Biophysical Journal
Article

Biophysical Society

Machine learning shows torsion angle preferences in left-handed and right-handed quadruplex DNAs

Kevin Li,[1] Liliya A. Yatsunyk,[1] and Stephen Neidle[2,*]
[1]Department of Chemistry and Biochemistry, Swarthmore College, Swarthmore, Pennsylvania and [2]UCL School of Pharmacy, University College London, London, United Kingdom

ABSTRACT Left-handed G quadruplexes (LHG4) have been recently discovered as a new class of G quadruplexes. The biological functions of LHG4s are still unknown, but they share a striking resemblance to Z-DNA in their helicity and jagged phosphate backbone. To further understand structural features of the LHG4s that define their left handedness, we have employed human-interpretable machine-learning methods to classify right- and left-handed G4s purely based on torsional angle analysis. Our results reveal the importance of the α, β, δ, and χ angles in left-handed structuring across both Z-DNAs and LHG4s. Our analysis may serve as the first step to understanding the conditions of formation for LHG4s and their potential biological relevance.

SIGNIFICANCE Our work explores left-handed G quadruplexes, a novel non-canonical DNA structure. Using machine-learning methods, we have demonstrated that certain backbone torsion angles (α, β, δ, and χ) can be used to differentiate between right- and left-handed G quadruplexes as well as other right- (B-DNA) and left-handed (Z-DNA) DNA structures. Our analysis may serve as the first step to understanding the conditions of formation for left-handed G quadruplexes and their potential biological relevance.

INTRODUCTION

Genomic DNA is much more than the encoded blueprint of life. It comprises canonical DNA, which is a right-handed (RH), double-helical (DH) structure, mostly B-DNA, that forms relevant to the majority of cellular DNA functionality (Fig. 1 *A*). DNA, however, can form a variety of non-canonical structures. In regions of high G-C repetition, it may form a DH left-handed (LH) DNA structure, termed Z-DNA (Fig. 1 *B*) (1). Z-DNA is characterized by a jagged phosphate backbone, a feature attributed to the alternating anti-syn dinucleotide steps going from G to C (2). Since its discovery in 1970, the biological relevance of Z-DNA has remained controversial, although in particular the characterization of the ADAR enzyme, an RNA-binding deaminase that binds Z-DNA with high specificity, has suggested a significant biological role for Z-DNA (3). Crystal structures of Z-DNA bound to the Zα domain of ADAR show contacts primarily along the phosphates in the jagged backbone, demonstrating that specificity arises from a Z conformation and not from the DNA sequence (4).

Another non-canonical DNA structural type, termed the G quadruplex (G4), is typically composed of four guanine-rich stretches associated into stacked guanine-tetrads, with RH helicity (Fig. 1 *C*) (5). Numerous RH G4 crystal and NMR structures have been reported (see, for example, refs. (6–10)). Sequences with the potential to form G4s are widely, though not randomly, distributed in many genomes, with their prevalence in the human genome being of special focus (5). Increasing evidence points toward the in vivo existence of G4s, where their presence is implicated in transcription, translation, replication, and telomeric stability (5). Interestingly, the Zα domain from ADAR displays binding affinity and stabilization of a specific conformation of G4—a parallel fold. In this fold, all four G-rich strands are oriented in the same direction (6).

In 2015, the first LH G quadruplex (LHG4; Fig. 1 *D*) was crystallized by the Phan laboratory (7). Its sequence is derived from the guanine-rich RHG4 aptamer AGRO100 (8). To date, eight crystal and two solution NMR structures of LHG4s have been reported, all deriving from the minimal LH motifs of $(GTG)_4$ or $(GGT)_3GTG$ (7,9–12). These minimal motifs not only fold into LHG4s but are also able to drive

Submitted May 13, 2022, and accepted for publication August 18, 2022.
*Correspondence: s.neidle@ucl.ac.uk
Editor: Tamar Schlick.
https://doi.org/10.1016/j.bpj.2022.08.021

Please cite this article in press as: Li et al., Machine learning shows torsion angle preferences in left-handed and right-handed quadruplex DNAs, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.08.021

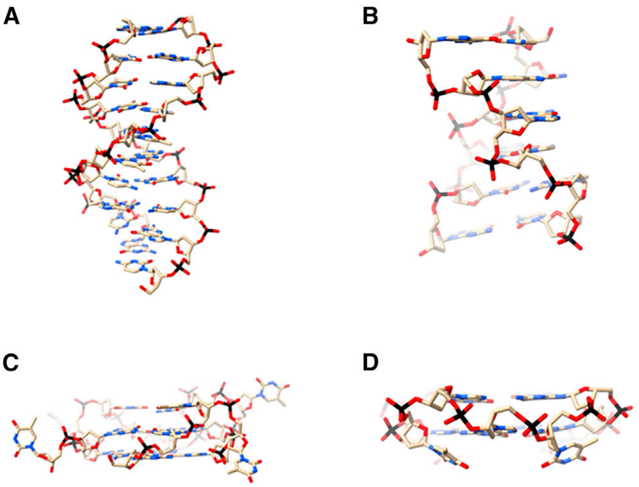

FIGURE 1 (*A–D*) Crystal structures of (*A*) B-DNA (PDB: 1BNA), (*B*) Z-DNA (PDB: 4HIF), (*C*) RHG4 (PDB: 6QJO), and (*D*) LHG4 (PDB: 6QJO). Crystal structure 6QJO contains both right- and left-handed G4s domains. In (*D*), only the LHG4 is shown. To see this figure in color, go online.

LHG4 formation in some related sequences that, alone, would adopt a RHG4 fold (7). The known LHG4 structures consist of two-tetrad G4s, stabilized in the crystal by dimerization with another two-tetrad G4 across a 5′-5′ interface (7). Structurally, LHG4s are characterized by a Z-like jagged backbone oriented in the same direction as its four neighbors (e.g., parallel G4); the thymines that separate each stretch of guanines cap the tops of the tetrads (7). However, unlike Z-DNA, all the nucleotides in LHG4 are oriented in the anti-conformation (7). Genomic searches in the human genome for the LHG4 minimal motifs $(GTG)_4$ have returned more than 10,000 hits, which is some two orders of magnitude greater than the expected random occurrence of a 12-bp motif (9). The biological relevance and role of LHG4s is still unknown, as are conditions and driving forces for their formation. Understanding these questions, however, is likely intertwined with understanding LHG4 structures.

We report here studies directed at understanding the structural differences between LHG4s and RHG4s. We have applied decision trees as a human interpretable machine-learning method to analyze torsion angles in both LHG4s and RHG4s. By generating a classifier that discriminates between LHG4s and RHG4s with accuracy of around 90%, we can interpret the produced decision tree to determine the principal torsion angles that distinguish LH and RH. Despite the decision tree never being trained on DH DNA, the methodology was capable of classifying the handedness of Z-DNA and B-DNA with greater than 86% accuracy. After supplementing the training set with Z/B-DNA samples, the accuracy of Z/B-DNA classification increased to 97%. This provides an important check on the ability of the algorithms to reproduce the experimental data and represents the initial part of our analyses.

MATERIALS AND METHODS

Data sets

The G4 data set consists of 125 nucleotides taken from seven parallel LHG4 structures and 88 nucleotides taken from six parallel RHG4 structures from the Protein Data Bank (Table 1). Apart from one NMR structure (2N3M), all G4 structures included here had been determined by X-ray crystallography. The DH data set consists of 76 nucleotides taken from eight Z-DNA crystal structures and 175 nucleotides taken from 12 B-DNA crystal structures. These are representative of the highest-resolution structures currently available (an initial study used several early B-DNA crystal structures, with resolutions in the range 1.90–3.00 Å) (Table S1). Each nucleotide sample contains a label of “0” if it originates from an LH structure or “1” from an RH structure. In addition, each nucleotide is characterized by six backbone torsion angles (α, β, γ, δ, ε, ζ) and the base/sugar glycosidic torsional angle (χ) (Fig. 2 *A*). These torsional angles were calculated using the X3DNA web server (36).

Principal-component analysis (PCA)

PCA is a widely used technique for dimensionality reduction, allowing the variance among the higher-dimensionality data set to be represented and visualized in a reduced number of dimensions. This is done by calculating principal components, which are directions in the data set along which the variation is maximized. Data points in the higher dimensionality data set are then projected onto the principal components, producing the dimensionality reduced data set. Plots of the first two principal components are useful in determining interesting clusters of examples that contain similar higher-dimensional characteristics with respect to the principal components.

Principal components were calculated from the torsional angle data set using the PCA class from the scikit-learn Python package, with *n*, the number of components, set to 2 (37). The resultant PCA graph was plotted using the Matplotlib Python package (38).

Experimental training and validation

Our models were trained on the G4 data set and evaluated using repeated, random holdout validation. Essentially, samples from the G4 data set were randomly split 70/30 into the training and testing sets, such that for the 213 nucleotides in the data set, 149 were randomly placed into the training set, and the remaining 64 were placed in the testing set. The 70/30 training/testing split is commonly used in machine learning (39). Each model then learned the classification by training on the training set and is scored based on its accuracy on the testing set, which contains samples the model has never seen. Accuracy here is defined as the number of samples classified correctly over the total samples in the testing set. This process was repeated 1000 times, each time with different nucleotides randomly split into the training and testing sets to produce a representative average accuracy that is not skewed due to sampling bias. Random holdout validation was performed using the score function from each respective classifier in the scikit-learn package (37).

ID3 decision trees

The scikit-learn ID3 decision tree is an algorithm that trains by splitting a data set recursively by feature (α, β, γ, δ, ε, ζ, χ) and across a threshold ($0 \le \theta \le 360$) until similarly classed data points (0 for LH, 1 for RH) are grouped together. The resulting decision tree produced can then be used to classify novel data points. To start, the algorithm finds a node, which is a feature and a threshold, that splits the data set into two resulting sets with the lowest entropy (40). Entropy, in this context, can be understood as a measure of class purity in a data set. The entropy of a data set is defined as $E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$, where *c* is the total number of classes (two in

TABLE 1 DNA crystal structures included in the data set

PDB ID	Structural type	DNA sequence	Resolution (Å)	References
6FQ2	LHG4	$(TG_2)_4T_2(GTG)_4T_2$	2.31	(11)
7DFY	LHG4	$(GTG)_4$	1.69	(9)
4U5M	LHG4	$G(TG_2)_3TGT_2(GTG)_4T$	1.50	(7)
6GZ6	LHG4	$G_2T_2G_2TGTG_2T_2G_2T\ (GTG)_4$	2.01	(11)
6QJO[a]	LHG4/RHG4	$G(GT)_3(GGT)_2(GTG)_4T_2$	1.80	(12)
7D5D	LHG4	$G(GT)_5G_2T(GTG)_4T_2$	1.18	(10)
7D5E	LHG4	$(TG_2)_4T_2(GTG)_4T_2$	1.30	(10)
7KLP	parallel RHG4	$A(GGGTTA)_3GGG$	1.35	(13)
6N65	parallel RHG4	$AG_3CGGTGTG_3AATAG_3AA$	1.60	(14)
3T5E	parallel RHG4	$A(GGGTTA)_3GGG$	2.10	(15)
6H5R	parallel RHG4	$TA(GGGTTA)_3GGGT$	2.00	(16)
4FXM	parallel RHG4	$A(GGGTTA)_3GGG$	1.65	(17)
2N3M[b]	parallel RHG4	$(GGT)_3TGTT(GTG)_3$	–	(18)
3P4J	Z-DNA	CG_3	0.55	(19)
4OCB	Z-DNA	CG_6	0.75	(20)
4FS6	Z-DNA	CG_3	1.30	(21)
4FS5	Z-DNA	CG_3	1.30	(21)
4HIG	Z-DNA	CG_3	0.75	(22)
4HIF	Z-DNA	CG_3	0.85	(22)
7JY2	Z-DNA	CG_3	1.00	(23)
7ATG	Z-DNA	CG_3	0.60	(24)
1BNA	B-DNA	$(CG)_2AATT(CG)_2$	1.90	(25)
2BNA	B-DNA	$(CG)_2AATT(CG)_2$	2.70	(26)
3BNA	B-DNA	$(CG)_2AATTC_{Br}GCG$	3.00	(27)
4BNA	B-DNA	$(CG)_2AATTC_{Br}GCG$	2.30	(27)
5BNA	B-DNA	$(CG)_2AATT(CG)_2$	2.60	(28)
1D60	B-DNA	CCAACITTGG	2.20	(29)
1SGS	B-DNA	$CGCTGGA_3T_3CCAGC$	1.60	(30)
1DC0	B-DNA	$CAT\ G_3C_3ATG$	1.30	(31)
1D8G	B-DNA	CCAGTACTGG	0.74	(32)
5DNB	B-DNA	CCAACGTTGG	1.40	(33)
436D	B-DNA	CGCGAA TAF TCGCG	1.10	(34)
4C64	B-DNA	CGCGAATTCGCG	1.32	(35)

[a] Structure is truncated. 6QJO was split into separate RH and LH components.

[b] An NMR structure. Only the two tetrad G4s in 2N3M were included.

this case, LH and RH) and p_i is proportion of class i data points over the total data set (40). For data sets of only two classes, entropy values are bounded between 0 and 1. A data set that is purely of one class produces the lowest entropy of 0 (assuming $\log_2 0 = 0$), while a data set that is evenly split between classes produces the highest entropy of 1. After producing two data sets, the algorithm is then recursively called on each resultant data set until the entropy drop between parent and child data sets diminishes below a preset threshold. The ID3 classifier was implemented using the DecisionTreeClassifier class from the scikit-learn Python package with the ccp_alpha parameter set to 0.04 (37). The resultant decision tree was visualized using the plot_tree function from scikit-learn (37).

RESULTS

Composition of the data set

The G4 data set consists of 125 nucleotides taken from seven parallel LHG4 structures and 88 nucleotides taken from six parallel RHG4 structures. Our study considers only parallel G4 structures (Table 1) in order to prevent introducing differences in the LH and RH data sets due to G4 topological differences—all the known LHG4 structures have parallel topology. Additionally, we include only guanines forming the tetrads and omit overhang and loop nucleotides since all current LHG4 structures are composed of single thymine loops while the RHG4 structures have longer loop lengths and also contain cytosines, adenines, and guanines in the loop regions. The DH data set consists of 76 nucleotides taken from eight Z-DNA crystal structures and 175 nucleotides from 12 B-DNA crystal structures.

Nucleotide angle analysis

We analyzed the assembly of backbone and glycosidic torsion angles (Fig. 2 *A*) using two-component PCA (Fig. 2 *B*). The resultant PCA graph captured 65.5% of the variance in the original data set. Overall, nucleotides from Z-DNA and LHG4 each show strong clustering, indicating either strong conformational homogeneity in LH folding nucleotides or low structural variability among the known crystal structures. In LHG4s, the guanosines segregated into two clusters, one populated by the first guanosine in every tract in the 5′-3′ direction and the other populated by the second guanosine (note, all known current LHG4s contain only two guanosines in a tract). Between these values lies the cluster

Please cite this article in press as: Li et al., Machine learning shows torsion angle preferences in left-handed and right-handed quadruplex DNAs, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.08.021

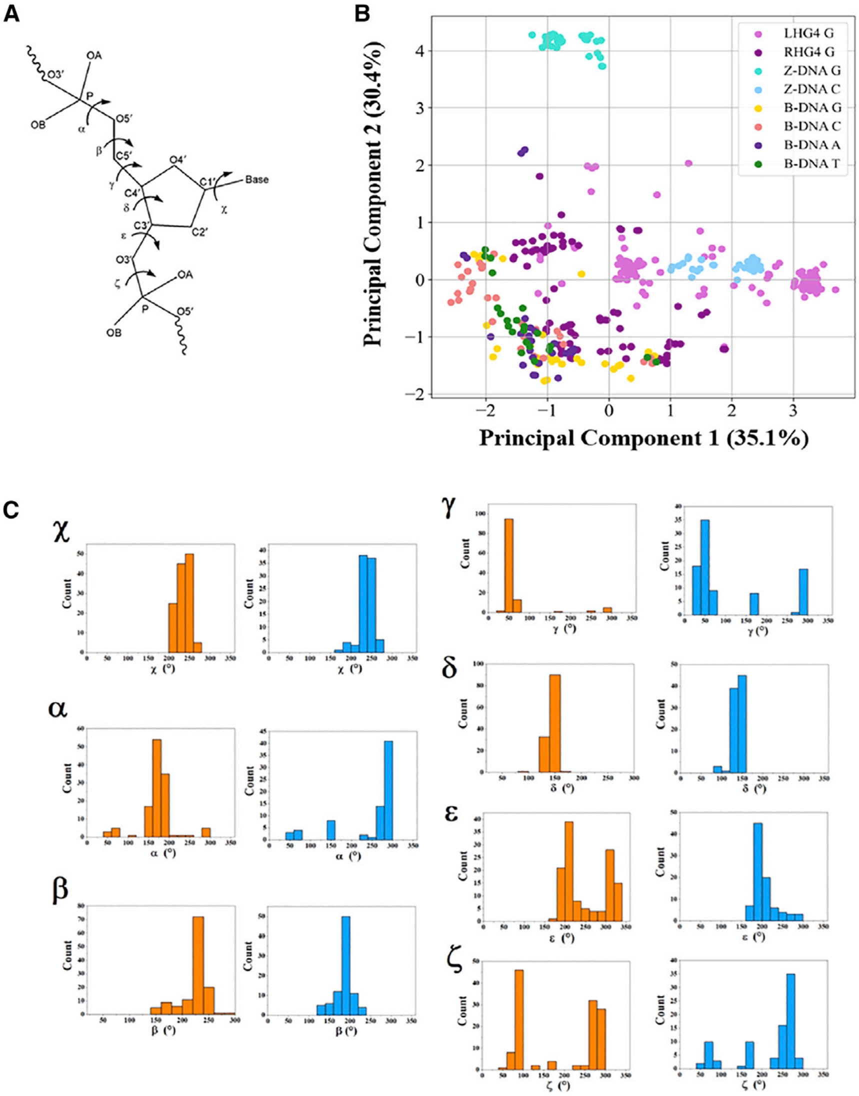

FIGURE 2 (*A*) Torsion angles considered in this work. (*B*) Two-component PCA of RH and LH DNA structures. Nucleotides are colored based on the structure they come from and their base. (*C*) Torsion angle distribution for guanosines in LHG4 (colored in orange) and RHG4 (colored in blue) across the classification data set. To see this figure in color, go online.

representing cytidines from Z-DNA. The torsional angles of Z-DNA cytidines closely match the guanosines in LHG4, with multiple nucleotides from both clusters of guanosines in LHG4 overlapping with the Z-DNA cytidine cluster. The guanosine Z-DNA cluster, however, is distantly segregated from the rest of the LH clusters. This observation can be explained by the 180° change in the guanosine glycosidic angles, a result of the syn-anti dinucleotide step in Z-DNA.

In contrast to the clustering of the LH nucleotides, RH structures show a large spread of observed torsional angles, with no clear clustering. Despite the wide distribution in RH structures, there is very little overlap between the torsion angles of RH- and LH-derived nucleotides on the PCA plot.

To further investigate the differences between LHG4 and RHG4, we plotted their torsional angle distributions across all the guanosines in the data set (Fig. 2 *C*). Of note, the ε and ζ angles in LHG4 guanosines populate a bimodal distribution unlike their counterparts in RHG4 guanosines. These two angles are principally related to the differences between the first and second LHG4 guanosine in each stretch. Between RHG4 and LHG4 guanosines, their differences in the α angle are most striking. LHG4 guanosines have α angle ranges between 150° and 200°, while α angle ranges for RHG4 guanosines accumulate above 250°.

Decision trees

To further understand how DNA torsional angles relate to handedness, we trained a decision tree on the LHG4 + RHG4 data set using the same 70/30 split. This decision tree achieved an accuracy of 89.3% over 1000 iterations. To determine the generalizability of the torsional angle thresholds to right versus left handedness, we then used the same trained model to classify B-DNA versus Z-DNA, which the model had never seen before. Our model achieved an accuracy of 85.6% over the same iterations. These results show a certain degree of transfer learning between the G4 and DH DNA data sets. Over 10 iterations, decision trees trained using only the G4 data set chose to first split on the α angle seven out of 10 times, with α values above 252°, on average, designated as the threshold for RH DNA (Fig 3 *A*). Of these resultant decision trees, all seven unanimously further split on the β angle at 150°. These decision trees perform remarkably well on the DH data set, scoring over 95%. In the remaining three iterations, decision trees chose to split first on the β angle at 215° followed by varying splits in the γ, δ, or ζ angles. These trees score below 65% accuracy on the DH data set, contributing to the 85.6% averaged accuracy.

Motivated by the possibility of transferred learning, we then asked whether adding samples from B- and Z-DNA to the training set would increase the accuracy of the decision tree in classifying LH versus RH DNA (both G4 and DH). Indeed, we observed an increased accuracy in classifying B- versus Z-DNA from 85.6% to 97.7% and LHG4 versus RHG4 from 89.3% to 92.2% over 1000 iterations (Table 2).

As more DH samples (B- or Z-DNA) are added to the training set, the decision trees produced become more likely to split first on the α angle at 252° followed by splitting on the β angle at 150°. After more than 20 DH samples have been randomly mixed into the training set, resultant decision trees unanimously split on the α and β angles at near 252° and 150°, respectively.

In the α and β angle decision tree, the α angle alone already achieves an accuracy of 87.2% in splitting the data set into homogeneous classes. To determine how well the other angles can individually be used to classify LHG4 from RHG4, we then trained decision trees using only values of a single torsional angle at a time (Fig 3 *B*).

Of the seven angles, α and β were most accurate in separating LHG4 and RHG4, scoring 87.2% and 85.2%, respectively. These scores must be compared with the baseline of 58.6% accuracy, which is the percentage of LHG4 in the G4 data set. In other words, arbitrarily guessing LH would still be correct more than half the time.

DISCUSSION

Patterns of clusters in nucleotide conformations across DNA and RNA structures have been previously identified using Euclidean clustering algorithms (41) and cluster-plot surveys of experimental structures (see, for example, refs. (23,24)). Unlike traditional algorithms where a predetermined set of rules are used to transform an input to the output, machine-learning algorithms are given the input and output of a problem in hopes of formulating the rules that connect them. These algorithms are a powerful tool to parse interesting patterns in a data set. Their application to the nucleic acid field is not a new concept. Particularly in the G4 field, high-throughput studies can generate the massive amounts of data necessary to predict G4 tertiary fold from sequence through neural networks (42). There is considerable current emphasis on deep learning, a branch of machine learning that employs neural networks and multiple layers of data. Deep learning, unfortunately, cannot be yet applied to LHG4 structures due to the lack of data on diverse folds. Only two closely related sequences have been characterized, and the general rules defining LHG4 structures remain to be elucidated. And despite the raw predictive power that deep learning can provide, it is a machine-learning model that sacrifices interpretability for predictive power. Understanding how a neural network comes to its conclusion is typically not at all obvious. Thus, in this study, where understanding how torsion angles correlate with handedness is more important than the prediction itself, we employ ID3 decision trees.

We have coupled our decision tree analysis with more straightforward approaches using PCA plots and histograms. These methods allow a more intuitive understanding of the data set and act as validation for the decision tree output. And while a direct analysis of conformational angles can also reveal trends, the classic approach does not cope well with multi-dimensional data sets, particularly when

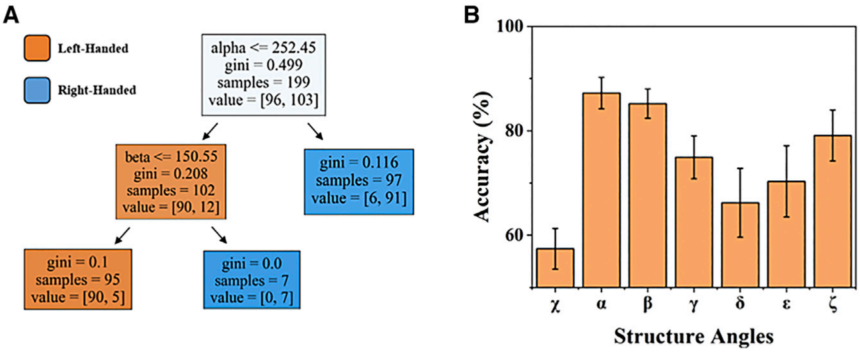

FIGURE 3 RH versus LH decision tree. (*A*) Decision tree values are an average of the most common decision tree over 10 iterations. Nodes are colored orange or blue based on whether the decision tree would classify the structure as LHG4 or RHG4 if the algorithm halted at that node. The first integer under “value” represents the number of LHG4s in the population, whereas the second integer refers to the number of RHG4s. (*B*) Decision tree accuracy in classifying RHG4 versus LHG4 using single torsion angles. Bar graph shows the average accuracy using each torsional angle over 1000 training-testing iterations. Error bars represent the standard deviation of each accuracy over the iterations. To see this figure in color, go online.

TABLE 2 Decision tree accuracy in classifying RH versus LH G4 and DH DNA (B- and Z-DNA)

DH samples added	G4 classifier (%)	DH classifier (%)
0	89.3	85.6
10	90.7	93.1
20	91.5	95.6
30	91.7	96.8
40	92.2	97.4
50	92.2	97.7

consideration of multiple angles at once produces a better prediction than single angle analysis. As further instances of LHG4 structures are discovered, a machine-learning approach may be easier to scale to either corroborate existing trends or reveal new ones.

Our decision tree analysis accurately classifies LHG4 versus RHG4 with higher than 92% accuracy by splitting on the α and β angles. Here, guanosines with α angles lower than 252° and β angles higher than 150° are classified as LH, and samples outside these values are classified as RH. These same thresholds, when applied to B- versus Z-DNA, achieve an accuracy up to 97%, demonstrating a strong generalizability between the torsional angles of similarly handed DNA folds.

Analysis of LHG4 versus RHG4 angle distributions corroborates the thresholds made by the decision tree. α values in LHG4s range primarily from 150° to 200°, while those in RHG4s range primarily from 250° to 325°, supporting the threshold of 252° set by the decision tree. The distribution of α angles in LHG4s centers around 180°. This produces a straight-line movement in the backbone progression, creating the signature zig-zag characteristic of the LH fold. Meanwhile, the distribution of α angles in RHG4 peaks at 300°, generating the even and steady curvature of the RH fold. Previous studies have also documented the angle distributions for B- and Z-DNA (41,43): α values of cytidine in Z-DNA closely cluster between 150° and 200°, peaking at 180°, while those of guanosine greatly differ due to the anti-syn step but are lower than 200°. α values in B-DNA match with RHG4, clustering between 240° and 360°. These results are consistent with the PCA plot, which shows overlapping between the RH samples as well as close clustering between cytidines in Z-DNA and the guanosines in LHG4.

Put together, our findings suggest that torsional angle difference between RH and LH DNA structures may be conserved—and, possibly, that the progression of the phosphate backbone may be intrinsically connected to the handedness of the structure. PCA plotting shows overlap between the torsional angles of LHG4 and Z-DNA, which has previously been observed in the analysis of the first LHG4 crystal structure (7). Given that the Zα domain of the ADAR enzyme preferentially binds to the jagged backbone shared between LH structures (Z-DNA and LHG4), it may be interesting to determine whether the Zα domain (4) can also bind LHG4 structures.

Additionally, PCA of the structures covered in this study shows a wide range of torsional angle values for RH structures, contrasting with the tighter clustering produced from LH angles. Although this may be a consequence of the limited diversity of current LHG4 structures, even within the same RHG4 structure, guanosine torsional angles may vary drastically more than within its LH counterpart. The wide spread of the RH torsional values demonstrates a greater degree of flexibility in the backbone, whereas the set of torsional angles where an LH structure can be stabilized appears to be very constrained and rigid, which may explain the low number of available LHG4 scaffolds. An energy landscape of DNA structures based on their torsional angles suggests wide valleys around the torsional values corresponding to RH structures. Meanwhile, the energy landscape around the mean torsional values of LH structures could be imagined as a sharp and narrow dip. For any unfolded DNA polymer rolling around on the energy landscape, it is more likely to take the RH fold as it cascades down the slope of the wide valley than it is to randomly drop into the rigid conformation of the LH fold.

These LH folds are only likely to manifest if conditions arise that make them more stable than their RH counterpart. For DH DNA, it has been shown that under conditions of, for example, negative super coiling, Z-DNA conformations are favored, as the LH helicity acts to relieve some torsional stress (44). Similarly, in sequences that fold into LHG4, it is possible that the LH form is favored because the sequence is unable to arrange the G tetrads in an RH manner. Even so, it is still unknown why LHG4-folding sequences are more adept at folding successive G tetrads into an LH form. These questions may be answered as we further explore the characteristic torsional angles of LH structures using more LHG4 crystal and NMR structures and explore the effect of overhangs on the ability of G4 DNAs to adopt RH or LH folds.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2022.08.021.

AUTHOR CONTRIBUTIONS

S.N. came up with the idea and wrote and edited the manuscript; K.L. designed research, performed research, analyzed data, and wrote and edited the manuscript; and L.A.Y. edited the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Brad Johnson (University of Pennsylvania) for critically reading the manuscript, Ben Mitchell (Swarthmore College) for his insight and for reviewing of the computational work performed and the presenting of the data, and Phillipe Hansen-Estruch (University of California, Berkeley) for his input in discussing these experiments and suggesting the use of decision trees as an interpretable classifier. This work is supported by the National Institutes of Health (1R15CA253134 to L.A.Y.) and by grants from The Wellcome Trust, the Medical Research Council, and the UCL Technology Fund to S.N. for support of quadruplex projects at UCL.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Rich, A., and S. Zhang. 2003. Z-DNA: the long road to biological function. *Nat. Rev. Genet.* 4:566–572.
2. Wang, A. H., G. J. Quigley, …, A. Rich. 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature.* 282:680–686.
3. Herbert, A. 2019. Z-DNA and Z-RNA in human disease. *Commun. Biol.* 2:7–10.
4. Schwartz, T., M. A. Rould, …, A. Rich. 1999. Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science.* 284:1841–1845.
5. Varshney, D., J. Spiegel, …, S. Balasubramanian. 2020. The regulation and functions of DNA and RNA G-quadruplexes. *Nat. Rev. Mol. Cell Biol.* 21:459–474.
6. Kang, H.-J., T. V. T. Le, …, H.-J. Park. 2014. Novel interaction of the Z-DNA binding domain of human ADAR1 with the oncogenic c-Myc promoter G-quadruplex. *J. Mol. Biol.* 426:2594–2604.
7. Chung, W. J., B. Heddi, …, A. T. Phan. 2015. Structure of a left-handed DNA G-quadruplex. *Proc. Natl. Acad. Sci. USA.* 112:2729–2733.
8. Girvan, A. C., Y. Teng, …, P. J. Bates. 2006. AGRO100 inhibits activation of nuclear factor-kappaB (NF-kappaB) by forming a complex with NF-kappaB essential modulator (NEMO) and nucleolin. *Mol. Cancer Therapeut.* 5:1790–1799.
9. Das, P., F. R. Winnerdy, …, A. T. Phan. 2021. A novel minimal motif for left-handed G-quadruplex formation. *Chem. Commun.* 57:2527–2530.
10. Das, P., K. H. Ngo, …, A. T. Phan. 2021. Bulges in left-handed G-quadruplexes. *Nucleic Acids Res.* 49:1724–1736.
11. Bakalar, B., B. Heddi, …, A. T. Phan. 2019. A minimal sequence for left-handed G-quadruplex formation. *Angew. Chem., Int. Ed. Engl.* 58:2331–2335.
12. Winnerdy, F. R., B. Bakalar, …, A. T. Phan. 2019. NMR solution and X-ray crystal structures of a DNA molecule containing both right- and left-handed parallel-stranded G-quadruplexes. *Nucleic Acids Res.* 47:8272–8281.
13. Li, K., L. Yatsunyk, and S. Neidle. 2021. Water spines and networks in G-quadruplex structures. *Nucleic Acids Res.* 49:519–528.
14. Ou, A., J. W. Schmidberger, …, N. M. Smith. 2020. High resolution crystal structure of a KRAS promoter G-quadruplex reveals a dimer with extensive poly-A π-stacking interactions for small-molecule recognition. *Nucleic Acids Res.* 48:5766–5776.
15. Collie, G. W., R. Promontorio, …, G. N. Parkinson. 2012. Structural basis for telomeric G-quadruplex targeting by naphthalene diimide ligands. *J. Am. Chem. Soc.* 134:2723–2731.
16. Guarra, F., T. Marzo, …, C. Gabbiani. 2018. Interaction of a gold(I) dicarbene anticancer drug with human telomeric DNA G-quadruplex: solution and computationally aided X-ray diffraction analysis. *Dalton Trans.* 47:16132–16138.
17. Nicoludis, J. M., S. T. Miller, …, L. A. Yatsunyk. 2012. Optimized end-stacking provides specificity of N-methyl mesoporphyrin IX for human telomeric G-quadruplex DNA. *J. Am. Chem. Soc.* 134:20446–20456.
18. Do, N. Q., W. J. Chung, …, A. T. Phan. 2017. G-quadruplex structure of an anti-proliferative DNA sequence. *Nucleic Acids Res.* 45:7487–7493.
19. Brzezinski, K., A. Brzuszkiewicz, …, Z. Dauter. 2011. High regularity of Z-DNA revealed by ultra high-resolution crystal structure at 0.55 Å. *Nucleic Acids Res.* 39:6238–6248.
20. Luo, Z., M. Dauter, and Z. Dauter. 2014. Phosphates in the Z-DNA dodecamer are flexible, but their P-SAD signal is sufficient for structure solution. *Acta Crystallogr. D Biol. Crystallogr.* 70:1790–1800.
21. Chatake, T., and T. Sunami. 2013. Direct interactions between Z-DNA and alkaline earth cations, discovered in the presence of high concentrations of $MgCl_2$ and $CaCl_2$. *J. Inorg. Biochem.* 124:15–25.
22. Drozdzal, P., M. Gilski, …, M. Jaskolski. 2013. Ultrahigh-resolution crystal structures of Z-DNA in complex with Mn(2+) and Zn(2+) ions. *Acta Crystallogr. D Biol. Crystallogr.* 69:1180–1190.
23. Harp, J. M., L. Coates, …, M. Egli. 2021. Water structure around a left-handed Z-DNA fragment analyzed by cryo neutron crystallography. *Nucleic Acids Res.* 49:4782–4792.
24. Drozdzal, P., M. Gilski, and M. Jaskolski. 2021. Crystal structure of Z-DNA in complex with the polyamine putrescine and potassium cations at ultra-high resolution. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* 77:331–338.
25. Drew, H. R., R. M. Wing, …, R. E. Dickerson. 1981. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. USA.* 78:2179–2183.
26. Drew, H. R., S. Samson, and R. E. Dickerson. 1982. Structure of a B-DNA dodecamer at 16 K. *Proc. Natl. Acad. Sci. USA.* 79:4040–4044.
27. Fratini, A. V., M. L. Kopka, …, R. E. Dickerson. 1982. Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. *J. Biol. Chem.* 257:14686–14707.
28. Wing, R. M., P. Pjura, …, R. E. Dickerson. 1984. The primary mode of binding of cisplatin to a B-DNA dodecamer: C-G-C-G-A-A-T-T-C-G-C-G. *EMBO J.* 3:1201–1206.
29. Lipanov, A., M. L. Kopka, …, R. E. Dickerson. 1993. Structure of the B-DNA decamer C-C-A-A-C-I-T-T-G-G in two different space groups: conformational flexibility of B-DNA. *Biochemistry.* 32:1373–1389.
30. Huang, D.-B., C. B. Phelps, …, G. Ghosh. 2005. Crystal structure of a free kappaB DNA: insights into DNA recognition by transcription factor NF-kappaB. *J. Mol. Biol.* 346:147–160.
31. Ng, H.-L., M. L. Kopka, and R. E. Dickerson. 2000. The structure of a stable intermediate in the A ↔ B DNA helix transition. *Proc. Natl. Acad. Sci. USA.* 97:2035–2039.
32. Kielkopf, C. L., S. Ding, …, D. C. Rees. 2000. Conformational flexibility of B-DNA at 0.74 A resolution: d(CCAGTACTGG)(2). *J. Mol. Biol.* 296:787–801.
33. Privé, G. G., K. Yanagi, and R. E. Dickerson. 1991. Structure of the B-DNA decamer C-C-A-A-C-G-T-T-G-G and comparison with isomorphous decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *J. Mol. Biol.* 217:177–199.
34. Tereshko, V., G. Minasov, and M. Egli. 1999. The Dickerson-Drew B-DNA dodecamer revisited at atomic resolution. *J. Am. Chem. Soc.* 121:470–471.
35. Lercher, L., M. A. McDonough, …, C. J. Schofield. 2014. Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chem. Commun.* 50:1794–1796.
36. Li, S., W. K. Olson, and X.-J. Lu. 2019. Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.* 47:26–34.
37. Pedregosa, F., G. Varoquaux, …, É. Duchesnay. 2012. Scikit-learn: Machine Learning in Python. https://scikit-learn.org.
38. Hunter, J. D. 2007. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* 9:90–95. https://ieeexplore.ieee.org/.
39. Gholamy, A., V. Kreinovich, and O. Kosheleva. 2018. Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. *Int. J. Intell. Technol. Appl. Stat.* 11:105–111.

40. Breiman, L., J. H. Friedman, …, C. J. Stone. 2017. Classification and Regression Trees. Routledge.

41. Černý, J., P. Božíková, …, B. Schneider. 2020. A unified dinucleotide alphabet describing both RNA and DNA structures. *Nucleic Acids Res.* 48:6367–6381.

42. Barshai, M., and Y. Orenstein. 2019. Predicting G-quadruplexes from DNA sequences using multi-kernel convolutional neural networks. *In* Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Association for Computing Machinery, pp. 357–365.

43. Schneider, B., S. Neidle, and H. M. Berman. 1997. Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers.* 42:113–124.

44. Rahmouni, A. R., and R. D. Wells. 1989. Stabilization of Z DNA in vivo by localized supercoiling. *Science.* 246:358–363.