Extracting reproducible simulation studies from model repositories using the CombineArchive Toolkit

Martin Scharm, Dagmar Waltemath

Department of Systems Biology and Bioinformatics
University of Rostock
D-18051 Rostock
martin.scharm@uni-rostock.de
dagmar.waltemath@uni-rostock.de

Abstract:
The COMBINE archive is a digital container format for files related to a virtual experiment in computational biology. It eases the management of numerous files related to a simulation study, fosters collaboration, and ultimately enables the exchange of reproducible research results. The CombineArchive Toolkit is a software for creating, exploring, modifying, and sharing COMBINE archives. Open model repositories such as BioModels Database are a valuable resource of models and associated simulation descriptions. However, so far no tool exists to export COMBINE archives for a given simulation study from such databases. Here we demonstrate how the CombineArchive Toolkit can be used to extract reproducible simulation studies from model repositories. We use the example of Masymos, a graph database with a sophisticated link concept to connect model-related files on the storage layer.

1 Reusability of simulation studies in systems biology

A study by the pharmaceutical company Bayer in 2011 showed that only 64% of published data in academic journals were replicable [PSA11]. This and similar observations led to initiatives and projects that aim at improved result reproducibility. For example, the Reproducibility Initiative[1] is a collaboration between Science Exchange, PLOS, figshare and Mendeley, and it identifies and rewards high quality reproducible research via independent validation of key experimental results. One example for a European project is FAIRDOM[2]. It is a collaboration of data management groups who will develop the necessary toolset to extend existing network services to the wider European Systems Biology community. Furthermore, funders require explicit data management strategies in grant applications. Trans-project data management systems such as the SEEK platform [WO+11] have become an integral part of the scientific landscape. Finally, various projects work towards reproducible experiments by re-thinking the publication process. For example, Research Objects [BDR+10] support the publication of data, code and other resources

[1] http://reproducibilityinitiative.org/
[2] http://fair-dom.org/

alongside the PDF paper. A call for Virtual Experiments [CVW14] highlights the actual benefits of generic simulation setups for the reusability of full behavioural repertoires of computational models.

Virtual experiments in systems biology projects comprise of a variety of files in heterogeneous formats [HWW14]: the encoding of the model, the associated simulation setups, the semantic description of the underlying biological and mathematical concepts, the graphical description of the modelled processes, reference publications, supporting figures, data tables for model parametrisation, result data sets, experimental data sets used to verify the simulation results etc. One achievement of the Computational Modeling in Biology Network (COMBINE, [WB+14]) is the implementation of guidelines and data formats for the standardised representation of this data. While the various types of model-related data remain heterogeneous, the data of each single type is well described. For example, models can be encoded in the Systems Biology Markup Language (SBML, [HF+03]), an XML format that is supported by over 100 software tools today [HB+11]. The semantics of a model can be specified through links into bio-ontologies [CJ+11], and in fact most publicly available SBML models are also curated and annotated. A format for the visual representation of a model is the Systems Biology Graphical Notation (SBGN, [LNH+09]). It defines standard glyphs for common biological objects, their properties and their interactions. Simulation setups can be described in the Simulation Experiment Description Markup Language (SED-ML, [WA+11]), enabling their immediate execution in all software tools that read the format.

In summary, the current situation is such that standards exist to encode model-related data, the data is publicly available in model repositories, it can be linked, and it can be queried. In this paper, we show how the retrieved data can be bundled as COMBINE archives. Our prototype implementation showcases a workflow that combines two existing software tools: A graph database for linked storage of model-related data (Masymos) and a web-based tool to generate COMBINE archives (CAT). Using M2CAT (Masymos to CAT) a model can easily be retrieved together with all other files that are necessary to understand the model and to reproduce the original results.

2 Exporting COMBINE archives from Masymos

The presented workflow integrates two of our previously developed software tools: Masymos, a graph database to store model files, simulation descriptions and other model-related data; [HWW14] and the CombineArchiveWeb tool [SW+14], an online application to manage COMBINE archives.

Masymos: Storing model-related data. Model management includes tasks that aim to optimise model storage, search, retrieval, provenance, visualisation, and comparison. The increased awareness of data management needs among researchers and funders requires novel ways of ensuring reproducibility, distribution, and management of modelling results. Masymos is a graph database of models and model-related data in computational biology [HWW14]. It is a Neo4J database that contains annotated models in SBML and

CellML formats, simulation descriptions in SED-ML format, and links to reference publications and bio-ontologies. Masymos implements a sophisticated link concept to connect the heterogeneous types of model-related data on the storage layer. We showed in earlier publications how our graph-based search over the linked data allows for novel types of queries, including structure queries, statistics, search for simulation setups etc [HWW14]. The data stored in Masymos is cloned from BioModels Database [LD+10] and from the CellML Model Repository [LL+08]. We also indexed the definitions of all bio-ontology concepts used in any of the models for subsequent ranked retrieval. The dominant ontologies are the *Gene Ontology*, the *Chemical Entities of Biological Interest* (ChEBI) and *UniProt*. Additionally, we integrated the Systems Biology Ontology, SBO [CJ+11], and the KInetic Simulation Algorithm Ontology, KiSAO [CJ+11].

CombineArchive Toolkit: Facilitating the transfer of simulation studies. Models need to be shared. The above standards guarantee interoperability of the encoded data, and the concept behind Masymos ensures easy access to simulation models and associated data. However, Masymos focuses on the management of files and on their long-term availability. It does not offer user-friendly data export. An archive-based tool for exporting and then sharing all files belonging to one scientific study is the CombineArchive Toolkit (CAT)[3]. It follows the COMBINE archive specification [BA+14] and generates zip-like file bundles.

The COMBINE archive is a zip container for files related to a simulation study. It uses the Open Modeling EXchange Format (OMEX) [BA+14] which mainly consists of a manifest file, metadata files, and all files necessary to reproduce a virtual experiment. The manifest lists all files belonging to the archive. It stores the format of each file, using a URI, and the file's location within the archive. In addition, files can be flagged "master", meaning those files should be evaluated first, whenever the archive is being loaded into a software tool. The metadata files may contain information encoded in common formats such as RDF, vCard, or Dublin Core Metadata. A list of tools that support the COMBINE archive is available from the specification [BA+14].

The COMBINE Archive Toolkit (CAT) is one implementation of the COMBINE archive format. It was built to support researchers in creating and exploring COMBINE archives. The CAT consists of a core library, a desktop application, and a web-based interface. The CombineArchiveLibrary is implemented in Java and offers all necessary methods to handle COMBINE archives (extract files, browse archive, add or remove files, rename or reorganise files, attach and retrieve metadata). The CombineArchiveWeb interface [SW+14] enables collaborative creation and modification of COMBINE archives. It offers RESTful services and can be used from other client applications. Currently, the CombineArchiveWeb interface has simple connections to BioModels Database and the CellML Model Repository, making it possible to retrieve models. The software is openly available and thus own instances can be installed if particular privacy interests apply.

[3] `https://sems.uni-rostock.de/projects/combinearchive/`

3 Implementation

Our software M2CAT enables researchers to retrieve models or simulation studies from Masymos and directly generates COMBINE archives from all relevant files (Figure 1). The data can then be either downloaded or opened in the CombineArchiveWeb interface.

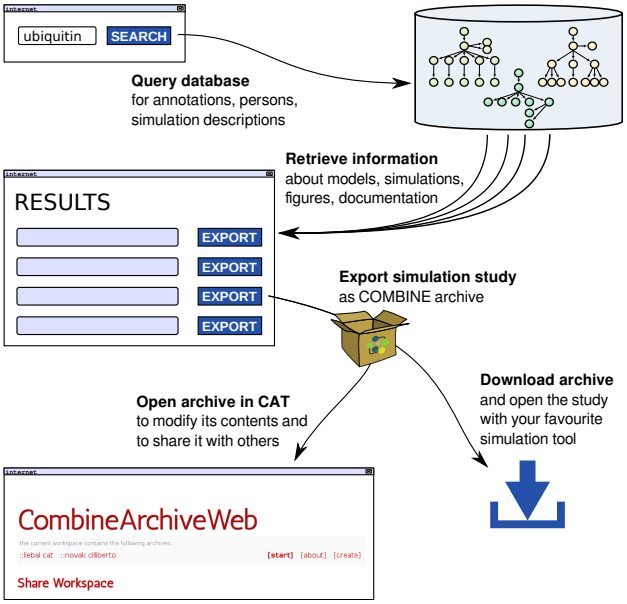

Figure 1: **The M2CAT workflow.** The keywords are used to search model files, simulation setups, and author names in Masymos. The search results are presented in a web interface. They are organised by model name, and displayed together with the additional resources. By clicking the *export* button, users can either download the COMBINE archive of an entry, or open the archive directly in the CombineArchiveWeb tool.

The workflow is implemented in a web interface at m2cat.sems.uni-rostock.de. Masymos can be queried via a simple search field. The keywords are translated into Cypher queries and sent to Masymos via Neo4j's RESTful service. First, M2CAT consults the annotation index, which contains all terms occurring in the semantic annotations of models (Query 1). The result is a list of models and associated documents.

```
START   res=node:annotationIndex('RESOURCETEXT:(ubiquitin)')
MATCH   res<-[rel:is]-(:ANNOTATION)-->(s)-[:BELONGS_TO]->(m:MODEL)-[:BELONGS_TO]->
        (d:DOCUMENT)
RETURN  distinct(m),d
```

Query 1: Return models `m` which contain elements annotated with *ubiquitin* and associated documents `d`.

Afterwards, M2CAT consults Masymos' Person nodes, which contain the names of all authors of reference publications, curators, and other contributors of model code (Query 2). The result is again a list of models and associated documents.

```
MATCH   (d:DOCUMENT)-[HAS_MODEL]->(l:MODEL)-[HAS_ANNOTATION]->
        (k:ANNOTATION)-[HAS_PUBLICATION]->(m:PUBLICATION)-[HAS_AUTHOR]->
        (n:PERSON{FAMILYNAME:"ubiquitin"})
RETURN  distinct(m),d
```

Query 2: Return models `m` which are annotated with publications written by *ubiquitin* and associated documents `d`.

Finally, M2CAT searches for additional resources for each entry in both lists. For example, Query 3 retrieves all simulation descriptions that are linked to a certain model document.

```
START   known=node(273)
MATCH   (known)-->(:MODEL)<--(:SEDML_MODELREFERENCE)<--(s:SEDML)<--(d:DOCUMENT)
RETURN  s,d
```

Query 3: For the document with id *273*: Return simulation descriptions `s` and associated documents `d`.

The sets of documents belonging to the simulation studies and which matched the keywords are then presented on the website. For each set M2CAT provides the model name and a list of available resources.

Using M2CAT it is now possible to export simulation studies as COMBINE archives. Masymos does not store the files themselves, but provides links to all necessary files through the retrieved document nodes. M2CAT resolves these links, retrieves the files, and utilises the CombineArchiveLibrary to create the archive. Specifically, the files are packed, the archive and its files are annotated with metadata about the contents and the creator, and the manifest file is written. The user can then either download the archive or open it in the CombineArchiveWeb tool (see again Figure 1). Downloaded archives simplify journal submissions, if the model code is required together with the manuscript. Opening the COMBINE archive in our web tools enables modellers to add further files to the archive, such as additional figures or supplemental descriptions. Further use cases of the COMBINE archive include easy sharing of files among collaborators, encapsulating datasets used for model development and validation, sharing consistent instances of a model, enabling automatic (machine-only) transfer of research results [BA+14].

4 Summary

The systems biology community has identified reusability of simulation studies as one major challenge in the field. We present here a workflow that combines existing tools for model management and provides a user-friendly interface for downloading reproducible simulation studies. The queried database Masymos already integrates several resources for model-related data. M2CAT gathers this data and generates COMBINE archives from it. Our instance of Masymos currently contains models in SBML and CellML formats and associated simulation experiments in SED-ML format. We think that the usefulness of the COMBINE archives generated by M2CAT increases, if more model-related data are included. Specifically, we wish to include graphical representations of models in SBGN format and simulation results. However, the described workflow is applicable to model repositories in general. It reveals its full power, if the model-related data are already linked, as demonstrated by Masymos.

References

[BA+14] F. Bergmann, R. Adams, et al. COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. *BMC bioinformatics*, 15(1):369, 2014.

[BDR+10] S. Bechhofer, D. De Roure, et al. Research objects: Towards exchange and reuse of digital knowledge. *The Future of the Web for Collaborative Science*, 2010.

[CJ+11] M. Courtot, N. Juty, et al. Controlled vocabularies and semantics in systems biology. *Molecular systems biology*, 7(1), 2011.

[CVW14] J. Cooper, J.O. Vik, and D. Waltemath. A call for virtual experiments: accelerating the scientific process. *Progress in Biophysics and Molecular Biology*, 10.1016/j.pbiomolbio.2014.10.001, 2014.

[HB+11] M. Hucka, F. Bergmann, et al. A profile of today's SBML-compatible software. In *2011 IEEE 7th International Conference on e-Science*, pages 143–150, 2011.

[HF+03] M. Hucka, A. Finney, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[HWW14] R. Henkel, O. Wolkenhauer, and D. Walthemath. Combining computational models, semantic annotations, and simulation experiments in a graph database. *DATABASE*, accepted for publication, 2014.

[LD+10] C. Li, M. Donizelli, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC systems biology*, 4(1):92, 2010.

[LL+08] C.M. Lloyd, J.R. Lawson, et al. The CellML model repository. *Bioinformatics*, 24(18):2122–2123, 2008.

[LNH+09] N. Le Novère, M. Hucka, et al. The systems biology graphical notation. *Nature biotechnology*, 27(8):735–741, 2009.

[PSA11] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712–712, 2011.

[SW+14] Martin Scharm, Florian Wendland, et al. The CombineArchiveWeb Application - A Web-based Tool to Handle Files Associated with Modelling Results. In *Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences, Berlin, Germany, December 9-11, 2014.*, 2014.

[WA+11] D. Waltemath, R. Adams, et al. Reproducible computational biology experiments with SED-ML-the simulation experiment description markup language. *BMC systems biology*, 5(1):198, 2011.

[WB+14] D. Waltemath, F. Bergmann, et al. Meeting report from the fourth meeting of the Computational Modeling in Biology Network (COMBINE). *Standards in Genomic Sciences*, 9(3), 2014.

[WO+11] K. Wolstencroft, S. Owen, et al. The SEEK: a platform for sharing data and models in systems biology. *Methods Enzymol*, 500:629–655, 2011.