Biophysical Reviews
https://doi.org/10.1007/s12551-024-01197-3

COMMENT

# MD-DATA: the legacy of the ABC Consortium

**Adam Hospital**[1] · **Modesto Orozco**[1,2]

Received: 5 March 2024 / Accepted: 26 April 2024


**Abstract**
The ABC Consortium has been generating nucleic-acids MD trajectories for more than 20 years. This brief comment highlights the importance of this data for the field, which triggered a number of critical studies, including force-field parameterization and development of new coarse-grained and mesoscopic models. With the world entering into a new data-driven era led by artificial intelligence, where data is becoming more essential than ever, the ABC initiative is leading the way for nucleic acid flexibility.

## Comment

Theoretical approaches to science aim to reduce the impact of serendipity on the advance of our knowledge. However, serendipity often guides the evolution of theoretical sciences. The Ascona B-DNA Consortium (ABC) is an excellent example of how new and unpredicted research objectives emerge when looking for very specific information. Thus, the first ABC round aimed to study the tetramer-dependent properties of B-DNA (Beveridge et al. 2004; Dixit et al. 2005), but unexpectedly the project helped to improve force-fields (FF; (Pérez et al. 2007)). The second round (Pasi et al. 2014) helped to describe bimodality in certain steps, but also resulted in the development of last-generation FFs (Ivani et al. 2016; Zgarbová et al. 2015). The third round aimed to reproduce DNA polymorphism (Dans et al. 2019), but as a side product, the analysis of data led to the description of the kinetics of DNA transitions and the development of a myriad of coarse-grained and mesoscopic models (Walther et al. 2020; López-Güell et al. 2023), which made it possible to move to the chromatin scale (Buitrago et al. 2021). We cannot predict what the impact of the ongoing HexABC project will be beyond characterizing the properties of the 2080 unique hexamers of DNA. The only clear statement that we can make is that stored ABC data will be crucial for it.

Theoretical science is moving from an algorithm-based to a data-driven paradigm. Artificial intelligence methods are anxiously waiting for high-quality data to derive predictive models (Barissi et al. 2022). In this new scenario, we should be careful in reporting MD data with FAIR (findable, accessible, interoperable, and reusable) standards, providing provenance of the trajectories obtained using community-accepted simulation standards and stored after passing severe quality controls (Hospital et al. 2020). We should be prepared to solve computational challenges beyond mere CPU usage and closer to those faced by data-intensive sciences. ABC is pioneering the field: the HexABC consortium has generated 380 validated trajectories covering the 2080 unique DNA hexamers obtained using community-accepted standards. Performing such simulations has been a major effort for the 13 groups involved, but the greatest challenge has been to move around 200 TB of data from production sites to the datacenters in Utah and Barcelona, checking the integrity of the trajectories and detecting potential artefacts in the simulations that require human inspection. Analyzing 500,000 files (200 TB) and storing all the information in a NoSQL database with remote programmatic access

✉ Modesto Orozco
modesto.orozco@irbbarcelona.org

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

[2] Department of Biochemistry and Biomedicine, University of Barcelona, 08028 Barcelona, Spain

Published online: 07 May 2024

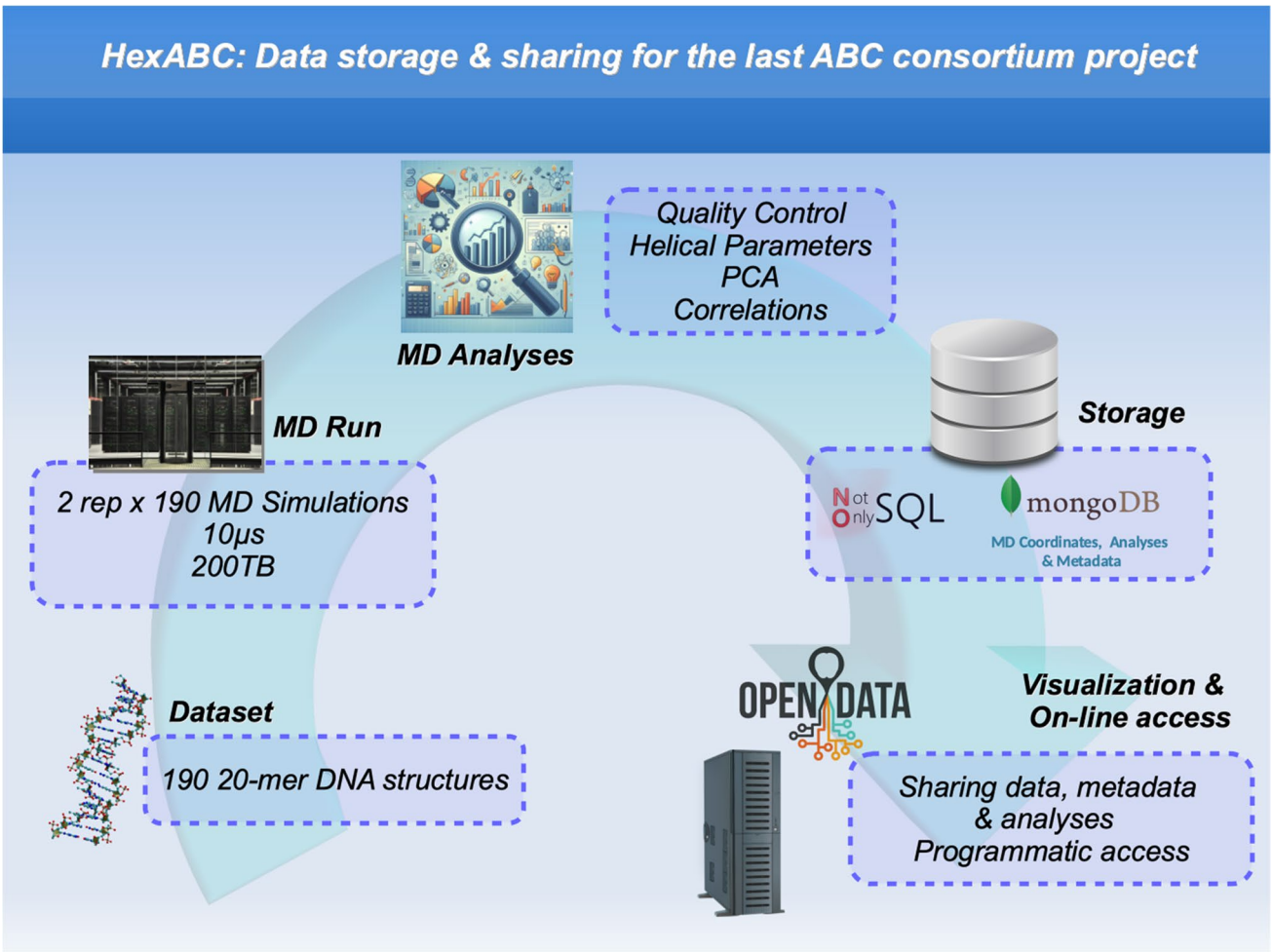


**Fig. 1** The main flow of HexABC data production and storage

represent an effort comparable to that of obtaining the trajectories. However, the final result: a validated database of B-DNA simulations will represent the best legacy of the ABC consortium (Fig. 1).

**Author contribution** M.O. wrote the main manuscript text and A.H. made substantial contributions to the conception or design of the work and reviewed the manuscript.

**Data availability** Dataset generated by the ABC consortium and mentioned in the manuscript file will be made available as open data via the MDDB repository (https://mddbr.eu/).

## Declarations

**Ethics approval and consent to participate** Not applicable.

**Competing interests** The authors declare no competing interests.



## References

Barissi S, Sala A, Wieczór M, Battistini F, Orozco M (2022) DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors. Nucleic Acids Res 50(16):9105–9114. https://doi.org/10.1093/nar/gkac708

Beveridge DL, Barreiro G, Byun KS, Case DA, Cheatham TE, Dixit SB, Giudice E, Lankas F, Lavery R, Maddocks JH, Osman R,

Seibert E, Sklenar H, Stoll G, Thayer KM, Varnai P, Young MA (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. Biophys J 87(6):3799–3813. https://doi.org/10.1529/biophysj.104.045252

Buitrago D, Labrador M, Arcon JP, Lema R, Flores O, Esteve-Codina A, Blanc J, Villegas N, Bellido D, Gut M, Dans PD, Heath SC, Gut IG, Brun Heath I, Orozco M (2021) Impact of DNA methylation on 3D genome structure. Nat Commun 12(1):3243. https://doi.org/10.1038/s41467-021-23142-8

Dans PD, Balaceanu A, Pasi M, Patelli AS, Petkevičiūtė D, Walther J, Hospital A, Bayarri G, Lavery R, Maddocks JH, Orozco M (2019) The static and dynamic structural heterogeneities of B-DNA: extending Calladine-Dickerson rules. Nucleic Acids Res 47(21):11090–11102. https://doi.org/10.1093/nar/gkz905

Dixit SB, Beveridge DL, Case DA, Cheatham TE, Giudice E, Lankas F, Lavery R, Maddocks JH, Osman R, Sklenar H, Thayer KM, Varnai P (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. Biophys J 89(6):3721–3740. https://doi.org/10.1529/biophysj.105.067397

Hospital A, Battistini F, Soliva R, Gelpí JL, Orozco M (2020) Surviving the deluge of biosimulation data. Wires Comput Mol Sci 10(3):e1449. https://doi.org/10.1002/wcms.1449

Ivani I, Dans PD, Noy A, Pérez A, Faustino I, Hospital A, Walther J, Andrio P, Goñi R, Balaceanu A, Portella G, Battistini F, Gelpí JL, González C, Vendruscolo M, Laughton CA, Harris SA, Case DA, Orozco M (2016) Parmbsc1: a refined force field for DNA simulations. Nat Methods 13(1):55–58. https://doi.org/10.1038/nmeth.3658

López-Güell K, Battistini F, Orozco M (2023) Correlated motions in DNA: beyond base-pair step models of DNA flexibility. Nucleic Acids Res 51(6):2633–2640. https://doi.org/10.1093/nar/gkad136

Pasi M, Maddocks JH, Beveridge D, Bishop TC, Case DA, Cheatham T, Dans PD, Jayaram B, Lankas F, Laughton C, Mitchell J, Osman R, Orozco M, Pérez A, Petkevičiūtė D, Spackova N, Sponer J, Zakrzewska K, Lavery R (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. Nucleic Acids Res 42(19):12272–12283. https://doi.org/10.1093/nar/gku855

Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE, Laughton CA, Orozco M (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J 92(11):3817–3829. https://doi.org/10.1529/biophysj.106.097782

Walther J, Dans PD, Balaceanu A, Hospital A, Bayarri G, Orozco M (2020) A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. Nucleic Acids Res 48(5):e29–e29. https://doi.org/10.1093/nar/gkaa015

Zgarbová M, Šponer J, Otyepka M, Cheatham TE, Galindo-Murillo R, 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(12):5723–5736. https://doi.org/10.1021/acs.jctc.5b00716