ARPHA Preprints

PENSOFT

Grant Proposal

Author-formatted document posted on 05/12/2023

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e116671

Tracing bio-structures with serial crystallography: Facilitating the access to high-throughput macromolecular x-ray crystallography techniques.

Miguel Angel Gonzalez, Rudolf Dimper, Patrick Fuhrmann, Gianluca Santoni, Jordi Bodera, Jayesh Wagh, Irina Safiulina, Arianna D'Angelo, Paolo Mutti, Paul Millar, Krisztian Pozsa, Leonardo Sala, Alun Ashton, Giuseppe La Rocca

ARPHA Preprints *Author-formatted document posted on 05/12/2023*. DOI: https://doi.org/10.3897/arphapreprints.e116671

Tracing bio-structures with serial crystallography: Facilitating the access to high-throughput macromolecular x-ray crystallography techniques.

Gonzalez, M., Dimper, R, Fuhrmann, P., Santoni, G., Bodera, J., Wagh, J., Safiulina, I., Mutti, P., D'Angelo, A., Millar, P., Pozsa, K., Sala, L., Ashton, A., La Rocca , G.

Abstract

Serial (femtosecond) X-ray-Crystallography (SFX) is a special variant of macromolecular X-ray crystallography aiming at rapid structural studies at room temperature. This highly innovative technology permits investigation of bio-structures not tractable with conventional X-ray crystallography, and is capable of studying fast in-situ biochemical processes. The method is still relatively new, but it is already one of the most prominent applications of free-electron lasers (FELs), and increasingly also of very brilliant synchrotron radiation sources. One of the unique characteristics of this type of experiments is the extremely high repetition rate combined with a quite moderate success rate. A crucial task in the rather complex data processing pipeline is the rapid and accurate classification of images: typically, only a few percent of the images contain a diffraction pattern suitable for subsequent integration and structure refinement. AI-supported image classification is hence particularly suited for drastic data reduction, saving precious storage space, compute cycles and processing time. The experimental techniques and methodologies are rapidly evolving, and the integration of emerging tools into the processing pipeline is an essential task. SFX data sets are big, require substantial storage, and computational power. The main goal of this SP is to establish and develop a data processing platform, which integrates services and developments from PaNOSC/ExPaNDS. The platform should provide integrated processing pipelines for well-established and cutting-edge applications, so that cross-disciplinary users with modest expertise gain rapid and convenient access to tools and documentation of newest developments. On the other hand, it should also provide convenient access to FAIR SFX-data, to foster developments and strengthen collaboration between experimentalists and developers of new algorithms and software implementations. This approach is of very high relevance for all PaN synchrotron and FEL facilities and their users.

Keywords

EOSC Future; Science Clusters; Science Projects; PaNOSC; Macromolecular Serial Crystallography

Description

Existing situation

The Photon and Neutron (PaN) user facilities serve a large variety of scientific user communities using x-rays and/or neutrons in their research projects in order to understand the structure and function of matter. The most important goal is hence the advancement of science through measures including accelerator and detector developments, evolution of experiment capabilities, or algorithmic and application development.

A number of projects like CALIPSOplus (Calipsoplus 2021), ExPaNDS (Expands 2023) or PaNOSC (Panosc 2022) already serve this approach for a small number of use cases in a complementary manner. The activities of the PaN user facilities (DESY 2023, ESRF 2023, ILL 2023, PSI 2023) in EOSC Future (EOSC 2023) will hence focus on key aspects in line with the initial call: (i) integration; (ii) consolidation; (iii) boosting data intense research; (iv) enabling researchers access to scalable environment; (v) widening the EOSC user base; and (vi) supporting cross-analysis of data from heterogeneous sources.

In the course of the COVID-19 crisis, the PaN User facilities have impressively demonstrated their ability to react quickly to new scientific and societal challenges. Many of the facilities initiated fast-track access for COVID-19 research, enabling a plethora of structural investigations, resulting in hundreds of thousands of datasets and numerous high-level publications. bioRxiv and medRxiv alone count more than 4300 publications on the topic. Despite the tremendous success, the approach also revealed a few deficits, which could have been avoided if the integration into EOSC had been more advanced. The publication of the raw underlying scientific data does not keep up with the publication of the results, which might need to be taken into consideration by facilities and projects implementing or updating their scientific data policies. Even if all data had been published, the reproducibility and validation of the scientific outcomes would have been nearly impossible. In many cases, in particular in very innovative, rapidly evolving research fields like serial crystallography, bleeding-edge processing pipelines are rarely commonly available, or require an expertise prohibitive for cross-disciplinary assessment. This Science Project (SP) will hence concentrate on serial macromolecular crystallography (Martin-Garcia and Basu 2020), which can have a strong impact on bio-medical or biochemical characterization of biological materials, with the aim to achieve a high level of reproducibility and automated validation by providing easy to handle cutting-edge processing pipelines integrated into EOSC and existing validation pipelines.

The serial crystallography (SX) approach can be done at synchrotron radiation sources such as ESRF (ESRF 2023) and DESY (DESY 2023). Monochromatic and pink beam experiments have demonstrated the feasibility of serial data collection using micro-crystals at numerous microfocus beamlines. Upcoming developments in beamline optics, detector technology and synchrotron sources by themselves will enable the use of even smaller micro-crystals (<1 µm), the use of larger macromolecules as well as the possibility of conducting mix-and-inject time-resolved studies (Martin-Garcia and Basu 2020). The development of SX has created many opportunities in structural biology, in particular in the field of atomic-resolution imaging of protein dynamics (Spence 2020). However, the novelty of the technique and the computational resources needed to handle the large amount of data generated by these experiments and apply the required powerful new algorithms to analyze them, restrict their usage to a few specialized groups. Therefore, there is a clear need to develop new tools and workflows that facilitate the access to SX data to a larger scientific community.

Objectives

The main objective is to provide an open infrastructure facilitating the access to the structural biology data sets measured by means of new serial (femtosecond) x-ray crystallography experiments, in particular at the beamlines involved in the project ID29 at the ESRF (ID29 2023) and P11 at DESY (P11 2023). One of the unique characteristics of this type of experiments is the extremely high repetition rate combined with a quite moderate success rate. A crucial task in the rather complex data processing pipeline is the rapid and accurate classification of images: typically, only a few percent of the images contain a diffraction pattern suitable for subsequent integration and structure refinement.

This goal requires:

1. Make the data measured on the SX beamlines openly accessible following FAIR principles by on-boarding on EOSC the data catalogs of the partner facilities (ILL Data 2023, ESRF Data 2023, PSI Data 2023).

2. Provide the software, workflows and computational resources needed to analyze efficiently the large data sets generated in this kind of experiments.

The approach foreseen consists in using the resources available at each of the facilities, in order to minimize the data transfer needs. Registered users will have full access to the open data, to the pre-installed data analysis software and to the needed computational means (including multi-CPU and GPU) via a virtual machine, e.g. based on VISA (VISA ESRF 2023). As the experimental techniques and methodologies are rapidly evolving, emerging tools will be continuously integrated into the processing pipeline. In this way the platform should provide integrated processing pipelines both for well-established and cutting-edge applications, so that cross-disciplinary users with modest expertise gain rapid and convenient access to tools and documentation of newest developments. On the other hand, it should also provide convenient access to FAIR SFX-data, to foster developments and strengthen collaboration between experimentalists and developers of new algorithms and software implementations.

Compliance to criteria developed by EOSC Future

This SP is part of the contribution of the PaNOSC cluster to the EOSC platform.

Eligibility

This SP has a cross-disciplinary character and involves three large-scale facilities providing access to EU scientists to a large number of x-ray beamlines. The proposed SP will constitute a practical demonstration of the tools and organization developed during the previous PaNOSC/ExPaNDS projects.

Contribution to EOSC

The SP will demonstrate how complex and highly specialized crystallographic data can be made available to and re-usable by a larger user community. The necessary services (data catalogs, virtual computational resources) will be on boarded into EOSC and provided to the whole community through the EOSC AAI federation.

Quality

The facilities involved are engaged in complying with FAIR principles. High-quality data and tools will be provided, making expert domain knowledge available for multidisciplinary research.

Relevance

SX data can provide unique information about the structure of proteins and other macromolecules, including ligand-protein complex conformations or favorable drug binding sites. As such, the new technical developments can accelerate the discovery of new drugs and health treatments. At present, the complexity of the technique makes it the realm only of specialized groups. This SP aims to open the technique to a much larger and crossdisciplinary community, involving not only physicists, but also chemists, biologist and medical researchers.

Implementation, Plan of work

The two main global jobs associated with the objectives mentioned above are:

1. On board the data catalogs of individual facilities into the EOSC Marketplace, making open data collected in the SX beamlines fully available to the whole community.

2. Create simplified pipelines allowing executing semi-automatically the processing of the large data sets measured in order to produce simple data files containing atomic information about the macromolecule's structures, which can be directly visualized and used for research and educational purposes.

This will need to execute a series of more specific tasks:

Tasks

● HDF5 data viewer available in EOSC (M6)

● Umbrella AAI federated with EOSC AAIs (M12)

● Decision on data transfer solution (PaNOSC/ExPaNDS vs EOSC) (M12)

● Data search API for domain specific searching of PaN data, in collaboration with

OpenAIRE (M12)

- Jupyter notebook prototype, PaN data analysis portal (M15)
- Jupyter notebook and PaN data analysis portal integrated with EOSC AAI, data transfer, and data search (M20)
- PaN learning platform with interactive content (M20)
- Data compression service (M20) (eventually from LEAPS-INNOV)

Use of resources

Initially, mainly internal resources will be exploited, profiting from the infrastructure recently put in place in the facilities to allow remote experiments and data treatment. In addition, internal scientific expertise will be used to define the pipelines to implement. Technical support from EGI and EOSC partners will be needed to on board the local services and integrate any web service that could be developed during the project.

Partners

European Synchrotron Radiation Facility (ESRF), Grenoble, France; Deutsches Elektronen-Synchrotron (DESI), Hamburg, Germany; Paul Scherrer Institut (PSI), Villigen, Switzerland.

Impact

Strategic

This SP will demonstrate how SX data can be made openly available together with the computational resources and tools needed to exploit them. By its own nature, this SP addresses key interdisciplinary scientific objectives of the participating PaN facilities (DESY 2023, ESRF 2023, ILL 2023, PSI 2023), fully exploiting the PaNOSC/ExPaNDS (Expands 2023, Panosc 2022) services in EOSC (EOSC 2023). The SP will demonstrate the innovative impact of the open-science data analysis in the EOSC framework and promote the application of FAIR principles (Wilkinson et al. 2016) for data stewardship. The proposed SP will lower the barrier of cross-disciplinary exploitation and advance innovative developments by providing user-friendly, readily available and cutting-edge data processing pipelines, accelerating the take-up of emerging technologies. This approach embraces FAIR-enabling services, many of which are available in the EOSC marketplace (EOSC-Marketplace 2023), thus increasing researcher's engagement in the EOSC implementation.

Scientific/User communities

The main scientific challenge consists in reducing the level of expertise required to analyze SX data, in order to make cross-disciplinary research truly possible. This will be achieved by providing semi-automatic pipelines and annotated notebooks simplifying the workflow to pass from the raw data to the final macromolecular structure.

From the technical point of view, current challenges include storage and rapid access to large data sets, definition of appropriate interfaces allowing the different software programs needed to communicate between them, and providing the needed computational resources to EOSC users.

Societal/Economic

One of the aims of this SP is to demonstrate the potential impact of SX in the bio-medical or bio-chemical characterization of biological materials, showing that it can accelerate the discovery of new drugs or health treatments.

Open data catalogs are already available and maintained by the partner facilities. Virtual infrastructures to perform remote data analysis have also been set in place during the recent COVID pandemics. They have become an essential tool in the way PaN sources work now, so they will be maintained by each of the individual facilities. Access to these resources through the EOSC marketplace should be ensured within the EOSC collaboration.

EU policies

The SP builds on the EU commitment to foster the advancement of synchrotron research, encouraging collaboration among member states and international partners, and establishing frameworks for access to these cutting-edge facilities by researchers from across Europe. In particular, the EU acknowledges that macromolecular crystallography is essential for understanding biological processes, developing new drugs, and advancing biotechnological applications. EU policies in this area focus on fostering collaboration among member states, providing access to high-quality synchrotron beamlines equipped for macromolecular crystallography, and ensuring training and education opportunities for scientists and researchers. Furthermore, the EU encourages the integration of macromolecular crystallography data into broader research initiatives and data-sharing platforms, enabling scientists across Europe to benefit from the wealth of structural information generated.

Engagement plan

Target groups

Researchers working in the domain of macromolecular crystallography. Scientists working in the fields of drug development, structural biology, biochemistry without previous experience in synchrotron/x-ray crystallography. Pharmaceutical and biotechnology industries.

Dissemination measures

Scientific publications and reports. Conferences, workshops and user meetings.

Acknowledgements

PaNOSC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852. ExPaNDS has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641. The EOSC Future project is co-funded by the European Union Horizon Programme call INFRAEOSC-03-2020 - Grant Agreement Number 101017536.

Funding program

EOSC Future project, WP 6.3, co-funded by the EU Horizon Programme call INFRAEOSC-03-2020 – Grant Agreement Number 101017536.

Conflicts of interest

The authors have declared that no competing interests exist.

References

[Calipsoplus (2021)] CALIPSOplus: Convenient Access to Light Sources Open to Innovation, Science and to the World. https://www.calipsoplus.eu/. Accessed on: 2023-6-27.

[DESY (2023)] Deutsches Elektronen-Synchrotron DESY. https://www.desy.de/index_eng.html. Accessed on: 2023-6-27.

[EOSC (2023)] EOSC Future. https://eoscfuture.eu/. Accessed on: 2023-6-27.

[EOSC-Marketplace (2023)] EOSC Marketplace Resources. https://marketplace.eoscportal.eu/. Accessed on: 2023-6-27.

[ESRF (2023)] ESRF The European Synchrotron. https://www.esrf.fr/. Accessed on: 2023-6-27.

[ESRF Data (2023)] ESRF Data Portal. https://data.esrf.fr/. Accessed on: 2023-6-27.

[Expands (2023)] ExPaNDS: European Open Science Cloud Photon and Neutron Data Service. https://expands.eu/. Accessed on: 2023-6-27.

[ID29 (2023)] ID29 SMX - Serial Macromolecular Crystallography. https://www.esrf.fr/id29. Accessed on: 2023-6-27.

[ILL (2023)] ILL Neutrons for society. https://www.ill.eu/. Accessed on: 2023-6-27.

[ILL Data (2023)] ILL Data Portal. https://data.ill.fr/. Accessed on: 2023-6-27.

[Martin-Garcia J, Basu S (2020)] Macromolecular Serial Crystallography. Crystals 10 (12). https://doi.org/10.3390/cryst10121079

[P11 (2023)] P11 - High-throughput Macromolecular Crystallography beamline. https://photon-science.desy.de/facilities/petra_iii/beamlines/p11_bio_imaging_and_diffraction/index_eng.html. Accessed on: 2023-6-27.

[Panosc (2022)] The Photon and Neutron Open Science Cloud (PaNOSC). https://www.panosc.eu/. Accessed on: 2023-6-27.

[PSI (2023)] Paul Scherrer Institut (PSI). https://www.psi.ch/en. Accessed on: 2023-6-27.

[PSI Data (2023)] PSI Public Data Repository. https://doi.psi.ch/. Accessed on: 2023-6-27.

[Spence JH (2020)] Serial Crystallography: Preface. Crystals 10 (2). https://doi.org/10.3390/cryst10020135

[VISA ESRF (2023)] VISA (Virtual Infrastructure for Scientific Analysis) ESRF. https://visa.esrf.fr. Accessed on: 2023-6-27.

[Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016)] The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18