Codefest *Report*

July 20-21, 2017

www.open-bio.org/wiki/Codefest_2017

What is Codefest?

- 2 day BOSC collaborative work session
- 60+ community members
- Learning and training
- Building relationships
- Writing code
- Everyone is welcome
- 8th successful year

How does Codefest work?

- Free
- You get wireless, power, space, coffee, food
- Open source collaborators
- Self organize around projects of interest
- Produce useful code and motivation
- Report on accomplishments

Thank you

- brmlab (https://brmlab.cz/)
- Matúš Kalaš, Heather Wiencko
- Repositive and Seven Bridges
- Institute of Organic Chemistry and Biochemistry
- OpenBio and BOSC Community

Themes from the Codefest

- New contributors
- Autonomy
- Last mile development
- Standards and coordination
- Fixing long standing and neglected bugs

MultiQC

Pull requests	18 opened	14 merged
Issues	9 opened	5 closed

New modules!
VCFtools, nonpareil, bcl2fastq(!), AfterQC

Table column ordering
Specify where columns should be placed in In modules, config and report

Module help texts
New drop-down texts above plots in reports to describe what's being shown

Scout integration
MultiQC reports embedded within Scout clinical genomics browser

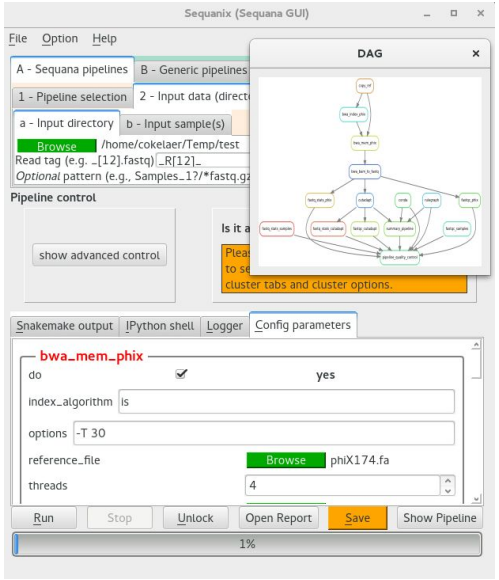
Collect software versions
Core MultiQC support for scraping software versions from logs

Module grouping
Just run modules related to a specific data type with new module tags.

Phil Ewels, Rickard Hammarén, Robin Andeer, Tim Booth, Dennis Schwartz, Dimitri Desvillechabrol, Amandine Perrin, Rowland Mosbergen, Murray Wham, Markus Ankenbrand, Raony Guimaraes, Tom Walsh

Snakemake, and "other Python things"

- Bug fixing (4 PRs merged, 3 in progress)
- Biopython architecture discussions
- Python 2 retirement by 2020
- Introduction to Snakemake + Sequanix



Seqanix GUI in PyQt for Snakemake pipelines; http://sequana.readthedocs.io

Members: Wibowo Arindrarto, Kai Blin, Spencer Bliven, Christian Brueffer, Peter Cock, Thomas Cokelaer, Joe Greener,

Protein structure

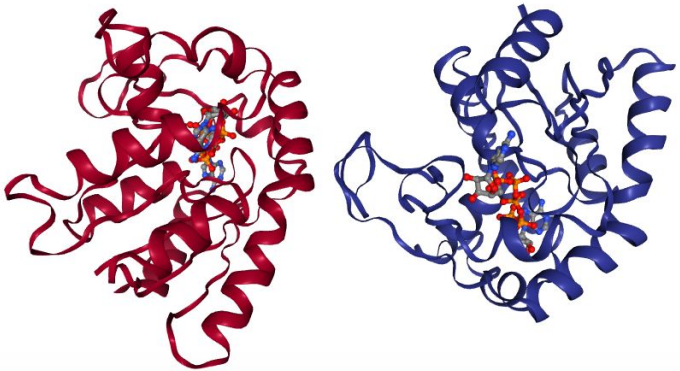
1. Integrate **Biopython** structural entities with **nglview** to allow interactive visualisation of protein structures in **Jupyter** notebooks

```
In [11]: import nglview as nv
         from Bio.PDB import PDBParser, MMCIFParser

In [12]: # Read in a PDB file with Biopython
         parser = PDBParser()
         structure = parser.get_structure("1AKE", "1AKE.pdb")

In [13]: view1 = nv.show_biopython(structure)
         view1
```

Atom: [GLU]204:B.CA



Members: Spencer Bliven, Alexander Rose, David Sehnal, Joe Greener

Protein structure

https://github.com/MolQL/molql

2. Molecular Query Language
   - Formal specification for general selection languages
   - Interchange Format to interconvert between existing query languages

User query

PyMol: `select chain A within 5 of resn HEM`

Jmol: `select within(5, [HEM]:A)`

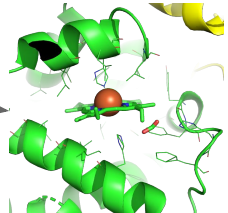
Abstract Syntax Tree

```
isCloseTo(
    chains('A'),
    residues('HEM'),
    5)
```

MolQL Interchange Format

```
{ kind:"isCloseTo",
    args: [
        { kind: "chains",
          options: ["A"] },
        { kind: "residues",
          options: ["HEM"]
    }
    ],
    options: { maxDistance:
    5 }
}
```

Query Result



nextflow

Exploring new uses and executor APIs

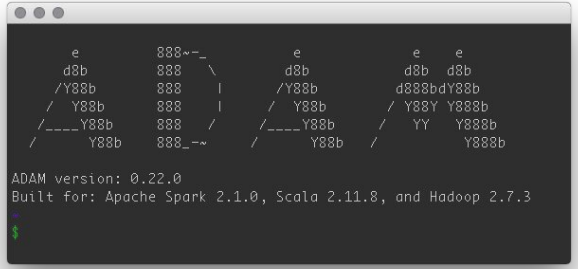
- Tested the experimental Kubernetes support in Nextflow
- Determined the issues in setting up a correctly configured Kubernetes cluster for use with a workflow
- Explored Google Cloud usage
- Discussed the incorporation of the Global Alliance for Genomics and Health (GA4GH) as a Nextflow executor
- Built the GA4GH API in Java using the protocol buffer definitions.
- Began exploring the API
- Group members:
  - Paolo Di Tommaso
  - Konstantinos Krampis
  - Kevin Sayers

kubernetes

Global Alliance for Genomics & Health

Workflows including Apache Spark based analyses

- Discussed supporting provisioning of Apache Spark clusters vs. delegating via environment profiles in CWL
- Bioconda recipe for ADAM, https://github.com/bigdatagenomics/adam
- Support for ADAM and Avocado (variant caller) in bcbio-nextgen in progress
- Demonstrated Apache Spark runtime configuration via profiles in Nextflow
- Group members: Michael Heuer, Brad Chapman, Roman Valls Guimerà, Paolo Di Tommaso



Reproducible software deployment

- Part of creating reproducible solutions is creating a work flow that uses persistent software resources
- Major packaging efforts are in Debian, GNU Guix and Bioconda
- Challenges in versioning, dependencies, reproducibility
- Solution:
  - Container (Docker) -> GNU Guix -> Guix packages -> BioConda -> Conda packages
  - Docker runs on CWL and Galaxy
- Problem solved. OK. Maybe. Work in progress…
- CWL and Galaxy projects are very interested
- Group members: Pjotr Prins, Steffen Möller

COMMON WORKFLOW LANGUAGE

Provenance

researchobject.org

JSON-LD

W3C® PROV

- `cwltool --provenance`
  generate **research object** (RO) w/ provenance of workflow execution
- **Model** of CWL RO structure → *BagIt archive*
  - Level 0: workflow job submission
  - **Level 1**: executing master workflow, in/out
  - Level 2: step execution
  - Level 3: nested workflows
- Capturing **input** data using *content-based addressing*
- **Rerunnable** and **portable** *master-job.json*
- Provenance as PROV JSON-LD

Farah Zaib Khan, Stian Soiland-Reyes, Tazro Inutano Ohta


```
{"@context": { "@base":
"app://2e1287e0-6dfb-11e7-8acf-0242ac11000/" },
"@id": "workflow/master-job.json#",
"@type": "WorkflowRun",
"workflow": "workflow/packed.cwl#main",
"inputs": [
  {"@id":
"data/5891b5b522d5df086d0ff0b110fbd9d21bb4fc7163af34d08286a2e
846f6be03",
   "describedByParameter": "workflow/packed.cwl#main/in1"} ],
"outputs": [
   {"@id":
"data/00688350913f2f292943a274b57019d58889eda272370af261c84e7
8e204743c",
     "describedByParameter": "workflow/packed.cwl#main/in1" }
],
"steps": [
   {"@id": "urn:uuid:4305467e-6dfb-11e7-885d-0242ac110002",
    "@type": "ProcessRun",
    "step": "workflow/packed.cwl#main/step1"},
   {"@id": "urn:uuid:c42dc36e-6dfd-11e7-bc24-0242ac110002",
    "@type": "ProcessRun",
    "step": "workflow/packed.cwl#main/step2"}
}
```


https://github.com/common-workflow-language/common-workflow-language/wiki/Research-Object-Proposal

Singularity support in CWL

github.com/johnfonner/cwltool/tree/feature-singularity

- CWL workflows using “dockerPull” can transparently use Singularity for container execution.
- 19 commits, ~150 lines of code

+

=

Members: Isak Sylvin, John Fonner

Rabix Suite

- https://github.com/rabix
- Fixing issues and creating a beta release of Rabix Composer
- Integrating Executor into Composer (prototype)
- Syncing Rabix Executor with CWL v1.0.1 errata and releasing v1.0
- Group members: Janko Simonović, Siniša Ivković, Luka Stojanović, Ivan Batić, Maja Nedeljković

Rabix

SevenBridges

CWLToil dynamic ResourceReqs

First Goal: Calculate resource requirements based on input files: number, sizes, and other metadata.

https://github.com/BD2KGenomics/toil/pull/1767

https://github.com/common-workflow-language/cwltool/issues/483

Final Goal: Calculate (computational || economic) costs **before** running a job on Toil/CWL, based on cores, input file sizes, memory, etc...

Michael Crusoe and Roman Valls Guimera

CWL SDKs

- **Create automatically from the specification of CWL some SDKs to handle the reading, manipulation, and writing of CWL files**

- Multiple “generic” approaches unsuccessful.
- Python SDK generation (direct from CWL spec) project started:
  - https://github.com/common-workflow-language/python-cwlmodel
- Ruby project started (using JSON schema):
  - https://github.com/njall/rubycwl
- Java:
  - https://github.com/StarvingMarvin/cwl-sdk
- *Pre-existing* TypeScript implementation from SBG:
  - https://github.com/rabix/cwl-ts

Members: Niall Beard, Kenzo Hillion, Hervé Ménager, Anton Khodak, Denis Yuen, Luka Stojanovic, Heather Wiencko with help from Ivan Batić, Maja Nedeljković, Michael Crusoe and Peter Amstutz

The Open Bioinformatics Community

Photo by Ntino Krampis: https://www.open-bio.org/wiki/Codefest_2017#Outcomes

Join us

- Welcome to BOSC
  - Birds of a Feather meet up
  - Lunch time today
- Come to Codefest next year
  - Training, conference, then Codefest
  - Multiple tracks and groups
  - Everyone is welcome

https://www.open-bio.org/wiki/Codefest_2017