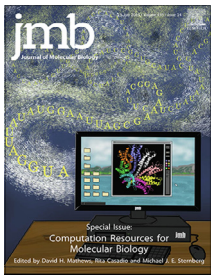
cgNA+web: a visual interface to the cgNA+ sequence-dependent statistical mechanics model of double-stranded nucleic acids

Sharma, R.; Patelli, A.S.; Bruin, L. de; Maddocks, J.H.



cgNA+web : A Visual Interface to the *cgNA+* Sequence-dependent Statistical Mechanics Model of Double-stranded Nucleic Acids

Rahul Sharma [1], Alessandro S. Patelli [1], Lennart De Bruin [2] and John H. Maddocks [1]*

*1 - **Laboratory for Computation and Visualisation in Mathematics and Mechanics,** Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland*
*2 - **Institute Lorentz for Theoretical Physics,** Leiden University, Leiden, The Netherlands*

***Correspondence to John H. Maddocks:** john.maddocks@epfl.ch (J.H. Maddocks)*
https://doi.org/10.1016/j.jmb.2023.167978
Edited by David Mathews

Abstract

The sequence-dependent statistical mechanics of double-stranded nucleic acid, or dsNA, is believed to be essential in its biological functions. In turn, the equilibrium statistical mechanics behaviour of dsNA depends strongly both on sequence-dependent perturbations in its ground state shape away from an idealised, uniform, double-helical configuration, and on its fluctuations as governed by its sequence-dependent stiffness. We here describe the *cgNA+web* browser-based interactive tool for visualising the sequence-dependent ground states of dsNA fragments of arbitrary sequences, as predicted by the underlying *cgNA+* coarse-grain model. Parameter sets are provided to model dsDNA, including the possibility of epigenetically modified *CpG* dinucleotide steps, dsRNA, and DNA:RNA Hybrid double helical fragments. The *cgNA+web* interface is specifically designed to compare ground state shapes of different sequences of the same dsNA, or analogous sequences of different dsNAs. The *cgNA+web* server is freely available at cgDNAweb.epfl.ch without any login requirement.

Introduction

The sequence-dependent mechanics of double-stranded nucleic acids (or dsNAs) is widely believed to be biologically important at length scales of tens of base pairs (bps) to several hundreds of bps, e.g. transcription factor protein binding sites, [1] phased A-tract sequences, [2] nucleosome positioning sequences, [3,4] effects of methylation or not within *CpG* islands, [5–7] physical consequences of SNPs, [8] gene regulation via RNA interference, [9] and the role of DNA:RNA Hybrids (DRHs) in DNA repair [10].

It is known that sequence dependence can strongly affect the configuration space equilibrium distribution of dsNA fragments interacting with a surrounding solvent. Such statistical mechanical properties of dsNA can be probed by a wide variety of experimental techniques, e.g. nuclear magnetic resonance, [11] minicircle cyclization, [12–15] X-ray crystallography [16] and a variety of atomic, or electron force, microscopy techniques [17–19]. Alternatively, the statistical mechanical equilibrium distribution for dsNAs can also be explored by appropriate time averaging over trajectories of fully atomistic molecular dynamics (or MD) simulations, such as the extensive simulations carried out by the Ascona B-DNA Consortium (or ABC) for dsDNA [20]. Similar MD investigations have also been performed for epigenetically base-modified dsDNA, [21,22] dsRNA, [23] and DRH [24,25]. However, most of these studies are done for a minimal number (often as few as 5 to 20) of short sequences of only a few tens of bp in length. Both of these limitations call into question the immediate applicability of direct MD simulation to the full scope

of the molecular biology of dsNAs, especially because it is now well-established that the physical properties of dsNAs are highly sequence-dependent, often with significant nonlocal sequence dependence [26–28,20,29–32].

While available computational power for atomistic MD simulation continues to increase, at decreasing cost, the vast sequence space of dsNA still means that for the foreseeable future direct MD simulation will only be able to address comparatively few cases of sequence variation, and primarily only at comparatively short length scales, say less than one hundred bp for a linear fragment. For this reason, there is considerable interest in developing sequence-dependent coarse-grain dsNA models of various kinds. The biological importance of such coarse grain models is attested to by the significant number of such models that have been developed. Nine works published before 2018 are detailed in, [33] which list can now be supplemented by Liebl et al., [34] Assenza et al., [35] and Walther et al. [36] There are also associated web servers, e.g. NAflex, [37] web3DNA, [38] oxDNA, [39] ThreaDNA, [40] and DNAshape [41]. A more detailed, relatively recent, survey of the field can be found, for example, in [42].

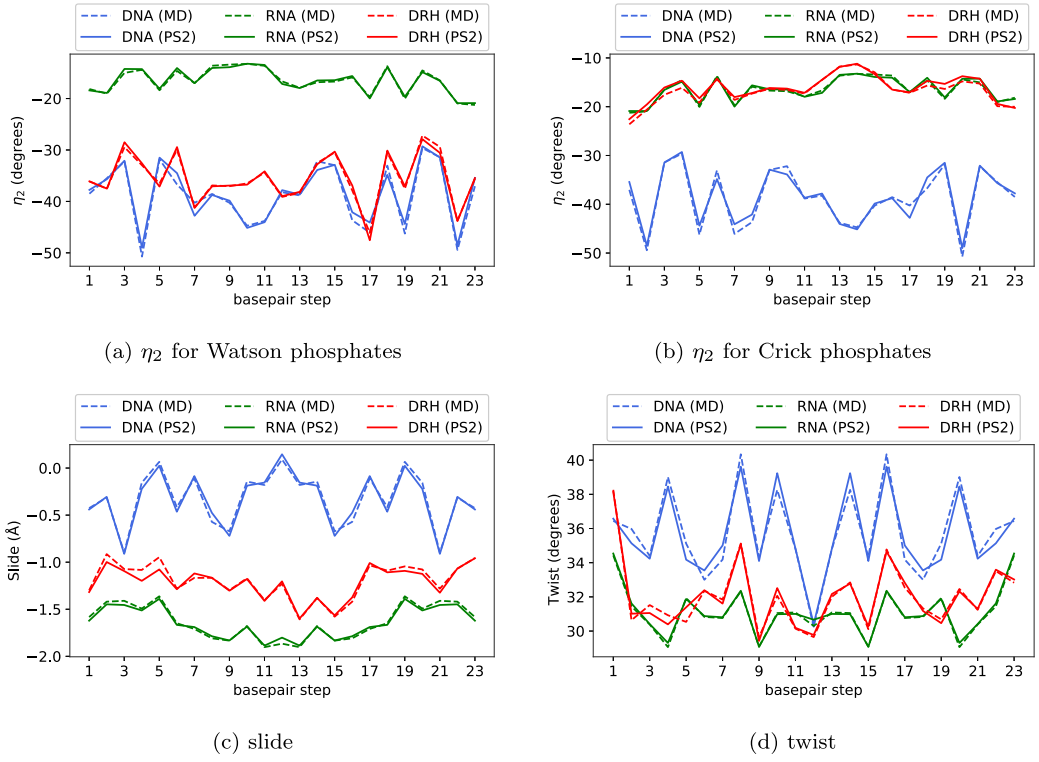
The goal of coarse grain models can be to achieve longer (model) simulation times for longer fragments or to directly predict equilibrium (or infinite time) probability distributions. The *cgDNA* family of models (https://lcvmwww.epfl.ch/research/cgDNA/ for further detail) are of this second type. Specific versions of *cgDNA* models have been implemented to address multiple, specific biological questions, e.g. sequence-dependent persistence lengths [43,31,32] of dsNAs, sequence-dependent unwrapping pathways [44] and energies [45] of dsDNA from the nucleosome core particle, the role of histone tails in nucleosome stability, [46] sequence-dependence of groove widths, [31,32] and to compute sequence-dependent shapes of DNA minicircles [47]. There is also a version *cgNA+loc* [48] where the marginal equilibrium distribution for a sub-sequence is constructed, which allows entire genomes to be scanned with a sliding window in order to identify physically exceptional sequence fragments, and thereby establish a link between statistical mechanics and bio-informatics.

In this article, we describe an evolution of the prior *cgDNAweb* [33] visual interface to the original *cgDNA* model [29,49] to the present *cgNA+web* interface to the updated *cgNA+* model [31,32]. The *cgNA+* model is a refinement of the original *cgDNA* model in two directions. First, [31] the level of coarse graining in *cgNA+* has an explicit description of configurations of each base (as in the original *cgDNA* model) plus an explicit description of the configurations of all of the phosphate groups, each approximated as a rigid body. This finer level of description was shown to be substantially more accurate even for the distributions (or probability density functions or pdfs) of the base configuration coordinates common to both models, see the *cgNA+* model panel on the webpage. Second, [32] parameter sets have been estimated for dsDNA sequences in extended alphabets, including epigenetically modified (methylated and hydroxymethylated) *CpG* steps, for dsRNA, and for DRH.

The bane of coarse-grain descriptions is estimating a sufficiently accurate model parameter set. Both *cgDNA* and *cgNA+* parameter sets are trained on statistics taken from large-scale MD simulations (using state-of-the-art protocols for fully atomistic MD, i.e. explicit solvent with counter ions) of a sequence training library comprising a relatively small number of short dsNA fragments. Some further details are given in the SI Section 4. For numerous test (*not* merely training library) sequences *cgNA+* predictions of ground state shapes have been verified [31,32] as having a negligible error when compared to statistics drawn directly from MD simulations, including the prediction of nonlocal sequence-dependence of ground states. Here negligible error implies that the difference between *cgNA+* model prediction and MD observed statistics is very small compared to the variation of the signal with sequence. Some typical data of this type are provided in Figure 1 and SI Figure 1.

The estimation of a full *cgNA+* model parameter set starting from scratch remains a quite computationally intensive task, both due to the number of multi-microsecond simulations required, and because a full *cgNA+* parameter set contains at least 21 K or more scalars to be fit (so that *cgNA+* has the distinctive flavour of a machine learning model). But there are now a number of pre-computed parameter sets available, and once a parameter set is known and fixed, using the *cgNA+* model is not at all computationally intensive. Currently, five parameter sets are provided within the *cgNA+web* interface, with the associated MD protocol description in the online *cgNA+web* documentation pane. As further new parameter sets become available, they will similarly be described there.

The *cgNA+* model was first implemented in Matlab/Octave and Python scripts that are publicly and freely available at https://github.com/rahul2512/cgNA_plus. These *cgNA+* scripts can be straightforwardly applied to ensembles of millions of sequences, each of several hundred bp lengths. And for intensive users, we would recommend investing the time to become familiar with one or other of these scripted implementations. In contrast the *cgNA+web* interface presented here is designed to allow a simple, fast and interactive access to visualisations of the predicted sequence-dependent ground state (i.e. the average or



(a) η_2 for Watson phosphates

(b) η_2 for Crick phosphates

(c) slide

(d) twist

Figure 1. On the ground states of a 24mer sequence *GCATTACGCTCCGGAGCGTAATGC*, with versions for dsDNA (blue), dsRNA (green) and DRH (red), plots of (a) η_2 rotation coordinate for Watson-strand phosphates, (b) η_2 for Crick-strand phosphates, (c) slide, and (d) twist, with solid lines being the *cgNA+* model predictions, and dashed line data being taken directly from atomistic MD simulation. In all cases the error between MD observations and *cgNA +* predictions is negligible compared to variation with sequence. (Data for other coordinates and other sequences entirely analogous.)

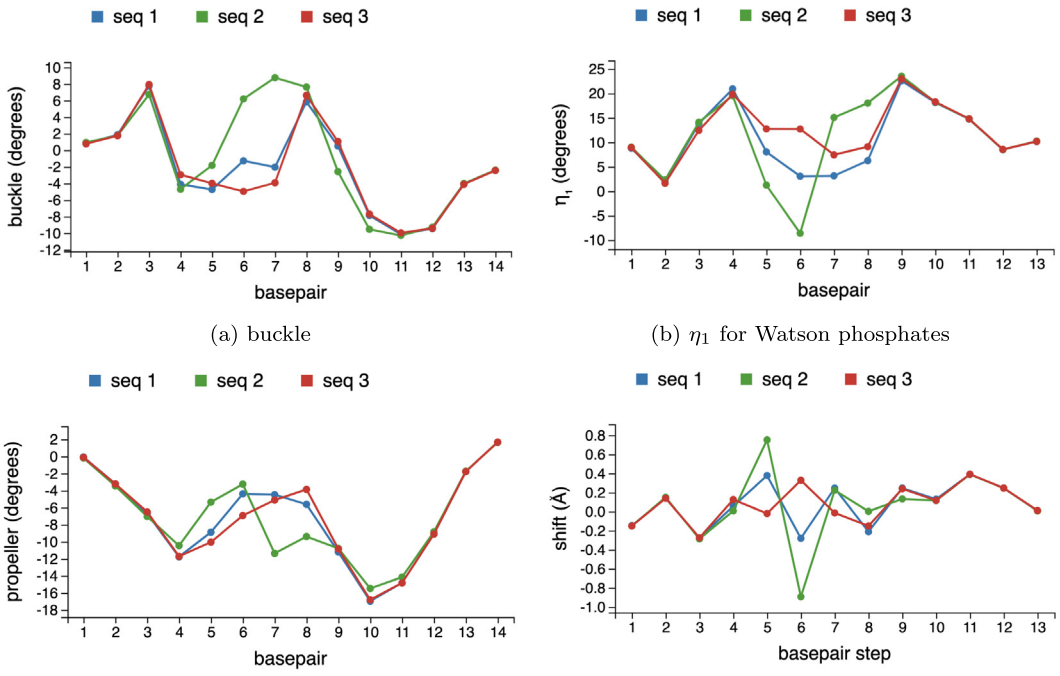
expected configuration) of a comparatively small number of dsNA fragments. The input to *cgNA +web* is a dsNA sequence (dsDNA in standard or epigenetically extended alphabets, dsRNA, or DRH) of length 10 bps to 3 K bps. Visual output is a computed, sequence-dependent, ground state (or minimum free energy) shape, provided as both 2D plots of coarse-grain coordinates along the ground state configuration (specifically an enhanced version of the standard Curves+ [50] internal coordinates), and as 3D visualisations of the sequence-dependent, approximately double helical, ground state structures, at various resolutions (and also as downloadable, standard-format, data files if desired). Up to four different sequences can be visualised simultaneously, with any sequence easily replaced by another choice interactively. *cgNA+web* is particularly useful to compare ground states of different sequences of the same type of dsNA (e.g. SNPs, or the effects of methylating or hemi-methylating *CpG* dinucleotide steps in dsDNA, cf. Figure 2 for 2D comparisons and Figure 3 panel a) for 3D visualisations), or to compare dsDNA, dsRNA and DRH ground states for directly analogous sequences, cf. Figure 1 for 2D comparisons of this type, and Figure 3 panels b)- e) for 3D visualisations.

The cgNA+ Model

cgNA+web can be used as a black box, but it is better to understand the key features of the underlying *cgNA+* model. (Full detail can be found in ref. [31, 32]) *cgNA+* is a predictive model for dsNAs that given a sequence $\mathcal{S}$ (along a designated reading strand) of length n bps, and a parameter set $\mathcal{P}$ provides a Gaussian, or multi-variate normal, pdf on configuration space coordinates $z \in \mathbb{R}^{24n-18}$ in the form:

$$\rho(z;\mathcal{S},\mathcal{P}) = \frac{1}{Z} exp\left\{-\frac{1}{2}(z-\mu)\cdot\mathcal{K}(z-\mu)\right\}. \quad (1)$$

The dimension of z is $24n-18$ because 24 corresponds to 4 sets of 6 coordinates, with each set of 6 describing the three rotations and three translations of the relative displacement between two rigid bodies. In the *cgNA+* model each interior (i.e. neither the first nor last) base pair level is approximated by four rigid bodies, specifically two (non-rigid) nucleotides each made up of a (rigid)



(a) buckle (b) η_1 for Watson phosphates

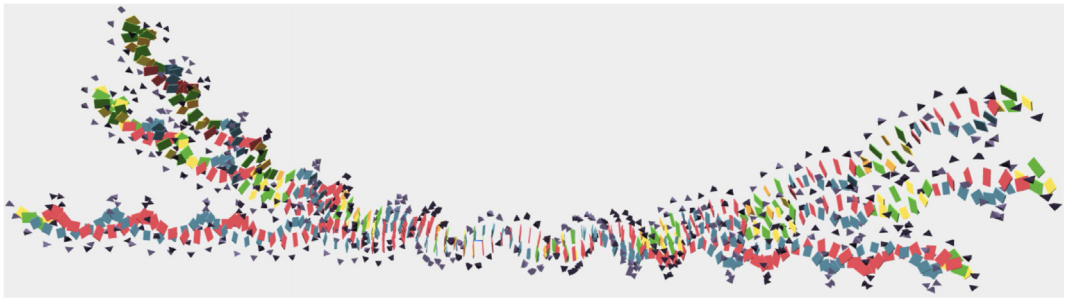
(c) propeller (d) shift

Figure 2. Plots of (a) buckle, (b) η_1 for Watson phosphates, (c) propeller, and (d) shift along the ground states of three dsDNA 14mer sequences differing in only the central base composition *GCGGTG[C/A/M]GCTTTGC* for reference *C* (seq. 1, blue), for point mutation *A* (seq. 2, green) and asymmetric methylation *M* (seq. 3 red). Changes in the ground state are nonlocal to the altered base pair at position 7.

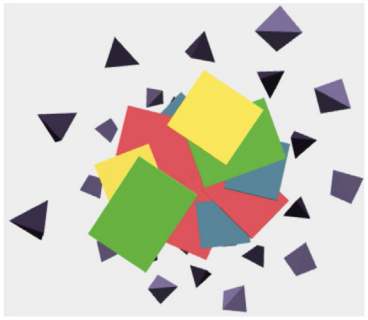
base and a (rigid) phosphate group (but the two 5′ phosphate groups are omitted from the first and last base pair levels). (The sugars are of course explicitly present in any MD simulation, but they are only implicit in the *cgNA+* coarse graining.) Each additional interior dinucleotide step, or base pair junction, between adjacent base pair levels, adds 24 additional coarse grain coordinates; twelve familiar and standard coordinates namely (a Curves+ [50] implementation of the Tsukuba convention [51] for) six intra base pair coordinates (buckle, propeller, opening, shear, stretch, stagger) and six inter base pair (or junction) coordinates (tilt, roll, twist, shift, slide, rise), plus two, less familiar, analogous sets of relative rotations and translations $(\eta_1, \eta_2, \eta_3, w_1, w_2, w_3)$ serving to locate each phosphate group, one in each backbone, with respect to the base within its nucleotide. Schematics of these coordinates are provided in the *cgNA+* model panel on the webpage. The *cgNA+* units of translations are Angstroms and for rotations fifth radians (or approximately 11.5 degrees). This nonstandard choice of unit for rotations is for reasons of good numerical scaling internal to the model, but for the sake of familiarity *cgNA+web* can also output rotational coordinates converted to degrees.

The core of the *cgNA+* model is a simple algorithm that given an input sequence $\mathcal{S}$ and parameter set $\mathcal{P}$ outputs the symmetric, positive-definite, stiffness (or inverse covariance or precision matrix) $\mathcal{K}(\mathcal{S},\mathcal{P}) \in \mathbb{R}^{24n-18\times 24n-18}$ and ground (or intrinsic or minimum energy) state configuration $\mu(\mathcal{S},\mathcal{P}) \in \mathbb{R}^{24n-18}$, which together define the quadratic (free) energy appearing in the argument of the exponential in the pdf equation (1). (Boltzmann's constant is absorbed into the entries of the stiffness matrix $\mathcal{K}(\mathcal{S},\mathcal{P})$ which are fit to MD simulations at a known constant temperature, and $\mathrm{Z}(\mathcal{S},\mathcal{P})$ is just the standard normalisation constant.).

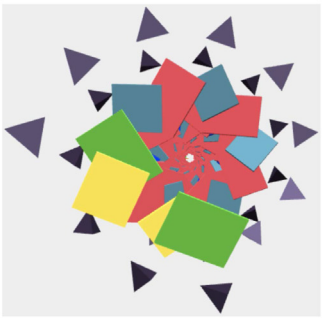
The two basic tenets of the *cgNA+* model are firstly that the total fragment free energy for a dsNA with *n* base pair levels is the sum over junctions of $(n-1)$ localised quadratic free energies in dimension 42, each representing all nearest-neighbour interactions between the eight rigid bodies making up the four nucleotides on either side of the *i*th junction. And secondly that the coefficients in the *i*th junction localised energy depend only on the flanking dinucleotide step in the sequence along the reading strand, and not on the actual value of *i*, i.e. not on the location of the junction within the fragment. The only exception to this last assumption is the first and last junction energies that have dinucleotide step sequence dependence that is different from the localised junction energy for the same dinucleotide sequence step arising at an interior junction. Both of these central assumptions can be checked against statistics taken directly from MD



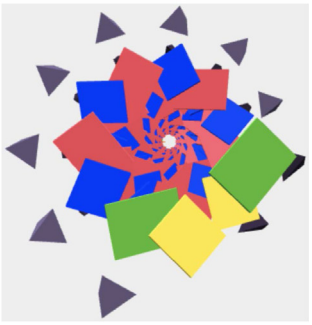
(a) Intrinsically bent phased A-tracts (unmethylated and methylated) and intrinsically straight poly(T)



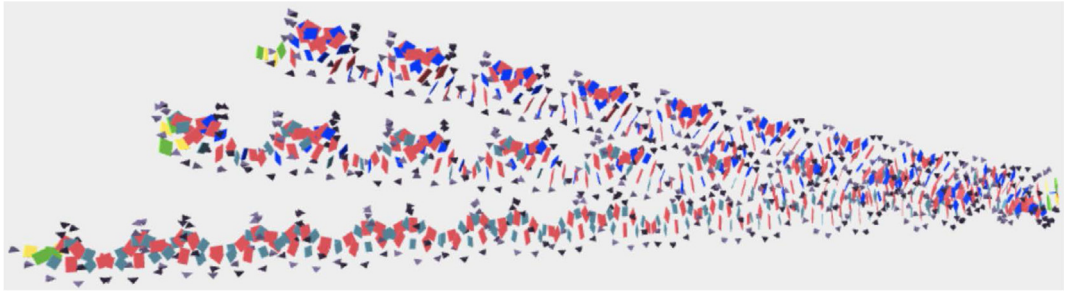
(b) Top view poly(AT) dsDNA



(c) Top view poly(AT) DRH



(d) Top view poly(AU) dsRNA



(e) 3D structure of poly(AT)

Figure 3. 3D visualisations of the ground state configurations of (a) two phased A-tract sequences in standard *GC* *(T_6GCGCG)$_9$GC*, and methylated *GC(T_6GMNMN)$_9$GC)* versions, compared with the intrinsically straight *GC*poly (*T*)*GC* of same length, all aligned at their central base pairs. Both A-tract sequences are highly curved, with the methylated version more bent. Top-views of (b) *GC*poly*(AT)GC* dsDNA, (c) *GC*poly*(AT)GC* DRH, (d) *GC*poly*(AU)GC* dsRNA. The radius of the B-form like helical structure for dsDNA is less than that of the A-like dsRNA structure, with DRH intermediate. (e) For the same three sequences, side view (with all of the first base pair frames aligned). The embedded length of the dsDNA structure (bottom) is longer than that of DRH (middle), which is longer than that of dsRNA (top). Interactive visualisation with *cgNA+web* recommended.

simulation of dsNA fragments, and they both hold to a remarkably high accuracy. The analogous two tenets were already assumed in the precursor *cgDNA* rigid base family of models, but the assumptions are a significantly better fit to observed MD simulation statistics in the finer grain *cgNA+* context, with its additional explicit phosphate degrees of freedom.

With these two simplifying approximations, the algebra inherent in summing localised junction quadratic energies implies that the stiffness matrix $\mathcal{K}(\mathcal{S},\mathcal{P})$ is banded with overlapping diagonal blocks and localised sequence dependence of the blocks. However in the algorithm for computing the ground state $\mu(\mathcal{S},\mathcal{P})$ a completion of squares step arises that involves the inverse stiffness (or covariance) matrix $\mathcal{K}^{-1}(\mathcal{S},\mathcal{P})$, which is dense with nonlocal sequence dependence of its entries. This in turn implies that the components of the ground state $\mu(\mathcal{S},\mathcal{P})$ also have nonlocal sequence dependence. Remarkably the nonlocality arising from this linear algebra within the coarse grain model closely matches the nonlocal sequence-dependence of ground states observed directly in MD simulation. The explanation is that the nonlocality of the ground state arises due to the

physical phenomenon of *frustration*. Because each base pair level comprises two deformable nucleotides, in order to minimise the total free energy of the fragment, the oligomer ground state configuration of the four rigid bodies (i.e. the two phosphate groups and two bases) comprising base pair level i has to be a compromise between the two base pair level configurations minimising the two local energies arising in nearest-neighbour interactions across junctions $(i-1,i)$ and $(i,i+1)$, and this negotiation propagates along the dsNA chain of dinucleotides.

Specific *cgNA+* parameter sets have been estimated for dsDNA (in both standard and epigenetically modified alphabets), for dsRNA, and for DRH. Each parameter set is found by fitting to an appropriate library of MD simulations of the dsNA at hand. The specific MD protocol captures, for example, the species and concentration of counter ions in the solvent, and also depends on the specific MD force field, water model, etc. A full description can be found in ref. [31, 32], but briefly each parameter set is made up of a) interior dinucleotide-dependent $\mathcal{K}^{XY}$ blocks $\in \mathbb{R}^{42\times42}$ plus $\mathcal{K}^{5'XY}$ and $\mathcal{K}^{XY3'}$ end blocks $\in \mathbb{R}^{36\times36}$, and b) interior dinucleotide-dependent stress-like vectors $\sigma^{XY} \in \mathbb{R}^{42}$ plus $\sigma^{5'XY}$ and $\sigma^{XY3'}$ end vectors $\in \mathbb{R}^{36}$. The parameter set estimation step is highly non-trivial. For example, for either dsDNA or dsRNA in their standard four letter alphabets with $X, Y \in \{A, C, G, T/U\}$, there are ten independent sets of 42×42 blocks corresponding to the ten independent dinucleotide sequence steps remaining once the Crick-Watson symmetry of choice of reading strand is considered. In the end, each such parameter set has approximately 21 K independent scalars to be estimated, which can be achieved due to a, somewhat sophisticated, numerical fitting algorithm exploiting the appropriate Fisher information matrix, all applied to the large amount of training MD simulation data (of either dsDNA or dsRNA) that can be generated straightforwardly.

The extension of the dsDNA model to allow C5 methylation or hydroxymethylation of cytosines in *CpG* steps in either hemi (only one strand modified) or fully modified forms, is accomplished by extending the sequence alphabet, with for example *MN* denoting a *CpG* step where both *C*s have been methylated, while *HG* and *CK* denote *CpG* steps where only one *C* has been hydroxymethylated (on respectively the reading and complementary strands). The number of independent parameter blocks grows rapidly with such extended alphabets, so as a practical matter to control the size of the parameter set we exclude some more esoteric sequence possibilities. For example we currently do not allow *CpG* end blocks to be epigenetically modified, and while *MNMN* tetramer sequences are allowed (to encompass fully methylated *CpG* islands), a methylated step cannot be immediately adjacent to a hydroxymethylated step, so tetramer sequences such as *MNHK* are not currently modelled.

For DRH we make the convention that the sequence is read along the DNA strand (as always in the 5′ to 3′ direction). We revert to the standard four letter alphabet, but now there are 16 independent interior XY parameter blocks as there is no Crick-Watson reading strand symmetry. Here we currently limit the size of the parameter set (and the number of necessary training MD simulations) by assuming that all sequence fragments have *GpC* dinucleotide steps at both ends (otherwise there would be an additional 30 independent end blocks in the parameter set leading to approximately 33 K scalar parameters).

cgNA+web

The objective of *cgNA+web* is to allow easy visualisation of ground state configurations of specific dsNA sequences of interest to the user. We believe the interactivity of the visualisation is important. Therefore, we strongly encourage the reader to be using *cgNA+web* while reading this and the next Examples Section (Section ‘Example sequences’).

The header of the *cgNA+* web page contains four independent sequence input forms (with switching between the four instances via moving the yellow button by clicking). The desired sequence can be entered either by copy and paste (including from FASTA files) or by direct entry (according to the syntax restrictions described in detail in the online documentation). As soon as a sequence is entered, a drop-down menu allows the selection of the *cgNA+* parameter set to be used, and the “Go” button then launches the *cgNA+* model reconstruction on the back-end server. We recommend choosing one of the parameter sets DNAPS2, RNAPS2 or DRHPS2 to make reconstructions for the indicated type of dsNA. These three parameter sets were trained on MD simulation libraries with compatible MD protocols. The DNAPS2 parameter set allows input sequences in the epigenetically extended alphabet. The two dsDNA parameter sets DNAPS1 and *cgDNA* are primarily provided for backward compatibility, and neither includes epigenetic modifications.

The input form border turns red if there is an unrecognised input syntax for the selected parameter set, or because of sequence length restrictions. The strict software limits on the length of sequences are that the input sequences must be at least 4 bp and less than 3 K bp long. However we advise the practical limits of between 10 bp and 1 K bp. The lower limit arises because end-effects can be observed to penetrate up to 5 bp from either end. For sequences longer than

1 K bp the 2D plot outputs rarely give interesting information, and the 3D visualisations can become unwieldy to rotate (depending on the graphics capability of the local machine).

There are seven tabs/fields in the *cgNA+web* page immediately below the header. The five to the right contain a basic online documentation (complementing this Section), some suggested example sequences (analogous to the next Section), citation information, an area for downloading predicted raw numerical data files from the server, and a brief discussion of the *cgNA+* model (complementing Section '*cgNA+web*'). The two most important tabs are on the left. They present the output of *cgNA+web* in the form of 3D and 2D visualisation windows. We describe them more in the context of the examples presented in the next Section.

Example Sequences

Our first example is a 24mer sequence *GCATTACGCTCCGGAGCGTAATGC* (not from the training library and containing all dinucleotide sequence steps at least once on the reading strand). The same sequence can be entered into three of the sequence input forms, but with three different parameter sets DNAPS2, RNAPS2 and DRHPS2 selected in the three instances. (Note that for the RNAPS2 parameter set an input of *T* is automatically read as *U*. For DRHPS2 the convention is to give the DNA strand sequence. And as remarked previously to keep the DRHPS2 parameter set reasonably small, the first and last dinucleotide steps must be *GpC*, which they are for this sequence. If they were not, the sequence would be extended by automatically adding additional *GpC* steps at either end.) The 2D pane now has 24 sub-windows, each plotting three superposed instances (in different colours for the different dsNAs) of one of the 24 *cgNA+* coordinates plotted along the three ground states. On hovering the cursor over a data point in the plot, a pop-up window displays the corresponding numerical data and the local context of the input sequences. Strong sequence dependence of all three ground states can be observed, as well as significant differences between the three types of dsNA. Figure 1 reproduces four of the 24 *cgNA+web* panels with the model predictions in solid line style. However in addition there is superposed data in dashed line style, which is the ground states observed directly from the appropriate averages over MD simulation trajectories. The MD data is not available within *cgNA+web*, because the whole point is that the necessary MD simulations are quite computationally intensive, while the necessary computations performed by the web server back-end to make *cgNA+web* predictions are in comparison trivial. For this example we have carried out the MD to show that the *cgNA+* model accurately captures the underlying sequence dependence in the MD ground states for all three types of dsNA, and in particular the error between model prediction and MD observation (solid versus dashed lines) is always significantly smaller than the variation along the sequence, as well as being smaller than the differences between the three types of dsNA. We have seen no sequence example to contradict this claim. More examples are considered in [31,32].

2D plots of *cgNA+* internal coordinates usually provide the most detailed information on the ground states for relatively short sequences such as these ones. Some notable features can be observed for the three different dsNAs. For instance, slide and twist for DRH are in between those of dsDNA and dsRNA. Between dsRNA and dsDNA both the $5'/3'$ (or Crick and Watson) phosphate coordinates η_2 are strikingly different, which might be attributed to the general A- and B-forms of dsRNA and dsDNA, respectively. Whereas the η_2 Watson phosphate rotational coordinate (i.e. the DNA strand of DRH) is close to the pure dsDNA case while η_2 for the Crick phosphates (i.e. the RNA backbone of DRH) remains close to the pure dsRNA values.

Unless your eye is well-trained, our experience is that for short fragments it is hard to perceive significant differences in the 3D interactive visualisations provided under the 3D pane. In this particular case it can be observed that the helical embedding of the dsRNA sequence is significantly shorter than that of the dsDNA, with DRH in between. Clicking on a base or phosphate opens a pop up window that helps to decipher which 3D configuration belongs to which sequence input. There is an arbitrariness in how a 3D configuration is embedded in the visualisation box. The default is that with multiple input sequences the first base pair frame of all sequences are aligned.

As already mentioned *cgNA+web* parameter sets are trained on extensive atomistic MD simulations using state-of-the-art MD simulation protocols. One difference between the DNAPS2 and DNAPS1 parameter sets is that they were trained using different protocols. The second difference is that DNAPS2 parameters encompass epigenetically modified bases. In SI Figure 1, we have plotted the ground state of a standard dsDNA sequence observed in two MD simulations using the two different MD protocols, along with the corresponding *cgNA+web* predictions using the associated parameter sets. It can be seen that for this particular sequence, which was chosen such that the average shape observed in the two MD simulations is noticeably different, the two corresponding *cgNA+web* predictions are almost indistinguishable from their underlying MD estimates. Thus, the coarse grain *cgNA+* model is sufficiently accurate to discern subtle differences between MD simulation protocols.

Another example involves sequences *GCGGTG[C/A/M]GCTTTGC* in three variants where [*C/A/M*] denotes a *C*, an *A*, or a methylated *C* i.e. an *M*. These three sequences can be reconstructed within *cgNA+web* using DNAPS2 in all three cases, and 2D plots of all 24 coordinates become available. Four of the 24 panels are shown in Figure 2. Arguably, the epigenetic methylation of *C* is a smaller chemical change in dsDNA than a point mutation. It can be observed that the change in the ground state due to either point mutation or asymmetric methylation of the *C* is significant and strongly nonlocal with the change due to point mutation larger than the change due to methylation of *C*. Additionally, in SI Figure 2, we have compared some of the internal coordinates for the ground state of *GCGGTG[CG]CTTTGC* with its two variants by symmetrically methylating the highlighted *CpG* step (in square brackets), i.e. *CpG* to *MpN*, and symmetrically hydroxymethylating i.e. *CpG* to *HpK*. On epigenetically modifying the cytosines of the central *CpG* step, there is a significant change in the local structure of the dsDNA, and more importantly, the change is strongly nonlocal (up to three bps both sides beyond the modified *CpG* step). It is interesting to note that the effect of methylation and hydroxymethylation of the *CpG* step on the ground state is very similar.

For our final example, in Figure 3, we have provided examples of 3D structure visualisations for various longer, periodic sequences using *cgNA+web*. For longer sequences the 2D plots are more difficult to understand. For periodic sequences (or tandem repeats) the 2D plots are very close to periodic away from the ends. For example, sequence fragments with repeating A-tracts exhibit a highly exceptional 2D signal. However the passage from 2D internal coordinates to 3D configurations of bases and phosphates is not entirely intuitive, and it is not easy to judge for example whether or not the A-tracts are appropriately phased to produce a large over all bend in the ground state. Accordingly the *cgNA+web* server implements the appropriate nonlinear reconstruction and provides associated interactive visualisations in the 3D tab. Then exceptionally straight, or exceptionally bent, ground states become very evident. And because the phosphate groups are explicitly modelled, the phasing of intrinsic bends with respect to major and minor grooves, as well as the groove widths themselves can be visually observed, at least qualitatively. In Figure 3(a) we have compared the poly(*T*) sequence embedded in *GC* ends (which has an intrinsically straight ground state) with a phased A-tract sequence $GC(T_6GCGCG)_9GC$ (which has an intrinsically bent ground state). The phased A-tract sequence is considered in both unmodified and methylated versions, where we observe that methylation increases the already substantial intrinsic bend. Note that in print we can only provided one viewpoint of the 3D structures projected onto a plane, but the real power of *cgNA+web* lies in the interactive rotations available in the 3D visualisation pane, which we strongly recommend to the user. On clicking a base or a phosphate group, a pop up window describes the object and the dsNA to which it belongs, which facilitates a better comparison of various dsNA fragments whose 3D visualisations overlap. Furthermore, we have compared *GC*poly*(AT)GC* for dsDNA, with *GC*poly*(AU)GC* for dsRNA, and the analogous DRH in Figure 3(b–e). In Figure 3 (b–d), we have shown the top cross-section view of 3D structures for poly*(AT)*/poly*(AU)* to highlight the different helical structures formed by the base pair placement for the three dsNA duplexes. For dsDNA, compatible with its B-form geometry, the base pairs are almost centred over the helical axis, while in contrast for dsRNA, which prefers A-form geometry, the base pairs are displaced away from the central helical axis and are closer to the major groove resulting in a helix with a more open cylindrical core. And DRH is intermediate to the dsDNA and dsRNA structures with an open, but narrower, cylindrical core. Then in Figure 3(d), we have plotted the side-view for the same three structures now aligned at their first base pair frames. All three 3D structures are intrinsically straight with different helical axis and are of different lengths. The overall length of the embedded 3D structures (for the same number of base pairs) is in the order dsDNA $>$ DRH $>$ dsRNA related to their B-form, mixed A-B form, and A-form helical geometries.

Discussion

We have presented *cgNA+web* which provides an easy-to-use visual interface to the coarse grain *cgNA+* model, which can in turn predict sequence-dependent ground state configurations of dsDNA, including with epigenetically modified *CpG* steps, dsRNA and DRH, all with trivial computational effort. We have shown examples which suggest that for all three dsNAs the strongly sequence-dependent predictions made by *cgNA+web* are in effect as accurate as any ground state prediction that could be made by a much more computationally demanding MD simulation. The only restriction is that the *cgNA+* parameter set that is used should be trained on a sequence library built with the same simulation protocol as the MD trajectories whose data you are trying to reproduce. Which MD protocol is in fact the most physically accurate is another question entirely. For example we do not yet have a *cgNA+* parameter set with any divalent counter ion in the solvent, because the accuracy of any non-polarisable MD force field for multi-valent ions is

still a controversial question. We do expect that new *cgNA+* parameter sets will continue to become available as new MD simulation libraries are run, and any new parameter sets will be placed on the *cgNA+ web* site as they become available.

For any input sequence the *cgNA+web* back end has to compute the sequence-dependent banded stiffness matrix as an intermediate step in the computation of the sequence-dependent ground state. These banded stiffness matrices are themselves of considerable interest, and they are available for download as data files, which can then be post-processed as per the user's interests. But the current version of *cgNA+web* does not have any option to visualise these stiffness matrices interactively. One reason is that it is not so evident what is actually interesting to visualise. One possibility is to plot spectra of all eigenvalues, but we judged that to still be too computationally intensive to routinely carry out on the interactive server for possibly rather long sequences. Some static examples of eigenvalue spectra are provided in the SI Figures 3–8 (in the scaling where translation and rotation variables are measured in Å and 1/5 radians, respectively). Another interesting stiffness related observable to consider is sequence-dependent persistence length in various senses as discussed in Mitchel et al., [43] Some static post-processed examples are described in the *cgNA+* model pane on the website. But computing persistence lengths requires a further Monte Carlo simulation, which we again judged to be too demanding to be run interactively on the web server, even though there is a highly efficient C++ code *cgNA+mc* [31,32] available, which exploits the bandedness of *cgNA+* model stiffness matrices.

CRediT author statement

RS: Code implementation including web site modifications, parameter set estimation, including MD simulation and data analysis, writing. **ASP:** Code implementation, model development, early parameter set estimation, including MD simulation and data analysis, writing early draft. **LDB:** Design and implementation of core Web server. **JHM:** Development of model and parameter estimation techniques, design of web site, writing, funding acquisition.

Data and Code availability

The *cgNA+web* code (both front- and back-end) is available at https://github.com/rahul2512/cgNA_plus_web. Scripted versions of the *cgNA+* model are also available as Python and Matlab/Octave packages.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding from the EPFL and the Swiss National Science Foundation, Grant No. 200020_182184.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jmb.2023.167978.

Received 25 November 2022;
Accepted 18 January 2023;
Available online 25 January 2023

Keywords:
dsRNA;
DNA:RNA Hybrid;
dsDNA mechanics;
sequence-dependence;
Coarse-grain modelling

References

1. Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., Honig, B., (2009). The role of DNA shape in protein–DNA recognition. *Nature* **461** (7268), 1248–1253.
2. Haran, T.E., Mohanty, U., (2009). The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.* **42** (1), 41–81.
3. Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., Widom, J., (2006). A genomic code for nucleosome positioning. *Nature* **442** (7104), 772–778.
4. Satchwell, S.C., Drew, H.R., Travers, A.A., (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191** (4), 659–675.
5. Kass, S.U., Pruss, D., Wolffe, A.P., (1997). How does DNA methylation repress transcription? *Trends Genet.* **13** (11), 444–449.
6. Cedar, H., Bergman, Y., (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10** (5), 295–304.
7. Pennings, S., Allan, J., Davey, C.S., (2005). DNA methylation, nucleosome formation and positioning. *Briefings Funct. Genomics* **3** (4), 351–361.
8. Shastry, B.S., (2002). SNP alleles in human disease and evolution. *J. Hum. Genet.* **47** (11), 561–566.
9. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S. E., Mello, C.C., (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391** (6669), 806–811.

10. Ohle, C., Tesorero, R., Schermann, G., Dobrev, N., Sinning, I., Fischer, T., (2016). Transient RNA-DNA hybrids are required for efficient double-strand break repair. *Cell* **167** (4), 1001–1013.
11. Nikolova, E.N., Bascom, G.D., Andricioaei, I., Al-Hashimi, H.M., (2012). Probing sequence-specific DNA flexibility in A-tracts and pyrimidine-purine steps by nuclear magnetic resonance 13C relaxation and molecular dynamics simulations. *Biochemistry* **51** (43), 8654–8664.
12. Manning, R.S., Maddocks, J.H., Kahn, J.D., (1996). A continuum rod model of sequence-dependent DNA structure. *J. Chem. Phys.* **105** (13), 5626–5646.
13. Geggier, S., Vologodskii, A., (2010). Sequence dependence of DNA bending rigidity. *Proc. Nat. Acad. Sci.* **107** (35), 15421–15426.
14. Rosanio, G., Widom, J., Uhlenbeck, O.C., (2015). In vitro selection of DNA s with an increased propensity to form small circles. *Biopolymers* **103** (6), 303–320.
15. Basu, A., Bobrovnikov, D.G., Qureshi, Z., Kayikcioglu, T., Ngo, T., Ranjan, A., Eustermann, S., Cieza, B., et al., (2021). Measuring DNA mechanics on the genome scale. *Nature* **589** (7842), 462–467.
16. Drew, H.R., Wing, R.M., Takano, T., Broka, C., Tanaka, S., Itakura, K., Dickerson, R.E., (1981). Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Nat. Acad. Sci.* **78** (4), 2179–2183.
17. Bednar, J., Furrer, P., Katritch, V., Stasiak, A., Dubochet, J., Stasiak, A., (1995). Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *J. Mol. Biol.* **254** (4), 579–594.
18. Wiggins, P.A., Van Der Heijden, T., Moreno-Herrero, F., Spakowitz, A., Phillips, R., Widom, J., Dekker, C., Nelson, P.C., (2006). High flexibility of DNA on short length scales probed by atomic force microscopy. *Nat. Nanotechnol.* **1** (2), 137–141.
19. Abels, J., Moreno-Herrero, F., Van der Heijden, T., Dekker, C., Dekker, N.H., (2005). Single-molecule measurements of the persistence length of double-stranded RNA. *Biophys. J.* **88** (4), 2737–2744.
20. Pasi, M., Maddocks, J.H., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham III, T., Dans, P.D., Jayaram, B., et al., (2014). μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **42** (19), 12272–12283.
21. Pérez, A., Castellazzi, C.L., Battistini, F., Collinet, K., Flores, O., Deniz, O., Ruiz, M.L., Torrents, D., et al., (2012). Impact of methylation on the physical properties of DNA. *Biophys. J.* **102** (9), 2140–2148.
22. Battistini, F., Dans, P.D., Terrazas, M., Castellazzi, C.L., Portella, G., Labrador, M., Villegas, N., Brun-Heath, I., et al., (2021). The impact of the hydroxymethylcytosine epigenetic signature on DNA structure and function. *PLoS Comput. Biol.* **17** (11), e1009547.
23. Noy, A., Pérez, A., Lankas, F., Luque, F.J., Orozco, M., (2004). Relative flexibility of DNA and RNA: a molecular dynamics study. *J. Mol. Biol.* **343** (3), 627–638.
24. Noy, A., Pérez, A., Márquez, M., Luque, F.J., Orozco, M., (2005). Structure, recognition properties, and flexibility of the DNA.RNA hybrid. *J. Am. Chem. Soc.* **127** (13), 4910–4920.
25. Cheatham, T.E., Kollman, P.A., (1997). Molecular dynamics simulations highlight the structural differences among DNA:DNA, RNA:RNA, and DNA:RNA hybrid duplexes. *J. Am. Chem. Soc.* **119** (21), 4805–4825.
26. Balaceanu, A., Buitrago, D., Walther, J., Hospital, A., Dans, P.D., Orozco, M., (2019). Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. *Nucleic Acids Res.* **47** (9), 4418–4430.
27. Beveridge, D.L., Barreiro, G., Byun, K.S., Case, D.A., Cheatham III, T.E., Dixit, S.B., Giudice, E., Lankas, F., et al., (2004). Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophys. J.* **87** (6), 3799–3813.
28. Dixit, S.B., Beveridge, D.L., Case, D.A., Cheatham 3rd, T. E., Giudice, E., Lankas, F., Lavery, R., Maddocks, J.H., et al., (2005). Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.* **89** (6), 3721–3740.
29. Petkevičiūtė, D., Pasi, M., Gonzalez, O., Maddocks, J.H., (2014). cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res.* **42** (20) e153–e153.
30. Da Rosa, G., Grille, L., Calzada, V., Ahmad, K., Arcon, J. P., Battistini, F., Bayarri, G., Bishop, T., et al., (2021). Sequence-dependent structural properties of B-DNA: what have we learned in 40 years? *Biophys. Revi.*, 1–11.
31. Patelli, A. (2019). A sequence-dependent coarse-grain model of B-DNA with explicit description of bases and phosphate groups parametrised from large scale Molecular Dynamics simulations, EPFL PhD thesis #9552.
32. Sharma, R. (2023). cgNA+: A sequence-dependent coarse-grain model of double-stranded nucleic acids, EPFL PhD thesis #9792.
33. De Bruin, L., Maddocks, J.H., (2018). cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res.* **46** (W1), W5–W10.
34. Liebl, K., Zacharias, M., (2021). Accurate modeling of DNA conformational flexibility by a multivariate ising model. *Proc. Nat. Acad. Sci.* **118** (15) e2021263118.
35. Assenza, S., Pérez, R., (2022). Accurate sequence-dependent coarse-grained model for conformational and elastic properties of double-stranded DNA. *J. Chem. Theory Comput.* **18** (5), 3239–3256.
36. Walther, J., Dans, P.D., Balaceanu, A., Hospital, A., Bayarri, G., Orozco, M., (2020). A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res.* **48** (5) e29–e29.
37. Hospital, A., Faustino, I., Collepardo-Guevara, R., Gonzalez, C., Gelpí, J.L., Orozco, M., (2013). NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.* **41** (W1), W47–W55.
38. Zheng, G., Lu, X.-J., Olson, W.K., (2009). Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.* **37** (suppl_2), W240–W246.
39. Poppleton, E., Romero, R., Mallya, A., Rovigatti, L., Šulc, P., (2021). Oxdna. org: a public webserver for coarse-grained simulations of DNA and RNA nanostructures. *Nucleic Acids Res.* **49** (W1), W491–W498.
40. Cevost, J., Vaillant, C., Meyer, S., (2018). ThreaDNA: predicting DNA mechanics' contribution to sequence

selectivity of proteins along whole genomes. *Bioinformatics* **34** (4), 609–616.
41. Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., Rohs, R., (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.* **41** (W1), W56–W62.
42. Dans, P.D., Walther, J., Gómez, H., Orozco, M., (2016). Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.* **37**, 29–45.
43. Mitchell, J.S., Glowacki, J., Grandchamp, A.E., Manning, R.S., Maddocks, J.H., (2017). Sequence-dependent persistence lengths of DNA. *J. Chem. Theory Comput.* **13** (4), 1539–1555.
44. Mauney, A.W., Tokuda, J.M., Gloss, L.M., Gonzalez, O., Pollack, L., (2018). Local DNA sequence controls asymmetry of DNA unwrapping from nucleosome core particles. *Biophys. J.* **115** (5), 773–781.
45. Giniūnaitė, R., Petkevičiūtė-Gerlach, D., (2022). Predicting the configuration and energy of DNA in a nucleosome by coarse-grain modelling. *PCCP* **24**, 26124–26133.
46. Bendandi, A., Patelli, A.S., Diaspro, A., Rocchia, W., (2020). The role of histone tails in nucleosome stability: An electrostatic perspective. *Comput. Struct. Biotechnol. J.* **18**, 2799–2809.
47. Glowacki, J. (2016). Computation and Visualization in Multiscale Modelling of DNA Mechanics, EPFL PhD thesis #7062.
48. Zwahlen, T. (2023). Landscapes of DNA mechanics and Genomes, EPFL PhD thesis #8784.
49. Gonzalez, O., Petkevičiūtė, D., Maddocks, J.H., (2013). A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.* **138** (5), 02B604.
50. Lavery, R., Moakher, M., Maddocks, J.H., Petkevičiūtė, D., Zakrzewska, K., (2009). Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **37** (17), 5917–5929.
51. Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., Heinemann, U., Lu, X., et al., (2001). A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.* **313** (1), 229–237.