Data in Brief 36 (2021) 107109

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A homogeneous dataset of polyglutamine and glutamine rich aggregating peptides simulations

Exequiel E. Barrera[a,b], Sergio Pantano[b,c,*], Francesco Zonta[c,*]

[a] Instituto de Histología y Embriología (IHEM) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CC56, Universidad Nacional de Cuyo (UNCuyo), Mendoza, Argentina
[b] Biomolecular Simulations Group, Institut Pasteur de Montevideo, Mataojo 2020, CP 11400 Montevideo, Uruguay
[c] Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, Shanghai 201210, China

ARTICLE INFO

Article history:
Received 10 March 2021
Revised 13 April 2021
Accepted 20 April 2021
Available online 30 April 2021

Keywords:
SIRAH
Coarse-grained simulation
Soluble oligomer
polyQ
Q-rich
Aggregation
Oligomerization
Molecular dynamics

ABSTRACT

This dataset contains a collection of molecular dynamics (MD) simulations of polyglutamine (polyQ) and glutamine-rich (Q-rich) peptides in the multi-microsecond timescale. Primary data from coarse-grained simulations performed using the SIRAH force field has been processed to provide fully atomistic coordinates. The dataset encloses MD trajectories of polyQs of 4 (Q4), 11 (Q11), and 36 (Q36) amino acids long. In the case of Q11, simulations in presence of Q5 and QEQQQ peptides, which modulate aggregation, are also included. The dataset also comprises MD trajectories of the gliadin related p31-43 peptide, and Insulin's C-peptide at pH=7 and pH=3.2, which constitute examples of Q-rich and Q-poor aggregating peptides. The dataset grants molecular insights on the role of glutamines in spontaneous and unbiased ab-initio aggregation of a series of peptides using a homogeneous set of simulations [1]. The trajectory files are provided in Protein Data Bank (PDB) format containing the Cartesian coordinates of all heavy atoms in the aggregating peptides. Further analyses of the trajectories can be performed directly using any molecular visualization/analysis software suites.

DOI of original article: 10.1016/j.csbj.2021.02.014

* Corresponding authors.

E-mail addresses: spantano@pasteur.edu.uy (S. Pantano), fzonta@shanghaitech.edu.cn (F. Zonta).

https://doi.org/10.1016/j.dib.2021.107109

Specifications Table

Subject	Biological Sciences.
Specific subject area	Protein Biophysics. Molecular dynamics simulations of aggregating peptides.
Type of data	Secondary Data. Molecular dynamics trajectories of multiple peptide systems.
How data were acquired	Hardware: CPU (Intel Core i7-5930K, 3.5 GHz) accelerated with a TitanX GPU. Software: Gromacs 2018.4 using the SIRAH 2.0 force-field for performing MD simulations and SIRAH Tools, along with AmberTools 2018 and Amber14SB force-field implemented in VMD 1.9.3 for backmapping.
Data format	Filtered.
Parameters for data collection	MD simulations were performed at 300K and 1 Bar for multiple microseconds. Full details of all simulations are reported in Table 1.
Description of data collection	Raw molecular dynamics data at coarse-grained level was filtered to maintain one every ten steps and and protein's heavy atoms were backmapped using SIRAH Tools. Simulation frames are reported every 1 ns of simulation.
Data source location	Primary Data was collected at the Uruguayan Center for Supercomputation (ClusterUY).
Data accessibility	Repository name: Mendeley Data Direct URL to data: https://data.mendeley.com/datasets/2tmsbchh42/2 Instructions for accessing these data: Data is freely accessible.
Related research article	The primary data source consists of a set of coarse-grained MD simulations. They are described in the associated manuscript "Dissecting the role of glutamine in seeding peptide aggregation" by E. E. Barrera, F. Zonta, and S. Pantano, Computational and Structural Biotechnology Journal, 2021, DOI: https://doi.org/10.1016/j.csbj.2021.02.014

Value of the Data

- Homogeneous sets of simulations on different aggregating peptides on multimicroseconds timescale are very rare in the literature. Analysis of this dataset can provide valuable insights obviating the lengthy process of generating the data from the scratch.
- Data of interest to computational biophysicist/biochemists studying peptide aggregation.
- Molecular coordinates can be read/analyzed with standard software for structural biology or molecular visualization.

1. Data Description

The dataset is deposited on Mendeley data with the doi: 10.17632/2tmsbchh42.2. It contains two .zip files (one for the polyglutamine peptides and another for the Q-rich peptides) enclosing separated files for each peptide trajectory. The peptide composition and specifics of each system, and name of individual data trajectories are reported in Table 1. This dataset contains eight files of molecular trajectories of different peptides in Protein Data Bank (pdb) format that can be visualized/analyzed with standard molecular visualization/simulation programs.

Table 1
Summary of the systems simulated.

Peptide	Monomers in the box	Box size (nm)	Peptide concentration (mM)	Protonation state in the termini	Length of the Simulations (µs)	Name of the file in Mendeley data
Q4	27	11.5 (cubic)	29.4	neutral	3	Q4_agg_5us.pdb
Q11	10	13.5 (cubic)	6.7	neutral	5	Q11_agg_5us.pdb
Q11 + Q5	20	13.5 (cubic)	13.4	neutral	5	Q11-QQQQQ_agg_5us.pdb
Q11 + QEQQQ	20	13.5 (cubic)	13.4	neutral	5	Q11-QEQQQ_agg_5us.pdb
Q36	3	13.5 (cubic)	2.1	neutral	5	Q36_agg_5us.pdb
p31-43	50	23 × 22 × 19 (octahedral)	8.4	neutral	5	p31-43_agg_5us.pdb
C-peptide	30	23 × 22 × 19 (octahedral)	5.1	zwitterionic	5	C-peptide_agg_pH7_5us.pdb
C-peptide	30	23 × 22 × 19 (octahedral)	5.1	N-terminal (+) C-terminal (neutral)	5	C-peptide_agg_pH3.2_5us.pdb

2. Experimental Design, Materials and Methods

2.1. Primary data

A detailed description of the protocol followed to generate the primary data is reported in the associated paper [1]. Briefly, for each system we started from fully atomistic peptide copies that were uniformly distributed in simulation boxes listed in Table 1. Systems were mapped to coarse-grain using SIRAH Tools [2], and solvated. In the simulations of the C-peptide at pH = 7 and pH = 3.2, KCl ions were added to a concentration of 150 mM. MD simulations were performed in the NPT ensemble at 300 K and 1 atm using the SIRAH force field version 2.0 [3] using GROMACS 2018.4 as simulation engine [4].

2.2. Secondary data

The secondary data consists of the trajectories of the peptides reported in Table 1 backmapped to fully atomistic representation. This will allow to interested scientist to run straightforwardly further analyses using standard simulation/structural biology tools obviating the significant computational cost associated to the generation of the data and facilitate the interpretation of the coarse-grained representation to non-experts. Backmapping was performed using SIRAH Tools [2]. To this aim we used a tcl script included in the distribution that can be loaded on the popular molecular visualization software named VMD 1.9.3 [5] Once the coarse-grained trajectories are loaded, they are processed one frame at the time. Since the simplified SIRAH representation preserves the position of a few atoms in each residue, individual simulation frames were taken separately and missing atoms were first added using internal coordinates residue by residue. The reconstructed molecules were then loaded to the tleap module of Amber18 [6] to generate individual topology and coordinates. Subsequently, these coordinates underwent an all-atoms energy minimization in vacuum with a cut off of 1.2 nm using the sander module of Amber18 and the Amber14SB force field [7]. We performed 50 steps of energy minimization using the steepest descent algorithm followed by 100 steps using conjugated gradient Finally, atomistic structures were concatenated and saved into one single trajectory files. Each of the trajectory files listed in Table 1 contains one frame per ns. To preserve the portability of the dataset, only the trajectories containing the heavy atoms of the peptides are reported in the database. It is important to notice that the above-described process is integrated in SIRAH tools and executed with a command line from the VMD console.

Ethics Statement

Not applicable.

CRediT Author Statement

E.E. Barrera: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Drafting the manuscript, Revising the manuscript critically for important intellectual content; **S. Pantano:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Drafting the manuscript, Revising the manuscript critically for important intellectual content; **F. Zonta:** Analysis and/or interpretation of data, Drafting the manuscript, Revising the manuscript critically for important intellectual content.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships, which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

This work was partially funded by FOCEM (MERCOSUR Structural Convergence Fund), COF 03/11 and the National Natural Science Foundation of China (NSFC grant 31770776 to FZ). EEB and SP belong to the SNI program of ANII. Simulations were performed on the National Uruguayan Center for Supercomputing, ClusterUY.

References

[1] E.E. Barrera, F. Zonta, S. Pantano, Dissecting the role of glutamine in seeding peptide aggregation, Comput. Struct. Biotechnol. J. 19 (2021) 1595–1602, doi:10.1016/j.csbj.2021.02.014.

[2] M. Machado, S. Pantano, SIRAH Tools: mapping, backmapping and vis- ualization of coarse-grained models, Bioinformatics 32 (2016) 2–3, doi:10.1093/bioinformatics/btw020.

[3] M.R. Machado, E.E. Barrera, F. Klein, M. Sóñora, S. Silva, S. Pantano, The SIRAH 2.0 force field: altius, fortius, citius, J. Chem. Theory Comput. 15 (2019) 2719–2733, doi:10.1021/acs.jctc.9b00006.

[4] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, et al., Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX 1–2 (2015) 19–25, doi:10.1016/j.softx.2015.06.001.

[5] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1996) 33–38, doi:10.1016/0263-7855(96)00018-5.

[6] D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, P.A. Kollman, AMBER 2018, University of California, San Francisco, 2018.

[7] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, C. Simmerling, ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB, J. Chem. Theory Comput. 11 (2015) 3696–3713, doi:10.1021/acs.jctc.5b00255.