Advances in Automation and Efficiency for the Exascale Era – Experiences from the Biomolecular Sciences
PASC Minisymposia, Basel, 3/07/2018

Workflow Automation and Efficiency for Macromolecular Simulations and Screening

Adam Hospital Gasch, IRB Barcelona, Spain

adam.hospital@irbbarcelona.org
• Design, deploy and make available solution-oriented biomolecular workflows.

• Excellence in **Usability**.

  • Ease of use
  • Availability
  • Reproducibility
  • Multi-platform
  • Multi-infrastructure
Meyer et al, Structure 2010, 18 (11), 1399-1409.

- ~2000 MD Simulations
- Amber, Namd, Gromacs
- Charmm22, Charmm27, Parm99, Parm99SB, Parm03, OPLS, etc.
- Web interface connected to a Database
- Automatic MD Setup, Run & Analysis

A. Hospital et al, Bioinformatics 2012, 28 (9), 1278-1279.
• **Ease of use:** Taverna
• **Availability:** Web Services
• **Interoperability:** Ontology

• BioMoby discontinued
• Multi-platform
• Multi-infrastructure
A distributed infrastructure for life-science information

ELIXIR Europe

Platforms

ELIXIR’s activities are divided into five areas called ‘Platforms’. These are Data, Tools, Interoperability, Compute and Training. The Platforms are managed by Platform leaders and the work is carried out by groups within the Platforms.

- **Data Platform**
  - Aims to identify key data resources across Europe and support the linkages between data and literature.

- **Tools Platform**
  - Helps researchers find the best software tools to analyse their data.

- **Compute Platform**
  - Develops services to make it easier to store, share and analyse large datasets.

- **Interoperability Platform**
  - Develops and encourages the adoption of standards to describe life science data.

- **Training Platform**
  - Helps scientists and developers find the training they need, and also provides that training.

Workflow and Tool interoperability

**Goal:** To create specifications that enable data scientists to describe analysis tools and workflows to make them interoperable across Use Cases.

**Leader:** Carole Goble (ELIXIR UK)

**Activities:**
- Enabling the reuse, extension, scaling, and reproducibility of scientific workflows, an Implementation Study working with the Marine Metagenomics Use Case on using the Common Workflow Language (CWL).
- Promoting industry standards for API documentation.
- Engaged with the Common Workflow Language (CWL) community group and BioExcel Centre of Excellence for BioMolecular Simulation.
• Push for **best practices** in **biomolecular workflows** development:

  • Producing **easy to use, reproducible, multi-platform** workflows
  • Following **FAIR principles** applied to software, together with **ELIXIR**

  • **Findable**
    • Registries, Metadata, Ontology.
  
  • **Accessible**
    • Infrastructure-agnostic, Workflow manager-agnostic.
  
  • **Interoperable**
    • Software architecture based on wrappers.
  
  • **Reusable**
    • CWL descriptions, registry, containerization.
Findability: Automatic discovery and retrieval of tools

PyMDSsetup version 0.2 Alpha

https://bio.tools/api/tool/?input=structure
Interoperability: Building blocks compatibility
• Common Workflow Language (CWL)

• Share & reuse workflows

• Make them portable and scalable across a variety of software and hardware environments

• Containerization
Accessibility: Repositories, Documentation, Installation

[GitHub logo]

https://github.com/bioexcel/biobb*

Accessibility
Availability
Usability

biobb_ist
- Jupyter Notebook
- Apache-2.0
- Updated 27 days ago

biobb_md
- Python
- Apache-2.0
- Updated 27 days ago

biobb_common
- Python
- Apache-2.0
- Updated 27 days ago

Pydinstall
latest

Introduction & Installation

API Documentation

- command_wrapper package
- configuration package
- ebi_api package
- gnuplot_wrapper package
- gromacs_extra package

gromacs_wrapper package

Submodules
- gromacs_wrapper.editconf module
- gromacs_wrapper.grompp module
- gromacs_wrapper.mdrun module
- gromacs_wrapper.pdb2gmx module
- gromacs_wrapper.rms module
- gromacs_wrapper.solvate module

Read the Docs

Create MDP

Launch

Read the Docs

gromacs_wrapper.grompp module

Python wrapper for the GROMACS grompp module

```
class gromacs_wrapper.grompp.Grompp(input_pro_path, input_top_zip_path, output_tpr_path, properties, input_cpt_path=None, **kwargs)

Parameters:

- input_gro_path (str) – Path to the input GROMACS structure GRO file.
- input_top_zip_path (str) – Path to the input GROMACS topology TOP and ITP files in zip format.
- output_tpr_path (str) – Path to the output portable binary run file TPR.
- input_cpt_path (str) – Path to the input GROMACS checkpoint file CPT.
- properties (dict) – mdp (str): MDP options specification.
```

Bases:

- object

Wrapper for the 5.1.2 version of the GROMACS grompp module. The GROMACS preprocessor module needs to be fed with the input system and the dynamics parameters to create a portable binary run input file TPR. The dynamics parameters are specified in the mdp section of the configuration YAML file. The parameter names and defaults are the same as the ones in the official MDP specification: [http://manual.gromacs.org/current/online/mdp_opt.html](http://manual.gromacs.org/current/online/mdp_opt.html)
Accessibility: Repositories, Documentation, Installation

https://github.com/bioexcel/biobb*

Accessibility
Availability
Usability
Accessibility: Repositories, Documentation, Installation
Benchmark Test

Infrastructures

Cloud

- OpenNebula
- OpenStack

HPC

Workflow managers

- Toil
- Galaxy
- jupyter
- pyCOMPSs
MD Bio Building Blocks

Workflow: Gromacs Full MD setup

Building Blocks: Gromacs Python wrapper
https://github.com/bioexcel/pymdsetup

Parallelization: PyCOMPSs

```
81  mutations_counter = 0
82  for mut in mutations:
83      if mutations_counter == mutations_limit: break
84      mutations_counter += 1
85      mut = mut if not mut.startswith('*') else mut.replace('*', 'ALL')
86      paths = conf.get_paths_dic(mut)
87      prop = conf.get_prop_dic(mut)
```
Use Case: Massive Protein Mutation

COMPSs

Input ID: @protein_name
Sequence Analysis
- Recover Protein Sequence
  - Ensemble
- Recover Protein Isoforms
  - Ensemble
  - Isoform 1
  - Isoform 2
- Recover Protein Variants
  - Ensemble
  - Variant 1
  - Variant 2
- Structure Analysis
  - Recover Protein Structure
    - PDB Mapping
      - Evalue4 - PDB Map
      - If PDB
        - Recover PDB/6
      - No
        - Model Protein Structure/Modeler
          - PDB Structures
- Structure Mutants
  - Model Protein Structures
    - Mutants
  - Protein
- Molecular Dynamics Setup
  - MDWEB / Grandala
- Molecular Dynamics Run
  - Grandala
- Molecular Dynamics Analysis
  - Grandala / MDWEB / MD Analysis / Reslve
- Output
  - Analysis

POD Code
- NMBSPOD_API
- NMBS LINPROT_API

WT PDB
- Mutations List

SCWRL 4
- Model Mutation
  - GMX PDB2GAM
- GMX EDITCONF
- GMX SOLVE
- Define Box

SCWRL 4
- Model Mutation
  - GMX PDB2GAM
- GMX EDITCONF
- GMX SOLVE
- Define Box

SCWRL 4
- Model Mutation
  - GMX PDB2GAM
- GMX EDITCONF
- GMX SOLVE
- Define Box

https://github.com/bioexcel/pymdsetup
Use Case: Massive Protein Mutation

- Pyruvate Kinase
- 4 x 500 residues (~395,000 atoms MD)
- Selected mutations from literature (200 annotated mutants from Uniprot, manually curated)
- 5ns MD-length

200 * 48 procs (1 node) 9600 procs

http://www.cell.com/biophysj/fulltext/S0006-3495(15)00606-2

https://github.com/bioexcel/pymdsetup/blob/master/workflows/pyruvateKinase_MN.py
Pyruvate Kinase

4 x 500 residues

Selected mutations from literature (200 annotated mutants from uniprot, manually curated)

5 ns MD-length

Use Case: Massive Protein Mutation

http://www.cell.com/biophysj/fulltext/S0006-3495(15)00606-2
https://github.com/bioexcel/pymdsetup/blob/master/workflows/pyruvateKinase_MN.py
Epidermal Growth Factor Receptor (EGFR) - (Kinase Domain)

EGFR mutations drive some types of cancers, like carcinoma and glioblastoma.

327 residues

Selected mutations from literature:
- T790M (gatekeeper) confers resistance to Erlotinib and Gefitinib by increasing ATP binding.
- L718Q, L844V kill Rociletinib
- M766T, L858R, L718A, T854A

http://pdb101.rcsb.org/motm/126

https://github.com/bioexcel/pymdsetup/blob/master/workflows/egfr_md.py
Use Case: EGFR (VS)

Virtual Screening Pipeline

Retrieval
- Compounds
- Decoys
- Receptors

Structures
- Structure Modeling
- Molecular Dynamics (flexibility)

Recognition
- Biomolecular docking
- Scoring & Analysis

http://bioexcel.eu/research/pilot-use-cases/use-case-virtual-screening/
• Structural Ensemble generation (using MD workflow)

• Structure clustering (gromos, Daura et al. 1999)

• Active Site identification (Fpocket)

• Compounds/Decoys retrieval (openPHACTS, DUD)

• Docking (Autodock Vina)

https://bioexcel.eu/webinar-openphacts/
https://github.com/bioexcel/virtualscreening
• Fully interoperable software library to build biomolecular workflows

• Developed following software best practices (ELIXIR project)

• Specified and documented using CWL

• Multi-platform and workflow manager-agnostic

• Being tested with HPC scientific studies