The sequence determinants of nucleosome positioning in mammals

Julien Mozziconacci , StrInG Lab, MNHN, INC
Pablo Navarro, EPIC Lab, Institut Pasteur, INSB

Nucleosomes are the elementary units of DNA folding into functional chromosomes in eukaryotes. These nucleo-proteic complexes are regularly spaced along genomes every ~200 bp and cover 90 % of the genome (Figure 1a). The histone proteins found in nucleosomes are among the most conserved architectural proteins in eukaryotes and can harbor conserved bio-chemical modifications that are the basis of cell differentiation during development. A remarkable fact about nucleosomes is that their position along the genome is very similar in a cell population. Mapping these positions using enzymatic digestion of the DNA found in between nucleosomes leads to regular density peaks corresponding to precise nucleosome positions (Figure 1b). **In all organisms, nucleosome positions reflect the position of genes and regulatory elements i.e. DNA sequences that are implicated in gene activity regulation. The key question is how are these positions established along the DNA molecule ?** In yeast, the question has been largely addressed and at least partially solved: the small intergenic regions contain poly(A) as well as poly(CG) motifs which are known to repel nucleosomes. The nucleosomes are thus found on gene bodies where an ensemble of nucleosome associated proteins organize them in regular and compact arrays. This organization has been deciphered over the years by the works of many groups. Quantitative modeling approaches have been developed to simulate the 1D nucleosomal density over the DNA sequence knowing the interaction energy term between nucleosomes and the energy landscape given by the sequence. In vitro, this energy landscape can be well approximated using the sequence dependent bending properties of DNA. In vivo however, the energy landscape depends on all the other DNA binding proteins present in the cell and can change in different conditions. While it is not possible to infer it from first principles, we have recently shown that it can be retrieved using deep learning models. Using a deep convolutional neural network, typically used in the context of image classification, we inferred the in vivo nucleosome density along the DNA sequence (Figure 1c). We re-discovered the well known sequence determinants of nucleosome positioning in yeast and assessed the effect on the nucleosome density of individually mutating every nucleotide in the yeast genome[1] .

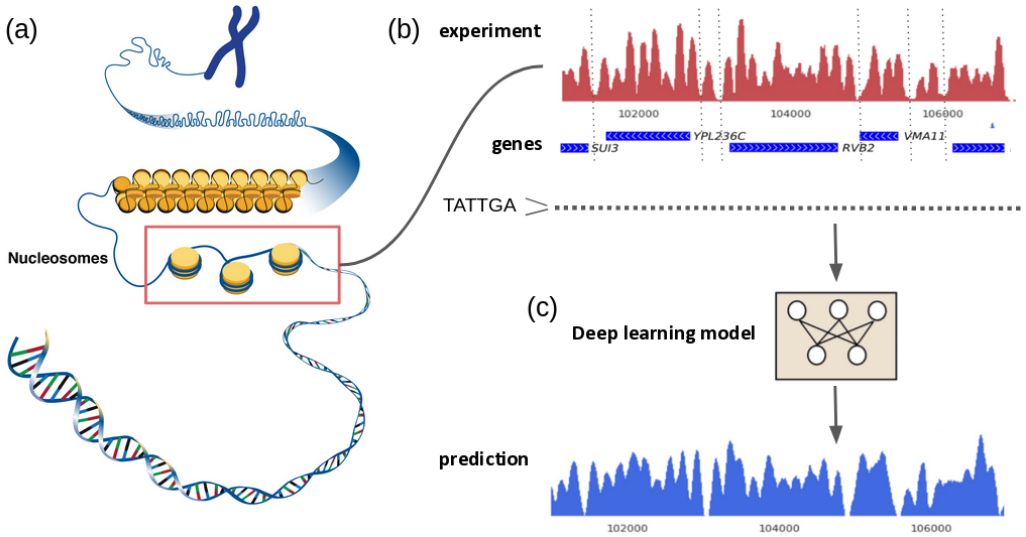

Figure 1: (a) The hierarchical structure of a chromosome (b) Experimentally determined nucleosome density along a portion of the yeast chromosome 16. Genes are indicated in blue. Intergenic regions (dashed lines) are depleted in nucleosomes. (c) Prediction of the nucleosome density obtained from the sequence alone using our deep learning model that has been trained on the other yeast chromosomes.

This simple picture does not hold for complex multicellular organisms such as mammals for which the genome is 100 times larger. While the number of genes of multicellular organisms is similar than in unicellular organisms, the intergenic sequences in these genomes are much longer, offering a potential for the differential regulation of genes and the apparition of different cell types. In mammals, the vast majority of these sequences (>70% [2]) is composed of repetitive DNA motifs of 10-3000 bp. The result of this expansion of repetitive sequences is that the distance between regulatory sites, which locally position nucleosomes, can exceed 100 kbp. However, it appears that nucleosomes are still well positioned within these large domains. Even more intriguing, the spacing between nucleosomes in mammals can vary along the genome, being longer for regions in which genes are repressed [3]. How this differential nucleosome spacing is achieved remains a mystery. Two recents results from our groups lead us to propose the present project and potentially solve this mystery. The first one is that we were able to train a very simple deep learning model on only 1 (out of 21) mouse chromosome and that this model is able to predict the nucleosome positions for all chromosomes with a correlation of 0.84, which is much higher than the best results it can attain in yeast (0.65). The second one is an aggregate map of nucleosome signals over the mouse DNA repeats showing a strong consistent signal over different instances of the same repeat. These two results point to a role of these elements in nucleosome positioning, as has been proposed by Alain Arneodo for Alu elements in humans and envisioned for other metazoans[4]. Debates regarding potential cellular functions of these elements have been long standing. Controversial references to 'junk' or 'selfish' DNA were originally put forward, implying that repetitive DNA segments are remainders from past evolution or autonomous self-replicating sequences hacking the cell machinery to proliferate, respectively [5,6]. With the advent of genomics studies, the influence of repetitive elements on DNA-related metabolic processes or genome evolution has been increasingly investigated. Comparative genomics studies have reinforced the idea that transposable elements have, or had, the ability to reshape gene regulatory networks of vertebrate's genomes over evolutionary times [7]. The potential new role of DNA repeats in nucleosome positioning along the genome would have important implications in the field of gene regulation in development and disease and in genome evolution.

Specific aims and methods: The PhD project will focus on nucleosome organization in mouse embryonic stem (ES) cells, which we routinely culture, are easily synchronized in mitosis, can differentiate into specific cell types and can be genetically modified. The PhD student will establish the DNA-sequence determinants of nucleosome positioning combining experimental approaches and computational methods in three different biological contexts: genetically polymorphic or transgenic ES cells, pure mitotic populations and differentiated cells populations.

Aim1 _ Determination of nucleosome organization principles. The PhD student will produce ultra-deep mono-nucleosome MNase-seq datasets. This data will be used as a training dataset for deep learning models and propose a first set of positioning rules. The different predictions will then be confronted to identify regions in which nucleosomes are found at the same position in all cells and regions exhibiting polymorphic positions. We next plan to use these extremely accurate maps of nucleosomes to analyze known active/inactive genetic elements in ES cells. Among different hypotheses, he will assess the roles of repetitive elements, insulators and regulatory elements as punctuation marks for nucleosome positioning.

Aim2_ Fine predictions of nucleosome organization. He will use the models trained in Aim1 to perform an in silico mutagenesis approach and assess the impact of mutations on nucleosome positioning. The approach will be validated using polymorphic ES cells (129Sv vs C57BL/6, which present ~1 single nucleotide polymorphism every 100 bp). If we observe major differences between 129Sv predictions and C57BL/6 data, he will use 129Sv Bacterial Artificial Chromosomes or Fosmids to integrate these large sequences in a C57BL/6 background and assess how the nucleosomes are organized across this foreigner DNA.

Aim3_ Discovery of regulatory activities altering nucleosome organization. We propose to use the high resolution nucleosome position maps obtained in aims 1 and 2 in order to discover de novo regions of particular interest, such as enhancers or insulators, which are supposed to be devoid of nucleosomes. The recruited PhD student will next produce standard MNase-seq datasets (400.10^6 reads) in mitotic ES cells and in cells differentiated into neuronal precursors (a cell type where many epigenomic datasets are available) and improve these maps to high resolution maps using transfer learning [8] . By exploiting the trained networks he will study the different sets of sequence rules that apply in different cell types and identify regions for which the predictions in different cell types change and use de novo motif discovery algorithms to find TF involved in gene regulatory inheritance in mitosis and differentiation. A subsequent validation by ChIP-seq will represent a major tour-de-force to transform the study of nucleosome positioning in mammals into a generator of data-driven hypotheses opening up new avenues to foster research in gene regulation.

Julien MOZZICONACCI (JM) at LPTMC, Sorbonne University . JM is a physicist at LPTMC and newly appointed professor at the Muséum National d'Histoire Naturelle and Institut Universitaire de France. He is an internationally recognized expert in integrative structural modeling in the field of chromatin and chromosome structure and dynamics. JM has a strong expertise in statistical analysis of biological data and is one of the pioneers of Hi-C data analysis [9]. JM notably showed using Hi-C data the pervasive correlation between genome fold and DNA repeated sequence positions in human, mouse and drosophila [10]. He recently developed within the team a group working on application of deep learning to DNA sequences [1,11,12]. He has supervised or co-supervised 7 PhD students and is a regular reviewer for high rank interdisciplinary journals (Cell, Nature Methods,…) and international funding agencies (ERC, "Veni Vedi Vici",...)

Pablo NAVARRO-GIL (PN) at Institut Pasteur, is a Director of Research at the Department of Developmental and Stem Cell Biology, where he leads a team of 14 scientists working on gene regulation in mouse ES cells, early mouse embryos and human cancer cell lines. A major focus over the last few years has been to decipher how TFs control nucleosome positions and how these regulations are altered or preserved during mitosis. PN pioneered the field of mitotic TF binding in ES cells and proposed the first mechanistic model of nucleosome-positioning-based inheritance during mitosis and replication [13,13–15]. Several members of the team, including permanent scientists and engineers, will be involved in this project, together with a team of bioinformaticians.

(1) **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. https://doi.org/10.1101/gr.264416.120.**

(2) de Koning, A. P. J.; Gu, W.; Castoe, T. A.; Batzer, M. A.; Pollock, D. D. Repetitive Elements May Comprise over Two-Thirds of the Human Genome. *PLoS Genet.* **2011**, *7* (12), e1002384. https://doi.org/10.1371/journal.pgen.1002384.

(3) Valouev, A.; Johnson, S. M.; Boyd, S. D.; Smith, C. L.; Fire, A. Z.; Sidow, A. Determinants of Nucleosome Organization in Primary Human Cells. *Nature* **2011**, *474* (7352), 516–520. https://doi.org/10.1038/nature10002.

(4) Brunet, F. G.; Audit, B.; Drillon, G.; Argoul, F.; Volff, J.-N.; Arneodo, A. Evidence for DNA Sequence Encoding of an Accessible Nucleosomal Array across Vertebrates. *Biophys. J.* **2018**, *114* (10), 2308–2316. https://doi.org/10.1016/j.bpj.2018.02.025.

(5) McClintock, B. The Significance of Responses of the Genome to Challenge. *Science* **1984**, *226* (4676), 792–801. https://doi.org/10.1126/science.15739260.

(6) Orgel, L. E.; Crick, F. H. Selfish DNA: The Ultimate Parasite. *Nature* **1980**, *284* (5757), 604–607. https://doi.org/10.1038/284604a0.

(7) Venuto, D.; Bourque, G. Identifying Co-Opted Transposable Elements Using Comparative Epigenomics. *Dev. Growth Differ.* **2018**, *60* (1), 53–62. https://doi.org/10.1111/dgd.12423.

(8) Novakovsky, G.; Saraswat, M.; Fornes, O.; Mostafavi, S.; Wasserman, W. W. Biologically Relevant Transfer Learning Improves Transcription Factor Binding Prediction. *Genome Biol.* **2021**, *22* (1), 280. https://doi.org/10.1186/s13059-021-02499-5.

(9) **Botta, M.; Haider, S.; Leung, I. X. Y.; Lio, P.; Mozziconacci, J. Intra- and Inter-Chromosomal Interactions**

Correlate with CTCF Binding Genome Wide. *Mol. Syst. Biol.* 2010, *6* (1), 426. https://doi.org/10.1038/msb.2010.79.

(10) **Cournac, A.; Koszul, R.; Mozziconacci, J. The 3D Folding of Metazoan Genomes Correlates with the Association of Similar Repetitive Elements. *Nucleic Acids Res.* 2016, *44* (1), 245–255. https://doi.org/10.1093/nar/gkv1292**.

(11) **Khodabandelou, G.; Routhier, E.; Mozziconacci, J. Genome Annotation across Species Using Deep Convolutional Neural Networks. *PeerJ Comput. Sci.* 2020, *6*, e278. https://doi.org/10.7717/peerj-cs.278.**

(12) **Routhier, E.; Bin Kamruddin, A.; Mozziconacci, J. Keras_dna: A Wrapper for Fast Implementation of Deep Learning Models in Genomics. *Bioinforma. Oxf. Engl.* 2021, *37* (11), 1593–1594. https://doi.org/10.1093/bioinformatics/btaa929.**

(13) **Festuccia, N.; Owens, N.; Papadopoulou, T.; Gonzalez, I.; Tachtsidi, A.; Vandoermel-Pournin, S.; Gallego, E.; Gutierrez, N.; Dubois, A.; Cohen-Tannoudji, M.; Navarro, P. Transcription Factor Activity and Nucleosome Organization in Mitosis. *Genome Res.* 2019, *29* (2), 250–260. https://doi.org/10.1101/gr.243048.118.**

(14) **Owens, N.; Papadopoulou, T.; Festuccia, N.; Tachtsidi, A.; Gonzalez, I.; Dubois, A.; Vandormael-Pournin, S.; Nora, E. P.; Bruneau, B. G.; Cohen-Tannoudji, M.; Navarro, P. CTCF Confers Local Nucleosome Resiliency after DNA Replication and during Mitosis. *eLife* 2019, *8*, e47898. https://doi.org/10.7554/eLife.47898.**

(15) **Festuccia, N.; Gonzalez, I.; Owens, N.; Navarro, P. Mitotic Bookmarking in Development and Stem Cells. *Dev. Camb. Engl.* 2017, *144* (20), 3633–3645. https://doi.org/10.1242/dev.146522**.