

CGeNArate: a sequence-dependent coarse-grained model of DNA for accurate atomistic MD simulations of kb-long duplexes

David Farré-Gil¹, Juan Pablo Arcon¹, Charles A. Laughton² and Modesto Orozco^{1,3,*}

¹Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, E-08028 Barcelona, Spain

²School of Pharmacy and Biodiscovery Institute, University of Nottingham, University Park, Nottingham NG7 2RD, UK

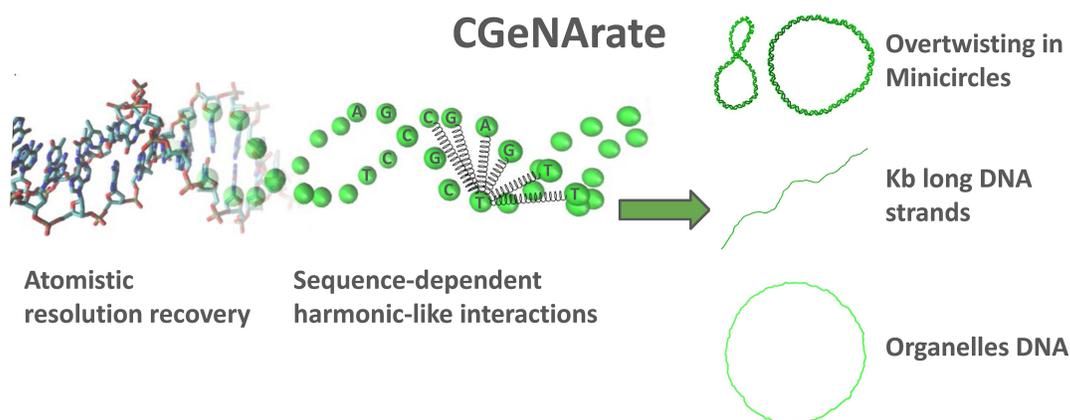
³Department of Biochemistry and Biomedicine, University of Barcelona, E-08028 Barcelona, Spain

*To whom correspondence should be addressed. Tel: +34 93 40 37156; Fax: +35 93 403 7157; Email: modesto.orozco@irbbarcelona.org

Abstract

We present CGeNArate, a new model for molecular dynamics simulations of very long segments of B-DNA in the context of biotechnological or chromatin studies. The developed method uses a coarse-grained Hamiltonian with trajectories that are back-mapped to the atomistic resolution level with extreme accuracy by means of Machine Learning Approaches. The method is sequence-dependent and reproduces very well not only local, but also global physical properties of DNA. The efficiency of the method allows us to recover with a reduced computational effort high-quality atomic-resolution ensembles of segments containing many kilobases of DNA, entering into the gene range or even the entire DNA of certain cellular organelles.

Graphical abstract



Introduction

DNA has been both a topic of interest and a challenge for theoreticians who faced the formidable problem of simulating a multiscale system ranging from the base-pair (bp; Å-scale) to the meter-long chromatin fiber of developed organisms (1). At the highest level of resolution, quantum mechanics (QM) theory provides electronic details of small DNA segments (2,3), but most atomistic information of DNA is obtained from the use of molecular dynamics (MD) coupled to classical force-fields (FFs). Last generation DNA FFs (4–6) have achieved a level of accuracy comparable with that of experiments (7), and more impressively, have shown predictive power in a variety of systems, even far from biological conditions (8–12). However, and despite their success, atomistic simulations are limited in

the size of the systems to be studied, as the total number of particles to be simulated scales roughly with the third power of the length of the duplex, making in practice impossible to simulate duplexes longer than c.a. 70–100 bp.

Alternatives to atomistic methods aim to reduce the cost of the calculation by using simplified solvent models, merging groups of atoms into beads and representing a Hamiltonian by very simple terms. Two families of approaches have emerged from these ideas: (i) mesoscopic models and (ii) Cartesian coarse-grained approaches. The mesoscopic models take advantage of a helical coordinate system that is the natural one to describe a DNA duplex. In the simplest version, bp step (bps) movements are described as 3 translational (rise, slide, shift) and three rotational (twist, roll, tilt)

Received: June 20, 2023. Revised: May 1, 2024. Editorial Decision: May 2, 2024. Accepted: May 14, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

degrees of freedom, and the energy is computed by using (bps) local harmonic Hamiltonians (13,14), which were fitted to a diverse set of experimentally-determined bps geometries. Second-generation models follow the same physical approaches, but were fitted to atomistic MD simulations (15) which allowed them to be parametrized for all the unique tetramers (i.e. three consecutive bps (16–18)). The latest versions have been extended to capture non-local effects (19–22), non-harmonic deviations (17), and even base pair distortions (20,21,23). All these mesoscopic methods are accurate and computationally efficient, which allow the simulation of medium sized chromatin fibers (24–26). However, they present some intrinsic caveats: (i) no backbone information is directly available from the ensembles; (ii) the mesoscopic methods do not couple well with MD algorithms; finally, (iii) non-bonded terms required to simulate long DNA duplexes are difficult to introduce. Recent approaches based on learning the connection between helical coordinates and backbone geometry in atomistic MD simulations (17) have partially solved the first problem (27), but facing the other two would require a very important development effort, with little guarantee that the resulting method will be still computationally efficient.

Cartesian Coarse-Grained (cCG) models simplify the DNA representation by grouping atoms into beads, whose interactions are treated by simple equations adapted to trace the most usual deformations of DNA. The solvent environment is largely simplified, and sampling is obtained through MD simulations. Broadly speaking, the myriad of cCG models available (reviewed in (1)) can be classified based on: (i) the energy functional, (ii) the number of beads per nucleotide, (iii) the way in which they account for solvent and (iv) the type of strategy used to refine the method. The energy functional can be very different considering the number of beads, the solvent model, and whether cCG is designed to capture near-equilibrium or large denaturing transitions. The number of beads per nucleotide is also very variable: for example, from just 1 bead in Vercauteren's model (28) or MRG-CG (29), 2 in OxDNA (30,31) or Aksimentiev's models (32), 3 of de 3SPN (33), MAD_{na} (34) or BioModi (35) and up to 6–8 beads of high resolution models such as SiRAH (36,37), MARTINI (38), UNRES (39) or HiRe-DNA (40). Despite the reduction in resolution and the large size of the beads, most DNA cCG studies tackle only medium sized (<10² bp) duplexes (1). The treatment of water and ions can be done explicitly (like in MARTINI or SiRAH models) or by means of a continuum model (like in HiRe-DNA or 3SPN, and some versions of SiRAH). Finally, the fitting of functional can follow two main paradigms (which can be combined (1)): (i) the top-down, exemplified by OxDNA or SiRAH, where parameters are refined to reproduce some macroscopic experimental observable, (ii) bottom-up, followed among others by Vercauteren's group or MAD_{na} developers who used atomistic MD simulations as reference. The top-down refinement guarantees accurate average polymer properties, but the lack of enough experimental reference data precludes careful consideration of sequence effects, and no guarantee exists on the accuracy of short-scale details. On the contrary, the bottom-up approach leads to energy functionals that can capture well short-scale details and sequence-dependent effects, but they rely on force-fields whose ability to reproduce polymer properties is not always granted. In summary, there is a plethora of methods available, and the end user should make a careful selection based, mainly, on the nature of the problem.

For biological applications the main challenge of these methods is to reproduce very long segments of DNAs (above kbase), with accurate sequence specificity. Sampling must be fast, but if required, the full atomistic description should be recoverable, allowing detailed representation of DNA interactions. Here we present CGeNArate, a new cCG method created to explore the dynamics of long segments of DNA, approaching those of interest for the representation of chromatin. The method uses implicit solvent, only 1 bead per nucleotide, a simple energy functional including up to 4th order bonded terms, coupled with simple long-range electrostatic and steric functionals. The method, implemented in a 'de novo' MD code, can easily manage very long oligomers (above kb scale) and has been parametrized from all-atom MD simulations following a bottom-up approach with a tetramer-level sequence specificity, but taking also into consideration global properties obtained from simulations of long oligomers. It shows an unexpected ability to reproduce mechanical and dynamical properties of a variety of oligomers which were not considered during the parametrization, including circular DNAs, kb long duplexes and even entire mitochondrial DNA. Additionally, the use of a novel machine learning (ML) approach trained with a large dataset of atomistic MD simulations, allows us to map with astonishing accuracy the cCG trajectories into atomistic ensembles for part or the entire duplex. We expect CGeNArate will become a valuable tool to describe segments of the chromatin fiber, even substituting state-of-the-art mesoscopic models(17,19,20,23).

Materials and methods

CGeNArate is intended to simulate duplex DNA, not extremely far from the equilibrium geometry (as it happens in chromatin). The method uses 1 bead per nucleotide located at the C1' atom position of the sugar, which facilitates the Machine-Learning back-mapping to the atomistic level and allows a reasonable description of DNA shape.

Hamiltonian definition

The energy functional is defined as the addition of sequential-dependent (bonded) and remote (non-bonded) terms as described in Eq. ((1)):

$$E = E_{seq} + E_{remote} \quad (1)$$

Following Savelyev and Papoian (41), the sequential contribution (E_{seq}) is computed considering 11-bead windows (Figure 1). This means that each bead i in the Watson strand (Figure 1) interacts with its neighboring beads ($i + 1$ and $i - 1$), with its paired bead (j) in the Crick strand, as well as with 5 beads upstream and 5 beads downstream the Crick's paired bead (i.e. $j + 1$ to $j + 5$ in one direction and $j - 1$ to $j - 5$ in the other), and the $i + 2i - 2$ bead through angle-dependent interactions (see below). The sequential term is divided into two contributions: one is tetramer dependent ($E_{seq-4mer}$) and is calibrated from atomistic MD simulations considering sequence-dependent properties of DNA (see 'Fitting the Hamiltonian' section), and the other, which accounts for distant interactions in the 11-bead window ($E_{seq-distant}$), is calibrated with sequence-averaged dynamic information of DNA (see Figure 1); Eq. ((2)):

$$E_{seq} = E_{seq-4mer} + E_{seq-distant} \quad (2)$$

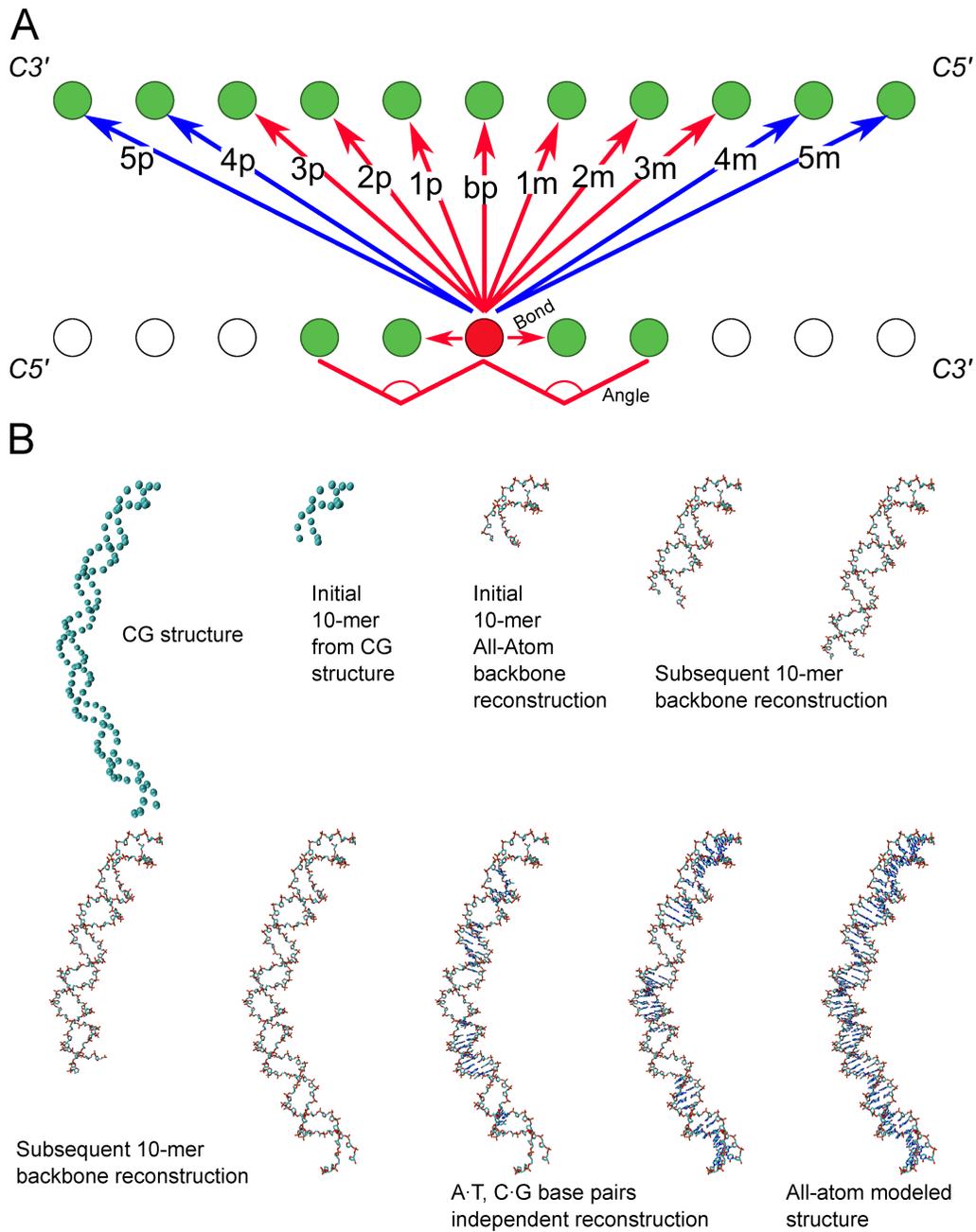


Figure 1. Methods (A) outline of the seq terms used in the model. Blue arrows correspond to distant terms, while red arrows represent 4mer terms. (B) Step by step evolution of the All-Atom reconstruction process from the CG duplex.

Note that interactions that are present in two tetramers are obtained by averaging between them (e.g. the angle interaction between beads $i - 1, i, i + 1$ is shared by the tetramers $i - 2, i - 1, i, i + 1$ and $i - 1, i, i + 1, i + 2$, see below).

Following again Savelyev and Papoian (41), we consider (Figure 1) *stacking* interactions (i.e. $i : i + 1$ and $j : j - 1$) and *angle* interactions ($i : i + 1 : i + 2$ and $j : j - 1 : j - 2$), affecting both Watson and Crick strands and the cross-interactions: *pairing* ($i : j$) and *fan* ($i : j - 1, \dots, i : j - 5$ and $i : j + 1, \dots, i : j + 5$). Each interaction is represented by a truncated polynomial expansion. The 2nd order term introduces a basal harmonicity which is modulated by the 3rd order term, while the 4th order term avoids large unrealistic distortions that might

happen under stress conditions; see Eq. ((3),(4)).

$$E_{stacking, pairing, fan} = \sum_{a=2}^4 K_a (l - l_0)^a \quad (3)$$

$$E_{angle} = \sum_{a=2}^4 K_a (\alpha - \alpha_0)^a \quad (4)$$

where K_α are the force constants of the interaction, l_0 is the equilibrium distance between beads and α_0 is the equilibrium angle (see ‘Fitting the Hamiltonian’ section for details on how they are derived). Note that, by construction, large deformations leading to base opening or kinks in the fiber are not al-

lowed, but further versions of the method, where harmonic terms would be substituted by Morse-like potentials could account for this type of extreme deformation.

Thus, $E_{seq-4mer}$ is determined as shown in Figure 1 and Eq. (5):

$$E_{seq-4mer} = \sum_{Tetramer} \left(\sum^2 E_{stacking} + \sum^2 E_{pairing} + \sum^4 E_{angle} + \sum^8 E_{fan} \right) \quad (5)$$

with the sequential long-term interaction defined as the remaining sequential interactions (see Figure 1), as described in a compacted form in (6):

$$E_{seq-distant} = \sum_{index = -5, -4, 4, 5} \left(\sum_i E_{fan} \right) \quad (6)$$

The remote term is divided as follows:

$$E_{remote} = E_{LJ} + E_{ele} \quad (7)$$

For further details on the remote terms, see Supplementary Methods (Remote term specifications). Note that to avoid double counting interactions, the ‘remote contribution’ is switched off for interactions between the neighboring beads within 5 bp in both strands (see Figure 1). Note also that alternative formalisms can be implemented to account for intermolecular interactions involving charged polyelectrolytes (42).

Fitting the Hamiltonian

The sequential tetramer (seq-4mer) parameters were refined in an iterative manner, taking equilibrium distances and angles from atomistic MD simulations. A first set of sequence-independent parameters was obtained by fitting force-constants in Eq. (4) to reproduce as close as possible variances and covariance of the different distances and angles included in the Hamiltonian definition. We perform then a 1st tetramer-based parametrization, where the *seq-4mer* terms are refined sequentially for each tetramer in the context of initial guesses for the parameters of the remaining tetramers (average parameter of all tetramers). Once 2nd iteration parameters are obtained for all the tetramers in the sequence, the process is repeated until convergence is achieved (typically, 3–4 iterations are required for convergence). Data used for fitting was obtained from the thirteen 18-mer duplexes of the miniABC dataset (16) stored in the BigNASim database (mmb.irbbarcelona.org/BIGNASim (43)). Parameters were refined using the constrained optimization by linear approximation (COBYLA) method (44), which allowed us to avoid overtraining artifacts that would lead to physically unrealistic parameters. (see Supplementary Methods Parameter fitting at tetranucleotide level). The overall optimization process leads to a significant improvement in the overlap between the distribution of CG and AA observables (see Supplementary Figure S1)

The *seq-distant* terms ($i \rightarrow j \pm 4$ and $i \rightarrow j \pm 5$) were obtained initially from equilibrium values of atomistic MD simulations, and then re-adjusted by fitting a 40-mer duplex. Standard annealing procedures increasing and decreasing the ‘effective temperature’ was used to refine the associated parameters, accordingly Metropolis-Hasting simulations with different seeds

were computed and the best of the sampled sets were refined by conjugate gradient minimizations. The end-to-end distance and the associated variance were used as merit functions in the fitting. The remote non-bonded electrostatic term (Eqs. (8) and (9)) was determined considering $q = -1$ in each bead, and dielectric and inverse distance parameters corresponding to 100 mM NaCl aqueous solution. Following previous works (17,45) Lenard Jones parameters σ and ϵ_{LJ} were set to 10 Å and 0.59 kcal/mol.

Integration of the equations of motion

The Hamiltonian above has been implemented in a *de novo* Langevin Dynamics code. Integration of the equations of motions was performed using velocity Verlet with an integration step of 0.1 ps, which guarantees stability in the trajectory for temperatures up to 500 K (see Supplementary Figure S2). The masses of the beads correspond to those of the nucleotides. Temperature was maintained constant using Langevin bath with standard coupling parameters (46). Friction terms corresponding to those of Brownian stochastic forces were generated following a Box–Muller transformation (47). As noted in Supplementary Figure S2, the method is quite robust to small perturbation of these parameters.

All atom rebuilding

The 1-bead per nucleotide trajectories were back-projected to all-atom resolution by using a machine learning (ML) approach that takes the CG coordinates through time and the sequence of the duplex as descriptors. The method is developed from the GLIMPS (27) approach, which was originally created to rebuild atomistic structures from mesoscopic descriptors. As shown in Figure 1, atomic resolution back-projection was done in two steps: (i) backbone reconstitution and (ii) generation of the A·T/T·A and G·C/C·G geometries. In all cases, training was done for all the 10-mer contained in 40-mer trajectories deposited in our BigNASim database (43). The all (heavy) atom Cartesian coordinates were the objective values, while the C1’ CG coordinates (from the same atomistic simulations) and the duplex sequence were the descriptors.

Once training is done, all atoms reconstitution is performed using in general 10-mer blocks. As described above, we first use one GLIMPS model to rebuild the sequence-neutral backbone atom positions for the decamer segment from the C1’-atom positions, then use a second, base-pair dependent, GLIMPS model at each of the 10 base pair steps to rebuild the base atom positions from the now-established sugar and phosphate positions. Finally, as some local distortions appear, a short steepest descent minimization is performed using the ParmBSC1 force-field. Structural changes introduced by these geometry optimization steps are small (typically tenths of Å), but steric clashes that might appear during the coarse-grained to all-atom decoding are removed (see Figure S3).

Validation analysis

All validation analysis was performed using different sequences to those used for the training of the method and using when possible both MD-derived and experimental values. Details of the different sequences and metrics used for validating the ensembles are detailed in Supplementary Material.

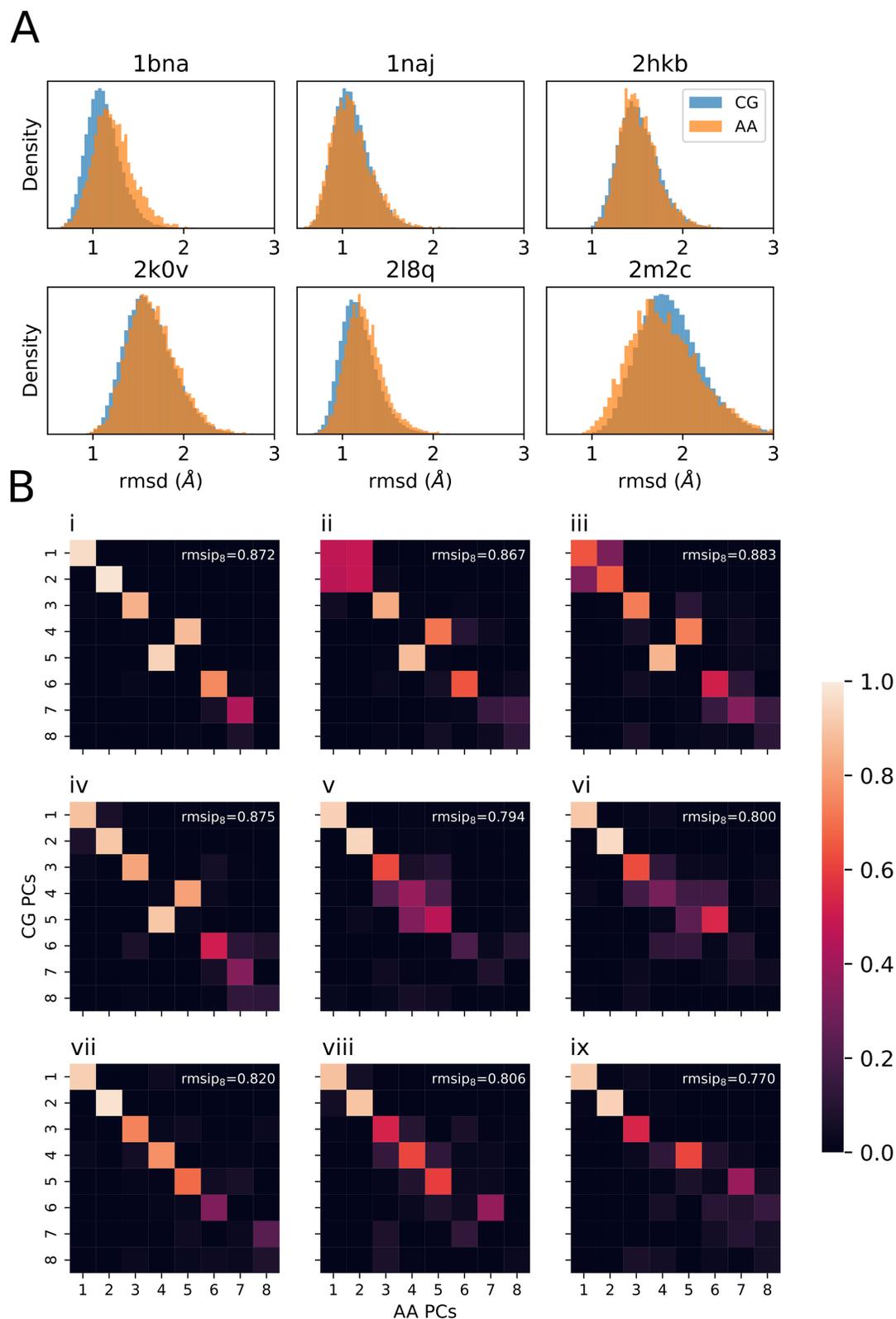


Figure 2. Evaluation metrics of the Coarse-Grained trajectories: **(A)** rmsd distribution between C1' atoms of simulated Coarse-Grained (CG, blue) and all-atom (AA, orange) trajectories against the pdb experimental reference structure indicated in each panel. **(B)** Root mean square inner product (rmsip) between CG and AA principal components of trajectories from the test set, ordered by decreasing variance. Each panel corresponds to simulations of the following structures (see Materials and methods). A: BigNASim Code CGTG, B: BigNASim Code AGCT, C: 1zgw, D: BigNASim Code AGCG, E: BigNASim Code CTAG_flex, F: 2lef, G: 1j5n, H: 2m2c, I: 1naj.

Table 1. Global deformation properties of the 13 sequences of the miniABC dataset (16)

	1	2	3	4	5	6	7	8	9	10	11	12	13
RMSIP AA to CG	0.87	0.87	0.86	0.89	0.86	0.87	0.88	0.86	0.88	0.88	0.88	0.87	0.87
PL AA	58	58	62	67	61	72	66	64	68	56	57	60	60
PL CG	50	50	59	52	55	64	67	56	62	52	49	54	54
EtE AA	56	56	57	56	57	57	57	56	57	57	57	57	56
Std	1.6	1.7	1.6	1.6	1.6	1.6	1.5	1.6	1.5	1.6	1.8	2.0	1.6
EtE CG	58	57	59	58	59	58	59	58	58	58	58	59	58
std	2.0	2.0	1.8	1.9	1.9	1.8	1.8	2.0	1.8	2.0	1.9	1.9	2.0

Global deformation properties of CG simulations.

RMSIP is the inner product between the all-atom (AA) and the coarse-grained (CG) trajectories. PL stands for the persistence length (in nm), EtE is the end-to-end distance (in Å) in AA or CG simulations with std being the associated standard deviation (another proxy of flexibility).

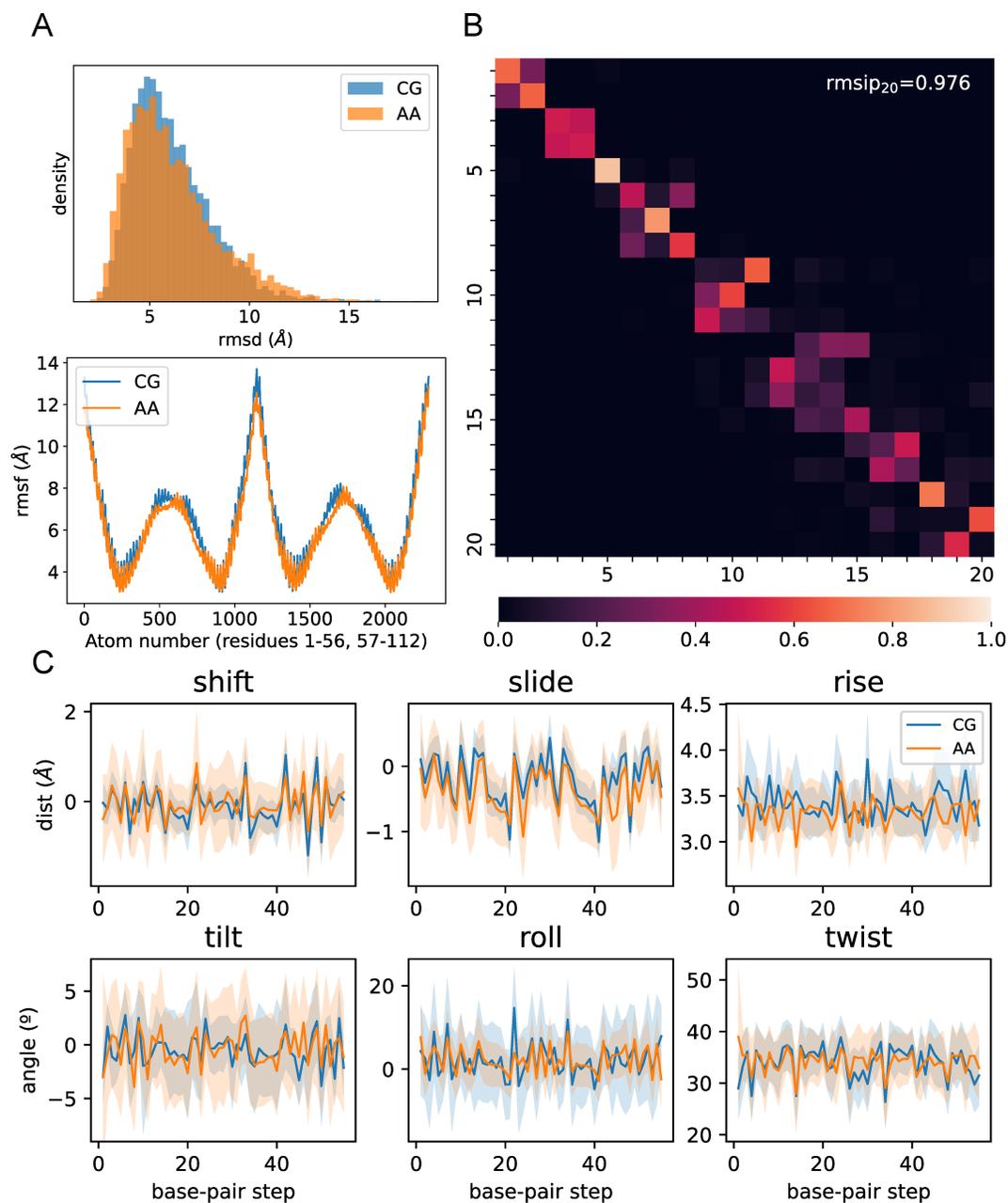


Figure 3. Results from the 56-mer simulations. **(A)** (Top) All-Atom RMSD distribution against AA average from the reconstructed CG structures (blue), and from the AA simulation (orange). (Below) Mean fluctuation per atom of reconstructed CG simulation (blue), and the AA simulation (orange). The atom number ranges through one strand in 5'-to-3' direction, then the other strand in the same direction. **(B)** Root mean square inner product matrix between CG and AA principal components extracted from the respective trajectories. **(C)** Helical parameters across each base-pair step for both CG and AA trajectories. Lines represent mean values, and shadows represent 1 standard deviation.

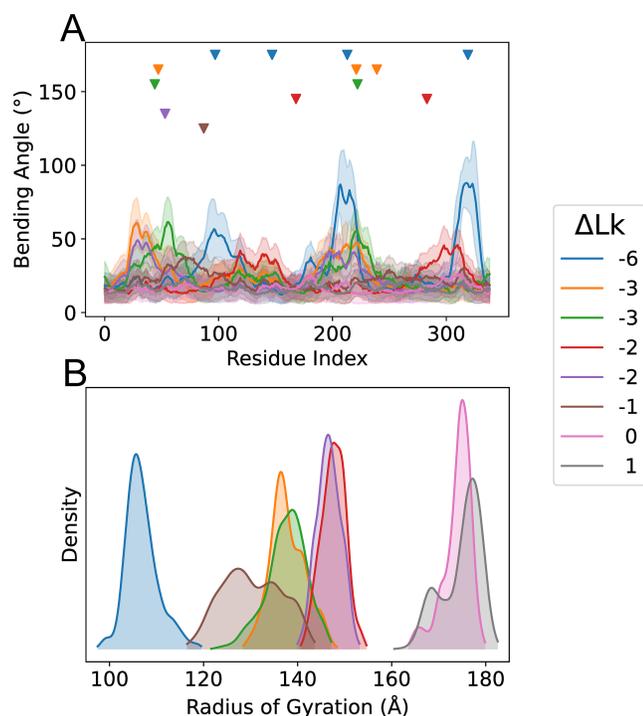


Figure 4. Results from simulations of Circular DNA. **(A)** Bending Angle per base pair from CG simulations, colored by Linking Number difference (ΔLk). Lines represent mean values, and shadows represent 1 standard deviation. Triangles at the top represent kinks/defects in the Atomistic simulations, with the same color coding. **(B)** Radius of Gyration distribution from the CG Simulation, colored by linking number difference.

Computational details

All-atom simulations were performed using PARMBSC1 force-field (4), explicit solvent (SPC/E (48) or TIP3P (49) water models) and 100–200 mM salt using state of the art simulation conditions (16,50) at room temperature and pressure (see Supplementary Information for additional details). Trajectories are stored in our BigNASim database (43). For comparison purposes, additional simulations were done using implicit solvent MD, using standard AMBER GB/SA implementation ((51); additional details in Supplementary Information).

Results and discussion

The coarse-grained trajectories

The CG method and its associated Hamiltonian presented here can provide ensembles which resemble very closely at the C1' level those derived from atomistic MD simulations obtained with *state-of-the-art* force-fields and explicit solvent representation (AA-trajectory). For the 18-mer duplexes included in the miniABC database(16) the RMSd (C1') between all-atom and CG MD averaged structures is 0.79 Å (i.e. 0.04 Å × bp), a very small value, within the standard deviation implicit to the average (around 1.7 Å). Note that this good fitting is remarkable considering that the training was done at the 4-mer level, not for the entire duplex, whose global structure was not considered at any point of the calibration. Even more impressively: the results obtained for duplexes out of the training set are also very accurate, not only in terms of the reference AA MD trajectory, but also of experimental structures (see Figure 2). Interestingly, the histograms of RMSd are quite similar in CG and AA simulations, suggest-

ing that the model captures flexibility well (Figure 2). This is confirmed by the inspection of essential deformation modes, which are almost identical in CG and AA simulations as noted in the global overlap between the first ten modes around 0.9 for the miniABC dataset (see Table 1 and selected examples of eigenvectors overlap in Supplementary Figure S4). The performance is maintained for duplexes not considered in the training process (Figure 2), which demonstrates that the essential deformation movement of DNA is very well recaptured by our method (see Supplementary Figure S5). Global flexibility descriptors such as the persistence length, or end-to-end distance (see Supplementary Methods) are also correct (see Table 1), fitting in fact better experimental values than the reference AA values, something that was also found in mesoscopic models of DNA (42). This probably occurs because of some fortuitous error cancellation related to the neglect of long-range anti-correlation effects that corrects a tendency of PARMBSC1 to overestimate DNA stiffness. Finally, and very encouragingly, the model can accurately capture sequence variability. This is shown in Figure S6, where the cross RMSds between 500 structures (CG or AA) of the thirteen 18-mer duplexes (i.e. 13 × 500 total structures) are reported. Not only are the lowest RMSd obtained along the diagonal, reproducing the AA simulations, but even the out of the diagonal similarities detected in AA-simulations are well reproduced in our CG simulations. This indicates that the model is reproducing sequence-dependent structural details with a quality similar to AA simulations. Note again that no training was done using global structural parameters of the duplex.

The AA reconstituted trajectories

When the decoding process is performed, the all-atoms reconstituted trajectories are surprisingly close to the original all-atom (AA) trajectories (all heavy atoms RMSd) around 0.102 Å per bp in the mini-ABC database and around 0.095 Å per bp for the different duplexes considered for validation set. The analysis of a 56-mer duplex available in the BigNASim database (diverse in composition and 3 times larger than the mini-ABC duplexes) demonstrate that the reconstituted trajectory (decoding the CG ensemble) and the AA trajectory (not considered at any point during the calibration of the model) are hard to distinguish (Figure 3, Supplementary Figure S7), not only in terms of general structural descriptors, but also of atomic fluctuations, essential deformation modes, and sequence-dependent helical parameters. A more in-depth exploration of Atomistic details, helical and backbone structures is provided in SI (Supplementary Figures S8 and S9), as well as further testing of helical parameters on a range of sequences (Supplementary Figures S10 and S11). Results suggest a quality beyond the expectations of a C1'-only model, which indicate that some fine details such as backbone geometries are somehow (at least partially) captured in the C1' geometries in a way that ML approaches can capture them. The most sizeable differences between AA and CG distribution is for the standard deviations of some helical parameters (see Figure 3), indicating the intrinsic shortcomings of the C1' representation. Correction of these deviations would require the addition of extra degrees of freedom.

Circular DNAs

In order to test the limits of our model we explored 339 bp circular DNA, which has been already studied by AA MD simulation, and for which electron microscopy images are

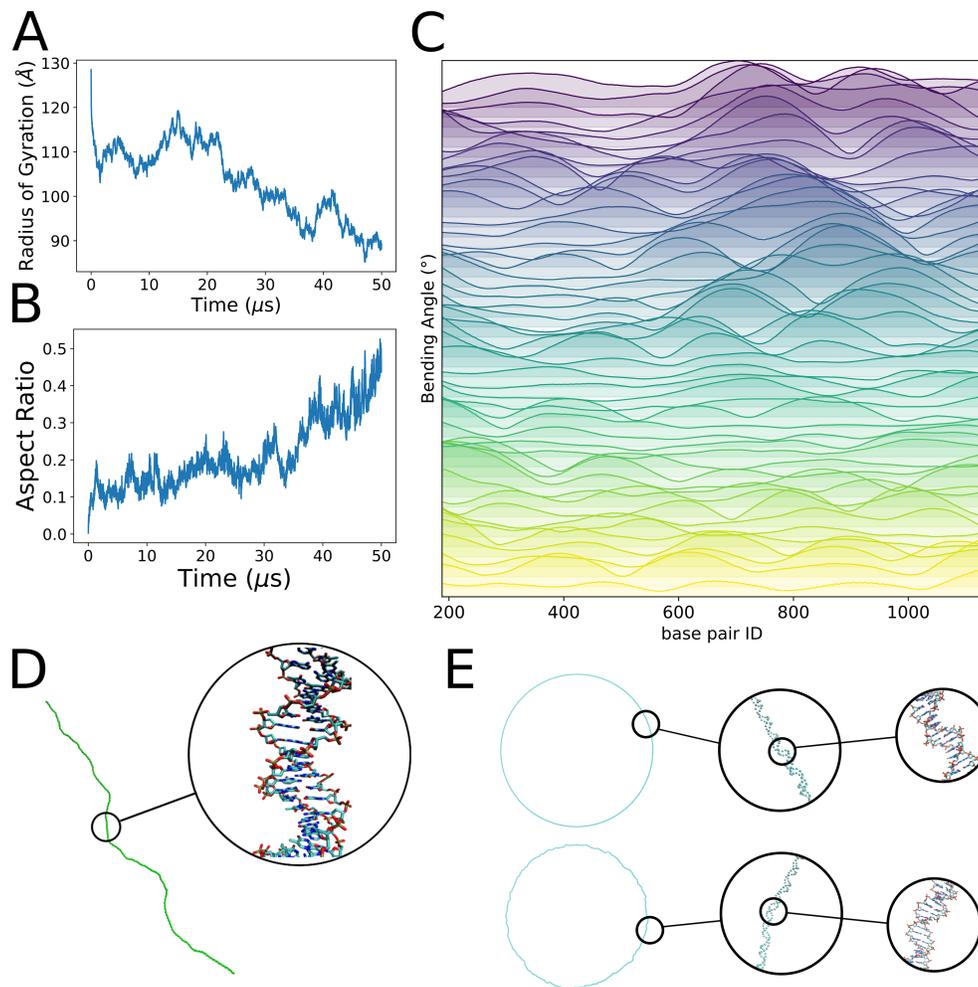


Figure 5. Coarse-grained simulation of yeast gene YCL020W containing 1317 bp, and Mitochondrial DNA. (A, B) Time evolution of the radius of gyration (A) and the aspect ratio (B) of the structure of the gene. (C) Bending angle per base pair of the gene structure, computed with a window of 196 bp, colored by simulation fragment, first plot corresponds to the first microsecond of simulation, second corresponds to the second microsecond and so on. The values correspond to the average over each trajectory block. Plots go from bottom (yellow) to top (purple), as the simulation blocks increase from block 1 to block 50. (D) Representative snapshot showing the all-atom reconstruction of the gene structure. (E) Human mitochondrial DNA structure at three different scales. Zoom at the top corresponds to a frame near the beginning of the simulation. Zoom at the bottom is after 15 μ s of simulation, where local distortions are evident.

available(52). Starting structures (see Supplementary Methods for generation scripts) from relaxed DNA quickly converged into the predicted supercoiling, leading to stable trajectories sampling a wide range of conformational space. By construction, the method is not able to capture kinks, but those regions where kinks appear in atomistic MD simulations (52) (for a given supercoiling) are those where higher bending angles are obtained (Figure 4), showing the ability of the model to identify softer and more flexible regions of the minicircle. The Global descriptors captured from CG simulations show diversity in structure, related to those collected from electron microscopy measures (see Figure 4, Supplementary Figure S12 and reference (52)) and a wide sampling of conformational shapes are obtained as visible in the oscillation of the radii of gyration along the trajectory (see Supplementary Figure S12). Finally, and quite surprisingly, the ML method, which was trained with linear DNA, maintains a good ability to back-map minicircle trajectories to AA resolution (see Supplementary Figure S12). In summary, our CG model and associated ML-reconstitution algorithm are not optimal to explore extremely stressed DNAs like those

in minicircles, where the elastic regime might not be valid, but they can be very useful to perform massive screenings to obtain reasonable atomic resolution ensembles from which AA MD simulations can be performed.

Very long systems

The objective of any CG representation of DNA is to expand the size of the systems accessible to simulation. We tested here the performance of the method in two large systems: i) a 1317-mer long duplex (2.6 kbases in terms of mass) (link, Supplementary Movie S1) bearing the YCL020W gene of *Saccharomyces cerevisiae* (the TYA retrotransposon coding for the TY1 virus-like particle) and ii) the human mitochondrial DNA (33 kbases) (link, Supplementary Movie S2). Any of these two systems is very far away from what is accessible to atomistic MD simulations, but can be simulated by our model even when using a single processor desktop computer, providing clues of the DNA flexibility in the polymeric range, which to our knowledge, were never described before. Figure 5 shows details of a 50-microsecond simulation of YCL020W

Data availability

The data that supports the findings of this study is openly available in the public repository <https://mmb.irbbarcelona.org/gitlab/dfarre/cgenarate-materials>. The executable for CGeNArate is also available in the public repository <https://mmb.irbbarcelona.org/gitlab/dfarre/cgenarate-materials>. Simulation files will be deposited in BigNASim (<https://mmb.irbbarcelona.org/BIGNASim/>). The list of accession numbers is provided in the Supplementary Data.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We are indebted to Dr F. Battistini for helping in the discussion of results. The authors acknowledge the support of Barcelona Supercomputer Center, the Spanish Ministry of Science [PDI2021-122478NB-I00 to M.O.], AGAUR, Generalitat de Catalunya, Grups de Recerca Consolidats, [2021 SGR 00863 to M.O.], Instituto de Salud Carlos Tercero [IMPACT IMP/00019 to M.O.], BioExcel-3: Centre of Excellence for Computational Biomolecular Research [101093290 HORIZON-EUROHPC-JU-2021-COE-01 to M.O.], HORIZON-INFRA-2022-DEV-01 [101094651-MDDB to M.O.], European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie [754510 to J.P.A.], PROBIST postdoctoral fellowship from the Barcelona Institute of Science and Technology [to J.P.A.], Access to ARCHER2 and JADE2 UK National Supercomputing Services was provided through the Engineering and Physical Sciences Research Council grant to the HECBioSim Consortium [EP/R029407/1 to C.L.]. AGAUR, Generalitat de Catalunya granted a FI fellowship to D.F.G, co-financed by European Social Fund Plus.

Funding

Generalitat de Catalunya [2021 SGR 00863]; Ministerio de Ciencia e Innovación [PDI2021-122478NB-I00]. Funding for open access charge: Spanish Ministry of Science [PDI2021-122478NB-I00].

Conflict of interest statement

None declared.

References

- Dans, P.D., Walther, J., Gómez, H. and Orozco, M. (2016) Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.*, **37**, 29–45.
- Hobza, P. and Šponer, J. (1999) Structure, energetics, and dynamics of the nucleic acid base pairs: nonempirical *ab initio* calculations. *Chem. Rev.*, **99**, 3247–3276.
- Šponer, J., Mládek, A., Špačková, N., Cang, X., Cheatham, T.E. and Grimme, S. (2013) Relative stability of different DNA guanine quadruplex stem topologies derived using large-scale quantum-chemical computations. *J. Am. Chem. Soc.*, **135**, 9785–9796.
- Ivani, I., Dans, P.D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A., *et al.* (2015) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
- Zgarbová, M., Šponer, J., Otyepka, M., Cheatham, T.E., Galindo-Murillo, R. and Jurečka, P. (2015) Refinement of the sugar–Phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *J. Chem. Theory Comput.*, **11**, 5723–5736.
- Hart, K., Foloppe, N., Baker, C.M., Denning, E.J., Nilsson, L. and MacKerell, A.D. (2012) Optimization of the CHARMM additive force field for DNA: improved treatment of the BI/BII conformational equilibrium. *J. Chem. Theory Comput.*, **8**, 348–362.
- 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.
- Arcella, A., Dreyer, J., Ippoliti, E., Ivani, I., Portella, G., Gabelica, V., Carloni, P. and Orozco, M. (2014) Structure and dynamics of oligonucleotides in the gas phase. *Angew. Chem. Int. Ed.*, **54**, 467–471.
- Portella, G., Germann, M.W., Hud, N.V. and Orozco, M. (2014) MD and NMR analyses of choline and TMA binding to duplex DNA: on the origins of aberrant sequence-dependent stability by alkyl cations in aqueous and water-free solvents. *J. Am. Chem. Soc.*, **136**, 3075–3086.
- Portella, G., Terrazas, M., Villegas, N., González, C. and Orozco, M. (2015) Can A denaturant stabilize DNA? Pyridine reverses DNA denaturation in acidic pH. *Angew. Chem. Int. Ed.*, **54**, 10488–10491.
- Porrini, M., Rosu, F., Rabin, C., Darré, L., Gómez, H., Orozco, M. and Gabelica, V. (2017) Compaction of duplex nucleic acids upon native electrospray mass spectrometry. *ACS Cent Sci*, **3**, 454–461.
- Kuzmanic, A., Dans, P.D. and Orozco, M. (2019) An In-depth look at DNA crystals through the prism of molecular dynamics simulations. *Chem.*, **5**, 649–663.
- Bhattacharya, D. and Bansal, M. (1988) A general procedure for generation of curved dna molecules. *J. Biomol. Struct. Dyn.*, **6**, 093–104.
- Bhattacharyya, D. and Bansal, M. (1989) A self-consistent formulation for analysis and generation of non-uniform dna structures. *J. Biomol. Struct. Dyn.*, **6**, 635–653.
- 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.
- Dans, P.D., Balaceanu, A., Pasi, M., Patelli, A.S., Petkevičiūtė, D., Walther, J., Hospital, A., Bayarri, G., Lavery, R., Maddocks, J.H., *et al.* (2019) The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res.*, **47**, 11090–11102.
- Walther, J., Dans, P.D., Balaceanu, A., Hospital, A., Bayarri, G. and Orozco, M. (2020) A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res.*, **48**, e29.
- Pasi, M., Maddocks, J.H., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., Dans, P.D., Jayaram, B., Lankaš, F., Laughton, C., *et al.* (2014) μ ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
- López-Güell, K., Battistini, F. and Orozco, M. (2023) Correlated motions in DNA: beyond base-pair step models of DNA flexibility. *Nucleic Acids Res.*, **51**, 2633–2640.
- De Bruin, L. and Maddocks, J.H. (2018) cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA. *Nucleic Acids Res.*, **46**, W5–W10.
- Gonzalez, O., Petkevičiūtė, D. and Maddocks, J.H. (2013) A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.*, **138**, 055102.
- Liebl, K. and Zacharias, M. (2021) Accurate modeling of DNA conformational flexibility by a multivariate ising model. *Proc. Natl. Acad. Sci.*, **118**, e2021263118.

23. 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.
24. Neguembor, M.V., Arcon, J.P., Buitrago, D., Lema, R., Walther, J., Garate, X., Martin, L., Romero, P., AlHaj Abed, J., Gut, M., et al. (2022) MiOS, an integrated imaging and computational strategy to model gene folding with nucleosome resolution. *Nat. Struct. Mol. Biol.*, **29**, 1011–1023.
25. Basu, A., Bobrovnikov, D.G., Cieza, B., Arcon, J.P., Qureshi, Z., Orozco, M. and Ha, T. (2022) Deciphering the mechanical code of the genome and epigenome. *Nat. Struct. Mol. Biol.*, **29**, 1178–1187.
26. Buitrago, D., Labrador, M., Arcon, J.P., Lema, R., Flores, O., Esteve-Codina, A., Blanc, J., Villegas, N., Bellido, D., Gut, M., et al. (2021) Impact of DNA methylation on 3D genome structure. *Nat. Commun.*, **12**, 3243.
27. Louison, K.A., Dryden, I.L. and Laughton, C.A. (2021) GLIMPS: a machine learning approach to resolution transformation for multiscale modeling. *J. Chem. Theory Comput.*, **17**, 7930–7937.
28. Naómé, A., Laaksonen, A. and Vercauteren, D.P. (2014) A solvent-mediated coarse-grained model of DNA derived with the systematic Newton inversion method. *J. Chem. Theory Comput.*, **10**, 3541–3549.
29. Savelyev, A. and Papoian, G.A. (2009) Molecular renormalization group coarse-graining of polymer chains: application to double-stranded DNA. *Biophys. J.*, **96**, 4044–4052.
30. Ouldrige, T.E., Louis, A.A. and Doye, J.P.K. (2010) DNA nanotweezers studied with a coarse-grained model of DNA. *Phys. Rev. Lett.*, **104**, 178101.
31. Doye, J.P.K., Ouldrige, T.E., Louis, A.A., Romano, F., Šulc, P., Matek, C., Snodin, B.E.K., Rovigatti, L., Schreck, J.S., Harrison, R.M., et al. (2013) Coarse-graining DNA for simulations of DNA nanotechnology. *Phys. Chem. Chem. Phys.*, **15**, 20395.
32. Maffeo, C., Ngo, T.T.M., Ha, T. and Aksimentiev, A. (2014) A coarse-grained model of unstructured single-stranded DNA derived from atomistic simulation and single-molecule experiment. *J. Chem. Theory Comput.*, **10**, 2891–2896.
33. 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.
34. Assenza, S. and Pérez, R. (2022) Accurate sequence-dependent coarse-grained model for conformational and elastic properties of double-stranded DNA. *J. Chem. Theory Comput.*, **18**, 3239–3256.
35. Markegard, C.B., Fu, J.W., Reddy, K.A. and Nguyen, H.D. (2015) Coarse-grained simulation study of sequence effects on DNA hybridization in a concentrated environment. *J. Phys. Chem. B*, **119**, 1823–1834.
36. Machado, M.R. and Pantano, S. (2015) Exploring LacI–DNA dynamics by multiscale simulations using the SIRAH Force field. *J. Chem. Theory Comput.*, **11**, 5012–5023.
37. Klein, F., Soñora, M., Helene Santos, L., Nazareno Frigini, E., Ballesteros-Casallas, A., Rodrigo Machado, M. and Pantano, S. (2023) The SIRAH force field: a suite for simulations of complex biological systems at the coarse-grained and multiscale levels. *J. Struct. Biol.*, **215**, 107985.
38. Uusitalo, J.J., Ingólfsson, H.I., Akhshi, P., Tieleman, D.P. and Marrink, S.J. (2015) Martini coarse-grained force field: extension to DNA. *J. Chem. Theory Comput.*, **11**, 3932–3945.
39. Maciejczyk, M., Spasic, A., Liwo, A. and Scheraga, H.A. (2014) DNA duplex formation with a coarse-grained model. *J. Chem. Theory Comput.*, **10**, 5020–5035.
40. Cragolini, T., Derreumaux, P. and Pasquali, S. (2013) Coarse-grained simulations of RNA and DNA duplexes. *J. Phys. Chem. B*, **117**, 8047–8060.
41. Savelyev, A. and Papoian, G.A. (2010) Chemically accurate coarse graining of double-stranded DNA. *Proc. Natl. Acad. Sci.*, **107**, 20340–20345.
42. Savelyev, A. and Papoian, G.A. (2008) Polyionic charge density plays a key role in differential recognition of mobile ions by biopolymers. *J. Phys. Chem. B*, **112**, 9135–9145.
43. Hospital, A., Andrio, P., Cugnasco, C., Codo, L., Becerra, Y., Dans, P.D., Battistini, F., Torres, J., Goñi, 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.
44. Powell, M.J.D. (1994) A direct search optimization method that models the objective and constraint functions by linear interpolation. In: *Advances in Optimization and Numerical Analysis*. Springer Netherlands, Dordrecht, pp. 51–67.
45. Nam, G.M. and Arya, G. (2014) Torsional behavior of chromatin is modulated by rotational phasing of nucleosomes. *Nucleic Acids Res.*, **42**, 9691–9699.
46. Leimkuhler, B. and Matthews, C. (2015) In: *Molecular Dynamics*. Springer International Publishing, Cham.
47. Box, G.E.P. and Muller, M.E. (1958) A note on the generation of random normal deviates. *Ann. Math. Stat.*, **29**, 610–611.
48. Berendsen, H.J.C., Grigera, J.R. and Straatsma, T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
49. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
50. da Rosa, G., Grille, L., Calzada, V., Ahmad, K., Arcon, J.P., Battistini, F., Bayarri, G., Bishop, T., Carloni, P., Cheatham, T. III, et al. (2021) Sequence-dependent structural properties of B-DNA: what have we learned in 40 years? *Biophys. Rev.*, **13**, 995–1005.
51. Onufriev, A.V. and Case, D.A. (2019) Generalized born implicit solvent models for biomolecules. *Annu. Rev. Biophys.*, **48**, 275–296.
52. Pyne, A.L.B., Noy, A., Main, K.H.S., Velasco-Berrelleza, V., Piperakis, M.M., Mitchenall, L.A., Cugliandolo, F.M., Beton, J.G., Stevenson, C.E.M., Hoogenboom, B.W., et al. (2021) Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and major groove recognition by triplex-forming oligonucleotides. *Nat. Commun.*, **12**, 1053.
53. Balaceanu, A., Pérez, A., Dans, P.D. and Orozco, M. (2018) Allostery and signal transfer in DNA. *Nucleic Acids Res.*, **46**, 7554–7565.
54. Pang, D., Thierry, A.R. and Dritschilo, A. (2015) DNA studies using atomic force microscopy: capabilities for measurement of short DNA fragments. *Front. Mol. Biosci.*, **2**, 1.
55. Dršata, T. and Lankaš, F. (2015) Multiscale modelling of DNA mechanics. *J. Phys. Condens. Matter*, **27**, 323102.
56. Battistini, F., Sala, A., Hospital, A. and Orozco, M. (2023) Sequence-dependent properties of the RNA duplex. *J. Chem. Inf. Model.*, **63**, 5259–5271.