$ end positions of the theoretical excluding nucleosome energy barriers.

genes. Hopefully, probably thanks to chromatin remodellers that are found all over yeast chromosomes likely increasing the effective temperature (Rippe et al. 2007), equilibrium statistical nucleosome ordering will be recovered along most genes with a characteristic time much shorter than the typical time separating the successive chromatin alterations induced by elongation. In a work under progress, we are revisiting the experimental *in vitro* nucleosome positioning data of Zhang et al. (2011) under the scope of our theoretical modelling with the idea that the 1 pN confining force previously found to account for the *in vivo* intra-genic nucleosome organization might well result from active remodelling at the $5'$ end of most yeast genes.

But what is remarkable in the results reported in Figure 23 (Chevereau et al. 2009; Vaillant et al. 2010) is the functional implications of the intra-genic chromatin structure of yeast genes. In agreement with the predictions of our thermodynamical modelling, we have been able to identify two main classes of *in vivo* intra-genic nucleosome organizations: crystal-like genes with regularly positioned nucleosomes and bi-stable genes with rather irregular nucleosome occupancy profiles resulting from the coexistence of two possible crystal-like states with different compaction levels, a weakly compacted n nucleosome state and a highly compact $n+1$ nucleosome state. As compared to crystal-like genes that present a constitutive expression level, bi-stable genes show a higher transcriptional plasticity and are more sensitive to chromatin regulators. Indeed, by means of a single nucleosome switching, bi-stable genes may drastically alter their expression level in response to external changes. In that context, a less intuitive result is the fact that the transcription rate tends to increase when the NRL decreases, so when the linear compaction level increases. A very similar trend has been also observed recently in human cells by Valouev et al. (2011). Several possible interpretations of this coupling between intra-genic chromatin and polymerase elongation process have been proposed (Vaillant et al. 2010) including the fact that a short linker size would rather lead to a loose 30 nm fiber favoring the accessibility and sequential action of components of the elongation machinery (Lesne and Victor 2006).

The nucleosome structure of promoters and its implications in transcription initiation has been studied in various organisms from yeast to human including the nematode and Drosophila (Bernstein et al. 2004; Lee et al. 2004; Yuan et al. 2005; Lee et al. 2007; Ozsolak et al. 2007; Mavrich et al. 2008a; Miele et al. 2008; Shivaswamy et al. 2008; Tirosh and Barkai 2008; Valouev et al. 2008; Arneodo et al. 2011). In Chevereau et al. (2009), and Vaillant et al. (2010), we have identified a new paradigm of transcriptional control mediated by the stability and the level of compaction of the intra-genic chromatin architecture. To what extent these chromatin mediated regulation processes generalize to other eukaryotic species is a very challenging question for future experimental and theoretical studies.

A sequence-dependent physical model of nucleosome occupancy

As shown previously with CTCF (subsection 'Infinite energy barriers', Figure 19), *in vivo* nucleosome occupancy and positioning can be locally controlled by external factors: the stable binding of proteins or protein complexes at a genomic locus act as an 'effective' vertical barrier that induces stretches of periodically distributed nucleosomes in agreement with a statistical ordering principle. 'Chromatin regulators' such as chromatin remodellers have

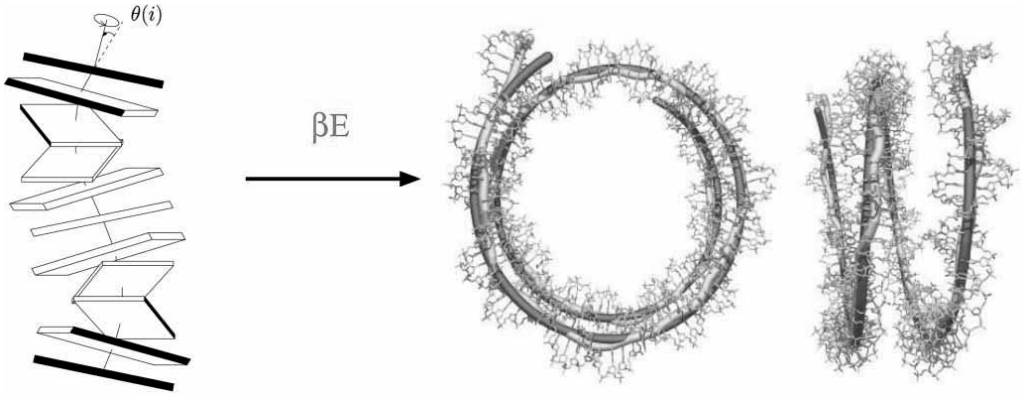

Figure 24. Our physical modelling consists of computing the energy cost to bend a DNA fragment of length l_w into almost two turns of the DNA double helix which are involved in the crystallized nucleosome particle (radius $R = 4.19$ nm, pitch $P = 2.59$ nm). Adapted with permission from Richmond and Davey (2003). Copyright 2003 by Nature Publishing Group.

a fundamental role in shaping *in vivo* primary structure. Some remodellers, in combination with DNA binding factors or/and chromatin modifiers, are targeted to specific genomic loci or histone epigenomic marks, to actively eject [e.g. RSC (Hartley and Madhani 2009; Wang et al. 2011b)] or maintain [e.g. RSC (Moreira and Holmberg 1999; Floer et al. 2010), Mit1 (Garcia et al. 2010)] nucleosomes. Besides these 'extrinsic' regulations of nucleosome positioning/occupancy, there may also exist an 'intrinsic' contribution from the sequence-specificity of histone–DNA interactions. The question of the sequence-specificity and of the role it plays *in vivo* has been extensively studied for the last 30 years. However there is still no consensus on it. According to recent reports (Peckham et al. 2007; Yuan and Liu 2008), in *S. cerevisiae*, no more than 20% of the *in vivo* nucleosome positioning above what is expected by chance is determined by intrinsic signals in the genomic DNA. As reported in '*In vivo* and *in vitro* genome-wide primary structure of chromatin', the recent availability of *in vivo* and *in vitro* genome wide nucleosome occupancy data (Figures 1, 2 and 10) (Lee et al. 2007; Kaplan et al. 2009) has led to a renewal of interest in modelling nucleosome organization along the 10 nm chromatin fiber. Thus a model of DNA-sequence dependent nucleosome positioning based on statistical learning (Field et al. 2008; Kaplan et al. 2009) was shown to be significantly predictive of the nucleosome organization *in vivo* in budding yeast as well as in other organisms like fly and human. In the spirit of the physical modelling developed in an early work (Vaillant et al. 2005, 2006), we have recently proposed a model which is based on the sequence-dependent DNA bending properties (Vaillant et al. 2007; Chevereau et al. 2009; Milani et al. 2009) and, as reported in this section, which performs as well as models based on statistical learning.

'Intrinsic' nucleosome formation energy landscape

To compute the free energy landscape associated with the formation of one nucleosome at a given position s along DNA, we will assume that (Vaillant et al. 2007) (i) DNA is an unshearable elastic rod whose conformations are described by the set of three local angles $\Omega_1(s)$ (tilt), $\Omega_2(s)$ (roll), $\Omega_3(s)$ (twist), and (ii) the DNA chain along the nucleosome at position s is constrained to form an ideal superhelix of radius $R = 4.19$ nm and pitch $P = 2.59$ nm as observed in the X-ray crystallographic nucleosome structure (Luger et al. 1997; Richmond and Davey 2003) over a total length l_w which fixes the distribution of angular deformations $(\Omega_i^{\mathrm{nuc}}(u-s))_{i=1,2,3}, u = s, \ldots, s+l$ (Figure 24). Within linear elasticity approximation, the energy cost for nucleosome formation is given by:

$$\beta E(s, l_w) = \int_s^{s+l_w} \sum_{i=1}^{3} \frac{A_i}{2} (\Omega_i^{\mathrm{nuc}}(u-s) - \Omega_i^o(u))^2 \, \mathrm{d}u, \tag{45}$$

where A_1, A_2 and A_3 are the stiffnesses associated with the tilt, roll and twist deformations around their intrinsic values Ω_1^o, Ω_2^o and Ω_3^o, respectively. Consistently with our previous works (Audit et al. 2002; Vaillant et al. 2005, 2006, 2007), we will use here the 'Pnuc' structural bending table (Goodsell and Dickerson 1994) which is mainly a trinucleotide roll coding table (Ω_2^o), with zero tilt ($\Omega_1^o = 0$) and constant twist ($\Omega_3^o = 2\pi/10.5$). Since the values of this bending table were arbitrarily assigned between 0 and $\pi/18$, we have performed the following affine rescaling $\Omega_2^{*o} = \gamma(\Omega_2^o - \eta)$ with $\eta = 0.15$ and γ a tuning parameter that controls the fluctuation range $\delta = \langle (E - \overline{E})^2 \rangle^{1/2}$ of the energy landscape. For example, for yeast, we have fixed $\gamma = 0.4$ which amounts to imposing $\delta = 2$ kT, a value which allowed us to get comparable overall nucleosome occupancy distributions as observed both *in vitro* and *in vivo* (Vaillant et al. 2007).

In Figure 25 is shown the theoretical nucleosome occupancy profile obtained along the yeast chromosome 2 when fixing the model parameters to provide a very good match, at a statistical level, with the experimental *in vitro* (Kaplan et al. 2009) and *in vivo* (Lee et al. 2007) data. The 2D map in Figure 25(a) actually represents the evolution of the nucleosome occupancy probability $P(s)$ when increasing the mean residual chemical potential $\tilde{\mu}$. In Figure 25(b)

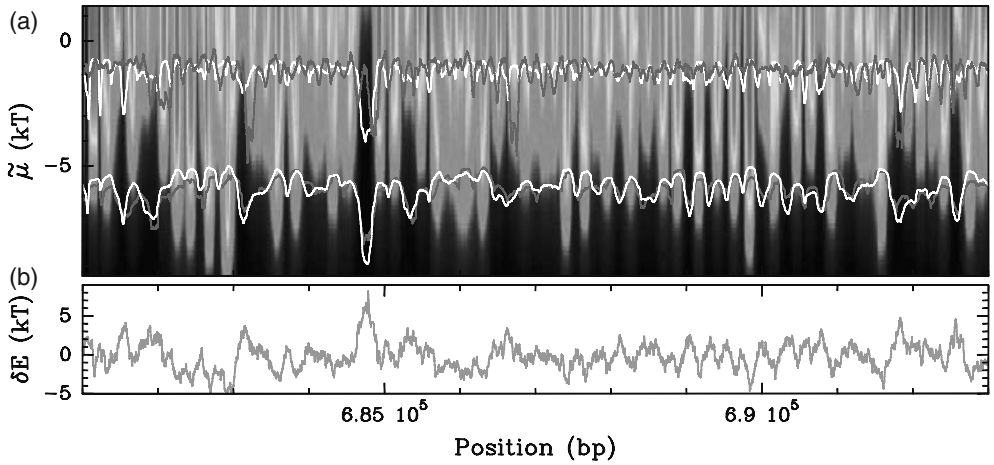

Figure 25. (a) 2D map representing the theoretical nucleosome occupancy probability $P(s)$ [Equation (37)] along a 12 kbp long fragment of the budding yeast chromosome 2 as a function of the residual chemical potential $\tilde{\mu} = \mu - \overline{E}$ (Chevereau et al. 2009): dark blue corresponds to low probability and red to high probability. The two white occupancy profiles are the theoretical profiles obtained for $\tilde{\mu} = -6\,\mathrm{kT}$ and $-1.3\,\mathrm{kT}$ that correspond to a genome nucleosome coverage of 30% and 75% as observed *in vitro* (Kaplan et al. 2009) and *in vivo* (Lee et al. 2007) respectively; the corresponding *in vitro* and *in vivo* experimental nucleosome occupancy profiles are shown in red for comparison. (b) The corresponding energy landscape $E(s)$ computed with the following parameter values: $\delta = \langle (E - \overline{E})^2 \rangle^{1/2} = 2\,\mathrm{kT}$ and $l_w = 125$ bp (see text).

is shown the predicted energy landscape when fixing the effective nucleosome wrapping length to $l_w = 125$ bp, a value which is smaller than the typical and well accepted 146 bp nucleosomal DNA length (Luger et al. 1997). This suggests that sequence-specificity is effectively dominated by the wrapping around the H3–H4 tetramer. Note that the hard-core length l that we consider for the computation of the nucleosome density remains $l = 146$ bp. Interestingly, Figure 25(a) enlightens the fundamental role of the energy landscape and its topography (amplitude, size and distribution of favorable and unfavorable regions) that entirely control the fluctuations in the nucleosome occupancy profile but in a non-trivial (nonlinear and non-local) manner that depends on the chemical potential.

As shown in Figure 26, the histone octamer sequence specificity can be estimated at a low value of the chemical potential $\tilde{\mu}$ $(= -6\,\mathrm{kT})$ for a diluted system where the ratio of the nucleosome densities at two different points s_1 and s_2 (here separated by 1 kbp) is given by $\rho(s_1)/\rho(s_2) \simeq \mathrm{e}^{-\beta(E(s_1)-E(s_2))}$. For a more concentrated system at a higher value of $\tilde{\mu}$ $(= 0\,\mathrm{kT})$, the difference in energy $\Delta E_{12} = E(s_1) - E(s_2)$ is no longer sufficient to specify the density difference between the two points as the consequence of the interactions between particles. Thus at high nucleosome density, a same nucleosome formation energy can lead to very different nucleosome densities due to different energetic environments.

The main feature in the nucleosome occupancy heat map shown in Figure 25(a) is the fact that the highest energy barriers present in the energy landscape (Figure 25b) correspond to regions that are robustly depleted in nucleosomes whatever the overall nucleosome density. At low density (low $\tilde{\mu}$ values), the confinement is weak and the nucleosomes distribute everywhere in between the highest energy barriers according to the energy landscape fluctuations.

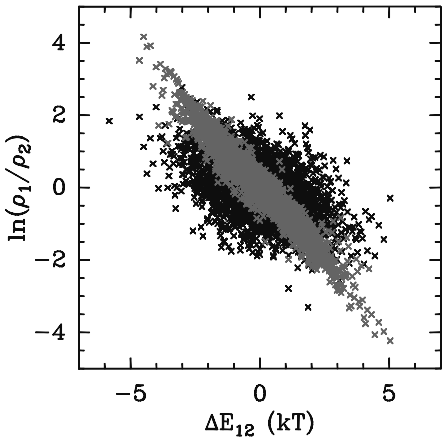

Figure 26. $\ln(\rho(s_1)/\rho(s_2))$ versus $\Delta E_{12} = E(s_1) - E(s_2)$, where $\rho(s_1)$ (resp. $\rho(s_2)$) is the nucleosome density (computed as explained in the text, from the budding yeast genome) and $E(s_1)$ (resp. $E(s_2)$) the nucleosome formation energy at the position s_1 (resp. s_2). The crosses correspond to two statistical samples in a diluted ($\tilde{\mu} = -6\,\mathrm{kT}$, red) and dense ($\tilde{\mu} = 0\,\mathrm{kT}$, black) non-uniform fluid.

When the nucleosome density is increased, ordering progressively appears leading to an overall organization with 'crystal-like' phases of regularly positioned nucleosomes confined near or in between excluding energy barriers, coexisting with 'fluid-like' phases where ordering is lost, in agreement with the statistical ordering principles described in the section 'Statistical positioning'.

Remark: As illustrated in Figure 25(b) and in Figure 43, see later on, the energy landscape computed in budding yeast presents a disordered topography as the consequence of the 'disordered' organization of the underlying genome; similar genome wide behavior is observed for Hemiascomyta yeasts as well as for *S. pombe*. As we shall see in the concluding section, at the genome scale, this genomic 'disorder' is characterized by (almost) gaussian statistics

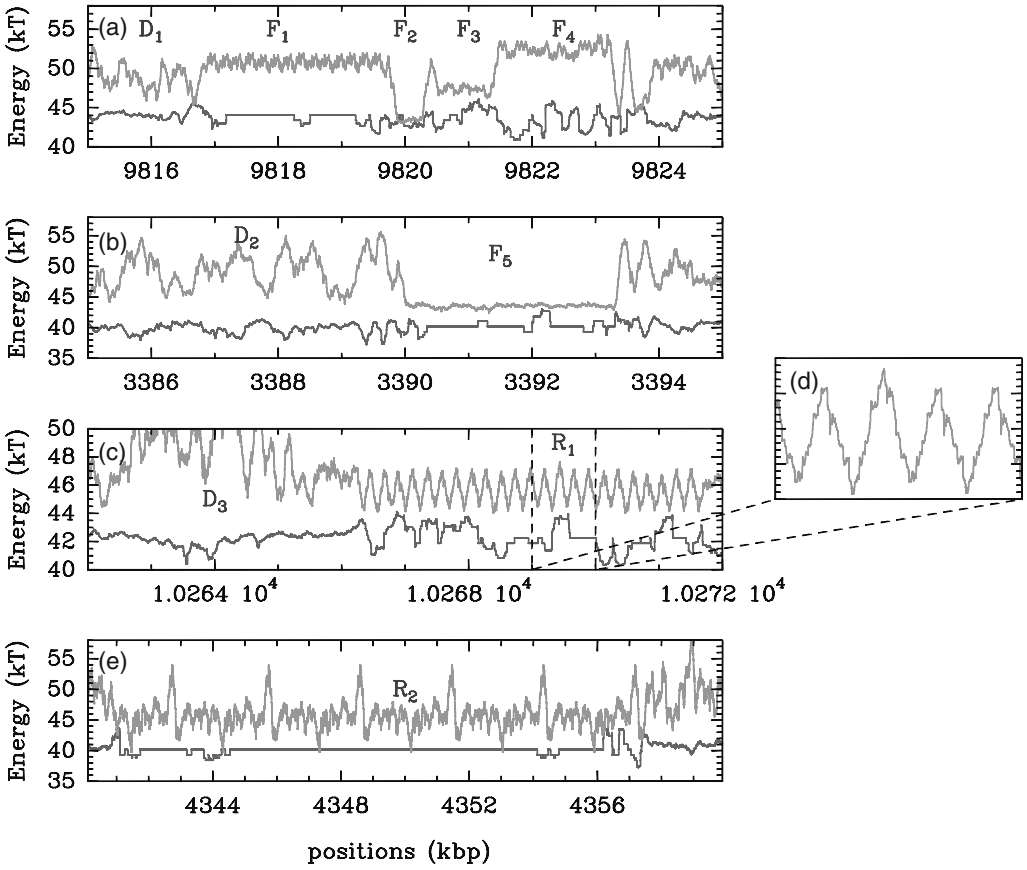

Figure 27. Energy landscape $E(s)$ computed with the following parameter values: $\delta = \langle (E - \overline{E})^2 \rangle^{1/2} = 2\,\mathrm{kT}$ and $l_w = 125\,\mathrm{bp}$ along 10 kbp fragments of the chromosome 1 of *C. elegans* (green). Disordered patterns (regions D_1, D_2 and D_3) are alternating with regular patterns, either quasi-flat (regions F_1, F_2, F_3, F_4 and F_5) or periodic with periodic stretches of barriers/wells (R_1 and R_2). In red are reported the experimental occupancy data $\delta Y(s)$ of Valouev et al. (2008).

and long-range correlations. However, besides this disordered organization, genomes of higher eukaryotes are largely composed by repeated sequences (Richard et al. 2008) that lead to 'periodic'/regular patterns in the energy landscape as shown in Figure 27 for *C. elegans*. The shape of the regular patterns depends on the repeated motif (in particular the motif's size) and interestingly, for short repeats ($<$10 bp) the $E(s)$ profile can become almost flat (see part F_5 in Figure 27b). However, due to their periodic nature, current biochemical mapping methods cannot provide any reliable chromatin pattern in these regions (see parts F_1, F_5 and R_2 in Figure 27).

Modelling of in vitro *nucleosome occupancy data in* S. cerevisiae

As shown in Figure 28 for 10 kbp fragments of various *S. cerevisiae* chromosomes, when adjusting the chemical potential $\tilde{\mu} = -6\,\mathrm{kT}$ to obtain the nucleosome density (30%) observed *in vitro* by Kaplan et al. (2009) (subsection '*In vitro* nucleosome occupancy profiles'), we get nucleosome occupancy profiles that reproduce quite impressively the data (Chevereau et al. 2009). The mean Pearson correlation computed along the 12 Mbp of the budding yeast genome is $\overline{r} = 0.74$, a result which is as good as the correlation value $\overline{r} = 0.74$ (resp. $\overline{r} = 0.89$) obtained with the Field et al. (2008) (resp. Kaplan et al. 2009) model based on statistical learning. Furthermore, this very satisfactory mean Pearson correlation value really reflects the pertinence and consistency of our physical model all along the *S. cerevisiae* chromosomes (Chevereau et al. 2009). As shown in Figure 29, the histogram of correlation values computed in 1 kbp sliding windows along the entire genome is mainly concentrated over a range $0.7 < r < 1$, with a well-defined maximum for a value as high as $r = 0.85$. For the sake of comparison, we have also reported the histogram of correlation values obtained between the predictions of our physical model and those of the Field et al. (2008) statistical model. This histogram is even more concentrated at very large r values with a rather sharp maximum for $r = 0.92$. This brings the demonstration that our model based on the structural and mechanical properties of the DNA double helix performs as well as rather sophisticated models requiring statistical learning (Field et al. 2008; Kaplan et al. 2009; Tillo and Hughes 2009). This is confirmed in Figure 30 where the experimental *in vitro* distribution of nucleosome occupancy values and auto-correlation function (Figures 10c and d) are quite well reproduced by our physical model. Importantly as observed in the data (Figure 10d), our physical model for low $\tilde{\mu}$ predicts (as dictated by the energy profile) no oscillatory modulation of the (power-law) decay of the auto-correlation function of the nucleosome occupancy

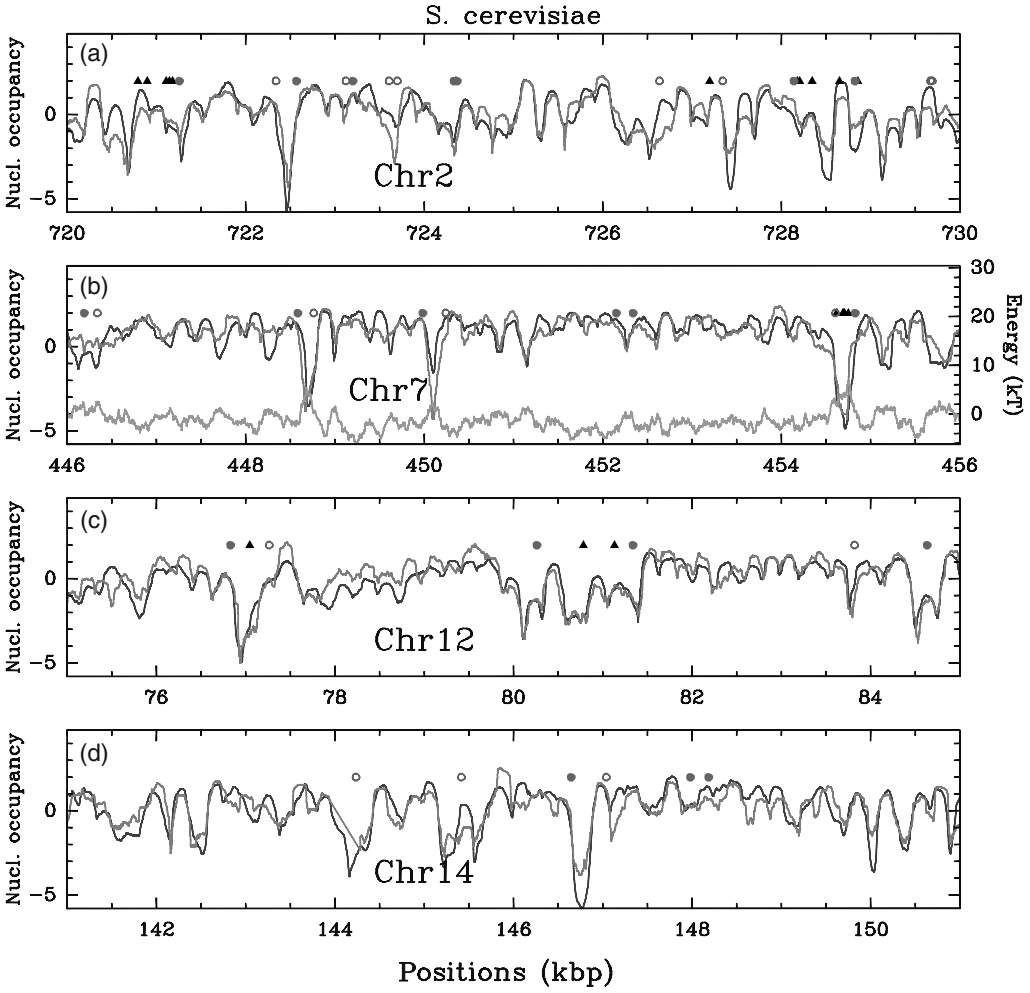

Figure 28. Comparison between the experimental occupancy profile from the *in vitro* MNase-seq experiment of Kaplan et al. (2009) (orange), the theoretical low-density occupancy profile (blue) and the energy landscape (green) (subsection "Intrinsic' nucleosome formation energy landscape') over regions of 10 kbp of several *S. cerevisiae* chromosomes. The theoretical predictions were obtained with the following parameter values: $\tilde{\mu} = -6\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$ and $l_w = 125\,\mathrm{bp}$.

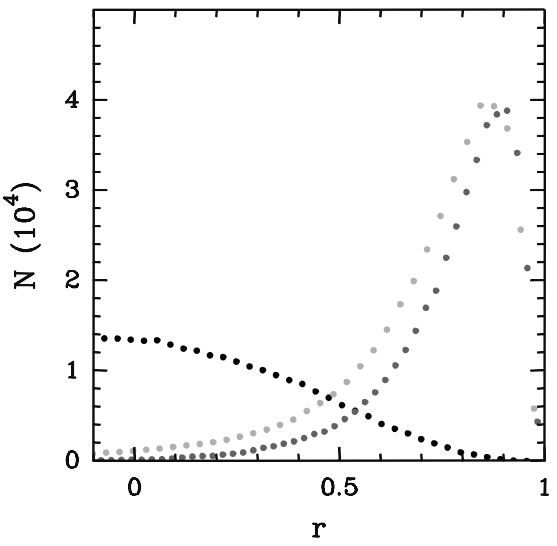

Figure 29. Histograms of Pearson correlation values r as measured in a 1 kbp sliding window between our physical modelling ($\tilde{\mu} = -6\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$ and $l_w = 125\,\mathrm{bp}$) and the Kaplan et al. *S. cerevisiae in vitro* MNase-seq data (Kaplan et al. 2009) (light blue), Field et al. statistical model (Field et al. 2008) (pink) and a random occupancy landscape (black).

thereby confirming the absence of statistical ordering at low nucleosome density.

Remark: Let us mention that other attempts to describe nucleosome positioning in *S. cerevisiae* based on sequence-dependent nucleosome score (N-score) models built from the learning on two training sets of sequences, one corresponding to nucleosome sequences and the other one to linker sequences, do not provide competitive predictions (Peckham et al. 2007; Yuan and Liu 2008). For example, the Peckham et al. model performance for Kaplan et al. *in vitro* MNase-seq data is rather modest as quantified by a mean Pearson correlation $\overline{r} = 0.48$. Concerning other physical models proposed using either different di- or tri-nucleotide coding tables (Miele et al. 2008) rather than the PNuc coding table used here or constructed *ab initio* (Tolstorukov et al. 2007; Morozov et al. 2009; Tolkunov and Morozov 2010), the obtained performances are even poorer, e.g. $\overline{r} = 0.38$ for the Miele et al. model, and $\overline{r} = 0.01$ for the Tolstorukov et al. model. However, very recent physical modelling based on a new *ab initio* computation of DNA

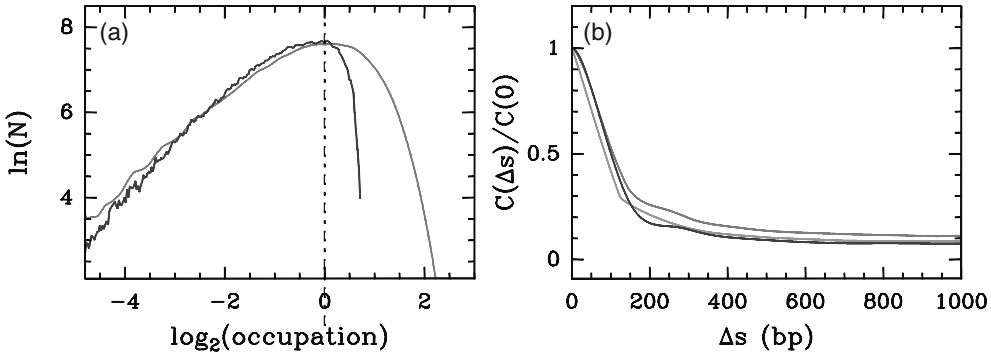

Figure 30. Comparing the predictions of our physical modelling ($\tilde{\mu} = -6\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$ and $l_w = 125\,\mathrm{bp}$) with the Kaplan et al. *S. cerevisiae in vitro* MNase-seq data (Kaplan et al. 2009). (a) Histograms of nucleosome occupancy $Y(s)$ values centered at their typical value: model (blue), *in vitro* data (orange). (b) Corresponding auto-correlation function $C(\Delta s) = \langle \delta Y(s) \delta Y(s + \Delta s) \rangle$; the green curve corresponds to the auto-correlation function of the theoretical nucleosome formation energy profile (see the chromosome 7 panel in Figure 28).

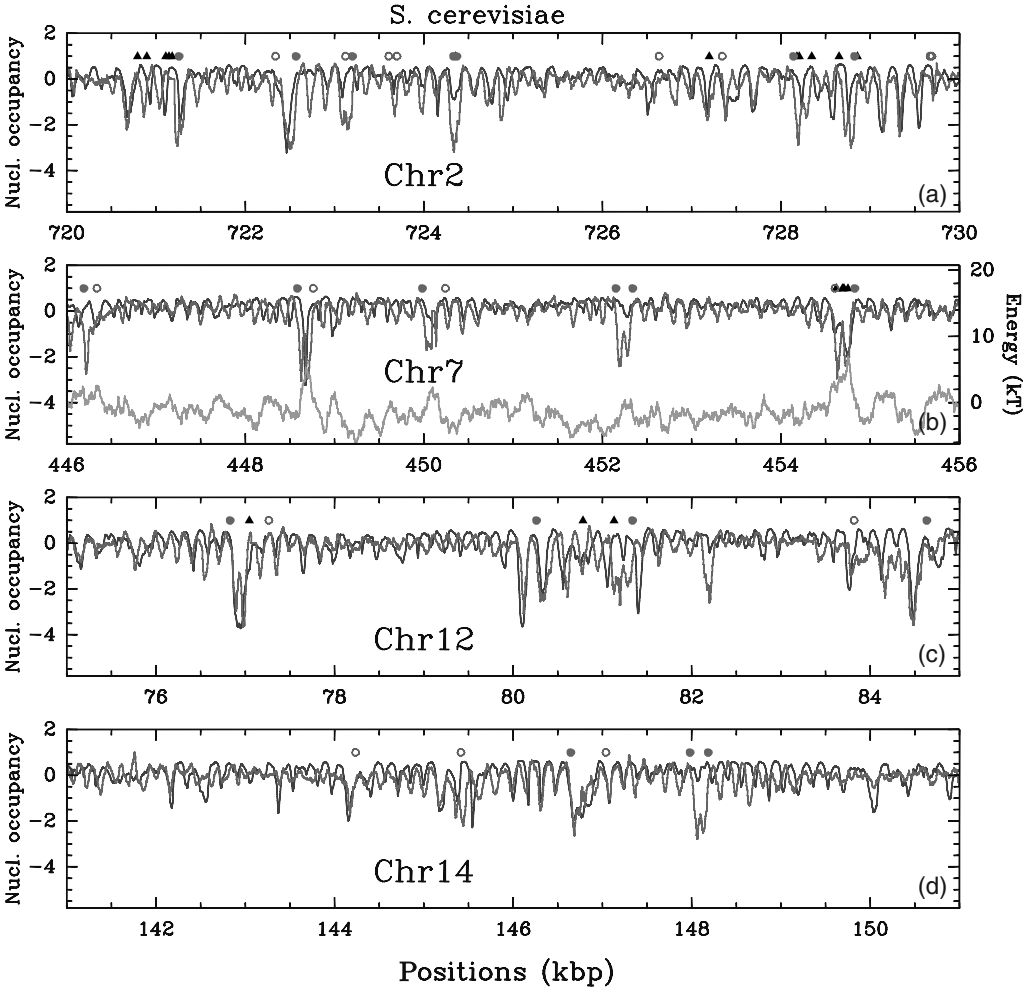

Figure 31. Comparison between the experimental occupancy profile from the *in vivo* MNase-chip experiment of Lee et al. (2007) (red), the theoretical high-density occupancy profile (blue) and the energy landscape (green) (subsection "Intrinsic' nucleosome formation energy landscape') over regions of 10 kbp of several *S. cerevisiae* chromosomes. The theoretical predictions were obtained with the following parameter values: $\tilde{\mu} = -1.3\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$ and $l_w = 125\,\mathrm{bp}$.

sequence-dependent elasticity provides a very good match at TSS and TTS of yeast genes (Deniz et al. 2011).

Modelling of in vivo *nucleosome occupancy data in* S. cerevisiae

As shown in Figure 31 for various budding yeast chromosomes (same 10 kbp contigs as in Figure 28), when increasing the chemical potential $\tilde{\mu}$ ($= -1.3\,\mathrm{kT}$) to reach the *in vivo* nucleosome density (75%), our physical model predicts nucleosome occupancy profiles that are still in good agreement with the experimental data (Chevereau et al. 2009). However, as reported in Figure 32(a), the histogram of Pearson correlation values is significantly shifted to lower values as compared to the one previously obtained at lower (*in vitro*) nucleosome density in Figure 29, with a mean value $\overline{r} = 0.33$ and a rather wide support. The weakest correlations observed with our model are also shared by other

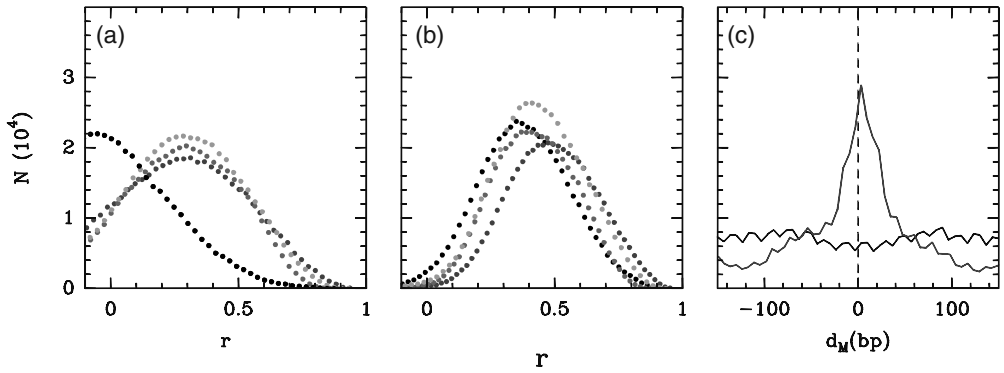

Figure 32. Histograms of Pearson correlation values r between the Lee et al. *S. cerevisiae in vivo* MNase-chip data (Lee et al. 2007) and our physical modelling ($\tilde{\mu} = -1.3\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$, $l_w = 125\,\mathrm{bp}$) (blue), Yuan and Liu (2008) model (pink) and a random occupancy landscape (black). The Pearson correlation was measured in a 1 kbp sliding window over the 16 yeast chromosomes: (a) no shift $d = 0$ between the theoretical and experimental signal; (b) for a shift d_M that maximizes the correlation; (c) histogram of optimal shift d_M values. In (a,b), the green dots correspond to the histogram of Pearson correlation values obtained between the *in vivo* data and the theoretical nucleosome formation energy profile (actually with the affinity $-E(s)$) (see the chromosome 7 panel in Figure 31).

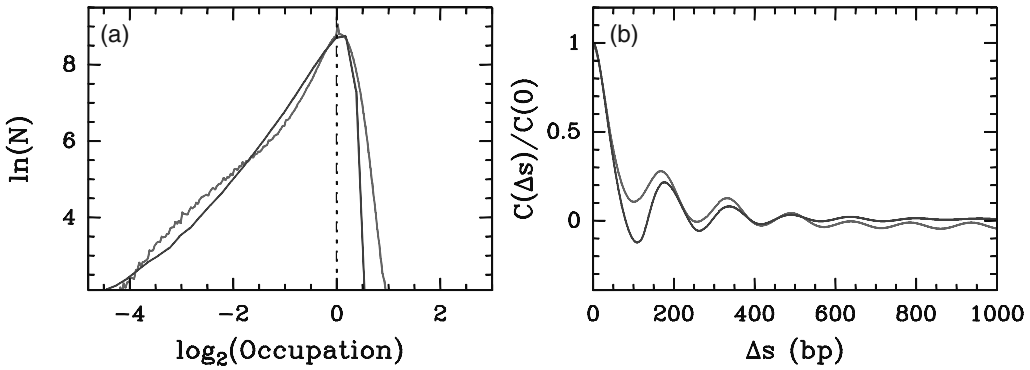

Figure 33. Comparing the predictions of our physical modelling ($\tilde{\mu} = -1.3\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$, $l_w = 125\,\mathrm{bp}$) with the Lee et al. (2007) *S. cerevisiae in vivo* MNase-chip data. (a) Histograms of nucleosome occupancy $Y(s)$ values centered at their typical value: model (blue), *in vivo* data (red). (b) Corresponding auto-correlation function $C(\Delta s) = \langle \delta Y(s) \delta Y(s + \Delta s) \rangle$.

models based on statistical learning like the Yuan and Liu (2008) model that yields a Pearson correlation histogram very similar to the one obtained with our physical model (Figure 32a), the Field et al. (2008) model ($\bar{r} = 0.39$), the Kaplan et al. (2009) model ($\bar{r} = 0.34$) and the Peckham et al. (2007) model ($\bar{r} = 0.22$). As a careful inspection of the theoretical and *in vivo* experimental nucleosome occupancy profiles in Figure 31 seems to indicate, these weakest correlations result from two main features, namely, (i) experimental NFRs that do not correspond to genomic energy barriers but more likely result from the action of external factors like transcription of other proteic factors (TF, Insulators, PIC…) and (ii) regions (up to 1 kbp) where the experimental nucleosomal pattern is shifted by a few tens bp with respect to the predicted nucleosomal pattern as the possible outcome of ATP consuming remodelling factors. We will come back to this point in more details in the next subsection.

Let us emphasize that at a statistical level, our physical model accounts very well for the *in vivo* distribution of nucleosome occupancy values obtained from the Lee et al. (2007) MNase-chip data (Figure 33a) as well as for the harmonic modulation with a period $l^* = 172\,\mathrm{bp}$ which is slightly larger than the NRL $l^* = 167\,\mathrm{bp}$ observed *in vivo* in the two-point correlation function (Figure 33b). As reported in Figure 34, the distribution of the local NRL values obtained from the predicted nucleosome profile is indeed shifted towards larger values (centered around $l^* = 172\,\mathrm{bp}$), but present a fluctuation range similar to the experimental distribution. Note that the NRL computed as the periodic modulation of the two-point correlation of the (one-point) nucleosome distribution is not *stricto sensus* equal to the NRL that we

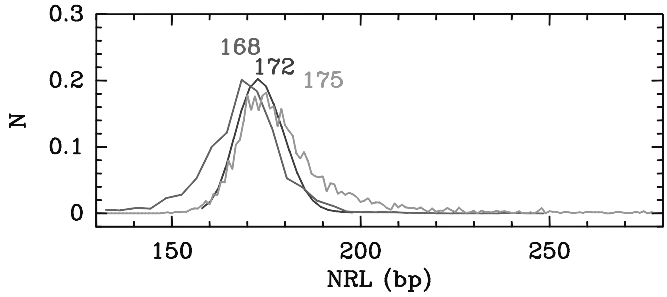

Figure 34. Histograms of local NRL: comparison of the predictions of our physical modelling ($\tilde{\mu} = -1.3\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$ and $l_w = 125\,\mathrm{bp}$) for budding yeast (blue) (the same histogram is obtained for *C. elegans*) with the *in vivo S. cerevisiae* MNase-chip data of Lee et al. (2007) (red) and *C. elegans* MNase-seq data of Valouev et al. (2008) (green).

have introduced and discussed in subsection 'Homogeneous energy profile $E = E_o$', as characterizing the spatial periodic modulation of the two-point nucleosome distribution. The equality holds in the particular case of the semi-confined fluids, i.e. around vertical energy barriers, where both density and pair distribution function indeed coincide. The extreme situation is in regions where the density (occupancy) is almost flat (or unorganized) and where, consequently, the two-point correlation function (or the Fourier spectrum analysis) cannot account for the internal periodic ordering of the nucleosomal array. These results confirm that the stretches of well-ordered nucleosomes observed *in vivo* but not *in vitro* (i.e. at high but not at low nucleosome density) are the direct consequence of the organizing role of effective nucleosome energy barriers that condition nucleosome ordering over rather long distances consistent with the statistical physics principles (see section 'Thermodynamical model of nucleosome assembly').

Remark: As noticed in the section '*In vivo* and *in vitro* genome-wide primary structure of chromatin', the fact that our physical modelling reproduces quite well the distribution of *in vivo* nucleosome occupancy values obtained in *S. cerevisiae* by Lee et al. (2007) (MNase-chip), questions the reliability of the *in vivo* Kaplan et al. (2009) (MNase-seq) data that surprisingly yields the same nucleosome occupancy histogram as obtained *in vitro* (Figures 7b and 10c) whereas the nucleosome density is more than twice bigger *in vivo* (75%) than *in vitro* (30%).

From in vitro *to* vivo*: 'Intrinsic' versus 'extrinsic' nucleosome positioning*

The small-scale chromatin structure, as defined by the local nucleosome occupancy, conditions the regulation of transcription in particular by modulating the accessibility of TFs to their cognate regulatory sites (Kornberg and Lorch 1999; Li et al. 2007; Morse 2007; Rando and Ahmad 2007; Segal and Widom 2009b). Actually, as seen in the previous sections, the nucleosome occupancy profile predicted, directly from the DNA sequence, by our physical model using a grand canonical description, accounts remarkably well for the nucleosome occupancy profile observed *in vitro* (see subsection 'Modelling of *in vitro* nucleosome occupancy data in *S. cerevisiae*') (Vaillant et al. 2007; Chevereau et al. 2009). However, the comparison with *in vivo* data reveals that the 'intrinsic' nucleosome positioning encoded in the sequence can be influenced and perturbed by the action of 'extrinsic' factors like TFs and ATP-dependent remodellers (Segal and Widom 2009; Radman-Livaja and Rando 2010).

Transcription factors

TFs can influence nucleosome positioning *in vivo* by competing with histones to access to their DNA target sites (Koerber et al. 2009). The outcome of this competition likely depends on the relative affinities of the nucleosomes and TFs to the underlying DNA sequence but also on their relative concentrations (Segal and Widom 2009c). As shown in Figures 35(a) and (b), when comparing the nucleosome occupancy profile predicted by our physical model

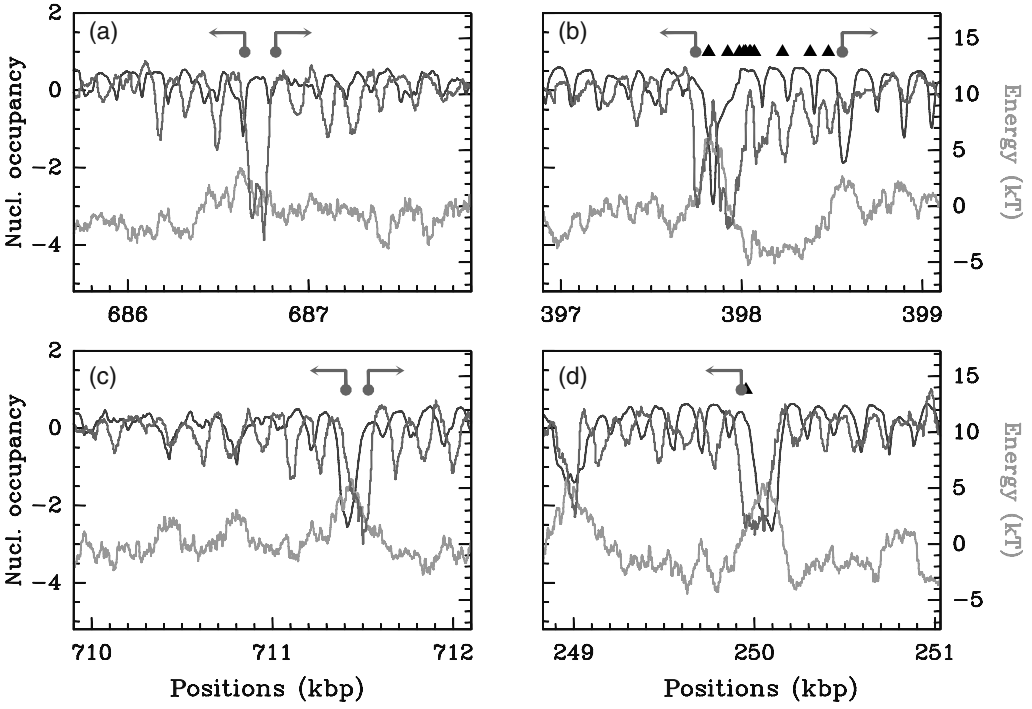

Figure 35. Nucleosome occupancy profiles observed *in vivo* (red) and predicted by our physical model for parameter values $\tilde{\mu} = -1.3$ kT, $\delta = 2$ kT and $l_w = 125$ bp (blue) along fragments of *S. cerevisiae* chromosome 2 (a), 7 (b), 2 (c) and 6 (d). For comparison also represented are the corresponding theoretical energy landscapes (green). The symbols represent the positions of TSS (red dots) and TFS (black triangle). The arrows at TSS indicate the transcription sense.

at high nucleosome density with the *in vivo* yeast data, we observe mainly two kinds of differences. There are locations where a NFR is observed *in vivo* but not *in vitro* as predicted by our physical model (Figure 35a). At many other locations where some nucleosome depletion is predicted by our model, the *in vivo* nucleosome occupancy profile displays a deeper and pronounced depleted region. At a larger scale, as seen in the intergenic region between the two divergent yeast genes in Figure 35(b), the high concentration of TFs coincides with a significant lowering of the *in vivo* mean nucleosome occupancy. Importantly, this lowering has not disturbed the regular nucleosome ordering predicted by the physical model consistent with the emerging view that nucleosome and TFs compete to occupy DNA in thermodynamic equilibrium (Raveh-Sadka et al. 2009; Segal and Widom 2009b, 2009c). Even more interesting, TFs reside in the predicted nucleosome depleted linker regions of the intergenic regular nucleosome array suggesting some cooperativity in TF binding (Adams and Workman 1995; Polach and Widom 1996; Vashee et al. 1998; Miller and Widom 2003) as the result of a collaborative competition against 'intrinsic' collective nucleosome ordering and not just of specific protein–protein interactions. Note that the theoretical NFR of the anti-sense gene at the left of Figure 35(b) has been disturbed (mainly shifted) by the binding of TFs that have induced a second NFR nearby possibly catalyzed by remodelling factors.

ATP-dependent chromatin remodelling factors

Nucleosome positioning can also be controlled by a family of enzymes that consume the energy from ATP hydrolysis to move nucleosomes to different locations along the DNA or even to disassemble nucleosomes (Tsukiyama and Wu 1997, Längst et al. 1999; Lorch et al. 1999; Travers 1999; Whitehouse et al. 1999, 2007; Peterson and Workman 2000; Hamiche et al. 2001; Angelov et al. 2003; Boeger et al. 2008; Shivaswamy et al. 2008; Hartley and Madhani 2009; Clapier and Cairns 2009). *In vitro* these molecular motors were shown to actively drive nucleosomes away from presumed equilibrium positions (Montel et al. 2007; Rippe et al. 2007) and this regardless of the underlying DNA sequence. As illustrated in Figures 35(c) and (d), some *in vivo* nucleosome occupancy patterns including the $5'$ NFR (resp. $3'$ NFR) and the flanking ordered nucleosomes inside the corresponding genes are globally shifted by $\sim$50–100 bp from their predicted positions by our DNA sequence directed grand canonical modelling. This observation is confirmed statistically in Figures 32(b) and (c) where we have re-computed the Pearson correlation histogram between our physical theoretical occupancy profiles and the *in vivo* experimental ones of Lee et al. (2007), when allowing a possible shift ($-200 < d < 200$ bp) between the numerical and experimental profiles. When optimizing the shift (d_M) for each 1 kbp sliding window, the histogram is significantly shifted towards higher values with a mean value $\bar{r} = 0.5$. Note that a similar histogram is now obtained with a random control where our 1 kbp theoretical profiles were compared to randomly chosen 1 kbp experimental profiles along the 16 yeast chromosomes. But as shown in Figure 32(c), whereas the distribution of optimal shift d_M values is rather flat for the control, it is narrowly peaked around $d_M = 0$ with a width $\simeq 60$ bp, confirming that up to some local shift of a few tens bp, this sequence-dependent model accounts remarkably well for *in vivo* nucleosome positioning data. Let us point out that these distances are typical of distances over which the ATP consuming remodelling factors are known to operate *in vivo* (Whitehouse et al. 2007; Shivaswamy et al. 2008; Hartley and Madhani 2009). This 'extrinsic' nucleosome positioning under the action of remodelling factors is likely to explain the strong phasing of the $5'$ NFR with respect to the gene TSS observed *in vivo* as compared to our physical model predictions as discussed in Chevereau et al. (2009), and Vaillant et al. (2010).

But as noticed by Rippe et al. (2007), some remodelling complexes only change the relative nucleosome occupancy without altering nucleosome positions. This suggests that equilibrium positioning can also be relevant even during remodelling activity. Since these chromatin remodellers are found all over the yeast chromosomes, they may contribute to increasing the effective temperature so that thermal equilibrium is attained much faster, in particular along most yeast genes as discussed in the subsection 'Bistability induced by statistical confining in between two energy barriers'. In that context, our grand canonical physical model can be used as a theoretical reference for *in vitro* nucleosome positioning whose comparison with *in vivo* data is likely to provide very instructive information for future modelling of both remodeller (Teif and Rippe 2009) and TF (Raveh-Sadka et al. 2009) driven 'extrinsic' nucleosome positioning.

Remark: Let us note that at high density, our 'intrinsic' model could be improved by allowing a short-range attraction between nucleosomes as the result of the folding of the nucleosomal array into higher-order chromatin structure (Chereji et al. 2011; Riposo and Mozziconacci 2012). Works in this direction are under progress.

What about other genomes?

S. kluyveri

As shown in Figure 36(a) on a 10 kbp fragment of *S. kluyverii* chromosome C, when using the same parameters ($\tilde{\mu} = -1.3$ kT, $\delta = 2$ kT and $l_w = 125$ bp) as used in the previous subsection to model the nucleosome occupancy profiles observed *in vivo* in *S. cerevisiae*, our physical model again predicts theoretical profiles that match rather well the experimental data of Tsankov et al. (2010) (Figure 3). When averaging over the eight *S. kluyveri* chromosomes, we get a mean Pearson correlation $\bar{r} = 0.32$ in good agreement with the value previously obtained for *S. cerevisiae* ($\bar{r} = 0.33$).

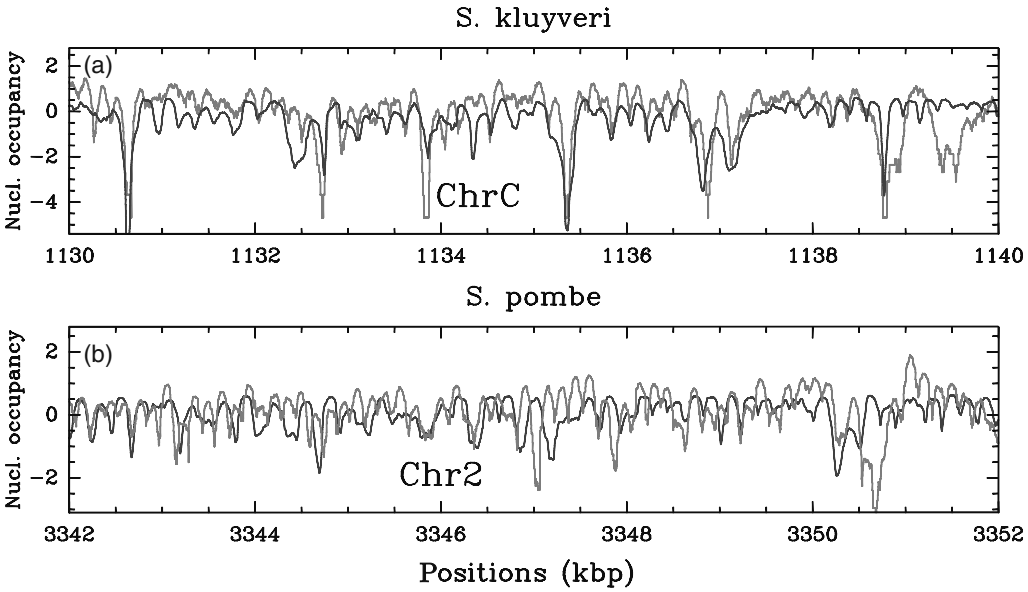

Figure 36. Comparison between our physical model predictions ($\tilde{\mu} = -1.3\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$, $l_w = 125\,\mathrm{bp}$) (blue) and *in vivo* nucleosome occupancy MNase-seq data (Tsankov et al. 2010, 2011) (orange): (a) *S. kluyveri*; 10 kbp fragment on chromosome C; (b) *S. pombe*; 10 kbp fragment on chromosome 2.

Let us point out that quite similar results are obtained for other Hemiascomycota fungi (data not shown).

S. pombe

The comparison of our physical model predictions with the nucleosome occupancy profiles observed *in vivo* in *S. pombe* (Figure 4) looks much less satisfactory as illustrated in Figure 36(b) on a 10 kbp fragment of chromosome 2. Indeed if the model parameters estimated in the previous subsection for *S. cerevisiae* are still rather optimal, the mean Pearson correlation obtained when averaging over the three *S. pombe* chromosomes $\overline{r} = 0.1$ is much weaker than for *S. cerevisiae* and *S. kluyverii*. Let us mention that a similar observation was reported by Lantermann et al. (2010) when comparing their *in vivo* MNase-chip data (Figure 4a) and the predictions of models based on statistical learning (Field et al. 2008; Kaplan et al. 2009).

C. elegans

As can be seen by a simple visual inspection of Figure 37, our physical model accounts much better for the *C. elegans in vivo* nucleosome occupancy MNase-seq data of Valouev et al. (2008) (see Figure 5) and this without requiring any change in the model parameters estimated from *S. cerevisiae* data. As shown in Figure 38, the histogram of Pearson correlation values computed in a 10 kbp sliding window along the entire *C. elegans* genome is rather wide with a maximum at $r^* = 0.51$ and a mean $\overline{r} = 0.43$. This mean value is significantly larger than the mean values previously obtained for the different yeast genomes. Again

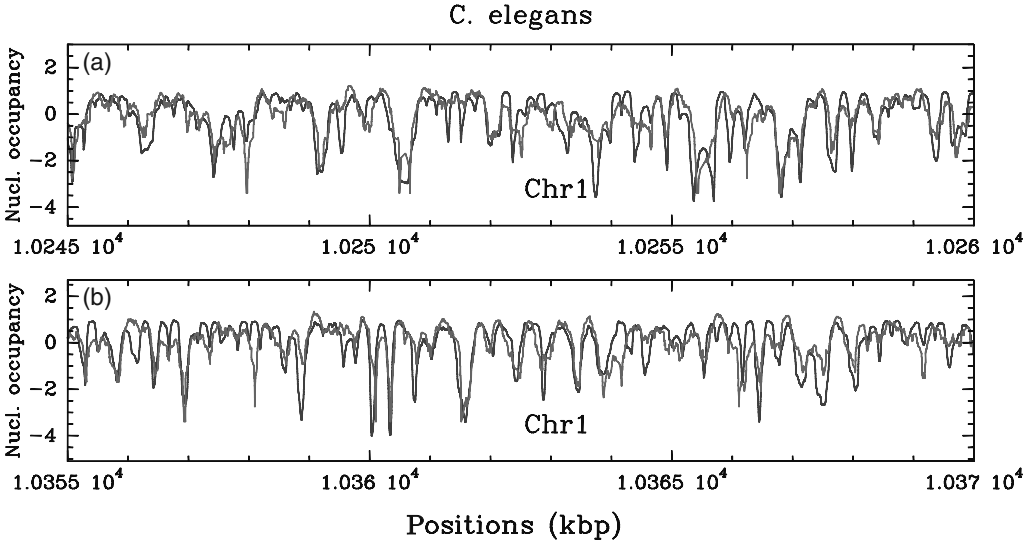

Figure 37. Comparison between our physical model predictions ($\tilde{\mu} = -1.3\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$, $l_w = 125\,\mathrm{bp}$) (blue) and the *C. elegans in vivo* nucleosome occupancy MNase-seq data (Valouev et al. 2008) (red): (a) and (b) correspond to two 15 kbp fragments of chromosome 1.

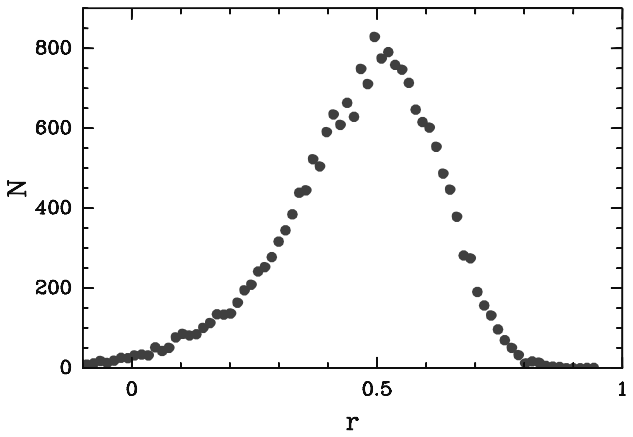

Figure 38. Histogram of Pearson correlation values r between our physical model ($\tilde{\mu} = -1.3\,\mathrm{kT}$, $\delta = 2\,\mathrm{kT}$, $l_w = 125\,\mathrm{bp}$) and the Valouev et al. MNase-seq *in vivo* nucleosome occupancy data (Valouev et al. 2008) (see Figure 37). The Pearson correlation r was measured in a 10 kbp sliding window over the six *C. elegans* chromosomes.

it compares and competes remarkably well with the mean correlation values obtained with models based on statistical learning, namely the Kaplan et al. (2009) model ($\bar{r} = 0.47$), the Field et al. (2008) model ($\bar{r} = 0.46$) and the Peckham et al. (2007) model ($\bar{r} = 0.29$). As previously observed for *S. cerevisiae*, the performances of formely proposed physical models are by far less satisfactory, e.g. the much weaker $\bar{r}$ values obtained with the Miele et al. (2008) model ($\bar{r} = 0.21$) and the Tolstorukov et al. (2007) model ($\bar{r} = -0.001$).

Human

As previously noticed in the subsection '*In vivo* nucleosome occupancy profiles' (Figure 6), the sequencing depth in the human *in vivo* nucleosome occupancy MNase-seq data obtained by Schones et al. (2008) is too weak to allow us to perform a quantitative comparison with the predictions of our physical model. However, as shown in Figure 39 on several 10 kbp fragments of the human chromosome 6, the agreement is good. In particular, when focusing on the nucleosome occupancy profiles observed around human gene TSS, the numerical and experimental mean profiles are quite consistent, in particular for CpG poor gene promoters (data not shown). A more quantitative comparison with the recent *in vivo* and *in vitro* nucleosome occupancy MNase-seq data of Valouev et al. (2011) is currently under progress.

Discussion

Rôle of the genomic sequence on nucleosome positioning: 10 bp periodicity versus long-range correlations?

10 bp periodicity

With the objective of finding consensus nucleosome positioning sequences, in a pioneering analysis of the original *S. cerevisiae in vivo* nucleosome occupancy data of Yuan et al. (2005), Ioshikhes et al. (2006) and Segal et al. (2006) have concluded that a large set of well-defined nucleosome positions could effectively be related to a 'genomic nucleosomal pattern' based on the 10 bp periodicity in the distribution of given dinucleotide steps (e.g. AA/TT) (Satchwell et al. 1986; Ioshikhes et al. 1996; Widom 1996, 2001). Indeed, using probabilistic models that take into account the matching of their patterns to the sequence and the steric hindrance between nucleosomes, they both obtained nucleosome occupancy profiles that correlate rather well with the *in vivo* experimental data, specially in regions of well-positioned nucleosomes. However, later studies by Peckham et al. (2007) and Yuan and Liu (2008) have seriously questioned the conclusions of Ioshikhes et al. (2006) and Segal et al. (2006). When using a HMM (Hidden Markov Model) prediction algorithm to compare the performances of their models to those of Ioshikhes et al. and Segal et al. models, they both found that no more than 20% of the *in vivo* nucleosome positioning above what is expected by chance is determined by intrinsic signals in the genomic DNA. As shown in Figure 40(a), when applying a similar HMM approach to the numerical nucleosome occupancy profiles predicted by our physical model and then comparing to the set of well-positioned nucleosomes obtained by Lee et al. (2007) on their *S. cerevisiae in vivo* data, we get performances quite similar to those obtained with the Yuan and Liu (2008) model. Indeed our physical model predicts 48.7% of true positive within a distance of 35 bp as compared to 42% by chance. This is actually nothing but the expression, at the level of (HMM derived) well-positioned nucleosomes, of the remodelling shifting effects previously discussed in the subsection 'ATP-dependent chromatin remodelling factors'. When focusing on well-positioned nucleosomes, the characteristic 'remodelling distance' (i.e. the 'shifting' distance above which the overlapping between predicted and experimental nucleosomes is decreasing) is about 35 bp (Figure 40a). The fact that our physical model performs as modestly on well-positioned nucleosomes as the Peckham et al. (2007) and Yuan and Liu (2008) models, is an indication that these

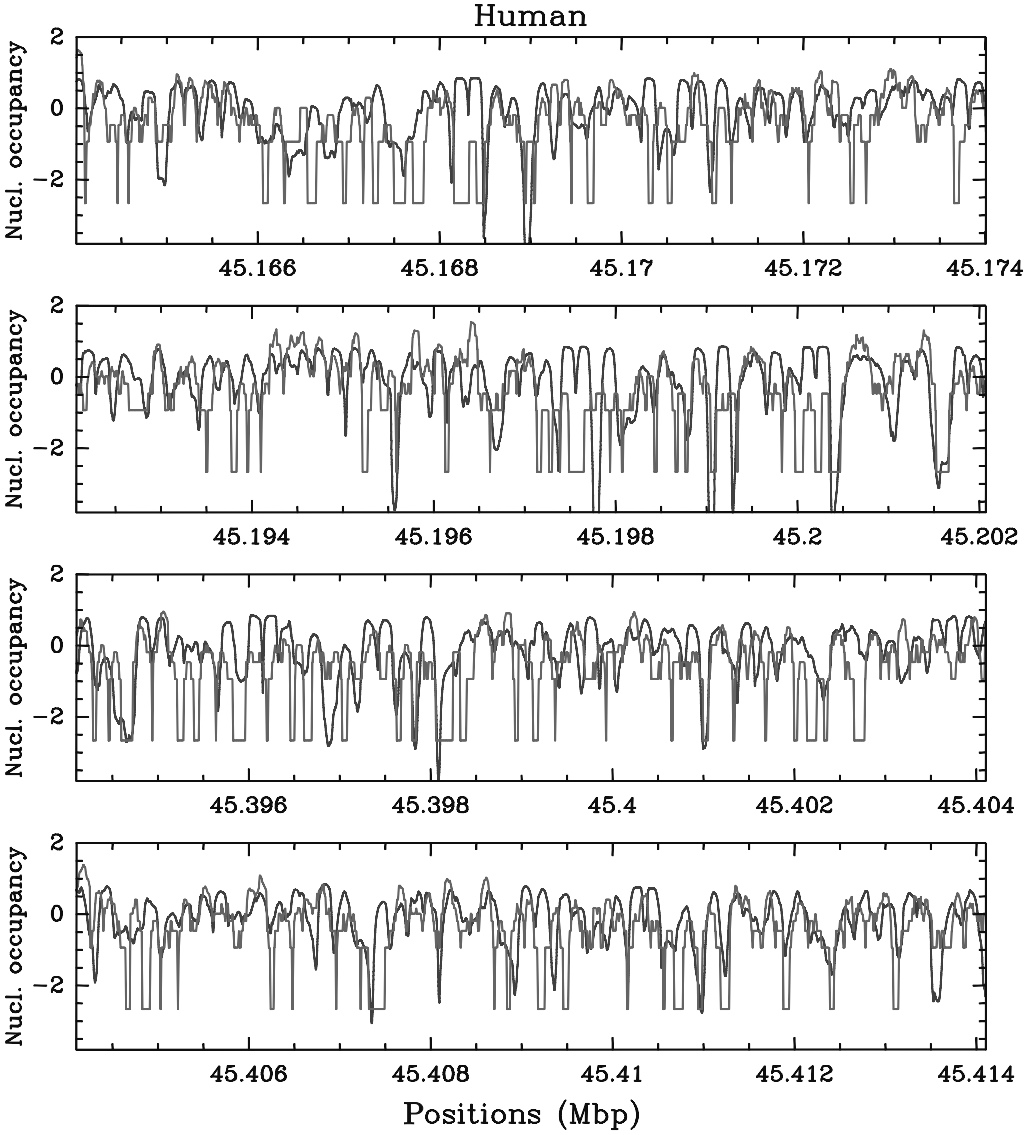

Figure 39. Comparison between our physical model predictions ($\tilde{\mu} = -1.3$, $\delta = 2\,\mathrm{kT}$, $l_w = 125\,\mathrm{bp}$) (blue) and the *in vivo* nucleosome occupancy MNase-seq data obtained by Schones et al. (2008) in human CD4$^+$ T cells (red). The four panels correspond to 10 kbp fragments of chromosome 6.

authors were right to moderate the too hasty statement made by Ioshikhes et al. (2006) and Segal et al. (2006) that 10 bp periodic dinucleotide patterns could explain a majority of well-positioned nucleosomes. However, let us point out that when computing the performance of the HMM predictions of our physical model in a 5 kbp sliding windows over the 16 *S. cerevisiae* chromosomes, we obtain a histogram of performance values that now strikingly deviates from random expectation (Figure 40b). If performance is rather homogeneously distributed along the genome around the expectation value for the random control (42%), there is a greater heterogeneity for our theoretical predictions with a large proportion of genomic regions where performance is much better than expected by chance, while in some other regions it is worse. Again, this reflects at the level of well-positioned nucleosomes, which is what we observed for the Pearson correlation distribution in Figure 32.

Actually the rather modest predictive power of methods based on a 10 bp periodicity of some di- or tri-nucleotides is not surprising since this periodicity has been established on strongly positioning sequences that present large variations in nucleotide contents (from 0.1 to 0.5, Figure 41a). Among these sequences that have an anomalously large affinity to the histone octamer to form the nucleosome, the 601 sequence (Lowary and Widom 1998; Thåström et al. 2004) was shown to have a gain in the formation energy of the DNA–histone complex of $\Delta E = -4.9 \pm 0.55\,\mathrm{kT}$ relative to a reference sequence (the sea urchin 5S rRNA gene sequence (Dong et al. 1990)). This 601 sequence was recently shown to prevent the nucleosome from sliding (Shlyakhtenko et al. 2009) which explains that, for obvious functional reasons, no organism actually possesses this sequence in its genome. As shown in Figure 41(a), the sequences that are known to bind to the histone octamers in eukaryotic organisms have a much weaker nucleotide content variability, typically 5–10% for the chicken and 3–12% for *S. cerevisiae*. As illustrated in Figure 41(b), this simply means that nucleosomes adapt themselves better on sequences that display a 10 bp periodicity with AA/TT/AT that oscillate in phase with each other and

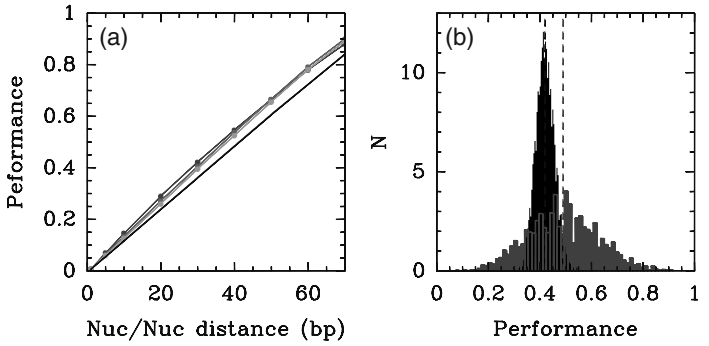

Figure 40. Performances of our physical model ($\tilde{\mu} = -1.3$ kT, $\delta = 2$ kT, $l_w = 125$ bp) and of Yuan and Liu (2008) N-score model in terms of well-positioned nucleosomes as obtained by HMM methods. The comparison is made against the set of well-positioned nucleosomes obtained by Lee et al. (2007) on their *S. cerevisiae in vivo* experimental data using a similar HMM algorithm. 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 versus the overlapping distance for the theoretical predictions of our physical model for the nucleosome occupancy profile (blue) and the energy landscape (green), and the Yuan and Liu N-score model (magenta). (b) Statistics of the performance values (at 35 bp accuracy) computed in a sliding window of size 5 kbp along the entire *S. cerevisiae* genome for our theoretical nucleosome occupancy predictions (blue) and for the random control (black). The vertical dashed lines (black and blue) indicate the corresponding mean values.

facing the minor groove, and out of phase with GC facing the major groove. As originally pointed out in Audit et al. (2001, 2002, 2004), when performing time–frequency analysis of eukaryotic genomic sequences using either dinucleotide codings or more elaborated di- or tri-nucleotide experimental tables coding for the structural and/or bending properties of the DNA double helix, there is no significant peak that emerges in the power spectrum at the frequency $1/10\,\mathrm{bp}^{-1}$. This confirms that if locally the 10 bp periodicity sketched in Figure 41(b) can help to phase and position some nucleosomes, at the genome scale this periodicity is clearly not exploited to position the majority of well-defined nucleosomes observed *in vivo*.

Long-range correlations

Actually when performing power-spectrum and correlation analysis on the *in vivo* nucleosome occupancy data of Lee et al. (2007) (Figure 1), we mainly reveal the existence of a mean period $l^* \sim 167 \pm 10$ bp that corresponds to regular arrays of well-ordered nucleosomes. This chacteristic NRL manifests as a bump in the power spectrum (Figure 42) at high frequencies ($1/167\,\mathrm{bp}^{-1}$) and not as a peak for strict periodicity, as the signature of some fluctuations in the NRL values. As previously reported in subsection 'Modelling of *in vivo* nucleosome occupancy data in *S. cerevisiae*' (Figure 33), this statistical nucleosome ordering can also be diagnosed from the periodic modulations observed in the auto-correlation function $C(\Delta s) = \langle \delta Y(s) \delta Y(s + \Delta s) \rangle$. But in addition and very importantly, when plotted in a logarithmic representation, the power spectrum displays a very convincing power law decay $S(k) \propto k^{-\nu}$, with exponent $\nu = 2H - 1 = 0.74 \pm 0.02$ ($H = 0.87$) that is likely to be a direct consequence of the large-scale (low frequency $k < 1/200$) LRC regime observed in the yeast DNA bending profile in Audit et al. (2001, 2002, 2004).

When reproducing this statistical analysis on the *in vitro* nucleosome occupancy data of Kaplan et al. (Figure 10), consistently with the disappearance of the periodic modulations in the auto-correlation function (Figure 10d), the power spectrum in Figure 42 no longer displays a bump

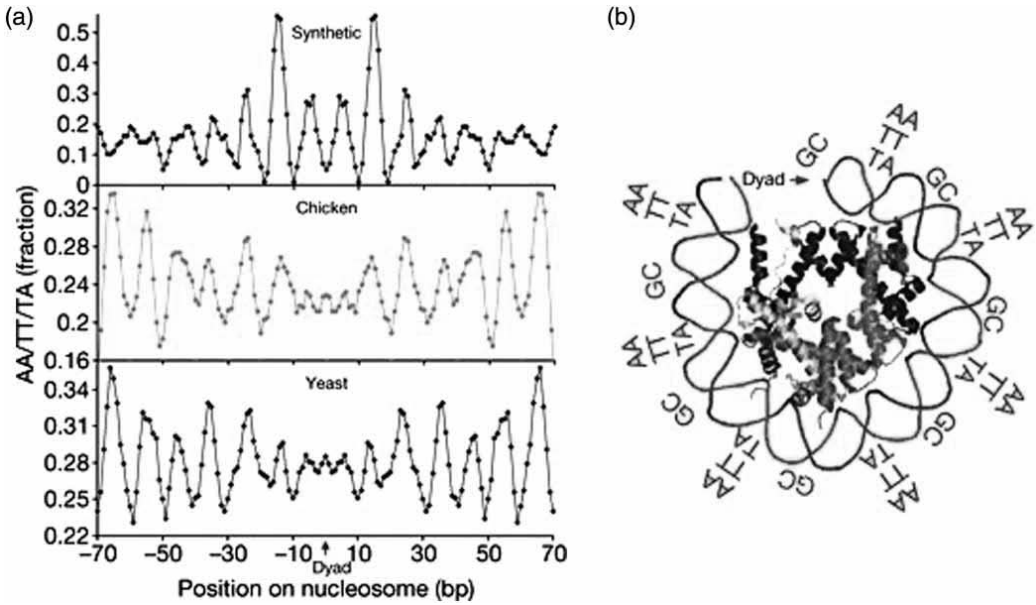

Figure 41. (a) Fraction in dinucleotides AA/TT/TA (3 bp moving average) at each position of centre aligned yeast, chicken and random chemically synthesized nucleosome-bound DNA sequences showing ~10 bp periodicity of these dinucleotides. (b) Key dinucleotides inferred from the alignment are shown relative to the three-dimensional structure of one-half of the symmetric nucleosome. Adapted from Segal et al. (2006).

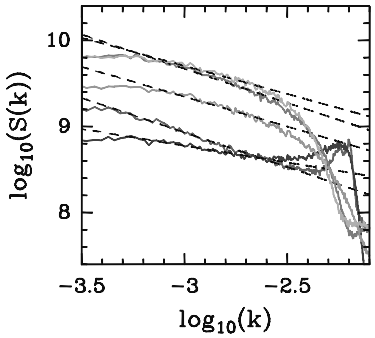

Figure 42. Power spectrum analysis of nucleosome occupancy profiles obtained from the *in vivo* data of Lee et al. (2007) (red), the *in vitro* data of Kaplan et al. (2009) (orange), the physical model described in the section 'A sequence-dependent physical model of nucleosome occupancy' for $\delta = 2\,\mathrm{kT}$ and low $\tilde{\mu} = -6\,\mathrm{kT}$ (cyan) and high $\tilde{\mu} = -1.3\,\mathrm{kT}$ (dark blue) nucleosome density. For comparison, the green curve corresponds to the power spectrum of the formation energy landscape. The dashed lines correspond to the power-spectrum scaling exponent values $\nu = 0.65, 0.74, 0.68, 0.74$ and 0.46 from top to bottom corresponding to the following Hurst exponent values $H = 0.82, 0.87, 0.84, 0.87$ and 0.77, respectively.

at high frequency $(l^*)^{-1} = (167\,\mathrm{bp})^{-1}$ as an indication of a significant weakening of periodic nucleosome ordering. But what is remarkable is the fact that the power spectrum (as well as the auto-correlation function) still presents a very convincing power-law behavior with exponent $\nu = 0.74$ corresponding to a Hurst exponent value $H = 0.85 > 1/2$, the hallmark of the presence of LRC.

In Figure 42 are also shown for comparison the power spectra of the nucleosome occupancy profiles predicted by our physical model at low ($\tilde{\mu} = -6\,\mathrm{kT}$) and high ($\tilde{\mu} = -1.3\,\mathrm{kT}$) nucleosome density. As previously observed for the auto-correlation functions in Figures 30(b) and 33(b) respectively, the theoretical power spectra are in good agreement with the experimental ones. In particular the power-law decay of the power-spectrum is well reproduced with an exponent $\nu = 0.65\,(H = 0.82)$ and $0.46\,(H = 0.73)$ respectively at low and high nucleosome densities which corroborates the existence of LRC in the numerical nucleosome occupancy profiles and further strengthens the relevance of these LRC in the experimental *in vitro* and *in vivo* data. But the most important result reported in Figure 42 is the fact that these LRC are also observed in the power spectrum of the nucleosome formation energy landscape with a power-law exponent $\nu = 0.68$ ($H = 0.84$), as likely dictated by the LRC encoded in the DNA sequence (Audit et al. 2001, 2002). Furthermore, as shown in Figure 43, when recomputing the energy landscape after randomly shuffling the DNA sequence, the obtained energy profile displays uncorrelated Gaussian fluctuations without anymore tail at large $\Delta E > 0$ values corresponding to the presence of excluding energy barriers in the genuine DNA sequence. This is a strong indication that the sequence signaling which prevails is excluding energy barriers that result from the presence of LRC in the DNA bending profile. As reported in the subsection 'Modelling of *in vitro* nucleosome occupancy data in *S. cerevisiae*', they explain the

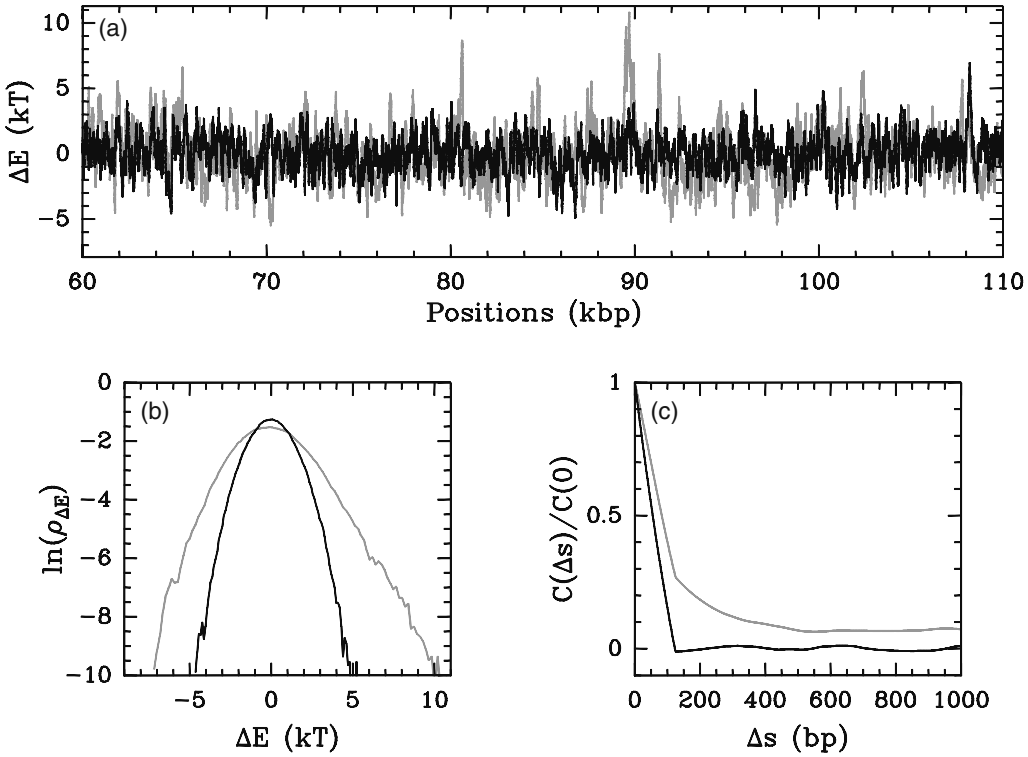

Figure 43. Energy landscape statistics ($\Delta E(s) = E(s) - \bar{E}$) computed with the following parameter values: $\delta = \langle (E - \bar{E})^2 \rangle^{1/2} = 2\,\mathrm{kT}$ and $l_w = 125\,\mathrm{bp}$. The colors correspond to the LRC genomic DNA sequence (green) and to its randomly shuffled uncorrelated version (black). (a) $\Delta E(s)$ along a 50 kbp long fragment of budding yeast chromosome 2. (b) Energy pdfs computed for the 16 yeast chromosomes. (c) Energy auto-correlation function $C(\Delta s)/C(0)$ vs Δs.

NFRs observed *in vitro* in *S. cerevisiae* as well as the regular nucleosome ordering observed nearby *in vivo* (subsection 'Modelling of *in vivo* nucleosome occupancy data in *S. cerevisiae*') as the result of statistical confining according to thermal equilibrium principles (Section 'Statistical positioning'). But as discussed in the subsection 'From *in vitro* to *in vivo*: 'Intrinsic' versus 'extrinsic' nucleosome positioning', there are much more (2 or 3 times) NFRs observed *in vivo* than *in vitro* and than predicted from the energy barriers encoded in the DNA sequence. These additional NFRs likely result from the action of external factors (TF, chromatin remodellers) and contribute to strengthen the collective nucleosomal ordering observed *in vivo* and specially in *S. cerevisiae* genes as shown in the subsection 'Bistability induced by statistical confining in between two energy barriers' (Figure 23).

(G + C) content drives nucleosome occupancy: experimental bias or reality?

As originally pointed out by Miele et al. (2008), the nucleosome occupancy profile observed *in vivo* in *S. cerevisiae* turns out to be significantly correlated to the local (G + C) content when estimated in a 125 bp sliding window. As shown in Figure 44, the (G + C) content provides an excellent prediction of the *in vitro* nucleosome occupancy MNase-seq data obtained by Kaplan et al. (2009) in *S. cerevisiae*. The mean Pearson correlation is as large as $\overline{r} = 0.78$ and is comparable to the performances of our physical model ($\overline{r} = 0.74$) and of more sophisticated models based on statistical learning (see the subsection 'Modelling of *in vitro* nucleosome occupancy data in *S. cerevisiae*'). *In vivo*, this correlation is still strong and comparable with the ones obtained with the nucleosome profiles predicted by other models (see the subsection 'Modelling of *in vivo* nucleosome occupancy data in *S. cerevisiae*'). This is particularly true for the *S. cerevisiae in vivo* MNase-seq data of Kaplan et al. (2009), $\overline{r} = 0.40$. A significantly smaller but still significant value $\overline{r} = 0.25$ is obtained with the *S. cerevisiae in vivo* MNase-chip data of Lee et al. (2007). Note that a similar correlation between (G + C) content and nucleosome occupancy is observed in the *C. elegans in vivo* data of Valouev et al. (2008): $\overline{r} = 0.42$, as compared to $\overline{r} = 0.43$ with our physical model. Overall, it turns out that in terms of the genomic rule, the (G + C) content is the one that appears to contribute most significantly to 'intrinsic' nucleosome occupancy. As a consequence, all the models which are based on an affinity of DNA to the histone octamer that strongly correlates to the (G + C) content are likely to predict nucleosome occupancy profiles that match quite well the *in vitro* experimental data (Tillo and Hughes 2009). This is typically the case of the 'Pnuc' structural bending coding table (Goodsell and Dickerson 1994; Gabrielian and Pongor 1996) used in our physical model (see subsection 'Intrinsic' nucleosome formation energy landscape) and of other di- or tri-nucleotide coding tables proposed in other studies (Anselmi et al. 2000; Miele et al. 2008; Santis et al. 2010). As previously mentioned, another well-established and major compositional factor is the poly(dA:dT) which are known to correspond to rigid DNA fragments that impair nucleosome formation (Yuan et al. 2005; Bao et al. 2006; Segal and Widom 2009a; Tillo and Hughes 2009) and favor nucleosome disassembly (Iyer and Struhl 1995; Suter et al. 2000) by increasing the DNA wrapping free energy cost. Overall, as estimated on synthetic 150 bp oligonucleotides (Kaplan et al. 2009; Tillo and Hughes 2009), over the range of (G + C) content from 20% to 60%, the DNA/histone affinities can be five-fold or greater. This means that the difference in nucleosome formation energy between the genomic low (G + C) unfavorable sequences and high (G + C) more favorable sequences is ~ 1.6 kT (maximum value ~ 3 kT). As far as poly-A are concerned, the depletion observed with respect to random sequences actually depends on their size, namely 2 (-0.7 kT) to 6 (-1.7 kT) fold for 5 to 15 bp fragments up to 30 fold (-3.4 kT) for 25 bp fragments. Thus according to these *in vitro* data (Kaplan et al. 2009) which are 'MNase independent' (see below), the range of affinities, or in other words, the energy variability for a genomic sequence to form the nucleosome is rather weak (as compared to strongly positioning artificial sequences, as the 601 sequence (Lowary and Widom 1998, Thåström et al. 2004)) but non-negligible as implemented in our physical model in the subsection "Intrinsic' nucleosome formation energy landscape' by fixing the parameter $\delta = \langle (E - \overline{E})^2 \rangle^{1/2} = 2$ kT that controls the fluctuation range in the energy landscape.

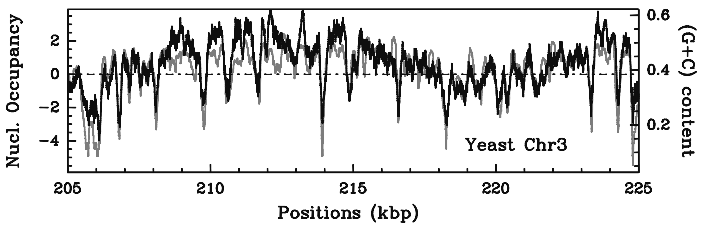

Figure 44. Comparison between the (G + C) content estimated in a 125 bp sliding window (black) and the *S. cerevisiae in vitro* nucleosome occupancy MNase-seq data of Kaplan et al. (2009) (orange). The horizontal line indicates the genome wide average (G + C) content value of 0.38.

Experimental artefact

The observed important correlation between the (G + C) content and *in vitro* (and to a less extent *in vivo*) nucleosome occupancy data raises the issue of a possible experimental artefact. Indeed, it is well known that the MNase presents a sequence specificity, cutting preferentially at AT steps (Dingwall et al. 1981; Hörz and Altenburger 1981). Recently, Chung et al. (2010) and Fan et al. (2010) have shown that MNase digestion profiles obtained on genomic naked DNA (*S. cerevisiae*) are indeed significantly correlated to the (G + C) content fluctuations as well as to the

chromatin digestion profiles obtained *in vitro* and *in vivo*. Similar observation at the promoters of yeast to fly genes was already reported in Miele et al. (2008). As performed in some experimental studies (Yuan et al. 2005; Lantermann et al. 2010), a way to overcome this possible bias consists in normalizing the chromatin MNase digestion data by the corresponding ones obtained on naked DNA. Hopefully, in an experimental study that just appeared (Allan et al. 2012), equivalent patterns of nucleosome positioning sites were obtained when using two nucleases that have notable differences in cleavage behavior, namely the MNase and the Caspase-Activated DNase (CAD). These recent results indicate that possible biases (e.g. content) in nucleosome positioning collected using MNase are likely to be insignificant. As noticed by Chung et al. (2010) this correlation with (G + C) content might be related to the reconstitution procedure (salt-gradient dialysis) that would naturally favor the sequence specificity of the tetramer $(H3/H4)_2$ which has been shown to preferentially bind to (G + C) rich sequences at high salt concentration. The physical origin of this sequence specificity (of the tetramer) and its effectiveness in the *in vivo* context remain elusive.

(G + C) content dependent DNA/histones interactions

Increasing experimental evidence seems to validate the central effect of the mean (G + C) content on the nucleosomal organization of the 10 nm chromatin fiber as recently reported by Valouev et al. (2011) for different types of human cells. This has led to a renewal in the modelling of DNA sequence evolution and specially of its (G + C) content across the eukaryotic kingdom in relationship with the availability of nucleosome occupancy data (Kenigsberg et al. 2010). Let us note that this correlation of the nucleosome occupancy profile with the (G + C) content is not revealed by most of the energetic models used in molecular dynamics to account for the detailed atomic

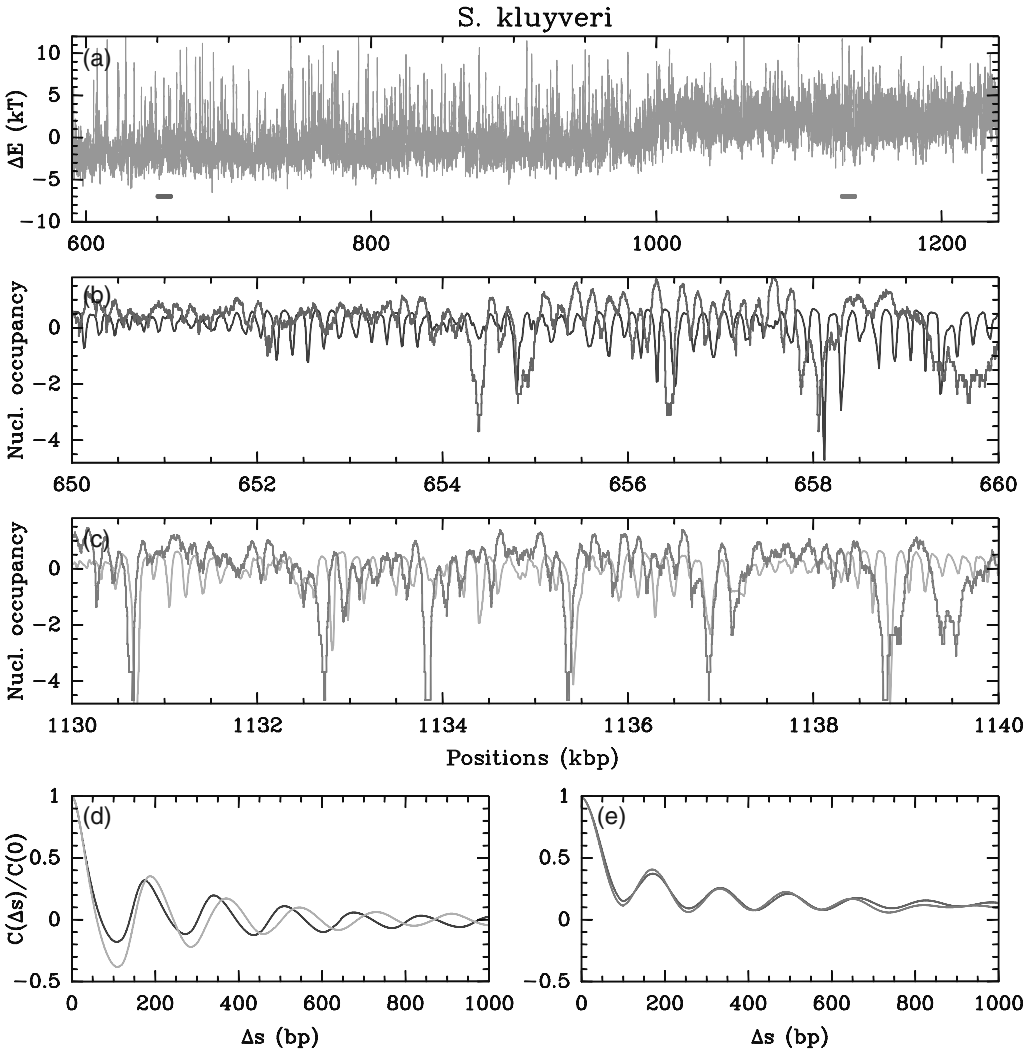

Figure 45. Comparison between the *S. kluyveri in vivo* nucleosome occupancy profile $\delta Y(s)$ of Tsankov et al. (2010) and the theoretical profile predicted by our physical model (see sections 'Statistical positioning' and 'A sequence-dependent physical model of nucleosome occupancy'). (a) Formation energy landscape $\Delta E(s) = E(s) - \overline{E}$ along the chromosome C computed with the following parameters $\delta = 2$ kT, $l_w = 125$ bp; the high (G + C) content (52%) contig corresponds to the first 1 Mbp of the chromosome; low (G + C) content (G + C = 40%) part corresponds to the last 250 kbp. (b,c) Comparison of the predictions of our physical modelling ($\tilde{\mu} = -1.3$ kT) (dark blue/cyan) with Tsankov et al. (2010) data (Figure 3) (red/orange) along a 10 kbp fragment of the high/low (G + C) content contig (indicated in (a) by the red/orange segments). (d,e) Corresponding auto-correlation functions $C(\Delta s) = \langle \delta Y(s) \delta Y(s + \Delta s) \rangle$.

interactions involved in the nucleosome complex (Morozov et al. 2009; Tolkunov and Morozov 2010). Indeed the physical models that perform as well as the one used all along in this paper (Vaillant et al. 2007; Chevereau et al. 2009) are based on energetic coding tables that were established from the local structural and mechanical properties of the DNA double helix, namely the 'Pnuc' coding table (Goodsell and Dickerson 1994; Gabrielian and Pongor 1996) in our study and the Anselmi et al. (2000) coding table in the Miele et al. (2008) model. In both these coding tables, the dominating contribution is given by the roll angle which definitely strongly correlates with the (G + C) content. *A posteriori*, it is therefore not so surprising that our physical model accounts so well for *in vitro* nucleosome positioning data in organisms like *S. cerevisiae* (see subsection 'Modelling of *in vivo* nucleosome occupancy data in *S. cerevisiae*') where the (G + C) content is rather homogeneous around 39% with a typical variability in the nucleosome formation energy profile $\delta = 2\,\mathrm{kT}$ (see the subsection 'Rôle of the genomic sequence on nucleosome positioning: 10 bp periodicity versus long-range correlations?'). But this raises the issue of modelling nucleosome occupancy data in higher eukaryotic organisms like in human where the so-called isochore structure manifests as large-scale domains (several 100 kbp) with uniform (G + C) content and appreciable scatter of the average (G + C) content when comparing domains (Bernardi 1989, 1995, 2000, 2001; Mouchiroud and Bernardi 1993; Lander et al. 2001; Li et al. 2003; Duret et al. 2006). Is the variability in the fluctuations in the energy landscape, namely large δ in (G + C) rich isochores and small δ in (G + C) poor isochores, sufficient to reproduce the experimental nucleosome densities and NRL in these regions? Do we need to change accordingly the chemical potential $\tilde{\mu}$ as the possible signature of compensatory co-evolutive mechanisms of regulation of the nucleosomal array? If yes, what are these sequence-dependent ((G + C) dependent) chromatin regulation mechanisms (Dekker 2008)? In Figure 45 are shown the results of a preliminary analysis of regions in the *S. kluyveri* genome that have a significantly different (G + C) content (Payen et al. 2009): the first 1 Mbp of the chromosome C has an average (G + C) content of 52.9% which is significantly higher than the 40.4% of the rest of the genome. As revealed by the auto-correlation functions of the experimental *in vivo* data of Tsankov et al. (2010) reported in Figure 45(e), the observed NRL in the high (G + C) domain is actually the same as the one observed in the rest of the genome, i.e. a typical value of $l^* = 167\,\mathrm{bp}$ (Figure 9). This contrasts with our physical model that clearly predicts a smaller NRL for the high (G + C) region ($l^* = 167\,\mathrm{bp}$) than for the rest of the genome (lower (G + C)) ($l^* = 172\,\mathrm{bp}$) (Figure 45d). This results from a lower nucleosome formation energy and thus a higher residual chemical potential in high as compared to low (G + C) regions (Figure 45a). Remarkably, the predicted value in the high (G+C) domain is exactly the one observed *in vivo* whatever the genomic content. This indicates that either our sequence-dependent model overestimates the influence of the (G + C) content variations or there is a mechanism that fully compensates/eliminates this 'intrinsic' nucleosome energy difference. A similar comparative analysis of our physical model predictions with the human *in vivo* nucleosome occupancy data of both Schones et al. (2008) and Valouev et al. (2011) at the genome scale is currently under progress. It is likely to provide new insight on the isochore stucture of mammalian genomes in relation with their primary nucleosomal chromatin structure.

Acknowledgements

We are very grateful to Y. d'Aubenton-Carafa, B. Audit, M. Barbi, C. Lavelle, R. Lavery, V. Miele, C. Thermes, A. Travers and J.M. Victor for very stimulating and fruitful discussions. This research was supported by the Conseil Régional Rhône-Alpes (project 'Le rôle de la séquence sur la structure et la dynamique de la chromatine', Emergence 2005) and the Agence Nationale de la Recherche 'Programme Physique et Chimie du Vivant' (project 'DNAnucl: Influence de la séquence ADN sur la structure et la dynamique du nucléosome', PCV2006).

References

Adams CC and Workman JL. 1995. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol Cell Biol. 15:1405–1421.

Albert I Mavrich TN, Tomsho LP Qi J, Zanton SJ, Schuster SC, Pugh BF. 2007. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. Nature. 446:572–576.

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. 2002. Molecular biology of the cell. 4th ed. New-York: Garland Publishing.

Allan J, Fraser R, Owen-Hughes T, Keszenman-Pereyra D. 2012. Micrococcal nuclease does not substantially bias nucleosome mapping. J Mol Biol. 417:152–164.

Angelov D, Molla A, Perche P-Y, Hans F, Cote J, Khochbin S, Bouvet P, Dimitrov S. 2003. The histone variant macroH2A interferes with transcription factor binding and SWI/SNF nucleosome remodeling. Mol Cell. 11:1033–1041.

Anselmi C, Bocchinfuso G, Santis PD, Savino M, Scipioni A. 2000. A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. Biophys J. 79:601–613.

Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton Carafa Y, Thermes C. 2011. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. Phys Rep. 498:45–188.

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

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

Audit B, Vaillant C, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2004. Wavelet analysis of DNA bending profiles reveals

structural constraints on the evolution of genomic sequences. J Biol Phys. 30:33–81.
Bao Y, White CL, 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.
Baxter R. 1982. Exactly solved models in statistical mechanics. London: Academic Press.
Bednar J, Horowitz RA, Grigoryev SA, Carruthers LM, Hansen JC, Koster AJ, Woodcock CL. 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 USA. 95:14173–14178.
Bernardi G. 1989. The isochore organization of the human genome. Annu Rev Genet. 23:637–661.
Bernardi G. 1995. The human genome: Organization and evolutionary history. Annu Rev Genet. 29:445–476.
Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene. 241:3–17.
Bernardi G. 2001. Misunderstandings about isochores. Part 1. Gene. 276:3–13.
Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL. 2004. Global nucleosome occupancy in yeast. Genome Biol. 5:R62.
Blank TA, Becker PB. 1995. Electrostatic mechanism of nucleosome spacing. J Mol Biol. 252:305–313.
Boeger H, Griesenbeck J, Kornberg RD. 2008. Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. Cell. 133:716–726.
Burcin M, Arnold R, Lutz M, Kaiser B, Runge D, Lottspeich F, Filippova GN, Lobanenkov VV, Renkawitz R. 1997. Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. Mol Cell Biol. 17:1281–1288.
Calladine CR, Drew HR. 1999. Understanding DNA. San Diego: Academic Press.
Chereji RV, Tolkunov D, Locke G, Morozov AV. 2011. Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions. Phys Rev E: Stat Nonlin Soft Matter Phys. 83:050903.
Chevereau G. 2010. Thermodynamique du positionnement des nucléosomes [Ph.D. thesis]. [Lyon (France)]: Université de Lyon-Ecole Normale Supérieure.
Chevereau G, Palmeira L, Thermes C, Arneodo A, 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 AE, Sperling SR, Vingron M. 2010. The effect of micrococcal nuclease digestion on nucleosome positioning data. PLoS One. 5:e15754.
Clapier CR, Cairns BR. 2009. The biology of chromatin remodeling complexes. Annu Rev Biochem. 78:273–304.
Davis HT. 1990. Density distribution functions of confined Tonks–Takahashi fluids. J Chem Phys. 93:4339–4344.
Dekker J. 2008. Mapping *in vivo* chromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction. J Biol Chem. 283:34532–34540.
Deniz O, Flores O, Battistini F, Pérez A, Soler-Lòpez M, Orozco M. 2011. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. BMC Genomics. 12:489.
Depken M, Schiessel H. 2009. Nucleosome shape dictates chromatin fiber structure. Biophys J. 96:777–784.
Dingwall C, Lomonossoff GP, Laskey RA. 1981. High sequence specificity of micrococcal nuclease. Nucleic Acids Res. 9:2659–2673.
Dong F, Hansen JC, van Holde KE. 1990. DNA and protein determinants of nucleosome positioning on sea urchin 5S rRNA gene sequences *in vitro*. Proc Natl Acad Sci USA. 87:5724–5728.
Dorigo B, Schalch T, Kulangara A, Duda S, Schroeder RR, Richmond TJ. 2004. Nucleosome arrays reveal the two-start organization of the chromatin fibers. Science. 306: 1571–1573.
Duret L, Eyre-Walker A, Galtier N. 2006. A new perspective on isochore evolution. Gene. 385:71–74.
Fan X, Moqtaderi Z, Jin Y, Zhang Y, Liu XS, 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 USA. 107: 17945–17950.
Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol. 4:e1000216.
Fire A, Alcazar R, Tan F. 2006. Unusual DNA structures associated with germline genetic activity in *Caenorhabditis elegans*. Genetics. 173:1259–1273.
Floer M, Wang X, Prabhu V, Berrozpe G, Narayan S, Spagna D, Alvarez D, Kendall J, Krasnitz A, Stepansky A, Hicks J, Bryant GO, Ptashne M. 2010. A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. Cell. 141:407–418.
Fu Y, Sinha M, Peterson CL, Weng Z. 2008. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. 4:e1000138.
Gabrielian A, Pongor S. 1996. Correlation of instrinsic DNA curvature with DNA property periodicity. FEBS Lett. 393: 65–68.
Garcia JF, Dumesic PA, Hartley PD, El-Samad H, Madhani HD. 2010. Combinatorial, site-specific requirement for heterochromatic silencing factors in the elimination of nucleosome-free regions. Genes Dev. 24: 1758–1771.
Giaquinta P. 2008. Entropy and ordering of hard rods in one dimension. Entropy. 10:248–260.
Gkikopoulos T, Schofield P, Singh V, Pinskaya M, Mellor J, Smolle M, Workman JL, Barton GJ, Owen-Hughes T. 2011. A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. Science. 333: 1758–1760.
Goodsell DS, Dickerson RE. 1994. Bending and curvature calculations in B-DNA. Nucleic Acids Res. 22: 5497–5503.
Hall MA, Shundrovsky A, Bai L, Fulbright RM, Lis JT, Wang MD. 2009. High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. Nat Struct Mol Biol. 16:124–129.
Hamiche A, Kang JG, Dennis C, Xiao H, Wu C. 2001. Histone tails modulate nucleosome mobility and regulate ATP-dependent nucleosome sliding by NURF. Proc Natl Acad Sci USA. 98:14316–14321.
Hansen J, McDonald I. 2006. Theory of simple liquids. London: Academic Press.
Hartley PD, Madhani HD. 2009. Mechanisms that specify promoter nucleosome location and identity. Cell. 137:445–458.
Hörz W, Altenburger W. 1981. Sequence specific cleavage of DNA by micrococcal nuclease. Nucleic Acids Res. 9: 2643–2658.
Ioshikhes I, Bolshoy A, Derenshteyn K, Borodovsky M, Trifonov EN. 1996. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J Mol Biol. 262:129–139.

Ioshikhes IP, Albert I, Zanton SJ, Pugh BF. 2006. Nucleosome positions predicted through comparative genomics. Nat Genet. 38:1210–1215.

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

Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ. 2006. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. Genome Res. 16:1505–1516.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 458:362–366.

Kenigsberg E, Bar A, Segal E, 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, Rippe K. 2008. Nucleosome geometry and internucleosomal interactions control the chromatin fiber conformation. Biophys J. 95:3692–3705.

Klenova EM, Nicolas RH, Paterson HF, Carne AF, Heath CM, Goodwin GH, Neiman PE, Lobanenkov VV. 1993. CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. Mol Cell Biol. 13:7612–7624.

Koerber RT, Rhee HS, Jiang C, Pugh BF. 2009. Interaction of transcriptional regulators with specific nucleosomes across the *Saccharomyces* genomes. Mol Cell. 35:889–902.

Kornberg RD, Lorch Y. 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosomes. Cell. 98:285–294.

Kornberg RD, Stryer L. 1988. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. Nucleic Acids Res. 16:6677–6690.

Lam FH, Steger DJ, O'Shea EK. 2008. Chromatin decouples promoter threshold from dynamic range. Nature. 453:246–250.

Lander ES et al. 2001. Initial sequencing and analysis of the human genomes. Nature. 409:860–921.

Längst G, Bonte EJ, Corona DF, Becker PB. 1999. Nucleosome movement by CHRAC and ISWI without disruption or trans-displacement of the histone octamer. Cell. 97:843–852.

Lantermann AB, Straub T, Strålfors A, Yuan G-C, Ekwall K, 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.

Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet. 36:900–905.

Lee JT. 2003. Molecular links between X-inactivation and autosomal imprinting: X-inactivation as a driving force for the evolution of imprinting? Curr Biol. 13:R242–R254.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. 2007. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet. 39:1235–1244.

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

Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. Cell. 128:707–719.

Li W, Bernaola-Galván P, Carpena P, Oliver JL. 2003. Isochores merit the prefix iso. Comput Biol Chem. 27:5–10.

Lieb E, Mattis D. 1966. Mathematical physics in one dimension. London: Academic Press.

Lionnet T, Dawid A, Bigot S, Barre F-X, Saleh OA, Heslot F, Allemand J-F, Bensimon D, Croquette V. 2006. DNA mechanics as a tool to probe helicase and translocase activity. Nucleic Acids Res. 34:4232–4244.

Lorch Y, Zhang M, Kornberg RD. 1999. Histone octamer transfer by a chromatin-remodeling complex. Cell. 96:389–392.

Lowary PT, Widom J. 1997. Nucleosome packaging and nucleosome positioning of genomic DNA. Proc Natl Acad Sci USA. 94:1183–1188.

Lowary PT, 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, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 389:251–260.

Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. 2008a. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. 18:1073–1083.

Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, Gilmour DS, Albert I, Pugh BF. 2008b. Nucleosome organization in the *Drosophila* genome. Nature. 453:358–362.

Mergell B, Everaers R, 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, Grange T. 2008. DNA physical properties determine nucleosome occupancy from yeast to fly. Nucleic Acids Res. 36:3746–3756.

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

Miller JA, Widom J. 2003. Collaborative competition mechanism for gene activation in vivo. Mol Cell Biol. 23:1623–1632.

Mito Y, Henikoff JG, Henikoff S. 2007. Histone replacement marks the boundaries of cis-regulatory domains. Science. 315:1408–1411.

Möbius W, Gerland U. 2010. Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites. PLoS Comput Biol. 6:e1000891.

Montel F, Fontaine E, St-Jean P, Castelnovo M, Faivre-Moskalenko C. 2007. Atomic force microscopy imaging of SWI/SNF action: mapping the nucleosome remodeling and sliding. Biophys J. 93:566–578.

Moreira JM, Holmberg S. 1999. Transcriptional repression of the yeast CHA1 gene requires the chromatin-remodeling complex RSC. EMBO J. 18:2836–2844.

Morozov AV, Fortney K, Gaykalova DA, Studitsky VM, Widom J, Siggia ED. 2009. Using DNA mechanics to predict in vitro nucleosome positions and formation energies. Nucleic Acids Res. 37:4707–4722.

Morse RH. 2007. Transcription factor access to promoter elements. J Cell Biochem. 102:560–570.

Moshkin YM, Chalkley GE, Kan TW, Reddy BA, Ozgur Z, van Ijcken WFJ, Dekkers DHW, Demmers JA, Travers AA, Verrijzer CP. 2012. Remodelers organize cellular chromatin by counteracting intrinsic histone-DNA sequence preferences in a class-specific manner. Mol Cell Biol. 32:675–688.

Mouchiroud D, Bernardi G. 1993. Compositional properties of coding sequences and mammalian phylogeny. J Mol Evol. 37:109–116.

Moukhtar J, Faivre-Moskalenko C, Milani P, Audit B, Vaillant C, Fontaine E, Mongelard F, Lavorel G, St-Jean P, Bouvet P, Argoul F, Arneodo A. 2010. Effect of genomic long-range

correlations on DNA persistence length: from theory to single molecule experiments. J Phys Chem B. 114:5125–5143.
Moukhtar J, Fontaine E, Faivre-Moskalenko C, Arneodo A. 2007. Probing persistence in DNA curvature properties with atomic force microscopy. Phys Rev Lett. 98:178101.
Moukhtar J, Vaillant C, Audit B, Arneodo A. 2009. Generalized wormlike chain model for long-range correlated heteropolymers. Europhys Lett. 86:48001.
Moukhtar J, Vaillant C, Audit B, Arneodo A. 2011. Revisiting polymer statistical physics to account for the presence of long-range-correlated structural disorder in 2D DNA chains. Eur Phys J E: Soft Matter. 34:119.
Noll M, Kornberg RD. 1977. Action of micrococcal nuclease on chromatin and the location of histone H1. J Mol Biol. 109:393–404.
Ohlsson R, Bartkuhn M, Renkawitz R. 2010. CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin. Chromosoma. 119:351–360.
Ozsolak F, Song JS, Liu XS, Fisher DE. 2007. High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol. 25:244–248.
Payen C, Fischer G, Marck C, Proux C, Sherman DJ, Coppée J-Y, Johnston M, Dujon B, Neuvéglise C. 2009. Unusual composition of a yeast chromosome arm is associated with its delayed replication. Genome Res. 19:1710–1721.
Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. 2007. Nucleosome positioning signals in genomic DNA. Genome Res. 17:1170–1177.
Percus JK. 1976. Equilibrium state of a classical fluid of hard rods in an external field. J Stat Phys. 15:505–511.
Percus JK. 1982. One-dimensional classical fluid with nearest-neighbor interaction in arbitrary external field. J Stat Phys. 28: 67.
Peterson CL, Workman JL. 2000. Promoter targeting and chromatin remodeling by the SWI/SNF complex. Curr Opin Genet Dev. 10:187–192.
Piasecki J, Peliti L. 1993. Harmonic properties of hard-spheres crystals: a one-demensional study. J Phys A: Math Gen. 26:4819–4825.
Polach KJ, Widom J. 1996. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. J Mol Biol. 258:800–812.
Pusarla R-H, Vinayachandran V, Bhargava P. 2007. Nucleosome positioning in relation to nucleosome spacing and DNA sequence-specific binding of a protein. FEBS J. 274:2396–2410.
Radman-Livaja M, Rando OJ. 2010. Nucleosome positioning: how is it established, and why does it matter? Dev Biol. 339:258–266.
Rando OJ, Ahmad K. 2007. Rules and regulation in the primary structure of chromatin. Curr Opin Cell Biol. 19: 250–256.
Raveh-Sadka T, Levo M, Segal E. 2009. Incorporating nucleosomes into thermodynamic models of transcription regulation. Genome Res. 19:1480–1496.
Rayleigh Lord (1891). On the virial of a system of hard colliding bodies. Nature London. 45:80–82.
Richard G-F, Kerrest A, Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 72:686–727.
Richmond TJ, Davey CA. 2003. The structure of DNA in the nucleosome core. Nature. 423:145–150.
Riposo J, Mozziconacci J. 2012. Nucleosome positioning and nucleosome stacking: two faces of the same coin. Mol Biosyst. 8:1172–1178.
Rippe K, Schrader A, Riede P, Strohner R, Lehmann E, Längst G. 2007. DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. Proc Natl Acad Sci USA. 104:15635–15640.
Robinson P, Fairall L, Huynh V, Rhodes D. 2006. EM measurements define the dimensions of the "30-nm" chromatin fibre: evidence for a compact, interdigitated structures. Proc Natl Acad Sci USA. 103:6506–6511.
Robledo A, Rowlinson J. 1986. The distribution of hard rods on a line of finite length. Mol Phys. 58:711–721.
Salsburg ZW, Kirkwood J, Zwanzig R. 1953. Molecular distribution functions in a one-dimensional fluid. J Chem Phys. 21:1098.
Santis PD, Morosetti S, Scipioni A. 2010. Prediction of nucleosome positioning in genomes: limits and perspectives of physical and bioinformatic approaches. J Biomol Struct Dyn. 27:747–764.
Satchwell SC, Drew HR, Travers AA. 1986. Sequence periodicities in chicken nucleosome core DNA. J Mol Biol. 191:659–675.
Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell. 132:887–898.
Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J. 2006. A genomic code for nucleosome positioning. Nature. 442:772–778.
Segal E, Widom J. 2009a. Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol. 19: 65–71.
Segal E, Widom J. 2009b. From DNA sequence to transcriptional behaviour: a quantitative approach. Nat Rev Genet. 10: 443–456.
Segal E, Widom J. 2009c. What control nucleosome positions? Trends Genet. 25:335–343.
Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol. 6:e65.
Shlyakhtenko LS, Lushnikov AY, Lyubchenko YL. 2009. Dynamics of nucleosomes revealed by time-lapse atomic force microscopy. Biochemistry. 48:7842–7848.
Solis FJ, Bash R, Yodh J, Lindsay SM, 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.
Strick TR, Dessinges MN, Charvin G, Dekker NH, Allemand JF, Bensimon D, Croquette V. 2003. Stretching of macromolecules and proteins. Rep Prog Phys. 66:1–45.
Suter B, Schnappauf G, 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.
Takahashi H. 1942. A simple method for treating the statistical mechanics of one-dimensional substances. Proc Phys Math Soc Jpn. 24:60.
Teif VB, Rippe K. 2009. Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. Nucleic Acids Res. 37:5641–5655.
Thåström A, Bingham LM, Widom J. 2004. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. J Mol Biol. 338:695–709.
Thåström A, Lowary PT, Widlund HR, Cao H, Kubista M, Widom J. 1999. Sequence motifs and free energies of selected natural

and non-natural nucleosome positioning DNA sequences. J Mol Biol. 288:213–229.
Tillo D, Hughes TR. 2009. G + C content dominates intrinsic nucleosome occupancy. BMC Bioinforma. 10:442.
Tirosh I, Barkai N. 2008. Two strategies for gene regulation by promoter nucleosomes. Genome Res. 18:1084–1091.
Tolkunov D, Morozov AV. 2010. Genomic studies and computational predictions of nucleosome positions and formation energies. Adv Protein Chem Struct Biol. 79:1–57.
Tolstorukov MY, Colasanti AV, McCandlish DM, Olson WK, Zhurkin VB. 2007. A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. J Mol Biol. 371:725–738.
Tonks L. 1936. The complete equation of state of one, two and three-dimensional gases of hard elastic spheres. Phys Rev. 50:955–963.
Travers A. 1999. An engine for nucleosome remodeling. Cell. 96:311–314.
Tsankov A, Yanagisawa Y, Rhind N, Regev A, Rando OJ. 2011. Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. Genome Res. 21:1851–1862.
Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. 2010. The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol. 8:e1000414.
Tsukiyama T, Wu C. 1997. Chromatin remodeling and transcription. Curr Opin Genet Dev. 7:182–191.
Vaillant C, Audit B, Arneodo A. 2005. Thermodynamics of DNA loops with long-range correlated structural disorder. Phys Rev Lett. 95:068101.
Vaillant C, Audit B, 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, Arnéodo A. 2006. Formation and positioning of nucleosomes: effect of sequence-dependent long-range correlated structural disorder. Eur Phys J E: Soft Matter 19:263–277.
Vaillant C, Palmeira L, Chevereau G, Audit B, d'Aubenton Carafa Y, Thermes C, 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 JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM. 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 SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. Nature. 474:516–520.
van Holde KE. 1988. Chromatin. New York: Springer-Verlag.
Vanderlick TK, Scriven LE, Davis HT. 1986. Solution of Percus equation for the density of hard-rods in an external-field. Phys Rev A. 34:5130–5131.
Vashee S, Melcher K, Ding WV, Johnston SA, Kodadek T. 1998. Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein–protein interactions. Curr Biol. 8:452–458.
Wang MD, Schnitzer MJ, Yin H, Landick R, Gelles J, Block SM. 1998. Force and velocity measured for single molecules of RNA polymerase. Science. 282:902–907.
Wang X, Bai L, Bryant GO, Ptashne M. 2011a. Nucleosomes and the accessibility problem. Trends Genet. 27:487–492.
Wang X, Bryant GO, Floer M, Spagna D, Ptashne M. 2011b. An effect of DNA sequence on nucleosome occupancy and removal. Nat Struct Mol Biol. 18:507–509.
Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N. 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res. 20:90–100.
Whitehouse I, Flaus A, Cairns BR, White MF, Workman JL, Owen-Hughes T. 1999. Nucleosome mobilization catalysed by the yeast SWI/SNF complex. Nature. 400:784–787.
Whitehouse I, Rando OJ, Delrow J, Tsukiyama T. 2007. Chromatin remodelling at promoters suppresses antisense transcription. Nature. 450:1031–1035.
Whitehouse I, Tsukiyama T. 2006. Antagonistic forces that position nucleosomes *in vivo*. Nat Struct Mol Biol. 13:633–640.
Widom J. 1996. Short-range order in two eukaryotic genomes: relation to chromosome structure. J Mol Biol. 259:579–588.
Widom J. 2001. Role of DNA sequence in nucleosome stability and dynamics. Q Rev Biophys. 34:269–324.
Wolffe AP. 1998. Chromatin Structure and Function, 3rd ed. London: Academic Press.
Woodcock CL, Skoultchi AI, 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, Liu JS. 2008. Genomic sequence is highly predictive of local nucleosome depletion. PLoS Comput Biol. 4:e13.
Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science. 309:626–630.
Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, 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 Z, Pugh BF. 2011. High-resolution genome-wide mapping of the primary structure of chromatin. Cell. 144:175–186.
Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, Pugh BF. 2011. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. Science. 332: 977–980.