Designing a Cloud and HPC Based M&S platform to Investigate the IVD Diseases Mechanisms

Maria Paola Ferri[*1#2], Laia Codo[*1], Josep Lluis Gelpi[*1#2]

[#1]*Life Sciences Department, Universitat de Barcelona, Address Including Country Name*

mferrife108@alumnes.ub.edu, gelpi@ub.edu

[*2]*Barcelona Supercomputing Center, Plaza de Usebi Guell 1-3, Barcelona, Spain*

maria.ferri@bsc.es, laia.codo@bsc.es, josep.gelpi@bsc.es

***Keywords*— BioBB, Intervertebral Disc Degeneration, Workflow Framework, Portable Platform**

I. EXTENDED ABSTRACT

The main objective of the PhD proposal consists of the creation of a platform for IVD Models & Simulations (M&S) tools and their integration into automated workflows, within the HORIZON MSCA Disc4All project.

Based on the European Open Science Cloud (EOSC) vision, the expected platform will be a Cloud-based one, furnished with a front-end, to guarantee reproducibility, accessibility and easy use for experts and non-experts.

The development of an automated and specialized platform can represent the best hybrid technology with perks on both healthcare data management and computational environments exploitation, given the use of Cloud infrastructures on healthcare software and databases. Though rendering automatic not only the database, but prediction and simulation models in a user-friendly integrated system, may facilitate a difficult diagnosis and forward therapy, especially considering the various forces at play in a multi-omics data analysis of its kind.

A. *Introduction*

LBP (Low Back Pain) is the largest cause of morbidity worldwide, yet there remains undetected the liable specific cause, leading to poor treatment options and prognosis. As for other highly multi-factorial musculoskeletal (MS) disorders, such as lumbar intervertebral disc (IVD) and Lumbar Disc Degeneration (LDD), the interplay of a wide range of factors is poorly understood. Some genetic factors have been identified as possible biomarkers of specific LDD phenotype, such as Interleukin (IL)-, ECM protein- and protease-related genes, as well as human genetic variants (GDF5, SKT, PARK2 and CHST3)[1]. Beyond molecular data, also nutritional, clinical, social conditions are at stake for such diagnosis: molecular profiling, medical imaging, lifestyle data, sex are crucial factors to determine the predisposition of a single patient to these multi-factorial disorders[2].

Unfortunately, the integration of such data into a holistic and rational map of degenerative processes and risk factors has not been achieved, requiring the creation of professional cross competencies training programs.

The HORIZON MSCA Disc4All project is proposing to provide for an interdisciplinary solution to make up for the lack of diagnosis tools and instruments, considering all the biological, medical and computational contributions. The curation and integration of all the heterogeneous data sets, as well as experimental and theoretical/computational template and algorithms, are going to be embedded to exploit the multidisciplinary and multiscale models and simulations tools, based on image analysis, biophysics and biology[14].

B. *Methods*

The ultimate purpose of this Disc4All final infrastructure, deployed on the biological, bioinformatics algorithms, image analysis and ML/AI models and tools, is to integrate and interpolate primary patient data and achieve a coherent attribution of a MS phenotype.

To get a homogeneous configuration between all the tools, the first and main aim of the project is to contain all the workflow tools into a similar framework. This would be reached through the use of the BioExcel Building Blocks (BioBBs)[3], a collection of Python wrappers on top of biomolecular/ bioinformatics simulation tools. With their mutual unique syntax, they offer a layer of interoperability between the wrapped tools, to make their interconnection easier in the building process of a biomolecular/bioinformatics workflow.

The BioBBs also allow singular deployment of each tool within a Docker bio-container, but will also be assessed a private Cloud resource and its own Disc4All instantiation.

Sharing this technology, the tools produced for the Disc4all project would be effortlessly combined in the most convenient and comprehensive way, through the CWL (Common Workflow Language) or any other convenient workflow manager language.

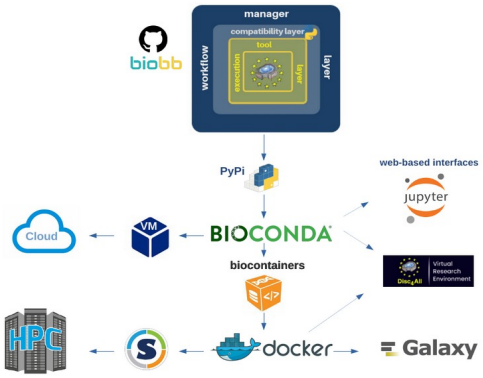

Fig. 1 BioBB's Automated deployment strategy[3]

In the end, thanks to the BioBBs adaptation layer, there are many possibilities for exposing the BioBB based workflow, such as Galaxy, Jupyter Notebooks, etc. One of the strategies for the front-end display is the web-server VRE (Virtual Research Environment)[3]. It will be exploited to construct the front-end of the infrastructure, so to provide for the non-expert users an easy-to-use platform to perform their analysis with the containerized pipeline.

C. Discussion and Results

The platform will provide the necessary environment to develop workflows using the tools developed in the Disc4All project.

The current objective for the development of the platform is the adaptation of the tools and software, given their computation and data requirements, into BioBBs. Since more of the half of the tools are in development, the best course of action is to start for each of them a single BioBB instantiation, so as to first define the basic pre-processing steps and environment. This introductory BioBBs are in production, one for Genomics data retrieval and the other one for a pre-Learning Active Cycle extracting features from imaging data.

Different modular Cloud containers are deployed for each BioBB, to have independent environments that could be possibly concatenated, following official FAIR workflow standards.

On one of them, the first draft for the final front-end platform is instantiated and exposed, in compliance with the openVRE environment. It is furnished with a simple case-use of the project (which is the aforementioned pre-processing BioBB data retrieval), that is up and working. This preliminary interface then is going to be completed with the rest of the tools, that could be called individually, and with the integral pipeline.

D. Future Enhancement

In the foreseeable future, the work on the adaptation to the BioBB framework will continue with the advancement of the tools, as well as the exploration of federated data access strategies with the furnished data model, for smooth flow of information through the platform, as well as exploring other data models adapters.

For the long term results, should be taken due account the different types of instantiations on Cloud-based and HPC environments, to build the computational infrastructure through Docker & EOSC cloud-based services, such as BSC's MareNostrum or StarLife (for users with granted HPC credentials), and to be able to probe the perks of hybrid workflows.

The front-end is a key point for the users (end-users and developers), to render available the workflow in the appropriate computational environments, so its finalization would be carried out in conformity with the progression of the study.

E. ACKNOWLEDGMENTS

This project is part of the Disc4All Training network to advance integrated computational simulations in translational medicine, applies to intervertebral disc degeneration and funded by Horizon 2020 (H2020-MSCA-ITN-ETN-2020 GA: 955735).

I'd want to convey my appreciation and gratitude to my supervisors and co-supervisor for providing me with the chance to work and learn from them in their welcoming group. I'd want also to thank my fellow colleagues and PIs from the Disc4All group, for their contribution and collaboration.

References

[1] Gantenbein B, May RD, Bermudez-Lekerika P, Oswald KAC, Benneker LM, Albers CE (2021) EGR2, IGF1 and IL6 Expression Are Elevated in the Intervertebral Disc of Patients Suffering from Diffuse Idiopathic Skeletal Hyperostosis (DISH) Compared to Degenerative or Trauma Discs. Applied Sciences 11(9): https://doi.org/10.3390/app11094072

[2] Eskola, P. J. et al. Gender difference in genetic association between IL1A variant and early lumbar disc degeneration: A three-year follow-up. Int. J. Mol. Epidemiol. Genet. 3, 195–204 (2012)

[3] BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows. Pau Andrio, Adam Hospital, Javier Conejero, Luis Jordá, Marc Del Pino, Laia Codo, Stian Soiland-Reyes, Carole Goble, Daniele Lezzi, Rosa M. Badia, Modesto Orozco & Josep Ll. Gelpi. Nature Scientific Data, 09/2019,Volume 6, Issue 1, p.169, (2019).

[4] Towards FAIR principles for research software, Lamprecht Anna-Lena, Garcia Leyla, Kuzak Mateusz, Martinez Carlos, Arcila Ricardo, Martin Del Pico Eva, Dominguez Del Angel Victoria, van de Sandt Stephanie, Ison Jon, Martinez Paula Andrea, McQuilton Peter, Valencia Alfonso. Harrow Jennifer. Psomopoulos Fotis, Gelpi Josep Ll., Chue Hong Neil, Goble Caroleu, Capella-Gutierrez, Salvador. 10.3233/DS-190026. Data Science, vol. Pre-press, no. Pre-press, pp. 1-23, 2019

[5] Felipe da Veiga Leprevost, Björn A Grüning, Saulo Alves Aflitos, Hannes L Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, Mingze Bai, Rafael C Jimenez, Timo Sachsenberg, Julianus Pfeuffer, Roberto Vera Alvarez, Johannes Griss, Alexey I Nesvizhskii, Yasset Perez-Riverol, BioContainers: an open-source and community-driven framework for software standardization, *Bioinformatics*, Volume 33, Issue 16, 15 August 2017, Pages 2580–2582, https://doi.org/10.1093/bioinformatics/btx192

Author biography

Maria Paola Ferri was born in Rome, Italy, in 1995. She graduated in 2018 in Biotechnologies, from the University of Rome La Sapienza, and in 2021 she obtained the Internation Master degree in Bioinformatics from the Alma Mater University of Bolumn. Her master thesis was focused on the creation of Cloud architectures designed to support and instantiate Bioinformatics applications, curing the portability and the user-friendly access to them. Right now, she is holding a position in the Life Science's Department at the BSC (Barcelona Supercomputing Center). Her role in the Disc4All project would be to extend the implementation of the IDD identification and prediction software in development into a Cloud and HPC environment, so to create in the end a portable and ready-to-use front-end workflow in the LBP (low back pain) investigation, for user and non-experts.