Topological mechanisms of DNA and DNA-protein interactions

A Data Management Plan created using DMPonline.be

Creator: Midas Segers

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / **URL:** 11O2323N

ID: 198259

Start date: 01-11-2022

End date: 31-10-2024

Project abstract:

The DNA in cells interact with a large number of proteins that exert forces and torques on it. These mechanical stresses induce topological changes in DNA, such as for example, the formation of supercoils, in which the DNA manifests a structure of intertwined entangled loops similar to those formed in overtwisted telephone cords. DNA supercoiling naturally occurs in cells, due to the torsional strain induced by the action of cellular enzymes, and it is relevant for many cellular processes. The aim of this project is to investigate topological aspects of DNA-protein interactions, by means of computer simulations and modeling. Experiments in the past few years have shown that DNA topology plays a very important role in its interaction with proteins, but they have often limited temporal and spatial resolutions. Modeling and simulations provide complementary inputs that are essential to deepen insights on proteins-DNA interactions.

Last modified: 25-04-2023

Topological mechanisms of DNA and DNA-protein interactions
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Topological mechanisms of DNA and DNA-protein interactions
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Topological mechanisms of DNA and DNA-protein interactions
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
PolyMC - code	C++ code	Reuse	Digital	Software	C++ code	< 100MB	
PolyMC - data	PolyMC output	New	Digital	Simulation Data	.txt	< 100 GB	
Plectoneme Dynamics - code	LAMMPS code	New	Digital	Software	LAMMPS code	< 100 MB	
Plectoneme Dynamics - data	LAMMPS output	New	Digital	Simulation Data	.txt	< 1 TB	
Protein compaction of DNA	LAMMPS code	New	Digital	Software	LAMMPS code	< 100 MB	
Protein compaction of DNA - data	LAMMPS output	New	Digital	Simulation Data	.txt	< 1 TB	
SMMC - code	LAMMPS code	Reuse + Adapt	Digital	Software	LAMMPS code	< 100 MB	
SMMC - data	LAMMPS output	New	Digital	Simulation Data	.txt	< 1 TB	
All-Atom Simulations - data	Gromacs	New	Digital	Simlation Data	Various Gromacs files	< 50 TB	
Data Analysis	Python scripts for analysis of simulations	New	Digital	Software	.py	< 1 GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

The PolyMC code is a C++ code was developed by Enrico Skoruppa within our group. His code is freely available on his Github: https://github.com/eskoruppa/PolyMC

SMCC model is a LAMMPS code developed by Stefanos Nomidis who was a PhD student in our group. His LAMMPS code is available on his Github page: https://github.com/sknomidis/SMC_LAMMPS

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Use within research group:

- **Software**: Software is stored on the KU Leuven server. The different software are saved in separate folders, e.g. *home/software_name/name_executable*
  Code files (C++, LAMMPS code and Python files) are annotated for clarity and reusability.
- **Data:** The data is stored in a subfolder of the folder containing the software. In addition different subfolders will be used for distinctive simulation procedures. Data storing files have filenames provide information on the relevant variables used during the simulation (variablename = VAR1, variablevalue=var1). If too many variables are needed we add an additional file with the same name but with a custom extension (.var). This file contains all variables in a text-format necessary to rerun a simulation. We use different (custom) file-extensions which identify the data contained in a file (e.g. .endlink, . in, ... for the PolyMC software).

  Datafile: home/software_name/results/procedure/data/VAR1_var1_VAR2_var2.ext
  Variablefile: home/software_name/results/procedure/data/VAR1_var1_VAR2_var2.var
  If deemed necessary a .txt file explaining the naming convention might be added.

Public:

- **Software**: Software will be made publicly available upon publication on my Github-page (https://github.com/SegersMidas). Software will be accompanied with a README.txt in addition to exemplary scripts illustrating possible different simulation procedures. Scripts will contain in-line comments to enhance clarity.
- **Simulation Data:** The data generated throughout this project mainly consist of molecular dynamics trajectories which are memory demanding (> TB). Due to memory requirements trajectories will not be made publicly available. However, relevant quantities are the result of statistical analysis on a multitude of such trajectories. The details of this analysis will be reported on in publications and if deemed useful the developed python libraries will be made available on my Github-page (https://github.com/SegersMidas). The python files will be annotated for readability.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

3. Data storage & back-up during the research project

Where will the data be stored?

Data is stored on the data-server of the department of Physics and Astronomy.

How will the data be backed up?

Back-ups are made daily. In case of memory loss, data can be recovered up to a month after the loss.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

The global departmental data server has storage capacity in the order of petabytes and is still expanding. There is no strict limit on data storage for individual members/groups of the department. Upon consultation with the departmental IT-service the storage required for this project was deemed possible and without any issues.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The departmental server has security at the user and research group level. Data at the user-level can only be accessed by the user, whilst data at the research group-level can be accessed by all members of the group, both require a user-login to avoid access by thirds.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The departmental data server is funded by the department of Physics and Astronomy.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Software developed during this project and simulation data thereof are stored on the departmental data-server. After the end of the project all data will be preserved for at least 10 years.
The developed software will also be stored in Github repositories (https://github.com/SegersMidas) which allow storage up to 5GB which is sufficient for this purpose.

Where will these data be archived (stored and curated for the long-term)?

Software developed during this project and simulation data thereof are stored on the departmental data-server.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The departmental data server is funded by the department of Physics and Astronomy.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, …)

Open Access repository: Software developed throughout the project will be made freely available in a Github repository (https://github.com/SegersMidas).

Restricted access repository: Simulation data stored on the departmental data server will be shared with all members of our research group. Access can be extended to other members of the department.

If access is restricted, please specify who will be able to access the data and under what conditions.

Restricted access repository: Simulation data stored on the departmental data server will be shared with all members of our research group. Access can be extended to other members of the department using the *share*-folder or in consultation with the project supervisor Enrico Carlon and the departmental IT-service.
For third-party members from outside the department or institution simulation data can be exchanged after publication and upon contacting with my supervisor and me for research purposes.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Software developed throughout the project will be made freely available in a Github repository (https://github.com/SegersMidas).
Simulation data can be made available to individuals after publishing upon contact.

When will the data be made available?

If applicable, data will be made available upon publication of research results.

Which data usage licenses are you going to provide? If none, please explain why.

Developed software is made freely available on Github for download and can be modified or shared. When used in scientific work the original creator ought to be credited. Hence the Creative Commons Attribution 4.0 International (CC BY 4.0) license seems suitable.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

Given that the data server is maintained by the department of Physics and Astronomy, we foresee no additional costs related to data sharing.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Midas Segers

Who will manage data storage and backup during the research project?

Midas Segers

Who will manage data preservation and sharing?

Midas Segers

Who will update and implement this DMP?

Midas Segers