Nucleic Acids Research, 2025, 1–7
https://doi.org/10.1093/nar/gkaf371
Web Server issue

CGeNArateWeb: a web server for the atomistic study of the structure and dynamics of chromatin fibers

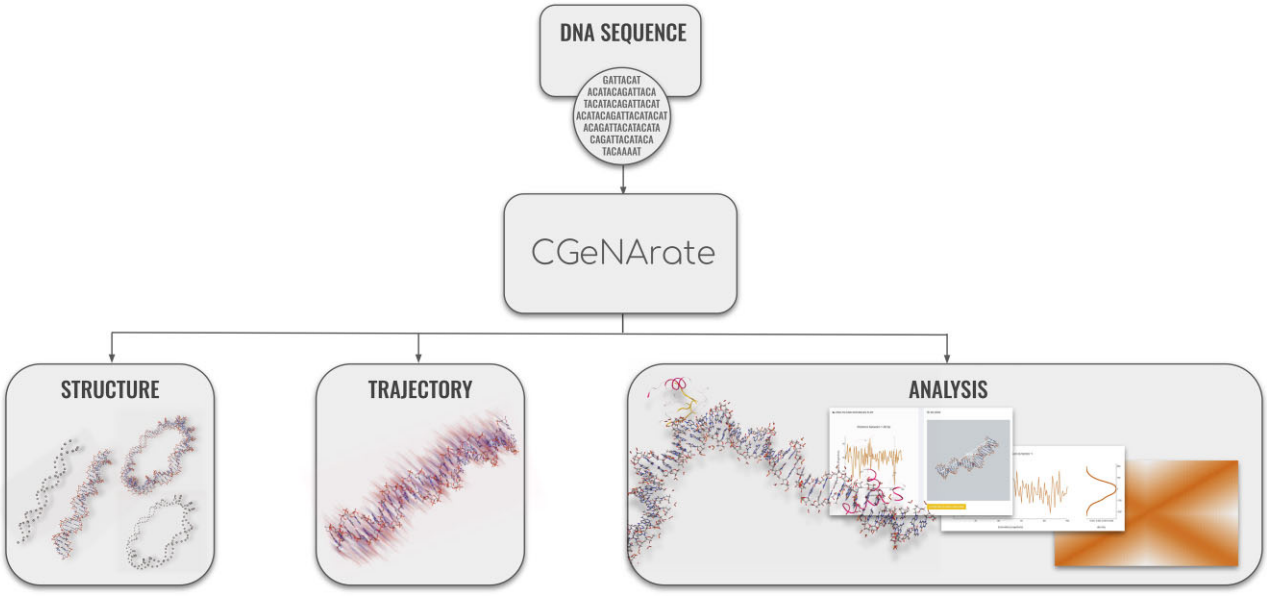
David Farré-Gil [1,2], Genis Bayarri [1], Charles A. Laughton [3], Adam Hospital [1], Modesto Orozco [1,4,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona 08028, Spain
[2]Department of Mathematics and Computer Science, University of Barcelona, Barcelona 08007, Spain
[3]School of Pharmacy and Biodiscovery Institute, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom
[4]Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona 08028, Spain
[*]To whom correspondence should be addressed. Email: modesto.orozco@irbbarcelona.org

Abstract

We present CGeNArateWeb, a new web tool for the three-dimensional simulation of naked DNA and protein-bound chromatin fibers. The server allows the user to obtain a dynamic representation of long segments of linear, circular, or protein–DNA segments thanks to a Langevin dynamics coarse-grained (CG) model working with a machine-learning (ML) fitted C1′-resolution Hamiltonian. The CG trajectories can be back-mapped to atomistic resolution using another ML algorithm trained on a large database of molecular dynamics (MD) simulations. The method allows the user to get structural and dynamic information on large (kilobase range) portions of both protein-bound and free DNA, to transform conceptual cartoons into structural and dynamical models. Trajectories are analyzed using an extensive set of nucleic acid-specific analysis tools, and the results are displayed using a powerful and flexible graphic interface. The web tool uses state-of-the-art technologies such as (i) Docker components orchestrated by Docker Swarm, with containers deployed on demand for computations, (ii) WebGL-programmed NGL molecular viewer and the JavaScript plotly library for interactive plots, and (iii) noSQL-MongoDB for storage. The server is accessible at https://mmb.irbbarcelona.org/CGNAW/. The web tool is free and open to all users, and there are no login requirements.

Graphical abstract



Introduction

Obtaining dynamic information on structural systems has become the new frontier in the post-AlphaFold era. Obtaining dynamical information on biological macromolecules is still very challenging. On one hand, experimental techniques use experimental setups designed to reduce flexibility, as a very

Received: March 24, 2025. **Revised:** April 13, 2025. **Editorial Decision:** April 17, 2025. **Accepted:** April 25, 2025

flexible structure will not provide net images, densities, or clear NMR signals. On the other hand, information-based techniques learn using multiple sequence alignments (MSAs) and structural information; i.e. they need to infer flexibility by fragmenting MSAs to provide alternative structures [1], which very often have little physical sense [2], and by construction, never define a Boltzmann ensemble of the macromolecular system. In this context, simulation techniques appear as the only alternative to provide information of macromolecular dynamics at the atomistic level under physiological-like conditions. Despite recent advances in accuracy and speed [3–5], traditional MD is still limited by the size of the system and the timescale of the process to study. Such scale problems arise from the large number of particles that need to be considered, as well as by the small (femtosecond range) integration time step. In this context, coarse-graining (CG) models, where many atoms are considered together as a single heavy bead and solvent is simplified as a continuum, emerge as an exciting approach to extend the range of applicability of MD simulations.

DNA is probably one of the most complex macromolecule to study, and a textbook example of a multiscale system moving from nanometers to meters in space scale and from picoseconds to hours in the timescale. Techniques such as X-ray crystallography, NMR spectroscopy, or atomistic MD, which provide good representations of small and rigid DNA fragments, are useless to study long segments, especially in the context of chromatin, where DNA is intimately bound to effector proteins. Recently (see [6] and references therein), a variety of techniques based on sequencing (MNAseq, CHIPseq, ATAC-seq, chromosome conformation capture, etc.) have emerged as the main source of low-resolution data, providing information on the accessibility of chromatin, its intra- and intermolecular contacts, and the region of protein–DNA interactions. However, transforming this low-resolution data into structures requires the use of CG models (see [7–15] and recent review in [4]). Currently, most of the available CG methods are created to study medium-size systems, neglect sequence variability in DNA properties, or cannot easily incorporate proteins and other effectors in the simulation. Furthermore, CG methods are implemented in expert-oriented simulation packages, requiring dedicated hardware and specialized analysis packages. The net result is that, currently, the typical "wet-lab" bioscientist lacking this expertise is unable to transform the experimental information on chromatin into structural models.

We present CGeNArateWeb (CGNAW), a new web tool incorporating an extended version of our ML-trained sequence-dependent C1′-based CG model of duplex DNA [7, 16] able to deal with short-time response dynamic representation of Kb segments of DNA, both naked (linear and circular) and protein bound. In addition to the CGeNArate [7] simulation engine, the server incorporates a modified version of the ML method GLIMPS [7, 16], which transforms CG samplings into atomistic ensembles that can be downloaded for *in situ* analysis or subjected to a set of DNA-adapted analysis tools incorporated in the server. CGNAW democratizes the use of MD techniques and fills the gap between atomistic and chromatin simulations. The tool is freely accessible at https://mmb.irbbarcelona.org/CGNAW/. A source code useful to perform simulations of naked DNA longer than those accessible from the server can be downloaded from https://github.com/mmb-irb/cgenarate-materials/tree/main.

Materials and methods

Simulation engine

The CGeNArate method represents the DNA geometry with one bead per nucleotide located at the C1′ position and a polynomial expansion of the intramolecular DNA energy [15, 17], which accounts for short-range intrastrand distance ($i \rightarrow i \pm 1$) and angle ($i \rightarrow i \pm 2$) term and interstrand distance-dependent terms ($i \rightarrow j$, $j \pm 1$,..., $j \pm 5$). The parameters in the equations are determined by fitting variance/covariance matrices derived from MD simulations of all unique tetramers of DNA [17] and our local database of PARMBSC1 [18, 19] atomistic trajectories of DNA duplexes [20]. The CGeNArate Hamiltonian includes also a Lennard Jones and a Debye–Hückel term to account for remote interactions, which facilitates the incorporation of interacting particles. The (mostly) local nature of the Hamiltonian facilitates parallelization in the stand-alone version, within the OPENMP protocol, which grants a very good scalability of the calculation. Finally, to achieve the ultra-fast sampling required for a server, the current implementation considers effector proteins as dummy particles whose interactions are projected into the DNA beads, and the internal dynamics of bound segments of DNA are kept constrained using Lagrange multipliers.

Back-mapping procedure

A modification of the GLIMPS algorithm [7, 16] is used to reconstitute atomistic coordinates of the DNA from C1′ beads, while the PDB structure is used to reconstitute protein atoms. GLIMPS uses an ML-powered two-step approach, first transforming CG coordinates to an atomistic representation of the sugar–phosphate backbone in 10-mer segments, and then in a second pass generating coordinates for the specific base pairs. As discussed elsewhere, the back-mapping method was trained using our PARMBSC1-BigNAsim library of MD simulations [20] and is able to capture with impressive accuracy not only the global DNA shape but also fine details at the base-pair step level and even the backbone geometry [7].

Computational efficiency

The stand-alone version of CGeNArate, which can be downloaded through github, is capable of simulating 1ns of dynamics of a 1kb DNA duplex within 50s, using a single Zenodo3 core. With an OPENMP parallel version, the time is reduced to 8s in 16 cores. Current web implementation, however, is restricted to 500 bp due to the inherent limitations of web-based visualization tools, which struggle with rendering larger datasets efficiently in a browser environment. Simulation time is restricted to 200 ns to provide results within a reasonable waiting time. Further versions implementing the parallelized version would allow reaching the multi-microsecond timescale with the fast response expected for a web interface. Longer simulations, e.g. those required to explore circularization propensities, could require downloading the source code (see above). Note that the sampling obtained with a CG model with implicit solvent is at least 10 times larger than that expected from an atomistic resolution with explicit solvent simulation [7]. The computational cost of back-mapping the CG simulation into the atomistic level is around 2 s per frame for a 500-bp sequence (20 000 atoms).

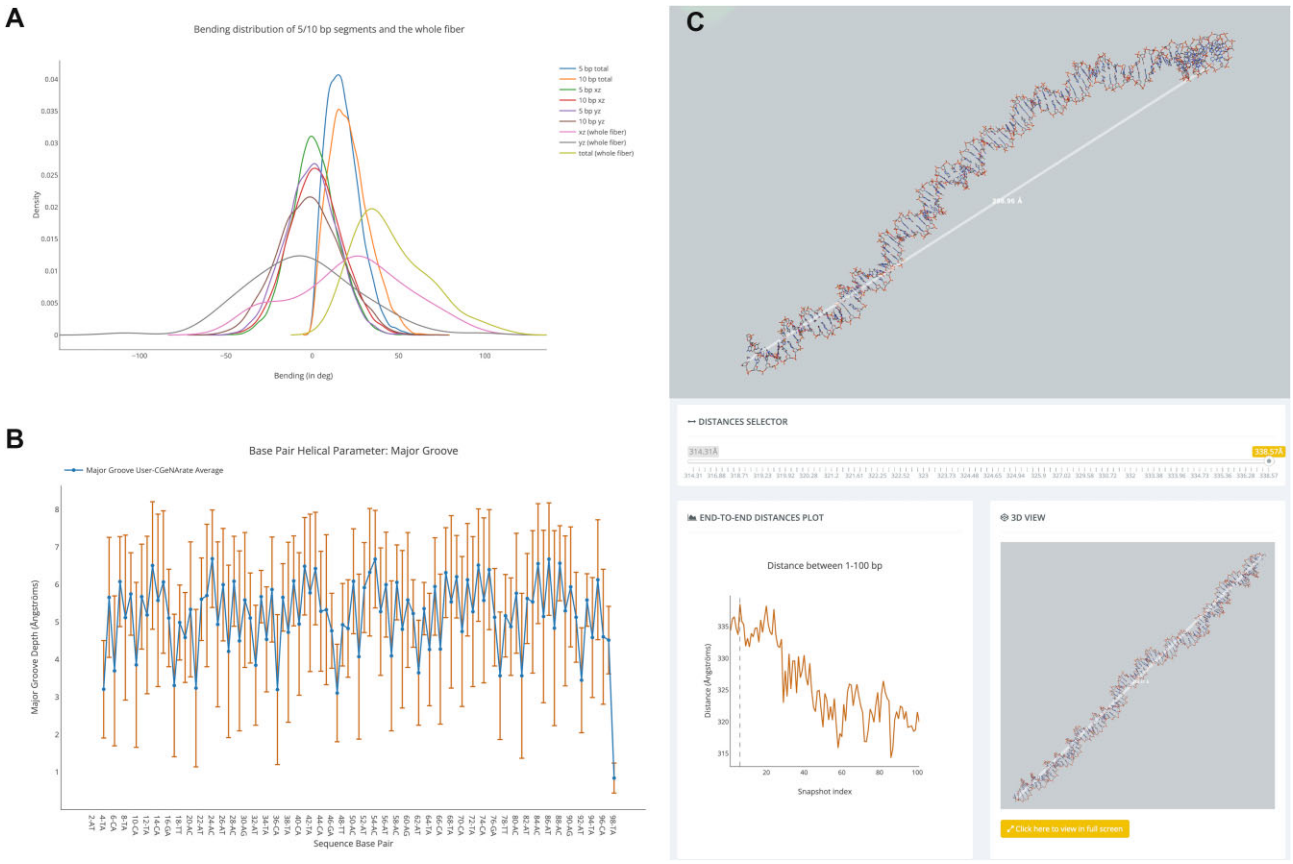


Figure 1. Structural and dynamical characteristics of a 100-nt linear DNA, simulated for 100 ns: (**A**) Bending distributions for 5- and 10-bp windows and the whole fiber, also separated by directional contributions. (**B**) Major groove depth showing binding pockets of around 6 Å. (**C**) End-to-end distance visualization for the most bent structure on top and most extended structure on the bottom, showing also the interactive distance–frame–structure interface.

Technology

CGNAW is a web portal implemented with Slim PHP micro-framework (https://www.slimframework.com/) following a model–view–controller architectural pattern, supported by a MongoDB noSQL database (https://www.mongodb.org). NGL molecular viewer [21] is used to visualize 3D structures and trajectories, and the plotly JavaScript package (https://plot.ly/) for modern data visualization was chosen to display all the analysis plots. Jobs are queued using SGE manager (http://gridscheduler.sourceforge.net/), and served in an on-demand processing model performed by Docker containers automatically deployed in a Docker Swarm stack for multi-tiered applications.

Input information

Starting from a DNA sequence, CGNAW offers the possibility to study three systems: naked linear DNA, circular DNA, and protein-bound DNA (see above). The user should select the desired level of resolution in the output (CG or atomistic) and the operations to be performed with the ensemble (up to 500 structures, collected from the total simulation time, of up to 100 ns), flexibility analyses, etc. (see examples below). Depending on the type of DNA considered (naked, protein-bound, or circular), additional input parameters are needed. For example, the linking number is required for circular DNAs, and the PDB IDs of the proteins bound to DNA need to be specified when simulating DNA–protein complexes. Note that in the present version PDB entry is not universal, but structure should be retrieved from a continuously extended list of PDB curated complexes. The server offers the user the possibility to place the protein(s) at specific site(s) or scan the DNA to find the places where deforming the DNA to adopt the bioactive conformation is easier. This is computed by determining the elastic energy required to deform the DNA from the equilibrium to the protein-bound conformation as described elsewhere [22, 23].

Upon selecting the submit button, the user receives a URL address where they will find all the results of the simulation once it is completed.

Output information

CGNAW results are divided into two main sections: (i) *summary*, which contains information about all the input parameters chosen for the job process, together with an NGL visualization of the generated structure and trajectory, and (ii) *trajectory flexibility analysis*, which contains a set of flexibility analyses done on the generated trajectory. The list of analyses varies depending on the selected method and resolution, and all together provide a full description of DNA flexibility. The inventory of analyses performed within the server includes helical parameters, stiffness energy constants, distance contact maps (for DNA and proteins), end-to-end distances,

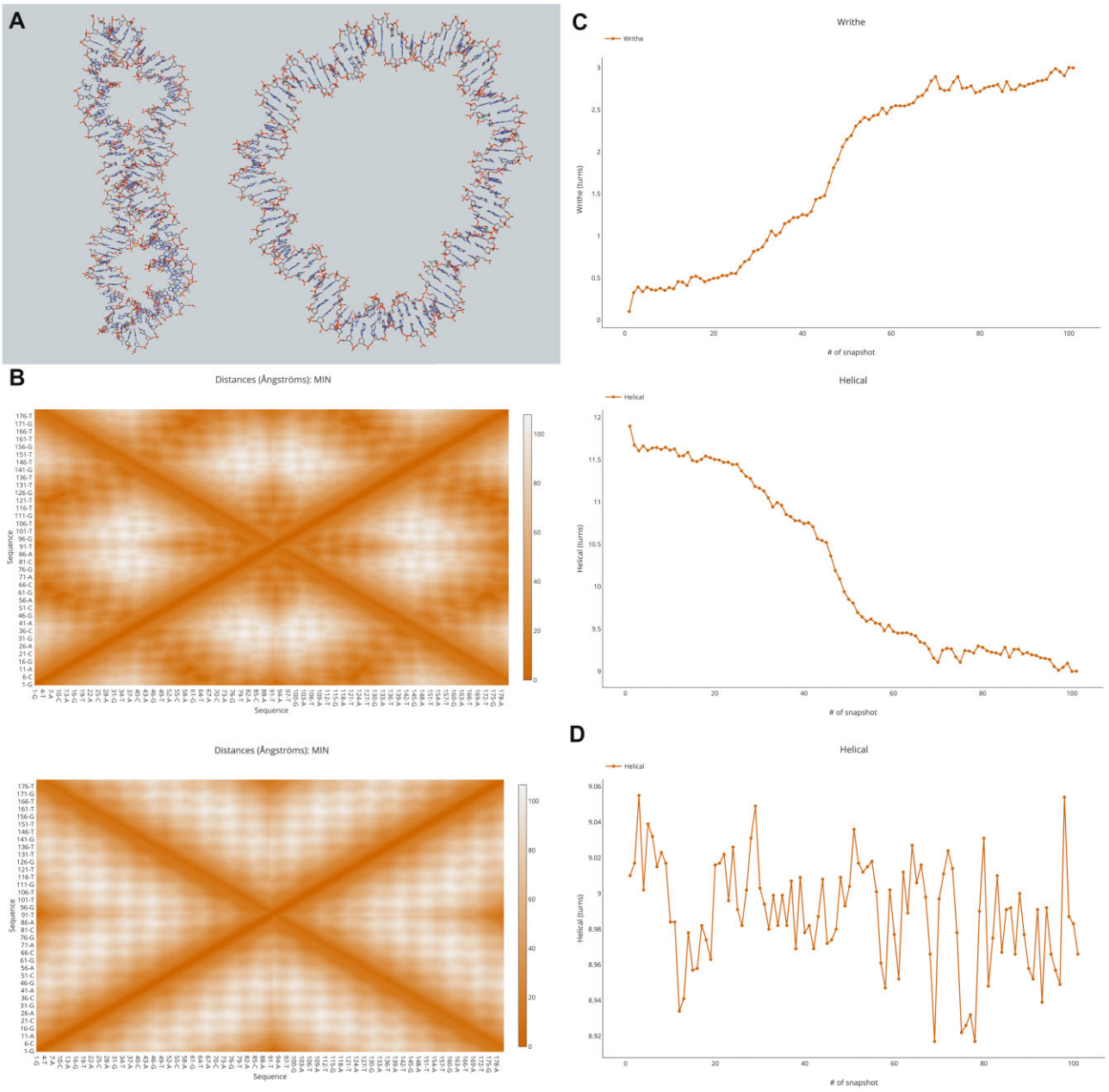


Figure 2. Structural impact of overtwisting in DNA, comparing +3 and 0 ΔLk for a 92-bp sequence. (**A**) Final structures of the simulations, left $\Delta Lk =$ +3; right, $\Delta Lk = 0$. (**B**) Minimum distances for nucleotide–nucleotide contacts: top, $\Delta Lk =$ +3; bottom, $\Delta Lk = 0$. (**C**) Writhe and helical values for the $\Delta Lk =$ +3 simulation, showing recovery of helical turns from 12 (+3) to canonical 9 (+0). (**D**) Helical values for the $\Delta Lk = 0$ simulation, stable at 9.

DNA bending, persistence length, elastic energies, and virtual DNA footprinting, among others. Results are presented in a very intuitive and friendly interface, exploiting interactivity when possible (see section below). A guided tour for each analysis tool helps the user to get started navigating through the analysis section.

Each of the results sections offers the possibility to download the specific analysis raw data in a compressed file for further analysis or as a starting point for atomistic MD simulations using local tools, standard packages such as CURVES [24], or our DNA-specific BioExcel Building Blocks workflows [25–27]. Access to the web server is free. Sample inputs and outputs are supplied to ease the process of getting familiar with the tool and its possibilities.

Results

The server includes a few examples of potential use of the tool to represent DNA ensembles (Figs 1–3) for both for naked and protein-bound DNA.

One example is the study of the dynamics of long pieces of DNA (100-mer in the example), for which the user can explore the general structural and dynamical characteristics, the groove geometries (to explore possible binding pockets), persistence lengths, and end-to-end distances to evaluate the circularization propensity of different sequences (Fig. 1). Here the server offers an interactive interface that displays the structure together with an end-to-end distance plot and an evaluation of the persistence length for rigid and flexible duplexes. Navigating through the top slider, the user can easily have a

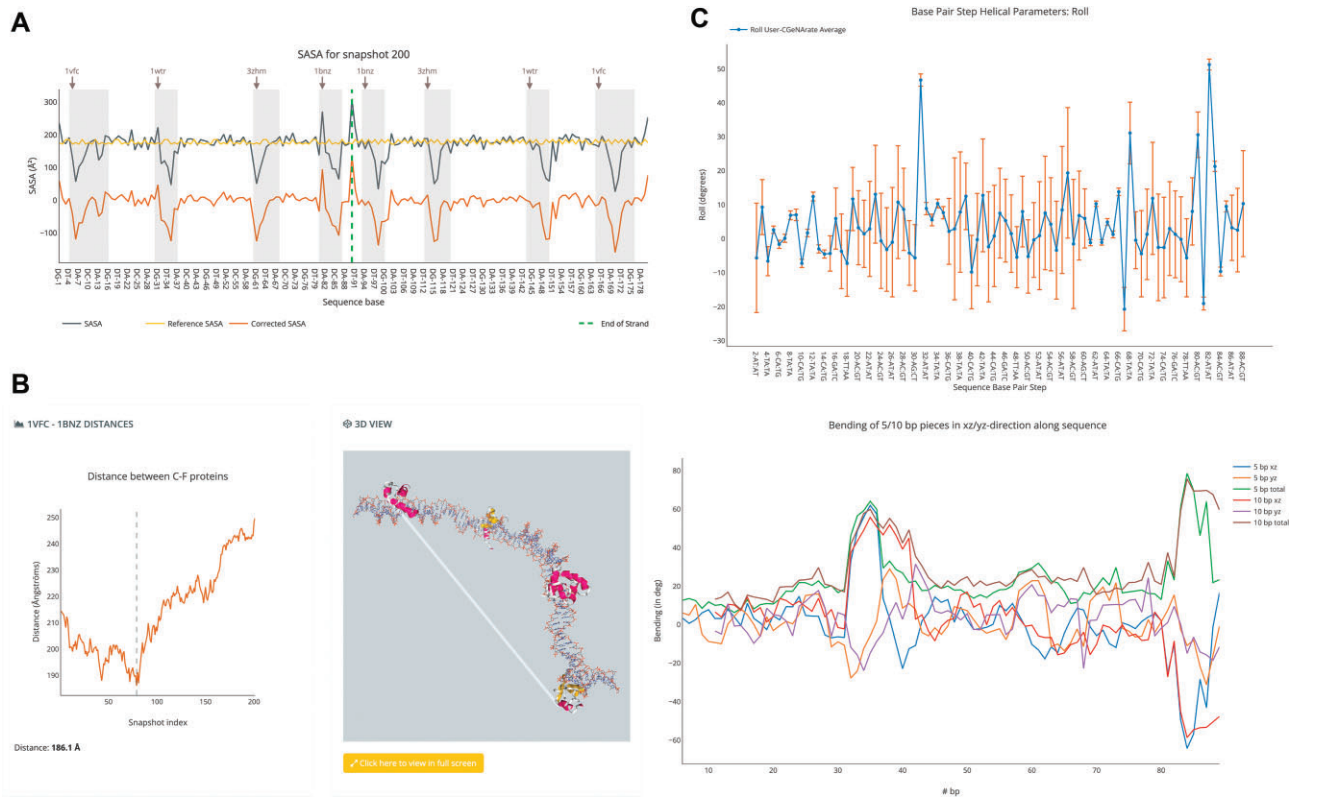


Figure 3. Structural results for a DNA fiber with four bound proteins: (**A**) Solvent accessible surface area computed for the final structure in the simulation. The DNA fragments where proteins are placed are highlighted with gray-shaded regions, correlating with decreased accessibility. (**B**) Interactive protein–protein contact evolution, showing the lowest distance between the chosen proteins, and the structure with most probable cross-talk between the proteins. (**C**) Distortion impact of the binding proteins toward the DNA. The top graph shows the helical parameter roll per base pair step, one of the best internal descriptors of DNA wrapping. The bottom graph shows DNA bending along the sequence, also identifying DNA wrapping with the more highly bent regions of the duplex.

3D view of the generated ensemble, from the most extended to the most bent structures (Fig. 1) for both duplexes.

A second example demonstrates the potential of CGNAW to study the dynamics of DNA mini-circles under different linking numbers. This is reflected in the predicted shapes, as well as the evolution of writhe and twist values, following trends from atomistic simulations [28]. Long-range contacts illustrate how the simulation is able to spontaneously detect superhelical formation resulting from a change in linking number and the tendency of DNA to recover its natural twist (Fig. 2).

A third example refers to the simulation of protein–DNA complexes, where we show how our simple method is able to provide abundant information on the nature of large chromatin fiber containing several bound proteins. The method is able to detect regions of DNA protected to degradation (shown as those with reduced Solvent accessible surface area (SASA)), providing information useful to rationalize footprinting experiments. The server is also able to detect the regions of large distortion in the duplex, the DNA-mediated protein–protein contacts, with the possibility to dynamically select the snapshot(s) with the highest probability of protein cross-talk or the end-to-end distance in the fiber (Fig. 3). Very interestingly, by playing with the protein–protein contacts, the user can explore the impact that dimerization domains might have on the system or the impact that the presence of rigid or flexible linkers might have in determining the structure of the global chromatin fiber.

Discussion

DNA is one of the most complex macromolecules and one of the few research fields where scientists with very different background, from experimental molecular biologists to theoretical physicists, converge. Molecular biologists often rely on very coarse data, coming in most cases from low-resolution biophysical or sequencing experiments. However, in order to derive testable hypothesis, this qualitative information should be transformed into structural models. Theoreticians have derived simulation tools that could be useful to derive structural and dynamical information, but the methods are computationally demanding and require large supercomputers and deep knowledge on simulation techniques and associated software. CGNAW aims to fill the gap between molecular biologists interested in understanding the properties of DNAs and protein–DNA complexes and molecular simulation techniques. We show that the server, which implements a very simple CG simulation engine providing C1′ resolution trajectories (mappable at the atomistic level), is able to provide the end user, at little computational cost and without any need for specific knowledge on simulation, information that can be of direct interest for molecular biologists.

Acknowledgements

We thank the ABC consortium for the atomistic MD simulations used to derive the parameters and all the colleges there

for many discussions on the simulation engine. We thank Prof. Agnes Noy for the atomistic simulations on mini-circles. M.O. is an ICREA Academia Fellow. D.F. was granted a Joan Oró Fellowship by AGAUR, Generalitat de Catalunya, cofinanced by European Social Fund Plus.

Conflict of interest

None declared.

Funding

European Regional Development Fund [ERFD Operative Programme for Catalunya]; Agència de Gestió d'Ajuts Universitaris i de Recerca; BioExcel-3 [101093290]; Dirección Nacional de Innovación, Ciencia y Tecnología [PCI2022-134976-2, PID2021-122478NB-I00]; MDDB (European Commision) [101094561]. This work was funded by the Spanish "Ministerio de Ciencia e Innovación" (PID2021-122478NB-I00), the Center of Excellence for European Commision. "BioExcel-2. Centre of Excellence for Computational Biomolecular Research" [European Union: REF: 101093290 & Ministerio de Ciencia e Innovación: PCI2022-134976-2, and the MDDB project founded by the European Commission: REF.: 101094561. This project is co-funded by the European Regional Development Fund under the framework of the ERFD Operative Programme for Catalunya, the Catalan Government AGAUR (Grups Recerca Consolidats), and the ISCIII "XNA-Hub project. Funding to pay the Open Access publication charges for this article was provided by BioExcel-3 Grant agreement ID: 101093290.

Data availability

The tool is freely accessible at https://mmb.irbbarcelona.org/CGNAW/. A source code useful to perform simulations of naked DNA longer than those accessible from the server can be downloaded from https://github.com/mmb-irb/cgenarate-materials/tree/main and https://doi.org/10.5281/zenodo.15245085.

References

1. Wayment-Steele HK, Ojoawo A, Otten R *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* 2024;**625**:832–9. https://doi.org/10.1038/s41586-023-06832-9
2. Schafer JW, Lee M, Chakravarty D *et al.* Sequence clustering confounds AlphaFold2. *Nature* 2025;**638**:E8–12. https://doi.org/10.1038/s41586-024-08267-2
3. Hospital A, Battistini F, Soliva R *et al.* Surviving the deluge of biosimulation data. *WIREs Comput Mol Sci* 2020;**10**:e1449. https://doi.org/10.1002/wcms.1449
4. Dans PD, Walther J, Gómez H *et al.* Multiscale simulation of DNA. *Curr Opin Struct Biol* 2016;**37**:29–45. https://doi.org/10.1016/j.sbi.2015.11.011
5. Goñi R, Orozco M. The yottaflop frontier of atomistic molecular dynamics simulations. In: Carbó-Dorca R, Chakraborty T (eds), *Theoretical and Quantum Chemistry at the Dawn of the 21st Century*. Oakville, Canada: Apple Academic Press, 2018, 573–92.
6. Marti-Renom MA, Almouzni G, Bickmore WA *et al.* Challenges and guidelines toward 4D nucleome data and model standards. *Nat Genet* 2018;**50**:1352–8. https://doi.org/10.1038/s41588-018-0236-3
7. Farré-Gil D, Arcon JP, Laughton CA *et al.* CGeNArate: a sequence-dependent coarse-grained model of DNA for accurate atomistic MD simulations of kb-long duplexes. *Nucleic Acids Res* 2024;**52**:6791–801. https://doi.org/10.1093/nar/gkae444
8. Naômé A, Laaksonen A, Vercauteren DP. A solvent-mediated coarse-grained model of DNA derived with the systematic Newton inversion method. *J Chem Theory Comput* 2014;**10**:3541–9. https://doi.org/10.1021/ct500222s
9. Freeman GS, Hinckley DM, Lequieu JP *et al.* Coarse-grained modeling of DNA curvature. *J Chem Phys* 2014;**141**:165103. https://doi.org/10.1063/1.4897649
10. Doye JPK, Ouldridge TE, Louis AA *et al.* Coarse-graining DNA for simulations of DNA nanotechnology. *Phys Chem Chem Phys* 2013;**15**:20395. https://doi.org/10.1039/c3cp53545b
11. Assenza S, Pérez R. Accurate sequence-dependent coarse-grained model for conformational and elastic properties of double-stranded DNA. *J Chem Theory Comput* 2022;**18**:3239–56. https://doi.org/10.1021/acs.jctc.2c00138
12. Markegard CB, Fu IW, Reddy KA *et al.* Coarse-grained simulation study of sequence effects on DNA hybridization in a concentrated environment. *J Phys Chem B* 2015;**119**:1823–34. https://doi.org/10.1021/jp509857k
13. Klein F, Soñora M, Helene Santos L *et al.* The SIRAH force field: a suite for simulations of complex biological systems at the coarse-grained and multiscale levels. *J Struct Biol* 2023;**215**:107985. https://doi.org/10.1016/j.jsb.2023.107985
14. Uusitalo JJ, Ingólfsson HI, Akhshi P *et al.* Martini coarse-grained force field: extension to DNA. *J Chem Theory Comput* 2015;**11**:3932–45. https://doi.org/10.1021/acs.jctc.5b00286
15. Cragnolini T, Derreumaux P, Pasquali S. Coarse-grained simulations of RNA and DNA duplexes. *J Phys Chem B* 2013;**117**:8047–60. https://doi.org/10.1021/jp400786b
16. Louison KA, Dryden IL, Laughton CA. GLIMPS: a machine learning approach to resolution transformation for multiscale modeling. *J Chem Theory Comput* 2021;**17**:7930–7. https://doi.org/10.1021/acs.jctc.1c00735
17. Dans PD, Balaceanu A, Pasi M *et al.* The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res* 2019;**47**:11090–102. https://doi.org/10.1093/nar/gkz905
18. Ivani I, Dans PD, Noy A *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat Methods* 2016;**13**:55–8. https://doi.org/10.1038/nmeth.3658
19. Dans PD, Ivani I, Hospital A *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* 2017;**45**:4217–30. https://doi.org/10.1093/nar/gkw1355
20. Hospital A, Andrio P, Cugnasco C *et al.* BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res* 2016;**44**:D272–8. https://doi.org/10.1093/nar/gkv1301
21. Rose AS, Bradley AR, Valasatava Y *et al.* NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* 2018;**34**:3755–8. https://doi.org/10.1093/bioinformatics/bty419
22. Walther J, Dans PD, Balaceanu A *et al.* A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res* 2020;**48**:e29. https://doi.org/10.1093/nar/gkaa015
23. Battistini F, Hospital A, Buitrago D *et al.* How B-DNA dynamics decipher sequence-selective protein recognition. *J Mol Biol* 2019;**431**:3845–59. https://doi.org/10.1016/j.jmb.2019.07.021
24. Lavery R, Moakher M, Maddocks JH *et al.* Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res* 2009;**37**:5917–29. https://doi.org/10.1093/nar/gkp608
25. Andrio P, Hospital A, Conejero J *et al.* BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows. *Sci Data* 2019;**6**:169. https://doi.org/10.1038/s41597-019-0177-4
26. Bayarri G, Andrio P, Hospital A *et al.* BioExcel Building Blocks REST API (BioBB REST API), programmatic access to interoperable biomolecular simulation tools. *Bioinformatics* 2022;**38**:3302–3. https://doi.org/10.1093/bioinformatics/btac316

27. Bayarri G, Andrio P, Hospital A *et al*. BioExcel Building Blocks Workflows (BioBB-Wfs), an integrated web-based platform for biomolecular simulations. *Nucleic Acids Res* 2022;**50**:W99–107. https://doi.org/10.1093/nar/gkac380

28. Noy A, Maxwell A, Harris SA. Interference between triplex and protein binding to distal sites on supercoiled DNA. *Biophys J* 2017;**112**:523–31. https://doi.org/10.1016/j.bpj.2016.12.034

Received: March 24, 2025. **Revised:** April 13, 2025. **Editorial Decision:** April 17, 2025. **Accepted:** April 25, 2025