MethodsX 9 (2022) 101607

Contents lists available at ScienceDirect

MethodsX

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

Sharing data, sharing methods, sharing science.

Sergio Pantano

Institut Pasteur de Montevideo, Montevideo, Uruguay

ARTICLE INFO
Method name:
Article history: Available online 14 December 2021

Introduction

Research and discovery in the most disparate areas of knowledge are constantly being pushed by continuous technological revolutions. The permanent waves of advances changed how scientific work is planned, produced, and communicated within the scientific community and made available to the general public. The "big-data revolution" impacted practically all scientific disciplines, increasingly imposing data-driven methodologies [1]. A significant outcome of this process is a shift towards open data initiatives (Budapest OS initiative, Plan S, EOSC, etc.). These initiatives have contributed to greater availability and accessibility of publicly funded scientific research, accompanied by the creation of freely available data repositories for large volumes of information.

Soon after these initiatives started, the necessity to improve data organization, and storage efficiency and effectiveness became evident. As a result, FAIR (findable, accessible, interoperable, reusable) guiding principles for scientific data management and stewardship were outlined [2,3].

The large-scale data accessibility powered a plethora of new analytical tools such as those commonly bundled under the generic denomination of artificial intelligence [4–8].

Obviously, the capacity to sort, analyze and integrate this overwhelming amount of information is extremely challenging and demands particular computational skills. Indeed, a minimal computational background is no longer an asset in science; it has become a must. However, achieving simultaneously a comprehensive knowledge of the fundamentals of a particular scientific discipline *and* highly specific computational skills is unlikely to become a standard. Modern science is carried out by interdisciplinary and complementary working teams that, at their time, thrive from knowledge achieved by larger scientific communities. These organizational dynamics require a second level of openness, the sharing of methods and computational analysis tools. Sharing software for data analysis saves the precious time needed for coding, testing, debugging, and documenting, obviating the need for highly specialized computational skills.

Just like the raw data, the nature of the software needed to process it may be of different nature and produced in many formats, languages, and operative systems. It may sometimes be a single software package or pipelines (namely, an array of processing elements organized so that the output of the previous step becomes the input of the next). Clearly, the interoperability of the software

https://doi.org/10.1016/j.mex.2021.101607

and well-documented instructions for compiling and running the software in suitable hardware architectures are mandatory in all these cases.

MethodsX aims to boost science openness at various levels by improving reproducibility, making methods, protocols, and the associated research more discoverable, opening doors for collaboration, and enhancing the research cycle. In line with these aims, MethodsX recently published a number of software packages and pipelines for data analysis to aid research at different levels and in various disciplines. The following paragraphs will briefly overview some specific examples and how they may help within and beyond the scientific community.

Software utilities for molecular simulations

The increasing computer power and consistently faster algorithms have taken molecular simulations to the level of computational microscopes that enable the study of molecular systems at an unprecedented level [9]. However, applications in material sciences and biology are still critically dependent on stating conditions, e.g., the initial 3D coordinates of the atoms that compose the molecular systems. Recently a couple of exciting applications published in our journal addressed this issue. The software *Nanosculpt* [10] provides a simple method to generate atomic coordinates of complex objects incorporating topological information taken directly from experimental data (as Cryoelectron microscopy/tomography) or user-created arbitrary shapes. In another publication, Gupta et al. [11] analyzed current problems associated with the initial thermalization of the molecular systems and provide an optimized procedure in which nanocrystalline materials are thermalized by coupling specific regions of the simulation box to separated thermostats at different target temperatures.

Although tested only for nanomaterials, both methods also harbor great potential for biological applications.

Bioinformatics and molecular biology

The wide availability of different sequencing techniques made advanced bioinformatic methods a pressing necessity to get the most out of the data.

The reconstruction of gene phylogenies is based on large sequence alignments. However, homologous sequences are often difficult to align unambiguously, or alignments may contain insufficient information to accurately model gene evolution, leading to incorrect gene trees and erroneous predictions of events of duplications and losses [12]. A workaround for this problem is the construction of longer "supergenes" that comprise sets of loci with putatively similar genealogical stories. However, validating the concatenation of several genes in one single supertree remains a difficult task. Adams and Castoe recently published a model-based protocol for assessing the accuracy of supergenes construction based on phylogenetic congruency that may validate supergene hypotheses [13].

The combination of chromatin immunoprecipitation (ChIP) with sequencing (ChIP-Seq) constitutes a powerful method for identifying genome-wide DNA binding sites. The technique is conceptually simple: DNA-bound proteins are co-immunoprecipitated, purified, and sequenced. However, identifying contiguous specific sequences that act as super-enhancers by binding different transcription factors is computationally highly demanding. Orlova et al. devised a cost-efficient strategy based on distributed computing in virtual machine cloud environments [14]. This method is particularly well suited for research centers with modest access to computational resources and, in principle, independent of the operative system used, making it robust and interoperable.

Another example of cost-efficient bioinformatic methodology but applied to pharmacology is provided by the work of Zidan et al. [15]. They introduced a computational pipeline for pharmacovigilance/pharmacogenomics named PHARMIP. It combines chemical structure and database-reported information about specific drug candidates to anticipate the genetic factors underlying the drug-reported adverse reactions. The implementation is available with a user-friendly interface that facilitates the use by non-experts and can provide valuable hints about genetic risk factors or propensities for specific drug candidates.

Imaging

The use of fluorescent probes for imaging has revolutionized our understanding of countless biological processes, as it provides non-invasive means to interrogate living organisms in real-time. Nevertheless, experiments may generate a considerable amount of data that need to be processed in different manners. A nice pipeline that offers a semiautomated treatment of time-lapse fluorescence microscopy images, quantification of individual cell signals, and a statistical analysis of the data was recently provided by Lévy et al. [16]. The pipeline combines routines in the popular platforms FIJI, R, and MATLAB packages, automatizing the analysis of a large number of samples and increasing its statistical robustness.

Fluorescent microscopy also grants the opportunity to follow single-molecule localization, unraveling the subcellular organization of different molecular species. Hoboth et al. presented a Single Molecule Localization Microscopy (SMLM) approach to follow the nuclear localization of phosphatidylinositol 4,5- bisphosphate using indirect immunofluorescence labeling [17]. They developed a tool within the free software ImageJ2, which is orthogonal but highly complementary to the more traditional biochemical and lipidomic analyses.

Horzum et al. devised another ImageJ application to process immunofluorescence images of vinculin, a well-characterized marker of cellular focal adhesion [18]. Their "Step-by-step quantitative analysis of focal adhesions" provided a practical approach to quantify a variety of fluorescent images improving the signal-to-noise relationship in systems with high background.

Finally, a pipeline incorporating a macro in the Fiji environment and R scripts to measure the intensity of intracellular staining was reported by Zonderland et al. [19]. This pipeline allows for simple measuring of staining intensities straightforwardly and independently of the cellular shapes or sizes.

Miscellaneous applications

The packages, pipelines, and methods outlined in the previous paragraphs constitute examples related to some specific areas of research. However, just to illustrate the wide range covered by our journal, we mention a few examples of software with diverse areas of application. The work by Cornish et al. [20] presents a script in phyton language to interrogate the PubMed database for different variations and permutations of eponyms and names given to genes, proteins, and chemical compounds. This simple method enables a fast and unambiguous search, sorting, and characterizing scientific citations by all PubMed's data fields.

A completely different example is provided by the contribution by Arenas-Castro and Gonçalves [21]. They offered a model-assisted method to forecast the suitability of crop production in different terrains as a function of climate change predictions. This machine learning approach written in the R package is freely available on GitHub and may provide helpful information for agricultural sciences and decision-makers about long-term farming politics.

The last example regards a machine learning approach for accurately estimating the mortality produced by pandemic or epidemic events. This is done by calculating the excess mortality from retrospective data using linear regression approaches [22]. Although applied to estimate mortality caused by Covid-19 in Italy in 2020, this approach can be generally applied to obtain valuable epidemiologic insights, keeping updated information about the progress and effect of successive waves of contagions and the effectiveness of containment or vaccination measures.

Conclusions and outlook

As evident from the articles bundled in this editorial piece, methods-sharing initiatives from scientists for scientists are fervent and extended to different areas. However, significant challenges remain. They are associated with the software's interoperability and robustness to different environments and operative systems, the standardization of data formats used as input and generated as output of the analyzes. Using "software containers" (e.g., software packages that provide applications, dependencies, system libraries, settings, and other binaries, and all the configuration files

needed to run) constitutes an effective workaround for the first problem. However, standardization remains a central problem in many areas and still needs to be addressed as a community effort.

Sharing analysis software and computational protocols may create positive feedback loops to progressively overcome these limitations and make high-quality methods more broadly accessible.

E-mail address: spantano@pasteur.edu.uy

References

[1] S. Klein, The World of Big Data and IoT, IoT Solutions in Microsoft's Azure IoT Suite, Apress, Berkeley, CA, 2017, doi:10.1007/978-1-4842-2143-3_1.
[2] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016), doi:10.1038/sdata.2016.18.
[3] J. Wise, A.G. de Barron, A. Splendiani, B. Balali-Mood, D. Vasant, E. Little, G. Mellino, I. Harrow, I. Smith, J. Taubert, K. van Bochove, M. Romacker, P. Walgemoed, R.C. Jimenez, R. Winnenburg, T. Plasterer, V. Gupta, V. Hedley, Implementation and relevance of FAIR data principles in biopharmaceutical R&D, Drug Discovery Today 24 (2019), doi:10.1016/j.drudis.2019.01.008.
[4] Y. Duan, J.S. Edwards, Y.K. Dwivedi, Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda, International Journal of Information Management 48 (2019), doi:10.1016/j.ijinfomgt.2019.01.021.
[5] D.E. O'Leary, Artificial Intelligence and Big Data, IEEE Intelligent Systems 28 (2013), doi:10.1109/MIS.2013.39.
[6] K. Kersting, U. Meyer, From Big Data to Big Artificial Intelligence? KI - Künstliche Intelligenz 32 (2018), doi:10.1007/s13218-017-0523-7.
[7] A.L. Oliveira, Biotechnology, Big Data and Artificial Intelligence, Biotechnology Journal 14 (2019), doi:10.1002/biot.201800613.
[8] A.M. Rahmani, E. Azhir, S. Ali, M. Mohammadi, O.H. Ahmed, M. Yassin Ghafour, S. Hasan Ahmed, M. Hosseinzadeh, Artificial intelligence approaches and mechanisms for big data analytics: a systematic study, PeerJ Computer Science 7 (2021), doi:10.7717/peerj-cs.488.
[9] E.H. Lee, J. Hsin, M. Sotomayor, G. Comellas, K. Schulten, Discovery Through the Computational Microscope, Structure. 17 (2009), doi:10.1016/j.str.2009.09.001.
[10] A. Prakash, M. Hummel, S. Schmauder, E. Bitzek, Nano: A methodology for generating complex realistic configurations for atomistic simulations, MethodsX. 3 (2016), doi:10.1016/j.mex.2016.03.002.
[11] A. Gupta, S.S. Rajaram, G.B. Thompson, G.J. Tucker, Improved computational method to generate properly equilibrated atomistic microstructures, MethodsX 8 (2021), doi:10.1016/j.mex.2021.101217.
[12] B. Boussau, G.J. Szollosi, L. Duret, M. Gouy, E. Tannier, V. Daubin, Genome-scale coestimation of species and gene trees, Genome Research 23 (2013), doi:10.1101/gr.141978.112.
[13] R.H. Adams, T.A. Castoe, Supergene validation: A model-based protocol for assessing the accuracy of non-model-based supergene methods, MethodsX 6 (2019), doi:10.1016/j.mex.2019.09.025.
[14] N.N. Orlova, O.v. Bogatova, A.v. Orlov, High-performance method for identification of super enhancers from ChIP-Seq data with configurable cloud virtual machines, MethodsX 7 (2020), doi:10.1016/j.mex.2020.101165.
[15] A.M. Zidan, E.A. Saad, N.E. Ibrahim, A. Mahmoud, M.H. Hashem, A.A. Hemeida, PHARMIP: An insilico method to predict genetics that underpin adverse drug reactions, MethodsX 7 (2020), doi:10.1016/j.mex.2019.100775.
[16] E. Lévy, F. Jaffrézic, D. Laloë, H. Rezaei, M.-E. Huang, V. Béringue, D. Martin, L. Vernis, PiQSARS: A pipeline for quantitative and statistical analyses of ratiometric fluorescent biosensors, MethodsX 7 (2020), doi:10.1016/j.mex.2020.101034.
[17] P. Hoboth, O. Šebesta, M. Sztacho, E. Castano, P. Hozák, Dual-color dSTORM imaging and ThunderSTORM image reconstruction and analysis to study the spatial organization of the nuclear phosphatidylinositol phosphates, MethodsX 8 (2021), doi:10.1016/j.mex.2021.101372.
[18] U. Horzum, B. Ozdil, D. Pesen-Okvur, Step-by-step quantitative analysis of focal adhesions, MethodsX 1 (2014), doi:10.1016/j.mex.2014.06.004.
[19] J. Zonderland, P. Wieringa, L. Moroni, A quantitative method to analyse F-actin distribution in cells, MethodsX 6 (2019), doi:10.1016/j.mex.2019.10.018.
[20] T.C. Cornish, L.J. Kricka, J.Y. Park, A Biopython-based method for comprehensively searching for eponyms in Pubmed, MethodsX 8 (2021), doi:10.1016/j.mex.2021.101264.
[21] S. Arenas-Castro, J. Gonçalves, SDM-CropProj – A model-assisted framework to forecast crop environmental suitability and fruit production, MethodsX 8 (2021), doi:10.1016/j.mex.2021.101394.
[22] D. Gibertoni, F. Sanmarchi, K.Y.C. Adja, D. Golinelli, C. Reno, L. Regazzi, J. Lenzi, Small-scale spatial distribution of COVID-19-related excess mortality, MethodsX 8 (2021), doi:10.1016/j.mex.2021.101257.