Accurate Prediction of B-form/A-form DNA Conformation from Primary Sequence: A Machine Learning and Free energy Handshake

Abhijit Gupta[1], Mandar Kulkarni[2] and Arnab Mukherjee[1,*]

[1] Department of Chemistry, Indian Institute of Science Education and Research, Pune, Maharashtra, 411008, India
[2] Currently in the Division of Biophysical Chemistry, Lund University, Chemical Center, P. O. B. 124, 22100 Lund, Sweden

* To whom correspondence should be addressed. Tel: +91-20-25908051; Fax: +91-20-25899790; Email: arnab.mukherjee@iiserpune.ac.in

ABSTRACT

DNA carries the genetic code of life. Different conformations of DNA are associated with various biological functions. Predicting the conformation of DNA from its primary sequence, although desirable, is a challenging problem owing to the polymorphic nature of DNA. Although a few efforts were made in this regard, currently there exists no method that can accurately predict the conformation of right-handed DNA solely from the sequence. In this study, we present a novel approach based on machine learning that predicts A-DNA and B-DNA conformational propensities of a sequence with high accuracy (~95%). In addition, we show that the impact of the dinucleotide steps in determining the conformation agrees qualitatively with the free energy cost for A-DNA formation in water. This method enables us to examine the genomic sequence to understand the prospective biological roles played by the A-form of DNA.

INTRODUCTION

The prediction of a DNA conformation from the mere knowledge of its sequence presents an opportunity to presume its role in specific biological processes. The biological processes, such as direct and indirect readout mechanisms in protein-DNA interactions, exploit the conformational flexibility exhibited by DNA. The reduction in relative humidity around DNA due to the presence of other solvents like ethanol(1) or the presence of protein molecules(2) causes B-DNA to A-DNA transition. The A-DNA conformation is shorter and more compact compared to B-DNA. During B $\rightarrow$ A transition, the phosphate groups protrude out, and minor groove becomes broad and shallow forming more water bridges in accordance with the theory of economy of hydration proposed by Saenger *et al.* (3).

The protein molecules such as transposase, endonuclease, and polymerase interact with B-DNA locally and convert a few dinucleotide steps to A-form in a whole DNA. (2) A-philic DNA segments exhibit low energy cost for deformation, and thus proteins bind to such hotspots during indirect recognition mechanism to commence the transcription process. (2) This mechanism is different from the direct recognition mechanism where protein interacts with a specific nucleotide sequence binding site. The A-form also participates in the protection of bacterial cells under extreme UV exposure. (4)

Whelan and co-workers have shown fully reversible B→A-DNA transition in living bacterial cells on desiccation and rehydration using FTIR spectroscopy. (5) Extremophiles like SIRV2 virus (*Sulfolobus islandicus rod-shaped virus 2*) survives at extreme temperatures of 80°C and acidity of pH 3 by adopting complete DNA in A-form, and aids protein to encapsidate DNA.(6) The motors that drive double-stranded DNA (dsDNA) genomes into viral capsids are among the strongest of all biological motors for which forces have been measured. DNA plays an active role in force generation.

The "scrunchworm hypothesis" holds that the motor proteins repeatedly dehydrate and rehydrate the DNA, which then undergoes cyclic shortening and lengthening motions. The protein components of the motor dehydrate a section of the DNA, converting it from the B- to A-form and shortening it by about 23%. The proteins then rehydrate the DNA, which converts back to the B form. (7)

Thus, it has become clear of late that A-DNA is merely not a non-functional conformation of DNA; it is an essential adaptation of DNA to survive harsh conditions. It is, therefore, intriguing to predict the sequence-structure relationship in DNA. Moreover, an understanding of sequence specificity of B-form → A-form transition and an apriori detection of the A-philic segment in the genome will unveil the possible hotspots of certain biological processes in specific genes of organisms. We have developed a method based on machine learning to realize this apriori prediction of conformational preference of a given DNA sequence towards A-form or B-form with high accuracy. We also relate this conformational preference to the free energy cost of a dinucleotide step to be converted to A-form. We can employ this approach for designing primers that are conformationally biased towards either A or B form and use them for studying their impact in different biological processes.

The polymorphic nature of DNA makes the DNA conformation's prediction a challenging task. The local or partial B-form to A-form transition of a small segment of DNA sequence always possesses the penalty of B-form/A-form junction formation on both 5' and 3' ends of a newly formed A-DNA segment in a whole sequence(8). Considering this aspect, we had previously performed rigorous umbrella sampling simulations to calculate this junction free energy values and characteristic local B-form to A-form free energy values for all ten unique dinucleotide steps. (9) The free energy values obtained therein are termed as "absolute free energy" values (ΔG_a) as they are devoid of any effects from flanking base pairs. We have used these absolute free energy values in our inference model for explaining the effect and relative contribution of each dinucleotide step towards the conformational preference of a DNA sequence.

There are only a few studies that attempted prediction of DNA conformation from its sequence. Basham and co-workers derived A-DNA propensity energy (APE)(10) based on the solvation free energy of trinucleotide steps to determine DNA structural preferences. However, APEs are unavailable for specific trinucleotide steps, thereby making this method inapplicable in general across a genomic DNA sequence. In a different approach, Tolstorukov and co-workers(11) formulated free energy models for all ten unique dinucleotide steps (D-12 model) and 32 individual trinucleotide steps (T-32 model) from experimental data of midpoints in B→A-DNA transition studied earlier by others.(12, 13) The T-32 model was found to be more accurate than the D-12 model. It inherently considers stereochemical

effects present along the B → A transition as it is based on three consecutive DNA base steps. However, the absence of the TAA/TTA free energy values limits the application of this dataset for a DNA structure prediction. Moreover, the accuracy of the above models is limited when applied to our present dataset, as mentioned below.

In our approach, we have focused on the development of a general and more accurate method based on machine learning (ML) approach that considers occurrences of 10 unique dinucleotide steps to predict the conformational preference of a given DNA sequence. We also have presented a method for predicting DNA conformation, albeit with less accuracy, based on the absolute free energy cost for $B \rightarrow A$ form conversion for the ten dinucleotide steps previously developed(9). In an ML-based approach, the inference is drawn based on observation alone. Therefore, although ML methods are suitable for prediction, the molecular origin behind the prediction remains unknown. To address this issue, we have built an explanatory model based on SHAP values(14) for interpreting and explaining our model output. The method based on free energy provides the molecular basis of the prediction based on A-philicity and B-philicity of the dinucleotide steps. Finally, we compare both the approaches to rationalize the prediction of DNA conformational preference.

MATERIAL AND METHODS

We have used both machine learning and free energy (FE)-based methods for the prediction of DNA conformation from the primary nucleotide sequence. We are going to discuss each method separately.

I. Machine learning Approach. First, we describe the steps that were followed in devising the machine learning approach for conformation prediction. Fig. 1 shows the schematic diagram of our approach.

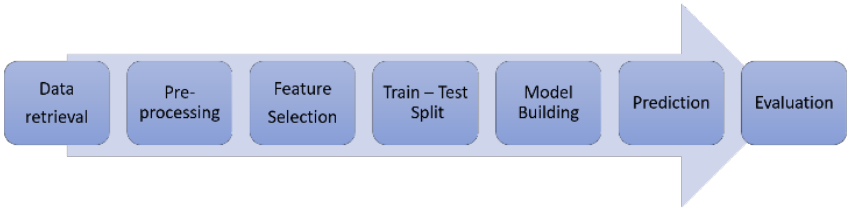

Figure 1: Outline of our machine learning approach for DNA conformation prediction.

(a) Data Collection for designing feature vectors. The first step in an ML approach is data collection. Since we use a supervised learning approach, we collected A- and B-DNA structures from the Nucleic Acid Database (NDB repository) (15, 16). The corresponding sequences were retrieved from RCSB PDB(17) database by a parser we wrote. We filtered out all redundant sequences along with all those sequences which had anything in addition to A, C, G, and T. Further, we have considered only the unbound double-stranded DNA structures. We removed all DNA sequences less than five base pair long from our analysis as they are too short to be deciding a particular conformation. Our curated dataset contained 187 samples, out of which 60 are A DNA sequences and 127 are B DNA sequences.
(b) Pre-processing and adjusting the class imbalance. Data pre-processing involves the transformations being applied to the data before feeding it to our algorithm. Particularly, for DNA

sequences and their respective conformations reported in the database, there was a significant class imbalance (32% A-DNA vs 68% B-DNA curated, non-redundant sequences) that became apparent during the preliminary analysis. To address class imbalance issue, in which training data belonging to one class outnumber the examples in the other, we used SMOTE + TOMEK method(18). We developed a framework that takes a sequence as an object and returns an “ordered dictionary” containing a count of different dinucleotide steps in it. Normalisation and SMOTE+TOMEK were applied in the pre-processing pipeline.

(c) Feature Selection. Feature selection is an important step in any ML approach. The characteristics of any object are called features. Incorrect and irrelevant feature selection may lead to undesirable and even wrong predictions. In a DNA sequence, the relevant features could be the length of the DNA, the number and types of dinucleotide steps, the number and types of tetranucleotide steps, etc. In this study, we have considered the count of all ten unique dinucleotide steps as our feature vectors (see Supporting Information (SI) section A, Supplementary data for detailed description). There are two main reasons: (i) first, the dinucleotide step is the smallest building block of DNA, and (ii) second, the absolute free energy values are to be used for an alternative prediction method is also based on the dinucleotide step.

(d) Splitting the train and test data. Here we split data into train and test sets for further analysis. 80% of the data was used for training, and the remaining 20% was left for testing. The StratifiedShuffleSplit (n_splits=1, test_size=0.2) function in Scikit-learn(19) was used for data splitting. The stratified splitting was employed to handle class imbalance between A and B DNA samples. It maintains the ratio of positive (A-DNA) and negative (B-DNA) cases of the total sample in train and test sets.

(e) Model Building. In this stage, machine learning models were selected for training. All classifiers in scikit-learn(19) use a fit (X, y) method to fit the model for a given train data X and train label y. We tested LightGBM(20)(Figure 3), XGBoost(21), SVM with “RBF’ kernel(22), and Logistic Regression(22) (Figure S1-S4, Supplementary data). For tuning hyperparameters of our models we used Optuna framework(23) for LightGBM and XGBoost and Randomized search with cross validation for other models. After testing a host of machine learning algorithms, we decided to use LightGBM as it gave the best compromise between accuracy and interpretability. A benefit of using gradient boosting is that after the boosted trees are constructed, it is relatively straightforward to retrieve importance scores for each attribute.

We have used Intel Distribution for Python and Python API for Intel Data Analytics Acceleration Library (Intel DAAL) - named PyDAAL(24)—to boost machine-learning and data analytics performance. Using the advantage of optimized scikit-learn (Scikit-learn with Intel DAAL) that comes with it, we were able to achieve faster training time and accurate results for the prediction problem.

(f) Prediction. During this stage, we use our trained model for predicting the output for a given input sequence based on its learning. That is, given an unlabelled observation X (here: DNA sequence), it returns the predicted label y (A-DNA or B-DNA).

(g) Evaluation. For assessment of the performance of our classification model, we have chosen accuracy and AUC (Area Under the Curve) and ROC (Receiver Operating Characteristics) as our primary evaluation metrics. AUC-ROC tells how much a particular model is capable of distinguishing

between different classes (A vs B). ROC represents a probability curve, and AUC represents the measure of separability between the two classes. Higher the AUC score, better the model is at distinguishing between A and B DNA samples. When there is a class imbalance, the accuracy alone cannot give an accurate assessment of the performance of a classification model. A classifier may proclaim all data points as belonging to the majority class and obtain a high accuracy score while performing poorly on the prediction of minority class samples. Owing to class imbalance, we have used additional metrics – precision, recall, f1-score and Matthews correlation coefficient (MCC) score to measure our model's performance on the test set (Table 2 a,b). Precision is defined as the ratio of true positives and true positives plus false positives. False positives are outcomes the model incorrectly labels as positive that are actually negative. In our example, false positives are B-DNA that the model classifies as A-DNA. While recall expresses the number of true positives divided by the number of true positives and the number of false negatives. In most problems pertaining to classification, one could give a higher priority to maximizing precision, or recall, depending upon the problem one is trying to solve. However, in general, there exists a more straightforward metric that takes into consideration both precision and recall. This metric is known as F1-score. It is the harmonic mean of precision and recall. Notably, MCC considers true and false positives and negatives and is generally regarded as a balanced measure that can be used when there is a class imbalance.(25) The formulae of these metrics are mentioned in Supplementary data, Section C.

II. Free Energy Based Approach. In this section, we describe the concept of absolute free energy values briefly (see ref. (9) for further details). Thermodynamically, the conformation of a particular structure depends on the free energetic stability. Therefore, the propensity of a sequence to adopt a particular conformation should depend on the overall free energy of the sequence in that conformation. Keeping that in mind, we had earlier calculated the free energy cost (Table 1) for the formation of A-form of each of the ten dinucleotide steps, as discussed below. (9)

Table 1. List of absolute energy values (ΔG_a) for all 10 possible dinucleotide steps.

Dinucleotide Steps	ΔG_a(kcal/mol)
AA/TT	2.34
GG/CC	0.86
AC/GT	1.91
CA/TG	2.40
AT/AT	2.29
TA/TA	1.59

AG/CT	0.67
GA/TC	0.84
CG/CG	3.06
GC/GC	1.33

* Please note, ΔG_J values were calculated only for homonucleotide steps and not heteronucleotide steps. $\Delta \boldsymbol{G_J}$ is 1.59 kcal/mol for AA/TT and 0.52 kcal/mol for GG/CC.

We used umbrella sampling simulations along a new reaction coordinate Z_p' and average Z_p' ($\overline{Z_p'}$) for 10 unique dinucleotide steps and a few trinucleotide steps embedded in the 13-mer DNA sequence.(8) These sequences, in general, can be presented as d(CGCGXXYYYCGCG)$_2$, where X/Y can be either A, T, C, or G. The presence of CG sequences on both termini reduces the possibility of base pairs fraying at the ends.(15) We showed earlier that creating an A-form in a B-DNA creates two B/A junctions. Therefore, the free energy obtained for the dinucleotide step XY (underlined in 13-mer sequence) from simulation can be written as,

$$\Delta G_{sim}(XY) = \Delta G_j(XX) + \Delta G_a(XY) + \Delta G_j(YY). \quad \textbf{Eq. 1}$$

At this stage, we are only aware of $\Delta G_{sim}(XY)$ value. We then performed simulations on di- and tri-homonucleotide sequences d(CGCGXXXXXCGCG)2 to find the junction and absolute free energy values for homo-dinucleotide steps. The free energy cost to convert XX step along Z_p' in sequence d(CGCGXXXXXCGCG)2 can be decomposed as,

$$\Delta G_{sim}(XX) = \Delta G_j(XX) + \Delta G_a(XX) + \Delta G_j(XX). \quad \textbf{Eq. 2}$$

Also, the free energy cost to convert XXX step in the same sequence d(CGCGXXXXXCGCG)$_2$ using an average Z_p' ($\overline{Z_p'}$) can be decomposed as,

$$\Delta G_{sim}(XXX) = \Delta G_j(XX) + 2\Delta G_a(XX) + \Delta G_j(XX) \quad \textbf{Eq. 3}$$

Subtracting Eq. 2 from Eq. 3, one can obtain absolute free energy value $\Delta G_a(XX)$ (which is devoid of any junction effect) for creating an A-form dinucleotide step within a B-DNA and eventually junction free energy values for homo dinucleotide steps AA, TT, GG, and CC. Table 1 lists these absolute and junction free energies. Using this junction free energy values ($\Delta G_j(XX)$ *or* $\Delta G_j(YY)$) one can calculate these absolute free energy values ($\Delta G_a(XY)$) for the rest of the hetero-dinucleotide steps. These values are also listed in Table 1. Note that, the junction effect comes only when a part of the DNA is converted from B-form to A-form. The full conversion of a B-DNA to A-DNA will depend only on the absolute free energy cost. That is the primary reason to calculate absolute free energy.

With these absolute free energy values, we constructed a model to predict the B- and A-DNA conformations from the sequence. This is similar to the earlier approach of Basham, who used the solvation free energy-based approach for trinucleotide step.(10) However, in Basham's results, all the trinucleotide steps were not considered. In our approach, we use the dinucleotide step and thereby can

take into account all possible sequence variations. Moreover, we believe that this is a direct approach where the free energetic stability dictates the propensity for a particular conformation. However, translating this free energy cost from a dinucleotide step to a full DNA can be accomplished in multiple ways. We adopted a simple approach where we calculated the average free energy cost (ΔG_{avFE}) for the conformational transition of a DNA sequence between B- and A-form defined as,

$$\Delta G_{avFE} = \frac{1}{N}\sum_{i=1}^{N} \Delta G_a(XY),$$

where N is the number of dinucleotide steps in a given sequence. This number is equal to one less than the length of DNA sequence and $\Delta G_a(XY)$ is absolute free energy value for a particular dinucleotide step, where $X, Y = A, C, G, or\ T$.

To create a classifier that can distinguish between A- and B-form of DNA, we calculated ΔG_{avFE} for those sequences in our training set for which both solvation free energy (ΔG_{Basham}(kcal/mol))(10) and absolute free energy values are calculatable.(Table S2, Supplementary data). Using these two features, we created a maximum margin classifier to best separate the A and B DNA samples. We found that the margin corresponding to $\Delta G_{avFE} \sim 1.71$ separates A and B classes (Figure 2). This corresponds to the following classifier that if ΔG_{avFE} is less than 1.71, it will be classified as A-DNA, else as B-DNA. This classification based on the absolute free energy will be referred hereafter as gAvg model.

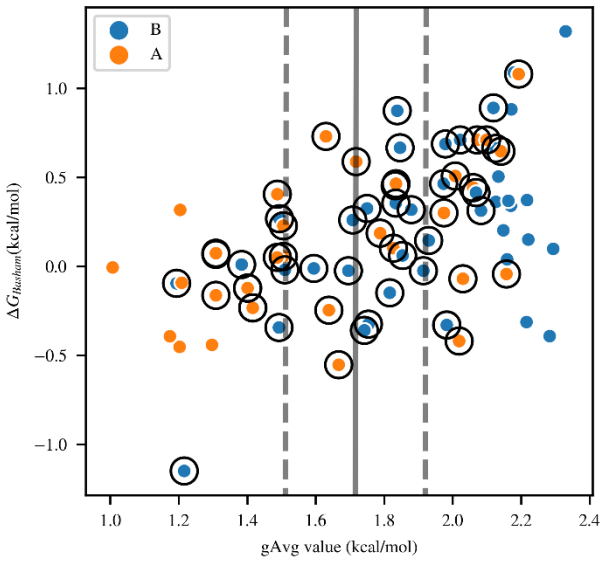

Figure 2: ΔG_{avFE} for sequences in the training set (Table S2). The point corresponding to $\Delta G_{avFE} = 1.71$ was chosen as the margin for classifying A and B DNA sequences. The dotted margins correspond to samples within ± 0.5 of the optimal margin value (1.71).

RESULTS

I. Prediction using absolute free energy cost

We have used the classifier mentioned above to predict whether a particular sequence will be B-DNA or A-DNA. The results of gAvg model are presented in Table 2a. We obtain an accuracy (see Sec C in SI) score of 53% on the test set (Table 2). We also report other parameters for evaluating the prediction such as recall, f1-score, and support as mentioned above (see SI for definition). The gAvg model achieves a moderate score in regard to precision, recall and F1. From the results obtained, it becomes evident that simplistic gAvg model cannot classify B-DNA samples correctly. The low precision score of 0.38 for A-DNA class tells us that this model wrongly classifies B-DNA samples as A-DNA (false positives). Perhaps it could be attributed to use of a rigid margin of 1.71 kcal/mol as $\Delta G_a$for classification. Similarly, we get a low recall score of 0.38 for B-DNA, which indicates that this model wrongly classifies many A-DNA samples as B-DNA, which again, could be attributed to the use of a rigid margin for classification. This gives us an insight that accurate classification of A and B DNA samples would require a more sophisticated model with a greater degree of freedom.

Table 2(a): Classification report for (a) gAvg model. The accuracy score for gAvg model is 0.53 and MCC score is 0.17

conformation	precision	recall	f1-score	support
(A DNA)	0.38	0.75	0.50	12
(B DNA)	0.79	0.42	0.55	26
average/total	0.66	0.53	0.53	38

Table 2(b): Classification report for ML model. The accuracy score for this model is 0.95 with MCC score of 0.89

conformation	precision	recall	f1-score	support
(A DNA)	0.86	1.00	0.92	12
(B DNA)	1.00	0.92	0.96	26
average/total	0.95	0.95	0.95	38

II Machine Learning Approach

We tested the performance of the machine learning approach, referred to as ML model hereafter, by performing 5-fold ($k = 5$) stratified cross-validation test. This approach involves randomly dividing the set of observations into k groups or folds of roughly equal size. Notably, stratification was used to maintain the distribution of A and B DNA samples in each fold, similar to those present in the training set. The first fold was treated as the validation set, and our classifier was fitted to the remaining $k - 1$ folds. The error was then calculated on the observations in the held-out set. We repeated this procedure k times; each time a different group of observations is treated as a validation set. We obtained a mean

ROC AUC score of 0.90 ± 0.02 (Figure 3), MCC score of 0.89 and an accuracy score of 94.7% on the test set (Table 2, Table S1(supplementary data)). In comparison, the gAvg model achieves MCC score of 0.17, which is significantly less than the ML model. We have shown the confusion matrix(25) (Figure 4(a)) that shows the classification results of our model. It contains the number of true positives, false positives, true negatives, and false negatives(see section C, figure S5 in supplementary data). True positives correspond to the number of accurate A-DNA samples predicted, true negatives correspond to the number of accurate B-DNA samples predicted, false positives correspond to the number of B-DNA samples that were misclassified as A-DNA, and false negatives correspond to the number of A-DNA samples that were misclassified as B-DNA. The model outputs the respective class probabilities for A-DNA and B-DNA samples. To convert them into class labels we chose the threshold given by the point that corresponds to minimum of False positive rate and False negative rate (Figure 4(b)).

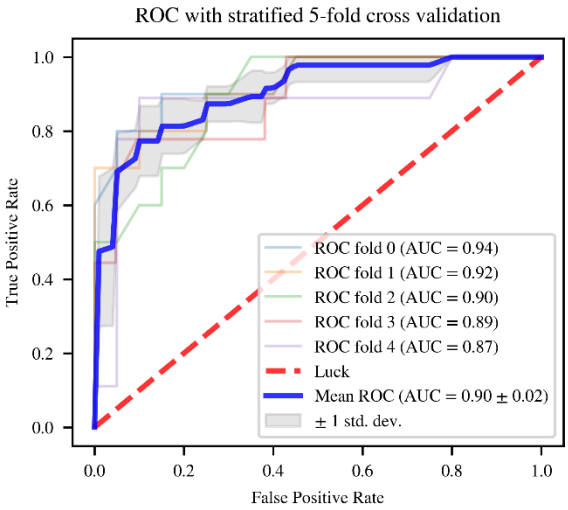

Figure 3: ROC curves for stratified 5-fold cross-validation scores. The mean ROC AUC (Area Under Curve) score for our model is 0.90±0.02.

(a) (b)

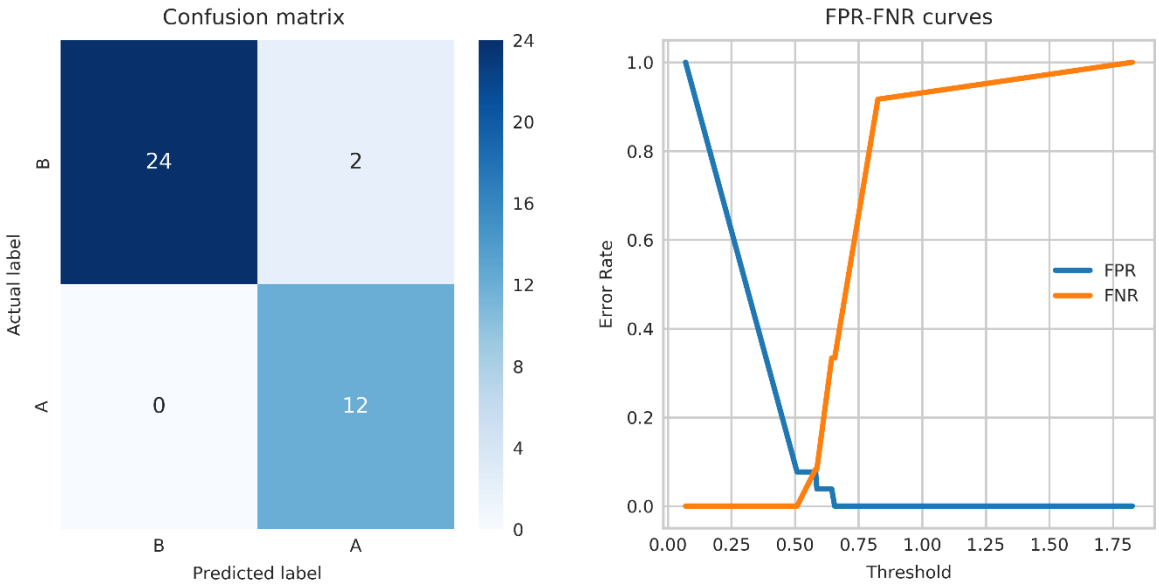

Figure 4: (a) The Confusion Matrix shows the classification results of our model. The overall accuracy of our model is 94.7 %, with MCC score as 0.89. **(b):** For choosing threshold for classification we have used the minimum point of False Positives Rate curve and False Negatives Rate curves at different thresholds.

To understand how individual dinucleotide steps affect the propensity of a sequence to assume a given conformation, we have used SHAP(14) (SHapley Additive exPlanations). SHAP is a unified approach for explaining the output of any machine learning model. It connects game theory with local explanations, uniting several previous methods, and representing the only possible consistent and locally accurate additive feature attribution method based on expectations(14). This explanation model uses simplified inputs, which are toggling features on and off rather than raw inputs to the original model. Figure 5 shows the schematic models of SHAP, where data is processed using the original model and using the SHAP criteria as mentioned above. $g(z')$ is a linear function of binary variables (ON or OFF), which determines the role of individual inputs of features in the prediction. SHAP builds model explanations by asking the same question for every prediction and feature: "How does prediction *i* change when feature *j* is removed from the model?", as mentioned above.

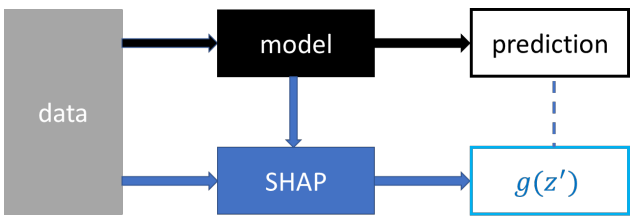

Figure 5: Schematics of SHAP model

To get an idea about which features are most important for our model, we have plotted the SHAP values of every feature for every sample. Figure 6 shows the SHAP summary plot, which sorts features by the sum of SHAP value magnitudes over all samples and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red means high impact, blue means low impact). We see that (AA/TT), a B-promoting dinucleotide step, and CG/CG, a B-promoting dinucleotide step, have the highest impact on our model prediction. The AA/TT step has the highest negative SHAP value, which corresponds to its highest contribution in predicting B-promoting DNA sequence. It is closely followed by CG/CG step. Similarly, the GG/CC and GC/GC have the highest positive SHAP value, which corresponds to their highest contribution in predicting A-promoting DNA sequences. It is interesting to note that there is a strong concordance between these inferences drawn from our ML model with the absolute free energy values (Table 1). We can also take the mean absolute value of the SHAP values for each feature to get a standard bar plot (Figure 7) which shows how each dinucleotide step(feature) contributes in the prediction of the propensity of A/B promoting DNA sequence.

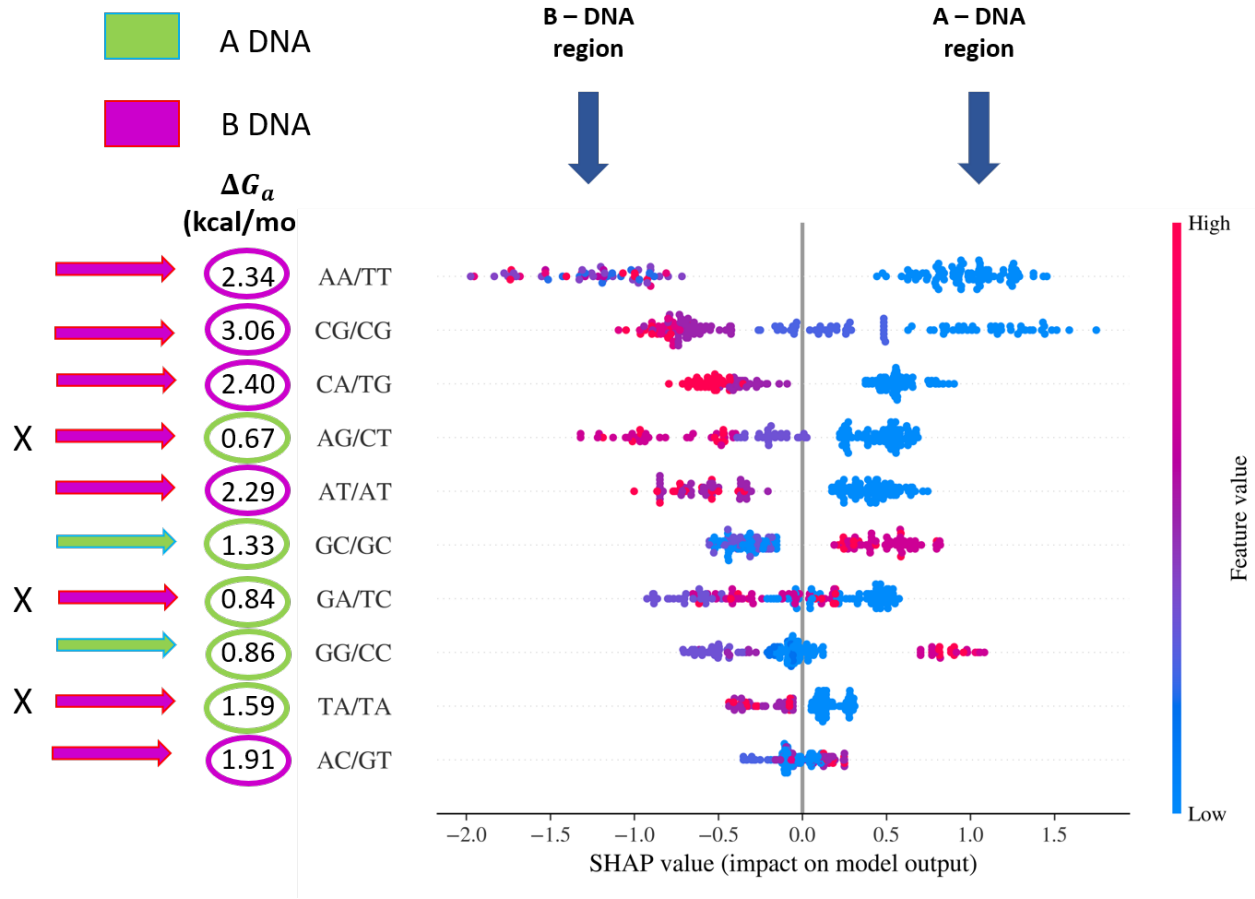

Figure 6: SHAP Summary Plot The plot above sorts features by the sum of SHAP value magnitudes over all samples and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red high, blue low). The crosses indicate disagreement between absolute free energy values and model interpretations.

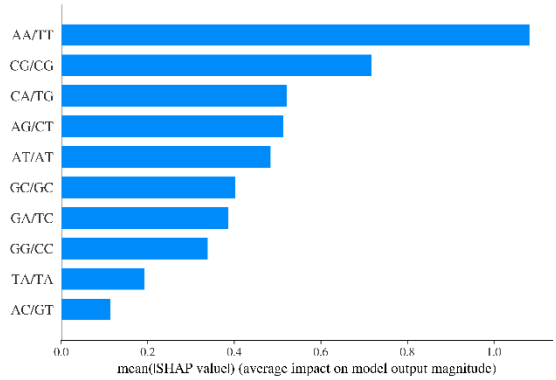

Figure 7: Mean of Absolute SHAP values show the average impact of each dinucleotide step in predicting whether a given sequence will attain A or B conformation.

DISCUSSION

In summary, we have developed two approaches to predict A-DNA or B-DNA conformation of a DNA sequence. The first approach, based on the B-to-A free energy scale of all ten unique dinucleotide steps, predicts conformation with 53 %. In the second approach, we have trained a machine-learning (ML) algorithm using a set of known A-DNA / B-DNA sequences. The ML approach provides better prediction with the correctness of ~ 95 %.

AA steps are highly B-philic due to the steric hindrance of their antisense counterpart TT step. A severe steric hindrance between protruding methyl groups of thymine base exists if it undergoes B→A transition and, thus, enhances the free energetic cost of the process. It is surprising to see that ML models can predict AA step as most B-philic step without the knowledge of the structure and interactions between the stacking base steps. In gAvg model, the most B-philic step is CG step and thus, underestimates B-philicty of AA step.

GG step is well-known to adopt or induce A-form in DNA sequences. The gAvg model predicts AG and GA steps as most A-philic steps, whereas GG as the third most A-philic step. This misplaced A-philicty in gAvg model might be related to the wrong prediction in some instances. Again, it is encouraging to note that the ML model can predict GG and GC as most A-philic steps without any structural information.

We have applied our method to 38 DNA sequences listed in NDB dataset (12 A-DNA and 26 B-DNA sequences) to predict A-DNA or B-DNA conformational preference and observed ~95% accuracy. Understanding why a model makes a specific prediction can be as crucial as the prediction's accuracy in many applications. It is crucial when we want to understand how each fundamental dinucleotide step contributes towards the conformation attained by a sequence. The highest accuracy for large modern datasets is often achieved by complex models that are difficult to interpret, such as ensemble or deep

learning models, creating a conflict between accuracy and interpretability. We have used LightGBM(20), an implementation of gradient boosting decision tree technique, which offers a balanced tradeoff between accuracy and interpretability, to address this problem. For gaining further insight into the interpretability of our model, SHAP analysis was employed with which we could come up with a consistent and locally accurate additive feature attribution method based on expectations. This study thus indicates that the conformational preference of a DNA lies in the fundamental free energetic driving force at a local dinucleotide level. Most of the DNA sequences used here, however, are short. Therefore, the cooperative effect may play a role in the case of longer DNA sequences, and an effort is underway to understand this.

At the moment, we are restricted by the paucity of a sufficient number of labelled DNA sequences. Out of 187 curated DNA sequences in the NDB dataset, 60 are A DNA sequences and 127 are B DNA sequences (dataset S1, S2, Supplementary data). Lack of enough data is one of the major challenges in any machine learning model. Furthermore, the severe class imbalance between A and B DNA is another limitation.

We believe that the proposed model can be implemented on other genomes to find unknown A-DNA promoter elements *a priori* and further study is underway to understand its application to eukaryotic genome analysis as well as to the genome of organisms that survive under stringent conditions using A-form of DNA.

DATA AVAILABILITY

DNA structure prediction from its sequence code available in the BitBucket repository (https://bitbucket.org/abhijit038/dna_structure_prediction_ml/). We intend to build a webserver for our program soon, where the user can provide raw sequences as the input and get the probabilities for them to attain A/B form conformation.

SUPPLEMENTARY DATA

Supplementary Data is available online as NAR_SI pdf file.

ACKNOWLEDGEMENT

We thank Dr. Leelavati Narlikar, National Chemical Laboratory, India for extensive discussions.

FUNDING

Department of Science and Technology (DST), Science and Engineering Board (SERB), Govt. of India (Grant EMR/2016/001069). It is also partially supported by Department of Biotechnology, India (BT/PR34215/AI/133/22/2019).

CONFLICT OF INTEREST

We declare no conflict of interest.

REFERENCES

1. Franklin,R.E. and Gosling,R.G. (1953) Molecular Configuration in Sodium Thymonucleate. *Nature*, **171**, 740–741.

2. Lu,X.-J., Shakked,Z. and Olson,W.K. (2000) A-form Conformational Motifs in Ligand-bound DNA Structures. *Journal of Molecular Biology*, **300**, 819–840.

3. Saenger,W., Hunter,W.N. and Kennard,O. (1986) DNA conformation is determined by economics in the hydration of phosphate groups. *Nature*, **324**, 385–388.

4. Mohr,S.C., Sokolov,N.V., He,C.M. and Setlow,P. (1991) Binding of small acid-soluble spore proteins from Bacillus subtilis changes the conformation of DNA from B to A. *Proceedings of the National Academy of Sciences*, **88**, 77–81.

5. Whelan,D.R., Hiscox,T.J., Rood,J.I., Bambery,K.R., McNaughton,D. and Wood,B.R. (2014) Detection of an en masse and reversible B- to A-DNA conformational transition in prokaryotes in response to desiccation. *Journal of The Royal Society Interface*, **11**, 20140454.

6. DiMaio,F., Yu,X., Rensen,E., Krupovic,M., Prangishvili,D. and Egelman,E.H. (2015) A virus that infects a hyperthermophile encapsidates A-form DNA. *Science*, **348**, 914–917.

7. Harvey,S.C. (2015) The scrunchworm hypothesis: Transitions between A-DNA and B-DNA provide the driving force for genome packaging in double-stranded DNA bacteriophages. *Journal of Structural Biology*, **189**, 1–8.

8. Jacobo-Molina,A., Ding,J., Nanni,R.G., Clark,A.D., Lu,X., Tantillo,C., Williams,R.L., Kamer,G., Ferris,A.L., Clark,P., *et al.* (1993) Crystal Structure of Human Immunodeficiency Virus Type 1 Reverse Transcriptase Complexed with Double-Stranded DNA at 3.0 Å Resolution Shows Bent DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 6320–6324.

9. Kulkarni,M. and Mukherjee,A. (2016) Computational Approach to Explore the B/A Junction Free Energy in DNA. *ChemPhysChem*, **17**, 147–154.

10. Basham,B., Schroth,G.P. and Ho,P.S. (1995) An A-DNA Triplet Code: Thermodynamic Rules for Predicting A- and B-DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 6464–6468.

11. Tolstorukov,M.Y., Ivanov,V.I., Malenkov,G.G., Jernigan,R.L. and Zhurkin,V.B. (2001) Sequence-Dependent B↔A Transition in DNA Evaluated with Dimeric and Trimeric Scales. *Biophysical Journal*, **81**, 3409–3421.

12. Minchenkova,L.E., Schyolkina,A.K., Chernov,B.K. and Ivanov,V.I. (1986) CC/GG Contacts Facilitate the B to A Transition of DMA in Solution. *Journal of Biomolecular Structure and Dynamics*, **4**, 463–476.

13. Ivanov,V.I., Minchenkova,L.E., Minyat,E.E. and Schyolkina,A.K. (1983) Cooperative Transitions in DNA with No Separation of Strands. *Cold Spring Harb Symp Quant Biol*, **47**, 243–250.

14. Lundberg,S.M. and Lee,S.-I. (2017) A Unified Approach to Interpreting Model Predictions. In Guyon,I., Luxburg,U.V., Bengio,S., Wallach,H., Fergus,R., Vishwanathan,S., Garnett,R. (eds), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 4765–4774.

15. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J*, **63**, 751–759.

16. Coimbatore Narayanan,B., Westbrook,J., Ghosh,S., Petrov,A.I., Sweeney,B., Zirbel,C.L., Leontis,N.B. and Berman,H.M. (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res*, **42**, D114–D122.

17. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z., *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res*, **33**, D233–D237.

18. Batista,G.E.A.P.A., Prati,R.C. and Monard,M.C. (2004) A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explor. Newsl.*, **6**, 20–29.

19. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

20. Ke,G., Meng,Q., Finley,T., Wang,T., Chen,W., Ma,W., Ye,Q. and Liu,T.-Y. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon,I., Luxburg,U.V., Bengio,S., Wallach,H., Fergus,R., Vishwanathan,S., Garnett,R. (eds), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 3146–3154.

21. Chen,T. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, New York, NY, USA, pp. 785–794.

22. Bishop,C.M. (2006) Pattern recognition and machine learning Springer, New York, NY.

23. Akiba,T., Sano,S., Yanase,T., Ohta,T. and Koyama,M. (2019) Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19. Association for Computing Machinery, New York, NY, USA, pp. 2623–2631.

24. Malakhov,A., Liu,D., Gorshkov,A. and Wilmarth,T. (2018) Composable Multi-Threading and Multi-Processing for Numeric Libraries. In. Austin, Texas, pp. 18–24.

25. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A.F. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

TABLE AND FIGURES LEGENDS

Figure 4(b): FPR: False Positives Rate, FNR: False Negatives Rate

Figure 6, 7: SHAP: Shapely Additive explanations