Machine-Learning Models to Recognize Patterns of Nucleosome and DNA Structures Positioning

Elen Tevanyan[1] and Maria Poptsova[1]

[1] National Research University Higher School of Economics, Moscow, Russia
mpoptsova@hse.ru

Abstract. Non-B DNA structures have a great potential to form and influence various genomic processes including transcription. One of the mechanisms of transcription regulation is nucleosome positioning. Even though only B-DNA can be wrapped around a nucleosome, non-B DNA structures can compete with a nucleosome for a genomic location. Here we used permanganate/S1 nuclease footprinting data on non-B DNA structures, such as Z-DNA, H-DNA, G-quadruplexes and stress-induced duplex destabilization (SIDD) sites, together with MNase-seq data on nucleosome positioning in the mouse genome. We found three types of patterns of nucleosome positioning around non-B DNA structures: a structure is surrounded by nucleosomes from both sides, from one side, or nucleosome free region. Machine learning models based on random forest and XGBoost algorithms were constructed to recognize DNA regions of 1kB length containing a particular pattern of nucleosome positioning for four types of DNA structures (Z-DNA, H-DNA, G-quadruplexes and SIDD sites) based on statistics of di- and trinucleotides. The best performance (94% of accuracy) was reached for G-quadruplexes while for other types of structures the accuracy was under 70%. We conclude that 1kB regions containing G-quadruplexes have distinct compositional properties, and this fact points to preferential locations of such pattern in the genome and requires further investigation. For other DNA structures a region composition is not a sufficient predictive factor and one should take into account other physical and structural DNA properties to improve nucleosome-DNA-structure pattern recognition.

Keywords: DNA structures, non-B DNA, G-quadruplexes, H-DNA, Z-DNA, triplex DNA, nucleosome positioning, machine-learning methods, random forest, gradient boosting, pattern recognition.

Currently there exist very few high throughput sequencing technologies revealing different types of DNA secondary structures at the genome-wide scale. These are Permanganate/S1 Nuclease Footprinting [1], G4-Seq [2], G4 Chip-Seq [3]. The first method estimates the genome potential to form various DNA structures, while the last two are designed specifically to recognize quadruplexes. Experimental methods to detect other structures such as triplexes, Z-DNA, cruciforms at a genome-wide scale are still limited. That is why the development of computer methods to recognize patterns of association between DNA structures and nucleosomes is of practical importance.

Here we used data from Permanganate/S1 Nuclease Footprinting of the mouse genome that revealed potential sites of unwound DNA where DNA structures were formed. For nucleosome maps we took MNase-Seq data for mouse genome from [1]. Computer annotation of DNA secondary structures were done with the following programs: Z-DNA – Zhunt [4], H-DNA - Inverted Repeats Finder [5], G-quadruplexes – QuadParser [6], SIDD – algorithm taken from [7]. The potential to

form DNA structures in mouse genome as it is inferred with computer methods, and the number of structure enriched in single stranded DNA (ssDNA) revealed in [1], are given in Table 1.

Table 1. Potential to form DNA structures in the mouse genome.

	Z-DNA	H-DNA	G-quadruplexes	SSID
Nu of computer predicted structures	249 752	320 585	263 167	416 444
Nu of predicted structures enriched in ssDNA	25 062 (10%)	17 109 (5.3%)	20 259 (7.7%)	15 298 (3.8%)

To analyse an association of nucleosomes and DNA structures we considered a region of 1kB centred on DNA structures. Nucleosome profiles around DNA structures revealed three types of patterns: a structure is surrounded by nucleosomes from both sides, from one side, or the region around a structure is nucleosome free (see Fig. 1). We did not find any pattern when a nucleosome position overlaps with a DNA structure, which is an expected result and in accordance with the current understanding of preferences for nucleosome positioning.

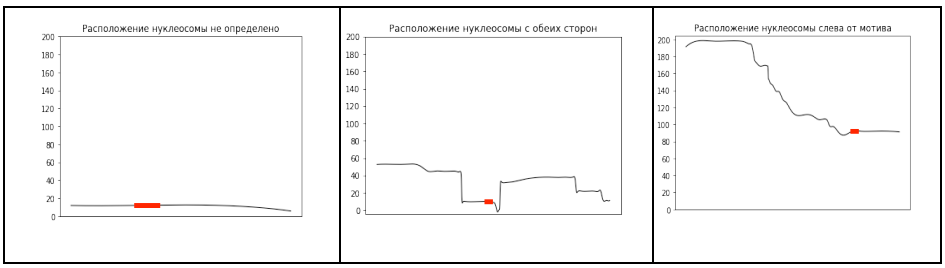

Fig. 1. Patterns of nucleosome positioning around DNA structures.

Next we built and train machine learning models to recognize regions containing a particular pattern based on di- and trinucleotide composition of 1 kB region containing the pattern. We tried two machine-learning methods – Random forest and XGBoost. Accuracy, precision and recall metrics are given in Table 2.

Low values or zeros for precision and recall for the pattern of two nucleosomes surrounding a DNA structure signify that the pattern is almost unrecognizable. Despite the low accuracy for predicting patterns with Z-DNA and H-DNA both Random forest and XGBoost are quite precise in predicting non-deterministic location in the considered region. G-quadruplexes are recognized best of all with both methods.

Table 2. Accuracy, precision and recall for all types of pattern recognition for all DNA structures.

Model\DNA structure	Z-DNA			H-DNA			G-quadruplexes			SSID		
Accuracy												
Random Forest	0.6			0.6			0.93			0.62		
XGBoost	0.61			0.66			0.93			0.68		
Class	1	2	3	1	2	3	1	2	3	1	2	3
Precision												
Random Forest	0.69	0.66	0.0	0.75	0.6	0.0	0.4	0.94	0.0	0.7	0.66	1.0
XGBoost	0.54	0.62	0.28	0.72	0.57	0.13	0.4	0.94	0.0	0.7	0.67	0.29
Recall												
Random Forest	0.72	0.55	0.0	0.78	0.49	0.0	0.01	0.99	0.0	0.72	0.56	0.0
XGBoost	0.18	0.93	0.0	0.72	0.61	0.04	0.02	0.99	0.0	0.76	0.64	0.02

The model can be improved by taking into account physical and chemical properties of DNA such as enthalpy, entropy, Gibbs energy, hydrophilicity, as well as helical structural properties of dinucleotides (Shift, Rall, Slide, Rise, Tild, Bend) available in DiProDB [8]. Even though it might not improve the model it could provide an understanding, which properties mainly contribute to the classification model. In future, analysis of other types of tissues could reveal tissue-specific regulation patterns.

References

1. Kouzine, F., Wojtowicz, D., Baranello, L., Yamane, A., Nelson, S., Resch, W., Kieffer-Kwon, K.R., Benham, C.J., Casellas, R., Przytycka, T.M., and Levens, D.: Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. Cell Syst 4(3), 344-356 e347 (2017).
2. Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P., and Balasubramanian, S.: High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nat Biotechnol 33(8), 877-881 (2015).
3. Hansel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D., and Balasubramanian, S.: Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. Nat Protoc 13 (3), 551-564 (2018).
4. Ho, P.S., Ellison, M.J., Quigley, G.J., and Rich, A.: A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. EMBO J 5 (10), 2737-2744 (1986).
5. Warburton, P.E., Giordano, J., Cheung, F., Gelfand, Y., and Benson, G.: Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res 14 (10A), 1861-1869 (2004).
6. Huppert, J.L., and Balasubramanian, S.: Prevalence of quadruplexes in the human genome. Nucleic Acids Res 33 (9), 2908-2916 (2005).

7. Wang, H., Noordewier, M., and Benham, C.J.: Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters. Genome Res 14 (8), 1575-1584 (2004).
8. Friedel, M., Nikolajewa, S., Suhnel, J., and Wilhelm, T.: DiProDB: a database for dinucleotide properties. Nucleic Acids Res 37 (Database issue), D37-40 (2009).