Published online 20 January 2020 *Nucleic Acids Research, 2020, Vol. 48, No. 5* ***e29***
doi: 10.1093/nar/gkaa015

A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level

Jürgen Walther[1], Pablo D. Dans[1,2,*], Alexandra Balaceanu[1], Adam Hospital[1], Genís Bayarri[1] and Modesto Orozco[1,3,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain, [2]Department of Biological Sciences, CENUR North Coast, University of the Republic, 50000 Salto, Uruguay and [3]Department of Biochemistry and Biomedicine, The University of Barcelona, 08028 Barcelona, Spain

Received July 05, 2019; Revised December 10, 2019; Editorial Decision January 02, 2020; Accepted January 04, 2020

ABSTRACT

We present a new coarse grained method for the simulation of duplex DNA. The algorithm uses a generalized multi-harmonic model that can represent any multi-normal distribution of helical parameters, thus avoiding caveats of current mesoscopic models for DNA simulation and representing a breakthrough in the field. The method has been parameterized from accurate parmbsc1 atomistic molecular dynamics simulations of all unique tetranucleotide sequences of DNA embedded in long duplexes and takes advantage of the correlation between helical states and backbone configurations to derive atomistic representations of DNA. The algorithm, which is implemented in a simple web interface and in a standalone package reproduces with high computational efficiency the structural landscape of long segments of DNA untreatable by atomistic molecular dynamics simulations.

INTRODUCTION

Under physiological conditions DNA behaves like a polymeric entity whose properties are dependent on the underlying sequence. Experimental approaches to the determination of sequence-dependent physical properties of DNA are impaired by their inability to deal with long and flexible polymers, which has fueled the development of theoretical simulation techniques (1), among them atomistic molecular dynamics (MD), a method that after recent improvements in force-fields (2,3) has shown extreme accuracy in describing the structural and dynamic properties of a variety of DNA structures (4–10). Unfortunately, the computational cost of MD simulation scales (roughly) with the third power of the length of the duplex, and a simple 100 bp duplex would require a simulation box containing more than 10^7 water molecules, a system for which reaching reasonable simulation times is nearly impossible.

Coarse grained (CG) methods are a cost-effective alternative to simulate very long segments of DNA, approaching the chromatin scale. In summary, two families of CG methods have been developed (1,11–14): the first ones (Cartesian CG) are based on reducing the atomistic representation of the nucleotides to a few beads whose interactions are defined by empirical potentials and whose movements are followed by means of (typically) Langevin-Brownian MD algorithms (15–17). The second family of methods (helical CG) reduces the degrees of freedom in DNA by considering the nucleobases or the base pairs (bp) as rigid planes whose movements are defined by three rotations and three translations. In this second family of methods the sampling is typically obtained by means of Monte Carlo (MC) simulation techniques. While the Cartesian CG methods have the advantage of universality, for physiological DNAs, helical CG methods are probably more efficient as helical coordinates are better suited to describe the essential movements of DNA (12,13).

Three crucial choices must be taken in defining a helical CG model. The first one is the level of resolution: nucleobases or base pairs. In nucleobase-resolution scheme the CG model should account for the movement of each nucleobase with respect to three neighbors in a simple base pair step (bps) (the paired one, one located at the 3′, and one at 5′ in the opposite strand), which sums up to 6^3 degrees of freedom per nucleobase. By combining nucleobase (intra base pair) and base pair step (inter base pair) helical coordinates the number of degrees of freedom can be significantly reduced (18–20). Simpler and more popular (21,22) are helical-CG methods that represent the DNA at the base pair level. In this case, movements at the base pair step level are limited to three translations and three rotations (shift, slide, rise, tilt, roll, twist), which at the expense of some loss of resolution, drastically simplifies the calculation and the parameterization of the model.

*To whom correspondence should be addressed. Tel: +34934037156; Email: modesto.orozco@irbbarcelona.org
Correspondence may also be addressed to Pablo D. Dans. Tel: +59891695145; Email: pablo.dans@unorte.edu.uy

The second important choice in building a helical CG model is the nature of the Hamiltonian (energy function) used to describe the dependence between the energy of the system and the change in helical coordinates. Most CG models rely on the use of a harmonic Hamiltonian (1,12–13,18–22), which assumes that under normal conditions the distributions of helical coordinates (at either nucleobase or base pair level) are Gaussian. Within this assumption the energy of the DNA can be easily described by means of a stiffness matrix and a deformation vector indicating the deviation of a helical coordinate from its equilibrium value (21). For the most common base pair resolution model this means that the energy is computed as shown in equation (1):

$$\mathrm{E}\,(\mathrm{X}) = \sum_{j=1}^{N} \frac{1}{2} \mathrm{K_j} \Delta \mathrm{X_j^2} \tag{1}$$

where E is the energy, N is the number of bps, $\mathrm{K_j}$ is the 6 × 6 stiffness matrix for bps j, and $\Delta\mathrm{X_j}$ is the 6-dimension deformation vector $\Delta\,\mathrm{X_j} = (\mathrm{X_j} - \mathrm{X_j^0})$, with $\mathrm{X_j}$ and $\mathrm{X_j^0}$ being the current conformation vector of bps j at a given point of the ensemble and the equilibrium vector respectively.

The last choice in the definition of a helical CG model is the origin of the parameters (stiffness matrix and the equilibrium vector $\mathrm{X_j^0}$ used to compute the deformation vector) defining the energy function. Original models developed by Olson & Zhurkin (21) extracted the parameters from the ten unique bps derived by inspection of the helical geometries of bps found in databases of crystal structures of DNA–protein complexes. Further refinements used MD simulation of different DNA duplexes containing the ten-unique bps as source of parameters (22,23). More recently, as the shortcomings of the bps scheme became evident, new nearest-neighbor harmonic models (NN) relying on inter base pair parameters adapted to all the different tetranucleotides emerged (1,6,24), with the corresponding parameters being fitted from atomistic MD simulations. These models showed a good ability to reproduce the conformational space of DNA duplexes, but were limited by two fundamental problems: (i) they were parameterized from the parmbsc0 force-field (2) which showed caveats in the representation of certain characteristics of the helix and (ii) they were based on the harmonic approximation, which is unable to reproduce multimodality shown both experimentally and theoretically in the distribution of inter base pair coordinates of certain bps (4,6–7,25–26).

We present here an evolution of the helical CG model which assumes a novel multi-normal model which accounts for the non-Gaussian nature of some inter base pair deformations and considers a flexible extended nearest neighbor model (eNN model), which reproduces very well the impact of remote neighbors in the definition of the deformability of bps. Parameters (stiffness and equilibrium values per state and shifting values between states) were derived from atomistic MD simulations using parmbsc1 force-field and state-of-the-art simulation procedures. Sampling is obtained by means of a highly efficient Metropolis Monte Carlo algorithm. The method has been implemented in a server (http://mmb.irbbarcelona.org/MCDNAlite/) which incorporates tools that, taking advantage of correlations between helical states and backbone conformation (25,27) allows the atomistic-level reconstitution of the DNA at the nucleobase and backbone level. The method produces MC ensembles that are difficult to distinguish from atomistic MD trajectories with a fraction of computational cost and reproduces well known experimental structures.

MATERIALS AND METHODS

Hamiltonian definition

A recent analysis of the dynamics of the 136 unique tetranucleotides of B-DNA performed by the ABC consortium (25) revealed that 80% of the 816 (136 × 6) unique inter base pair distributions cannot be correctly described using a single normal distribution (http://mmb.irbbarcelona.org/miniABC/ (25)). As described elsewhere (4) in many cases the peaks of the fitted normal distributions are close, and a single unimodal function can reasonably describe the real distribution. However, in 4% of the cases at least a bimodal distribution must be used to obtain a reasonable fit to the real distribution. Bimodality can be seen in slide (several tetranucleotides containing the central d(GpG) step), shift (typically in a few tetranucleotides containing d(YR) central step), and twist (very often in tetranucleotides containing central d(CG) or d(AG) steps). Certain tetranucleotides, such as d(CpTpApG) show especially complex distributions (26) impossible to describe by a single Gaussian. In summary, the normality assumption on which the harmonic model is based should be revisited for more realistic representations of DNA flexibility.

We propose here a new Hamiltonian inspired by empirical valence bond theory (28), where we assume that the distribution of inter base pair parameters (shift, slide, rise, tilt, roll, twist) underlies a Boltzmann-averaged combination of Gaussian distributions. The Hamiltonian leading to such a distribution can be derived as shown in equation (2):

$$\mathrm{E}\ \ (\mathrm{X}) = -\mathrm{k_B T} \sum_{j=1}^{N} \ln \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{e}^{-\frac{1}{\mathrm{k_B T}}\left(\frac{1}{2}\mathrm{K_{ij}}\Delta \mathrm{X_{ij}}{}^{2} + \mathrm{E_{ij}}\right)} \tag{2}$$

where $\mathrm{k_B}$ is the Boltzmann constant, T is the temperature, N is the number of bps, n is the number of states in which the distribution of inter base pair parameters of a given bps (in its sequence environment) can be decomposed (see below), K is the stiffness matrix associated to the state i in step j; ΔX is the deformation vector (with equilibrium values dependent on step j and state i) and $\mathrm{E_{ij}}$ is the relative energy of state i at bps j (shifting values between states). Note that for a single unimodal distribution equation (2) leads to the classical harmonic model shown in equation (1). Also note that due to sequence end effects single state dimer stiffness parameters are used for the first and last bps.

Definition of the states

Equation (2) implies that the energy is computed from a set of stiffness matrices and deformation vectors which are not only dependent on the step, but also on the state. In principle, if there are m states for each inter base pair distribution, we should expect $\mathrm{m^6}$ states at the bps level (i.e.

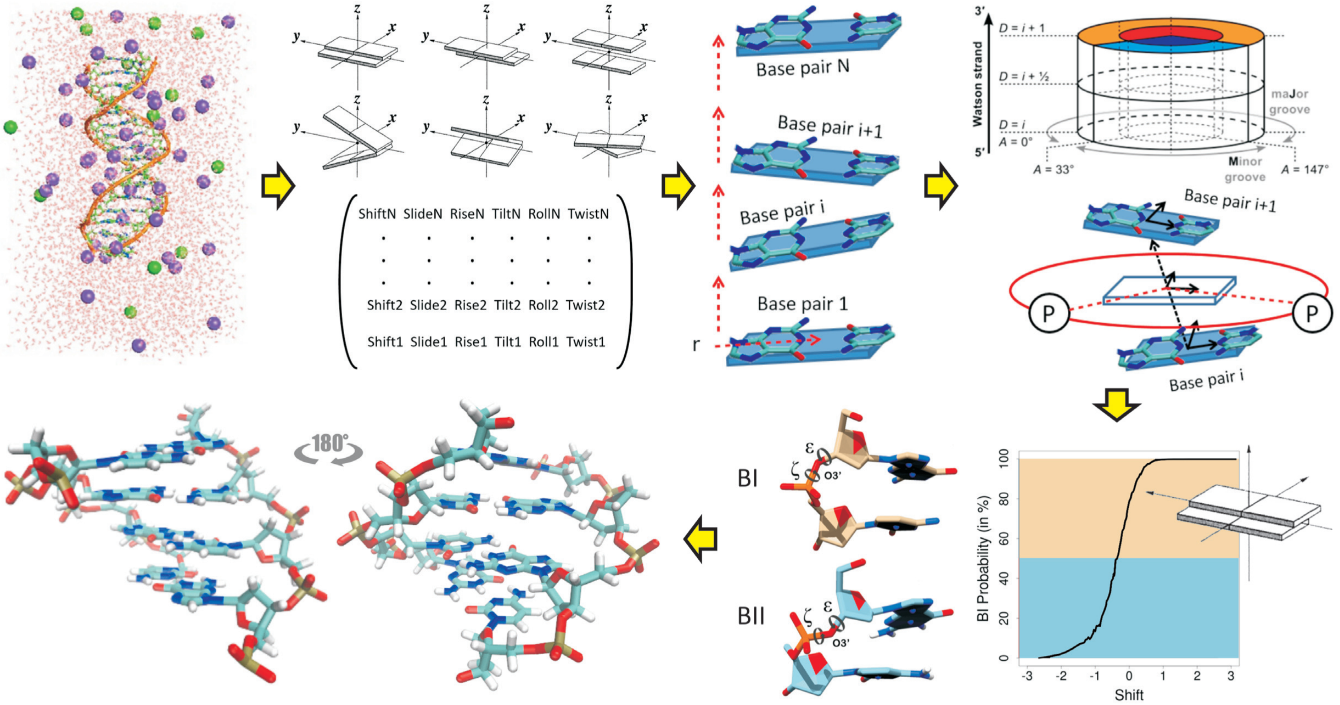

Figure 1. Workflow of the MC-eNN model. The model is parameterized by MD simulations of a sequence set of all unique 136 tetramers (see Table 1 for sequences). MC sampling in the inter base pair parameter space based on the new Hamiltonian (see equation (2)) of a structure with N+1 base pairs yields a set of 6 x N x T inter base pair coordinates (T is the number of structures sampled). For a single structure, atomistic coordinates of the nucleobases are derived using the SCHNArP algorithm (31) and the position of the phosphates relative to the helical axis using Lavery's rules (30) are determined. Using correlations of inter base pair parameters and backbone torsions the backbone states are classified to either BI or BII. For each central bps of a tetranucleotide the inter base pair coordinate showing the highest correlation with the backbone state is used as a classifier of the backbone state (see Supplementary Table S1 and 'Materials and Methods' section for more details). Average BI and BII backbone conformations for each of the 16 dimers were fit to the nucleobase position defined by the inter base pair coordinates. A short restrained steepest descent optimization relaxes mismatched local geometries resulting in the final structure (for more details see 'Materials and Methods' section).

Table 1. DNA sequences containing all 136 unique tetranucleotides used to parameterize the coarse grained model (miniABC library)

Seq. number	Watson strand (5′-3′ direction)
1	GCAACGTGCTATGGAAGC
2	GCAATAAGTACCAGGAGC
3	GCAGAAACAGCTCTGCGC
4	GCAGGCGCAAGACTGAGC
5	GCATTGGGGACACTACGC
6	GCGAACTCAAAGGTTGGC
7	GCGACCGAATGTAATTGC
8	GCGGAGGGCCGGGTGGGC
9	GCGTTAGATTAAAATTGC
10	GCTACGCGGATCGAGAGC
11	GCTGATATACGATGCAGC
12	GCTGGCATGAAGCGACGC
13	GCTTGTGACGGCTAGGGC

for bimodality $m = 2$ we could expect 64 different stiffness matrices and equilibrium vectors for each bps). Fortunately, the number of unique helical states is smaller as some inter base pair parameters are correlated and others show a purely uninormal-unimodal distribution. To assign in a systematic manner the number of states to describe a given bps we process μs-long parmbsc1 MD simulation of a large number of duplexes (see Table 1) containing the 136 unique tetranucleotides (data can be downloaded from http://mmb.irbbarcelona.org/BigNASim/ (29)). To this end, we transform the original inter base pair coordinates of the central bps of each tetranucleotide in a new set of dimensionless parameters using Lankaš transformation (30); see equation (3):

$$\gamma_i^* = \delta\gamma_i + (1-\delta)\,10.6\gamma_i \qquad (3)$$

where γ and γ^* are normal and dimensionless inter base pair parameters and δ is a Heaviside step function equal to 1 if γ is a translational parameter (measured in Å) and is equal to 0 when it is a rotational parameter (measured in degree).

Principal component analysis (PCA) is then performed to reduce the coordinate space where a certain number of components (those explaining at least 80% of variance) are kept (usually 3). Original trajectories projected in this reduced space are subjected to clustering following a Gaussian finite mixture model (31). The MD ensemble is then divided into several sub-ensembles for which the equilibrium vector (X_0) is determined. The covariance matrix in the original inter base pair parameter space is defined and inverted (22) to obtain the stiffness matrix specific for a given state of a bps in a certain tetranucleotide environment. Finally, all the harmonic models defining the global energetics of the tetranucleotide are combined by using equation (2).

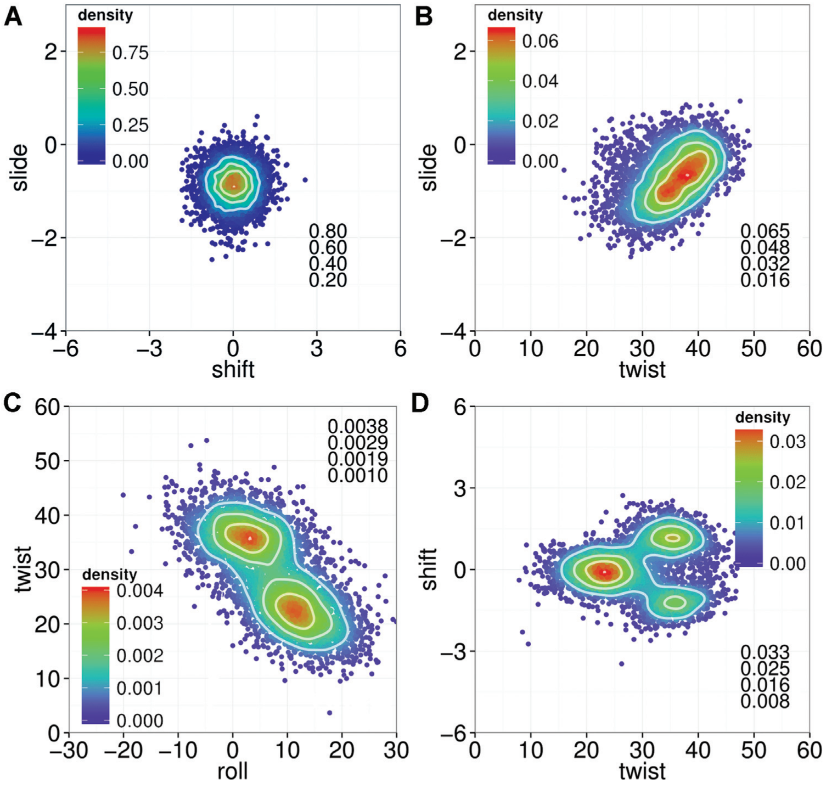

Figure 2. Examples of the four different scenarios of bi-dimensional inter base pair parameter distributions found in the BigNASim database. (**A**) Two uncorrelated and uninormal distributions show Gaussian behaviour (tetramer AATT in MD simulation with BigNASim ID 'DDD_800ns'). (**B**) Unimodal distributions which are correlated show elipsoidal shaped pattern (tetramer AAGC in MD simulation with BigNASim ID 'miniabc_K_12'). (**C**) Two hotspots appear when at least one of the two parameters contains two separate peaks (third appearance of tetramer CTAG in MD simulation with BigNASim ID 'AGCT'). (**D**) A complex multi-peaked bi-dimensional map is obtained when both inter base pair parameters are multimodal and correlated (tetramer TCGA in MD simulation with BigNASim ID 'miniabc_K_10'). The four isodensity lines equal to 100, 75, 50 and 25% of the maximum density and the corresponding values are shown in each plot.

Monte Carlo simulations

Simulation of the movements of the DNA at the CG level were performed using equation (2) (or for comparison equation (1)) implemented in a MC sampling algorithm, where movements in the inter base pair parameter space are attempted and accepted or not based on the Metropolis algorithm. For each MC move one to four inter base pair parameters are randomly selected to be modified. The strength of the change is determined by two values: a scaling factor which is dependent on the diagonal entry of the stiffness matrix of the inter base pair parameter and which is scaled to guarantee ∼40% acceptance rate. The output of an MC run is a long file of 6×N×T (N number of bps, T number of snapshots) inter base pair coordinates, which can be partially or totally transformed into Cartesian coordinates as described below. The sampling algorithm is implemented in a simple web interface (http://mmb.irbbarcelona.org/MCDNAlite) and ready to download as a stand-alone version via the web interface (http://mmb.irbbarcelona.org/MCDNAlite/standalone).

Atomic detail reconstitution

The inter base pair coordinates collected from the MC algorithm above were transformed to derive Cartesian representations of the DNA (Figure 1), as in many cases this is the level of detail required to understand DNA functionality. For a given set of inter base pair coordinates the positions of the phosphates were derived from helical axis by using Lavery's rules (see Figure 1 (32)). Atomistic coordinates of the nucleobases were derived using the SCHNArP algorithm (33), and backbone torsions were reconstituted using the correlations between inter base pair coordinates and backbone states (BI or BII) found in a recent ABC study (25). Thus, for each tetranucleotide the inter base pair coordinate showing the highest correlation with the backbone state is used as a classifier of the backbone state (typically shift; see Supplementary Table S1). The accuracy of the backbone state prediction is typically in the range of 80–90% (see Supplementary Figure S1). Average BI and BII backbone conformations for each of the 16 dimers were extracted from the meta-trajectory of all the occurrences of the dimers in a recent ABC simulation set (see Table 1) and fit to the nucleobase position defined by the inter base pair coordinates (see Figure 1). A short restrained steepest descent optimization relaxes mismatched local geometries without altering state definition. The mesoscopic MC-eNN ensemble using full atomistic reconstruction can be analyzed with any common MD analysis tool (links to NaFlex (34) are included in the web interface), which highly increases the usability of the model.

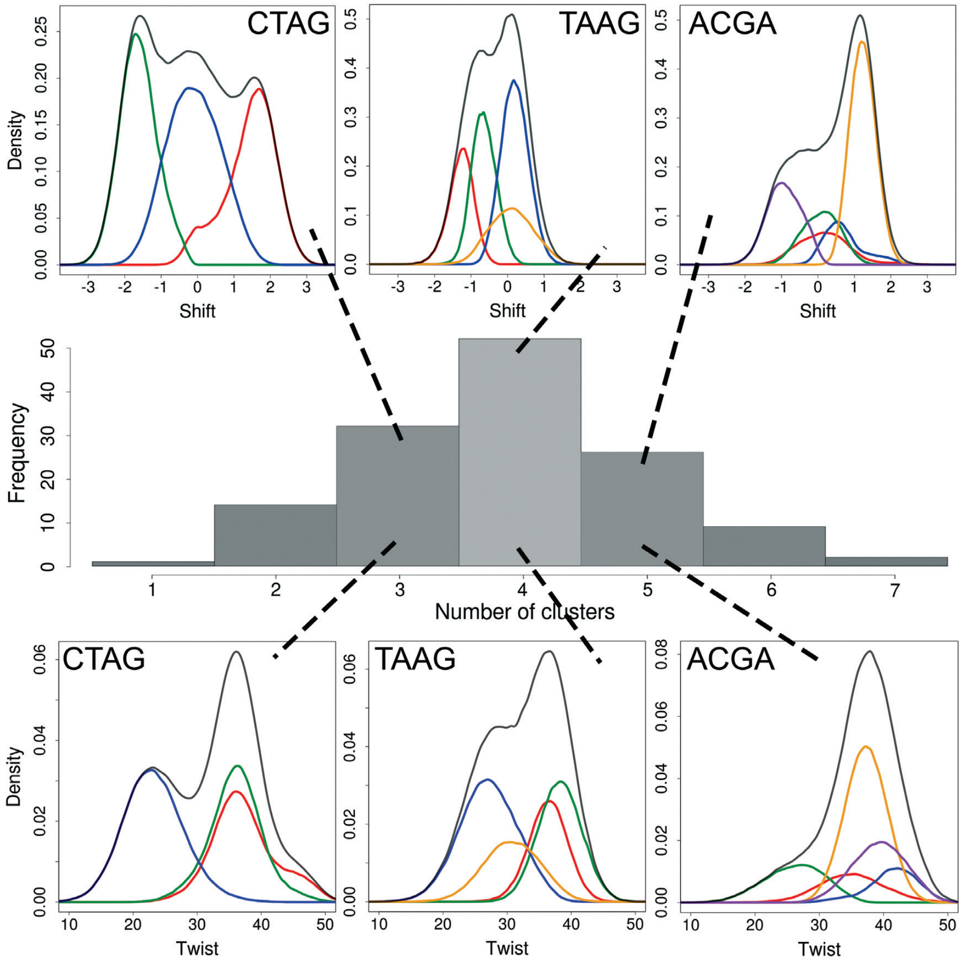

Figure 3. Histogram of the number of clusters to represent the six-dimensional inter base pair parameter space of the 136 unique tetramers (middle). Examples for the division of inter base pair parameter distributions into multiple states for the most common number of clusters are shown for Shift (top) and Twist (bottom) for the tetramers CTAG (3 clusters), TAAG (4 clusters) and ACGA (5 clusters). The inter base pair parameter distributions (gray) are clustered into several distributions shown in green, blue and red for 3 clusters; green, blue, red and orange for 4 clusters and green, blue, red, orange and purple for 5 clusters.

Data and analysis tools

Original trajectories were obtained in previous works using parmbsc1 force-field (3) and standard simulation protocols used by the ABC consortium (6) (individual simulation times at least 1 μs; data deposited at BigNASim (29) database; ID 'miniABC_K'). DNA inter base pair parameters, groove widths and backbone torsion angles were measured and analyzed with the Curves+ and Canal programs (32,35). PCA in Cartesian space was done using pcasuite (http://mmb.pcb.ub.es/software/pcasuite/pcasuite.html). Essential dynamics of simulated trajectories were obtained using the Boltzmann's averaged absolute similarity index (36). BIC (Bayesian Information Criterion) was used to determine the normal (one Gaussian) or multi-peaked nature of the distributions of inter base pair parameters (see Supplementary Methods and references (37,38)). For multi-peaked distributions we used an extension of the Helguerro's theorem (39,40) to distinguish those cases where the Gaussians are very close (unimodal) from those where they are significantly separated. Clustering was done using the mclust library (41) in R 3.1.2. The same software package was used to perform all the statistic studies and graphics in this work.

RESULTS AND DISCUSSION

The inter base pair parameter space from MD simulations

All of the 136 tetranucleotides and 80% of the 136 × 6 inter base pair distributions can be classified as multi-peaked, but only 20% of the tetranucleotides and 4% of individual inter base pair distributions are multi-modal based on Helguerro's theorem. However, these numbers mask the complexity of the coupling between inter base pair coordinates. This is illustrated by inspection of normalized bidimensional distributions (Figure 2 for examples), which show the existence of four major scenarios: (i) the inter base pair parameters are uncorrelated and show uninormal distributions leading to clear 2D Gaussian distributions, (ii) the two parameters show unimodal distributions, but are correlated leading to ellipsoidal shaped distributions, (iii)

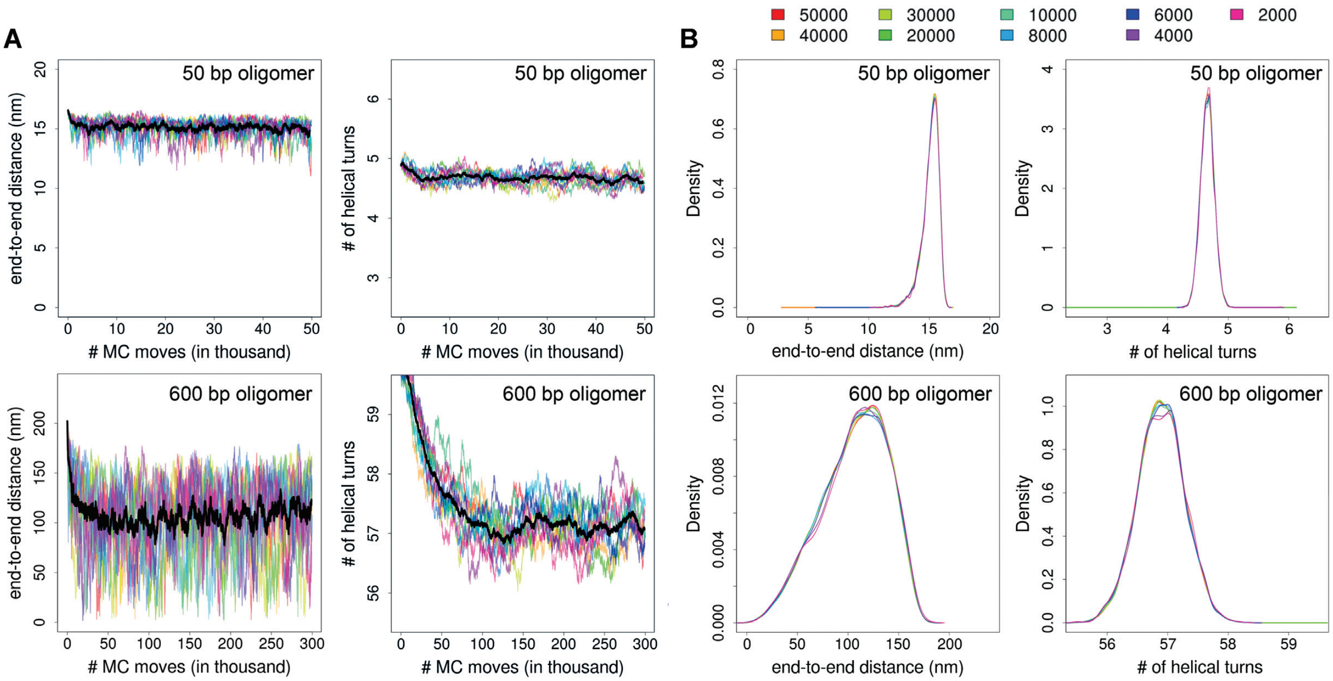

Figure 4. Equilibration and convergence of the MC-eNN simulation. (**A**) Number of MC moves needed for fiber equilibration obtained by investigating end-to-end distance (left) and number of helical turns (right) of 10 individual simulations of a fiber of random sequence of 50 bp (top) and 600 bp (bottom) in length. The 10 individual simulations are shown in different colors and a black line illustrates the average of the 10 simulations. Equilibration is obtained when the number of MC moves equals the length of the oligomer multiplied by 200 (see main text). (**B**) Convergence rules were achieved by comparing the length of the ensemble needed to obtain converged distributions. Distributions of end-to-end distance (left) and number of helical turns (right) of 2000–50 000 configurations of a fiber of random sequence of 50 bp (top) and 600 bp (bottom) in length show that a small number of configurations is sufficient for good sampling of sensitive global fiber parameters. Note: the maximum of the scale of the axis of end-to-end distance is calculated as 4 Å × fiber length (in base pair).

at least one of the two parameters is double-peaked resulting in two hotspots in the bi-dimensional map and finally, (iv) multiple peaks in two inter base pair parameters and correlation between them lead to a complex bi-dimensional probability distribution. Certainly, by moving to higher dimensions more complex probability distributions impossible to represent by combining 1D distributions would be encountered. To define unambiguously the number of states required to define the preferentially sampled regions we performed a clustering algorithm (see 'Materials and Methods' section), finding that most tetranucleotides can be represented by 3–5 clusters (Figure 3). The need to use more than five clusters is found in >10% of the cases (Figure 3), but those tetranucleotides where a single state is enough to represent the sampling are even less common.

As expected from previous studies (6,25–26), shift and twist are the main drivers for the multiplicity of states (see Supplementary Table S2). Note that no assumption on unimodality is made for the derivation of the different states, which means that an inter base pair parameter distribution of an individual state may be classified as multimodal. However, when Bayes–Helguerro's analysis is done at the state level, in only 0.8% of the clustered distributions (3192 in total) unimodality is not satisfied and overall multi-normality decreases from 80 to 20%. This means that the dimension reduction and clustering process outlined here reduces dramatically the problem of multi-normality and multimodality (see examples in Figure 3 and Supplementary Figure S2) and produces a robust protocol to define the number of states where a harmonic behavior is granted, the basic assumption required to use equation (2). A possibility to enhance the robust protocol would be to limit the maximum number of states in the clustering procedure to a lower number such as 3. On one side, this procedure would allow to produce similar quality results using less states per tetranucleotide (see Supplementary Figure S3), however ∼10% (13 out of 136) of the tetranucleotides experience multimodal behavior in at least one of the clustered inter base pair distributions (see Supplementary Figure S4) and consequently applying equation (2) would lead to a less accurate parameterization for those tetranucleotides (see Supplementary Figure S5). In any case it is very encouraging that the number or required states is much smaller than 2^6 that could be expected if bimodality was independently and universally found for all the helical coordinates.

Equilibration and convergence of Monte Carlo simulations

Before analyzing the performance of the eNN method we evaluate the expected length of the simulation required to obtain reasonably converged ensembles. To this end we performed several MC simulations (room temperature) of duplexes of random sequence and lengths ranging from 10 to 1000 bp using Arnott's fiber data to generate the starting structures. As Arnott's parameters are known to overesti-

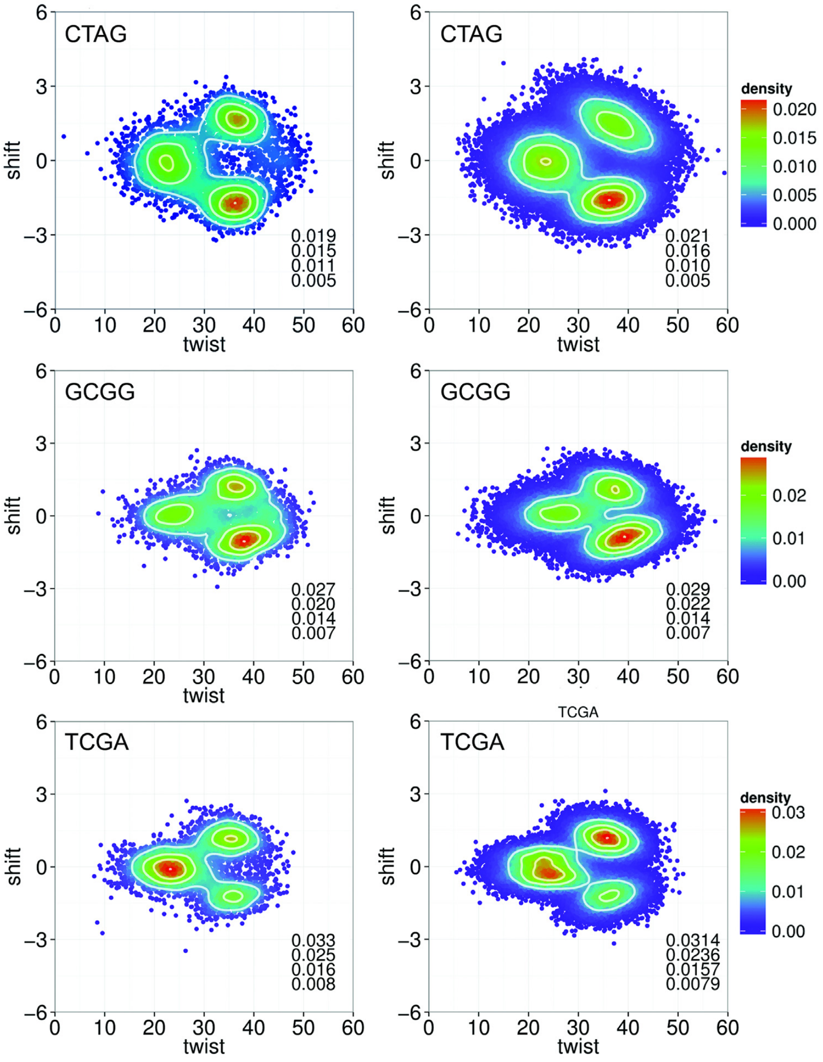

Figure 5. Bi-dimensional inter base pair parameter maps of Twist-Shift of three tetramers CTAG (top), GCGG (middle) and TCGA (bottom) of MD simulations of the parmbsc1-ABC dataset (left) and MC-eNN simulations (right) of the same sequences (see Table 1). For each tetramer there is a different color legend. The four isodensity lines equal to 100, 75, 50 and 25% of the maximum density and the corresponding isodensity values are shown in the bottom right of each plot. The bi-dimensional inter base pair parameter distributions of MD and MC-eNN simulations are indistinguishable even when correlated in a highly non-linear manner which is impossible to capture by a standard harmonic model.

mate twist by 1–2 degrees (42) we can evaluate the performance of the MC method to relax and equilibrate an incorrect structure. Results in Figure 4A (and Supplementary Figure S6) indicate that for the most sensitive parameter (the number of helical turns) equilibration is achieved when the number of collected configurations equals the length of the oligomer multiplied by 200 (for other parameters such as end-to-end distance convergence is faster, i.e. around 100 × length). Thus, for the largest oligomer considered here (1000 bp) equilibration is achieved after 100 000–200 000 MC steps. For oligomers of a size compatible with atomistic MD simulations (∼50 bp) equilibration is so fast that it is not visible in the plots (Supplementary Figure S6).

Once the rules for the equilibration time were clear we evaluated the length of the ensemble required to obtain converged distributions of local and global DNA properties. Results in Figure 4B (and Supplementary Figure S7) show that in general good sampling for sensitive global parameters such as the helical turns is obtained after a reasonably small number of configurations selected after equilibration (around 10 000–20 000 configurations). Irrespectively of the length of the duplex convergence in local geometry takes from 10 000 to 40 000 configurations depending on the complexity of the tetrad accessible inter base pair parameter space (see examples in Supplementary Figures S8 and 9). When comparison is possible, MC-convergence is faster than that obtained from MD simulations (see Supplementary Figure S9 and discussion below).

MC-eNN calculations reproduce well atomistic MD trajectories

We compare ensembles obtained for several medium-sized DNA duplexes (Supplementary Table S3) using our MC-

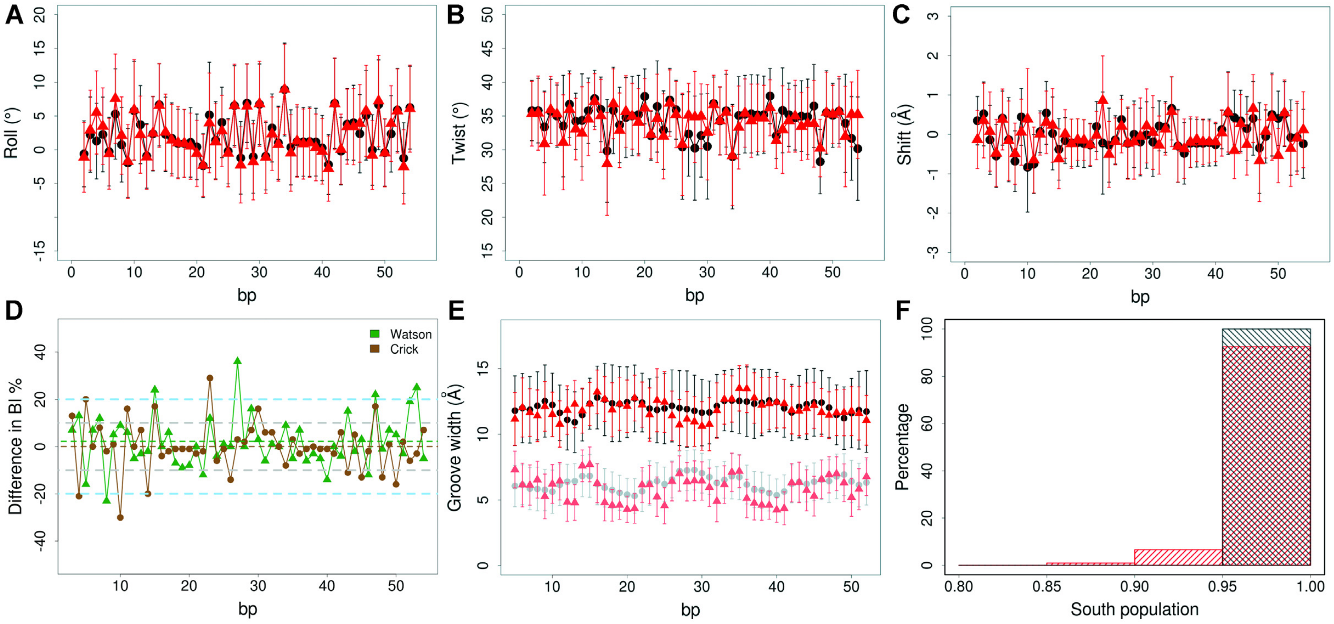

Figure 6. Comparison of MC-eNN (black) and MD simulations (red) of the longest naked DNA duplex in the BigNASim database (56 bp in length, sequence see Supplementary Table S5). (**A**) Roll distribution in degrees, (**B**) Twist distribution in degrees and (**C**) Shift distribution in Angstrom of the central 53 bps. (**D**) Difference in BI Percentage of backbone states of MC-eNN–MD in Watson (green) and Crick (brown) strand of the central 54 bp. The green and brown dashed line show the average difference in BI percentage in the Watson (2.2%) and Crick (0.1%) strand, the gray horizontal dashed lines illustrate the 10% margin corresponding to the accuracy of determining the backbone state in NMR experiments and blue horizontal dashed lines represent 20% difference in BI population similar to the average discrepancy of backbone state population estimations of MD simulations compared to NMR experiments. (**E**) Major (top, in bold) and minor (bottom, transparent) groove width. (**F**) Histogram of the population of South pucker (Phase angle of 120–210°) of all the South/North (Phase angle of 340–40°) pucker conformations of the central 54 bp. All the error bars of Figure 6A–D represent the standard deviation.

eNN protocol and 0.5–2 μs long atomistic MD simulations (using parmbsc1 force-field). Supplementary Figure S10 shows that MC and MD trajectories for the same sequence are nearly indistinguishable. Auto-similarity indexes (diagonal in Supplementary Figure S10) are always larger than cross-similarity index (for a common set of equal atoms) which indicates that the MC-eNN method reproduces very well the sequence-specific details of the deformability of DNA. Interestingly, global similarity of the standard harmonic model is only slightly lower than for MC-eNN (see Supplementary Table S4) suggesting that the influence of the accurate parameterization of MC-eNN is mostly at the local level. Indeed, local (Figure 5) inter base pair distributions obtained from MC-eNN calculations are impossible to differentiate from those derived from atomistic MD simulations, even in those cases where the inter base pair probability distributions are correlated in a highly non-linear manner, impossible to capture by a standard harmonic model (see Supplementary Figure S11). To test the limit of the method we compared MC-eNN and MD ensembles for the longest naked DNA duplex in the BigNASim database (56 bp in length, see Supplementary Table S5). The essential dynamics obtained from MC and MD samplings are nearly indistinguishable (absolute similarity index of 0.88; see Supplementary Figure S12) and the same level of agreement is found when looking to sequence-dependent inter base pair properties (Figure 6A–C and Supplementary Figure S13). In addition, even local and fine details, such as compensatory changes in neighboring steps, or the inter base pair distributions at highly structural polymorphic sites are well captured by the MC-eNN model.

The reconstitution protocol provides reasonable backbone conformations, leading to ‘atomistic’ reconstitutions that are hard to distinguish from the atomistic MD simulations. For example, for the 56-mer duplex the RMSd (using all heavy atoms as reference) of the ensemble versus the MD-averaged structure is around 0.09 Å × bp, while the RMSd increases to only 0.11 Å × bp when the MC-eNN ensemble is compared with the MD-averaged structure. Groove dimensions and many other subtle structural details such as the distribution of BI/BII states or the puckering of the sugar are well reproduced by the method (Figure 6D–F) reflected by an average difference in groove widths between MC-eNN and MD of 0.28 Å ± 0.68 and a linear correlation coefficient of 0.85 of BI population along the sequence of MC-eNN versus MD, significantly higher than when older force-fields (parmbsc0$_{\varepsilon\zeta\text{OLI}}$ and CHARMM36) were compared with nuclear magnetic resonance (NMR) experiments (43) (0.45 in average). The difference in backbone populations of MC-eNN and MD 1.1% ± 10.8 lies within the experimental accuracy of backbone state determination (43) (10%) in more than 70% of the cases, compared to 53% when older atomistic force-fields and experiments are compared (43). Both MC-eNN and MD experience a South versus North pucker population of 0.95–1.00 in over 90% of the cases with overall mean Phase angle of $P = 161° \pm 19$ for MC-eNN compared to $P = 149° \pm 30$ in MD. The accuracy of the ‘atomistic’ structures derived

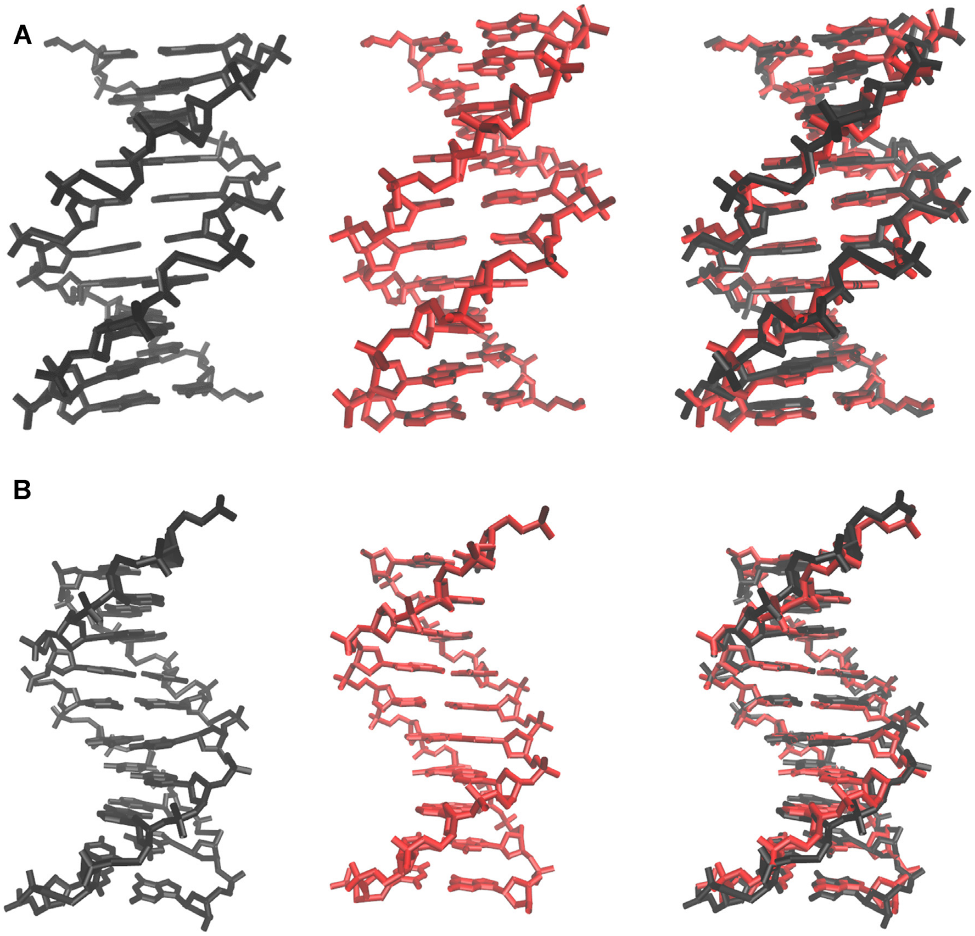

Figure 7. Comparison of the experimental structures in PDB with the reconstructed structures of MC-eNN simulations. The reconstructed MC-eNN structure (left) with the lowest RMSD to the first model of the experimental structure (middle) and the overlay of both structures (right). Only the heavy atoms without flanking base pairs are shown for better visualization. (**A**) 1ILC (resolved by X-ray; lowest RMSD to MC-eNN 0.34 Å/bp). (**B**) 424D (resolved by X-ray; lowest RMSD to MC-eNN 0.34 Å/bp). See Supplementary Table S6 for more details on the experimental structures.

from MC-eNN calculations seems to be good enough as to be used to discuss specific protein-binding to the DNA. An example of a bps where the nearest neighbor parameterization might need to be extended is CG in the GCGC context (as it appears in the 56-mer oligomer studied herein, see Figure 6). It is known that YR steps are very flexible and that especially CG experiences highly polymorphic behavior (7). In all the appearances of GCGC at bps 26, 28 and 30, the low and high twist state have different populations (see Supplementary Figure S14) probably arising from different hexamer contexts, a phenomenon already observed for other YR steps (26).

MC-eNN calculations reproduce well experimental structures

We performed an exhaustive comparison of MC-eNN ensembles with experimental (X-Ray or NMR) structures in PDB (Figures 7 and 8; Supplementary Figures S15-16 and Table S6). Our structures at T = 300 K show average RMSd around 0.3 Å × bp (using all heavy atoms as reference; see Figure 7 for examples) from the known experimental structure, a value that is close to those found in atomistic MD trajectories performed at the same temperature (see Supplementary Table S6) and not far from the RMSd generated by thermal noise (around 0.1 Å × bp, see previous section). The performance of the MC-eNN calculations is such that we can detect regions where experimental structures might need to be revisited. For example, large compensatory twist oscillations likely originated from the refinement protocol (1DN9, 1HQ7 in Supplementary Figure S16D and E), or regions where anomalous inter base pair parameter values (low Roll in last bps and very high twist in bps 4 for PDB id 2JYK in Figure 8B) occur.

MC-eNN performs well compared to other coarse grained models

When comparing the performance of MC-eNN with the standard harmonic approach and other coarse-grain models such as cgDNA we find good agreement in the average properties of inter base pair parameter distributions among the methods (see Supplementary Figures S17–19), however complex local features are only captured by MC-eNN (see Supplementary Figures S11 and 20). Another global measure of DNA flexibility is the persistence length. Using the

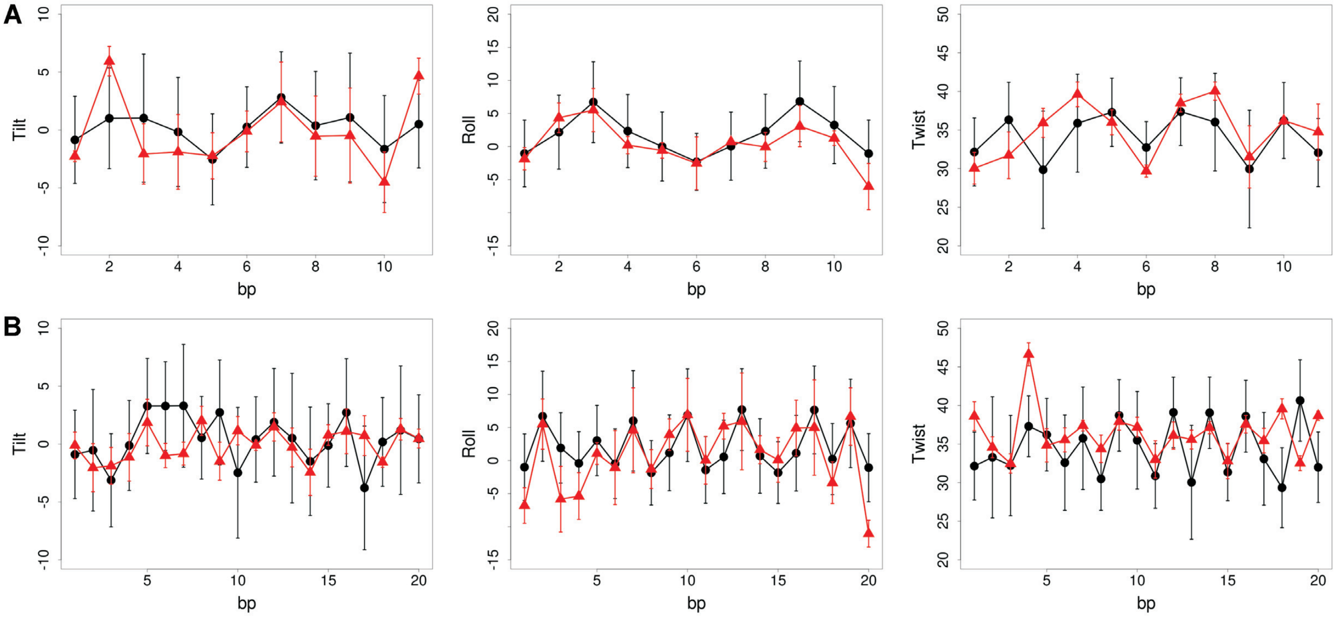

Figure 8. Comparison of the rotational inter base pair parameter distributions Tilt (left), Roll (middle) and Twist (right) of MC-eNN simulations (black) with experimental structures in PDB (red). Error bars represent the standard deviation of the MC-eNN simulation or the different models of the experiment, respectively. (**A**) PDB ID: 1ILC (12 bp, resolved by X-ray). (**B**) PDB ID: 2JYK (21 bp, resolved by NMR). The translational inter base pair parameter distributions are compared in Supplementary Figure S9 and more examples are depicted in Supplementary Figure S10 (see Supplementary Table S6 for more details on the experimental structures).

well-established procedure of calculating dynamic and apparent persistence length (Equations 6 and 9 in (44)), MC-eNN captured the same sequence-dependent persistence length pattern reported with cgDNA simulations (see Supplementary Figure S21 and Table S7), although with lower absolute values most likely due to the different definition of the tangent (see Supplementary Data).

Computational performance

The MC-eNN method is very efficient from a computational point of view. To obtain converged complex inter base pair distributions (see Figure 4B; Supplementary Figures S6 and 8) atomistic MD simulation of a 56-mer duplex (∼550 000 atoms) would require more than 500 days in one of our 64-core cluster (400 ns of trajectory), while to obtain equivalent sampling (as determined from the convergence rate) would require only 12 min in the same machine using the MC-eNN method outperforming MD by a factor of $\sim 10^5$ (see Supplementary Figure S22). The difference in computer performance between MD and MC-eNN calculations increases for larger duplexes, as the cost of MD simulations scales with the third power of the length of the DNA, while MC-eNN simulation time increases only linearly with the length of the duplex. Furthermore, contrary to atomistic MD, MC-eNN scales perfectly with the number of processors, which facilitates its use in supercomputers.

The MC-eNN web server

The MC-eNN simulation method is distributed as a stand-alone executable version for MacOS and Linux systems (see Supplementary Information; source code is available upon request), but it is also accessible as a web server http://mmb.irbbarcelona.org/MCDNAlite/ (the stand-alone version can be downloaded via the web server http://mmb.irbbarcelona.org/MCDNAlite/standalone) which requires just the sequence of the duplex as input and provides as output a limited number of alternative conformations, selected to capture the most probable configurations according to the states at tetranucleotide level. All results can be viewed directly in the web interface and downloaded for further local analysis. A direct link in the webserver to our NAFlex tool (34) constitutes a user-friendly way for deeper online analysis of the DNA structures.

CONCLUSION

We present a new mesoscopic model for the representation of the structure and dynamics of naked DNA structures, which integrates all the information acquired from the analysis of B-DNA dynamics from the latest efforts published by the ABC consortium. The method maintains the simple bps model, but tackles rigorously the multi-modality of inter base pair distributions and their dependence on nearest neighbors, allowing an accurate representation of complex polymorphisms in DNA. The mesoscopic ensembles provided by our algorithm can be transformed to atomistic models of DNA with a high accuracy even in local details, something beyond the expectations of a mesoscopic model. The method is extremely efficient, making it possible to simulate long fibers of DNA that will be unreachable for atomistic MD simulation in the next decades. It is implemented in simple tools that can be used by non-experts aiming to obtain a more complete picture of DNA than that derived from the inspection of canonical average structures.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the ABC consortium for the atomistic MD simulations used to derive the parameters and all the colleagues there for many discussions on the simulation engine. We are also indebted to Prof. R. Lavery for Curves+ and for his phosphate-location approach. We also thank Dr F. Battistini for providing MD simulations to benchmark the MC-eNN method and for inspiring discussions.

Author contributions: M.O., P.D.D. and J.W. designed the method. J.W. developed and coded the method and did all the analysis with support of P.D.D. A.B. provided average backbone states of all nucleobases. J.W. and A.H. developed the backend of the webserver while G.B. was responsible for setting up the web page. J.W., P.D.D. and M.O. discussed the analysis and wrote the manuscript with contributions from all the co-authors. M.O. and P.D.D. directed the work.

FUNDING

Spanish Ministry of Science [BFU2014-61670-EXP, BFU2014-52864-R]; Catalan SGR, the Instituto Nacional de Bioinformática; European Research Council (ERC SimDNA), European Union's Horizon 2020 Research and Innovation Program [676556]; Biomolecular and Bioinformatics Resources Platform co-funded by the Fondo Europeo de Desarrollo Regional (FEDER) [ISCIII PT 13/0001/0030 to M.O.]; MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona); La Caixa PhD Fellowship (UB and IRB Barcelona, Spain) (to J.W., A.B.); PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) (to P.D.D.); SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) (to P.D.D.); ICREA (Institució Catalana de Recerca i Estudis Avançats) (to M.O.). Funding for open access charge: European Union's Horizon 2020 Research and Innovation Program [676556].

Conflict of interest statement. None declared.

REFERENCES

1. Dans,P.D., Walther,J., Gómez,H. and Orozco,M. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29–45.
2. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
3. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
4. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
5. Balaceanu,A., Pasi,M., Dans,P.D., Hospital,A., Lavery,R. and Orozco,M. (2017) The role of unconventional hydrogen bonds in determining BII propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**, 21–28.
6. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
7. Dans,P.D., Faustino,I., Battistini,F., Zakrzewska,K., Lavery,R. and Orozco,M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320.
8. Dans,P.D., Ivani,I., Hospital,A., Portella,G., González,C. and Orozco,M. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.*, **45**, 4217–4230.
9. Dans,P.D., Danilāne,L., Ivani,I., Dršata,T., Lankaš,F., Hospital,A., Walther,J., Pujagut,R.I., Battistini,F., Gelpí,J.L. *et al.* (2016) Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.*, **44**, 4052–4066.
10. Balaceanu,A., Pérez,A., Dans,P.D. and Orozco,M. (2018) Allosterism and signal transfer in DNA. *Nucleic Acids Res.*, **46**, 7554–7565.
11. Dršata,T. and Lankaš,F. (2015) Multiscale modelling of DNA mechanics. *J. Phys. Condens. Matter*, **27**, 323102.
12. Orozco,M., Noy,A. and Pérez,A. (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, **18**, 185–193.
13. Orozco,M., Pérez,A., Noy,A. and Luque,F.J. (2003) Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.*, **32**, 350–364.
14. Gómez,H., Walther,J., Darré,L., Ivani,I., Dans,P.D. and Orozco,M. (2017) *Computational Tools for Chemical Biology*. In: Martín-Santamaría,S (ed). Molecular Modelling of Nucleic Acids. Royal Society of Chemistry, London, UK. pp. 165–197.
15. Dans,P.D., Zeida,A., Machado,M.R. and Pantano,S. (2010) A coarse grained model for Atomic-Detailed DNA simulations with explicit electrostatics. *J. Chem. Theory Comput.*, **6**, 1711–1725.
16. Ouldridge,T.E., Šulc,P., Romano,F., Doye,J.P.K. and Louis,A.a. (2013) DNA hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic Acids Res.*, **41**, 8886–8895.
17. Freeman,G.S., Hinckley,D.M., Lequieu,J.P., Whitmer,J.K. and de Pablo,J.J. (2014) Coarse-grained modeling of DNA curvature. *J. Chem. Phys.*, **141**, 165103.
18. Petkevičiūtė,D., Pasi,M., Gonzalez,O. and 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**, e153.
19. Dršata,T., Zgarbová,M., Špačková,N., Jurečka,P., Šponer,J. and Lankaš,F. (2014) Mechanical model of DNA allostery. *J. Phys. Chem. Lett.*, **5**, 3831–3835.
20. Lankaš,F., Gonzalez,O., Heffler,L.M., Stoll,G., Moakher,M. and Maddocks,J.H. (2009) On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, **11**, 10565–10588.
21. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–11168.
22. Lankaš,F., Šponer,J., Langowski,J. and Cheatham,T.E. (2003) DNA Basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.*, **85**, 2872–2883.
23. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, **38**, 299–313.
24. Dixit,S.B., Beveridge,D.L., Case,D.a., Cheatham,T.E., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,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**, 3721–3740.
25. Dans,P.D., Balaceanu,A., Pasi,M., Patelli,A.S., Petkevičiūtė,D., Walther,J., Hospital,A., Lavery,R., Maddocks,J.H. and Orozco,M. (2019) The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res.*, **47**, 11090–11102.
26. Balaceanu,A., Buitrago,D., Walther,J., Hospital,A., Dans,P.D. and Orozco,M. (2019) Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. *Nucleic Acids Res.*, **47**, 4418–4430.
27. Zgarbová,M., Jurečka,P., Lankaš,F., Cheatham,T.E., Šponer,J. and Otyepka,M. (2017) Influence of BII backbone substates on DNA

Twist: a unified view and comparison of simulation and experiment for all 136 distinct tetranucleotide sequences. *J. Chem. Inf. Model.*, **57**, 275–287.

28. Colizzi,F. and Bussi,G. (2012) RNA unwinding from reweighted pulling simulations. *J. Am. Chem. Soc.*, **134**, 5173–5179.
29. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Gõni,R., Orozco,M. *et al.* (2016) BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
30. Dršata,T., Pérez,A., Orozco,M., Morozov,A.V, Sponer,J. and Lankaš,F. (2013) Structure, stiffness and substates of the Dickerson-Drew dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.
31. DAY,N.E. (1969) Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.
32. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2412–2423.
33. Lu,X.J., El Hassan,M.A. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: Rebuilding program (SCHNArP). *J. Mol. Biol.*, **273**, 681–691.
34. Hospital,A., Faustino,I., Collepardo-Guevara,R., González,C., Gelpí,J.L. and Orozco,M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.*, **41**, W47–W55.
35. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
36. Pérez,A., Blas,J.R., Rueda,M., López-Bes,J.M., de la Cruz,X. and Orozco,M. (2005) Exploring the essential dynamics of B-DNA. *J. Chem. Theory Comput.*, **1**, 790–800.
37. Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.
38. Kass,R.E. and Raftery,A.E. (1995) Bayes Factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
39. Schilling,M.F., Watkins,A.E. and Watkins,W. (2002) Is human height bimodal? *Am. Stat.*, **56**, 223–229.
40. Helguero,F. (1904) Sui massimi delle curve dimorfiche. *Biometrika*, **3**, 85–98.
41. Fraley,C., Raftery,A.E., Scrucca,L., Brendan Murphy,T. and Fop,M. (2016) Gaussian mixture modelling for Model-Based clustering,classification, and density estimation [R package mclust version 5.4.2]. *R J.*, **8**, 205–233.
42. Arnott,S. and Hukins,D.W.L. (1972) Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.*, **47**, 1504–1509.
43. Ben Imeddourene,A., Elbahnsi,A., Guéroult,M., Oguey,C., Foloppe,N. and Hartmann,B. (2015) Simulations meet experiment to reveal new insights into DNA intrinsic mechanics. *PLOS Comput. Biol.*, **11**, e1004631.
44. Mitchell,J.S., Glowacki,J., Grandchamp,A.E., Manning,R.S. and Maddocks,J.H. (2017) Sequence-dependent persistence lengths of DNA. *J. Chem. Theory Comput.*, **13**, 1539–1555.