In silico design of DNA sequences for in vivo nucleosome positioning

Ethienne Routhier, Alexandra Joubert, Alex Westbrook, Edgar Pierre, Astrid Lancrey, Marie Cariou, jean-baptiste Boulé, Julien Mozziconacci

Published online DD MM YYYY

Nucleic Acids Research, YYYY, Vol. xx, No. xx ***1–8***
doi:10.1093/nar/gkn000

In silico design of DNA sequences for *in vivo* nucleosome positioning

Ethienne Routhier [1], Alexandra Joubert [2] Alex Westbrook [2] Edgar Pierre [1] Astrid Lancrey [2] Marie Cariou [3] Jean-Baptiste Boulé [2,*] Julien Mozziconacci [1,2,3,4 *]

[1]Laboratoire de Physique Théorique, CNRS, Sorbonne Université, Paris, France de la Matière Condensée, CNRS, Sorbonne Université, Paris, France and [2]Structure et Instabilité des Génomes, Museum National d'Histoire Naturelle, CNRS, INSERM, Paris, France [3] Acquisition et Analyse de données pour l'histoire naturelle, Museum National d'Histoire Naturelle, CNRS, Paris, France [4]Institut Universitaire de France, Paris, France

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

The computational design of synthetic DNA sequences with designer in vivo properties is gaining traction in the field of synthetic genomics. We propose here a computational method which combines a kinetic Monte Carlo framework with a deep mutational screening based on deep learning predictions. We apply our method to build regular nucleosome arrays with tailored nucleosomal repeat lengths (NRL) in yeast. Our design was validated in vivo by successfully engineering and integrating thousands of kilobases long tandem arrays of computationally optimized sequences which could accommodate NRLs much larger than the yeast natural NRL (namely 197 and 237 bp, compared to the natural NRL of ∼165 bp). This method delineates readily the key sequence rules for nucleosome positioning in yeast and should be easily applicable to other sequence properties and other genomes.

INTRODUCTION

Recent biotechnology techniques such as CRISPR/Cas9 and DNA oligonucleotide *in vivo* assembly have opened ways to precisely and extensively modify genomes. Taking advantage of these technologies, several projects have been launched with the aim to partially or completely design and assemble synthetic genomes (1). Nevertheless, controlling chromatin assembly and gene expression on a synthetic genome remains a challenge. Decisive efforts have been recently made for designing promoter sequences that produce controlled levels of mRNA in yeast (2, 3) or the activity of enhancers in a Drosophila cell line (4). While the control of gene expression is now within our grasp, there is yet no efficient way to control by sequence design the positioning of nucleosomes along a synthetic DNA cassette inserted in an eukaryotic genome. Nucleosome positioning is however of crucial importance as it influences DNA accessibility to DNA binding factors involved in DNA replication and transcription, thus adding a supplementary level of control on top of the DNA sequence (5, 6). In a seminal experiment, Lowary and Widom used a SELEX approach to isolate a DNA sequence with the highest affinity to a histone octamer among a set of more than 5×10^{12} sequences (7). The resulting 147 bp sequence, known as the 601-sequence, has been extensively used to reconstruct regular nucleosomal arrays *in vitro* (e.g. (8)). Given the success in using the 601 sequence for nucleosome positioning in *vitro*, we recently addressed the affinity of nucleosomes to the 601 sequence *in vivo* using insertions in the *S. cerevisiae* genome of three long arrays of approximately 50 nucleosomes, with three different spacing between nucleosomes (167, 197 and 237 bp), also known as nucleosomal repeat length (NRL). In this former study we found that in sharp contrast with the *in vitro* experiments, the affinity of nucleosomes to the 601 sequence was very low *in vivo* (9) suggesting that this sequence is not an adequate tool to control nucleosome positions in synthetic genomics approaches. Computational tools are a good alternative to optimize the design of synthetic sequences from *in vivo* data. Among the available computational methods, deep learning has been widely applied to building predictive models that relate DNA sequences to genomic functions (10, 11). The ability of deep neural networks to predict annotations resulting from variations of a sequence is now used for *de novo* design of genomic sequences, including tailored alternative poly-adenylation sites (12, 13) or human 5'UTR sequences (14, 15).

Building on these previous studies, we use here a nucleosome occupancy predictor (16) together with a kinetic Monte-Carlo framework in order to design three sequences (of 167 bp, 197 bp and 237 bp) that lead to a regular nucleosome positioning when assembled into tandem repeated arrays *in vivo* (Fig.1). We extend here previous work (17) and aim at making longer arrays with variable NRL. The three NRL chosen encompassed the natural NRL of *S.cerevisiae* (167 bp) and the longest NRL known in eukaryotic species (237 bp, found in the sea urchin sperm (18)).

*To whom correspondence should be addressed. Tel: +33 000 0000000; Fax: +33 000 0000000; Email: julien.mozziconacci@mnhn.fr. Correspondence can also be addressed to jean-baptiste.boule@mnhn.fr

MATERIALS AND METHODS

Deep learning model training

The deep learning model, based on previous work (16), is a CNN architecture taking as input a one-hot encoded DNA sequence of 2001bp and outputing the nucleosome occupancy at the center of the sequence. The input passes through 3 convolutional layers with respectively 64, 16 and 8 kernels of size 3, 8 and 80 and reLU activation. After each convolution is applied a max-pooling layer with pooling size of 2, to reduce dimensionality, followed by batch-normalization and dropout with rate 0.2. Finally, the output of the last layer is flattened and passed through a dense layer with reLU activation for the the final prediction. The loss function combines Pearson's correlation (corr) and the mean absolute error (MAE) (loss = MAE[ŷ, y] + 1 - corr[ŷ, y], with ŷ being the model prediction and y the target). The modelwas trained in a supervised manner on the YPH499 genome with the nucleosome profiles from our previous study on 601 sequences (9), while ignoring the synthetic repeats. The profile was truncated to the 99th percentile of the distribution, and then normalized between 0 and 1. We train the model on all 2001bp-long sequences and their reverse complement, with the central nucleosome occupancy as label.

Synthetic sequence initialization

The initial sequences were determined at random. Each of the four (A,T,G,C) nucleotides was sequentially picked N times with a probability proportional to its abundance within the genome. This procedure insures that starting sequences have a GC content similar to the natural GC content of *S.cerevisiae*.

Synthetic sequence mutation

Our mutation and selection strategy for optimizing the sequence is inspired from the kinetic Monte-Carlo (k-MC) method originally designed for Ising spin systems ((20)). Once the sequence of length N is initialized, it is duplicated $3\times N$ times in order to create $3\times N$ new sequences which will each harbor one of the $3\times N$ single mutations of the sequence (Fig.1a). We then associate an energy term with each of the mutated sequences (see below). For every sequence i a selection probability Γ_i is then defined as follow:

$$\Gamma_i = \frac{1}{Z} e^{-E_i/T} \quad (1)$$

with Z being the normalisation factor ($Z = \sum_i e^{-E_i/T}$) and T the temperature, a broadening factor chosen by the user and corresponding to a classical temperature in a Maxwell-Boltzmann distribution. The next sequence is then randomly chosen according to this probability distribution. The i^{th} configuration is selected - with a probability Γ_i - and the process continues for a given number of steps (*e.g.* 100). Each chosen configuration is saved at each step, and the configuration with the minimal energy is chosen at the end.

Synthetic sequence energy term

An energy is associated to every sequence, representing how far the sequence is from having the desired nucleosome positioning characteristics. The energy can be divided into four parts (Fig.1 b), described below.

Nucleosome occupancy energy on the direct strand E_{reg}
In a former work we used a convolutional neural network (CNN) to predict the nucleosome occupancy at the center nucleotide of a 2001 bp long DNA sequence (16). In order to evaluate the nucleosome occupancy over our synthetic array, we first create a $2001+N$ bp long sequence by repeating the monomeric synthetic sequence. Then, this sequence is cut in N sequences of 2001 bp long used as inputs of the network thus providing the predicted nucleosome occupancy over the whole synthetic sequence. The nucleosome occupancy related energy is the distance between the predicted nucleosome occupancy on the synthetic sequence and a target occupancy. It forces the predicted occupancy to converge to the target occupancy (Fig.1 c) The distance is defined as :

$$d(x,y) = 1 - corr(x,y) + \frac{1}{N}\sum_i |x_i - y_i| \quad (2)$$

where $corr$ stands for the Pearson correlation and N is the length of the sequence. The target signal y^{target} starts with the predicted nucleosome position (Gaussian coverage distribution) and end with the predicted linker (uniform, lower coverage distribution) :

$$\begin{aligned} y_i^{target} &= a.e^{-\frac{(i-73)^2}{2\sigma^2}} + b \text{ if } i \leq 147 \\ y_i^{target} &= a.e^{-\frac{(147-73)^2}{2\sigma^2}} + b \text{ if } i > 147 \end{aligned} \quad (3)$$

The parameters a,b and σ are chosen in order to create a realistic target (i.e which shape is similar to canonical nucleosomal occupancies found on the genome). The results are shown for $a=0.4$, $b=0.2$ and $\sigma = \frac{147}{4}$.

GC content related energy E_{GC}
As said previously, there are 4^N different sequences of a given length N. The CNN learned to position nucleosomes for sequences within the genome of *saccharomyces cerevisiae*. As a result, one wants to get a synthetic sequence which has a rather similar GC content as natural yeast. In order to do so, a constraint energy was first calculated as follow:

$$E_{GC} = \sqrt{(GC_{synt} - GC_{nat})^2} \quad (4)$$

with GC_{synt} the GC content of the synthetic sequence and GC_{nat} the GC content of the natural yeast (in that case,

$GC_{nat} \approx 0.38$).

Enhanced sampling mutation energy E_{mut}
It is common that such Monte Carlo procedure can get stuck into a local minimum of the energy, for example by flipping between the same two mutations. To avoid this potential pitfall, we added a term to the energy term that penalizes sequences that were already generated during the optimization process.

Nucleosome occupancy energy on the reverse strand E_{rev}
This energy is calculated in the same way as the nucleosome occupancy related energy on the direct strand, but using the reverse complement strand instead. On most sequences this term is not relevant since the nucleosomal occupancy predicted by the network is, due to the training process, the same for both strands. However, we added this term to add an extra penalty to sequences for which the network would not work as expected and would predict different occupancies for the two strands.

E_i in Eq.**1** can then be written as:

$$E_i = \alpha E_{GC_i} + \beta E_{reg_i} + \gamma E_{rev_i} + \delta E_{mut_i} \quad (\mathbf{5})$$

with α, β, γ and δ the relative weights for each energy component, chosen to get the same order of magnitude for each energy term. In our case energies are within the same range and we used a value of 1 for each weight.

Interpretation of the mutations selected during the optimization process.

Positions of the mutations along the sequence
A thousand independent k-MC optimization processes were executed and stopped after 20 steps, corresponding to the typical number after which the energy remains stable. At each time step, the energy associated with each possible mutation was computed and stored. To quantify the importance of a given nucleotide in the positioning capacity of the sequence, the average absolute energy variation created by a mutation at this specific position was computed.

Typical mutation logos
For each monomer length (167, 197 and 237 bp), we collected all 5-bp long motifs centered at each mutation site during 1000 independent optimization processes stopped after 20 steps. We then selected motifs for which a mutation induces an absolute energy variation $\Delta E = E_j - E_i$ (where j is the mutated sequence) higher than 0.5 (see **Supplementary Figure S1** for the distribution of the absolute energy changes). We then grouped the collected motifs during these 1000 independent optimisations into four categories depending on the sign of the energy modification and on their position in the linker region or in the dyad region (respectively after 147 bp and between 50 bp and 100 bp). We ended up with 14938 negative mutations in the linker, 16165 negative mutations in the dyad, 52603 positives mutations in the dyad and 158422 positive mutations in the linker. For each of these four categories we extracted significant logos of 5 bp using STREME (19, 21) with the argument w=5. We also provided a list of control sequence to STREME by passing the list of 5 bp motifs associated with mutations that modify the energy by less than 0.03. Finally, we selected logos that matched more than 10 % of the input sequences and represented them before and after the mutation.

In vivo assembly of the computationally designed sequences

Strains, plasmids, reagents and media
In vivo genomic assembly of synthetic sequences nucleosome positioning (SSNP, **Supplementary Table 1**) DNA repeats was performed in the yeast strain YPH499. Strains derived from this study are listed in **Supplementary Table 2**. Plasmids used for the *in vivo* expression of spCas9 and the guide RNA targeting the YMR262 gene have been previously described (22) and are listed in **Supplementary Table 3**. Strains were grown at 30°C in yeast extract Peptone Dextrose 2% media (YPD) or in the appropriate synthetic complete Dextrose 2% media (SCD) minus relevant amino acids necessary to maintain plasmid borne auxotrophic markers. All media reagents were purchased from Formedium and used as recommended. Oligonucleotides used in this study were synthetized by Eurogentec. Enzymes for nucleic acids modification were purchased from New England Biolabs. Zymolyase 20T was purchased from Amsbio.

Assembly of synthetic nucleosome positioning DNA repeats in the yeast genome
In vivo assembly using CRISPR/Cas9 and overlapping oligonucleotides in the Chromosome XIII of *S. cerevisiae* strain YPH499 was performed as previously described (9, 22). In summary, donor DNA containing the left (YMR/SSNP) and the right (SSNP/YMR) genomic junction were amplified by PCR using primers couples O-1/O-2(O-3/O-4/O-5) and O-10/O-6(O-7/O-8/O-9) respectively. This donor DNA results, upon recombinational assembly, in the deletion of the region -129 to 232 bp of the YMR gene. All oligonucleotides used for *in vivo* repeat assembly are listed in **Supplementary Table 4**. YPH499 was transformed with the Cas9 expressing plasmid pRS413-Cas9-His (AJ-P1). The resulting strain was transformed using the LiAc technique (23) with 1 µg of gRNA expressing plasmid targeting YMR262 (AJ-P2), 100 pmol of each of the four or six appropriate SSNP-oligonucleotides, and 10 pmol of both YMR/SSNP left and right junction PCR. After transformation cells were plated on SCD-His-Ura to select cells carrying both CAS9 and gRNA expressing plasmids. The in vivo assembly of synthetic repeated arrays was verified by left and right junction PCR amplification (O-1/O-2 ; O-3 ; O-4 ; O-5 and O-10/O-6 ; O-7 ; O-8 ; O-9) and Sanger sequencing. The correct locus and the size of the SSNP assembly was confirm by analyzed recombinant clones by southern-blotting. Genomic DNA from recombinant strains were digested with BamHI and DraI, which cut at each side of the insertion locus. Digested DNA was electrophoresed in 1% agarose and transferred by capillarity onto a nylon membrane (Hybond N+, GE healthcare). Membranes were hybridized in Church buffer at 68°C with a Cy5-labelled genomic probe ((24)). The genomic probe was a 1 kb DNA fragment amplified by PCR from genomic DNA using primers O-33/O-34 and Cy5-dCTP (see **Supplementary Figure S4**). Membranes were scanned using a FLA 9500 GE healthcare.

Mononucleosome preparation using Micrococcal Nuclease digestion and sequencing (MNase-Seq)
Each strain was grown to a cell density of $0.8*10^7$ cells/mL in 250 mL SCD media at 30°C with 200 rpm shaking. Cultures were treated with a final concentration of 1.85 % formaldehyde for 30 min at 30°C. Cross-linking was stopped by addition of 105 mM Glycine (final concentration). Cell pellets (6500g, 10 min) were washed and resuspended in 50 mL of 1 M Sorbitol, 10 mM Tris pH 7.5 supplemented with 10 mM ß-mercaptoethanol and 15 mg of Zymolyase 20T, and incubated 1h at 30°C with 50 rpm shaking. Spheroplasts were pelleted (6500g, 10 min), and lysed in 2.4 mL solution containing 1M Sorbitol, 50 mM NaCl, 10 mM Tris pH 7.5, 5 mM MgCl2, 1 mM CaCl2 and 0.75 % Igepal CA630 freshly supplemented with 1 mM ß-mercaptoethanol, 500 µM spermidine and 3000 units of MNase. The spheroplasts/MNase mixture was then incubated at 37°C for 30 min and stopped by adding 600 µL of 1% SDS, 10 mM EDTA. Reversal of crosslink and protein removal was achieved by adding 0.6 mg of Proteinase K (ThermoFisher) and overnight incubation at 65°C. DNA was purifed using phenol/chloroform extraction and salt precipitation method. Mononucleosomal DNA was isolated on a 1.8 % agarose electrophoresis gel and purified using QIAquick Gel extraction Kit. DNA concentration was determined using a Qubit Fluorometer and samples were processed and sequenced on a NovaSeq 6000 S4 PE150 XP by the Eurofins sequencing platform (NGSelect Amplicons).

RNA libraries preparation and sequencing Total RNA was extracted by starting from 25 mL culture of each strain grown to an OD_{600} of 0.5 in SCD media at 30°C with shaking at 200 rpm. Total RNA was then purified using hot acidic phenol method (25). After a rRNA depletion step using respectively the Ribominus™ Transcriptome Isolation Kit (Invitrogen, K1550-03) and the RiboCop rRNA Depletion Kit for Yeast (Lexogen, 190), respectively for the first and the second biological replicate, RNA concentration and quality was determined using respectively Qubit™RNA HS assay kit and Qubit™RNA IQ Assay kit with a Qubit®fluorometer and standard calibration and assay protocols provided by the manufacturer. RNA sequencing was achieved by Eurofins following their NGSelect RNA protocol. cDNA libraries (300 bp) were then sequenced on a NovaSeq 6000 S4 PE150 XP (2x150bp) by Eurofins.

Construction of synthetic reference genomes
For the three synthetic sequences we constructed a reference genome with the assembly of YPH499 and an additional scaffold carrying 7 tandem repeats with its 4kb flanking regions on each side.

Reads alignment
After the removal of barcodes with cutadapt (version 3.4) with parameters -m 50 -O 1, paired-end reads of 151 (Syn) or 66 bp (601) were mapped against the appropriate reference genome using Bowtie2 (version 2.2.5) (26, 27). We allowed a maximum fragment size of 250 bp corresponding to the maximum length of purified fragments (-X 250). Read pairs are not filtered for single alignments so that Bowtie2 assigns a random repeat to each pair.

Reconstruction of sequence coverage over DNA repeats
The nucleosome occupancy along the genome was reconstructed using bamCoverage from the deeptools package (version 3.5.1) at base-resolution (parameter --binSize 1) from full fragments (--extendReads) with lengths between 140 and 170 bp (--minFragmentLength 140 --maxFragmentLength 170) to consider only fragments in the mono-nucleosomal band. It was also normalized in Count Per Million reads (--normalizeUsing CPM). Due to potential edge effects on the first and last repeats, the nucleosome occupancy on the DNA repeats is extracted from the 5 middle repeats.

Prediction of Open Reading Frames over the DNA repeats
Augustus (28) predicts genes ab initio from a eukaryotic genomic sequences based on a generalized Hidden Markov Model. For each insertion, we predicted gene positions on sequences corresponding to seven repeats of the insertion and 4kb flanking the insertion site (4kb en 5' and 4kb en 3') using Augustus and training annotation files for *saccharomyces cerevisiae* (other default parameters).

RESULTS AND DISCUSSION

Our computational model predicts the inefficiency of the Widom-601 sequence *in vivo*

In order to test the *in vivo* ability of the Widom-601 sequence to form regular nucleosome arrays of various NRLs we recently synthesize and integrated such arrays within a yeast chromosome (9). Our results unambiguously showed that the ability of the 601 sequence to strongly position nucleosomes is lost in *S.cerevisiae* (Fig.2 b). We used our nucleosome occupancy predictor (11) on the 601 sequence arrays used in this previous study and questioned whether it is able to predict the nucleosome occupancy on a completely exogenous sequence, namely the Widom-601 sequence. In agreement with the experimental nucleosome occupancies, the predictor anticipates that the Widom-601 sequence is unable to position nucleosomes *in vivo* (Fig.2 b). The predicted occupancy is relatively flat over the 167 bp long sequence, showing no evidence of nucleosome localization on the first 147 bp where the Widom-601 sequence is placed. Moreover, for both the 197 bp and the 237 bp long sequence, the predictor anticipated that the preferential position of the nucleosomes is around the linker, in agreement with the experimental results (Fig.2 c). The ability of our model to predict correctly the preferential positions of nucleosomes on exogenous sequences prompted to use it for a design purpose.

Optimization of the monomer sequences for regular nucleosome positioning in tandem arrays

Starting from a random sequence, we computed the effect that would have all possible single mutations on the predicted occupancy (Fig.1). We then selected one of these mutations in order to obtain a nucleosomal occupancy corresponding to nucleosomes preferentially occupying the first 147 bp while being excluded from the rest of the linker region (the target occupancy). We devised an energy function (Fig.1 b) based on the distance between the predicted nucleosomal occupancy on the tandem arrays and the target occupancy. We picked a (in

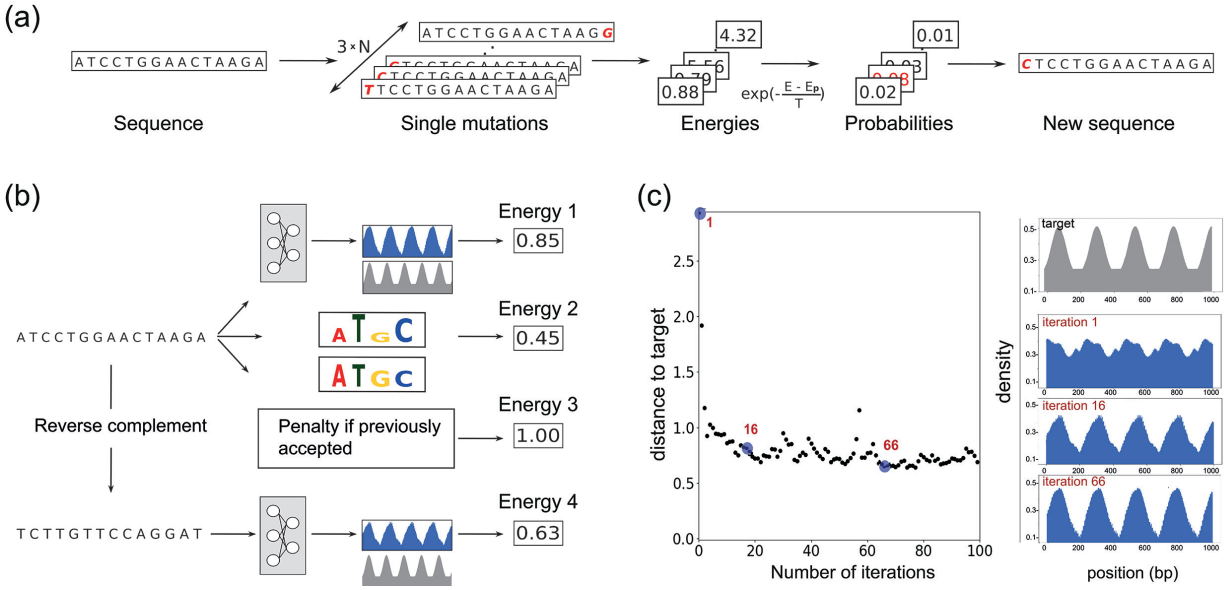

Figure 1. Sequence optimization via kinetic Monte-Carlo. (a) General principle of the method. (b) The energy associated to a sequence is the sum of four terms: (1) the distance between the nucleosome occupancy predicted by the model and the target nucleosome occupancy (2) the distance between the GC content of the sequence and the natural GC content of the yeast DNA (3) a penalty if the sequence has been already sampled (4) the distance between the occupancy predicted on the reversed complemented sequence and the target occupancy. (c) Evolution of the predicted nucleosome occupancy over several repeats (blue occupancies on the right) and of the corresponding energy for one repeat (corresponding points highlighted in blue) during the optimization process. The target occupancy is displayed at the top (grey).

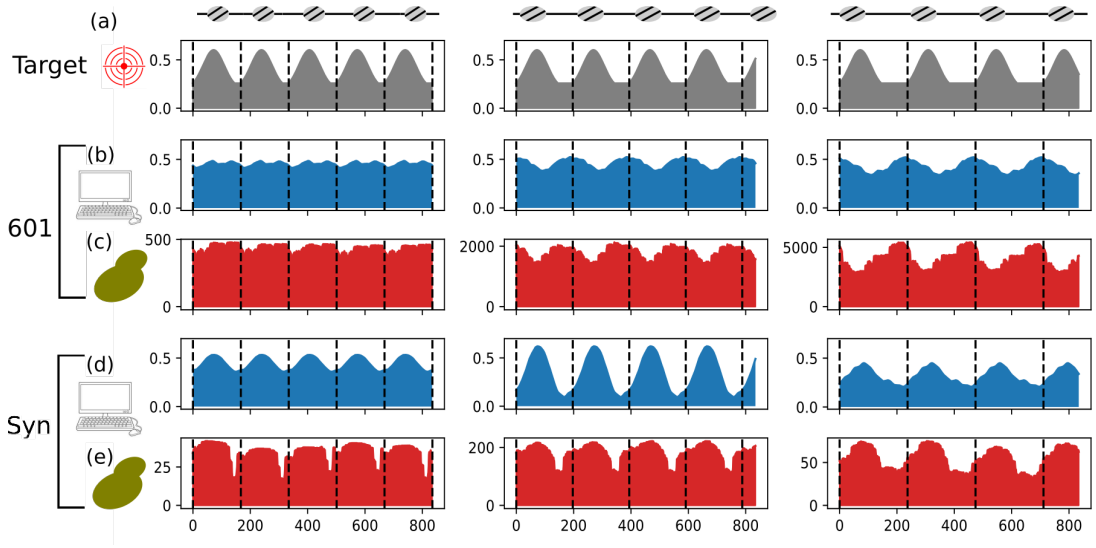

Figure 2. Target, predicted and experimental nucleosome occupancies on a 1000 bp long subset of Widom-601 repeats and of synthetic sequence repeats. (a) Target nucleosome occupancies corresponding (from left to right) to one nucleosome every 167 bp, 197 bp and 237 bp. (b,c) Normalized predicted (b) and experimental (c) nucleosome occupancy on a 1000 bp region of 601 repeats. (d,e) Normalized predicted (d) and experimental (e) nucleosome occupancy on the 1000 bp subset of our synthetic sequence repeats. Occupancies are normalized so that the area under the curve is equal to the area under the curve of the target occupancies

grey on Fig.1c). The overall energy diminishes during the first 20 to 50 steps of the k-MC optimisation process and is stable afterwards (Methods and Fig.1). The distance to the target was rapidly divided by a factor 3 during the 5 first mutations steps, then slowly converged and finally stabilised around a value of 0.6. The predicted nucleosome occupancies corresponding to three time points sampled during the optimisation process illustrate the convergence of the predicted occupancy towards the target function (Fig.1 c, right panels).

Interestingly, for all initial random sequences we used, 5 to 20 changes out of 167 197 or 237 bp were always sufficient to provide a sequence with nucleosome occupancy predicted close to our target sequence (**Supplementary figure S2**). Starting from random DNA seeds and after many round

of sequence optimisation, we could produce hundreds of positioning sequences with a predicted nucleosome occupancy that resembles the target function used (**Supplementary figure S3**). Three sequences corresponding to the three target occupancy we wished to impose on our tandem arrays were selected for further study.

Experimental validation of the positioning capabilities of the synthetic sequences

In order to validate *in vivo* the positioning efficiency of our three *in silico* optimized sequences, we performed MNase-seq experiments on three *S.cerevisiae* strains in which repeated arrays containing 167 bp, 197 bp or 237 bp long monomers were engineered into the non essential YMR262 gene in chromosome XIII (**Supplementary Figure S4**, Supplementary Methods and (9)).

We performed two technical replicates using the same strain and a biological replicate for which we re-selected a different strain, with a different number of insertions. While technical replicates show highly similar mononucleosomal length (**Supplementary figure S5**) and dyad positioning over the repeats (**Supplementary figure S6**), the two biological replicates show some degree of variations, potentially due to the different number of insertions but that nonetheless lead to similar conclusions exposed below.

The experimental nucleosomal occupancies (predicted and experimental) over the synthetic arrays, the Widom-601 arrays and the corresponding predicted occupancies are shown on Fig.2. These results show that there is a blatant similarity between the predicted occupancies and the measured occupancy in vivo with MNase digestion analysis (Fig.2 d,e). For all repeats, we observed a preferential positioning of the nucleosome in the first 147 bp of the synthetic sequence, with presence of a flat peak indicating that nucleosome dyads (i.e. the mid-points of the sequenced fragments) are sharply distributed around the center of the first 147 bp (**Supplementary Figure S6**). For the 197 bp long synthetic sequence, we also find a nucleosome well positioned on the first 147 bp of the sequence and a precisely positioned dyad (**Supplementary Figure S6**). For the 237 bp repeat, the experimental occupancy exhibited a peak on the first 150 bp followed by a constant low occupancy over 90 bp, similarly to the target. The central position of sequenced fragments distribution (**Supplementary Figure S6**) showed a strong enrichment in the center of the first 147 bp for the 167 and 197 bp repeats. For the 237 repeats, fragments centers are more broadly distributed so that the overall nucleosome occupancy is similar to the target we chose. All the linker regions exhibited low nucleosome occupancy, as predicted by our occupancy predictor.

Analysis of the optimisation process of the synthetic sequences

In order to understand the specific features selected by the optimization process, we performed a quantitative analysis of the mutations. We first identified which regions of the sequence were the most important to influence positioning by performing 1000 k-MCMC independent, 20 steps long optimisations and recorded at each step the energy changes associated with each of the possible mutation. High impact

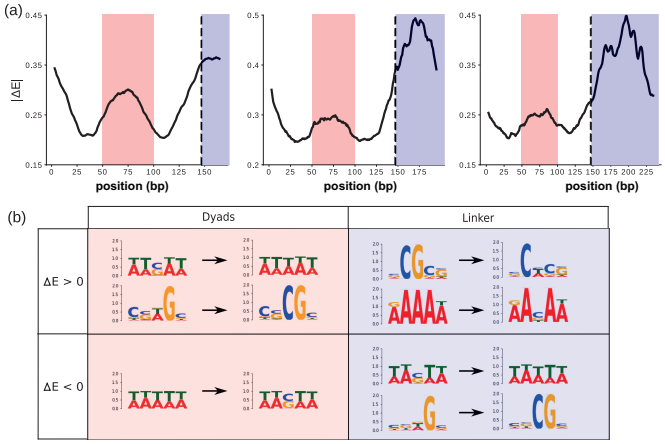

Figure 3. Positions on the sequence and typical motifs around important mutations during the optimisation process. (a) Average absolute change in energy due to mutations at each position along the 167 bp, 197 bp and 237 bp monomer (from left to right). (b) DNA motifs enriched around important mutations before and after running the optimisation process (respectively on the left and right of the black arrows).

mutations are associated with higher energy variations, either positively or negatively. The absolute value of the average ΔE along the sequence suggests that the most important regions for nucleosome positioning are the linker region (in blue in Fig.3 a) and the 50 bp region surrounding the dyad axis of the nucleosome (in red in Fig.3). In order to go further into the sequence rules that define nucleosome occupancy on our sequences, we studied the motifs corresponding to the most important mutations (Online Methods). We split mutations into four groups: positioned either in the linker or in the dyad region and either increasing or decreasing the energy term. We then searched for typical 5 bp long DNA motifs within each group using STREME (19). Two types of mutations could be extracted: mutations that construct one of the two known nucleosome repelling motives (poly(dA-dT) and poly(dCG)) and mutations that destruct these motives (16). The sign of the resulting energy change is dictated by the position of the nucleotide within the monomer sequence: in the linker region, destructive mutations increase the energy while constructive mutations decrease the energy. In the dyad region the opposite is true. These results are consistent with previous studies suggesting that nucleosome attractive sequences do not exist in *S.cerevisiae* genome and explain the inefficiency of the 601 sequence to position nucleosomes *in vivo* (5, 16).

Transcription of the synthetic arrays

Finally, we looked into the transcriptional status of the synthetic sequences to check if any of them could act as a promoter and sustain measurable transcription. We performed whole genome strand specific RNA-seq of the three strains and counted the transcripts initiated from the repeat, in comparison to surrounding genes. We found that only the 197 bp repeat yielded a high transcription output, leaking in the forward direction into the leftover of the YMR262 gene in which the array was inserted (Fig. 4 and **Supplementary Figure S4**). The 167 bp and 237 bp synthetic repeats, with respectively shorter and longer inter nucleosomal distance compared to the 197 bp repeat, do not show significant transcription. Although

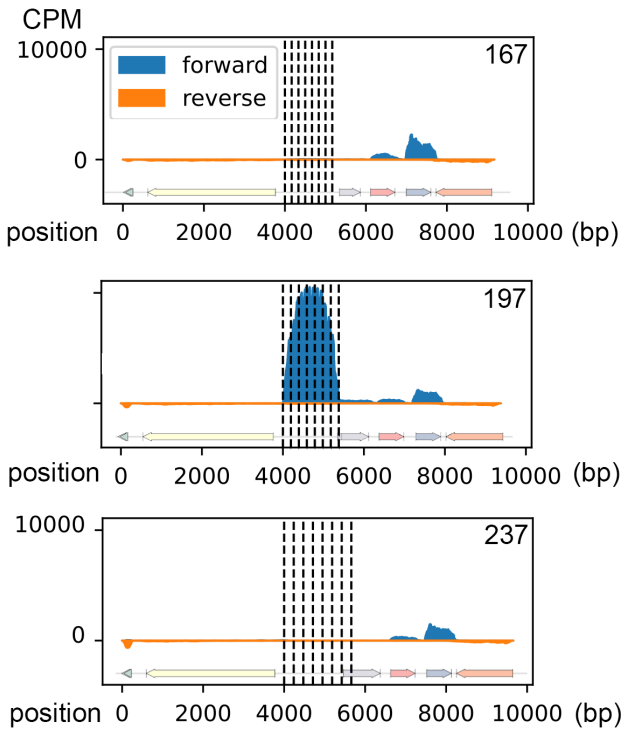

Figure 4. T**Transcription of the synthetic repetitive DNA arrays.** RNA fragments counts around the repeated region from one biological replicate of YPH499 containing 167, 197 and 237 bp repeats Augustus predictions of open reading frames are shown below the top tracks. The second biological replicate is presented in **Supplementary Figure S7**.

we can not rule out that pervasive transcription is present on the arrays (32), our results indicate that transcription is not strictly determined by an increase of NRL on the synthetic repeats. More importantly, the transcription of the 197 bp repeat shows that yeast transcription does not affect the predicted (and observed *in vivo*) nucleosome occupancy on the 197 bp array. To go further, we predicted open reading frames (ORFs) on the arrays and neighboring regions using the Augustus software (28). Using default parameters, Augustus predicted ORFs originating in both the 197 and 237 bp arrays. We hypothesize that the absence of transcripts in the 167 and 237 bp strain could be due to codon suboptimality of the transcribed sequence (33), or in a non exclusive manner due to lack of binding of any transcription factor. Analysis of the three sequences shows that they carry several start and stop codons, which makes them susceptible for nonsense-mediated mRNA decay (NMD) through recognition of premature stop codons (**Supplementary Figure S7**, (33)). To verify that the absence of detectable transcription is not due to the degradation of the RNAs produced, transcription could be measured in further experiments in Xrn1Δ or Rrp6Δ strains (34). Using YeasTract+ (35), we looked for potential transcription factor binding sites in the three monomer sequences. The results presented in **supplementary Figure S7** show that each repeat harbor several sites of potential binding sites for positive and negative factors. Interestingly, the 197 bp monomer contains a higher density of potential transcription binding sites, a first degree observation compatible with increased transcriptional activity on this repeat.

CONCLUSION

In this study, we used the k-MC heuristic combined with a deep learning nucleosome occupancy predictor to design an array of nucleosome positioning sequence inserted in the *S.cerevisiae* genome. By assembling DNA tandem repeats of these synthetic sequence in yeast, we validated that they were able to preferentially position a nucleosome on their first 147 bp and lead to the formation of arrays of nucleosomes positioned statistically following the computational prediction. The sequence rules that we extracted showed that this positioning is mainly induced by the creation of nucleosome repelling motifs within the linker region and the destruction of these motifs in the dyad region,compatible with established nucleosome positionning rule is S. cerevisiae (5). The resulting nucleosomal arrays, with altered repeat length, are thus expected to behave differently than regular arrays of 163 bp NRL found on gene bodies, which are generated by the spacing activities of chromatin remodelers (29, 30). Interesting questions that can be asked with our constructs include the positioning of nucleosomes on the array in the absence of remodeling factors and the potential effect of transcription across these arrays on NRL length. Our method, which combines a deep learning predictive method and the k-MCMC methodology, can be adapted to design DNA sequences with other designer characteristics, like transcription activity, and we expect it to play a growing role in the emerging field of synthetic genomics. Amongst the existing limitations, it is known that MNase digestion has sequences biases (31) that are learned by our model and therefore can affect the prediction output. A more precise training model could be achieved with alternative nucleosome positioning sequencing techniques (36). Also, the network is trained on a wild type strain in a specific growth medium (YPD) and cannot be used to predict the outcome in a different strain or growth condition.

ACKNOWLEDGEMENTS

We thank members of the Structure and Instabilités des génomes and LPTMC laboratories for constructive comments during the realization of this work. Work was supported by core funding from CNRS, INSERM and MNHN. ALJ and ER were supported by a doctoral fellowship from the french ministry for Education, Research and Technology. EP was funded by grant ANR-15-CE11-0023-03 (HiResBac).

AVAILABILITY OF DATA AND MATERIALS

Synthetic strains developed in this work are available upon request to JBB. Raw sequencing reads for MNase-Seq experiments are available at the following address: https://dataview.ncbi.nlm.nih.gov/object/PRJNA863754

The code developed in this work is available on github at the following address:

https://github.com/Alexwestbrook/nuc_sequence_design_clean

Conflict of interest statement. None declared.

REFERENCES

1. Ostrov N, Beal J, Ellis T, Gordon DB, Karas BJ, Lee HH, Lenaghan SC, Schloss JA, Stracquadanio G, Trefzer A, Bader JS, Church GM, Coelho CM, Efcavitch JW, Güell M, Mitchell LA, Nielsen AAK, Peck B, Smith AC, Stewart CN Jr, Tekotte H. Technological challenges and milestones for writing genomes. Science. 2019 Oct 18;366(6463):310-312
2. Hossain A, Lopez E, Halper SM, Cetnar DP, Reis AC, Strickland D, Klavins E, Salis HM. Automated design of thousands of nonrepetitive parts for engineering stable genetic systems. Nat Biotechnol. 2020 Dec;38(12):1466-1475
3. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat Biotechnol. 2020 Jan;38(1):56-65
4. de Almeida BP, Reiter F, Pagani M, Stark A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat Genet. 2022 May;54(5):613-6244
5. Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol. 2010 Jul 6;8(7):e1000414
6. Hughes AL, Jin Y, Rando OJ, Struhl K. A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. Mol Cell. 2012 Oct 12;48(1):5-15
7. Lowary P, Widom J (1998) New dna sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. Journal of molecular biology 276(1):19–42
8. Robinson PJ, Fairall L, Huynh VA, Rhodes D. EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. Proc Natl Acad Sci U S A. 2006 Apr 25;103(17):6506-11
9. Lancrey A, Joubert A, Duvernois-Berthet E, Routhier E, Raj S, Thierry A, Sigarteu M, Ponger L, Croquette V, Mozziconacci J, Boulé JB. Nucleosome Positioning on Large Tandem DNA Repeats of the '601' Sequence Engineered in Saccharomyces cerevisiae. J Mol Biol. 2022 Apr 15;434(7):167497
10. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019 Jan;51(1):12-18
11. Routhier E, Mozziconacci J (2022) Genomics enters the deep learning era. PeerJ 10:e13,613
12. Bogard N, Linder J, Rosenberg AB, Seelig G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. Cell. 2019 Jun 27;178(1):91-106.e23
13. Linder J, Bogard N, Rosenberg AB, Seelig G. A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences. Cell Syst. 2020 Jul 22;11(1):49-62.e16
14. Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, Fields S, Seelig G. Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. Genome Res. 2017 Dec;27(12):2015-2024
15. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, Seelig G. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. Nat Biotechnol. 2019 Jul;37(7):803-809
16. Routhier E, Pierre E, Khodabandelou G, Mozziconacci J. Genome-wide prediction of DNA mutation effect on nucleosome positions for yeast synthetic genomics. Genome Res. 2020 Dec 18;31(2):317–26
17. González S, García A, Vázquez E, Serrano R, Sánchez M, Quintales L, Antequera F. Nucleosomal signatures impose nucleosome positioning in coding and noncoding sequences in the genome. Genome Res. 2016 Nov;26(11):1532–1543
18. Zalenskaya IA, Pospelov VA, Zalensky AO, Vorob'ev VI. Nucleosomal structure of sea urchin and starfish sperm chromatin. Histone H2B is possibly involved in determining the length of linker DNA. Nucleic Acids Res. 1981 Feb 11;9(3):473-87
19. Bailey TL (2021) Streme: accurate and versatile sequence motif discovery. Bioinformatics 37(18):2834–2840
20. Bortz A, Kalos M, Lebowitz J (1975) A new algorithm for monte carlo simulation of ising spin systems. Journal of Computational Physics 17(1):10–18
21. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W202-8
22. Lancrey A, Joubert A, Boulé JB. Locus specific engineering of tandem DNA repeats in the genome of Saccharomyces cerevisiae using CRISPR/Cas9 and overlapping oligonucleotides. Sci Rep. 2018 May 8;8(1):7127
23. Hill J, Donald KA, Griffiths DE. DMSO-enhanced whole cell yeast transformation. Nucleic Acids Res. 1991 Oct 25;19(20):5791
24. Church GM, Gilbert W (1984) Genomic sequencing. Proceedings of the National Academy of Sciences of the United States of America 81(7 I):1991–1995
25. Green MR, Sambrook J (2021) Total RNA Extraction from Saccharomyces cerevisiae Using Hot Acid Phenol. Cold Spring Harb Protoc. 12:523-525
26. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25 10(3):R25
27. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nature methods 9(4):357
28. Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research 33(suppl2):W465-W467
29. Ocampo J, Chereji RV, Eriksson PR, Clark DJ. The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. Nucleic Acids Res. 2016 Jun 2;44(10):4625–35
30. Gkikopoulos T, Schofield P, Singh V, Pinskaya M, Mellor J, Smolle M, Workman JL, Barton GJ, Owen-Hughes T. A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. Science. 2011 Sep 23;333(6050):1758–60
31. Chereji RV, Ocampo J, Clark DJ (2017) MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers. Molecular Cell. 65(3):565-577.e3
32. Jensen TH, Jacquier A, Libri D (2013) Dealing with pervasive transcription. Molecular Cell 52:473-84.
33. Wu, Q, Bazzini, A (2023) Translation and mRNA Stability Control. Annual review of biochemistry, 92: 227-245.
34. Tisseur M, Kwapisz M, Morillon A (2011) Pervasive transcription – Lessons from yeast. Biochimie, 93: 1889-1896.
35. Teixeira MC, Viana R, Palma M, Oliveira J, Galocha M, Mota MN, Couceiro D, Pereira MG, Antunes M, Costa IV, Pais P, Parada C, Chaouiya C, SáCorreira I, Monteiro PT (2023) YEASTRACT+: a portal for the exploitation of global transcription regulation and metabolic model data in yeast biotechnology and pathogenesis. Nucleic Acids Research, advance access (doi:10.1093/nar/gkac1041)
36. Lieleg C, Krietenstein N, Walker M, Korber P (2015) Nucleosome positioning in yeasts: methods, maps, and mechanisms. Chromosoma 124:131-51. doi: 10.1007/s00412-014-0501-x. Epub 2014 Dec 23. PMID: 25529773.