MaSyMoS: Finding hidden treasures in model repositories

Ron Henkel and Dagmar Waltemath

Department of Systems Biology and Bioinformatics, University of Rostock, Germany

Abstract. Ensuring reproducibility of simulation studies in computational biology poses a challenge to data management solutions. One specific problem is the organisation and linking of all files related to a simulation study at the storage level. Such a study comprises of a number of files, and additional links to third party resources. MaSyMoS is a graph-based approach to representing the links between the files necessary to reproduce a simulation study. It is based on a Neo4J database and designed specifically to handle files in COMBINE standard formats.
Availability: https://sems.uni-rostock.de/projects/masymos/

Introduction

Computational models in open repositories are available for reuse and thus reduce time and effort for modelers. However, the reproducibility of modeling results has repeatedly been questioned. Improvements have been made by developing standard exchange formats for models, simulations and more recently data [1]. The curation pipeline of repositories such as BioModels Database [2] ensures the validity of the contained models. Additional semantic annotations with concepts from bio-ontologies furthermore enhance the understanding of the models. Finally, a recently developed archive format allows modelers to download bundles of files related to a model [3]. While many developments contribute to the preservation of modeling results, only few address the problem of retrieving the valuable information contained in model repositories. In many cases it is simply necessary to find the most suitable model for a particular question in minimum time. Here, concepts from Information Retrieval research help identify the best model for a given query [4]. However, so far only the model files themselves were in the focus of search engines. Those model files are unfortunately not sufficient to reproduce a scientific result. Many other types of data need to be retrieved and their relation to the model of interest needs to be clearly defined. For example, the setup of the simulation experiments available for the model, result data, initial and alternative parametrisation of the model, model semantics, etc. are valuable information. Most of this information is available in standard formats hosted by the COMBINE community [1]. However, incorporating the connections between the single files and formats in a search engine is not trivial, specifically with traditional storage concepts such as relational databases [5]. An alternative to store and query models is based on RDF and SPARQL [6]

Results

The *Management System for Models and Simulations* (MaSyMoS) uses a graph database in the back-end [5]. It stores all information about a modeling result, using COMBINE standards [1] and semantic annotations [7,8,9,10]. Specifically, the published instance of MaSyMos[1] contains all models from BioModels Database [2] and the CellML model repository [11] (October 2014). MaSyMoS contains 1,211 model files in SBML format, 820 model files in CellML format, 38 simulation setups in SED-ML format, and concepts from four bio-ontologies: SBO 613, Kisao 261, GO 41,952, ChEBI 53,143. In total, the graph database contains 2,127,740 nodes and 13,298,567 relationships - a network of semantic knowledge in computational biology. MaSyMoS enables new kinds of queries, including statistical queries on models; retrieval of models and associated simulations; retrieval of simulation setups and associated models; retrieval of models containing a specific semantic annotation; or retrieval of models relating to a publication.

Statistics on available model data MaSyMoS generates statistics on models. The main resources of such models are BioModels Database and the CellML model repository. Using Cypher [12], MaSyMoS can be asked for most frequent entities in all models, based on the semantic annotations, or on entity names; for largest models in the system; for models with most associated experiments; or for counting participant roles in reactions. One could ask how many entities had been annotated with the concept SBO:0000247 (the Systems Biology Ontology term representing a simple chemical) or one of its children (Q1)? SBO:0000247 (simple chemical) has two children, SBO:0000327 (non-macromolecular ion) and SBO:0000328 (non-macromolecular radical). Interestingly, the simple chemical annotation is used 126,003 times, the non-macromolecular ion is used 37 times and the non-macromolecular radical is not used in any model. The three top-most annotations in curated and non-curated models are SBO:0000176 (biochemical reaction, 129,727 times), the aforementioned SBO:0000247 (simple chemical, 126,003 times) and EC 3.6.1.14 (Adenosine-tetraphosphatase, 108,732 times). A quick query reveals that all EC 3.6.1.14 annotations are encoded in 34 non-curated models based on [13]. For example, in BIOMID MODEL1310110042 this specific annotation is used multiple times for the same reaction – likely an error during the automated model generation process of this non-curated model.

We can also compare models by number of annotations (Q2). This query has proven helpful for our research on methods for annotation-based feature extraction from arbitrary sets of SBML models [14]. The model yeast metabolic network by Smallbone *et al.* (BIOMID BIOMD0000000473) contains 3028 annotations. In general, we found that the curated SBML models (Q3) in BioModels database contain a maximum of 3028 annotations, a minimum of four annotations, and on average 103.3 annotations. All models together contain a maximum of 50648 annotations, a minimum of one annotation, and on average 6809.4 annotations.

[1] https://sems.uni-rostock.de/projects/masymos/

Retrieving files by publications. MaSyMoS contains the models' references to publications describing the models (and experiments). The predominant link is a PubMed ID. Sometimes, more than one model file is linked to an entry in PubMed. MaSyMoS can easily retrieve all three versions of the model belonging to this specific publication, together with the simulation setups in SED-ML format, and all semantic annotations attached to those files. We found (Q4) that some publications are linked to more than one model. For example, the review of Calvin-Benson cycle models by Arnold *et al.* [15] resulted in 11 single models being published through BioModels Database (BIOMIDs 383-93). We found one more publication that has 10 curated models linked to it, and many more examples of publications with 7 or less models.

Retrieving simulation setups for a model. For a modeling result to be reproducible, the analyses run on that model must be repeatable. A prerequisite is the availability of all files describing the analyses, mostly simulations. We imagine that future search interfaces will rank models higher, if the simulation setup is attached to the model. MaSyMoS links models and simulation setups in SED-ML format and thus allows the immediate retrieval of complete studies that can be loaded and run without further tweaking of the model. Query 5 retrieves experiments associated to the CellML model Novak1997 (two SED-ML files representing Figures 2a and 2b of [16]).

Retrieving complete simulation studies by keyword. Sometimes the goal is not to retrieve a specific simulation study. Rather, it is to explore the information that is available about a biological question. In those cases it is desirable to retrieve all information about a certain biological fact, and explore the results. Here, MaSyMoS offers queries by semantic annotation [4]. These queries will typically access index structures and combine the results with property restricted graph matching. However, they are perfect to explore the content of the database. In Q6 we demonstrate a combination of all above mentioned techniques. The task is to retrieve all models dealing with a specific substance. In addition the species (representing an *m-phase inducer phosphatase*) must take part as a reactant in a reaction, and it must be observed by a simulation experiment.

Summary

We present MaSyMoS, a query interface for models and associated information. It runs on a Neo4J database in the back-end and incorporates models in SBML and CellML format, simulation setups in SED-ML format, and various ontologies for the mathematical description of the model, the simulation algorithms, and the semantic description of biological entities. MaSyMoS extends the capabilities of current search engines in existing model repositories, allowing models to retrieve complete and reproducible simulation studies as well as exploring the model space. A prototype implementation of MaSyMoS supports the search implemented for the CellML model repository. We furthermore are in the process of developing export functionality for the COMBINE archive format [3] and we aim to include version information in the next release of MaSyMoS.

```
Q1:
MATCH (i:SBOOntology {id:"SBO_0000247"})<-[:isA*0..]-(o)-[:ENTITY_TO_SBO]-(r:RESOURCE)
WITH r as Resource match (Resource)-[:BELONGS_TO]->(a:ANNOTATION)
RETURN Resource, count(a) as Num order by Num desc;
Q2:
MATCH  (m:SBML_MODEL)<-[:BELONGS_TO*1..2]-(a:ANNOTATION)<-[:BELONGS_TO]-(r:RESOURCE)
WITH   m as Model, count(r) AS NumberOfAnnotation
RETURN  Model, NumberOfAnnotation ORDER BY NumberOfAnnotation Desc limit 1;
Q3:
MATCH  (m:SBML_MODEL)<-[:BELONGS_TO*1..2]-(a:ANNOTATION)<-[:BELONGS_TO]-(r:RESOURCE)
WITH   m as Model, count(r) AS NumberOfAnnotation
RETURN  max(NumberOfAnnotation), min(NumberOfAnnotation),
      avg(NumberOfAnnotation), stdev(NumberOfAnnotation);
Q4:
MATCH (r:RESOURCE)<-[:isDescribedBy]-(a:ANNOTATION)-[:BELONGS_TO]-(m:CURATED)
RETURN r.URI as Publication, count(m) as RelatedModels
ORDER BY RelatedModels DESC limit 10
Q5:
MATCH (m:SBML_MODEL)-[:REFERENCES_SIMULATION_MODEL]-ref-[:BELONGS_TO*2]->(sed:DOCUMENT)
WHERE m.NAME='Novak1997 - Cell Cycle'
RETURN  m.NAME AS Model, m.ID as ModelID, ref.MODELSOURCE as ModelSource, sed.FILENAME as SEDMLFile
Q6:
START  res=node:annotationIndex('RESOURCETEXT:(m-phase inducer phosphatase)')
MATCH  res<-[rel:is]-(a:ANNOTATION)-->(s:SBML_SPECIES) <-[:OBSERVES]-o-[:BELONGS_TO*]->(doc:DOCUMENT)
WITH  doc,res,s  MATCH ()<-[:IS_REACTANT]-s-[:BELONGS_TO]->m
RETURN DISTINCT doc.FILENAME AS SEDML, collect(distinct m.NAME) AS Model,
         collect(distinct res.URI) AS Resource, collect(distinct s.NAME) AS Species
```

Example queries.

References

1. Waltemath *et al.*: Meeting report from the fourth meeting of the Computational Modeling in Biology Network (COMBINE). *SIGS* 9.3, 2014.
2. Li *et al.*: BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 2010.
3. Bergmann *et al.*: COMBINE archive: One File To Share Them All. *arXiv*, 2014.
4. Henkel *et al.*: Ranked retrieval of computational biology models. *BMC Bioinformatics* 11.1:423, 2010.
5. Henkel *et al.*: Combining computational models, semantic annotations, and associated simulation experiments in a graph database. *PeerJ Preprints* e376v1, 2014.
6. Wimalaratne *et al.*: BioModels linked dataset. *BMC systems biology* 8.1:91, 2014.
7. Courtot *et al.*: Controlled vocabularies and semantics in systems biology. *Molecular systems biology* 7.1, 2011.
8. Ashburner *et al.*: Gene Ontology: tool for the unification of biology. *Nature genetics* 25.1, 2000.
9. Degtyarenko *et al.*: ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research* 36.suppl 1, 2008.
10. *UniProt Consortium*: The universal protein resource (UniProt). *Nucleic acids research* 36.suppl 1, 2008.
11. Lloyd *et al.*: The CellML Model Repository. *Bioinformatics*, 2008.
12. Robinson *et al.*: Graph Databases. *O'Reilly Media*, 2013
13. Thiele *et al.*: A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* 31(5): 419-425, 2013
14. Alm *et al.*: Annotation-Based Feature Extraction from Sets of SBML Models. *Data Integration in the Life Sciences*, Springer International Publishing, 2014.
15. Arnold *et al.*: A quantitative comparison of CalvinBenson cycle models. *Trends in plant science* 16.12:676-683, 2011
16. Novak and Tyson:Modeling the control of DNA replication in fission yeast. *Proceedings of the National Academy of Sciences* 94.17: 9147-9152, 1997.