AN INTEGRATED WEB-BASED INTERACTIVE DATA PLATFORM FOR MOLECULAR DYNAMICS SIMULATIONS

HRACHYA ASTSATRYAN, WAHI NARSISIAN, ELIZA GYULGYULYAN*, ARMEN POGHOSYAN †, AND YEVGENI MAMASAKHLISOV ‡

Abstract. The main aim of the article is to introduce an integrated web-based interactive data platform for molecular dynamic simulations using the datasets generated by the Armenian life science communities. The suggested platform, consists of data repository and workflow management services, is vital for current and future scientific discoveries. We will focus on interactive data visualization workflow service as a key to perform more in-depth analyzing of research data outputs, help to understand the problems efficiently and to consolidate the data into one collective illustration platform. The integrated data platform will be presented as an advanced integrated environment to capture, analyze, process and visualize the research data.

Key words. molecular dynamics simulations, high-performance computing, persistent identifier, Web based interactive visualization, DNA, biological systems.

1. Introduction. Modelling and numerical simulation - considered to be the third pillar of the science [1] after theory and experimentation is at the heart of multiple domains, which are not only scientific, but also societal (e.g., energy, health, environment), economic, financial (e.g., industrial competitiveness), and life ethics (e.g., biology). They also appear increasingly as decision-making tools for critical cases like global warming, natural disasters, etc. Since modeling and simulation are essential for many scientific advances, the control of all the aspects of high-performance computing (HPC) - as well as the capacity to exploit the masses of data to tackle the solution of these complex models - is inescapable.

The explosion in computational power help the biologists and life science researchers to conduct more advanced experiments and in its turn lead to a rapid increase in the amount of experimental data, research outputs, etc. [2]. As computational experiments, molecular dynamics (MD) simulations are widely used in the domain of life sciences to evaluate the equilibrium nature of classical many-body systems [3, 4]. The study of systems with a large number of atoms in long trajectory intervals (from nano to milliseconds) is required to explore a broad range of exciting phenomena, which is undoubtedly unfeasible without using appropriate HPC resources and storage facilities to manage and visualize these data.

In Armenia, several communities from life science domain use HPC resources and generate a significant amount of research outputs by storing them in local repositories. Usually, such local datasets are incomplete, and there is need to cluster them into the central repository by managing this data using appropriate metadata and identifiers. Because there is no centralized repository to hold all these data, the data sharing between these communities is almost impossible or a challenge. This also leads to being not able to have a single platform to process and visualize this data. The volume, complexity, and heterogeneity of data originating from these organizations have created challenges to have a complete understanding of complex biological processes and systems [5].

*Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia, 1, P. Sevak str., 0014 Yerevan, Armenia (hrach@sci.am).

†International Scientific Educational Center of the National Academy of Sciences of the Republic of Armenia, 24D, M. Baghramyan ave., 0019 Yerevan, Armenia.

‡Yerevan State University, 1 A. Manoogian str., 0015 Yerevan, Armenia.

The main aim of the article is to introduce an integrated web-based interactive data platform for molecular dynamic simulations using the datasets generated by the Armenian life science communities. The suggested platform, consists of data repository and workflow management services, is vital for current and future scientific discoveries. We will focus on interactive data visualization workflow service as a key to perform more in-depth analyzing of research data outputs, help to understand the problems efficiently and to consolidate the data into one collective illustration platform. The integrated data platform will be presented as an advanced integrated environment to capture, analyze, process and visualize the research data.

The remainder of this paper is divided into the following sections: section 2 introduces the life science communities and applications in Armenia widely using MD simulations, section 3 represents the integrated data platform, section 4 workflow services and finally section 5 is the conclusion.

2. Life science communities and applications in Armenia. The life science communities in Armenia produce a significant amount of data and widely use HPC resources. We will concentrate on the scientific outputs and datasets producing by the Bioinformatics Group of the International Scientific and Educational Center of the National Academy of Sciences of the Republic of Armenia (Bioinformatics group) and the Molecular Physics department of the Yerevan State University (MolPhys YSU). The complex systems, including surfactant, polymer, and protein, can be found in everyday life, as well as in many industrial applications, such as in pharmaceuticals, food processing or agrochemicals [6]. The interaction between surfactant and protein/polymer plays a critical role in many physical processes and opens up a wide range of commercial applications, including cosmetic formulations [7] or drug delivery systems. Moreover, such kinds of complex systems have been extensively investigated for many years as model systems for biological membranes [8] being of vital importance in studies of cell membranes. The state-of-the-art simulations can play an important role by offering a detailed picture of the structure and dynamics of complex systems by improving our knowledge and understanding of many interesting phenomena. The efficiency of such kind of longtime simulations can be reached via HPC platforms, as many interesting phenomena occur at nano-to-milliseconds time scale, and require massively large systems with many atoms to mimic real system. Although, it is also essential and necessary to analyze MD simulations with experimental data for validity ensuring. A series of papers have been published by the Bioinformatics group, where complex systems were investigated using classical MD simulation method [9, 10, 11] getting a deeper insight into dynamic processes occurring on longtime scale range. Note that the mentioned investigations have been done using various large-scale HPC platforms. The nucleic acids, e.g. double stranded DNA and single stranded RNA molecules play an important role in the living systems functionality, biomedical research, biological sensors development, etc. For example, hybridization process is a keystone of many essential processes, including transcription, replication, polymerase chain reaction, DNA sensors functioning, etc. Thermodynamics and kinetics of the nucleic acids hybridization have been extensively investigated both on the surface and in bulk [12, 13, 14]. The all-atom and coarse-grained simulations can give a substantial impact on the understanding of hybridization thermodynamics and kinetics. Such large-scale simulations require HPC platform and massively parallelized MD computations. A series of papers concerning single and double-stranded nucleic acids have been published by MolPhys YSU [15, 16, 17], which require validation using MD simulations. The combination of experimental and various computational methods

can give us a deep understanding of the complex processes, behind the nucleic acids hybridization.

3. Integrated Web-based interactive data platform. The suggest integrated web-based interactive data platform gives users a possibility to have a single entry point to different services. First, it enables to store and manage the output data of diverse molecules simulation and assign meta-data to each output to increase the data reliability. Also, the platform allows searching any required molecules type and downloading it. For the visualization, an interactive web-based solution is provided to use the web browser directly to have a quick look at the molecule structure and based on that decide to use it or not. And finally, the public link is provided to enable data downloading if the user does not have an account on iRODS (Integrated Rule-Oriented Data System) [18].

The platform consists of the following layers, which are illustrated in fig. 3.1.

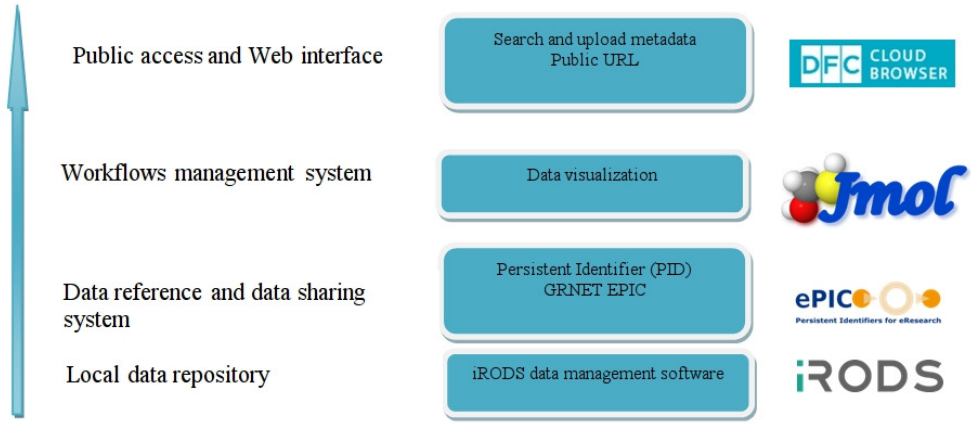

FIG. 3.1. *Diagram of the integrated web-based interactive data platform.*

3.1. Local data repository. The local data repository is the most common place for preserving digital research data. A repository is an online database service, which archives and stores the data, also provides a possibility to discover and access the data. The repository offers several features and benefits:

- To preserve the data for future work.
- To assign metadata and persistent identifier for each data which in its turn increase its reliability when others will use it.
- To increase the data discovery over the net.
- To prevent the users from maintaining the data by themselves, because the professional administrators will maintain the data.
- To enable data sharing between different communities.

The most significant benefit of these repositories is sharing the data opening a lot of new roads for research, collaboration, increase the research visibility, usage of high-quality data collected by other researchers, etc. Some repositories have restricted constraints and policies about how to use or store the data, and the user needs to register to be able to use the stored data. The data repository platform has been developed using the iRODS, which is an open-source data management software. iRODS provides a rule-based system management capabilities making data replication more

comfortable and provides extra data protection. The metadata system of iRODS is comprehensive and allows users to customize their application level metadata, instead of the metadata supplied by traditional file systems. The types of datasets mainly generated by the Bioinformatics and MolPhys YSU groups include:

- **trr file format** - the trajectory of a simulation including all the coordinates, velocities, forces and energies.
- **gro file format** - a molecular structure in Gromos87 format. gro files, as trajectory by simply concatenating files.
- **xtc file format** - a portable format for trajectories. It uses the xdr routines for writing and reading data which was created for the Unix file system.
- **pdb file format** - molecular structure files in the protein databank file format. The protein databank file format describes the positions of atoms in a molecular structure. Coordinates are read from the ATOM and HETATM records until the file ends or an ENDMDL record is encountered.
- **psf file format** - contain atoms, residues, segment names, residue types, atomic mass and charge, and the bond connectivity.

All in all the metadata is stored in Mysql database, and all related data is stored in the replicated storage to ensure the availability of the resources in case of any damage or failover problem.

3.2. Data reference and data sharing. As the amount of the digital data rises exponentially the relations between them becoming more and more essential and as data repositories are various by their size and format, it can be vital every data in data repositories to change its physical location which can cause a loss of the link to that data. It is commonly known that the Uniform Resource Locator (URL) of a data is not a permanent link to the location of the data and when the physical location of the data is changed the URL also should be changed everywhere, which can lead to a miss reaching of that particular data. For this reason, scientific institutions need a long-term preservation of resources with long-term accessibility. Persistent Identifier (PID) [19] is used for is a long-lasting data reference and sharing, which has two components: a unique identifier; and a service that locates the resource over time also when it's location changes. PID guarantees reliability in citing sources even if the URL changes. PID consists of a prefix that is globally unique within the context of the system providing the PID and suffix, which is unique within the local organization. The 21.15104 unique prefix is used for the platform and a suffix started by ASNET is automatically generated per dataset. The PIDs are generated and registered by data centers enabled through European Persistent Identifier Consortium (EPIC). EPIC provides PID services using handle systems for the European Research Community to allocate and to resolve persistent identifiers. The handle system of GRNET (Greek Research and Technology Network) is used to generate PIDs per each simulation with the prefix provided by GRNET Handle Restful web service. As the EPIC API supports the automatic generation of a local name, the suffix of PID is generated and executed automatically with a POST HTTP request in curl. This enables the opportunity to reach our data even if we change the physical address. We also use EPIC PID for the visualization described in the section below.

3.3. Workflow management system. Research and scientific processes in biology heavily use workflow systems to have all necessary steps to address the complexity of any scientific experiment and to enhance the discovery of new methods and solutions based on the execution of complex algorithms together with the access and analysis of experimental data. All this help to produce more accurate and reliable

results and outcomes, which can confirm real experiments or provide a proper and deep understanding of several processes. With the availability of HPC resources and different cloud services [20], the biologists in Armenia can run complex workflows that integrate programs, methods, and data from various resources and run different simulations in a single consolidated platform. Using different scientific computing methods with the support of scientific discovery development, a new area of scientific methods arise with new data analysis strategies enabled which are called e-science paradigm. The ultimate goal of the deployed platform is to provide a complete solution of life science scientists to conduct different workflows based on their experiments output. The suggested solution is to have a molecular visualization to be able to check the shape and the construction of the molecule before conducting any experiment. The service enables the users to upload their molecules, or use the sample uploaded on iRODS, or even use any public URL, which refers to some molecule. This will enable users to have more insight and understanding about the molecule and use it in any similar experiment.

3.4. Public access and Web interface. At a top layer, the data visualization layer is used, which is a web interface for molecules visualization (see fig. 3.2). This interface is based on Jmol, which is an open-source browser-based HTML5 viewer and stand-alone Java viewer for chemical structures in 3D [21]. Since it is written in the Java programming language, it is compatible with all major operating systems and, in the applet form, with most modern web browsers. Two beneficial features

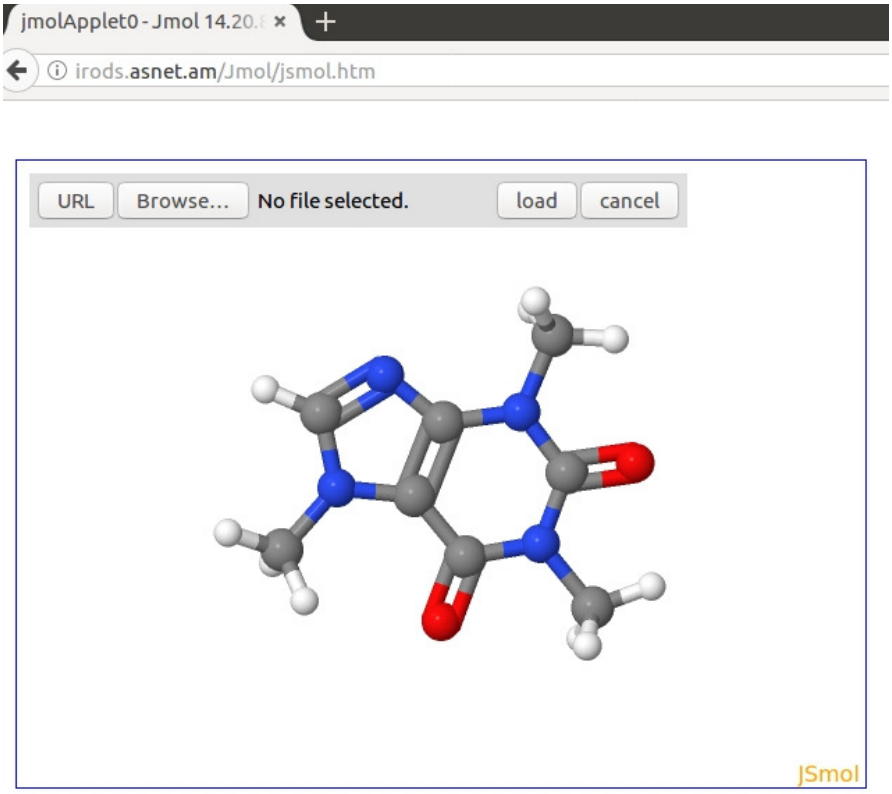

FIG. 3.2. *Data visualization interface.*

of this system have been customized, which allows users directly check the molecules stored in the iRODS repository using its public URL link, and upload the molecules from the personal computer and visualize it. The interface enables to colorize the molecules based on their type, animate them, add labels etc. At the top level, there is a web interface, which is a Cloud browser developed by DICE group [22]. The web interface simplifies the researchers work to not think about where to store the data and corresponding metadata, the researcher need only to upload the data and then fill the corresponding sections of the metadata. A new template is provided for each community containing the relevant fields in order to gather or input all required information for each uploaded data. Then is also a separate service that can be accessed from the above-mentioned web interface. It is a Public link to all stored data on iRODS. This link will enable any user to use the data on iRODS repository without a need to do a registration, which in its turn increase the usage of the data.

4. Demonstration and discussion. To perform data analysis in biology workflows are very useful. The workflows management systems help biologists and other users to create and run different experiments without concerning about the programming and what is going on at the back-end of the system. In bioinformatics, these systems usually concentrate more on a visual representation of molecules using graphical user interface, which in its turn enable scientists to run several scientific tasks in parallel and visualize the results in order to get more details and proper information about the experiments outcome. The 3D presentation or visualization of studied systems is an important key to better understand the intra-molecular structure and get insight from MD simulation. Many complex biological systems have been studied using the platform, for instance, a system, consist of a cationic poly (dially ldimethy lammonium chloride) (PDADMAC)/ sodium dodecyl sulfate (SDS)/Decanol and aqueous solution. Such kinds of systems are widely used in different fields, such as pharmaceutics or cosmetics, as well as for the synthesis of nanostructured material [23]. The mentioned system was extracted from our MD simulation. The visualization capabilities of the given system are illustrated in fig. 4.1. The visual presentation

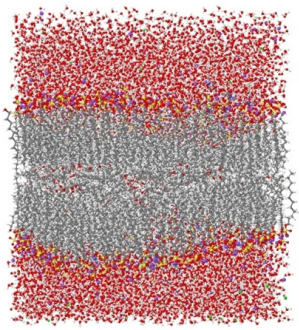

FIG. 4.1. *PDADMAC/SDS/Decanol in water bulk.*

makes it possible to reveal the polymer absorption features on SDS bilayer, as well as, the information coming from the decanol molecules' orientation. One can track that the decanols, which are located between the SDS methyl groups, are mostly in upright position. The vital information coming from visualization is that the PDADMAC molecules in two layers have different conformations, i.e. a more folded and a more flat conformation. Note that the MD results are in full agreement with our

experimental findings. Thus, the visualization of the systems gives us an information about the coexistence of two lamellar phases in surfactant-based systems induced by polyelectrolyte., as well as, about the lamella features, such as undulations, etc.

Using a present platform such an important processes as force pulling and nucleic acids hybridization have been addressed. Pulling simulation of helical B-DNA with the sequence d(CGCAAATTTCGC)2 has been performed. We addressed system, containing 417688 atoms, including dsDNA, water molecules, and 100 mM NaCl. We also considered the same system, containing MgCl2 instead of NaCl. The system under consideration is presented in fig. 4.2. During MD simulation the double stranded

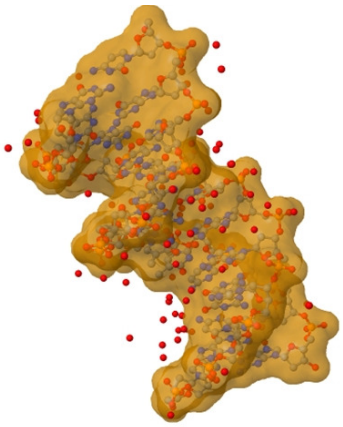

FIG. 4.2. *d(CGCAAATTTCGC)2 in presence Na+ and Cl- ions, Water molecules are hidden.*

DNA molecule was pulled by external force and the free energy of double stranded DNA was measured directly. The typical final confirmation of the DNA molecule is presented in fig. 4.3. The visual presentation makes it possible to observe the single strands separation of the double stranded DNA caused by external pulling.

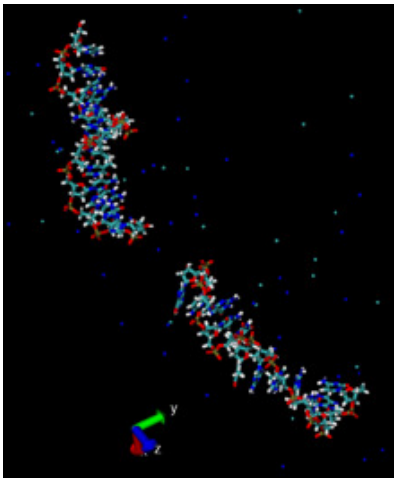

FIG. 4.3. *The strands of d(CGCAAATTTCGC)2 are separated*

5. Conclusion. Workflow systems is a basic model or pattern that provide support for research and scientific experiments by containing the all necessary steps of discovery based on the execution of different simulations and having the possibility to visualize the results in order to get a better understanding for the outputs. The deployed infrastructure gives the biologists in Armenia and beyond the possibility to increase the visibility of the actual laboratory processes, help them in examining the processes' impact on each other and which activities have more influence on the whole process. The system also gives a possibility to understand the relationship of the small processes in a larger system and how they interact with each other. Having the single point for all distributed data in Armenia enable researchers to collaborate more easily and to share their knowledge. Using the system help to unfold the complexity of any scientific problem and their domain, also to identify the redundancy of the conducted steps in order to avoid them in the future. There is a future plan to expand the system in order to contain pre-defined workflows for different activities in the domain of biology, also to enhance the system with more visualization features such as a visual comparison between two elements, display different animation etc.

Acknowledgments. The research leading to these results has been co-funded by the European Commission under the H2020 Research Infrastructures contract no. 675121 (project VI-SEEM) and the "Persistent Identifier Services for the life science community in Armenia" contract no. 653194 (RDA Europe).

REFERENCES

[1] X. Yang, L. Wang,G. von Laszewski, *Recent Research Advances in e-Science, Cluster Compute 12*, (2009), pp. 353–356.
[2] Z. Yin, H. Lan, G. Tan, M. Lu, A. V.Vasilakos, W. Liu, *Computing Platforms for Big Biological Data Analytics*, Perspectives and Challenges, Computational and Structural Biotechnology Journal, Volume 15 (2017), pp. 403–411.
[3] B. Alder, T. Wainwright, *Studies in Molecular Dynamics. I. General Method*, J. Chem. Phys. Vol. 31 (1959), pp. 459.
[4] M. Tuckerman, G. Martyna, *modern molecular dynamics methods: Techniques and Applications*, J. Chem. Phys. Vol. 104 (2000), pp. 159–178.
[5] E. Deelman, D. Gannon, M. Shields, *An overview of workflow system features and capabilities. Future Generation Computer Systems* , Workflows and e-science (2009), pp. 528–540.
[6] J.C.T Kwak, *Polymer-Surfactant System* , Marcel Dekker; New York, Surfactant Science Series volume 77 (1998)
[7] M.M. Rieger, L.D. Rhein, *In Surfactants in Cosmetics* , Marcel Dekker; New York, Surfactant Science Series volume 68 (1997)
[8] V. Luzzati, *X-ray diffraction studies of lipid-water systems*, In Biological Membranes, ed. by D. Chapman. Academic Press, London. 1 (1968), pp. 71–123.
[9] A. Poghosyan, L. Arsenyan, H. Astsatryan, *Dynamic Features of Complex Systems*, A Molecular Simulation Study - Modeling and Optimization in Science and Technologies. Vol. 2 (2014), pp. 117–121.
[10] A. Poghosyan, L. Arsenyan, A. Shahinyan, J. Koetz, *Polyethyleneimine Loaded Inverse SDS Micelle in Pentanol/Toluene Media*, Colloids and Surfaces A: Physicochemical and Engineering Aspects (2016), pp. 402–408.
[11] A. Poghosyan, H. Astsatryan, A. Shahinyan, *Parallel Peculiarities and Performance of GROMACS Package on HPC Platforms*, International Journal of Scientific and Engineering Research (2013), pp. 1755–1761.
[12] E. Arslan, J. Laurenzi, *An efficient algorithm for the stochastic simulation of the hybridization of DNA to microarrays*, BMC Bioinformatics 10 (2009), pp. 411–427.
[13] I. Wong, N. Melosh, *An Electrostatic Model for DNA Surface Hybridization*, Biophys. J 98 (2010), pp. 2954–2963.
[14] T. Schmitt, B. Rogers, T. Knotts, *Exploring the mechanisms of DNA hybridization on a surface*, J. Chem. Phys 138 (2013), pp. 1755–1761.

[15] A. Karapetian, Z. Grigoryan, Y. Mamasakhlisov, M. Minasyants, P. Vardevanyan, *Theoretical treatment of helixcoil transition of complexes DNA with two different ligands having different binding parameters*, J. Biomol. Struct. Dyn. 34 (2016), pp. 201–205.

[16] G. Hayrapetyan, F. Iannelli, J. Lekscha, V. Morozov, R. Netz, Y. Mamasakhlisov, *Cold melting of RNA with quenched sequence randomness*, Phys. Rev. 113 (2014).

[17] Y. Mamasakhlisov, Sh. Hayryan, V. Morozov, C. Hu, *Kinetics of the long ssRNA: Steady state*, Europhys. Lett. 106 (2014).

[18] Xu, Hao and Russell, Terrell and Coposky, Jason, *iRODS Primer 2: Integrated Rule-Oriented Data System*, Morgan and Claypool (2017).

[19] W. Cockshot, M. Atkinson, K. Chisholm, P. Bailey, R. Morrison, *Persistent object management system*, Softw: Pract. Exper., 14 pp. 49–71.

[20] H. Astsatryan, V. Sahakyan, Y. Shoukourian, P. Cros, M. Dayde, J. Dongarra, P. Oster, *Strengthening Compute and Data intensive Capacities of Armenia*, IEEE Proceedings of 14th RoEduNet International Conference - Networking in Education and Research (2015), pp. 28–23.

[21] A. Herrez, *Jmol to the rescue, Biochemistry and Molecular Biology Education*, Journal of Colloid and Interface Science(2011), pp. 255–261.

[22] https://github.com/DICE-UNC/irods-cloud-browser

[23] A. Poghosyan, L. Arsenyan, J. Koetz, A. Shahinyan, *Molecular dynamics study of poly diallyldimethylammonium chloride (PDADMAC)/sodium dodecyl sulfate (SDS)/decanol/water system*, The Journal of Physical Chemistry B.(2009), pp. 1303–1310.