INTERFACE FOCUS

royalsocietypublishing.org/journal/rsfs

Discussion

Cite this article: Riccardi E, Pantano S, Potestio R. 2019 Envisioning data sharing for the biocomputing community. *Interface Focus* **9**: 20190005.
http://dx.doi.org/10.1098/rsfs.2019.0005

Accepted: 4 March 2019

One contribution of 15 to a theme issue 'Multi-resolution simulations of intracellular processes'.

Subject Areas:
biophysics, computational biology, biochemistry

Keywords:
FAIR, open science, open data, reproducibility, data sharing

Author for correspondence:
Enrico Riccardi
e-mail: enrico.riccardi@ntnu.no

THE ROYAL SOCIETY PUBLISHING

Envisioning data sharing for the biocomputing community

Enrico Riccardi[1], Sergio Pantano[2] and Raffaello Potestio[3,4]

[1]Department of Chemistry, Norwegian University of Science and Technology, Høgskoleringen 5, 7491 Trondheim, Norway
[2]Institut Pasteur de Montevideo, Mataojo 2020, CP 11400 Montevideo, Uruguay
[3]Department of Physics, University of Trento, via Sommarive 14, 38123 Trento, Italy
[4]INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, 38123 Trento, Italy

ER, 0000-0003-1890-7113; SP, 0000-0001-6435-4543; RP, 0000-0001-6408-9380

The scientific community is facing a revolution in several aspects of its *modus operandi*, ranging from the way science is done—data production, collection, analysis—to the way it is communicated and made available to the public, be that an academic audience or a general one. These changes have been largely determined by two key players: the *big data* revolution or, less triumphantly, the impressive increase in computational power and data storage capacity; and the accelerating paradigm switch in science publication, with people and policies increasingly pushing towards open access frameworks. All these factors prompt the undertaking of initiatives oriented to maximize the effectiveness of the computational efforts carried out worldwide. Taking the moves from these observations, we here propose a coordinated initiative, focusing on the computational biophysics and biochemistry community but general and flexible in its defining characteristics, which aims at addressing the growing necessity of collecting, rationalizing, sharing and exploiting the data produced in this scientific environment.

1. Introduction

The power of computational methods in the study of living systems has immensely grown in the past few decades. The size of the systems that can be studied by means of computer-based approaches has boosted from a few to millions of atoms [1–6], while the accuracy of force fields has systematically improved thanks to *ab initio* calculations and the integration of experimental data, making these *in silico* models predictive. A prominent role has been and is being played by coarse-grained models [7–10], that is, simplified representations of complex biological systems whose level of resolution is lower than atomistic (crossing all scales up to the continuum) but enable the simulation of larger objects for longer times. Small yet powerful computer stations, GPU cards, small-sized computer clusters and ever-improving algorithms on top of everything have made cutting-edge research in computational biophysics amenable to groups at all scales, from the single person to the hundred-unit teams. As of now, 90% of the data that existed in the beginning of 2018 were created in the last 2 years [11].

All these advancements have tremendously contributed to push forward our understanding of the numerous, intricate, intermingled and multi-scale processes and phenomena that can be encircled in the broad definition of *life*. From the water self-protonization reaction spontaneously turning water into its charged constituents [12] to the flux of red blood cells in the vascular stream [13], computer models are now numerous enough and sufficiently accurate, in spite of the obvious limitations and shortcomings deriving from the approximations they entail, so as to permit a remarkably deeper insight into the functioning of biological entities. Above all, these models are becoming increasingly predictive.

The cosmic inflation of computational biophysics naturally has its inevitable downsides. We identify here the most prominent in the following issues:

— The distribution of immense treasures of data (molecular dynamics trajectories to name just the most obvious) which would otherwise remain confined in the laboratories that produced them.
— The storage of data in a compact, reliable, secure protocol, to contrast the data loss due to hardware failures and outdated software.
— The development of common procedures to rationalize data storage, avoiding data overflow and limiting the costs of backups.
— The limitation of many redundant research efforts, due to the necessity of a given group to re-do the work of another just for the sake of obtaining those data which are needed as 'input' for further investigation rather than representing an objective themselves.
— The reduction of the plurality of standards for input and output files, metadata, algorithms, etc., which often degenerate in a detrimental incompatibility between data and codes that would be otherwise sensible and useful to put in contact. This phenomenon creates quasi-closed communities of single-software users and prevents researchers from creating simple and effective pipelines or assemblies of algorithms to obtain new results.
— The mitigation of research opacity, a consequence of the difficulty to access the raw data, metadata, input files and detailed documentation of procedures and algorithms of many works, as well as to reproduce their results.

This last aspect is particularly worrisome, as it does not consist of a limitation of current approaches or a gap with respect to an otherwise achievable optimum; rather, it bears the risk of distorted, misguided and even intentionally falsified scientific behaviour. Indeed, alarming reports indicated that 'bad science' is becoming a prominent problem due to the current publishing format [14], in that the latter incentivizes publication for its own sake, e.g. by bringing ground-breaking claims to prominence while scarcely selecting against false positives and poorly designed experiments and analyses. The phenomenon is particularly grave if it is not unintentional, rather it originates from deliberate cheating or loafing. Normalized misbehaviour [15] is generating an increasing deviance from ethical norms; furthermore, science is endangered by statistical misunderstanding [16].

One of the worst consequences of dubious or opaque work is that it undermines the credibility and integrity of research as a whole [17], a phenomenon with dramatic societal consequences made even more critical by *intentional* scientific misconduct, for the detection of which a framework is currently missing [18–20]. An intuitive yet simplistic measure of the dimension of scientific misconduct is given by the number of retracted papers, yet it is at best conservative [21–25], thus making the quantification of the cost of this phenomenon, for the academic community as well as for society at large, extremely difficult, if not impossible [26].

Consequently, there is a growing urge for improving meta-research, that is, for instruments to evaluate and ameliorate research methods and practices. As is natural for a young and growing field, efforts to date are uncoordinated and fragmented [27]. Even with the employment of 'good practices', a vast quantity of research output is not fully exploited or even goes wasted [28], with the *file drawer problem* [29]—that is, to refrain from publishing evidence contrary to an author's hypotheses—being one of the most representative issues. The lack of protocols, standards and procedures to grant wide data accessibility leads to substantial costs (in terms of personnel time, facility resources, research funds) [30].

A lively discussion between researchers and policy makers is ongoing at several intranational and international levels in order to reverse this trend. Among the most recent initiatives undertaken at the European level, it is worth mentioning the creation of the European Open Science Cloud (https://www.eosc-portal.eu), whose objective is *to provide a safe environment for researchers to store, analyse and re-use data for research, innovation and educational purposes*.

Performing a very restrictive selection from a heterogeneous literature, we here consider and take the moves from two proposals in particular: *Reformulating Science* (methodological, cultural, structural reforms) [31,32] and *Science Utopia* [33,34]. With the suggestions therein in mind, and limited to the research fields of computational biophysics and chemistry, we consider a strategical priority to coordinate at a supranational level the availability of scientific data and software in order to increase research efficiency, reproducibility and openness. This is certainly not a novel idea and several successful examples of curated databases integrating biological information can be found in life sciences, such as the Genebank [35], UniProt [36] and the Worldwide Protein Data Bank (PDB) [37]. An initiative like the one proposed here is the NoMaD project [38], which maintains one of the largest open repositories for computational material science. We believe that a global effort has to be undertaken in order to rationalize the complex ecosystem of software and the goldmine of information that is emerging from the collective, albeit often independent, work of a steadily growing research community. Moreover, the availability of the data would also contribute to boost the scientific progress in developing countries, where the scarcity of resources impairs the training of highly specialized researchers.

2. Envisioning data sharing in biocomputing

The proposal we put forward in this paper is the creation of a platform devoted to putting in fruitful contact three instances: the data, the users and the institutions.

The first is the central objective of our attention: these are the valuable outcome of intense research activity, whose lifespan cannot be limited to the time interval from their production to the publication of a paper. Indeed, it is our opinion that scientific data should be made freely available not only to guarantee transparency, rather also to maximize the gain that can be obtained from their analysis, with the far-reaching goals of reducing useless replication and pushing forward our understanding of living matter. The second instance, the users, are the producers of data as well as their beneficiaries; however, these two roles should not necessarily coincide. The possibility of accessing other researchers' data would boost the effectiveness of one's work and expand the scope of the former's, while at the same time encouraging the application of good scientific practices such as the accurate documentation of codes and output data. Thirdly, the explicit and collective involvement

of institutions would lift from the researcher the burden of storage and maintenance, provide an independent quality control and foster collaboration and cross-fertilization.

In order to achieve these goals, we propose to create a framework to implement and promote the accessibility of data and software pertaining to the computer-based study of biological systems.

2.1. Data

The core idea of this initiative is to make computational biophysics data publicly available, searchable and downloadable, adhering to the FAIR principles [39]. Researchers who have produced data such as large macromolecular models, molecular dynamics trajectories or any other similar type of relevant material shall make these publicly available, together with a detailed documentation and metadata. While it is already possible to associate a digital object identifier, or DOI, to data [40], via a framework inspired by the DataCite Metadata Scheme [41], a protocol should be shaped according to the specific needs of the target scientific community.

The most natural infrastructure to this end is a website where links to the various data made available by researchers and institutions will be listed, and all information regarding the material present in the list at a given moment will be searchable. In this first phase, those willing to provide a link to their data will be allowed to do so only for material employed in published research. Such a limitation will represent a rough filter to guarantee a quality minimum for what is listed on the website. Clearly, it is difficult to even roughly estimate the amount of data storage required for this initiative to be successful. However, in the early phases of the project, contributors could only upload a one-page document like a PDB file header specifying: title, authors, affiliation, e-mail contact and content, along with a link from where compressed data can be downloaded. Authors/institutions (universities, laboratories, research centres) shall guarantee that these data are available, properly stored and backed-up and provide links from which they can be retrieved. In subsequent stages of the initiative, international institutions will be asked to contribute to the website by storing and making available the data produced by their researchers. This will guarantee the permanent availability of the material. A not negligible issue regards the format of the data, as many formats associated with specific software exist. Although it would be desirable to have one single standard, this goal is not realistic in the early stages of the initiative. This is considered a relatively minor limitation as several open source programs and even servers can efficiently convert coordinates and trajectories files; however, standard uniformation should be looked at as a prominent long-term goal.

2.2. Users

To increase the chances of success of the initiative, it is strategic that potential users perceive our proposal as a simple, accessible and clearly advantageous practice. Open data is and shall appear as a good practice that will lead to research environment and research effectiveness improvement. All possible strategies to promote the applications of the FAIR principles, from dedicated DOIs to grants opportunities, shall be implemented. Obviously, all researchers, regardless of their institution, country, gender, belief or nationality will have the opportunity to freely contribute and download data to the website. Eventually, if the amount of data offered is exceedingly large to be kept accessible on a permanent basis, a verified contact should be provided to ask privately the contributors.

2.3. Institutions

Once the initiative will have achieved a sufficient size, we envision that academic institutions could contribute to the website by storing and making available the data produced by their researchers and could be asked to do the same for the data of other contributing institutions. This should promote the usage of large amounts of data to carry out curiosity-driven, basic research, resulting in the development of new computational tools to process the available information. The definition and usage of standardized protocols will greatly enhance the capability of validating and reproducing research results, thus minimizing the risk of inadvertently or intentionally erroneous scientific claims. The institutional involvement is valuable because it can provide rewarding schemes that can further foster the adoption of open science in general and of our unified approach in particular. Incentives to guide the researcher towards best practices shall be favoured with respect to top-down impositions.

2.4. An exemplary initiative: NoMaD

We consider the NoMaD project [42] to be the most notable and relevant initiative with respect to the proposal formulated in this opinion. NoMaD was launched in 2018 thanks to the support of the European Union in the framework of the Centres of Excellence, and provides a FAIR data-driven platform with a focus on material science. The NoMaD initiative represents not only an example of clarity, efficiency and openness in data storage and conservation, rather also a potential partner in the construction of a similar platform for computational biophysics.

3. Conclusion

In the present opinion article, we discussed a critical aspect of research communication, consisting of a lack of reproducibility and transferability of research results between groups, from the perspective of computational biophysics and biochemistry. Consistently with the open science movement and considering the impressive increase in computational power and data storage capacity, we proposed an initiative that would facilitate, within computational biophysics and chemistry, (i) the adoption of open science approaches, (ii) the rationalization of data storage, (iii) the increase of research efficiency through the reduction of replication, (iv) the increase of data validation and reproducibility. Clearly, a number of issues remain open. Just to name a few, the precise format of the data stored and its management, intellectual property issues, funding possibilities to ensure the continuity of this initiative, among others, should be addressed in due time.

Within this framework, both users (i.e. researchers) and institutions will benefit from increased reliability of research outputs and output reproducibility.

Data accessibility. This article has no additional data.

Competing interests. We declare we have no competing interests.

Funding. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 758588).

Acknowledgements. We thank Sarah Harris, Ali Hassanali, Radek Erban and Ana Slavec for a critical reading of the manuscript and insightful discussion. The authors gratefully acknowledge the support provided by the Royal Society in the funding and organization of the Theo Murphy international scientific meeting *Multi-resolution simulations of intracellular processes* which took place on 24–25 September 2018 at Chicheley Hall, Buckinghamshire (UK), during which inspiring and fruitful conversations took place.

References

1. Alder BJ, Wainwright TE. 1959 Studies in molecular dynamics. I. General method. *J. Chem. Phys.* **31**, 459–466. (doi:10.1063/1.1730376)
2. Karplus M. 2002 Molecular dynamics simulations of biomolecules. *Acc. Chem. Res.* **35**, 321–323. (doi:10.1021/ar020082r)
3. Anderson JA, Lorenz CD, Travesset A. 2008 General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* **227**, 5342–5359. (doi:10.1016/j.jcp.2008.01.047)
4. Stone JE, Phillips JC, Freddolino PL, Hardy DJ, Trabuco LG, Schulten K. 2007 Accelerating molecular modeling applications with graphics processors. *J. Comput. Chem.* **28**, 2618–2640. (doi:10.1002/(ISSN)1096-987X)
5. Fan Z, Chen W, Vierimaa V, Harju A. 2017 Efficient molecular dynamics simulations with many-body potentials on graphics processing units. *Comput. Phys. Commun.* **218**, 10–16. (doi:10.1016/j.cpc.2017.05.003)
6. Shaw DE *et al.* 2009 Millisecond-scale molecular dynamics simulations on anton. In *Proc. Conf. High Performance Computing Networking, Storage and Analysis*, SC '09, pp. 39:1–39:11. New York, NY: ACM.
7. Takada S. 2012 Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol.* **22**, 130–137. (doi:10.1016/j.sbi.2012.01.010)
8. Noid WG. 2013 *Systematic methods for structurally consistent coarse-grained models*, pp. 487–531. Totowa, NJ: Humana Press.
9. Saunders MG, Voth GA. 2013 Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **42**, 73–93. (doi:10.1146/annurev-biophys-083012-130348)
10. Potestio R, Peter C, Kremer K. 2014 Computer simulations of soft matter: linking the scales. *Entropy* **16**, 4199–4245. (doi:10.3390/e16084199)
11. insideBIGDATA. 2018 The 2018 state of data management. See https://insidebigdata.com/2018/09/21/2018-state-data-management/.
12. Moqadam M, Lervik A, Riccardi E, Venkatraman V, Alsberg BK, van Erp TS. 2018 Local initiation conditions for water autoionization. *Proc. Natl Acad. Sci. USA* **115**, E4569–E4576. (doi:10.1073/pnas.1714070115)
13. Gutierrez M, Fish MB, Golinski AW, Eniola-Adefeso O. 2018 Presence of rigid red blood cells in blood flow interferes with the vascular wall adhesion of leukocytes. *Langmuir* **34**, 2363–2372. (doi:10.1021/acs.langmuir.7b03890)
14. Smaldino PE, McElreath R. 2016 The natural selection of bad science. *R. Soc. open sci.* **3**, 160384. (doi:10.1098/rsos.160384)
15. De Vries R, Anderson MS, Martinson BC. 2006 Normal misbehavior: scientists talk about the ethics of research. *J. Empir. Res. Hum. Res. Ethics* **1**, 43–50. (doi:10.1525/jer.2006.1.1.43)
16. Colquhoun D. 2017 The reproducibility of research and the misinterpretation of p-values. *R. Soc. open sci.* **4**, 171085. (doi:10.1098/rsos.171085)
17. John LK, Loewenstein G, Prelec D. 2012 Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23**, 524–532. (doi:10.1177/0956797611430953)
18. Martinson BC, Anderson MS, De Vries R. 2005 Scientists behaving badly. *Nature* **435**, 737–738. (doi:10.1038/435737a)
19. Fanelli D. 2009 How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE* **4**, e5738. (doi:10.1371/journal.pone.0005738)
20. Jha A. 2012 False positives: fraud and misconduct are threatening scientific research. *The Guardian*, 13 September 2012.
21. Wang J, Ku JC, Alotaibi NM, Rutka JT. 2017 Retraction of neurosurgical publications: a systematic review. *World Neurosurg.* **103**, 809–814.e1. (doi:10.1016/j.wneu.2017.04.014)
22. Al-Ghareeb A, Hillel S, McKenna L, Cleary M, Visentin D, Jones M, Bressington D, Gray R. 2018 Retraction of publications in nursing and midwifery research: a systematic review. *Int. J. Nurs. Stud.* **81**, 8–13. (doi:10.1016/j.ijnurstu.2018.01.013)
23. Yan J, MacDonald A, Baisi L-P, Evaniew N, Bhandari M, Ghert M. 2016 Retractions in orthopaedic research. *Bone Joint Res.* **5**, 263–268. (doi:10.1302/2046-3758.56.BJR-2016-0047)
24. Ribeiro MD, Vasconcelos SMR. 2018 Retractions covered by retraction watch in the 2013–2015 period: prevalence for the most productive countries. *Scientometrics* **114**, 719–734. (doi:10.1007/s11192-017-2621-6)
25. Van Noorden R. 2011 The trouble with retractions. *Nature* **478**, 26–28. (doi:10.1038/478026a)
26. Fang FC, Casadevall A. 2011 Retracted science and the retraction index. *Infect. Immun.* **79**, 3855–3859. (doi:10.1128/IAI.05661-11)
27. Ioannidis JP, Fanelli D, Dunne DD, Goodman SN. 2015 Meta-research: evaluation and improvement of research methods and practices. *PLoS Biol.* **13**, e1002264. (doi:10.1371/journal.pbio.1002264)
28. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. 2014 Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* **383**, 267–276. (doi:10.1016/S0140-6736(13)62228-X)
29. Rosenthal R. 1979 The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638–641. (doi:10.1037/0033-2909.86.3.638)
30. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R. 2014 Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383**, 166–175. (doi:10.1016/S0140-6736(13)62227-8)
31. Casadevall A, Fang FC. 2011 Reforming science I. Methodological and cultural reforms. *Infect. Immun.* **80**, 891–896. (doi:10.1128/IAI.06183-11)
32. Fang FC, Casadevall A. 2011 Reforming science II. Structural reforms. *Infect. Immun.* **80**, 897–901. (doi:10.1128/IAI.06184-11)
33. Nosek BA, Bar-Anan Y. 2012 Scientific utopia: I. Opening scientific communication. *Psychol. Inq.* **23**, 217–243. (doi:10.1080/1047840X.2012.692215)
34. Nosek BA, Spies JR, Motyl M. 2012 Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* **7**, 615–631. (doi:10.1177/1745691612459058)
35. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012 Genbank. *Nucleic Acids Res.* **41**, D36–D42. (doi:10.1093/nar/gks1195)
36. Consortium U. 2014 Uniprot: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212. (doi:10.1093/nar/gku989)
37. Berman H, Henrick K, Nakamura H, Markley JL. 2006 The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**, (D301–D303. (doi:10.1093/nar/gkl971)
38. Abramson MA, Audet C, Couture G, Dennis Jr JE, Le Digabel S, Tribes C. 2011 The NoMaD project.
39. Wilkinson MD *et al.* 2016 The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018. (doi:10.1038/sdata.2016.18)
40. Brase J. 2009 Datacite—a global registration agency for research data. In *4th Int. Conf. on Cooperation and Promotion of Information Resources in Science and Technology, Beijing, China, 21–23 November 2009*, pp. 257–261. (doi:10.1109/COINFO.2009.66)
41. Starr J, Gastl A. 2011 Iscitedby: a metadata scheme for DataCite. *D-Lib Mag.* **17**. (doi:10.1045/january2011-starr)
42. Draxl C, Scheffler M. 2018 Nomad: the fair concept for big data-driven materials science. *MRS Bull.* **43**, 676–682. (doi:10.1557/mrs.2018.208)