Machine Learning Applications for Genomic Pattern Recognition Problem*

Elen Tevanyan and Maria Poptsova

National Research University Higher School of Economics,
Myasnitskya str. 20, 101000, Moscow, Russia
etevanian@hse.ru; mpoptsova@hse.ru

Abstract. DNA secondary structures are important functional elements that may influence cellular processes. One of their possible functions is regulation of nucleosome positioning. Here MNAse-seq and ssDNA-seq data were used to define patterns of positional relationship of DNA structures such as Z-DNA, H-DNA and G-quadruplexes with nucleosomes. Three types of patterns were found: a structure is surrounded by nucleosomes from both sides, from one side, or nucleosome free region. Machine-learning models based on Random forest algorithm and XGBoost were trained to recognize DNA region of 500 bp length containing a pattern of nucleosome positioning for three types of DNA structures (Z-DNA, H-DNA and G-quadruplexes) based on DNA sequence compositional properties. The best performance (more than 86% for ROC-AUC, accuracy, recall and presicion scores) was reached for G-quadruplexes. 500 bp regions containing G-quadruplexes have distinct compositional properties and point to the preferential locations of the defined patterns, which regulatory functions require further investigation. For other DNA structures a region composition is less powerful predictive factor and one should take into account other physical and structural DNA properties to improve nucleosome-DNA-structure pattern recognition.

Keywords: DNA structures, nucleosome positioning, machine-learning methods, random forest, xgboost

1 Introduction

Machine learning is widely applied to problems in genomic research[1]. Computational methods successfully annotate genomes with functional elements, such as transcription start sites [1], splice-sites [2], alternative splicing [3], promoters, enhancers [4]. Recent advancements in computational performance enable to predict nucleosome positioning using models based on neural networks [5]. However, it remains a challenging task to detect non-B-DNA structures and determine their function with machine learning algorithms due to the absence of experimentally confirmed genome-

wide data on many types of structures. As a result, patterns of DNA structures and their positioning with respect to other elements are hard to detect as well. Despite the limitation, non-B-DNA structures might influence chromatin reorganization by governing nucleosome positioning, thus, regulating transcription that makes pattern recognition of DNA structures and nucleosomes positioning an important task.

A model of right-handed double helix DNA molecule known as B-DNA was first proposed in [6]. Although B-DNA conformation is widely spread and is considered as canonical, more than ten types of DNA secondary structures are discovered: A-DNA, Z-DNA, H-DHA, V-DNA, stem-loops, G-quadruplexes, i-motif, buldge-DNA, etc [7]. Due to the scarcity of experimental data, the research is focused on three types of DNA structures: Z-DNA, H-DNA, and G-quadruplexes.

To begin with, left-handed double helix called Z-DNA is among the most studied DNA conformation and found both *in vitro* [8] and *in vivo* [9]. Z-DNA has the potential to be involved in transcription. It was shown [9] that three regions near the promoter of gene C-MYC have adopted Z-DNA conformation while the gene was actively transcribed. As for the in silico detecton, Z-Hunt algorithm is mainly used to describe genomic region's potential to form a left-handed helix.

The next structure of interest is H-DNA. Triple helix consists of a usual double helix, which is connected to separate single DNA strand either from its part or from another molecule [10]. The existence *in vitro* [11] was confirmed a short while after Watson and Crick discovery while *in vivo* proofs were found later [12]. Researches point out H-DNA involvement in replication, transcription reparation and homologous recombination [13]. For example, the study implies H-DNA acts as a barrier for replication [14]. The detection of H-DNA motives is based on the primary sequence content: algorithms search for inverted repeats.

As for G-quadruplexes, they also exist both *in vitro* [15] and *in vivo* [16]. *In silico* detection uses particular motif composition to classify a region as G-quadruplex adopting [17]. The biological function of G-quadruplexes is actively investigated. Recent studies highlight G-quadruplexes regulate transcription and replication [18]. In contrast to the absence of genome-wide data on other secondary structures, a technique called G4-seq is developed which maps genomic regions in G-quadruplex conformation.

DNA molecules are compactly packed in the cells and are organized into chromatin. Double-stranded DNA wraps around histone proteins forming nucleosomes [19]. When a nucleosome is formed, the underlying DNA region is inactive, it cannot be transcribed as transcription factors cannot bind DNA. Nucleosomes positioning is influenced by many factors: DNA sequence itself, histone modifications, remodeling complexes, transcription, replication.

As stated above, secondary DNA structures may influence transcription in cells by regulating nucleosome positioning. It was experimentally confirmed the only B-DNA can wrap around nucleosome which makes impossible for non-B-DNAs to bind histones [20]. There is an evidence that Z-DNA and H-DNA govern nucleosome positioning acting as a barrier [21] while G-quadruplexes are formed in nucleosome-free regions [22].

Machine learning methods are applied to determine nucleosomal profiles. The availability of genome-wide data on nucleosome positioning has led to the development of different models. The first papers in this field use simple techniques of statistical analysis to calculate the probability of nucleosome formation [23] with performance quality of 50%. Later studies describe nucleosome positioning as machine learning classification task and apply SVM and Random Forest algorithms, achieving the prediction power more than 80%. Recent studies focus on convolutional neural networks [5], which achieve accuracy, precision, and recall of more than 90%.

Classical machine learning approach supposes the sample to be described with features. The DNA sequence consists of only 4 units known as bases: A, C, G, T. Due to the specific nature of a sample in genomic studies, a region of DNA sequence is considered as a string and methods of feature extraction similar to a text analysis are applied. One of the simplest and effective strategies is to examine k-tuple nucleotide composition where k usually varies from 1 to 6. Another powerful approach is to describe DNA sequence with physical and chemical properties of base pairs which are presented in the publicly available databases. K-neighbors characteristics are used as features as well.

The literature analysis reveals several facts which are important for this paper. First, high-throughput techniques are developed only for G-quadruplexes, so data is available only for this type of structure, for other types of DNA structures computational methods are needed. Second, non-B-DNA structures are involved in the main cell processes like transcription and replication. What is more important, they prevent forming nucleosomes. Third, machine learning methods are used to define nucleosomal profiles.

To our knowledge, machine learning algorithms are trained to detect either nucleosomes or secondary DNA structures. This paper aims to recognize patterns of DNA structures and nucleosomes positioning. The results may lead to better understanding of chromatin remodeling mechanisms and how transcription is regulated by non-B-DNA.

2 Methods

2.1 Genome Computational Annotation

To analyze patterns of DNA secondary structures and nucleosome positioning the data on mouse genome is used. The mm9 version of genome is available at UCSC Genome Browser [24].The genome is annotated with three types of structures: Z-DNA, H-DNA and Q-quadruplexes. The table below represents the software used for each type of structures.

Table 1.

Structure	Software
Z-DNA	Z-Hunt [25]

H-DNA	Inverted Repeats Finder [26]
Q-quadruplexes	QuadParser [17]

2.2 Genome In Vivo Annotation

In vivo detection of secondary structures is a challenging task. However, the study presents [27] the design of ss-DNA-seq experiment on mouse B-cells to obtain genome-wide locations of DNA secondary structures. The data is available at Laboratory's Research Page [28]. The reads are aligned to mouse genome with Bowtie software, version 0.12.7 [29].

Then both computational annotations and *in vivo* detected structures are intersected to define non-putative motives of DNA secondary structures. The intersection is done with Bedtools [30] software of version 2.27.1.

2.3 Nucleosome Data

The nucleosome positioning profile is the result of MNAse-seq on mouse B-cells data analysis provided by the study [27]. All the details are described in the paper's methods while the data is available at NCBI under the SRA identifier SRA072844. The data is preprocessed according to Illumina Analysis pipeline. The reads are aligned to the mouse genome with Bowtie software, version 0.12.7.

After the alignment each read is lengthened up to 146 base pairs in 3' direction and is considered as a nucleosome forming region in the particular cell line.

2.4 Patterns of DNA structures and nucleosomes

The region of interest is a sequence of 500 bp length centered on the secondary structure. For that region the coverage with MNAse preprocessed data is calculated to discover the coverage density. The average coverage of the genome is computed based on randomly selected 200 000 regions. Any region of interest, which is covered by more than the average coverage, is further inspected for the type of pattern. The average coverage is compared with t-test. Regions, which fail the test, are considered as nucleosome-free (pattern 0).

The region of interest is split into three parts: the center (DNA secondary structure), the right side (250 bp), the left side (250 pb). The maximum coverages within each part are compared with each other. If all of them have close values, then the region is classified as nucleosome-free (pattern 0). If the peaks on both right and left sides are higher than that in the center, then the structure is surrounded by two nucleosomes (pattern 1). Following the same procedure, the pattern with a nucleosome on one side is defined (pattern 2). For the simplicity reason pattern 1 and pattern 2 are merged into one category.

2.5 Machine Learning Task

Let **x** be the sample representing a region of interest – a sequence of 500 bp length centered on a secondary structure. Let **y** be the pattern which the region is associated with and let y be considered as the class of the sample. The aim is to train a classifier which can predict the pattern of any particular region. In other words, the model defines the type of pattern of the region.

2.6 Feature extraction

It is a common problem to express a genomic region via feature vectors which can be handled by classical machine learning algorithms. As for the task in this paper, the sample represents a string of length of around 530 letters as the genome sequence consists of 4 elements: A, C, G, T. K-tuple nucleotide composition with k equal to 2 and 3 is used as the feature extraction strategy. In other words, each sequence is described with 80 features: 16 for the quantity of a particular dinucleotide and 64 for the triplets. One feature as GC-content is added to the dataset.

2.7 Machine Learning Algorithms

Two algorithms are used for the classification task: Random Forest [31] and XGBoost [32]. For both algorithms the following is true: the dataset is split into the training set and the test set in the proportion of 70-30% . The training set is used to validate algorithms with the 5-fold cross validation strategy. Different parameters are tested with randomized search strategy.

Random Forest Classifier

Algorithm available in the scikit-learn library [33] of version 0.20.01 was used in this study for the classification task. To find the best model, the number of trees is varied from 10 to 100.

XGBoost

Open-source library [32] for Python is used with parameters varied:

- λ from 0 to 1 with the step 0.1
- θ from 0 to 1 with the step 0.1
- η from 0 to 0.5 with the step 0.25

2.8 Model evaluation

A set of quality measures are used to evaluate models:

- Accuracy
- Recall

- Precision
- ROC-AUC

During the algorithms' optimization ROC-AUC score was used as the scoring function.

3 Results and Discussion

The investigation of the role of DNA secondary structures on nucleosome positioning and pattern search requires data on structures and nucleosome maps. The computational annotations of the mouse genome with DNA secondary structures were combined with ssDNA-seq data on B-cells. As it can be seen from table, it results in many putative motives of non-B-DNA (Table 2). The results are not surprising because the formation of alternative structure requires a set of conditions, wereas some genomic regions can be tightly packed.

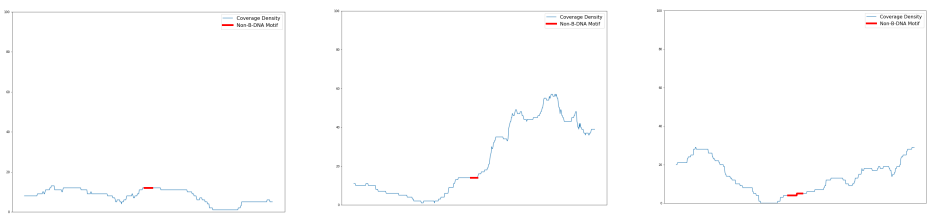
Table 2. Results of computer annotations of the mouse genome with DNA secondary structures and enrichment with ssDNA-seq data.

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

Then the motives enriched with ssDNA-seq data are determined that are used to find patterns of association with nucleosomes. The analysis of 500 bp regions centered on a DNA secondary structure determined based on MNAse-seq data together with ssDNA-seq reveals three types of patterns:

1) The region is nucleosome free (pattern 0)
2) The structure is surrounded by one side with nucleosome (pattern 1)
3) The structure is surrounded by both sides (pattern 2)

The patterns are illustrated on fig.1.



A B C

Fig. 1. Three types of patterns: A) nucleosome-free B) structure surrounded with one nucleosome C) structure surrounded by two nucleosomes

The most important observation is that nucleosome is never located on a structure in actively transcribed cells. The biological hypothesis is that the pattern 1 and the pattern 2 are involved in regulation processes. Secondary structures may act as barriers preventing nucleosome formation or blocking nucleosome movement. Evidences of that kind of behavior are reported in the literature. For example, the chromatin remodeling complex freed DNA from histone proteins and left the DNA in the condition which favor Z-DNA structure [9].

For the reason of simplicity, the pattern 1 and the pattern 2 are united into one which is further designated as the class 1, while nucleosome free regions are denoted as class 0.

The distribution of classes among different types of structures is shown in table 3.

Table 3. Distribution of Classes

Heading level	Z-DNA	H-DNA	G-quadruplexes
Nucleosome Free Region (Class 0)	12 889 (51%)	9 816 (57%)	13 173(65%)
Nucleosome From One Side or From Two Sides (Class 1)	12 173 (49%)	7 293 (43%)	7 086(35%)

The aim of a machine learning application for this task is to distinguish genomic regions with DNA secondary structures with regulation pattern from non-regulative structures. For this purpose classifiers are trained for each type of structures. The results are presented in table 4.

Table 4. Results of Machine Learning Models Training

Algorithm	Measure	Z-DNA	H-DNA	G-quadruplex
Random Forest	ROC-AUC	0.67	0.81	0.87
	Accuracy	0.67	0.82	0.88
	Recall	0.76	0.9	0.93
	Precision	0.64	0.81	0.89
XGBoost	ROC-AUC	0.67	0.81	0.86
	Accuracy	0.66	0.82	0.88
	Recall	0.79	0.88	0.93
	Precision	0.62	0.81	0.89

To begin with, both algorithms show almost the same results. Moreover, models for G-quaruplexes and H-DNA show good performance with prediction quality higher than 80%. This corresponds to the results of researches which aim to predict nucleo-

some positions., and the best results are demonstrated by the models based on neural network models. In addition, the poorest quality are demonstrated by the classifier which distinguishes Z-DNA regulatory pattern. The possible reason is the feature set used for these models. It consists of 2-tuple and 3-tuple nucleotide compositions. G-quadruplexes and H-DNA are formed in specific sequences, so it is natural to expect they are well predicted based on sequence content, while Z-DNA has more complex formation preferences.

Nevertheless, all the constructed models are significantly better than a random guessing leading to the idea that more complicated models may result in a better classification.

4 Conclusion

DNA exists in many forms. Non-B-DNA conformations may be involved in main molecular processes such as transcription and replication. One of the mechanisms is the governance of nucleosome positioning. To evaluate the existence of positional relationship between DNA structures and nucleosomes the data on nucleosome and DNA structure maps were combined and then machine learning models were trained to predict the patterns for a genomic region. Both Random Forest classifier and XGBoost classifier showed good performance on G-quaruplexes and H-DNA while the quality of the model for Z-DNA is not high.

The practical applications of the obtained results could arise from the abilities of non-B DNA structures serve as targets for drugs, and in this respect it is important to understand the extent of the distribution of patterns involving DNA secondary structures across the entire genome. Thus, controlling the formation of non-B DNA structures may promote or inhibit production of harmful proteins including oncoproteins.

Using G-quadruplexes as targets for drugs is widely discussed in literature [34-36]. Specifically, many quadruplexes are found in promoters of oncogenes, and targeting quadruplexes by small ligands is considered as potential anticancer therapy [34]. Also, G-quadruplexes are found in regulatory regions of viral genomes, and it opens a possibility to use them as targets in antiviral therapy. Z-DNA is also found in genomic regulatory regions and there are proteins that bind specifically Z-DNA [37]. Increased transcription of some oncogenes was associated with Z-DNA formation. H-DNA, or triplex DNA, is a form where RNA binds directly to double-stranded DNA. The regulatory potential of RNA is huge, and therapeutic potential is also high including anticancer therapy [38]. Overall, all the classes of non-B DNA structures can potentially be used in biomedical applications, and developing computational approaches could help in the design of experiments.

References

1. Libbrecht, M., Noble, W.: Machine learning applications in genetics and genomics. Nature Reviews Genetics. 16, 321-332 (2015).

2. Degroeve, S., De Baets, B., Van de Peer, Y., Rouze, P.: Feature subset selection for splice site prediction. Bioinformatics. 18, S75-S83 (2002).
3. Barash, Y., Calarco, J., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B., Frey, B.: Deciphering the splicing code. Nature. 465, 53-59 (2010).
4. Heintzman, N., Stuart, R., Hon, G., Fu, Y., Ching, C., Hawkins, R., Barrera, L., Van Calcar, S., Qu, C., Ching, K., Wang, W., Weng, Z., Green, R., Crawford, G., Ren, B.: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics. 39, 311-318 (2007).
5. Zhang, J., Peng, W., Wang, L.: LeNup: learning nucleosome positioning from DNA sequences with improved convolutional neural networks. Bioinformatics. 34, 1705-1712 (2018).
6. Svozil, D., Kalina, J., Omelka, M., Schneider, B.: DNA conformations and their sequence preferences. Nucleic Acids Research. 36, 3690-3706 (2008).
7. Widom, J.: The Genomic Code for Nucleosome Positioning. Biophysical Journal. 98, 608a (2010).
8. Wang, A., Quigley, G., Kolpak, F., Crawford, J., van Boom, J., van der Marel, G., Rich, A.: Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature. 282, 680-686 (1979).
9. Rich, A., Zhang, S.: Z-DNA: the long road to biological function. Nature Reviews Genetics. 4, 566-572 (2003).
10. Frank-Kamenetskii, M., Mirkin, S.: Triplex DNA Structures. Annual Review of Biochemistry. 64, 65-95 (1995).
11. Felsenfeld, G., Davies, D., Rich, A.: Formation of a Three-Stranded Polynucleotide Molecule. Journal of the American Chemical Society. 79, 2023-2024 (1957).
12. Zain, R., Sun, J.: Do natural DNA triple-helical structures occur and function in vivo?. Cellular and Molecular Life Sciences. 60, 862-870 (2003).
13. Jain, A., Wang, G., Vasquez, K.: DNA triple helices: Biological consequences and therapeutic potential. Biochimie. 90, 1117-1130 (2008).
14. Hoyne, P., Maher, L.: Functional Studies of Potential Intrastrand Triplex Elements in the Escherichia coli Genome. Journal of Molecular Biology. 318, 373-386 (2002).
15. Gellert, M., Lipsett, M., Davies, D.: Helix Formation by Guanilic Acid. Proceedings of the National Academy of Sciences. 48, 2013-2018 (1962).
16. Zhao, J., Bacolla, A., Wang, G., Vasquez, K.: Non-B DNA structure-induced genetic instability and evolution. Cellular and Molecular Life Sciences. 67, 43-62 (2009).
17. Huppert, J.: Prevalence of quadruplexes in the human genome. Nucleic Acids Research. 33, 2908-2916 (2005).
18. Hänsel-Hertsch, R., Di Antonio, M., Balasubramanian, S.: DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. Nature Reviews Molecular Cell Biology. 18, 279-284 (2017).
19. Epstein, R.: Human molecular biology. Cambridge University Press, Cambridge (2003).
20. Garner, M., Felsenfeld, G.: Effect of Z-DNA on nucleosome placement. Journal of Molecular Biology. 196, 581-590 (1987).
21. Westin, L., Blomquist, P., Milligan, J., Wrange, Ö.: Triple helix DNA alters nucleosomal histone-DNA interactions and acts as a nucleosome barrier. Nucleic Acids Research. 23, 2184-2191 (1995).
22. Hänsel-Hertsch, R., Beraldi, D., Lensing, S., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M., Tannahill, D., Balasubramanian, S.: G-quadruplex structures mark human regulatory chromatin. Nature Genetics. 48, 1267-1272 (2016).

23. Widom, J.: The Genomic Code for Nucleosome Positioning. Biophysical Journal. 98, 608a (2010).
24. UCSC Genome Browser Downloads, http://hgdownload.cse.ucsc.edu/downloads.html.
25. Champ, P., Maurice, S., Vargason, J., Camp, T., Ho, P.: Distributions of Z-DNA and nuclear factor I in human chromosome 22: a model for coupled transcriptional regulation. Nucleic Acids Research. 32, 6501-6510 (2004).
26. Inverted Repeats Finder Download Page, http://tandem.bu.edu/irf/irf.download.html.
27. Kouzine, F., Wojtowicz, D., Baranello, L., Yamane, A., Nelson, S., Resch, W., Kieffer-Kwon, K., Benham, C., Casellas, R., Przytycka, T., Levens, D.: Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. Cell Systems. 4, 344-356.e7 (2017).
28. Teresa Przytycka Research Page, https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#nonbdna.
29. Bowtie, http://bowtie-bio.sourceforge.net/index.shtml.
30. bedtools: a powerful toolset for genome arithmetic — bedtools 2.27.0 documentation, https://bedtools.readthedocs.io/en/latest.
31. Breiman, L.: Random Forests. Machine Learning. 45, 5-32 (2001).
32. XGBoost Documentation — xgboost 0.81 documentation, https://xgboost.readthedocs.io/en/latest.
33. scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation. https://scikit-learn.org/0.20.
34. Duchler, M.: G-quadruplexes: targets and tools in anticancer drug design. J Drug Target, 20, (5), 389-400 (2012).
35. Hurley, L.H., Wheelhouse, R.T., Sun, D., Kerwin, S.M., Salazar, M., Fedoroff, O.Y., Han, F.X., Han, H., Izbicka, E., and Von Hoff, D.D.: G-quadruplexes as targets for drug design. Pharmacol Ther. 85, (3), 141-158 (2000).
36. Ruggiero, E., and Richter, S.N.: G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. Nucleic Acids Res. 46, (7), 3270-3283 (2018).
37. Shin, S.I., Ham, S., Park, J., Seo, S.H., Lim, C.H., Jeon, H., Huh, J., and Roh, T.Y.: Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. DNA Res. (2016)
38. Jain, A., Wang, G., and Vasquez, K.M.: DNA triple helices: biological consequences and therapeutic potential. Biochimie. 90, (8), 1117-1130 (2008).