Management of simulation studies in computational biology

Dagmar Waltemath
e:Bio Junior Research Group SEMS
Systems Biology and Bioinformatics, Institute of Computer Science, University of Rostock, Germany
dagmar.waltemath@uni-rostock.de

1 On the need for data management in computational biology projects

Data management is a well defined task in computer science which investigates methods for organising and controlling the information generated during (research) projects. It comprises several tasks, including data storage, search, retrieval, version control and provenance. Effective data management strategies for computational biology are needed to handle the increasing amount of data that is being generated and processed: High-throughput experiments generate large amounts of data; computational models become complex; novel methods for model coupling enable researchers to combine models into even larger systems; increasing computational power allows for complex simulations; and the availability of data at different scales demands clever integration techniques. However, recent studies showed that the rate of reproducibility of scientific results in the life sciences, including computational biology, is not acceptable [Ioa14]. As a consequence, efforts have been launched to improve reusability and reproducibility of biomedical results (e. g., [M$^+$14, Ioa14]), and results of simulation studies in particular [W$^+$11a, B$^+$14, C$^+$15]. Today, paths towards improved data management are discussed by funders and publishers, in large scale projects and by individual researchers. For example, funders established policies such as the *ERASysAPP Data Management Guidelines*; the German Network for Bioinformatics Infrastructure, de.NBI (http://www.denbi.de) has dedicated data management centers; and projects are funded to develop support for sustainable data management, e. g., FAIR-DOM (http://fair-dom.org).

The junior research group SEMS (http://sems.uni-rostock.de) focuses on the management of specific data: It develops methods and tools for the management of simulation studies in computational biology.

2 Methods and tools for the management of simulation studies

SEMS focuses on models encoded in XML standard formats and annotated with terms from bio-ontologies. More specifically, we work with models encoded in the *Systems Biology Markup Language* (SBML [H$^+$03]) and CellML [C$^+$03]. The majority of these models are mathematical models describing biological and physiological processes. The execution of these models can be described using the *Simulation Experiment Description Markup Language* (SED-ML [W$^+$11b]). While these three formats encode for the necessary information to run models [W$^+$11a], additional semantic annotations are needed to capture the biology [C$^+$11b], for example annotations to the Gene Ontology. Graphical representations of the networks can be standardised using the Systems Biology Graphical Notation (SBGN [LN$^+$09]). Together with ongoing developments of standards for data representation and ontologies to express the behavior and dynamics of a model, a whole plethora of data is collected when performing a simulation study. Figure 1 summarises how SEMS supports the management and integration of that data: The displayed reaction is part of a model reproducing the mitotic oscillator involving Cyclin and cfc2 kinase. The model itself was published in an article by Goldbeter in 1991 [Gol91]. Its SBML encoding, together with the graphical network in SBGN, is provided through BioModels Database, a rich resource of curated and annotated models [L$^+$10]. A standardised drawing of the interactions in the network enables researchers to quickly grasp the essence of what the model encodes. It is particularly useful to discuss different versions of the model with collaborators in large projects. Models may be simulated in different ways, with the actual setup of the experiments depending on the specific question asked.

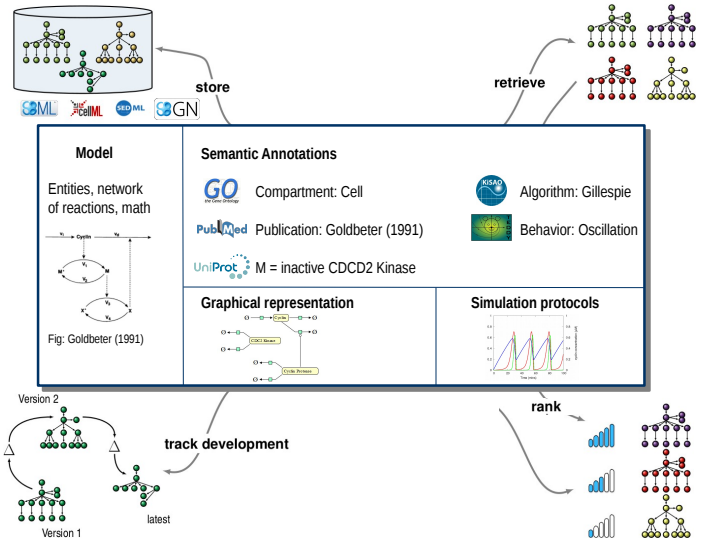

Figure 1: **Overview of integrated, model-related data and developed model management solutions.** The inner box shows the different types of model-related data that we integrate on the storage layer: model code (SBML, CellML), reference publications (e. g., PubMed), semantic annotations (bio-ontologies), graphical representations (SBGN), and simulation protocols (SED-ML). The outer ring shows our contributions to four major model management tasks: storage, search and retrieval, ranking, and version control.

These variations can be stored in SED-ML files, together with information on the simulation algorithm to use, or with links to parametrisation files and result data. An ontology of simulation algorithms is the *Kinetic Simulation Algorithm Ontology* (KiSAO [C$^+$11b]). Each result can be linked to a defined behavior encoded in the *Terminology for the Description of Dynamics* (TEDDY, [C$^+$11b]). Finally, the reference publication, typically in PDF format, is linked through a document object identifier (DOI).

The junior research group SEMS works towards better reproducibility of simulation studies, to lower the effort of reusing existing work, and to make scientific investigations more open and transparent. As depicted in Figure 1, we apply methods from data management on the problem of model management in computational biology. We are currently supported by three BMBF grants: The e:Bio junior group SEMS itself, one project in the German infrastructure for Bioinformatics (de.NBI), and another e:Bio project on SBGN-ED, an editor for SBGN maps. In all three projects, we collaborate closely with developers of model repositories (BioModels Database, *Physiome Model Repository* (PMR2 [Y$^+$11])) and data management systems such as SEEK [W$^+$15]. With our methods and tools, we aim to easen the findability, comparison, exploration, and understanding of simulation studies in computational biology.

Model search and retrieval The large number of published models necessitates smart search engines that incorporate the context of the model, semantic annotations, structural information and even allow for sub-model search. The results of a search need to be ranked according to user preferences. In 2010, we introduced the *ranked retrieval* engine for models, MORRE [H$^+$10]. It is a method to search and retrieve models from a given set, and to rank the results using state-of-the-art Information Retrieval methods. We showed how such a system improves the search in BioModels Database and PMR2. The first version of MORRE already considered model encoding and semantic annotations. We recently extended the method to enable clustering of models based on annotations [A$^+$15]. Furthermore, we investigate how similarity can be determined by transforming the network structure into a bipartite graph and detecting subgraph isomorphisms [RW14].

Model version control and provenance Models evolve over time, e. g., during model design, model publication, curation and reuse. A version of a model may fit a purpose while another may not, and

a model update may lead to modifications in the obtained simulation results. Model version control is therefore a necessary feature of all tools offerering model code for reuse. It allows users to track and understand the changes in a model [W+13]. To this end, we developed an algorithm for difference detection in versions of models, BiVeS [S+15]. It takes two versions of an SBML- or CellML-encoded model and calculates the differences. These differences can be exported in human-readable format, as an XML diff file, or they can be displayed visually. Since PMR2 integrated BiVeS, users can explore the changes of model code between different exposures. In the functional curation framework [C+11a], the BiVeS webservice is used to display differences in versions of CellML models.

Integration of model-related data We investigated the use of graph databases for the management of model-related data files. Our prototype system, MASYMOS [H+15], exemplifies how model-related data can be stored and linked, thereby enabling the retrieval of complete simulation studies. As most data are already encoded in XML, they can easily be converted into graph-like representations. Graph databases, in addition, enable flexible linking of data items. MASYMOS can thus reflect facts such as that a model is linked to several experiments, or that particular model entities are observed in a simulation. Ultimately, the application of graph concepts enables novel types of queries, for example for sub-models. This again can have a positive effect on the results of the ranked retrieval.

Exchange of reproducible simulation studies MASYMOS and MORRE contribute to the storage and retrieval of model-related data. How can one now export the extracted studies efficiently, without loosing important files, nor extracting files in wrong versions? Over the past years we contributed to the development of the COMBINE archive [B+14]. It serves as a container for all files necessary to reproduce a simulation study. Using the archive, simulation studies can thus be shipped as one single file. We developed a set of tools that read and modify archives; generate archives from data in MASYMOS; or enable sharing of archives online [SW15]. Another contribution of the group is the implementation of support for the COMBINE archive in the functional curation framework, where users can compare how different models handle a specific simulation task [Mir15].

Contribution to standards development SEMS actively contributes to the development of community standards, Minimum Information guidelines and ontologies through the *Computational Modeling in Biology Network* (COMBINE [W+14]). Specifically, our group members are editors of SBML, co-founders of SED-ML, active developers of the COMBINE Archive standard and coordinators of the COMBINE Network. We help with organising the annual community meetings, we support grant applications and outreach activities. For example, we teach students how to transform their modeling results into standard-compliant, reproducible simulation studies, and how to publish their data openly and sustainably. Finally, our group is part of the systems biology node for data management within the de.NBI network. Here we work towards integrating our model management tools into SEEK.

3 Summary: Promoting reproducible and open science

Reproducibility of scientific results is a major challenge in computational biology. The problem is manifold and can be addressed from different angles. In SEMS, we focus on the data management aspect: Only studies that are findable, verifiable, curated and well documented can be reproduced. A prerequisite is the availability of all necessary data and in interoperable formats. To this end we develop novel methods and tools for model management, specifically for search, retrieval, ranking, version control, and integration of model-related data. Furthermore, we are actively engaged in standards development and community efforts. Goals for the forthcoming years are to integrate SEMS tools in existing data management platforms, to raise the awareness for standards and reproducible science, and to integrate further types of data, specifically biomedical and clinical data.

References

[A+15] R Alm et al. Annotation-based feature extraction from sets of SBML models. *Journal of Biomedical Semantantics*, 6(1):20, 2015.

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

[C+03] A Cuellar et al. An Overview of CellML 1.1, a Biological Model Description Language. *SIMULATION*, 79(12):740–747, 2003.

[C+11a] J Cooper et al. High-throughput functional curation of cellular electrophysiology models. *Progress in Biophysics and Molecular Biology*, 107(1):11–20, 2011.

[C+11b] M Courtot et al. Controlled Vocabularies and Semantics in Systems Biology. *Molecular Systems Biology*, 7, 2011.

[C+15] J Cooper et al. A call for virtual experiments: accelerating the scientific process. *Progress in Biophysics and Molecular Biology*, 117(1):99–106, 2015.

[Gol91] Albert Goldbeter. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proceedings of the National Academy of Sciences*, 88(20):9107–9111, 1991.

[H+03] M Hucka et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *BIOINFORMATICS*, 19(4):524–531, 3 2003.

[H+10] R Henkel et al. Ranked retrieval of computational biology models. *BMC Bioinformatics*, 11(1):423, 2010.

[H+15] R Henkel et al. Combining computational models, semantic annotations and simulation experiments in a graph database. *DATABASE*, 2015:bau130, 2015.

[Ioa14] J Ioannidis. How to make more published research true. *PLoS Medicine*, 11(10), 2014.

[L+10] C Li et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4:92, Jun 2010.

[LN+09] N Le Novère et al. The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741, 8 2009.

[M+14] M Macleod et al. Biomedical research: increasing value, reducing waste. *The Lancet*, 383(9912):101–104, 2014.

[Mir15] G Mirams. Introducing the 'Cardiac Electrophysiology Web Lab'. https://mirams.wordpress.com/2014/05/09/web-lab/ (last accessed 2015-06-30), 2015.

[RW14] C Rosenke and D Waltemath. How Can Semantic Annotations Support the Identification of Network Similarities? In *Proceedings of the 2014 Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*, 2014.

[S+15] M Scharm et al. An algorithm to detect and communicate the differences in computational models describing biological systems. *BIOINFORMATICS, accepted for publication*, 2015.

[SW15] M Scharm and D Waltemath. Extracting reproducible simulation studies from model repositories using the CombineArchive Toolkit. In Norbert Ritter et al., editors, *Datenbanksysteme für Business, Technolog und Web (BTW 2015) - Workshop*, volume P-242, pages 137–44, 2015.

[W+11a] D Waltemath et al. Minimum information about a simulation experiment (MIASE). *PLoS Computational Biology*, 7(4):e1001122_1–e1001122_4, 2011.

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

[W+13] D Waltemath et al. Improving the reuse of computational models through version control. *BIOINFORMATICS*, 29(6):742–748, 2013.

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

[W+15] K Wolstencroft et al. SEEK: a systems biology data and model management platform. *BMC Systems Biology*, 9(1):33, 2015.

[Y+11] T Yu et al. The physiome model repository 2. *BIOINFORMATICS*, 27(5):743–744, 2011.