# Unsupervised Learning of Transient Structural Motifs of ssDNA Using Translationally and Rotationally Invariant Features

**Karis Kungsamutr**[a], **Yan Jie**[a], **N.Duane Loh**[a]

[a] *National University of Singapore, Singapore 119077* e1101902@u.nus.edu, phyyj@nus.edu.sg, duaneloh@nus.edu.sg

* Presenting author

## 1. Introduction

Single-stranded DNA (ssDNA) plays an essential role in processes such as recombinant DNA repair and transcription, which are vital for cell survival. They are also commonly used as aptamers, synthetic strands of nucleic acids designed for binding specific target molecules, functioning as drug molecules or drug delivery systems [1]. Understanding how these functions are performed requires studying the specific structural properties of ssDNA and how they can be changed under different conditions.

The functions of ssDNA involve interactions with other molecules which place it under force, and it is well established that the reaction of ssDNA to external force is more complex than unstructured polymer models can explain. For example, when applying a force to extend a polyA strand of ssDNA, rather than the force simply increasing with extension, force plateaus over a specific range of extensions can be observed where the force required to maintain the extension remains constant regardless of an increase or decrease in extension [2]. More recently, the binding and dissociation of ssDNA to other molecules have also been shown to occur only under specific force ranges [3]. Understanding these responses to force requires identifying the atomic pathways and quasi-stable states involved.

Molecular dynamic (MD) simulations can give insights into such pathways and quasi-stable states, but identifying them can be difficult due to the large number of atoms in the molecule and number of time-steps produced by any given simulation. It is also necessary to distinguish those atomic pathways which significantly affect the dynamics of the ssDNA from thermal noise. This is where unsupervised machine learning can help identify prevalent and persistent spatial-temporal motifs formed by ssDNA.

## 2. Are there structural motifs in ssDNA?

Our understanding of the function of double-stranded DNA is already based on the stable motif of the double helix formed as a result of complementary base-pairing between DNA strands. Apart from the canonical Watson-Crick base pairing, non-canonical motifs such as the alternative Hoogsteen base pairing and the G-quadruplex observed between 4 guanine residues also play significant roles in cell biology. MD simulations have successfully been used to study the atomic pathways leading to these motifs, and by extension should also be applicable to study such motifs in ssDNA.

In order to extract those persistent motifs relevant to the function of ssDNA, suitable featurizations of atomic clusters in these simulations must be identified. For such features to be effective, they should be translationally and rotationally invariant while still capturing the atomic configuration of the target molecule. An example of one such featurization previously used in MD simulations are the dihedral angles between key atoms in simulations of both proteins and nucleic acids. Applying Principal Component Analysis (PCA) to these featurizations have been shown to successfully distinguish between the different conformations the molecule can take (a procedure referred to as dPCA) [4]. As such, identifying the featurization which best captures the most significant movements of molecules ssDNA under force should also pick out the relevant motifs.

## 3. Generating Molecular Dynamic Trajectories

In order to study how ssDNA changes in response to external mechanical force, MD simulations of 16 base ssDNA strands (including polyA, polyT, polyC, polyG and a sequence of randomly chosen bases) were used to generate trajectories of the polymers under force. For each strand, the ssDNA was first stretched to its full contour length in order to break any secondary structures that might have formed. A spring force was applied to each end in order to maintain the extension along the z-axis (as shown in Fig.A2) and the simulation was run for 200 ns. The extension would then be compressed by about 1 nm, and the spring force would be adjusted to hold the ssDNA at the new extension for another 200 ns. The process was then repeated until the full range of force-dependent extensions were sampled.

Since the spring is set to hold the ssDNA at a fixed set of extensions, the force exerted by the ssDNA at each extension could be measured by the resulting compression of the spring. The average force and extension from each simulation could then be used to map out a force-extension curve, identifying which simulations involve conformations that require force to maintain and which exhibit unique behaviors such as the force-plateau. The atomic positions from the simulations could then be extracted in order to identify the relevant motifs.

## 4. Discussion: complex clusters of structures at different extensions

Handcrafted features were used in order to coarse-grain the ssDNA based on the atoms most likely to contribute to significant motion. The atoms chosen were based on the commonly used set of fiducial atoms used to describe dPCA [5]. Those atoms ignored involve rigid bonds which only affect thermally driven local dynamics less consequential to overall conformational changes. Regular dPCA assumes the variations due to dihedral angles outweigh the effect of stretching bonds, but since this might not be true for simulations where the molecule is extended under force, we featurized the structure of the entire ssDNA at each time point using the list of pairwise distances between its aforementioned fiducial atoms ($\mathbf{D}(t)$).

Transforming the set of time-dependent features of a single-stranded polyG (ssPolyG) into its UMAP embedding [6], we observed clear separation between structural motif clusters (right-column of Fig. 1). Applying DBSCAN clustering [7] to the embedding, we can then match the points in each cluster to the simulation time and the corresponding end-to-end distance of the ssDNA strand at that time-point (left-column of Fig. 1).

Its UMAP embedding show that the initial highly-extended states of the ssPolyG comprise multiple clusters of structures, each of which highly localized in simulation time. As the extension is abruptly reduced, the ssPolyG tends to hop between these smaller clusters. The hopping often occurs within a given extension while other clusters span multiple simulations, suggesting inherent quasi-stable structures which exist when the ssPolyG is overstretched and cannot be identified based on measures of the overall structure such as end-to-end distance alone.

The force plateau as shown in Fig. A1 occurs during the last four discrete steps of the end-to-end distance over time (also highlighted by the semi-transparent band in 1), during which the ssPolyG continues to hop between different clusters before settling into a single large cluster once it reaches a stable helical conformation. This then suggests the plateau is characterized by dynamic transitioning between multiple configurations, as suggested by previous theoretical studies [2].

Upon reaching the more stable configuration of a right-handed helix, the end-to-end distance then fluctuates about a constant value, occasionally collapsing to a shorter configuration distinguishable in UMAP space as smaller clusters. Curiously, even as the spring extension is lowered, very little force on average is measured from the spring despite the ssPolyG seeming to resist any change in end-to-end distance. By the final simulation, the entire structure has collapsed into a more compact structure, which the UMAP embedding is able to distinguish from the transient collapsed states during the helical structure by setting them into clearly separate clusters.

## 5. Conclusion

Using translationally and rotationally invariant features of ssDNA that capture its configuration, unsupervised machine learning methods can identify the stability of transient motifs that would not have been detectable by average properties of the overall structure. This demonstrates the utility of machine-learning methods for studying transient motifs in ssDNA dynamics.

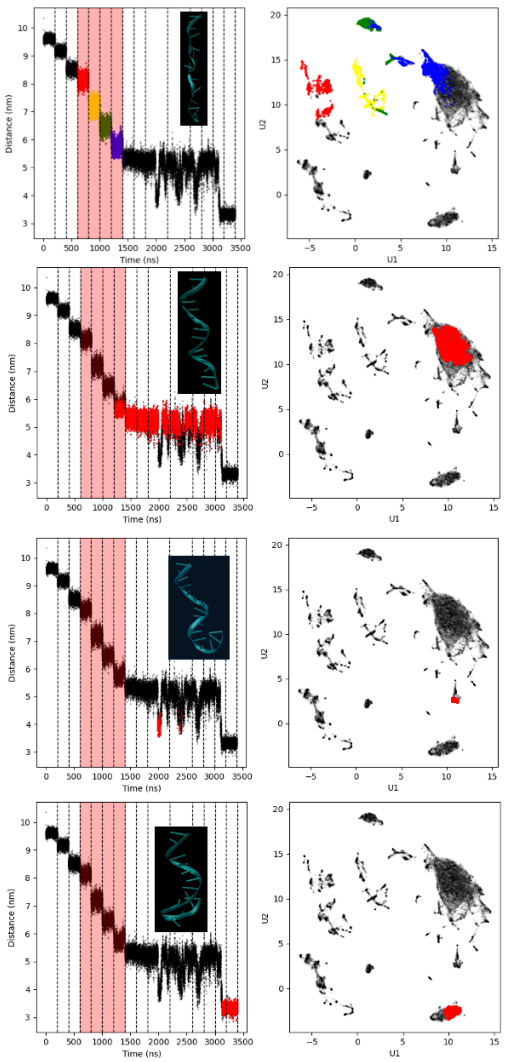


Fig. 1: Pair-wise distance features captures the transient, frustrated folding states of single-stranded polyG, as the end-to-end distance is abruptly reduced. Left column: shows the end-to-end distance of polyG as a function of simulation time. In-sets show the characteristic fold of the polyG strand, vertical lines separate the individual simulations, and the transparent band highlights simulations exhibiting the force-plateau shown in Fig. A1. Right column: each point represents its UMAP embedding of the pair-wise distance features of the entire polyG.

**Acknowledgments**

We thank the IT facility team at the Mechanobiology Institute (MBI) and Center for Bioimaging Sciences (CBIS) at NUS for the use of their compute resources.

**References**

[1] Citartan Marimuthu, Thean-Hock Tang, Junji Tominaga, Soo-Choon Tan, and Subash C. B. Gopinath. Single-stranded dna (ssdna) production in dna aptamer generation. *Analyst*, 137:1307–1315, 2012.

[2] Sanjay Kumar and Garima Mishra. Stretching single stranded dna. *Soft Matter*, 7:4595–4605, 2011.

[3] Jin Chen, Qingnan Tang, Shiwen Guo, Chen Lu, Shimin Le, and Jie Yan. Parallel triplex structure formed between stretched single-stranded dna and homologous duplex dna. *Nucleic Acids Research*, 45(17):10032–10041, 2017.

[4] Juliana Palma and Gustavo Pierdominici-Sottile. On the uses of pca to characterise molecular dynamics simulations of biological macromolecules: Basics and tips for an effective use. *ChemPhysChem*, 24(2):e202200491, 2023.

[5] Shozeb Haider, Gary N. Parkinson, and Stephen Neidle. Molecular dynamics and principal components analysis of human telomeric quadruplex multimers. *Biophysical Journal*, 95(1):296–311, 2008.

[6] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

**Appendix A.**

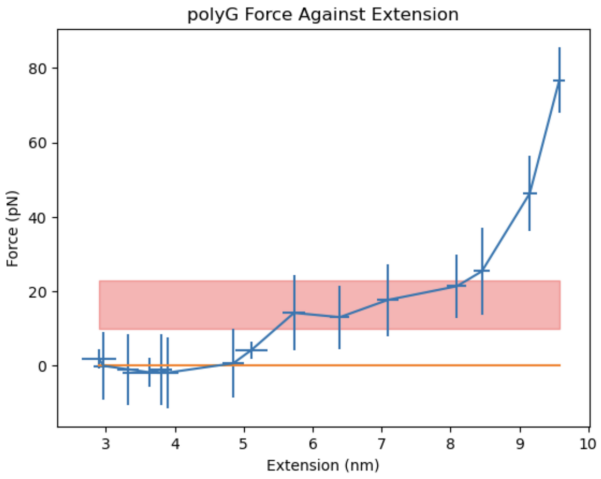


Fig. A1: The average force against the average extension of the polyG sequence, a force plateau is observed at extensions highlighted in red corresponding to the last four discrete steps of the end-to-end distance plots before the stable helix shown in the left-column graphs of Fig. 1

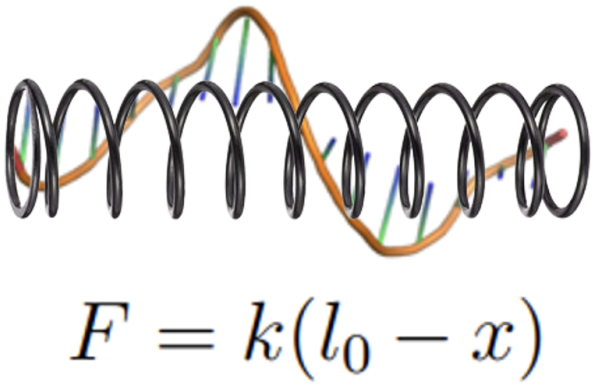


Fig. A2: Schematic of the basic setup of the simulation, where a spring force holds the terminal nucleotides of the ssDNA at a set extension by setting the equilibrium length l0 of the spring to that extension. Given the k-constant of the spring, the resulting extension x of the ssDNA can then be used to measure the force F due to the contraction of the spring. By allowing the system to equilibrate for 200 ns at each extension, the average forces on the spring at each extension and the average extension of the ssDNA can be used to plot out the curve shown in Fig. A1