PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2020, **22**, 19254

SerraNA: a program to determine nucleic acids elasticity from simulation data†

Victor Velasco-Berrelleza,[a] Matthew Burman,[a] Jack W. Shepherd,[a] Mark C. Leake,[ab] Ramin Golestanian [cd] and Agnes Noy [a]*

Received 19th May 2020,
Accepted 12th August 2020

DOI: 10.1039/d0cp02713h

rsc.li/pccp

The resistance of DNA to stretch, twist and bend is broadly well estimated by experiments and is important for gene regulation and chromosome packing. However, their sequence-dependence and how bulk elastic constants emerge from local fluctuations is less understood. Here, we present SerraNA, which is an open software that calculates elastic parameters of double-stranded nucleic acids from dinucleotide length up to the whole molecule using ensembles from numerical simulations. The program reveals that global bendability emerge from local periodic bending angles in phase with the DNA helicoidal shape. We apply SerraNA to the whole set of 136 tetra-bp combinations and we observe a high degree of sequence-dependence with differences over 200% for all elastic parameters. Tetramers with TA and CA base-pair steps are especially flexible, while the ones containing AA and AT tend to be the most rigid. Thus, AT-rich motifs can generate extreme mechanical properties, which are critical for creating strong global bends when phased properly. Our results also indicate base mismatches would make DNA more flexible, while protein binding would make it more rigid. SerraNA is a tool to be applied in the next generation of interdisciplinary investigations to further understand what determines the elasticity of DNA.

Introduction

For genomes to function properly, chromosomes need to fold into a hierarchy of structures, causing, for example, expression correlation of genes located within the same topological domain.[1] Besides, it is widely known that DNA looping is a fundamental structure for gene regulation that facilitates long-range communication between a promoter and its distal regulatory elements.[2,3] Moreover, DNA can be subjected to forces up to tens of pN approximately in cells due to the activity of protein motors.[4] And finally, on the shortest scale, DNA distortion has been detected as determining the formation of diverse DNA:protein complexes like nucleosomes, some transcription factors or bacterial nucleoid association proteins.[5] Therefore, it is important to measure the mechanical response of DNA to bending, stretching and torsion, which is well established to have average values close to 50 nm for the persistence length,[6–10] between 1100–1500 pN for the stretch modulus[11,12] and ranging from 90 to 120 nm for the torsion elastic constants[8,13,14] (for a good summary of experimental values see Lipfert *et al.*[15]).

What is less clear from experimental data is the spread of elastic properties depending on sequence and which local elements build up the bulk flexibility of long DNA fragments. There have been several attempts to deduce the particular values associated to a sequence from cyclization probabilities, although these methodologies are not unambiguous and require the use of theoretical models.[16,17] In addition, it has been very difficult to identify the mechanisms through which some short sequence motifs, like A-tracts, originate extraordinary bending.[18,19] On these matters, molecular dynamics (MD) simulations at atomic resolution have become an impressive source of new important information,[20] that have provided (i) systematic analysis at the dinucleotide level,[21,22] (ii) an evaluation of the influence of nearest flanking base-pairs (bp) up to the tetranucleotide level[23,24] and, among others, (iii) an explanation of contradictory stiffness data on A-tracts.[25] On a more coarse-grained level, Monte Carlo (MC) simulations have found that most of sequence-dependence variability is originated at the level of static curvature.[26]

Previously, we designed the Length-Dependent Elastic Model (LDEM) for describing how bulk elastic properties emerge from bp fluctuations using the sampling obtained by nucleic acids simulations.[27] The LDEM revealed that the crossover from local to global occurs typically within one helical turn

[a] *Department of Physics, University of York, York, YO10 5DD, UK. E-mail: agnes.noy@york.ac.uk*
[b] *Department of Biology, University of York, York, YO10 5NG, UK*
[c] *Max Planck Institute for Dynamics and Self-Organization (MPIDS), Göttingen, 37077, Germany*
[d] *Rudolf Peierls Center for Theoretical Physics, University of Oxford, Oxford OX1 3PU, UK*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0cp02713h

of DNA[27] as has been confirmed by others.[28,29] In terms of torsion elasticity, we observed a transition from dinucleotide values of 30–50 nm to the long-range elastic constants of 90–120 nm in agreement with experimental data.[8,13,14,30] The model also revealed that stretch modulus changed as a function of molecular length in a non-monotonic way on shorter scales followed by a stabilization to similar values of force-extension measurements (1100–1500 pN).[11,12,27,29] Highly soft stretch modulus measured by SAXS experiments on short oligomers[31] was observed to be caused mainly by end effects.[27] For the persistence length, we found that the periodic tangent–tangent correlation reflected the "crookedness"[32] of the static curvature of the DNA helix[26,27,29,30,32–35] and, without considering these modulations, the decay was close to the consensus value of 50 nm.[6–10] Thus, the LDEM is suitable for describing the average mechanical properties of DNA and, from this perspective, it was applied to test the DNA force-field for atomic simulations, Parmbsc1.[36]

Here we present SerraNA, which is an open-source, versatile and integrated implementation of the LDEM, that allows fast simulation analysis and detection of emergent sequence effects. It calculates the overall elastic constants of helical nucleic acids (NA) and the elastic structure profiles for every possible sub-length (*serra* from Latin means "mountain range"). To our knowledge, there is no other program that estimates bulk flexibility constants from ensembles obtained by numerical simulations and that uncovers systematically how these properties emerge from local sequence-dependence fluctuations.

The paper describes the theoretical background behind the LDEM and it provides estimations for the different elastic constants by using MD simulations over a series of DNA fragments between 32 to 62 bp. Then, SerraNA is used to determine how bulk elastic constants emerge from local fluctuations using bendability as an example. We also apply the program to perturbed DNAs due to a series of factors like base mismatch and protein binding, in particular, the nucleosome and the GCN4 transcription factor. Finally, the program is applied to the ABC trajectory database,[23] which contains all the possible tetra-bp combinations, for exhaustively evaluating the dependence on sequence of the elasticity of DNA.

The length-dependent elastic model (LDEM)

Geometric description for different fragments lengths

The two bending angles, roll and tilt, and the rotational angle twist at the bp-step level are adapted to evaluate the relative orientation of a pair of bp spaced by an increasing number of nucleotides. The vertical displacement, which is associated with stretch, is characterized by end-to-end distance but a fragment's contour length is also calculated for a more comprehensive description of the polymer structure (see below and Fig. 1 for further details).

The spatial configuration of a bp i is specified by giving the location of a reference point ($\hat{r}_i$) and the orientation of a right-handed orthonormal reference triad (T_i) following the mathematical procedure of the 3DNA program,[37] where $\hat{y}_i$ points to the backbone of first strand, $\hat{x}_i$ points to the major groove and $\hat{z}_i$ marks the molecular direction at that particular point (Fig. 1b). Then, the CEHS scheme is applied for obtaining the molecular twist and the roll tilt contributions to bend[38,39] (Fig. 1). The algorithm is used to calculate the mid-step triad T_{mst} between bp i and j that define an oligomer whose length ranges from 2 bp to N (Fig. 1b). N is the total number of bp in the DNA fragment minus the two for each end, which have been discarded in order to avoid temporary loss of base pairing and other end effects.

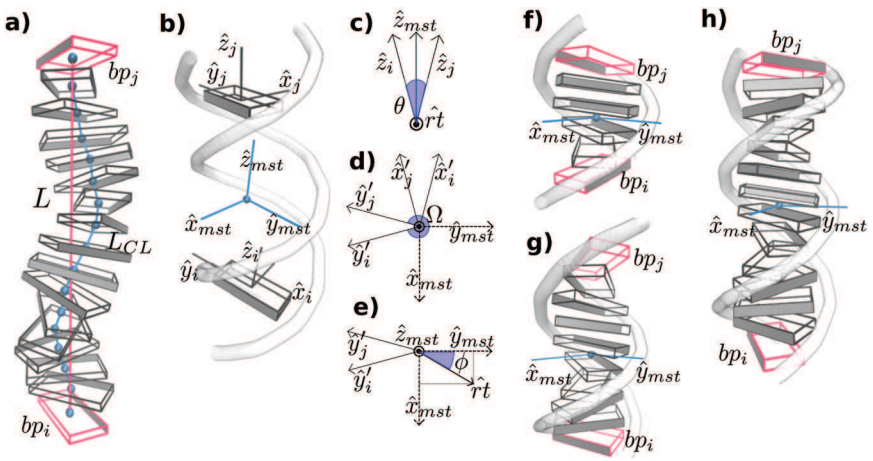

Fig. 1 Schematic diagrams of the algorithm implemented in SerraNA for calculating the geometric parameters at different fragment lengths. (a) Vertical displacement is characterize by end-to-end distance (in red) and contour length (in blue). (b) Twist and bend angles between bp i and j are obtained *via* the mid-base triad (T_{mst}) positioned at the mid-point. (c) Bending angle θ and bending axis $\hat{rt}$ are defined by directional vectors $\hat{z}_i$ and $\hat{z}_j$. Co-planar vectors $\hat{y}_i'$, $\hat{y}_j'$, $\hat{x}_{\mathrm{mst}}$ and $\hat{y}_{\mathrm{mst}}$ define twist angle Ω (d) and roll and tilt bending angles (e). (f)–(h) T_{mst} between bp (in red) separated by 4, 8 and 12 bp, respectively. Roll and tilt consist on bending towards grooves and backbone, respectively, at the fragment midpoint for the different sub-fragment lengths. Angles and T_{mst} are highlighted in blue.

The bending angle θ is obtained directly from the direction correlation ($\theta = \cos^{-1}(\hat{z}_i \cdot \hat{z}_j)$) and the corresponding bending axis $\hat{rt}$ is calculated by $\hat{rt} = \hat{z}_i \times \hat{z}_j$ (Fig. 1b). Next, T_i and T_j are rotated around $\hat{rt}$ by half of θ for obtaining $T_i' = R_{rt}(+\theta/2)T_i$ and $T_j' = R_{rt}(-\theta/2)T_j$, where the transformed $x-y$ planes are now parallel with each other and their z-axes coincide (see Fig. 1c and d). T_{mst} is directly built by averaging and normalizing T_i' and T_j'. The corresponding 3 rotations (tilt τ, roll ρ, twist Ω) are defined as:

$$\Omega = \cos^{-1}(\hat{y}_i' \cdot \hat{y}_j'); \ \rho = \theta \cos\phi; \ \tau = \theta \sin\phi \quad (1)$$

where ϕ is the angle between $\hat{rt}$ and the $\hat{y}_{mst}$ (Fig. 1e). Note that roll and tilt variables in lengths longer than a dinucleotide denote bending towards grooves and backbone direction, respectively, according the T_{mst} *i.e.* the fragment midpoint (see Fig. 1).

For each DNA sub-fragment, end-to-end distance (L) and contour length (L_{CL}) are defined as:

$$L = |r_j - r_i|; \quad L_{CL} = \sum_i^{j-1} |r_{i+1} - r_i|. \quad (2)$$

For completeness, the three rigid-body translation variables at the dinucleotide level (shift $X_{i,i+1}$, slide $Y_{i,i+1}$ and rise $Z_{i,i+1}$) are calculated by:

$$[X_{i,i+1}\ Y_{i,i+1}\ Z_{i,i+1}] = (r_{i+1} - r_i)T_m \quad (3)$$

and the extrapolation to longer scales can be designated by:

$$[\,X_0 \quad Y_0 \quad Z_0\,] = \sum_i^{j-1} [\,X_{i,i+1} \quad Y_{i,i+1} \quad Z_{i,i+1}\,], \quad (4)$$

where added-shift X_0, added-slide Y_0 and added-rise Z_0 can be interpreted structurally as the three pseudo components of L_{CL}.

For better comparison with experiments, only end-to-end distance L, twist Ω, roll ρ and tilt τ are utilized for the calculation of DNA elastic constants. SerraNA outputs the 'structural_parameters.out' file with the complete set of structural variables (including total bending angle, directional correlation, contour length and added shift, slide and rise) at all lengths with the idea of providing a full conformational illustration of the whole molecular stretch (see flowchart in Fig. 2 for more details).

Note that for severely bent DNAs ($\theta > 180$ degrees), $\hat{rt}$ points to the wrong opposite direction, being this one of the limitations of the algorithm (Fig. 1c). For the measurement of strong bends, we recommend our other software, SerraLINE (https://github.com/agnesnoy/SerraLINE), which considers only a global molecular contour[33] and which can be applied for comparison to microscopy images.

The length-dependent model of DNA elasticity

Under the assumption that distribution of values adopted by a variable X is fully Gaussian and non-correlated with the rest of deformation parameters, the corresponding elastic constant K for a particular length can be easily derived from its variance Var(X) estimated during a MD trajectory:[27,40]

$$K = k_B T b N \frac{1}{\mathrm{Var}(X)} \quad (5)$$

where fragment length or sub-lengths are specified by N dinucleotide steps with rise b = 0.34 nm. High quantile-quantile correlations ($R^2 > 0.98$) indicates this premise is reasonably good with the exceptions of twist bimodality[22,24] at short length scales and end-to-end skewness at long sub-fragment lengths (Fig. S1 and S2, ESI†).

However, the four distortion variables chosen to describe DNA flexibility on this model (roll, tilt, twist and end-to-end) are non-orthogonal. This effect is specifically taken into consideration by determining elastic constants as the diagonal terms of the inverse covariance matrix V^{-1} or elastic matrix F:[41]

$$F = k_B T b N\, V^{-1}, \quad (6)$$

Correspondingly, the diagonal terms of V^{-1} can be understood as the reciprocal of the partial variances, ($1/\mathrm{Var}_p(X)$). $\mathrm{Var}_p(X)$ is a measure of the residual variance associated with a deformation after removing the linear effects caused by other variables.[27,42] All terms from the different Fs calculated using all possible sub-fragments are printed in the 'elastic_parameters.out' output file for a complete dynamic description of the NA molecule (see Fig. 2).

Estimation of bulk twist elastic constant

The twist elastic constant for a singular sub-fragment k (C_k) is the diagonal term of F_k corresponding to twist, F_k being the elastic matrix associated to that particular DNA sub-fragment. Then, the twist elastic modulus as a function of length (C_N) is calculated by averaging all sub-fragments k with the same number of dinucleotide steps N: $C_N = \langle C_{k,N} \rangle$ (Fig. 3). Because the transition from bp level to the global elastic behavior occurs within one helix turn, values at lengths longer than 12 bp can already be considered good estimations of bulk twist elastic modulus C (Fig. 3a). Global C of an individual DNA fragment is calculated as the overall average of the series of C_N:

$$C = \sum_{11}^{N^*} \frac{C_N}{N^* - 11} \quad (7)$$

where N ranges from 11 bp-steps to N^*, N^* being the maximum sub-fragment length considered. By default SerraNA discards the ten longest sub-fragment lengths for counting N^* in order to have at least ten different values in averaging C_N, but this is an option that can be modified in the program (see Fig. 2).

Estimation of the long-range persistence length with its dynamic and static contributions

SerraNA calculates the persistence length A of a particular DNA fragment by means of (i) the linear fitting of the directional correlation decay or (ii) the inverse-covariance matrix method.

Mimicking the worm-like chain model (WLC), A is quantified by the linear approximation of the directional correlation

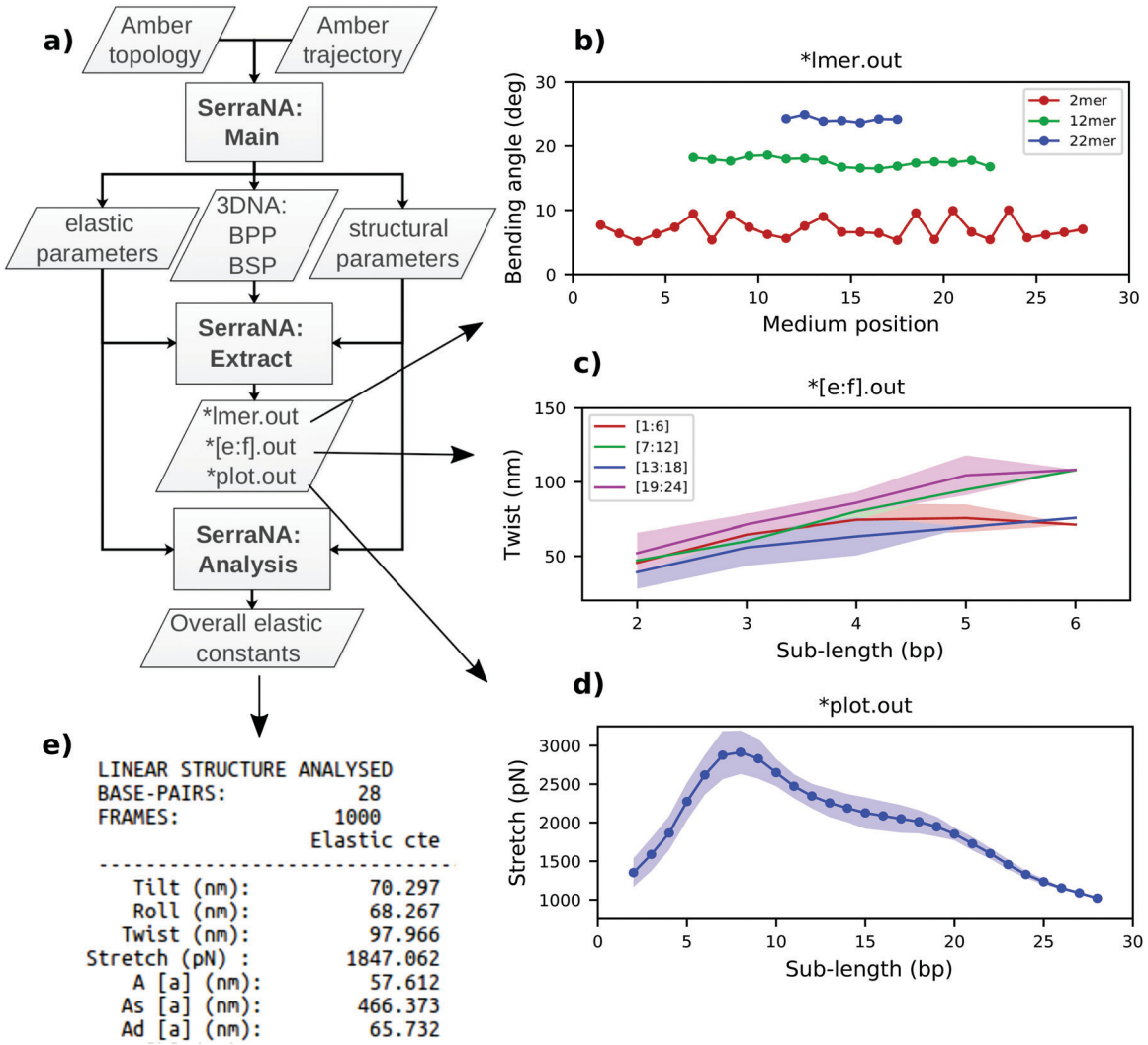

Fig. 2 General workflow of SerraNA using 32mer as an example. (a) Main program outputs bp and bp-step parameters (BPP and BSP, respectively), together with structural and elastic parameters at different lengths. Extract tool creates simple files for (b) plotting profiles along the molecule for a sub-length l (*lmer.out) and for (c) plotting the length-dependence from bp e to f (*[e:f].out) or (d) from the whole fragment (*plot.out). (e) Analysis tool extracts the overall elastic constants from a NA molecule.

decay between two bp tangent vectors, $\hat{z}_i$ and $\hat{z}_j$ separated by an increasing number of bp steps N with a distance rise b = 0.34 nm along the DNA:[27]

$$\langle \cos\theta_{i,j} \rangle \cong 1 - \frac{1}{2}\langle \theta_{i,j}^2 \rangle \equiv 1 - \frac{bN}{A}, \quad (8)$$

assuming a sufficiently weakly bending rod and where N ranges from 1 to N^* nucleotides, N^* being the longest sub-fragment considered on the fitting (see above paragraph) (Fig. 4a). The static and dynamic contributions to $\langle \theta^2 \rangle$ can be partitioned by $\langle \theta^2 \rangle = \langle \theta_s^2 \rangle + \langle \theta_d^2 \rangle$, where $\langle \theta_s^2 \rangle$ is originated from random distribution of sequence-dependent static bends and $\langle \theta_d^2 \rangle$ comes from the thermal fluctuations. $\langle \theta_s^2 \rangle$ are obtained through the DNA structure rebuilt[37] from the average base-pair step parameters. Then, the static and dynamic persistence length (A_s, A_d) are estimated by fitting the linear directional decay $1 - \frac{1}{2}\langle \theta_s^2 \rangle$ and $1 - \frac{1}{2}\langle \theta_d^2 \rangle$, respectively (Fig. 4b and c). A_s and A_d are combined using $1/A = 1/A_s + 1/A_d$[43] to obtain A again, which should be compatible with the direct linear fit to the full bending angle correlation decay.

The inverse-covariance method provides a second estimation of the dynamic persistence length (A_d') by directly combining the diagonal terms of F corresponding to the tilt and roll elastic constants (A_τ and A_ρ, respectively) for any pair of bp (Fig. 3):

$$\frac{1}{A_d'} = \frac{1}{2}\left(\frac{1}{A_\tau} + \frac{1}{A_\rho}\right). \quad (9)$$

Then, the global A_d' emerged from the entire DNA fragment is calculated following the methodology used for C (see above):

$$A_d' = \sum_{11}^{N^*} \frac{A_{d,N}'}{N^* - 11} \quad (10)$$

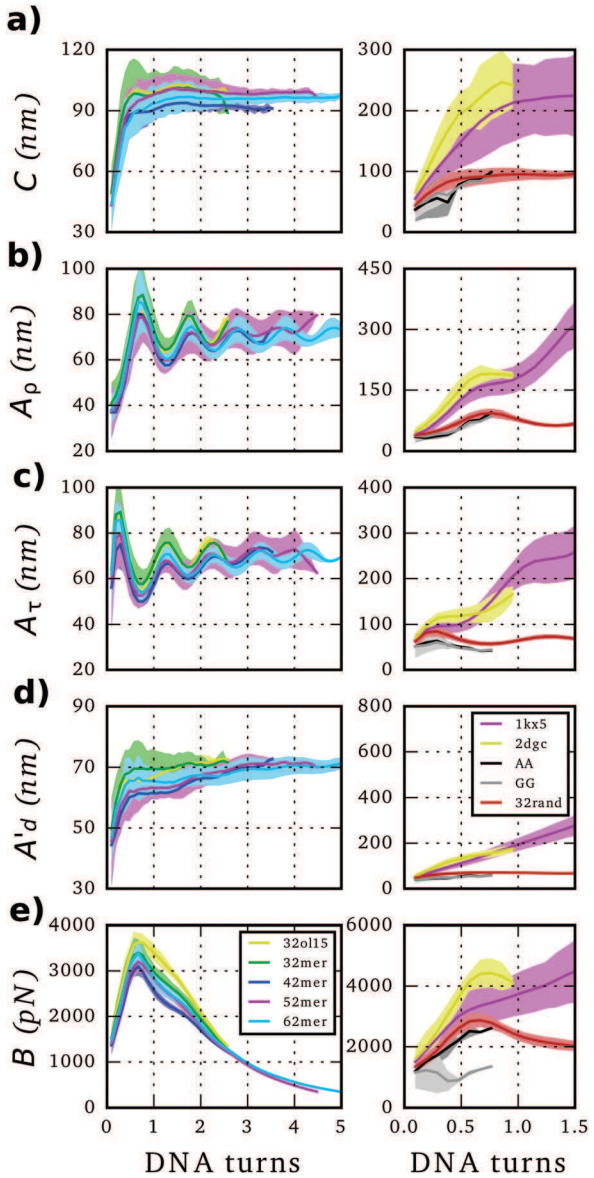

Fig. 3 Elastic constants associated with twist (C), roll (A_ρ), tilt (A_τ) and stretch (B) obtained through the inverse-covariance matrix method at different lengths, together with the dynamic persistence length (A_d') obtained *via* A_ρ and A_τ combination (see text for more details). Values reported here are averages over all possible sub-fragments with a particular length and the corresponding standard deviations are given as shade areas. Left, simulations on canonical relaxed DNAs run for this study. Right, simulations extracted from BIGNASim database containing different types of distortions like the nucleosome (PDB 1kx5), DNA bound to GCN4 transcription factor (PDB 2dgc) and oligomers with an AA or GG mismatch, where 32rand trajectory can be used as a reference for canonical behavior (see methods).

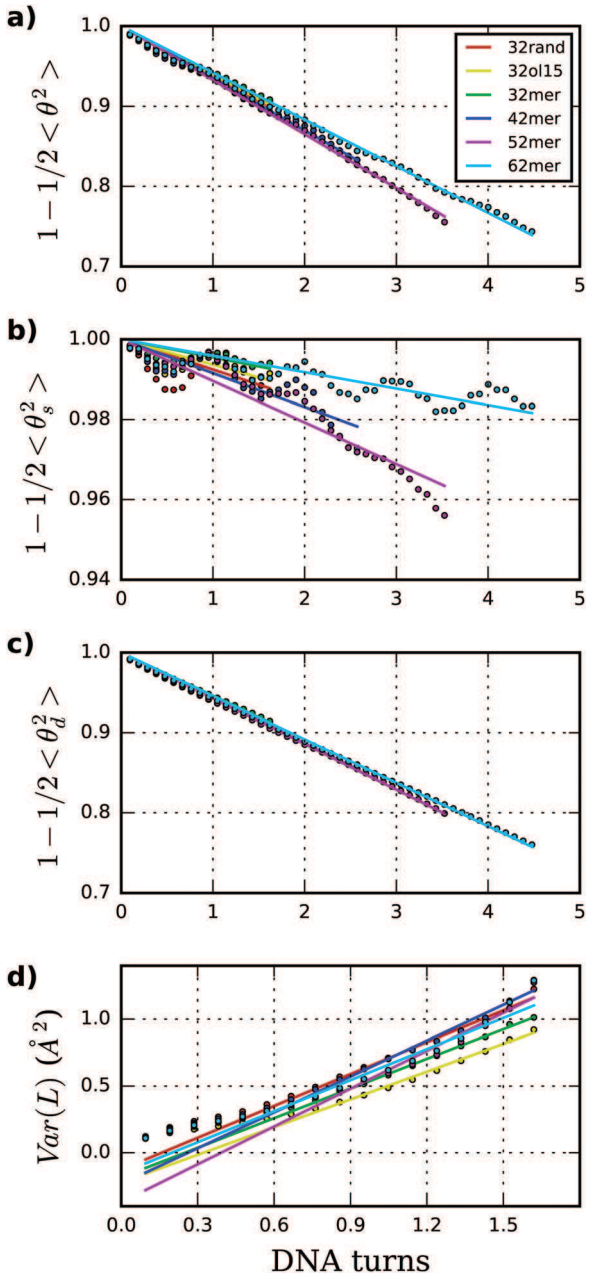

Fig. 4 (a)–(c) Persistence length (A) together with its static (A_s) and dynamic (A_d) contributions are obtained through the linear fit of directional decays associated to $\langle\theta^2\rangle$, $\langle\theta_s{}^2\rangle$ and $\langle\theta_d{}^2\rangle$, respectively. Values reported here are averages over all possible sub-fragments at a particular length, discarding ten longest lengths. (d) Stretch modulus (B) is obtained through linear fit of end-to-end partial variance using central 18mer.

where $A_{d,N}'$ are averages at a particular sub-fragment length with N bp-steps ranging from 11 to N^*, as the crossover from local to global dynamics occurs within the first DNA-turn (Fig. 3d).

A_d' provides higher values compared with the direct decay-fitting (A_d) as it just considers the partial variances associated with tilt and roll ($1/\mathrm{Var}_p(\tau)$ and $1/\mathrm{Var}_p(\rho)$, see above) after removing their linear correlations with the other deformation variables of F. A_d' is combined with the previously calculated A_s to obtain a second prediction of persistence length (A') using the expression $1/A' = 1/A_s + 1/A_d'$. In like manner, A' is stiffer than A as this value dismisses contributions from twist and stretch.

To account specifically for the asymmetry between minor and major grooves as it was stated by Marko and Siggia,[44] we

introduce the effect of twist-bend coupling (G) in eqn (9) for a new calculation of the dynamic persistence length (A_d''):[28,45]

$$\frac{1}{A_d''} = \frac{1}{2}\left(\frac{1}{A_\tau} + \frac{1}{A_\rho - G^2/C}\right). \quad (11)$$

Values for A_d'' are very similar to A_d' (Table S1, ESI†) indicating the importance of other cross-terms at the short length scales.

Estimation of bulk stretch modulus

In a similar way to twist, stretch moduli for all sub-fragments k (B_k) are acquired from the corresponding F_k's diagonal term associated to the end-to-end distance. As described before,[27] the stretch elastic profile as a function of length presents a complex behavior due to the prevalence of stacking interactions on the shortest oligomers and the appearance of extended end-effects softening the longest DNA parts (Fig. 3e and 5). In consequence, the bulk stretch modulus (B) is evaluated by considering only the end-to-end distances of the central 18mer and discarding oligomers shorter than 9 bp. Due to the limited number of points, the global S measure from a whole DNA molecule is obtained by fitting the linear increase of $\mathrm{Var_p}(L)$ within this length range, instead of averaging the equivalent B_N as in the previous sections (see Fig. 4d).

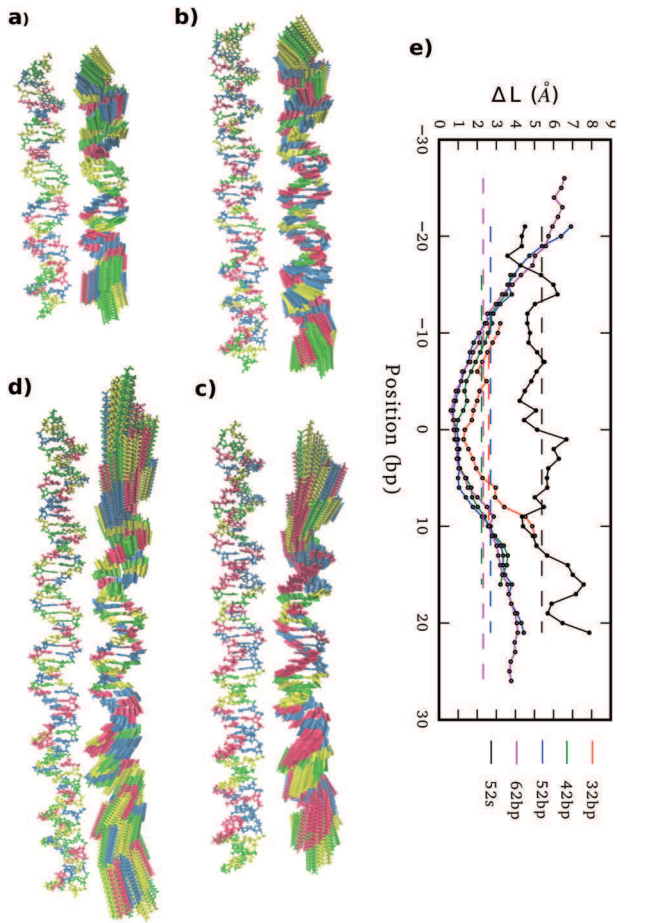

Fig. 5 (a)–(d) 32mer, 42mer, 52mer and 62mer averaged structures together with the corresponding end-effect essential modes. (e) Molecular position dependence of end-to-end distance local increments (ΔL) caused by end-vibrational modes from relaxed simulations and by the 52 bp pulling simulation with a maximum extension of 5% (52 s), using in all cases sub-fragments of 5 bp length.

Methods

Molecular dynamics simulations of linear DNA fragments

Linear DNA sequences of 32 bp (CGACTATCGC ATCCCGCTTA GCTATACCTA CG), 42 bp (CGCATGCATA CACACATACA TACACATACT AACACATACA CG), 52 bp (CGTATGAACG TCTATAAACG TCTATAAACG CCTATAAACG CCTATAAACG CG) and 62 bp (GCAGCAGCAC TAACGACAGC AGCAGCAGTA GCAGTAATAG AAGCAGCAGC AGCAGCAGTA GC) were extracted from the sequences 170–200 bp-long γ3, γ1, γ4 and γ2 as analyzed on Mitchell *et al.*,[26] which also correspond to the sequences NoSeq, CA, TATA and CAG on Virstedt *et al.*,[46] respectively. DNA duplexes were built using NAB module implemented in Amber16,[47] AMBER parm99 force-field[48] together with parmbsc0 and parmbsc1 corrections.[36,49] Fragments are named as 32mer, 42mer, 52mer and 62mer for the rest of the article. The 32-bp oligomer was also constructed using parmOL15[50,51] (named 32ol15 from now on). Structures were solvated in 200 mM Na^+ and Cl^- counter-ions[52] and in TIP3P octahedral boxes[53] with a buffer of 1.2 nm. Systems were energy-minimized, thermalized (T = 298 K) and equilibrated using standard protocols.[54,55] The final structures were subject to 1 μs of productive MD simulation at constant temperature (298 K) and pressure (1 atm)[56] using periodic boundary conditions, particle mesh Ewald[57] and an integration time step of 2 fs.[58] Principal component analysis was done with pyPcazip[59] and fast Fourier transforms were done with an in-house program written in python.

Trajectories obtained from BIGNASim and ABC simulation databases

Extra simulations were obtained from the BIGNAsim database[60] and analyzed together with the above. All simulations were run for 1 μs with bsc1 parameters,[36] TIP3P water model and neutralizing monovalent ions unless the contrary is stated:[36] (i) a DNA oligomer with 32 bp random sequence (ATGGATCCAT AGACCAGAAC ATGATGTTCT CA, labelled as 32rand from now on); (ii) nucleosome run for 500 ns (PDB 1kx5); (iii) DNA bound to the transcription factor GCN4 run with SPCE water (PDB 2dgc); and (iv) short oligomers with one A:A or G:G mismatch run for 500 ns (CCATACAATACGG, labelled as AA; CCATACGATACGG, labelled as GG, respectively).

Elastic properties for all distinct 136 tetranucleotides were obtained by analyzing MD simulations from the ABC consortium,[23] which are constituted of 39 oligomers of 18 bp, modeled for 1 μs, using parmbsc0 force-field,[49] SPC/E water[61] and 150 mM K^+Cl^- ion pair concentration.[62]

MD simulation of DNA pulling

The 52 bp oligomer was stretched on a series of umbrella sampling simulations in explicit solvent following the protocol developed by Shepherd *et al.*[63] Polymer length was increased in steps of 1 Å, which is in the range of thermal fluctuations for

unconstrained DNA,[27] thus, getting an almost instant equilibration after perturbation.[63] DNA was pulled by a total of 8 Å, resulting in a relative extension of just approximately 5%. This early-stage stretching regime is characterized by the maintenance of all canonical interactions on the double helix (hydrogen bonding and stacking), allowing a consistent comparison with the rest of trajectories run on relaxed DNA. Each umbrella sampling window was simulated for 1 ns making a simulation 8 ns long in total.

Linear regression and confidence intervals of elastic constants

Linear regression of directional memory and end-to-end partial variance ($\mathrm{Var_p}(L)$) on DNA length (N) are used to estimate bulk persistence lengths A and stretch modulus B, respectively, from gradients $\beta_A = -b/A$ and $\beta_B = k_\mathrm{B}Tb/B$. Confidence intervals of β_A and β_B ($\Delta\beta_A$ and $\Delta\beta_B$) are calculated with a confidence level of 70% in SerraNA using the student t-distribution for getting an almost direct comparison with other parameters where variability is estimated by standard deviation. Because A and B are non-linear functions ($y = f(x)$) of their respective gradients Δy or $f(x + \Delta x) - f(x)$, a confidence interval can be obtained approximately by $\frac{\partial f(x)}{\partial x}\Delta x$. Thus, confidence intervals for A and B (ΔA and ΔB) are calculated by:

$$\Delta A = \frac{b}{{\beta_A}^2}\Delta\beta_A \quad \Delta B = \frac{k_\mathrm{B}Tb}{{\beta_B}^2}\Delta\beta_B. \tag{12}$$

Results and discussion

SerraNA is a program written in Fortran that is freely accessible at https://github.com/agnesnoy/SerraNA under GNU Lesser General Public Licence and whose general workflow is shown in Fig. 2. The program builds upon the LDEM described by Noy and Golestanian[27] and it streamlines the procedure of calculating the persistence length, twist and stretch modulus of a DNA molecule or other double-stranded, helicoidal nucleic acids using an ensemble generated by MD or MC simulations.

Torsion elastic modulus

Elastic profiles as a function of length for the whole set of simulations are presented in Fig. 3. The calculated torsional modulus for all oligomers shows a crossover from the relatively soft value of around 30–60 nm at the single base-pair level to a large-scale asymptotic value between 90 and 100 nm (see Table 1), which is in agreement with previous study.[27] While softer values at short length scales are consistent with fluorescence polarization anisotropy measurements,[64,65] small-angle X-ray scattering (SAXS),[30] analysis of crystallographic DNA structures[41] and many calculations from MD,[29,36,40,66] stiffer magnitudes concur with single-molecule experiments[8,13,14] and other modeling estimations[20,29,67,68] at longer length scales. Values calculated for the whole segment also fall within the long-scale range between 90–100 nm (see Table 1), achieving an overall good convergence on the microsecond-long trajectories (Fig. S3, ESI†)

Persistence length

Persistence length (A), as well as its static (A_s) and dynamic (A_d) components, were deduced following the principles of the WLC model. Persistence lengths calculated by the fitting of directional decays are in general higher than the corresponding experimental data[46] and coarse-grained modeling[26] (see Table 1), although it should be noted that our magnitudes are obtained

Table 1 Bulk elastic constants estimated from unconstrained MD trajectories over linear DNA fragments[a]

DNA	C[b] (nm)	A[c] (nm)	A_s[c] (nm)	A_d[c] (nm)	A_d'[b] (nm)	A'[b] (nm)	B[c] (pN)
32rand	94.8 ± 0.8	57.4 ± 1.6	473 ± 87	65.4 ± 0.6	68.3 ± 1.0	59.7	1696 ± 15
32mer	100.1 ± 1.0 *101.4 ± 1.2*	61.1 ± 1.1 *58.7 ± 1.4* 56.3 **50.5 ± 2.1**	789 ± 93 *562 ± 74*	66.2 ± 0.7 *65.5 ± 0.8*	69.9 ± 0.5 *69.0 ± 1.1*	64.2 *61.4*	1920 ± 18 *2207 ± 41*
42mer	92.8 ± 0.7	54.8 ± 0.6 54.8 **45.5 ± 0.5**	422 ± 24	63.0 ± 0.6	64.7 ± 2.3	56.1	1705 ± 12
52mer	99.2 ± 0.8	52.9 ± 0.2 51.5 **45.5 ± 0.8**	344 ± 10	62.6 ± 0.2	67.8 ± 2.9	56.6	1843 ± 27
62mer	96.1 ± 0.6	61.2 ± 0.3 51.5 **41.7 ± 0.5**	869 ± 36	65.8 ± 0.2	68.2 ± 1.8	63.3	1731 ± 9
Average	96.6 ± 2.7	57.5 ± 3.3	579 ± 210	64.6 ± 1.5	67.8 ± 1.7	60.0 ± 3.3	1779 ± 88

[a] Elastic constants obtained using OL15 force-field are in italics. Persistence lengths on sequences over 100 bp, from which short fragments has been extracted from (see Methods), are in underlined text when they come from MC simulations[26] and in bold when they come from experiments.[46] [b] Overall averages and standard deviations for elastic constants obtained through the inverse-covariance method (twist C and persistence length A' and A_d') are calculated using the different sub-fragment lengths between 11 bp-steps to N^*, N^* being the maximum number of bp considered (see Methods). [c] Overall values and confidence levels at 70% for persistence lengths following the WLC model (A, A_s and A_d) and stretch modulus (B) are obtained through linear fits (see Methods).

with much shorter DNA molecules. Our average across sequences gives an overall stiffer estimation (57 ± s.d. 3 nm) compared with the range of experimental measurements (45–55 nm)[6–10] but in general agreement with estimations from simulations.[26,29,67] Part of this difference might be originated from the fact that our simulations are obtained with fully controlled ionic solutions (200 mM NaCl), without containing Mg^{2+} [46] and other buffers like Hepes, Tris or EDTA[7–10] known to affect DNA flexibility.[6,29,69] This variation could also be caused by inaccuracies in the modeling methods, although it is difficult to assess without comparing exactly the same sequences and with such a limited number of oligomers.

Fig. 4 shows tangent–tangent correlations arisen from A_s (*i.e.* from intrinsic curvature) exhibit modulations in phase with DNA-turn periodicity, in contrast to the decay originated from A_d (*i.e.* from thermal fluctuation).[27] Our calculations indicate A_s is much stiffer than A_d, even though A_s is the main source of variability (A_s = 576 ± 191 nm; A_d = 64.7 ± 1.4 nm; see Table 1 and Fig. 4). This trend was already observed on MC simulations[26] and it would explain the difficulty of arriving to a consensus description by experiments ($A_s > 1000$ and $A_d \approx$ 50 nm;[70] $A_s \approx 130$ and $A_d \approx 80$ nm[71]). For atomistic simulations, the small and oscillating decay together with the limited molecular length make the estimation of A_s (and as a consequence A) challenging and sometimes imprecise. These sources of error are exposed by the broad confidence intervals of A_s compared to A_d (Table 1) and the relative lack of convergence in some of A_s measurements *e.g.* for the 62mer (Fig. S3, ESI†). Another example is the discrepancy of A obtained by two different DNA force-fields (BSC1 and OL15), which is mainly caused by A_s and not A_d (see Table 1), being complicated to judge whether error comes from force-fields or the linear fit.

The inverse-covariance method yields an increased dynamic persistence length A_d' of 68 nm, and a resulting persistence length A' of 60 nm, as it only considers fluctuations not correlated with other deformation variables (*i.e.* partial variances, see methods)... is calculated through the combination of roll and tilt elastic constants, A_ρ and A_τ, which produce periodic and *anti*-symmetric profiles as a function of fragment length due to bending anisotropy towards grooves and backbone (see Fig. 3). On lengths containing half and complete helical turns, A_ρ and A_τ are equivalent because grooves and backbone face equitably towards both bending axes, whereas, at intermediate lengths, there is an imbalance between them (see Fig. 1).

Stretch modulus

Stretch modulus deduced from all unconstrained simulations present a non-monotonic dependence on length similar to the one previously described by Noy and Golestanian[27] and reproduced by Wales and co-workers[29] (see Fig. 3e). Base-stacking interactions cause stiffening at short scales up to 7 bp length as elastic constants present similar values associated with contour-lengths (see Fig. S4, ESI†). For longer sub-fragments, cooperativity emerges due to coordinated motion, softening the stretch modulus in two stages: (i) towards a plateau that would correspond to the regime captured by force-extension experiments[11,12] after incorporating an internal mode 13 bp long[27] and (ii) towards much more flexible magnitudes originated by long-ranged end-effects.[27] Principal component analysis reveals a mode that essentially captures vibration from edges and that produces a proportionate influence over the different oligomers containing gradually more bp (see Fig. 5). This fact shows that the characteristic length of the stretching-end mode is longer than five DNA-turns, still not reached for our atomistic simulations. In contrast, L increments are uniformly distributed along the molecule in the simulation where DNA is actively pulled (see Fig. 5), which shows that the end-stretching motion is just a vibrational mode not relevant for extracting the intrinsic stretch modulus of DNA.

We estimate stretch modulus *via* linear fitting of $\mathrm{Var_p}(L)$ just using the central 18 bp, since they constitute the molecular domain significantly unaltered by end-effects (see Fig. 5). Results give an overall average of 1779 ± 88 pN (Table 1), which is reasonably close to the experimental value *ca.* 1500 pN.[12]

From local to global elastic behavior

By analyzing elastic and structural length-dependence, SerraNA can also reveal how global elastic constants build up from the dynamics of smaller scales. For example, Fig. 6 compares the length-evolution on bending angles of the more bendable fragment (52mer) with the less one (62mer). Interestingly, bending is comparable between the two sequences at the single bp-step level (7.1 ± 1.5 and 7.2 ± 1.1 degrees, respectively), but are able to cause distinct values at the longer scale of 38 bp (35.6 ± 1.6 and 33.0 ± 0.7 degrees). The main difference at intermediate lengths (8, 16 and 28 bp) is the higher degree of periodicity, which is in phase with DNA helicoidal shape, presented by the curved oligomer compared to the straight one (see Fig. 6). Our data suggests that for creating a regular pattern characteristic of the curved fragment, a frequency with an exact number of cycles per DNA-turn at the single bp-step level is needed (3 cycles per DNA-turn for the 52mer in front of 3.5 for the 62mer, see Fig. 6), so local bends can couple for building up a significant curvature. Our results are in the same line of others that highlighted the importance of periodicity[32,72,73] for understanding the special mechanical properties of A-tracts[19] or nucleosome-positioning sequences.[74–76]

Protein-DNA and sequence mismatch

SerraNA has the capacity to deal with perturbed DNA molecules caused by a series of factors like sequence mismatch or protein binding. Although these singular cases might not comply with the harmonic approximation (Fig. S1, ESI†), the program can still provide indicative measurements of how the different type of perturbations affects the elasticity of DNA.

The introduction of a single A:A or G:G mismatch in the middle of an oligomer is enough to alter the structural parameters and to soften the corresponding elastic constants, not only at the dinucleotide level,[77] but also at the global molecular length (Fig. 3 and Fig. S5, ESI†). On the contrary, attachment of DNA molecules to proteins seems to constraint its dynamics, as the stiffer elastic constants suggests in Fig. 3. This effect is the

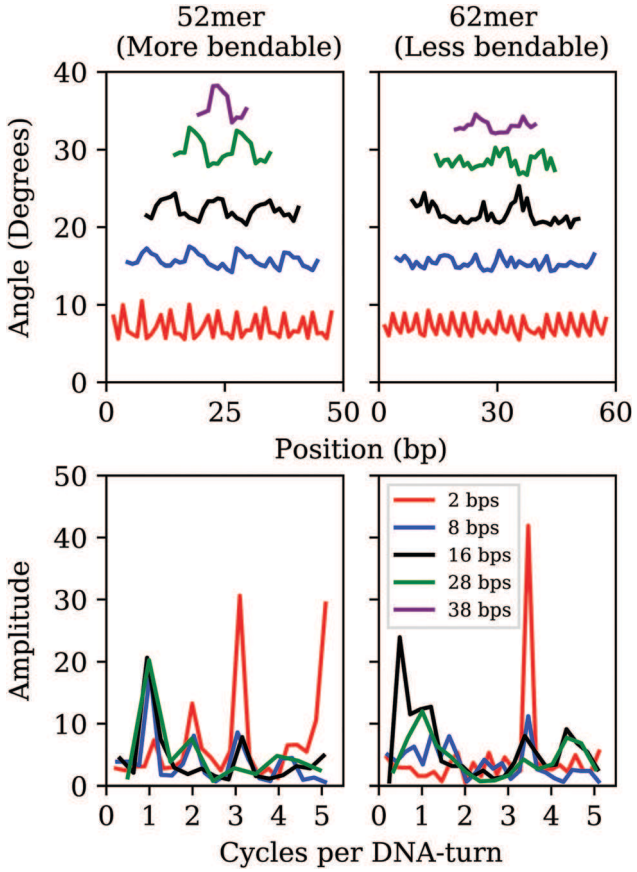

Fig. 6 Top: Length-evolution of bending angle profiles along the sequence for the most (left) and less (right) curved oligomers (52mer and 62mer, respectively). Bottom: Frequencies (in cycles per DNA-turn) obtained after applying fast Fourier transforms to bending positional data.

same for the two protein–DNA complexes selected in this study, despite their distinct character: nucleosome bends DNA strongly, while GCN4 keeps DNA uncurved. These preliminary results suggest that a role of protein recognition could be the confinement of DNA into one selected conformation from all configurational space. However, the analysis of more cases would be necessary for a more definite conclusion.

Tetranucleotide elastic constants from ABC database

Lastly, we analyzed how different DNA elastic properties depend on sequence. To this end, we applied SerraNA to the ABC simulation database, which contains the whole set of 136 tetranucleotide sequences in 39 different oligomers[23] (see Fig. 7). In general, we can observe a high degree of variability with flexible sequences twice as soft as rigid ones for all elastic constants.

The static persistence length is the most variable parameter in sequence space, spanning almost two orders of magnitude: from <25 nm in the case of TGGG, TGCA and CATG to >1000 nm for AATT (see Tables S2–S11, ESI†). In general we observe that the majority of the tetramers are very flexible and just 13 sequences (9%) have values >200 nm. The less curved tetramers involve central AA or AT steps, with AATT and AAAA being the top two with 1267 and 970 nm, respectively (Fig. 7 and Tables S2 and S7, ESI†). This is in agreement with previous studies and with the idea of A-tracts being so stiff that they impair nucleosomes wrapping[78] but facilitate looping and gene regulation when they are placed in phase.[78–80] It's worth mentioning that the extremely low values presented by most of the sequences are characteristic of this particular length (4 bp) as there is an accumulation of bending towards the major groove on one DNA side.[32] This behavior is reflected in the oscillations of the directional curvature correlation[27] (see Fig. 4b) and is exploited in fundamental processes like protein:DNA recognition[32,81] and the formation of DNA loops.[82]

When looking at the effect of thermal fluctuations on bendability (Fig. 7), we recover a scenario in agreement with previous crystallographic and modeling studies[21,23,66] where sequences containing the maximum number of "hinges" YR bp-steps (YRYR) are the most flexible and sequences with just RR and RY steps (RRYY and RRRY) the most rigid (being Y pyrimidines and R purines). The same tendency is observed on both bending degrees of freedom, roll and tilt (see Fig. S6, ESI†). Within the last two types of tetramers, sequences presenting central AA or AT steps are especially stiff (55 ± 4 nm, see Tables S4 and S7, ESI†) due to the influence of curvature, whereas YRYR tetramers containing TA and CA are specially flexible (17 ± 1 nm, see Table S8, ESI†).

The two different estimations of dynamic persistence lengths provide similar patterns on the sequence space in spite of their different ranges. We observed that although static persistence length presents more disparate values than the dynamic component, the latter is also important in determining the relative elasticity across all oligomers.

Fig. 7 shows that torsional moduli ranges from approximately 40 nm, which are characteristic of dinucleotides, up to 90 nm, which is a value typical of long scales. This is because 4 bp constitute an intermediate length in the transition from local to bulk (see Fig. 3), so the levels of correlation between bp-step fluctuations tend to diverge (see Fig. 2c), making sequence-dependence analysis very convoluted. Broadly, the most rigid sequences for this parameter are the ones with a central YR step (Fig. 7), which strikingly is the most flexible bp-step type at the dinucleotide level (see Fig. S7, ESI†) in agreement with previous studies.[21,54,66] We also observe that sequences with a bimodal behavior in the central step[22,24] don't show any special flexible feature. These facts demonstrate the remarkable importance of flanking bases in building up overall fluctuations and the very complex interplay between dinucleotide steps.[83,84]

Stretch modulus at the tetra-bp length are relatively high (see Fig. 7 and Tables S2–S11, ESI†) compared with experiments at the long-range scale. The remarkable similarity between other distance definitions (*i.e.* end-to-end, contour length or added-rise, Fig. 7 and Fig. S8, ESI†) suggests stretch stiffness at this length is mainly influenced by the strong stacking interactions. There is also an important degree of variability among sequences with some steps like AGGG, AGGA and AAGG presenting stretch modulus <1400 pN, which are twice as flexible as others such as CGAC, TTGC and CCGG (>2700 pN). In general, we observe YYRR and RRYY steps be the most rigid and

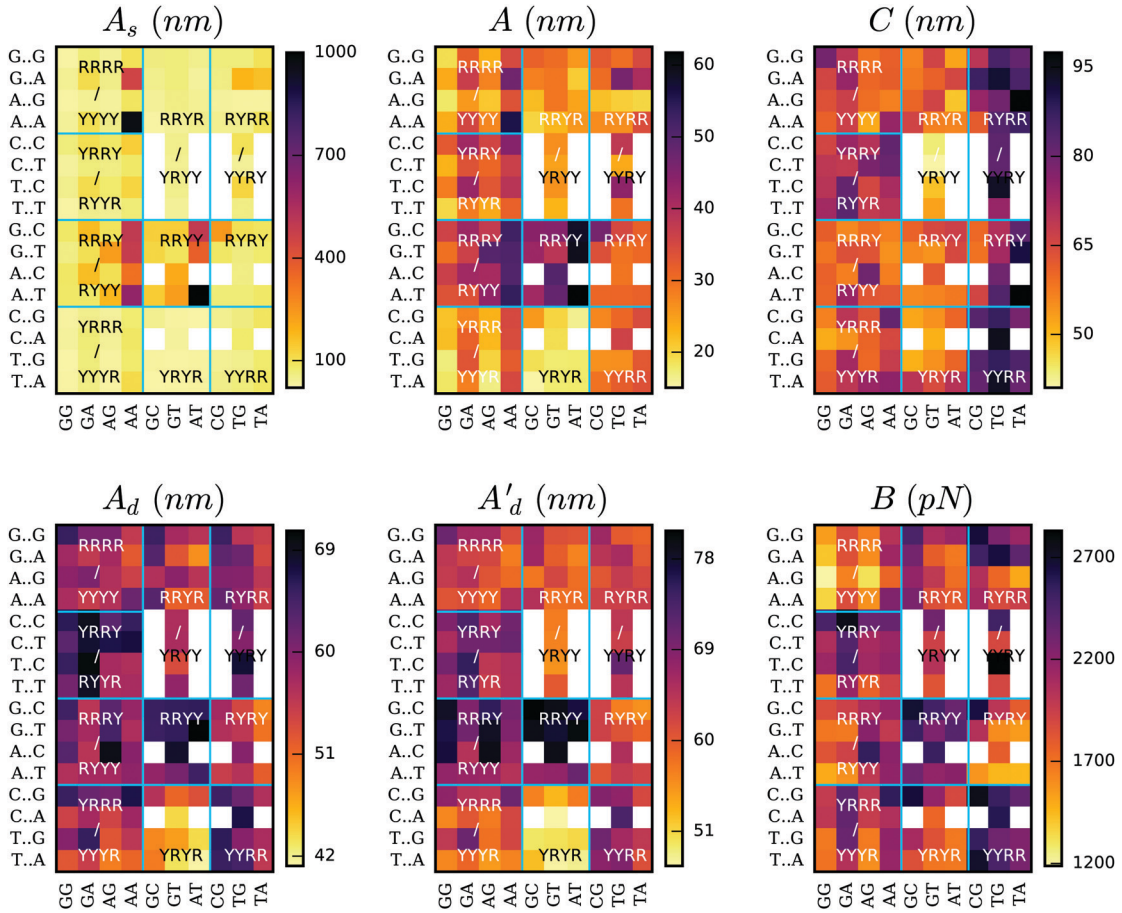

Fig. 7 Elastic constants at the length of 4 bp for the whole set of 136 tetra-nucleotide sequences obtained from ABC simulation database. Total persistence length together with its static and dynamic components (A, A_s and A_d, respectively) are calculated using the directional decay at the tetramer length. Twist (C), stretch modulus (B) and the second estimation of dynamic persistence length (A_d') are obtained directly from the inverse-covariance matrix for tetranucleotides. Vertical axis indicates middle steps, and horizontal axis flanking bases. Horizontal and vertical lines organize sequences according purine (R) or pyrimidine (Y) type. Sequence duplication is excluded through the use of white squares. AATT's A_s is off the palette with a value of 1267 ± 144 nm (see Table S7, ESI†).

RRRR the most flexible for this parameter, being determined mainly by the vertical component (added-rise) but also with some influence from lateral displacements, in particular from slide direction. AAAA sequence is an exception of RRRR type of tetramer by presenting a relatively stiff stretch modulus (2241 ± 88, see Table S2, ESI†), in reasonably good agreement with recent experimental data (~2400 pN).[19]

The analysis of ABC database makes clear that there is a flexibility dependence on the sequence of DNA and that reasonably extends to sequences larger than 4 bp. Regarding tetranucleotides elastic constants, rigidity tends to increase in regions composed by RRYY, YYRR, RRRY and YRRY, whereas sequences made with YRYR, RRYR are in general flexible, although this classification strongly depends on the type of elastic parameter.

Conclusions

In this article we present SerraNA, which is an open code that describes the elastic properties of nucleic-acids molecules with a canonical helicoidal shape (B- or A-form) using ensembles obtained from numerical simulations. We apply the program to analyze a series of atomistic MD simulations over DNA fragments and compare the extracted elastic values with available experiments.

We find reasonably good agreement on stretch and torsional modulus between our estimations (97 ± 3 nm and 1778 ± 88 pN) and experimental values (around 100 nm and 1500 pN, respectively). The calculation of stretch modulus is especially challenging because of the end-stretching vibration that masks the thermal fluctuations characteristic of the experimental stretch modulus at the range of kbp. As atomistic simulations are done over relatively short DNA molecules (tens to a hundred of bp), SerraNA approximates the calculation of this elastic parameter using only the two central DNA turns. In spite of all approximations, we find remarkable agreement between the only sequence experimentally measured, the A-tract, (~2400 pN)[19] and the modeled AAAA tetramer (2241 ± 88).

In the case of persistence length, simulations provide a slightly more rigid measure (57 ± 3 nm) than the generally

accepted value of 50 nm, although it's hard to discern whether it is due to intrinsic problems of force-fields, to non-identical ionic conditions with experimental buffers or to trouble in measuring the static persistence. Modulations on the tangent–tangent decay caused by DNA intrinsic shape and the relative shortness of the simulated DNA fragments makes the calculation of the static component of persistence length peculiarly complicated. Moreover, our simulations indicate that DNA curvature is the main source of variability on bendability between sequences (510 ± 210 nm), compared with 64.6 ± 1.5 nm caused purely by thermal fluctuations. When we analyze the whole set of tetrabases sequences from ABC database, we observed again a higher degree in variation on the static persistence length (s.d. 159 nm) in contrast to dynamic persistence length (s.d. 5.9 nm) (see Table S12, ESI†).

SerraNA also indicates how global elasticity emerges from local fluctuations by analyzing the change of mechanical properties as the length of considered fragments is systematically increased. Because the crossover from single base-pair level to bulk elastic behavior occurs typically within one helical turn of DNA, relatively short DNA fragments, like the ones simulated here, are already useful for uncovering this effect. In the case of persistence length, our results show that periodic patterns in phase of the DNA helical turn are particularly advantageous for developing significant bendability at longer scales.

We have demonstrated SerraNA can handle simulations where DNA is perturbed by protein-binding and mutational mismatch. However, they do not always satisfy the harmonic approximation, being one of the potential limits of our approach. Keeping this in mind, our results suggests mismatches would increase DNA flexibility, while protein binding would restraint its dynamics. Because we have only considered four examples here, one of which is the extreme case of the nucleosome, more cases would be necessary for a more definite result.

Finally, the systematic analysis of the whole set of 136 tetranucleotides reveals big differences with some sequences doubling others in all elastic parameters and, as a consequence, indicates the importance of sequence in determining DNA elastic properties. YRYR are the most flexible sequences compared with RRYY and RRRY, which are the most rigid. Particularly, AT and AA are the bp-steps causing less bendability, due to its straight natural configuration, in contrast to the highly flexible TA and CA bp-steps. RRYY and RRRY tetramers containing AT and AA steps present a persistence length 38 nm higher than YRYR tetramers with TA and CA steps. This demonstrate the role of AT-rich motifs in defining opposite mechanical properties, which can build up global deformability on longer sequences when they are regularly phased with the helicoidal shape. We thus see that SerraNA can shed light on the reasons behind the different emerging mechanical properties between AT and GC-rich long sequences[16,19] and, consequently, how their different biological functions might occur.[85]

In general, thought, we observe a complicated dependence for the different type of tetra steps compared with the dinucleotide level, showing the relevance of flanking sequences and the complex interplay between the different bp-steps. We expect that the use of SerraNA will help to clarify further how DNA elasticity can be modulated as a function of sequence, having important implications in understanding fundamental processes like DNA–protein recognition, DNA looping or packing inside the cell.[86] In particular, we anticipate using SerraNA in a range future experimental investigations[87] which will help us to unravel new physical properties of DNA at the single-molecule level.[88,89]

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank John Maddocks, Richard Lavery, Modesto Orozco and the rest of ABC consortium for giving us accessibility to all simulations of the database. This work was supported by the EPSRC grant EP/N027639/1 and by the Leverhulme Trust grants RPG-2019-156 and RPG-2017-340. V. V.-B. was funded by CONACYT agency from Mexican government (scholarship no 291163) and M. B. by EPSRC (EP/R513386/1). Computational time was secured on ARCHER and JADE *via* the UK High-End Computing Consortium for Biomolecular Simulation, HECBioSim, supported by EPSRC grant EP/R029407/1 and on Cambridge Tier-2 system funded by EPSRC Tier-2 capital grant EP/P020259/1. We also thank Tier 3 High Performance Computing (HPC) facilities at York (Viking and YARCC clusters) for additional computational resources.

Notes and references

1 J. H. Gibcus and J. Dekker, *Mol. Cell*, 2013, **49**, 773–782.
2 Y. Liu, V. Bondarenko, A. Ninfa and V. M. Studitsky, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 14883–14888.
3 D. G. Priest, S. Kumar, Y. Yan, D. D. Dunlap, I. B. Dodd and K. E. Shearwin, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, E4449–E4457.
4 S. Liu, G. Chistol and C. Bustamante, *Biophys. J.*, 2014, **106**, 1844–1858.
5 N. M. Luscombe, S. E. Austin, H. M. Berman and J. M. Thornton, *Genome Biol.*, 2000, **1**, 1–37.
6 C. G. Baumann, S. B. Smith, V. A. Bloomfield and C. Bustamante, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 6185–6190.
7 J. R. Wenner, M. C. Williams, I. Rouzina and V. A. Bloomfield, *Biophys. J.*, 2002, **82**, 3160–3169.
8 J. Lipfert, J. W. Kerssemakers, T. Jager and N. H. Dekker, *Nat. Methods*, 2010, **7**, 977–980.
9 E. Herrero-Galán, M. E. Fuentes-Pérez, C. Carrasco, J. M. Valpuesta, J. L. Carrascosa, F. Moreno-Herrero and J. R. Arias-González, *J. Am. Chem. Soc.*, 2013, **135**, 122–131.
10 A. K. Mazur and M. Maaloum, *Nucleic Acids Res.*, 2014, **42**, 14006–14012.
11 S. B. Smith, Y. Cui and C. Bustamante, *Science*, 1996, **271**, 795–798.

12 P. Gross, N. Laurens, L. B. Oddershede, U. Bockelmann, E. J. G. Peterman and G. J. L. Wuite, *Nat. Phys.*, 2011, **7**, 731–736.
13 Z. Bryant, M. D. Stone, J. Gore, S. B. Smith, N. R. Cozzarelli and C. Bustamante, *Nature*, 2003, **424**, 338–341.
14 F. Mosconi, J. F. M. C. Allemand, D. Bensimon and V. Croquette, *Phys. Rev. Lett.*, 2009, **102**, 078301.
15 J. Lipfert, G. M. Skinner, J. M. Keegstra, T. Hensgens, T. Jager, D. Dulin, M. Köber, Z. Yu, S. P. Donkers, F.-C. Chou, R. Das and N. H. Dekker, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 15408–15413.
16 Y. Zhang and D. M. Crothers, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 3161–3166.
17 S. Geggier and A. Vologodskii, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 15421–15426.
18 A. Barbič, D. P. Zimmer and D. M. Crothers, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 2369–2373.
19 A. Marin-Gonzalez, C. L. Pastrana, R. Bocanegra, A. Martín-González, J. G. Vilhena, R. Pérez, B. Ibarra, C. Aicart-Ramos and F. Moreno-Herrero, *Nucleic Acids Res.*, 2020, 5024–5026.
20 A. Marín-González, J. G. Vilhena, R. Pérez and F. Moreno-Herrero, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 7049–7054.
21 F. Lankaš, J. Šponer, J. Langowski and T. E. Cheatham, *Biophys. J.*, 2003, **85**, 2872–2883.
22 P. D. Dans, A. Balaceanu, M. Pasi, A. S. Patelli, D. Petkevičūtė, J. Walther, A. Hospital, G. Bayarri, R. Lavery, J. H. Maddocks and M. Orozco, *Nucleic Acids Res.*, 2019, **47**, 11090–11102.
23 M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, P. D. Dans, B. Jayaram, F. Lankaš, C. Laughton, J. Mitchell, R. Osman, M. Orozco, A. Pérez, D. Petkevičūtė, N. Spackova, J. Šponer, K. Zakrzewska and R. Lavery, *Nucleic Acids Res.*, 2014, **42**, 12272–12283.
24 P. D. Dans, I. Faustino, F. Battistini, K. Zakrzewska, R. Lavery and M. Orozco, *Nucleic Acids Res.*, 2014, **42**, 11304–11320.
25 T. Dršata, N. Špačková, P. Jurečka, M. Zgarbová, J. Šponer and F. Lankaš, *Nucleic Acids Res.*, 2014, **42**, 7383–7394.
26 J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning and J. H. Maddocks, *J. Chem. Theory Comput.*, 2017, **13**, 1539–1555.
27 A. Noy and R. Golestanian, *Phys. Rev. Lett.*, 2012, **109**, 228101.
28 E. Skoruppa, M. Laleman, S. K. Nomidis and E. Carlon, *J. Chem. Phys.*, 2017, **146**, 214902.
29 S. Xiao, H. Liang and D. J. Wales, *J. Phys. Chem. Lett.*, 2019, **10**, 4829–4835.
30 X. Shi, D. Herschlag and P. A. B. Harbury, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E1444–E1451.
31 R. S. Mathew-Fenn, R. Das and P. Harbury, *Science*, 2008, **322**, 446–449.
32 A. Marín-González, J. G. Vilhena, F. Moreno-Herrero and R. Pérez, *Phys. Rev. Lett.*, 2019, **122**, 048102.
33 T. Sutthibutpong, S. A. Harris and A. Noy, *J. Chem. Theory Comput.*, 2015, **11**, 2768–2775.
34 C. Gu, J. Zhang, Y. I. Yang, X. Chen, H. Ge, Y. Sun, X. Su, L. Yang, S. Xie and Y. Q. Gao, *J. Phys. Chem. B*, 2015, **119**, 13980–13990.
35 H. Dohnalová, T. Dršata, J. Šponer, M. Zacharias, J. Lipfert and F. Lankaš, *J. Chem. Theor. Comput.*, 2020, **16**, 2857–2863.
36 I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpí, C. González, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case and M. Orozco, *Nat. Methods*, 2015, **13**, 55–58.
37 X.-J. Lu and W. Olson, *Nucleic Acids Res.*, 2003, **31**, 5108–5121.
38 M. A. El Hassan and C. R. Calladine, *J. Mol. Biol.*, 1995, **251**, 648–664.
39 X.-J. Lu, M. A. El Hassan and C. A. Hunter, *J. Mol. Biol.*, 1997, **273**, 668–680.
40 A. Noy, A. Pérez, F. Lankaš, F. Javier Luque and M. Orozco, *J. Mol. Biol.*, 2004, **343**, 627–638.
41 W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock and V. B. Zhurkin, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 11163–11168.
42 J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley Publishing, 2009.
43 J. A. Schellman and S. C. Harvey, *Biophys. Chem.*, 1995, **55**, 95–114.
44 J. F. Marko and E. D. Siggia, *Macromolecules*, 1994, **27**, 981–988.
45 S. K. Nomidis, F. Kriegel, W. Vanderlinden, J. Lipfert and E. Carlon, *Phys. Rev. Lett.*, 2017, **118**, 217801.
46 J. Virstedt, T. Berge, R. M. Henderson, M. J. Waring and A. A. Travers, *J. Struct. Biol.*, 2004, **148**, 66–85.
47 D. Case, R. Betz, D. Cerutti, T. Cheatham, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Götz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.-S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko and P. Kollman, *Amber 16*, University of California, San Francisco, 2016.
48 W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, J. Kenneth, M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, 1995, **17**, 5179–5197.
49 A. Pérez, I. Marchán, D. Svozil, J. Šponer, T. E. Cheatham, C. A. Laughton and M. Orozco, *Biophys. J.*, 2007, **92**, 3817–3829.
50 M. Zgarbová, F. J. Luque, J. Šponer, T. E. Cheatham, M. Otyepka and P. Jurečka, *J. Chem. Theory Comput.*, 2013, **9**, 2339–2354.
51 M. Zgarbová, J. Šponer, M. Otyepka, T. E. Cheatham, R. Galindo-Murillo and P. Jurečka, *J. Chem. Theory Comput.*, 2015, **11**, 5723–5736.
52 D. E. Smith and L. X. Dang, *J. Chem. Phys.*, 1994, **100**, 3757–3766.
53 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.
54 A. Noy and R. Golestanian, *J. Phys. Chem. B*, 2010, **114**, 8022–8031.
55 T. Sutthibutpong, C. Matek, C. Benham, G. G. Slade, A. Noy, C. Laughton, J. P. Doye, A. A. Louis and S. A. Harris, *Nucleic Acids Res.*, 2016, **44**, 9121–9130.
56 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

57 T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.
58 J. P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.
59 A. Shkurti, R. Goni, P. Andrio, E. Breitmoser, I. Bethune, M. Orozco and C. A. Laughton, *SoftwareX*, 2016, **5**, 44–50.
60 A. Hospital, P. Andrio, C. Cugnasco, L. Codo, Y. Becerra, P. D. Dans, F. Battistini, J. Torres, R. Goñi, M. Orozco and J. L. Gelpí, *Nucleic Acids Res.*, 2015, **44**, D272–D278.
61 H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *J. Phys. Chem.*, 1987, **91**, 6269–6271.
62 L. X. Dang, *J. Am. Chem. Soc.*, 1995, **117**, 6954–6960.
63 J. W. Shepherd, R. J. Greenall, M. Probert, A. Noy and M. Leake, *Nucleic Acids Res.*, 2020, **48**, 1748–1763.
64 B. S. Fujimoto and J. M. Schurr, *Nature*, 1990, **344**, 175–178.
65 P. J. Heath, J. B. Clendenning, B. S. Fujimoto and M. J. Schurr, *J. Mol. Biol.*, 1996, **260**, 718–730.
66 A. Pérez, F. Lankaš, F. J. Luque and M. Orozco, *Nucleic Acids Res.*, 2008, **36**, 2379–2394.
67 F. Lankaš, J. Šponer, P. Hobza and J. Langowski, *J. Mol. Biol.*, 2000, **299**, 695–709.
68 K. Liebl and M. Zacharias, *J. Phys. Chem. B*, 2017, **121**, 11019–11030.
69 S. Guilbaud, L. Salomé, N. Destainville, M. Manghi and C. Tardin, *Phys. Rev. Lett.*, 2019, **122**, 028102.
70 M. Vologodskaia and A. Vologodskii, *J. Mol. Biol.*, 2002, **317**, 205–213.
71 J. Bednar, P. Furrer, V. Katritch, A. Stasiak, J. Dubochet and A. Stasiak, *J. Mol. Biol.*, 1995, **254**, 579–594.
72 S. K. Nomidis, M. Caraglio, M. Laleman, K. Phillips, E. Skoruppa and E. Carlon, *Phys. Rev. E*, 2019, **100**, 022402.
73 F. Mohammad-Rafiee and R. Golestanian, *J Phys.: Condens. Matter*, 2005, **17**, S1165–S1170.
74 T. E. Shrader and D. M. Crothers, *Proc. Natl. Acad. Sci. U. S. A.*, 1989, **86**, 7418–7422.
75 F. Mohammad-Rafiee and R. Golestanian, *Phys. Rev. Lett.*, 2005, **94**, 238102.
76 S. Balasubramanian, F. Xu and W. K. Olson, *Biophys. J.*, 2009, **96**, 2245–2260.
77 G. Rossetti, P. D. Dans, I. Gomez-Pinto, I. Ivani, C. Gonzalez and M. Orozco, *Nuc. Acids Res.*, 2015, **43**, 4309–4321.
78 T. Raveh-Sadka, M. Levo, U. Shabi, B. Shany, L. Keren, M. Lotan-Pompan, D. Zeevi, E. Sharon, A. Weinberger and E. Segal, *Nat. Genet.*, 2012, **44**, 743–750.
79 M. Tolstorukov, K. Virnik, S. Adhya and V. Zhurkin, *Nucleic Acids Res.*, 2005, **33**, 3907–3918.
80 A. R. Haeusler, K. A. Goodson, T. D. Lillian, X. Wang, S. Goyal, N. C. Perkins and J. D. Kahn, *Nucleic Acids Res.*, 2012, **40**, 4432–4445.
81 J. Li, J. M. Sagendorf, T.-P. Chiu, M. Pasi, A. Pérez and R. Rohs, *Nucleic Acids Res.*, 2017, **45**, 12877–12887.
82 M. Pasi, K. Zakrzewska, J. H. Maddocks and R. Lavery, *Nucleic Acids Res.*, 2017, **45**, 4269–4277.
83 A. Balaceanu, A. Pérez, P. D. Dans and M. Orozco, *Nucleic Acids Res.*, 2018, **46**, 7554–7565.
84 A. Balaceanu, D. Buitrago, J. Walther, A. Hospital, P. D. Dans and M. Orozco, *Nucleic Acids Res.*, 2019, **47**, 4418–4430.
85 A. E. Vinogradov and O. V. Anatskaya, *Mamm. Genome*, 2017, **28**, 455–464.
86 A. Noy, T. Sutthibutpong and S. A. Harris, *Biophys. Rev.*, 2016, **8**, 233–243.
87 A. Wollman, H. Miller, Z. Zhou and M. Leake, *Biochem. Soc. Trans.*, 2015, **43**, 139–145.
88 M. Leake, *Phil. Trans. R. Soc., B*, 2013, **368**, 20120248.
89 A. Bates, A. Noy, M. Piperakis, S. Harris and A. Maxwell, *Biochem. Soc. Trans.*, 2013, **41**, 565–570.