Sequence dependence DNA curvature in atomistic simulations and informative complexity measures for nucleosome occupancy*

Hector Zenil†, Samuel Demharter, Alberto Hernández-Espinosa and Peter Minary

Department of Computer Science, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

Abstract

Knowing the precise locations of nucleosomes in a genome is key to understanding how genes are regulated. We investigate the correlation of sequence complexity and DNA structure by using a prevailing model of DNA curvature prediction and recent experimental data on sequence curvature affinity. We reproduced the curvature experimental results with state-of-the-art atomistic simulations (*MOSAICS*) and found correlations across universal measures of complexity ranging from classical information to algorithmic complexity, in agreement with literature on regulation dependence of DNA curvature similar and beyond GC-content prediction. We validated these ideas with experimentally validated data of nucleosome occupancy from a Yeast chromosome thereby potentially contributing to positioning prediction. We introduce methods and tools written in a software library for DNA conformational analysis written in the *Wolfram Language* as an interface between the atomistic simulation software and mainstream DNA structural analysis software (*3DNA* and *Curves+*) for the analysis of local and global properties of nucleic acids.

Keywords: DNA structure; DNA complexity; nucleotide curvature; DNA bending; nucleosome occupancy

*HZ, PM and SD conceived and designed the experiments, SD and HZ performed the experiments, HZ, PM and AH performed the analysis. HZ and AH wrote *MOSAICA*.

†Corresponding author: hectorz@labores.eu

1 Introduction

Structural properties are fundamental in gene regulation, for example, bending may provide different degrees of accessibility to gene promoters, the switch regions that turn genes on, by bending or looping the DNA. For example, the transcriptional repressor CTCF, which besides regulating transcription (e.g. repressing the insulin-like growth factor 2), also physically binds to itself to form homodimers [12], which causes the bound DNA to form loops and parts of the DNA to make contact to other farther regions of DNA putting in physical contact genes that would otherwise be not regulated according to the symbolic sequence alone [10]. While we are moving towards sequencing at epigenetic resolution [4] and 3D genome sequencing [5] most analytics are performed at the level of the DNA sequence rather than its structure. Another example is the variant papillomavirus E2 protein binding sites that display intrinsic shapes supported by experimental data from X-ray crystallography analysis. The data suggest that the sequence containing 5′-AATT is intrinsically curved toward the center of the narrowed minor groove. Additional intrinsic curvature in the flanking major grooves gives rise to an overall helix axis deflection of 10 degrees. In contrast, the sequence containing 5′-ACGT and 5′-GTAC are straight [13].

We report on synthetic experiments and simulations to estimate intrinsic local and global properties (mainly natural curvature and bending) of each three test sequences validated experimentally and several others. Our DNA curvature estimates correlate with the published order of E2 binding site affinities for HPV-16 E2 protein [13]. The results are in agreement with the hypothesis that E2 binding affinity reflects the intrinsic shape or flexibility of variant E2 binding sites. Specifically, the degree of intrinsic DNA curvature (or predisposition to curvature) as found in the E2 binding affinity. The most curved according to the simulation was the E2 site containing the 5′-AATT subsequence (see Results) also in accordance with the prediction by the Liu-Beveridge dinucleotide model). The simulating and general results are all in agreement with the literature, the theoretical expectations, the Liu-Beveridge model and the validated experiments [13].

2 Methodology

The chart in Fig. [**?**] shows a summarized diagram of the methodology and results.

2.1 In silico replication of DNA experimental curvature

Reported in recent literature is the indirect readout of certain proteins that bind to a DNA sequence based on its intrinsic three-dimensional shape or curvature of a DNA binding sequence apart from direct protein contact with DNA base pairs. DNA bending has been observed in X-ray structures. In [13] the differing affinities of human papillomavirus (HPV) E2 proteins for different E2 binding sites have been validated to reflect indirect readout. The sequences with different affinities induced by the subsequence AATT were investigated and for the first 3 experimentally validated [13]:

1. 5'-ACCGAATTCGGT (high affinity)
2. 5'-ACCGTTAACGGT (medium affinity)
3. 5'-ACCGACGTCGGT (low affinity) and
4. 5'-GCGCGCGCGCGC (no affinity)

The three first sequences have all the same prefix and suffix and any distinction in their structural properties can only therefore be due to the affinity patterns (4 nucleotides). The fourth sequence is for control against a GC repetitive sequence only and will not always be considered for comparison because of its incomparable nature relative to the others. We performed atomistic Monte-Carlo simulations on these three sequences.

Initial formulation of models for DNA bending was prompted by recognition that DNA must be bent for packaging into nucleosomes. Bending in nucleosomes should occur by roll alternately toward both major and minor grooves, in preference to tilting toward one of the strands; this is now the generally accepted view for protein-induced DNA curvature. Trifonov and Sussman [11] developed the idea of a "wedge" contributed by independent dinucleotide steps in DNA and presented evidence for periodic repetition of particular dinucleotides (including AA) as facilitators of bending and hence nucleosome formation in eukaryotic DNA sequences.

Different angles of deformation are possible when one base stacks on another. One angle is *tilt angle*, which is in the direction of the hydrogen bonding. Another is the *roll angle*, which occurs at 90 degrees to the direction of the hydrogen bonding. The tilt angle opens in the direction of the phosphate backbone whereas roll can open towards the major or the minor groove (see Fig 3).

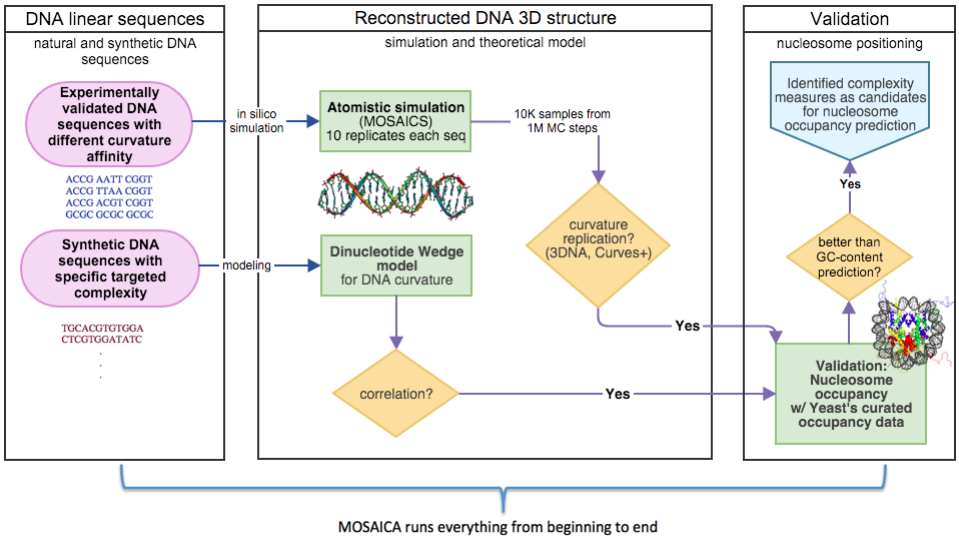

Figure 1: Flow chart of experimental setting from linear sequence to reconstruction of 3D structure (mainly curvature) through simulation, comparison, calibration, validation from two experiments that can be seen as undertaken separately (atomistic simulation and Dinucleotide Wedge modelling) suggest a common validation with nucleosome occupancy leading to the identification of candidate measures for positioning prediction better than GC-content alone.

2.2 Atomistic simulation by natural moves

Natural Move Monte-Carlo (NMMC) is a highly efficient conformational sampling method that uses generalised coordinates to approximate motions in macromolecular structures. The algorithms are implemented in *MOSAICS* (Methodologies for Optimisation and Sampling in Computational Studies), a software publicly available (http://www.cs.ox.ac.uk/mosaics/) and released under open license capable of nanoscale simulation achieved by customizable hierarchical degrees of freedom called 'natural moves', thereby circumventing limitations of conventional molecular modelling [7, 8].

2.3 Dinucleotide Wedge model

The wedge model suggests that bending is the result of driving a wedge between adjacent base pairs at various positions in the DNA. The model

assumes that bending can be explained by wedge properties attributed solely to an AA dinucleotide (8.7 degrees for each AA) [11]. No current model provides a completely accurate explanation of the physical properties of DNA such as bending [2] but the Wedge model (just as the, more basic, so-called Junction model which is also less suitable for short sequences and less general [3]) reasonably predict the bending of many DNA sequences [9].

2.4 Analysis of structural properties of DNA

Analysis software may fail for short sequences with small individual bends are all directed in the same direction or in the same plane. *3DNA* (available at `http://x3dna.org/`) is an integrated software system initially developed at Rutgers for the analysis of three-dimensional nucleic-acid local structures. The software determines a wide range of conformational parameters, including the identities and rigid-body parameters of interacting bases and base-pair steps, the nucleotides comprising helical fragments, the area of overlap of stacked bases and so on. *Curves+*, in turn, is a software for analysing the helical, backbone and groove parameters of nucleic acid structures, hence rather global properties of DNA.

Linear symbolic sequences of nucleotides were threaded onto an original straight (regular) B-form DNA and a Monte-Carlo knowledge-driven simulation with *MOSAICS* taking 10 000 equidistributed samples out of a million steps was run. For the affinity experiment 20 replicates were generated and averaged at every step both for *3DNA* and *Curves* analysis. The following are the results.

3 Results

3.1 Curvature complexity dependence in the Dinucleotide Wedge Model

In agreement with synthetic experiments (see Fig. 2), Fig. 3 (top) of short sequence complexity-driven curvature, exons as coding regions from the human genome were assigned smaller curvature than non-coding regions inside (introns) and outside genes (intragenic) according to the Dinucleotide Wedge model. This was taking a sample of 10 *average* genes (i.e. genes of about 10K bps) trimmed in subsequences of size 12 for which their curvature according to the Dinucleotide Wedge model was calculated. Exons had the lowest curvature followed by intragenic and introns with the highest curvature. The chosen genes are C13orf30, CKAP4, KIAA1394, LOC100128594,

LOC100131258, LOC100133163, MKI67IP, NRAS, PRSSL1, RHBDD2. There is some bias towards median genes located in fewer chromosomes than by chance, but they were reasonably distributed along all chromosomes.

The most interesting observation from Fig. 2A is that all complexity measures negatively correlate with curvature and curvature angles but positively correlates with the Logical Depth-based measure. Logical Depth is the only measure of increasing structural complexity rather than increasing randomness. It suggests that the greatest the Logical Depth of a sequence the lower curvature/angle (hence possibly regions of higher evolutionary pressure). The greatest correlation found was between curvature and the Logical Depth-based estimation (see Methodology and Sup Info).

Moreover, Fig. 2D shows that inclination and tip and therefore bending angle as calculated from *Curves+* corroborate the replication of the natural curvature differences among the experimental validated sequences. The one-way analysis of variance (ANOVA) test on the results in Fig. 2B and C show that the differences between the median values of the simulated curvature and the experimentally validated curvatures are in the same direction.

Fig. 2D shows that the average minor groove closes in one strand while the major groove displays a slight opening in the opposite strand as expected and experimentally found [13] and twisting as also suggested by *3DNA* in Fig. 4 in the Sup. Mat.

An investigation of sequence complexity (information-theoretic) curvature and curvature angle dependence by GC content, Shannon entropy, entropy rate, Compress, Bzip2, Kolmogorov by means of CTM and BDM and Logical Depth approximations (see Sup Info for description of all complexity measures) was undertaken both for artificially generated sequences ranging a wide spectra of complexity values including extremes (lower and highest complexity) for each measure and of real DNA sequences (from the human genome). Results are reported in Figs. 3 (top) and 2.

3.2 Structural atomistic simulations

In Fig. 4 trends of all local parameters investigated show different weak sequence dependencies after a multiple simulation (20 replicates each) of the short DNA sequences with different AATT affinities.

Groove geometry is measured for any nucleic acid with more than one strand. WIJ and DIJ indicate the widths and depths measured between strands I and J. In a canonical right-handed duplex, with the first strand oriented $5'3'$, W12 and D12 correspond to the minor groove and W21 and D21 to the major groove as measured with *Curves+*. Figs. 2, shows reproduced

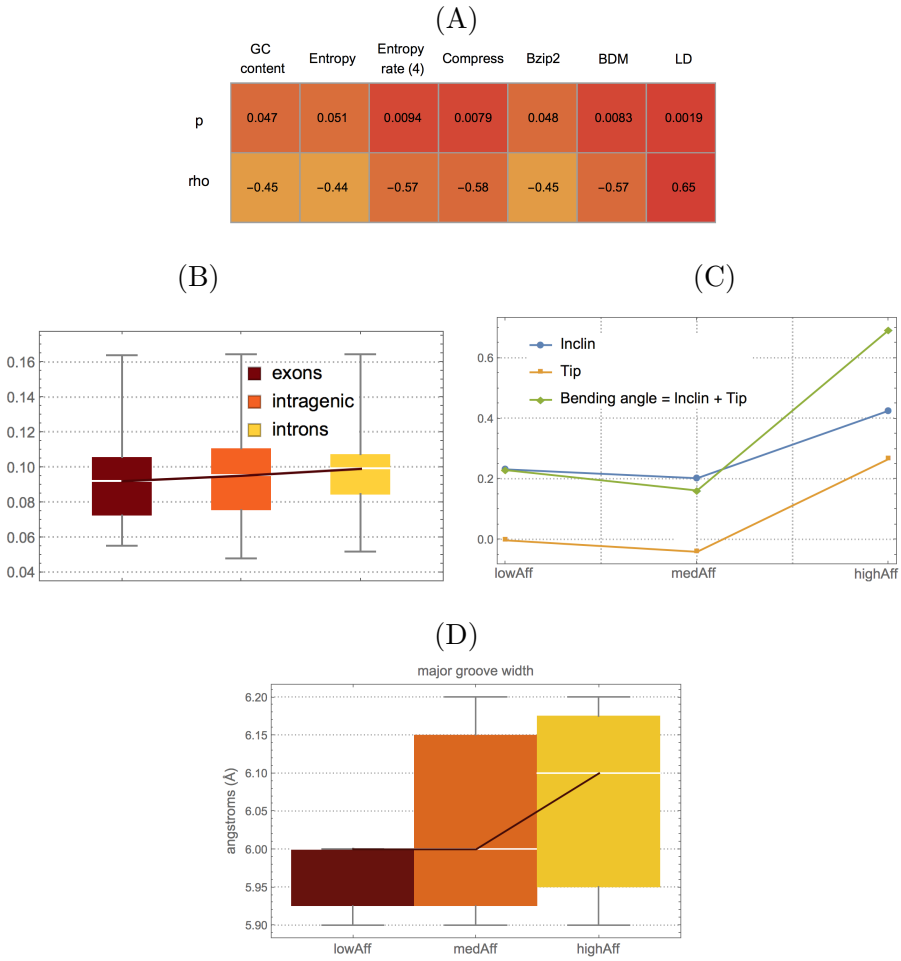

Figure 2: (A) Correlation of artificially generated DNA sequences with different complexity and the calculated natural curvature according to the Dinucleotide Wedge model (see Sup Info). (B) Natural DNA curvature/bending of real DNA regions from "average" length coding, non-coding and intragenic regions showing the expected curvature according to the Dinucleotide Wedge model as it would be packed in nucleosome arrangements (joined data points are for illustration purposes only). (D) Replicated natural curvature (values correspond only to the central 4 nucleotides determining the curvature) of different affinity DNA sequences by atomistic simulation (*MOSAICS*) with the major groove opening and greater bending (C) for high (AATT) affinity as calculated with *Curves+*

Pearson p-values	**In vivo**	**In vitro**
In vivo	0	4.6^{-81}
In vitro	4.6^{-81}	0
GC-content	1.66^{-13}	1.38^{-222}
1-Entropy	1.74^{-74}	1.65^{-203}
2-mer Entropy	2.13^{-67}	2.86^{-124}
Total k-mer Entropy rate	7.1^{-20}	3.56^{-55}
Compress	1.94^{-16}	3.92^{-16}
BDM	2.66^{-32}	6.94^{-21}
LD	1.15^{-43}	4.05^{-101}

Table 1: Correlation of nucleosome in vivo and in vitro occupancy with various complexity measures found to be correlated to DNA natural curvature and potential candidates to contribute to GC-content traditional nucleosome positioning prediction.

trends in the curvature of the three E2 binding site variants with different AATT affinities and a minor groove that slightly closes for one strand and other opens for the major groove in agreement with experimental data.

The ANOVA tests on the various structural differences shows that these differences are statistically significant after the atomistic simulations as estimated by both 3DNA and *Curves+*, and that the DNA sequences are bending in the predicted fashion according to the experimental evidence.

With the curvature theoretical so-called wedge model we calculated the predicted theoretical curvature of a set of sequences to see how they correlated to various information-theoretic measures. We found that all correlations were negative and none of them much better than the correlation with GC content alone, except for logical depth as theoretically expected, a measure of "decompression time" that one can identify with biological evolution [1].

In Table 1, we show how some complexity measures are informative about nucleosome occupancy (in agreement with the natural curvature and bendability results).

The values were calculated for natural DNA sequences that are the result of sliding a window along a single 20 000 bp DNA sequence from Yeast Chromosome 14th starting from position 187 000 that is well studied and for which there is in vivo and in vitro experimental data indicating and predicting nucleosome occupancy.

4 Conclusion

We emulated *in silico* curvature behaviour of DNA experimentally predicted by their sequence affinity. We investigated the correlation of several algorithmic and information-theoretic complexity measures with structural properties of DNA, that of curvature and bending, as theoretically modelled, experimentally validated and numerically simulated. We introduced various techniques of investigation that connects *in silico* simulation and estimations of sequence complexity. We release a set of functions that we have call the *MOSAICA* library, written in the *Wolfram Language* to contribute helping future investigations and automating aspects of the simulation and data analysis freely available for public use.

References

[1] Bennett, C.H., Logical Depth and Physical Complexity. In Herken, R., *The Universal Turing Machine: a Half-Century Survey*, Oxford U. Press, pp. 227-257, 1988.

[2] Structural details of an adenine tract that does not cause DNA to bend, AM Burkhoff, TD Tullius, *Nature* 331 (6155), 455–457.

[3] Donald M. Crothers, Tali E. Haran, and James G. Nadeau, Intrinsically bent DNA., Minireview. *The journal of biological chemistry.* Vol. 265, No. 13, Issue of May 5, pp. 7093-7096,1990.

[4] Johannes, Frank., Colot, Vincent., Jansen, Ritsert, C. Epigenome dynamics: a quantitative genetics perspective. *Nature Reviews*, 9: 883–889, 2008.

[5] Lieberman-Aiden, E. van Berkum, N.L. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science*, 326, 2009.

[6] Xiang-Jun Lu, 3DNA (v1.5) — A 3-Dimensional Nucleic Acid Structure Analysis and Rebuilding Software Package, 2003.

[7] Minary P, Levitt M (2010) Conformational optimization with natural degrees of freedom: A novel stochastic chain closure algorithm. *J Comput Biol* 17:993–1010.

[8] Sim, A. Y. L.; Levitt, M.; Minary, P., Modeling and design by hierarchical natural moves, *Proceedings of the National Academy of Sciences*, vol. 109, issue 8, pp. 2890-2895.

[9] RR. Sinden, *DNA Structure and Function*, Academic Press, 1994.

[10] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, Erez Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159, 2014.

[11] Ulanovsky, LE, and Trifonov, EN, Estimation of wedge components in curved DNA, *Nature*, 326, pp 720–722, 1987.

[12] Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species, *Mol. Cell* 13 (2): 291—8, 2004.

[13] Jeff M. Zimmerman and L. James Maher, III, Solution measurement of DNA curvature in papillomavirus E2 binding sites, *Nucleic Acids Res.* Sep 1; 31(17): 5134–5139, 2003.

Supplementary Information

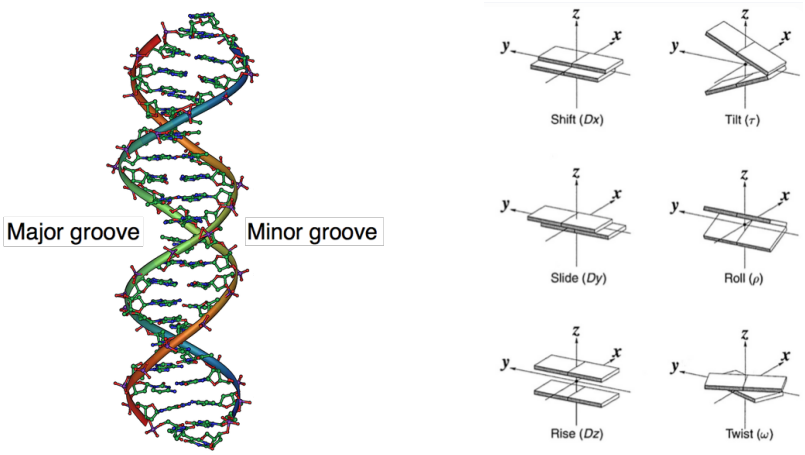

Figure 3: Left: DNA sequence (ACCGAATTCGGT) illustrating the major and minor grooves. Right: Sequential base pair (dinucleotides) structural parameters (adapted from the *3DNA* documentation) that are investigated in connection to the DNA symbolic sequence complexity.

The ANOVA tests among the different replicates plotted in the histograms in Fig. 4 indicate that the differences are unlikely by chance: for shift (F ratio 676.2 and p value 6.3^{-429}), slide (F ratio 1392.71 and p value 4.6^{-862}), rise (F ratio 1032.35 and p value 6.8^{-647}) and twist (F ratio 1219.78 and p value 1.8^{-759}).

4.1 Synthetic sequences

The 20 short artificial DNA sequences generated ranging a wide range of different patterns, repetitions, random looking strings and extremely simple ones were:

AAAAAAAAAAAA	*ATATATATATAT*	*AAAAAATTTTTT*
AAAAAAAAATAA	*AAAAAAAACAAT*	*AAGATCTACACT*
ATAGAACGCTCC	*ACCTATGAAAGC*	*TAGGCGGCGGGC*
TCGTTCGCGAAT	*TGCACGTGTGGA*	*CTAAACACAATA*
CTCTCAGGTCGT	*CTCGTGGATATC*	*CCACGATCCCGT*
GGCGGGGGGTGG	*GGGGGGGCGGGC*	*GGGGGGCCCCCC*
GCGCGCGCGCGC	*GGGGGGGGGGGG*	

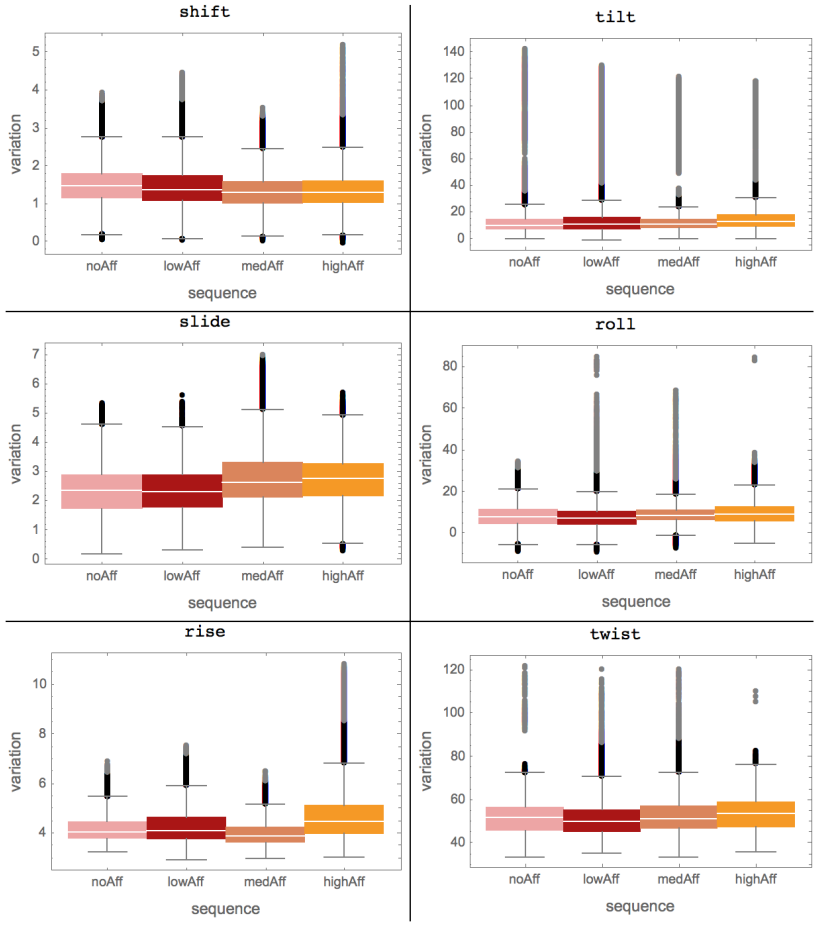

Figure 4: Structural parameters display some trends indicating sequence dependency from the multiple simulation (20 replicates each) of short DNA sequences with different AATT affinities. The bending angle between two base-pairs is defined [6] by $\sqrt{Roll^2 + Tilt^2}$ which displays a fair fit with distinguishable averages while $\sqrt{Roll^2 + Tilt^2}$ and distance between base-pair centres [6] by $\sqrt{Shift^2 + Slide^2 + Rise^2}$ is close to zero indicating that the length between end points of the sequences remain the same despite their different natural curvature.

Complexity measures

4.1.1 Classical Entropy and Entropy rate

4.1.2 Lossless compression algorithm

4.1.3 Algorithmic probability estimations

4.1.4 Logical Depth-based measures