Management of Data, Models, Analyses and Code for Reproducible Research: Present and Future

Matthias König

LiSyM Systems Medicine of the Liver
Humboldt University Berlin, Institute for Theoretical Biology
livermetabolism.com
Computational Modeling Liver

König M., Marchesini G., Vilstrup H., and Holzhütter HG. 
*A Multiscale Computational Model Predicts Human Liver Function From Single-Cell Metabolism* 
2016, [manuscript in preparation]  
https://github.com/matthiaskoenig/multiscale-galactose
Abshagen*, König*, et al. (* equal contribution)
Pathobiochemical signatures of cholestatic liver disease in bile duct ligated mice
2015
https://github.com/matthiaskoenig/bdl-analysis
cy3sbml: A Cytoscape app for SBML
2016, manuscript in preparation
https://github.com/matthiaskoenig/cy3sbml

König, Rodriguez, and Dräger
cy3sabiork: A Cytoscape app for visualizing kinetic data from SABIO-RK, 2016 [version 1; referees: 2 approved with reservations]
http://f1000research.com/articles/5-1736/v1
https://github.com/matthiaskoenig/cy3sabiork

Somogyi, Bouteiller, Glazier, König, Medley, Swat, Sauro.
libRoadRunner: a high performance SBML simulation and analysis library.
Bioinformatics. 2015
https://github.com/sys-bio/roadrunner
What data do you have?

- **Computational modeling**
  - **Code** (model creation, workflows, simulations, data processing, ...)
  - **Data** (parameters, time courses, ...)
  - **Models** (ODE, regression, classification)
  - **Literature & Documentation**
  - **Images**

- **Bioinformatics analyses**
  - **Code** (workflows)
  - **Data** (parameters, timecourses, ...)
  - **Models**
  - **Documentation**
  - **Images**

- **Software development**
  - **Code**
  - **Data**
  - **Documentation**
Present
The Present: Repositories

GitHub is a code hosting platform for version control and collaboration. It lets you and others work together on projects from anywhere.

- Version control
  - Diffs & Branches
- Collaborative editing
  - Pull requests
- Continuous integration
  - Unit tests
  - Commit hooks
- Releases & snapshots (citable code)
- Issue tracker
- Work anywhere & offline

https://github.com/matthiaskoenig/cy3sbml/
https://github.com/matthiaskoenig/cy3sbml/commit/986ebbdd5e77d403cf031ff0a5230de4d28fc284
“Steps must be taken to **preserve and cite** software in a sustainable, identifiable and simple way.”

“Best practice for citing a digital resource like code is to refer to a **digital object identifier (DOI)** for it whenever possible. This is because DOIs are persistent identifiers that can be obtained only by an agency that commits to the **obligation to maintain a reliable level of consistency in and preservation of the resource.**”

- domain agnostic, free, open-access research repository, Zenodo **targets the needs of the 'long tail' of research results**
- CERN Data Centre
- special commitment to sharing, citing and preserving data and code

**Zenodo**

https://github.com/matthiaskoenig/libsbgn-python

https://zenodo.org/record/17197

https://www.openaire.eu/intro-researchers
The Present: Open source, open access, open data, open review
Dynamic report generation!

**Jupyter notebooks**

- Web application that allows to create and share documents that contain live code, equations, visualizations and explanatory text.
- Open source, interactive data science and scientific computing across over 40 programming languages.
- Complete analysis, examples, test cases

**R: Knitr**

- Elegant, flexible and fast dynamic report generation with R
- BDL example
  - [https://github.com/matthiaskoenig/bdl-analysis](https://github.com/matthiaskoenig/bdl-analysis)
The Present: Standard Formats

- Encoding of information in standard formats where possible
  - Models in SBML
  - Simulations in SED-ML
- Annotations to ontologies
  - Knowledge integration
  - Documentation
- Reproducibility
  - Reproducibility of results (roadrunner, COPASI, JWS)
- Model quality
  - Passes library tests (e.g. unit tests, ...)
- Visualization
  - Iterative model cycle

König, Rodriguez, and Dräger
cy3sbml: A Cytoscape app for SBML 2016, manuscript in preparation
https://github.com/matthiaskoenig/cy3sbml
The Present: Archive Formats

Packaging of selected content with metadata

- Model, simulation, data, results (COMBINE archive)
- Investigation, Study, Assays, Models (ResearchObjects)

Enabling reproducible, transparent research.

http://www.researchobject.org/

http://co.mbine.org/documents/archive
The Present: Fairdom & SEEK

- Research projects are represented in SEEK and Fairdom
  - ISA (investigation, study, assay, model, publication)
- Online representation of a ResearchObject
- Orthogonal to repository content
  - Intention, idea, project overview
- Overlap to repository content
  - Model, datasets (assay), descriptions

https://www.fairdomhub.org/people/678/items
https://www.fairdomhub.org/investigations/96
Future
The Future: SEEK repository integration

- **Lossless import & export of ResearchObjects in SEEK**
  - Standard exchange format
  - Different software

- **Create ResearchObjects from repository**
  - Subset of repository is ResearchObject
  - Sufficient to have data files, metadata, repository definition file (which assets in repository)
  - NO upload of every single asset
  - Only single place of change (no managing in repository & SEEK)
  - Changes are version controlled
  - Collaborative update of files & metadata
  - Branches!
  - Continuous integration
    - Tests can be run on repository, publishing as part of
  - No overhead: Integration in existing workflow

- **Transfer between SEEK instances**
  - Registering multiple SEEK instances to one repository

- **ArchiveIndex**
  - DataFiles
  - Metadata

Under version control

- On release
- On commit
- Within CI

Create RO

- Export
- Import
- Update
The Future: Metadata for Figures & Datasets

Nightmare of digitized datasets

- 0% of data available from literature
- Digitize experimental data from publications
  - publication (PDF) → images (PNG) → Spreadsheets (XLSX) → CSV → modeling pipeline
- Need for annotation, metadata generation and dissemination
  - Ontologies
  - SourceData integration

![Metadata Generation Diagram]

Metadata
- annotation of csv columns
- Pdf, xls, ...

Create RO

SEEK
The Future: Docker container

Major challenge to create reproducible execution environments
- Operating Systems, library versions, ...
- Requirement for reproducible analysis

Docker: Build, ship, run.
- Docker containers wrap a piece of software in a complete filesystem that contains everything needed to run: code, runtime, system tools, system libraries – anything that can be installed on a server.
- This guarantees that the software will always run the same, regardless of its environment

Build docker containers which automatically reproduce the results
- Setup environment
- Pull repositories
- Run full analysis
- From any computer the exact same
- Distribute to nodes in cluster
The Present: Open source, open access, open data, open review

- Software and modeling in continuous flow
  - Publications are only snapshots in time

From idea to publication in 4-6 weeks

- F1000Research is an Open Science publishing platform for life scientists, offering immediate publication of articles, posters and slides without editorial bias. All articles benefit from transparent refereeing and the inclusion of all source data.
Pro & Cons: Repositories

A repository contains everything which is required to reproduce all results from scratch. I.e. a person can checkout the repository and recreate the complete analysis done in the paper.

• Advantages
  – Works
    • System which is approved in
  – Large user base & community acceptance
  – No overhead
    • No additional libraries, no XML/RDF/ISA/RO creation necessary
    • No annotations
  – Provenance
    • Who changed what when in which branch with which commit message
  – Contains everything

• Disadvantage
  – Only contains what is necessary to create the results and the history
    • Does not necessarily reflect the scientific question (ISA: investigation, study, assay)
    • What is the larger context (investigation)? Which experiments were performed?
  – Not for every project
    • ’long tail’ of research results.