H2020 EINFRA-5-2015

www.bioexcel.eu

Project Number 675728

D2.3 – User Feedback and Future Roadmap

WP2: Portable Environments for Computing and Data Resources

Document Information

Deliverable Number	D2.3
Deliverable Name	User Feedback and Future Roadmap
Due Date	2017-10-31 (PM24)
Deliverable Lead	IRB
Authors	Adam Hospital (IRB), Josep Lluís Gelpí (BSC), Jose A. Dianes (EMBL-EBI), Pau Andrio (BSC), Darren J. White (UEDIN), Emiliano Ippoliti (Juelich), Adrien Melquiond (UU), Vytautas Gapsys (MPG)
Keywords	Feedback, Cloud Computing, HPC, Tools, Workflows
WP	WP2
Nature	Report
Dissemination Level	Public
Final Version Date	2017-10-30
Reviewed by	A. Bonvin (UU), J. Dianes (EBI), I. Harrow (IHC), B.d. Groot (MPG), E. Laure (KTH)
MGT Board Approval	2017-10-30

Document History

Partner	Date	Comments	Version
IRB, BSC, UU, Jülich, EPCC, EMBL-EBI	2017-10-02	First draft	0.1
IRB, BSC, MPG, EPCC	2017-10-18	Second draft	0.2
IRB,EPCC,EMBL-EBI,UU,IHC	2017-10-24	Third draft	0.3
BSC, Jülich	2017-10-25	Fourth draft	0.4
IRB	2017-10-30	Comments from PMB review addressed	0.5
KTH	2017-10-30	Final styling	0.6

Executive Summary

This deliverable presents the first internal (partners) and external (collaborators, users) feedback for the transversal workflow *Model Protein Mutants* and for the project pilot use cases. The transversal workflow, used as a prototype to test the designed workflows development process in the project and also the computational infrastructure, has been the main source of feedback. An update of the technical work behind the five project pilot use cases is given, emphasizing the feedback received.

Future roadmaps for the BioExcel Cloud Portal and for the workflows and computational infrastructures are presented, which showcase the work planned for the rest of the project. User experience will be the main focus for the cloud portal that is expected to offer a growing number of tools and deployable VMs in the coming months. Two kinds of benchmarks, a technical one (different computational infrastructures) and a scientific one (real scientific studies in an HPC exascale approach) are proposed using the transversal workflow.

Contents

1 Introduction

During the first half of the BioExcel project, the *portable environments for computing and data resources* work package worked on two different blocks:

- **Definition & Design**: Study of the state of the art e-infrastructures to be used in the center of excellence, identification and collection of a set of tools (building blocks) to be used in the construction of future biomolecular workflows, and definition of workflow prototypes (work presented in D2.1).

- **Development**: Design and deployment of biomolecular workflows, following a set of best practices (presented in ELIXIR EXCELERATE project [1]), with verification and benchmarking, easy to be found, deployed and executed (work presented in D2.2).

The second block is an iterative process, once a particular workflow is deployed and tested, user feedback needs to be collected to identify strengths and weaknesses, possible bugs or issues, and comments in general. This feedback should be then used to improve the workflow. BioExcel WP2 is currently starting the first round of this iterative process, collecting feedback from users (internal and external), and studying, for each of the pilot use cases and for the transversal workflow prototype, which are the next steps to follow.

Providing easy access to BioExcel computing and data resources through a range of workflow environments is one of the main responsibilities of WP2. For that reason, a couple of specific computational infrastructures have been designed and established: A testbed infrastructure at BSC and a production infrastructure at EMBL-EBI. Both infrastructures (EMBL-EBI already and BSC expected to begin in 2018) are providers for ELIXIR compute infrastructure and fully aligned with the forthcoming European Open Science Cloud (EOSC) infrastructure and standards, participating in several of the current pilot projects. This alignment will assure the enrolment of BioExcel with the new scenario of European e-infrastructures. The testbed infrastructure is used for development and testing of our workflows and virtual machines (VMs). The production infrastructure (BioExcel Cloud Portal, at EMBL-EBI) is the central point for users to find, deploy and execute the services provided by BioExcel partners. This production infrastructure is linked to the ELIXIR life science tools and data services registry bio.tools[2] and the European Grid Infrastructure Application Database (EGI AppDB). The portal is already on-line, accessible using an ELIXIR Authorization and Authentication Infrastructure (AAI) credential. Automatic downloading and deployment of the VMs registered in EGI AppDB under the BioExcel Virtual Organization (BioExcel VO, supported by ELIXIR VO) are already developed and tested. User accessibility (connection and login), data volumes associated, and VM monitoring are the main points to be addressed in the second part of the project.

The first feedback received from internal and external users as well as from collaborators is described in section 2, divided in feedback for the *Model*

Protein Mutants transversal workflow, and for the 5 pilot use cases. A future roadmap for the *Portable Environments for Computing and Data Resources* work package from now till the end of the project is presented in section 3, followed by conclusions in section 4.

2 User Feedback

2.1 Transversal Workflow Unit: *Model Protein Mutants*

The transversal workflow "*Model Protein Mutants*", extensively presented in the previous D2.2, has been used as a prototype to test not just the designed workflows development process in the project but also the whole computational infrastructure. Consequently, it has been the main source of feedback for the portable environments for computing and data resources BioExcel package.

2.1.1 ELIXIR

2.1.1.1 ELIXIR partnership

BioExcel has a strong relationship with ELIXIR, as demonstrated by the numerous links that can be found in the WP2 deliverables (bio.tools, EGI Virtual Organization, AAI authentication, etc.). This strong relationship led us to establish an ELIXIR and BioExcel partnership agreement where ELIXIR Tools and Interoperability platforms adopted BioExcel as a use case for the Tools and Workflows discovery & interoperability project. The project, led by Josep Lluís Gelpí (BSC-ES) and Carole Goble (UNIMAN-UK) intends to put together recommendations for the registration and specification of bioinformatics tools and workflows, enabling data scientists to properly describe analysis tools and workflows to make them interoperable across use cases. It is one of the *Interoperability Platform Implementation Studies* in ELIXIR EXCELLERATE WP5 (Interoperability platform). The work in BioExcel includes the description of workflows using Common Workflow Language (CWL)[3], the specification of software libraries using openAPI recommendations, and the registration and appropriate annotation of BioExcel Tools with EDAM ontology[4] (what in turn includes the extension of EDAM to include the description of biosimulation operations). In addition, provenance metadata, and test and reference data will be packaged using the research-objects approach[5]. The proposal was presented in ELIXIR All-hands (Rome - March 2017). The complete description of the Protein Mutants workflow prototype using CWL (described below), was chosen as the initial demonstration example. The Python library used in the workflow building blocks is being specified using OpenAPI. It is worth noting that the partnership between BioExcel and ELIXIR brings structural bioinformatics as a new use case for ELIXIR, which currently places heavy emphasis on genomics.

2.1.1.2 ELIXIR tools registries and EDAM ontology

In the first period of the BioExcel project, all of the tools collected in the catalogue of tools presented in D2.1 were registered in *ELIXIR Tools & Data Service Registry* bio.tools with a BioExcel tag. From that moment till the end of the project, these entries are being modified and curated (sometimes even removed). Moreover, the new Virtual Machines (VMs) produced by the project, implementing biomolecular workflows, are also being included in ELIXIR

registries (bio.tools, and openEBench). The first BioExcel VM to be registered in bio.tools was the one implementing the *Model Protein Mutants* workflow, which can be found in EGI AppDB. This process produces a bidirectional communication: in some of the cases, we discover missing information in our registered tools; in other cases, such as the VMs registries, we discover missing fields in bio.tools, that we find extremely important in these particular cases, such as the VM image link/URL or the image format. Other issues such as the possibility to share owner permissions with a group (e.g. BioExcel) for a particular entry or the addition of missing EDAM [4] ontology classes relevant to representation of structural biomolecular computation have been also discussed. The later point is important for BioExcel and also for all the projects working in the bio-structural fields, as EDAM ontology, as well as the whole ELIXIR project (as pointed before), is highly biased towards genomics data and tools, while EDAM is becoming the reference ontology for bioinformatics operations. The correct registry of structural tools, with properly defined input and output EDAM classes is crucial for the tools findability and interoperability, two of the main points of the FAIR Data principles[6] promoted by ELIXIR end the EC policies.

This feedback is produced in direct collaboration with Dmitry Repchevsky (BSC) and Jon Ison (DTU) from ELIXIR bio.tools.

2.1.1.3 Common Workflow Language

The software development process of biomolecular research workflows defined and presented in the D2.2 includes a complete workflow description using the Common Workflow Language (CWL)[3]. CWL is a specification for describing workflows and tools that is portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high performance computing (HPC) environments. CWL is an independent community-led effort, with implementations being developed for more than 9 workflow engines, 3 of which are already meeting conformance tests. CWL is being adopted by the ELIXIR's Interoperability platform, as the recommended language to describe workflows. The ELIXIR and BioExcel partnership is further developing the CWL specification language.

Working together with CWL co-founder Michael Crusoe, BioExcel partners at the University of Manchester have developed a graphical and interactive tool to represent workflows described in CWL. The CWL Viewer (https://view.commonwl.org/) is a richly featured web visualization suite, which graphically presents and lists the details of CWL workflows with their inputs, outputs and steps. It also packages the CWL files into a downloadable Research Object Bundle including attribution, versioning and dependency metadata in the manifest, allowing it to be shared easily. The tool operates over any workflow held in a GitHub repository. Other features include: Path visualization from parent and child nodes; nested workflows support; workflow diagram download in a range of image formats; a gallery of previously submitted workflows; and support for private Git repositories and public GitHub including live updates. The *Model Protein Mutants* workflow can be accessed through this link:

https://view.commonwl.org/workflows/github.com/bioexcel/pymdsetup/blob/master/cwl/mutations.cwl (reproduced in Fig. 1, September 2017 version), which is directly connected to the BioExcel GitHub repository. A poster presenting the CWL Viewer received the *F1000 Best Poster* award, at BOSC 2017 (ISMB/ECCB) conference.

The visualization of the first versions of the *Model Protein Mutants* prototype described in CWL identified an issue with the code modularity (interoperability, see D2.2 section 2.2.2). The way the building blocks composing the full workflow were designed implied the definition of a high number of parameters. This is a natural consequence of simulation tools being designed for command-line usage, but ended up with a large number of connections in the workflow diagram, that made that visualization unusable. The workflow components definition was then redefined, enclosing the set of required parameters in a single configuration file resulting in just a single connection in the diagram. It should be noted here that care is needed when grouping together sets of input/output objects, as different workflow managers (e.g. PyCOMPSs) use these files to automatically create a dependency graph, and thus they are required as independent objects. The new version has been tested successfully in different workflow managers (PyCOMPSs[7], toil[8], Galaxy[9]) and the representation can be seen in Fig.1 and in the CWL viewer.

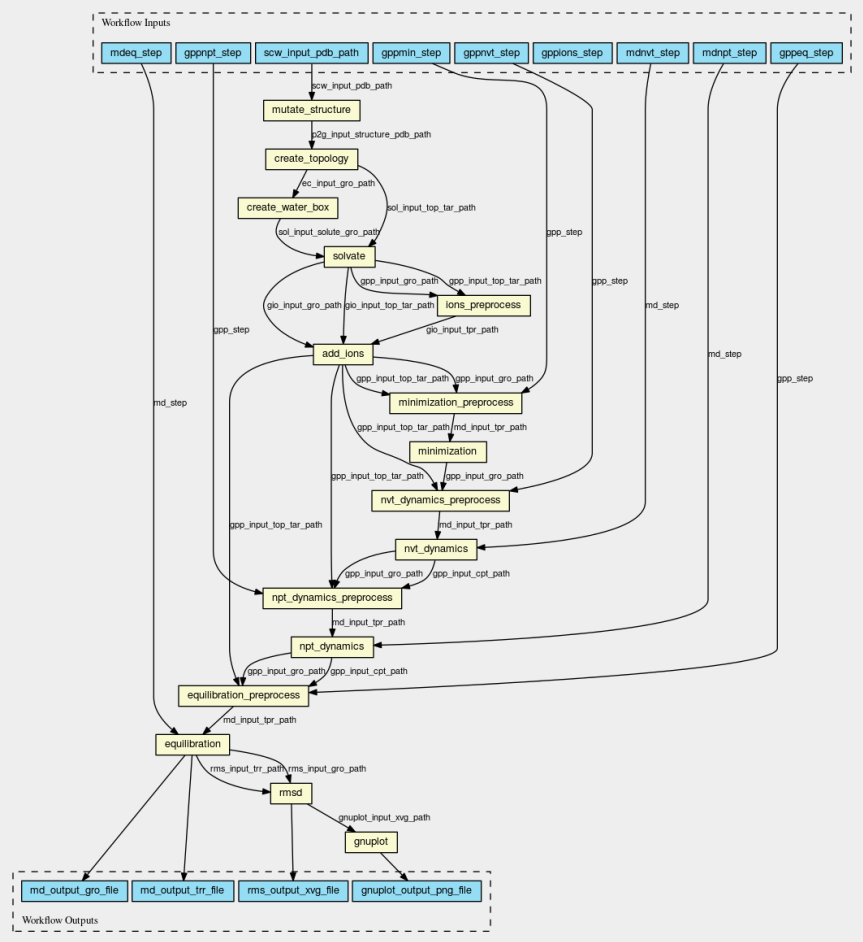

Fig. 1: Model Protein Mutants CWL diagram automatically generated by https://view.commonwl.org

This work has been done in direct collaboration with Michael Crusoe (CWL co-founder), whose feedback has been very helpful.

2.1.2 Cloud infrastructures

2.1.2.1 Virtual Machines, EGI and BioExcel VO.

The ultimate goal of BioExcel WP2 is to provide users with easy access to computing and data resources. To reach this goal it is important to demonstrate the feasibility to deploy and run the implemented workflows in several different infrastructures. A first attempt towards that was presented in the previous deliverable D2.2, with an initial benchmarking of the *Model Protein Mutants* workflow prototype run in an Open Nebula [10] cloud environment, in an Open Stack [11] cloud environment (EMBASSY Cloud), in the EGI grid infrastructure [12] and in the Marenostrum supercomputer (BSC). This benchmarking process kicked off another helpful feedback for this project package that is still ongoing. We have been collaborating with Enol Fernández and Gergely Sipos, from the EGI Foundation, who helped us to identify a couple of issues with our virtual appliances.

The first issue, solved already thanks to our partners in the BSC, was the necessity to implement cloud-init (http://cloudinit.readthedocs.io/) in BioExcel developed Virtual Machines. Cloud-init is a multi-distribution package able to handle early initialization of a cloud instance. Basically, it allows the VM to be deployed in any cloud infrastructure (Open Nebula, Open Stack, AWS, etc.) without the need for manual tuning. The inclusion of this package in our VMs allows us to directly upload them to EGI grid infrastructure (through the EGI AppDB, as described already in D2.2) and to start deploying them instantly. It also allows us to directly transfer VMs from our Open Nebula test bed at BSC in Barcelona, ES to our production Open Stack cloud infrastructure (EMBASSY Cloud) at EMBL-EBI in Cambridge, UK.

The second issue raised by the EGI Foundation was the requirement for BioExcel, as a Virtual Organization, to make available an associated set of service providers, through a group of data centers where the VMs can be deployed. Again, our strong link with the ELIXIR project is helping us with that, and our partners at EMBL-EBI, IRB and the EGI Foundation are working together to share the set of service providers with the ELIXIR EGI VO (vo.elixir-europe.org).

2.1.2.2 Cloud usage needs survey

The BioExcel Portal will present to researchers a list of life-science software supported by the project, which is obtained from the BioExcel tagged entries in the Elixir Tools Registry. From the BioExcel Portal a user will be able to select the service directly (if it is offered online as a service), find the HPC centres where it is already installed and available for use and how they can access the software, or to retrieve VMs from a repository and deploy the virtual machine or

container across different types of cloud infrastructure through the EBI Cloud Portal onto a cloud provider. The *Model Protein Mutants* workflow VM was the first one to be uploaded and deployed in the portal in the EMBASSY Cloud.

After this first experience, our BioExcel partners at EMBL-EBI conducted a *Cloud usage needs* survey as part of the Research Operations (ResOps) training workshops. The survey goal was to better understand the backgrounds and needs of users interested in different aspects of cloud computing. Additionally, they wanted to know if users were potentially interested in using web platforms that help with the process. Respondents have been part of the *ResOps* training which needs to be taken into account when making assumptions about the population sample (sample size of the survey was 23 users) and interpreting the results. It was clear that there is an implicit interest in the cloud and associated technologies among the attendants, and they have been shown a working version of the BioExcel EBI Cloud Portal.

The assumption is that, by better understanding our potential users backgrounds, we can focus on them and better understand their needs. That will eventually improve the quality of the products we offer them. The complete survey together with plots representing the results for each of the questions addressed can be found in the **Appendix** section.

Until we have more responses, we can't really arrive to any solid conclusion. There are some reasons to start believing that reproducibility is actually a big issue, and that people that get the ResOps training consider the cloud as a good tool to solve them. There might be also reasons to invest time and effort in promoting training, tools, and develop new solutions.

We might also have identified *bioinformaticians* as a good target for studying data scalability problems, but in this area we also see that the cloud might not be perceived as good a solution as we might expect. This might be due to lack of training or some specifics to the user problems that can't be solve easily with cloud computing. Other factors that might have an effect on this result could be the available bandwidth and workplace settings and security policies (e.g. firewalls).

2.1.3 Nostrum Biodiscovery

One of the most helpful sources of feedback received so far in BioExcel WP2 comes from Nostrum Biodiscovery. It is a BSC spin-off which aims to collaborate with pharmaceutical and biotech companies dedicated to the development of drugs and molecules of biotechnological interest, with the main focus based on helping these companies to maximize the success of their drug discovery and development process and consequently, increasing their market success. Nostrum Biodiscovery is collaborating directly with us in the Pilot use case 5: *Virtual Screening*, presented in a following section, but it has shown interest also in our workflow prototype *Model Protein Mutants*.

The *Model Protein Mutants* pipeline, thoroughly described in D2.2, is an automatic protocol to generate structures for protein variants detected from genomics data. These structures are then prepared, run, and analysed using Molecular Dynamics simulations. Nostrum Biodiscovery helped us in the first rounds of testing of this workflow, and proposed a case study that is being examined currently. The case study is a really well known case of interest in the pharmaceutical field, the Epidermal Growth Factor Receptor (EGFR) [13, 14]. EGFRs are transmembrane receptors located on the cell membrane. They have an extracellular binding domain, to which the peptide Epidermal Growth Factor (EGF) binds, a transmembrane domain and an intracellular tyrosine kinase domain. EGFRs play an important role in controlling normal cell growth, apoptosis and differentiation. Mutations of EGFRs can lead to abnormal activation and signal transduction causing unregulated cell division and ultimately driving some types of cancers, including carcinoma[15, 16] and glioblastoma[17, 18]. Thus, dysregulation of EGFR activity has been implicated in the oncogenic transformation of various types of cells and represents an important drug target[19]. A massive study of protein mutations, and their effect in the dimerization possibility can be performed using our workflow. As a vast amount of experimental information is available for this particular case, we consider it a strong use case to validate our workflow prototype. Moreover, this study reaffirms the importance of being able to run the pipeline in the exascale regime. The workflow, together with PyCOMPSs workflow manager[7], is designed to be run in thousands of processors, distributing the MD simulations for each mutation among the available machines, and automatically dealing with the different dependencies. Proving the viability and correctness of this case study with our developed workflow will of course open the door to new and bigger pharmacological studies involving protein mutations.

2.2 Pilot Use Cases

2.2.1 Pilot Use Case 1: High Throughput Workflow for Cancer Genome Sequencing Data

Working with partners in the Institute for Genetic and Molecular Medicine (IGMM) at the University of Edinburgh, we are in the process of developing an automated pipeline for rapid turnaround cancer analysis of high throughput sequencing data that we hope will improve the speed, robustness and ease of use of the current workflow. For example, at present, there exist several stages at which human interaction is required before proceeding with further sections of the pipeline. There are also several stages at which more sensible management of computing resources on HPC systems (such as Cirrus, a Tier 2 HPC system managed by EPCC) are required. The workflow we are developing exists in two stages: Sequence Quality Control (SeqQC) and Alignment.

- **Sequence Quality Control Stage - Workflow**

The current workflow (Fig. 2) currently consists of several stages that primarily use individual software: summarizing quality of samples (*FastQC*), and trimming adapter and poor quality reads (both using *cutadapt [20]*). In the original workflow, each read summary output from FastQC is examined by a human, who then decides on a course of action, such as passing on to the next stage of the workflow, running trim steps and/or a possible second run of FastQC and trimming. This is now automated in stage 1, and has been developed in such a way that implementation into larger workflows should be relatively easy to achieve for partners and other users.

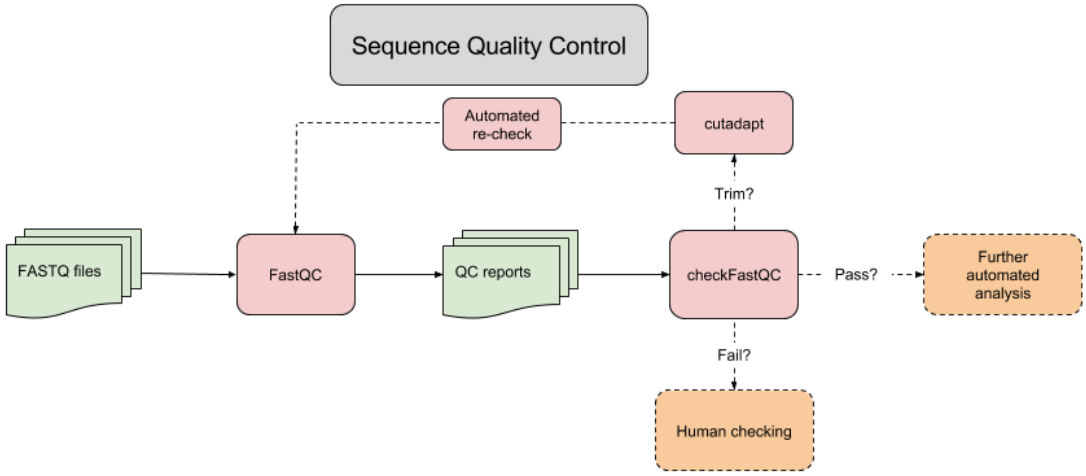

Fig. 2: Sequence Quality Control stages workflow

- **Sequence Quality Control Stage - Feedback**

The majority of the communication with UC1 collaborators has been to specify decision-making steps within the *checkFastQC* portion of the workflow in order to minimize manual user interaction. As of late September 2017, this stage

of the workflow is at a user-testable stage, and further changes will be made depending on the outcome of large-scale testing.

The primary issues we encountered revolve around the suitability of current workflow management solutions commonly used by the BioExcel community to these types of workflows. After early tests, and discussions with both our partners and Brad Chapman (creator of *BCBio [21]*) it was decided that a more bespoke workflow would be required, due to several issues raised. BCBio does not provide suitable flexibility for more bespoke workflows. Future versions of BCBio will slowly integrate *Common Workflow Language (CWL)* compliance, allowing user to create workflows in CWL that leverages parts of BCBio as needed. However, the current CWL specification does not provide conditional workflow paths (such as being able to check output states and re-run/loop back over previous steps). To do so would require developing a bespoke CWL runner that handles this internally, which is outside the initial scope of this use case at this time. However, future implementations of the workflow may take advantage of CWL if deemed suitable. However, current implementation is a series of Python wrappers controlling the execution of the workflow stages as needed. We also had some issues accessing suitable test data, due to the security requirements surrounding human genome sequencing data, but this is now resolved.

The work done so far, while only being small workflow, allowed us to get to understand how the larger, more complex parts of the pipeline can be best developed to work with the types of computing systems and expected data throughput required. This should speed up development of the later stages. Immediate future plans for this stage is to ask partners to run several tests of the workflow and provide feedback to direct future development. There are also several improvements that can be made to the current Sequence Quality Control workflow, given time:

- Create easier method of allowing user to alter *checkFastQC* flags/decision making.
- Find/test available multithreaded *cutadapt* and *FastQC* implementations.
- Create CWL-compliant wrappers for each stage of the workflow.
- Implement CWL-compliant management of workflow execution

- **Alignment Stage - Workflow**

The current Alignment stage workflow (Fig. 3) consists of several mapping, alignment and sorting tools working in parallel: BWA-MEM[22], Samblaster[23], Samtools[24] View/Sort/Index, and *GATK[25]* BaseRecalibrator/ PrintReads tools. Work on this portion of the workflow is at an early stage, but prior experiences gained from Sequence Quality Control development means future progress should be smoother.

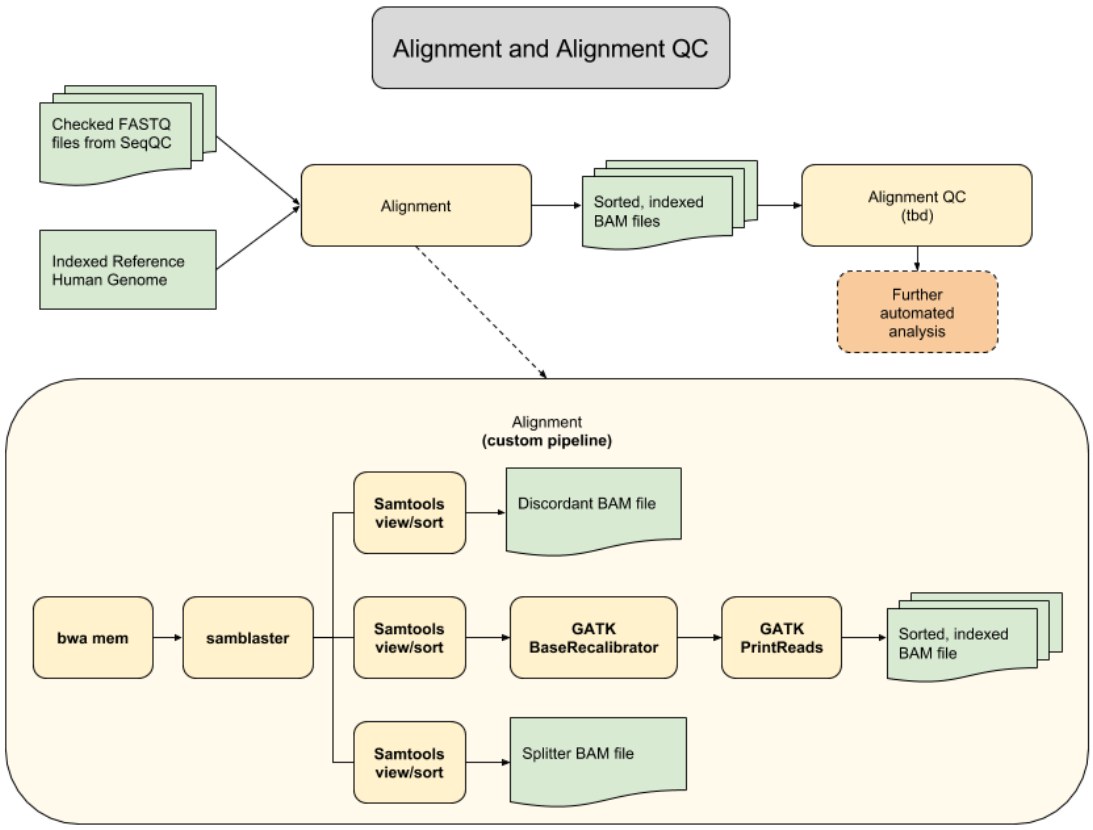

Fig. 3: Alignment workflow stages

- **Alignment Stage – Feedback**

To date, collaborator discussions have centered around how the current workflow (in the form of a bash script) performs on their current system compared to Cirrus. We have also discussed the suitability of the Halvade [26] implementation for WGS pipelines, build on Hadoop and using Map/Reduce parallelization. Future discussions on the Alignment Quality Control stage are planned.

Testing so far has been to investigate how thread allocation performs on a Cirrus-like HPC system for this type of workflow. Since BWA-MEM is a non-MPI multi-threaded program (using pthread for thread creation and control), there were some issues with the standard documentation and Cirrus job control system incorrectly placing all threads on a single core. A workaround is known for workflows like ours, so testing can continue unaffected. This should be fixed in future updates to the system. We have performed initial testing to investigate how the current Alignment stage as a whole handles threading. The main step involves pipes/redirects between 12 different instances of the programs listed above. It may be a case that the system attempts to place all multi-threaded processes across all available cores/threads when executed as described by partners. This could result in those threads constantly switching between processes, creating a significant bottleneck. This can create further issues since the current maximum run time for jobs on Cirrus is 96 hours, which is common

amongst current similar HPC systems. This will likely require a more detailed investigation of each step in the Alignment stage to best understand how to distribute threads for maximum performance. Testing of the Halvade tool is also planned, and could resolve many issues if early testing by EPCC partners scales to our use case.

2.2.2 Pilot Use Case 2: Free energy simulations of biomolecular complexes

Free energy gradients underlie all biomolecular motions, thus governing the processes at molecular level that are essential to maintaining life, such as protein folding, DNA recognition by transcription factors, substrate binding to enzymes etc. It is the identification of these free energy differences that allows understanding and, subsequently, controlling properties such as the affinity of a drug binding to its target, a protein's resistance to extreme temperatures or a precise antibody recognition of an intruding antigen. Molecular dynamics (MD) based simulations are particularly suited to investigate free energy gradients, as the simulations with physical rigour incorporate both entropic and enthalpic contributions to free energy for a well defined statistical ensemble. While there are many flavours of approaches to extracting free energies from simulations, in the current BioExcel Use Case we concentrate on the alchemical methods.

At its core the method of computational alchemy rests on the notion that thermodynamic properties like the free energy are path independent, so even pathways that are in practice inaccessible, such as alchemical transitions, yield meaningful quantities from simulation. To establish a link to experiments, thermodynamic cycles are frequently employed. For example, cycles can be used to efficiently compute the change in protein stability upon an amino acid mutation or the change in protein-DNA affinity upon a base mutation in the DNA. Instead of computing the cumbersome folding/unfolding or binding/unbinding, the alchemical mutation free energy can be computed efficiently. Naturally, the setup of such specific alchemical simulations differs from a standard MD setup and in fact can become highly technically involved. To facilitate the procedure of free energy calculation setup and the subsequent simulations we are developing a software package called pmx, developed in collaboration with Boehringer Ingelheim.

- **Workflows**
    - **pmx for proteins**

pmx readily allows setting up free energy calculations for the amino acid mutations. All the canonical amino acid combinations have been collected in a special set of libraries that enable an easy to use automation of the setup procedures. As the MD simulations rely on empirical force fields, the pmx based mutation libraries are also compatible with a number of contemporary molecular mechanics force fields. The approach provides

means to perform large scale mutation scans to assess changes in protein thermodynamic stability, protein-protein or protein-ligand interactions.

- **pmx for DNA**

Similarly to the amino acid mutation setup, pmx also supports nucleic acid mutations in DNA. A large scale nucleotide mutation scan over a number of protein-DNA complexes has demonstrated that alchemical pmx based free energy calculations are capable of capturing correct trends in DNA interactions with various transcription factors and nucleases. Furthermore, we are working on providing support for nucleotide mutations in RNA as well.

- **pmx for ligands**

Automation of the alchemical ligand modifications entails an additional challenge: in contrast to amino and nucleic acids a library of mutations cannot be pre-generated for an arbitrary set of molecules. For that purpose we are developing an algorithm, which could identify an optimal atom mapping for any pair of organic molecules. Furthermore, the prototype of this approach is readily capable of suggesting best suited pairs of ligands to be modified, this way providing an efficient way to navigate in a chemical library guided by the alchemical free energy calculations.

- **Webserver**

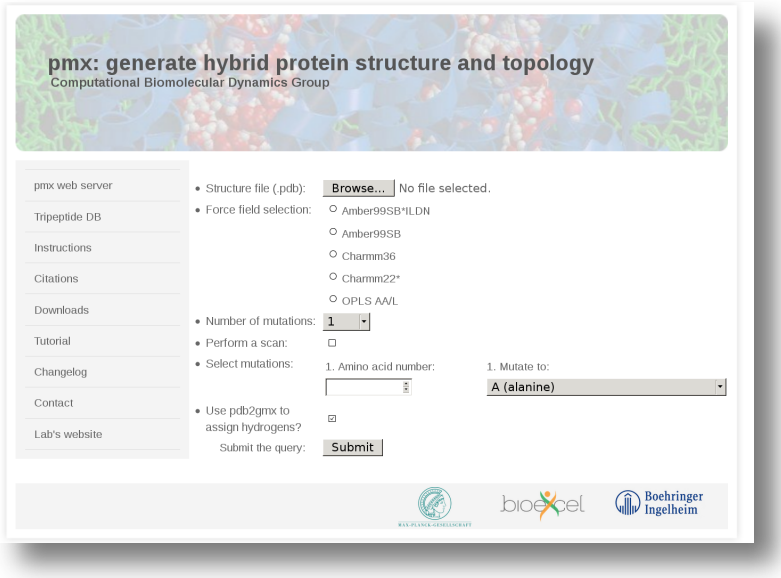

Fig. 4: pmx web server interface

The pmx utilities can be used as a command line. While such an approach may be mainly attractive to a power-user who needs to perform large scale

mutation scans, we also provide support for the hybrid structure/topology generation via a web-based interface. Both amino and nucleic acid mutations can be generated online: http://pmx.mpibpc.mpg.de. The webserver interface not only allows performing single point mutations, but also enables setting up scans over a protein by a selected amino acid or a full scan of a DNA chain.

- **User Feedback**

Two main channels for user feedback are currently available. Firstly, in one year since the launch of the pmx webserver, more than 100 unique IPs have used hybrid structure/topology generation utilities. Subsequently, users have provided us with the feedback which could be divided into three categories:

- **Bug reports**: e.g. user feedback helped identifying shortcomings in the approach where all bonds are constrained during free energy calculations;

- **Feature requests**: Based on these requests we have generated additional force field libraries and incorporated proline mutations;

- **Usage questions**: In these cases users mainly requested information about the actual free energy calculation protocols. Such enquiries have prompted us to start developing a user friendly automated framework for the full free energy calculation procedure and result analysis complemented by a set of detailed tutorials.

The Free Energy Calculation workshop in London, 2017, served as another important platform for user feedback. This two-day workshop allowed for a face-to-face communication with the users and other developers in the field:

- In a round of 30 participants involved in software development a number of important issues were covered by discussing best practices in free energy calculations, software accessibility, user friendliness of the developed tools and various technical aspects.

- In a larger round of 100 people, questions of a more general scope were discussed covering future directions of the field.

Additional feedback followed the London Free Energy workshop from the developers of free energy calculation methods in the USA. Following the success of the London workshop, prospects of a potential collaborative conference are being considered.

There are several major branches for the pmx future development. Firstly, the core algorithms have been updated to allow for an extended functionality, e.g. proline mutations, constraints on hydrogen atoms involving bonds only.

Naturally, this development requires re-building protein and DNA mutation libraries, as well as recalculating tripeptide free energy database. Secondly, the ligand modification utilities are to be officially released and made openly accessible to users in a stable pmx branch. Finally, in addition to the hybrid structure/topology generation, we are developing a framework for the non-equilibrium free energy calculation setup and subsequent result analysis.

2.2.3 Pilot Use Case 3: Multi-scale modeling of molecular basis for odor and taste

Use Case 3 is a collaboration between the JUELICH unit of BioExcel with the European Human Brain Project (HBP) initiative (http://www.humanbrainproject.eu/en/) on the investigation of the enzymatic reaction of the adenylyl cyclase (AC) bound to the G-protein (Gsa) and in complex with an ATP strand, which stimulates the cAMP synthesis amplifying signal transduction in the brain. The enzymatic reaction involves a (complex) chemical reaction and therefore it requires to explicitly describe the electronic degrees of freedom and in turn a quantum mechanical treatment. A full quantum description however is far beyond the current computational capabilities for systems of such a size. The JUELICH unit has the expertise in hybrid quantum mechanical/molecular mechanics approaches required to study those large systems. In addition, the JUELICH unit is interested in investigating a complex problem like the one proposed by HBP in order to later use such molecular dynamics simulations and corresponding results as reference for the comparison with the outcomes of the application of the new QM/MM interface, currently under development, to the same case, in order to have a strong validation of the new interface with a real biological case.

The HBP research group involved in this study is one of the five groups dedicated to molecular simulation in the Brain Simulation Platform (i.e. sub-project 6 or SP6) of HBP. While other partners in the molecular simulation team perform protein diffusion simulations in crowded environments or coarse-grain simulations of protein-protein complexes, the task of this research team is to provide atomistic information of enzyme-catalyzed chemical reactions and of protein-ligand binding and unbinding processes that govern neuronal cascades.

As molecular modelers, their interest in AC is to understand how the binding of Gas (Gai) stimulates (inhibits) cAMP synthesis. The study is hampered by the fact that only the X-ray structures of AC bound to Gas are available. Actually a structure with an ATP mimic is also available, though the ATP is not in a reactive conformation. HBP proposed to use the last one as starting structure and we performed an initial classical molecular dynamics in order to bring the system towards a reactive conformation before changing the level of theory to QM/MM and describe the chemical reaction at DFT level through the original QM/MM interface of the CPMD code.

- **Workflow**

The pipeline of the proposed iterative multistep approach is the following:

1. Establishing of the quantities/properties that have to be used to test the quality of the modeling
2. Initial modeling of the system
3. Equilibration step through traditional force field based level of theory
4. Intermediate QM/MM modeling by employing a lower QM level of theory (e.g. the computationally inexpensive semiempirical methods).
5. Test of this QM/MM modeling on the relevant identified properties: if the test fails, we go back to step 2.
6. QM/MM modeling by employing a more accurate QM level of theory (e.g. DFT method).
7. Test of this QM/MM modeling on the relevant identified properties: if the test fails, we go back to step 2.
8. Full QM/MM simulations at the higher level of theory.

Step 4 could be in principle divided in additional more intermediate steps, each one at an increased quantum level of theory. However, we do not explore this possibility in Use Case 3.

- **User Feedback**

Since this Use Case is a test case for the above proposal, the current user feedback is represented by the interaction with the HBP research group in the attempt to get a reliable model of the AC system and perform through this iterative approach a satisfactory investigation of the energetics of the enzymatic mechanism. We initially tried to model the reaction starting from the protonated nucleophile (O3'H). We experience several problems in controlling the proton transfer because several pathways can be in principle followed by the proton to reach the gamma-phosphate (that triggers the release of the pyrophosphate) but the relevance of each one is unknown and therefore none of them can be excluded a priori. In spite of these difficulties, we managed to have a preliminary estimate of the barrier, which turned out to be higher than the experimental data. For this reason, we speculate that a stepwise mechanism could be in place: deprotonation of the nucleophile, then nucleophilic attack. Studying the deprotonation of the nucleophile is inherently difficult, and moreover other studies from HBP groups and in literature indicate that it is not a rate limiting step. Therefore, we decided not to model directly this step but we moved the proton from O3'H to the gamma-phosphate and we started modeling the nucleophilic attack. With the first attempt, the system turned out not to be stable: one of the two magnesium ions that are involved in the reaction mechanism loses a ligand and its 6-coordination. In order to better understand if this instability origins from a problem with our modelling or if it is an inherent chemical instability we performed a lower detailed analysis by employing a computational less expensive semiempirical level of theory in the description of the quantum part. The results show that the structure is stable at that level of

theory and consequently we are brought to think that our original modelling of the system with the proton on the gamma-phosphate is not accurate enough. We have then developed a new modelling for the deprotonated case (O3') that is computationally slightly more expensive and that should eliminate the drawbacks of the first unsuccessful modelling. The corresponding simulations are currently running.

2.2.4 Pilot Use Case 4: Biomolecular recognition

Biomolecules are hardly monogamous, therefore studying their interaction at an atomic resolution is fundamental to the understanding of their functions, to design inhibitors or drugs that can modulate their activity and to rationalize the effect of a genetic mutation. High-throughput experimental techniques generate a wealth of qualitative and quantitative data, but the structural dimension is often missing, which calls for complementary modelling approaches. Moreover, the large number of interactions translates into an even larger amount of data, which require HPC and HTC solutions, automated workflows and cutting-edge technologies for the interactive and integrative manipulation, analysis and visualization of the data.

- **Workflow**

The biomolecular recognition use case relies on two well-established methods in computational biomolecular science, namely molecular docking and molecular dynamics (MD) simulations:

- Molecular docking refers to the prediction of three-dimensional structure of biomolecular complexes from their constituents. In the group of Prof. Alexandre Bonvin (Computational Structural Biology group, Utrecht University), we develop HADDOCK[27] (High Ambiguity Driven protein-protein DOCKing), a holistic and versatile information-driven docking software. It distinguishes itself from other approaches by its ability to integrate various information sources derived from biochemical, biophysical or bioinformatics methods to enhance sampling, scoring, or both. HADDOCK also allows direct and flexible modelling of large assemblies consisting of up to six different molecules, which, together with its rich data support, provides a truly integrative modelling platform.

- MD simulations are essential to study at an atomic level the motions of large and complex biomolecular systems. This technique is used to investigate the thermodynamic ensemble of the system in realistic environments at room temperature, understand its dynamic and can predict free energies of binding for protein-small ligands complexes. The implementation of this use case requires to combine HADDOCK with molecular dynamic simulation suites, such as Gromacs[28], into a functional workflow that would allow for modelling, simulation and data analysis. With this, we are paving the way for the systematic and automated modelling of biomolecular interactions, leveraging the

performance of HPC/HTC infrastructures supported by BioExcel. To this regard, it represents a first step toward reaching exascale computing for interactome modelling, one of the major challenges in the field of computational structural biology.

- **User Feedback**

In this context, we have signed a collaborative partnership with Dr. Daan Geerke (molecular toxicology group, VU Amsterdam), which formalizes our joint effort to combine HADDOCK with MDstudio, a microservice-based molecular dynamics workflows developed in his group with the support of the Dutch e-Science center. MDstudio relies on crossbar.io, a Web Application Messenger Protocol (WAMP) that allows building distributed systems out of application components that are loosely coupled and can communicate in real-time. In their current setting, both protein-small ligand docking and binding affinity prediction workflows are implemented in MDstudio, using both Gromacs and PLANTS[29] (a renowned software for protein-ligand docking). HADDOCK has been added to their initial design as a new service, via a remote procedure call on our server and the development of a specific python module to handle the communication. We are now entering the test phase and, after performance tuning and optimization in collaboration with the Gromacs core developers, we hope to deploy this service in Q1 2018.

In the meantime, we already started benchmarking two different applications of this project:

- Rescoring of the poses generated by HADDOCK
- Testing the stability of the best cluster representatives (post-docking analysis)

The input data for HADDOCK are protein, peptide or nucleic acids structure coordinates, either coming from a PDB code or manually uploaded. After modelling, we systematically run for each cluster representative a 50ns MD simulation in a solvated box using GROMACS 2016. The first application consists of improving the scoring function of HADDOCK by taking into account Gromacs energy terms or other parameters such as the number of hydrogen bonds at the complex interface. This only requires very short MD simulations (no longer than a few picoseconds) and could easily be systematically applied to the ~75-100 job submissions we receive daily on our HADDOCK web server.

The second application, which is also related to scoring, aims at testing the stability of the predicted clusters by running MD simulations for longer timescales. Scoring is not an absolute number but rather a relative metric to compare the different solutions of a same docking run, we want to investigate whether MD can help getting insights into the “likeliness” of a predicted complex, by discriminating between stable and unstable interfaces. Cases were selected from the protein-protein docking benchmark 5.0 and specific CAPRI targets for which the scoring was particularly challenging.

Finally, we would like to take advantage of MD simulations to sample the conformational space of peptides and use these conformational ensembles as an input for HADDOCK.

2.2.5 Pilot Use Case 5: Virtual screening

Use case 5 will allow users to run ensemble docking using Open PHACTS to obtain pharmacological compounds in combination with the Gromacs[28] MD engine to prepare MD ensembles and HADDOCK[27] / Seabed[30] to run biomolecular docking.

- **Workflow**

The pipeline behind this workflow has been divided in different steps, which are being implemented in parallel (see D2.2 for details and figure):

1. Recover Protein Structure/s and prepare MD ensemble.
2. Enhance Sampling.
3. Biomolecular Recognition.
4. Open PHACTS integration.

The complete pipeline is available as separate tools, and is being progressively adapted to our Python wrapping schema in order to easily integrate its functionalities together.

- **User Feedback**

The main sources of feedback for the Virtual Screening pilot use case are Nostrum Biodiscovery (presented in section 2.1.3) and Open PHACTS foundation.

Nostrum Biodiscovery, similarly to what had been done for the *Model Protein Mutants* workflow, proposed a possible use case to validate the *Virtual Screening* pipeline. The validation system proposed for this workflow is again EGFR, because of its great interest in the pharmaceutical industry. Currently, there are two therapeutic approaches hitting EGFR. One of them is based on monoclonal antibodies which bind to the extracellular domain of the receptor, antagonizing either the interaction with its cognate ligand (EGF) or its homo or hetero dimerization. The second therapeutic approach is knocking down its tyrosine-kinase activity. This is also a very interesting option as there are several therapies with marketing authorisation approvals that target its kinase domain. Approved small molecule drugs in this category are ATP competitive inhibitors, either reversible or covalent. An example are: Gefitinib, Vandetanib, Lapatinib, Erlotinib and Afatinib[31, 32]. Although structurally related, some of them require conformational changes in the receptor and thus bind to EGFR kinase domain with some degree of induced fit. Importantly, the administration of this treatments imposes a selection pressure on the cancer cells which eventually

develop mutations in the kinase domain that lead to resistance. One of the most prevalent mutations found in treated patients is the T790M mutation. This change is located in the so-called “gatekeeper” residue, in the interior of the ATP binding site. The replacement of a small threonine amino acid for a much bulkier methionine precludes or partially hinders the binding of the ATP competitive treatments listed above. This problem has spawned the development of a next-generation of ATP competitive inhibitors that target the T790M mutant, such as osimertinib, rociletinib, HM61713, ASP8273, EGF816 and PF-06747775[33]. Thus, a whole number of first and last-generation small molecule inhibitors is available for this target, some of them hitting the wild type sequence, others specifically designed for hitting mutant variants and having no activity on the WT. This whole body of knowledge can be exploited for setting up and fine-tuning the VS workflow, for testing its performance and reliability in a real target that is nowadays exploited in the clinics.

Open PHACTS platform is helping us in the first steps of the pipeline: the retrieval of a library of compounds and receptors of interest for this particular study. Discussions are still ongoing, but we are receiving a nice feedback from Alasdair Gray and Nick Lynch, from Open PHACTS project about the way we can work with the Open PHACTS API, and the information we can retrieve from the platform. From these discussions, and in collaboration with Open PHACTS platform, a BioExcel webinar was presented showing the main points of the VS use case: BioExcel and OpenPHACTS: Building pharmacological workflow blocks for Virtual Screening.

3 Future Roadmap

BioExcel WP2 future roadmap is described in the next sections, divided in two main blocks: **Cloud Portal** and **Workflows & Computational infrastructures**. Timeline roadmap diagrams are presented for each of the blocks, where grey arrows represent work already done, green arrows represent ongoing work, and blue arrows represent future work.

3.1 BioExcel Cloud Portal

The BioExcel cloud portal has gone through two major developments in the last six months. First of all, usage tracking of cloud resources deployed through the portal has been implemented through an ELK (Elastic-Search, Kibana, Logstash) system. Second, there has been important developments in User Experience including both, user research through a Cloud Usage needs (described in 2.1.2.2 and Appendix sections) and User Interface improvements that include a simplified deployment workflow (Fig. 5, here in the context of the user steps to perform a cloud deployment).

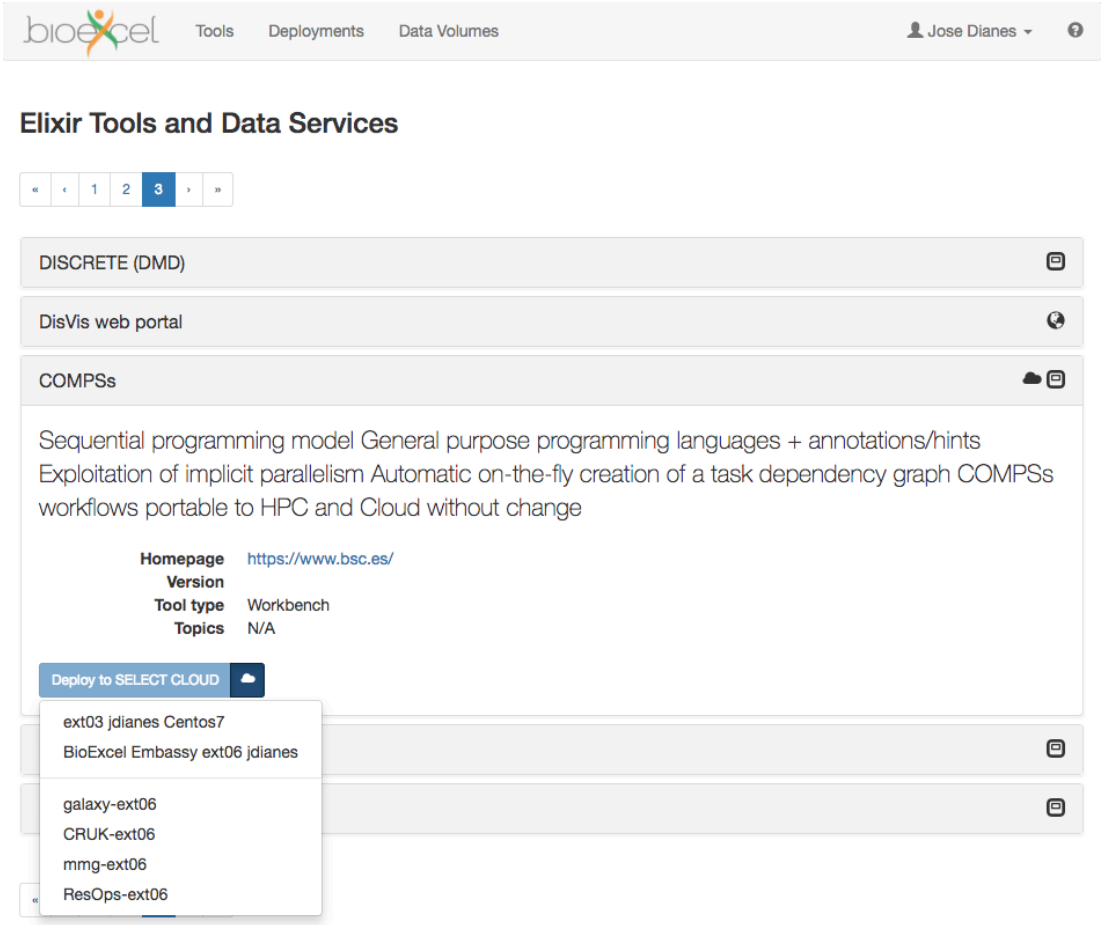

Fig. 5: UI improvements: pre-defined deployment configurations

The next six months will carry on with the User Experience work, including new iterations of the questionnaire described in section 2.1.2.2, usability testing, and interviews. Future work will also involve intensive work on developing virtual infrastructure to be deployed through the BioExcel cloud portal. This includes both, Virtual Machines but also Data Volumes that will be used to persist data and results after the computing part is finished.

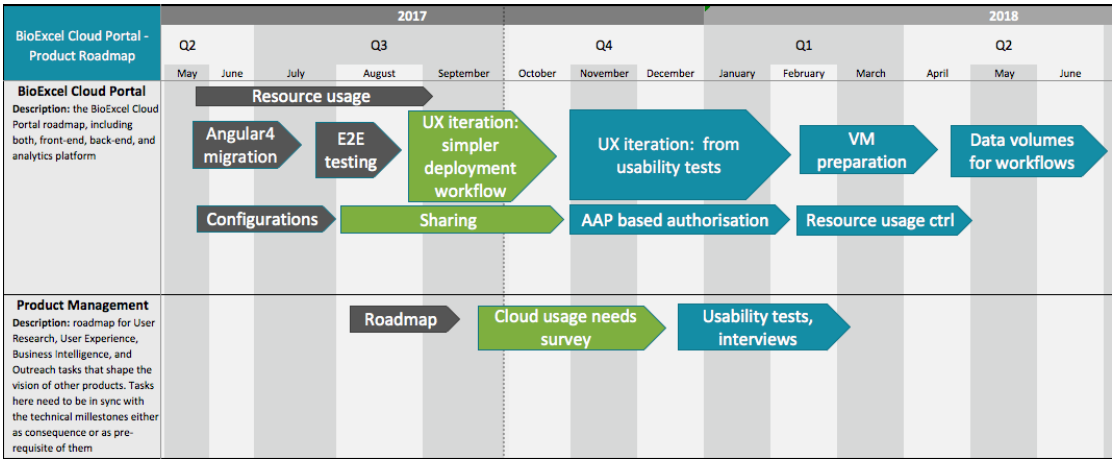

Fig. 6: BioExcel Portal roadmap.

The roadmap reflects all this. Current work is centered around the UX work already mentioned, together with being able to share configurations in order to deploy BioExcel VMs at shared cloud resources. In the meantime, we will continue working on the cloud usage needs questionnaire.

In the next months until the end of the year, we can start doing user interviews and observations in order to improve the user experience. The VM development will bring more usage and more opportunities to carry on with UX improvements.

Finally, we will implement mechanisms that, using the usage tracking capabilities already mentioned, will allow controlling the usage of shared cloud resources when deploying VMs and data volumes.

3.2 Workflows & Computational Infrastructures

Roadmap timeline for the workflow and computational infrastructures is shown in figure 7 below.

The work on the reference transversal workflow and the different pilot use cases will continue on different levels, depending on the use case (see section 2.2). Specification and description of the *Model Protein Mutations* workflow is ongoing and will be presented when finalized at ELIXIR project as a use case for Tools and Workflows Interoperability project. CWL description is done and available from the cwl-viewer tool, OpenAPI specifications for the workflow Python modules are being written, and EDAM ontology is being updated with new structural classes.

The work in computational infrastructures will continue with the implementation, register and deployment of new VMs that will be made available from the Cloud Portal. On the other hand, the transversal workflow, which is currently being used to generate a wide technical benchmark (proving the ability to run in different infrastructures, including large scale HPC resources), will be used to develop real scientific studies in HPC environments.

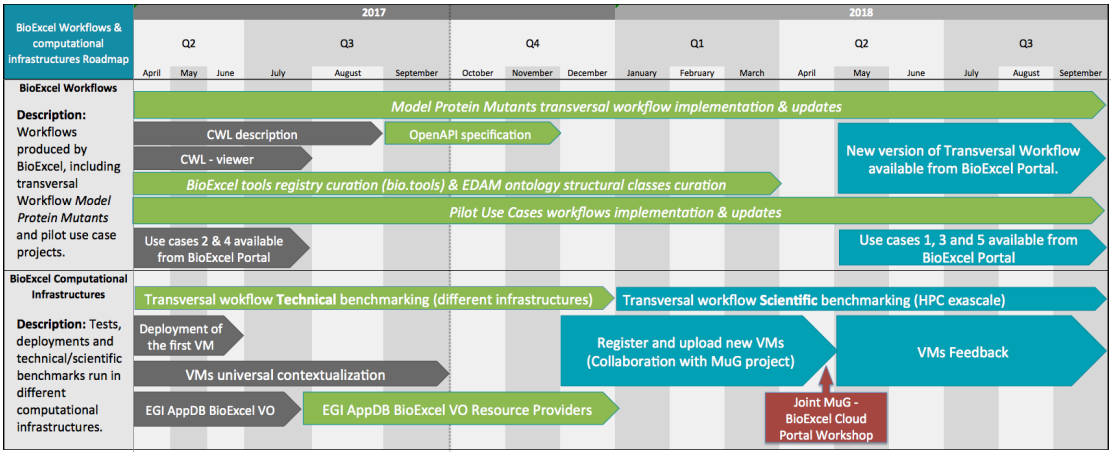

Fig. 7: BioExcel Workflows and Computational Infrastructures roadmap

The main points represented in the timeline are being described in more detail in the following sections.

3.2.1 Cloud Environments

As commented in the section 2.1.2.1: *Virtual Machines, EGI and BioExcel VO*, the problem of Virtual Machines contextualization needed to be addressed before going ahead with the plans for this work package. VMs in BioExcel were created following standard recommendations from static templates containing a basic configuration to achieve elasticity. However, during the deployment phase, instance specific settings (subject to the cloud environment) must be injected into the VM to correctly turn it on. This process is known as contextualization, and, although is a well-known task, it has not been easy to solve, and it is an essential point to automatically deploy our VMs in different cloud environments, without the need of a manual tuning. Now with this point solved, the next steps

in the coming months for the project is to start registering and deploying new VMs into the infrastructure, including the correct registry of the Virtual Appliance (VA) into the EGI AppDB database, followed by the automatic deployment into the Embassy Cloud through our BioExcel portal. Among the expected VMs we plan to deploy tools from the MuG (Multi-scale complex genomics VRE) project that are already being tested at the Embassy cloud as part of the MuG own roadmap. A possible joint MuG-BioExcel workshop using these VMs is being planned for Q2 2018. The inclusion of other VM based deployments coming from the ELIXIR catalog is being studied in collaboration with ELIXIR tools platform.

Having the contextualization correctly configured in our VMs allows us to also deploy them in EGI grid infrastructure. Again, as described in section 2.1.2.1, the work done in collaboration with EGI foundation and ELIXIR Compute Platform will authorize BioExcel partners to deploy and test our VAs registered into the BioExcel Virtual Organization with the same centers subscribed in the ELIXIR VO (EMBL-EBI, IN2P3, GRNET, CESNET). Tests done so far in EGI for BioExcel VMs used the EGI Federated cloud, which is the EGI catch-all allocation for start up communities. Moving to the ELIXIR VO providers is a step forward in the project, embarking on the life science clouds. Work planned for the next months includes the final confirmation of this enrollment, and the corresponding tests & benchmarks of our VAs in the new service provider's machines. The alignment with EGI will empower BioExcel to interact with the forthcoming EOSC (European Open Science Cloud) initiative and adopt a full integration with their uses and standards. As part of our collaboration with ELIXIR, workflows developed here will be made available (as CWL specifications) and registered in the appropriate sites.

3.2.2 HPC & Exascale

The software model designed to be used in BioExcel workflow components consists of a series of building blocks organized as a library of modules, encapsulating the necessary functionalities (see D2.2). These modules are being generated as configurable Python modules wrapping the original software. Interaction with the underlying software will be managed through command line execution, or, when appropriate, through a specific Python API provided by the software. This ensures that the original software can be kept untouched, minimizing installation and configuration issues. Also the underlying software can be upgraded as new releases become available with only need of updating the interface, keeping the external API untouched. Besides, parallelization strategies already available in such applications can also be used when appropriate. In general terms, wrappers will expose tasks and their dependencies, such that the underlying computational infrastructure can optimize their execution. Our task-based strategy for parallelism is commonly used in a number of runtime environments for high-performance computing (for example, see the RADICAL Pilot project[34] or the Extasy project[35]). This particularity, together with the power offered by PyCOMPSs easing the development of parallel applications for HPC infrastructures, should be exploited

through the implemented workflows. The first attempt was done with the *Model protein Mutants* prototype in Marenostrum and Minotauro BSC supercomputers. The maximum number of processors tested in parallel was 192, but the same code could be run using thousands of processors. This is currently being tested in the brand new Marenostrum 4, and will be exported to other supercomputer infrastructures such as ARCHER in the EPCC or Jureca in Jülich, both installed in BioExcel partners premises.

The embarrassingly parallel regime of workflows such as the *Model Protein Mutants*, where thousands of different mutations can be studied at the same time, each of them using hundreds of processors for the MD simulation (see workflow description and diagram in D2.2) makes them perfect candidates for an exascale computing system. Having real use cases of interest in the pharmaceutical field, as the ones presented in the first part of this deliverable, makes this point even more attractive, and BioExcel is going to push forward in this area.

3.2.3 Testing & Benchmarking

A process that has already been initiated with the *Model Protein Mutants* workflow prototype together with Nostrum Biodiscovery Company is the scientific testing of our pipelines. Technological testing is being done in the BSC testbed, and was exhaustively described in the D2.2.

As presented in the first part of this deliverable, enterprises such as Nostrum Biodiscovery or Open PHACTS are helping us identifying real use cases, studies of interest for the pharmaceutical field that can be exploited using our recently developed workflows. The use cases that are being evaluated are the high throughput analysis of EGFR mutations, and their connection to the protein stability and dimerization feasibility (*Model Protein Mutations* workflow*)*, and, using the results of the previous study, investigate the different compounds that could be used to knock down EGFR tyrosine-kinase activity, hitting the wild type sequence, or a mutant variant (*Virtual Screening* workflow). We think that these real use cases are the next step needed to validate our work, and we will work on that from now to the end of the project.

On the other hand, real studies such as the ones presented in the last paragraph are perfect candidates to be used in a computational benchmarking. The mutations example, as commented in the previous section, permits the usage of thousands of processors in parallel, and the possibility to easily install and run our pipeline in different supercomputers makes the implementation of a benchmarking feasible. This kind of benchmarking is useful to learn the optimal number of processors to ask for in such a big computational studies. Benchmark done in BioExcel so far is demonstrating the capability of our workflow prototype to be run in different architectures without the need of specific fine tuning. Infrastructures tested to date are supercomputers like BSC Marenostrum and BSC Minotauro, Virtual Machines in private (OpenNebula) and public (EGI) cloud environments, and workstations. The following table shows preliminary results for the time needed to compute 1ns-length MD simulations for 10

mutations in a small system (PDB 2JQ3, *Human Apolipoprotein C-III*) using PyCOMPSs workflow manager.

Infrastructure	AVG Time (minutes)	Scale-up
MareNostrum **1 core**	381.58	1
MareNostrum **16 cores**	32.57	11.71
MareNostrum **48 cores**	13.02	29.31
MareNostrum **80 cores**	6.60	57.77
MareNostrum **160 cores**	3.43	111.14
MinoTauro (GPUs) - Serial **1 GPU**	68.28	--
Workstation **1 core**	393.92	1
Workstation **8 cores**	101.33	3.88
OpenNebula VM **1 core**	315.00	1
OpenNebula VM **16 cores**	29.75	10.58

Table 1: Model Protein Mutants Workflow Prototype Benchmarking (update).

The table illustrates the capacity of BioExcel workflow prototype to run in different parallel environments (OpenMP or MPI) in different infrastructures. Work in the coming months, as explained in the previous section, will allow us to fill the benchmark with a higher number of processors, different supercomputers, and a real scientific use case.

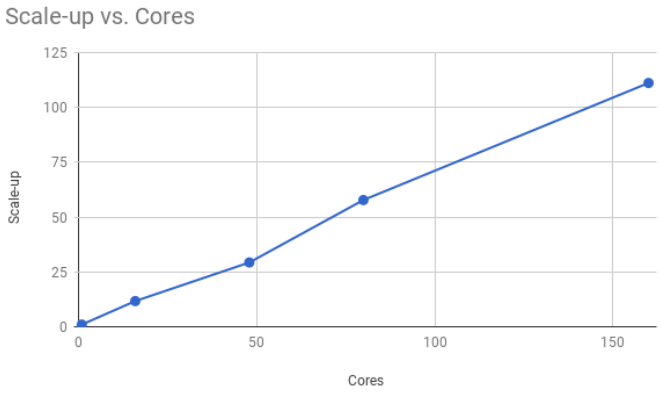

Fig. 8: Model Protein Mutants Workflow Prototype benchmarking speed up (BSC Marenostrum).

4 Conclusions

BioExcel project is starting to obtain feedback from their work in workflows and computational infrastructures. At this point in the project, most of this feedback is coming from internal tests and external collaborations. External user feedback is expected to start to be collected in the coming months, when the different use cases and Virtual Machines will be made available through the BioExcel cloud portal.

The work done with the transversal workflow unit *Model Protein Mutants* has triggered useful collaborations, mainly with ELIXIR project (CWL, tools and interoperability platforms, bio.tools) and Nostrum Biodiscovery. Expertise gained from these associations will be applied in the coming months to update the workflow and to use it in real scientific cases.

Different type of feedback has been gathered from the project pilot use cases, depending on the current workflow phase and collaborators. It is clear that all the use cases will benefit from this collected information, regardless of the provenance. Updates, branches, and even remodeling of the workflow pipeline have derived from them, and are clearly guiding us towards a better final product.

The future roadmap for the BioExcel Cloud Portal, workflows and computational infrastructures for the third year of the project is already defined. The Cloud Portal is up and running and is evolving to include more functionalities such as data volumes and usage of shared cloud resources. The next months will be mainly focused on increasing the user experience and usability.

A final protocol to generate contextualized Virtual Machines able to be easily deployed in different cloud environments has been defined, integrating all expertise learned during the first phase of the project. Packaging of different tools and workflows in VMs and the following register and upload to the described platforms to make them available through the Cloud Portal are the determined next steps.

The workflow development process presented in the project allows the generated workflows to run in completely different computational infrastructures. A technical benchmark is being produced, using the same workflow in VMs, workstations and supercomputers. The last part of the project will be focused in exploiting the ability of our workflow to be run in an exascale regime. For that, the *Model Protein Mutants* transversal workflow is going to be used to generate a scientific benchmark. A real scientific case (EGFR), which will serve as a definitive proof of concept for this workflow, has been chosen and will be studied using thousands of processors in parallel.

5 References

1. Crosswell, L.C. and J.M. Thornton, *ELIXIR: a distributed infrastructure for European biological data.* Trends in Biotechnology. **30**(5): p. 241-242.
2. Ison, J., et al., *Tools and data services registry: a community effort to document bioinformatics resources.* Nucleic Acids Research, 2016. **44**(D1): p. D38-D47.
3. Peter, A., et al., *Common Workflow Language, v1.0*. 2016, Figshare.
4. Ison, J., et al., *EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats.* Bioinformatics, 2013. **29**(10): p. 1325-32.
5. Bechhofer, S., et al., *Why linked data is not enough for scientists.* Future Generation Computer Systems, 2013. **29**(2): p. 599-611.
6. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* 2016. **3**: p. 160018.
7. Tejedor, E., et al., *PyCOMPSs: Parallel computational workflows in Python.* International Journal of High Performance Computing Applications, 2015.
8. Vivian, J., et al., *Toil enables reproducible, open source, big biomedical data analyses.* Nat Biotech, 2017. **35**(4): p. 314-316.
9. Afgan, E., et al., *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.* Nucleic Acids Research, 2016. **44**(W1): p. W3-W10.
10. Rubén, S.M., M.-V. Rafael, and M.L. Ignacio, *IaaS Cloud Architecture: From Virtualized Datacenters to Federated Cloud Infrastructures.* Computer, 2012. **45**: p. 65.
11. Jackson, K., *OpenStack Cloud Computing Cookbook*. 2012: Packt Publishing. 318.
12. Kranzlmüller, D., J.M. de Lucas, and P. Öster, *The European Grid Initiative (EGI)*, in *Remote Instrumentation and Virtual Laboratories: Service Architecture and Networking*, F. Davoli, et al., Editors. 2010, Springer US: Boston, MA. p. 61-66.
13. Littlefield, P., et al., *Structural analysis of the EGFR/HER3 heterodimer reveals the molecular basis for activating HER3 mutations.* Science signaling, 2014. **7**(354): p. ra114-ra114.
14. Zhang, N., et al., *HER3/ErbB3, an emerging cancer therapeutic target.* Acta Biochimica et Biophysica Sinica, 2016. **48**(1): p. 39-48.
15. Papanastasiou, A.D., et al., *RANK and EGFR in invasive breast carcinoma.* Cancer Genetics. **216**: p. 61-66.
16. van Vugt, V.A., et al., *Neurological improvement of perineural and leptomeningeal spread of squamous cell carcinoma treated with intrathecal chemotherapy and systemic EGFR inhibition.* CNS Oncology, 2017.
17. Francis, J.M., et al., *EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing.* Cancer discovery, 2014. **4**(8): p. 956-971.
18. Liu, F., et al., *EGFR Mutation Promotes Glioblastoma Through Epigenome and Transcription Factor Network Remodeling.* Molecular cell, 2015. **60**(2): p. 307-318.
19. Joseph, S., et al., *Dysregulation of Epidermal Growth Factor Receptor in Actinic Keratosis and Squamous Cell Carcinoma*. Vol. 46. 2015. 20-7.

20. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis, 2011.
21. Guimera, R.V., *bcbio-nextgen: Automated, distributed next-gen sequencing pipeline.* EMBnet.journal; Vol 17: Supplement B, 2012.
22. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. 2013: q-bio.GN.
23. Faust, G.G. and I.M. Hall, *SAMBLASTER: fast duplicate marking and structural variant read extraction.* Bioinformatics, 2014. **30**(17): p. 2503-2505.
24. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
25. McKenna, A., et al., *The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Research, 2010. **20**(9): p. 1297-1303.
26. Decap, D., et al., *Halvade: scalable sequence analysis with MapReduce.* Bioinformatics, 2015. **31**(15): p. 2482-2488.
27. van Zundert, G.C.P., et al., *The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes.* Journal of Molecular Biology, 2016. **428**(4): p. 720-725.
28. Abraham, M.J., et al., *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* SoftwareX, 2015. **1–2**: p. 19-25.
29. Korb, O., T. Stützle, and T.E. Exner, *Empirical Scoring Functions for Advanced Protein–Ligand Docking with PLANTS.* Journal of Chemical Information and Modeling, 2009. **49**(1): p. 84-96.
30. Fenollosa, C., et al., *SEABED: Small molEcule activity scanner weB servicE baseD.* Bioinformatics, 2015. **31**(5): p. 773-5.
31. Gabay, M.P., et al., *Oral Targeted Therapies and Central Nervous System (CNS) Metastases.* CNS Drugs, 2015. **29**(11): p. 935-952.
32. Yang, Z., et al., *Comparison of gefitinib, erlotinib and afatinib in non-small cell lung cancer: A meta-analysis.* International Journal of Cancer, 2017. **140**(12): p. 2805-2819.
33. Wang, S., S. Cang, and D. Liu, *Third-generation inhibitors targeting EGFR T790M mutation in advanced non-small cell lung cancer.* Journal of Hematology & Oncology, 2016. **9**: p. 34.
34. Merzky, A., et al., *RADICAL-Pilot: Scalable Execution of Heterogeneous and Dynamic Workloads on Supercomputers.* CoRR, 2015. **abs/1512.08194**.
35. Balasubramanian, V., et al., *ExTASY: Scalable and Flexible Coupling of MD Simulations and Advanced Sampling Techniques*. Vol. abs/1606.00093\. 2016.

Appendix: *Cloud usage needs survey*

A.1 Questionnaire

The following questions are part of the survey:

- What is your background?
- Did your data grow significantly in the last 1-2 years?
- Did your computational needs grow significantly in the last 1-2 years?
- Are you considering cloud computing to solve scalability issues?
- Are you considering cloud computing to solve reproducibility issues?
- How important is solving scalability problems for your work?
- How important is reproducibility for your work?
- How are you planning to use the cloud?

A.2 Data and compute growth needs

The following figure shows the distribution of responses for the question *Did your data grow during the last year?* A majority of the respondents answered yes.

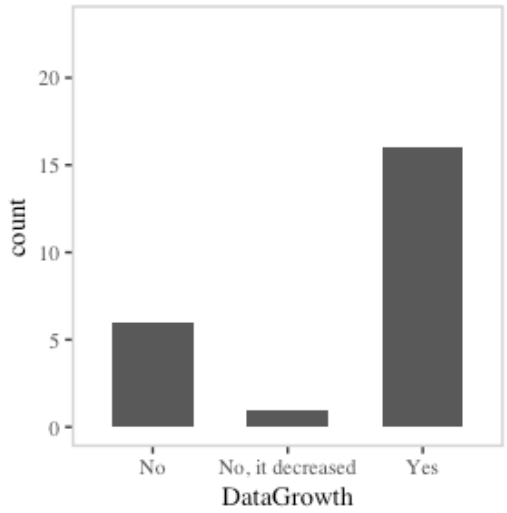

Fig. A1: Cloud usage needs survey: Did your data grow during the last year?

We can split that by *background* which reveals that, among developers, the perception of data growth is not that strong. It seems that the stronger perception towards *yes* comes from bioinformaticians.

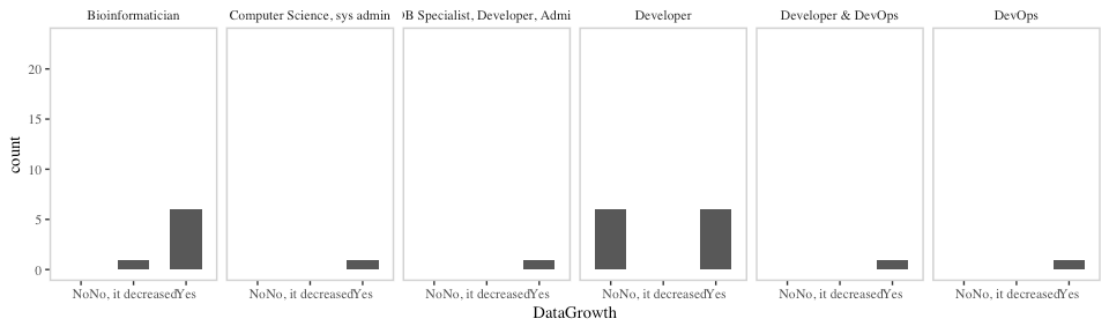

Fig. A2: Cloud usage needs survey: Did your data grow during the last year (by fields)?

Similar responses are found for the question *Did your computing needs grow during the last year?*

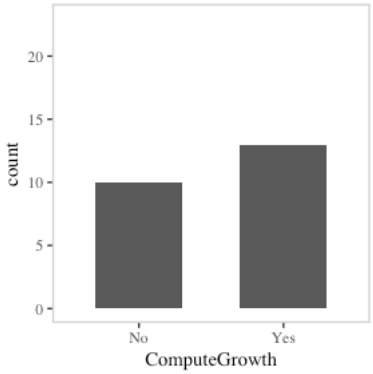

Fig. A3: Cloud usage needs survey: Did your computing needs grow during the last year?

We can repeat the previous split. Something interesting here is that *bioinformaticians* might not perceive compute needs that increased as data needs, while *developers* do more than in the case of data growth.

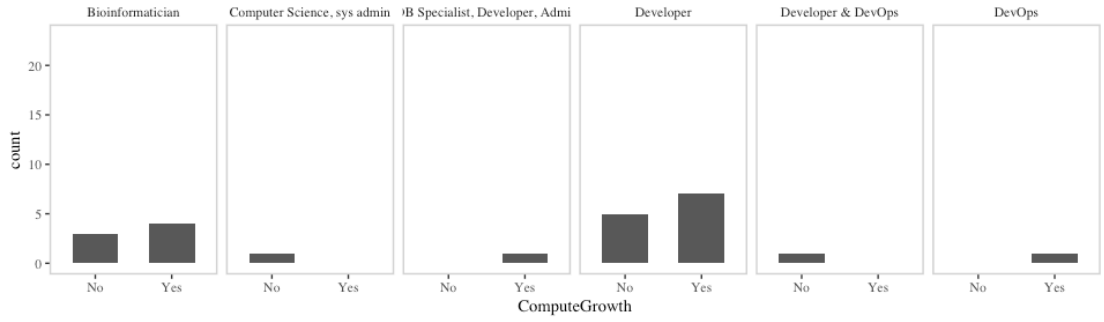

Fig. A4: Cloud usage needs survey: Did your computing needs grow during the last year (by field)?

Definitely we need a bigger sample size, but this could mean that when providing scalability solutions for data storage, we might benefit more from asking *bioinformaticians* about the nature of the data problem while talk to *developers* about computation needs. The underlying problem is probably similar for both, but the perception and insight can change from one group to another.

A.3 Scalability and cloud

An important question in the survey refers to the preference for using the cloud to solve scalability problems.

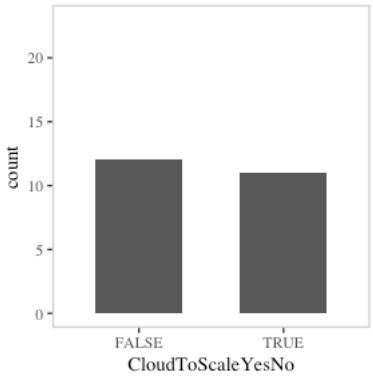

Fig. A5: Cloud usage needs survey: Are you considering cloud computing to solve scalability issues?

With this samples size half of the responses are interested in using it. But there is another question there about the perceived importance of scalability to the respondents job.

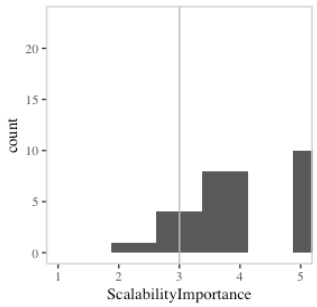

Fig. A6: Cloud usage needs survey: How important is solving scalability problems for your work?

It seems that more responses agree that scalability is important for their job. However this represent a dichotomy with the fact that just half of the people considers using the cloud to solve scalability problems. A reason for this can be that respondents do not consider the cloud as solution for scalability, either because they don't understand the technology or know how to use it (something that can be solved with training) or because they don't consider it an appropriate one for their scalability problems (something to research indeed).

These two last charts can be worth it to explore by respondent background, as a way to identify who to talk in order to train or to research new solutions for scalability problems.

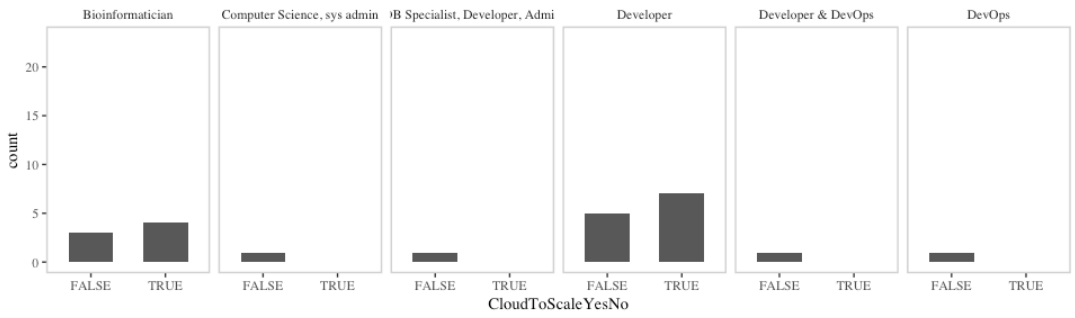

Fig. A7: Cloud usage needs survey: Are you considering cloud computing to solve scalability issues (by field)?

Fig. A8: Cloud usage needs survey: How important is solving scalability problems for your work (by field)?

Again, we need a bigger sample, but it looks like that developers show less dichotomy and are more consistent with the assumption than the cloud will be perceived as a solution to scalability problems. Is this a sign of them understanding the technology better? Or is it that they don't understand the scalability problem properly?

A.4 Reproducibility and cloud

A similar analysis can be performed for reproducibility problems and cloud usage.

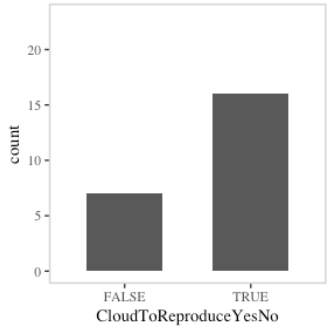

Fig. A9: Cloud usage needs survey: Are you considering cloud computing to solve reproducibility issues?

If we compare these results with the inclination to use the cloud for scalability issues, there is a clearer tendency to use the cloud to improve reproducibility. Let's have a look at how important reproducibility is for our respondents.

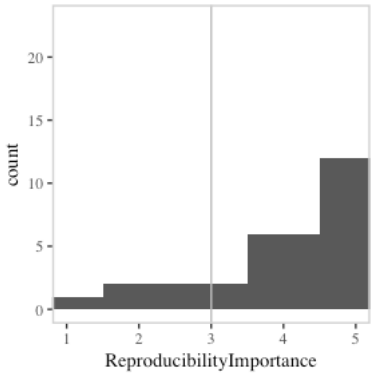

Fig. A10: Cloud usage needs survey: How important is reproducibility for your work?

In this case the view is quite consistent. Most respondents consider reproducibility important and consider the cloud a good way to solve their problems.

A.5 Preferences to get into the cloud

One of the last questions in the survey tries to understand preferences in ways to do cloud computing.

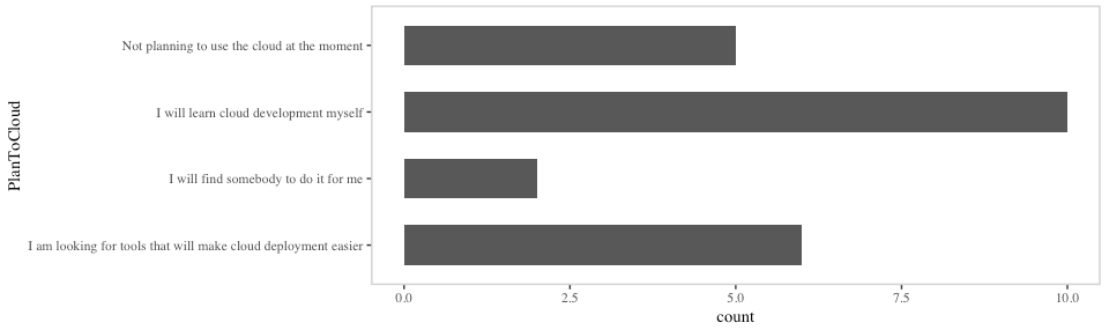

Fig. A11: Cloud usage needs survey: How are you planning to use the cloud?

Most users either want to do cloud themselves after getting the training (specially developers, see Fig. A12) or they consider tools shown during the training to assist them. Not many respondents will look for help from somebody with cloud knowledge, although we should not take these conclusions firmly due to the small and specific sample queried.

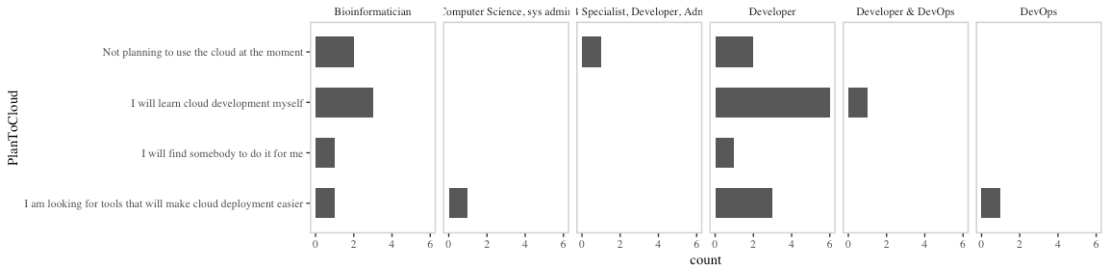

Fig. A12: Cloud usage needs survey: How are you planning to use the cloud? By background