Possibilities for Integrating Model-related Data in Computational Biology

Dagmar Waltemath[1] and Olaf Wolkenhauer[1,2] and Nicolas Le Novère[3] and Michel Dumontier[4]

[1] University of Rostock, Dept. for Systems Biology and Bioinformatics, Germany
[2] STIAS, Wallenberg Research Centre, Stellenbosch University, Stellenbosch, SA
[3] Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
[4] Department of Biology, Carleton University, Ottawa, Ontario, Canada

Abstract. Computational models play an important role in framing our existing knowledge. Such a formal framework is capable of testing scientific hypotheses about biological systems. Increased numbers of annotated models are being developed in order to facilitate communication, search, comparison, retrieval and validation. A key part of model description relies on consistent representation of information which has been addressed through structured file formats and community guidelines. As we move towards a more comprehensive representation of model-related information, it becomes ever more important to understand how these information are most easily integrated in order to satisfy complex use scenarios. Here, we review key model-related data and emerging methods for facile integration and analysis. More effective approaches for knowledge representation are essential to decrease the time and effort involved in data integration and create new opportunities for model-based science.

1 Introduction

The number, size and complexity of computational models in biology are continuously increasing. Consequently, efficient model reuse is fundamental to progressing computational biology research [16, 23]. Scientists may search for models of interest, aim to compare virtual experiments, or aim to build large network models. They may also seek to infer additional knowledge in a field, to archive their works, or simply aim to make their work publicly available and reproducible by others. Here the effectiveness of reusability depends on several factors including the availability of models through the Web and open repositories; the supplementary information explaining what the model encodes; and the level of explanation provided about results associated to the models [28]. Open repositories such as BioModels Database [18] or the Physiome Model Repository PMR2 [30] grant researchers access to model code and associated meta-data. These models and their components are generally provided in standard format and annotated with terms from shared vocabularies. Shared vocabularies may be lists of allowed terms (a controlled vocabulary), hierarchically organized terminologies (a taxonomy) or more formally defined ontologies. Models are furthermore equipped with links to external information.

Standardized model representation formats are the Systems Biology Markup Language (SBML) [13], or CellML [20]. Their declarative nature makes models independent of the code-level implementation. Model code is provided together with links to reference publications and graphical representations of the models' network structure. In few cases, standardized simulation descriptions are available, and links to simulation tools capable of running the models are given.

Obviously, just making data digital – even if in standard formats – is not sufficient for facile integration [15]. We argue that additional effort is required to ensure full interoperability of knowledge put into and arising from computational modeling and simulation.

2 State of the Art

Model repositories provide access to already published models in different standard formats and at different levels of granularity. They also provide model-related information that may be mandatory to reuse a model, or helpful to understand it. To date, no common platform exists to query models across these repositories, e.g. searching for SBML *and* CellML models on the Cell Cycle. As a consequence, model retrieval remains a tedious process. Not only that different repositories have to be queried, it may even become necessary to compare implementations of the same model in different resources.

Models in standard format are generally equipped with additional knowledge. The abovementioned model representation formats use the Resource Description Framework (RDF) [5] for annotation. Annotations are attached to model constituents that represent objects in the biological world, mathematical concepts such as mathematical equations, modeling concepts such as Michaelis-Menten enzyme kinetics etc. Thus annotations link the computational model to the knowledge stored externally in databases and knowledge bases. Prominent ontologies in the field are the Gene Ontology [2], the NCBI Taxonomy [29], and the Systems Biology Ontology [7]. Annotated models form the basis for tasks such as model comparison, merging, search, or display [19, 24].

However, the standardized representation of models is only the first step towards reproducible e-science. A second and equally important aspect is that the methods and simulation parameters used to generate numerical results reported in a publication must also be adequately described [26]. SED-ML is being developed as an XML-based exchange standard for capturing the necessary elements to reproduce simulation [27]. Simulation descriptions in SED-ML include references to appropriate simulation algorithms from the Kinetic Simulation Algorithm Ontology (KiSAO) [7]. The resulting behavior of a model in a particular simulation experiment can then be formally described with annotations linking to terms from the TErminology for the Description of DYnamics (TEDDY) [7]. Information about models is complemented by data from the literature (e.g., reference publications), graphical representations of models (e.g., in the Systems Biology Graphical Notation (SBGN) [17]), or pointers to simulation results (e.g., in costly simulations). Further data relevant for model reuse include experimental data and setups that form the basis for a computational representation of a biological system, initial assumptions and considerations during the modeling process, or constraints of the model when reusing it. In this paper we focus on the model-related data summarized in Table 1.

Type	**Format**
publication	text
model	XML (SBML, CellML, NeuroML); RDF/OWL (SIO)
simulation	XML (SED-ML); OWL (KiSAO)
results	XML (NuML); RDF/OWL (SIO)
controlled vocabularies	OBO, RDF/OWL
experimental data	XML (MAGE-ML, PSI-MI, MzML); RDF/OWL (Bio2RDF, SIO, OBI)

Table 1. Selection of model-related data and their availability in standard formats.

Research projects concerned with the integration of data for the life sciences are manifold [1] and based on different technologies such as workflow systems, semantic web technologies or graph-based approaches [21, 4]. Remarkably, these projects focus on the experimental data itself and therefore have well identified problems with integrating Bioinformatics data. More recently, the value of frameworks for the integration and simulation of computational models in physiology has been demonstrated [9]. The authors present a method to integrate models of different scales based on ontology knowledge. Similarly, possibilities for the integration of models with simulation descriptions have recently been outlined [10]. Following up on these efforts, we argue that, similarly to the ISA project [22] which integrates and makes interoperable resources linked to experimental data, integration methods must now be developed for model-related information. In this paper, we review some ongoing developments in computational biology.

3 Possibilities for the Integration of Model-related Data

Data integration is fundamental to improved interoperability of computational models and may have different purposes: To transport a complete model from one point to the other, e.g. in collaborative modeling projects, or when submitting a model for publication in an open database. Another purpose is to integrate the data to search it. One may want to limit a search to a particular set of models that contain the parameter estimation step, or one may look for models that are based on a specific set of experimental data. Finally, the integration of model-related data allows to extend the current knowledge of a system using inference. In the following, we describe three already developed approaches to integrate model-related data: First, all data may be stored in one COMBINE archive. Second, a graph database can be implemented to link all resources on the storage level. Third, the data can be converted into linked RDF-data, as exemplified for a great number of resources by the BIO2RDF project [3].

3.1 The COMBINE Archive

The initial idea of an archive to exchange model and simulation descriptions had been proposed to support the use of SED-ML. Based on that idea, the COMBINE archive provides researchers a single-file format for storing all model-related information (`http://co.mbine.org/documents/archive`). A single file is easier to share among research partners, and it is easier to store. A COMBINE archive is a zip-file containing the various documents necessary for the description of one or several models and all associated data and procedures for running these models. For instance, a COMBINE archive may contain models in SBML, simulation experiment descriptions in SED-ML, and graphical representations in SBGN-ML. Furthermore the archive may include publications, figures, result data, annotation files and others. A manifest file, encoded in XML, lists the location and type of each file within an archive. In the current version of the COMBINE archive, all the files described must be included in the archive itself. However, it is envisioned that in the future the manifest could list files located elsewhere, using valid and resolvable URIs, following the IDENTIFIERS.org scheme [14]. The archive's metadata file then contains clerical information about the various files contained in the archive, and the archive itself. The use of the COMBINE archive format is expected to improve the exchange of computational studies of biological systems. For example, BioModels Database curators welcome the submission of COMBINE archives instead of sole model code as archives help to quickly reproduce the results that the model says to generate and thereby to curate and publish the models in BioModels Database.

3.2 Graph-based Model Storage

An alternative way to integrate model-related information is through a graph-based storage approach. The great amount of meta-information associated with today's models, and the fact that models represent network structures make graph databases attractive for model storage [10]. NoSQL approaches have already been successfully used in other Life Science applications [25], for example in projects like Bio4J (`http://bio4j.com/`) which is a graph-based database for bio-ontologies. In the context of computational biology, a graph-based model store can be built by transforming the model's XML structure into a graph consisting of nodes (for model constituents and annotations) and edges (relations between the nodes). Edges may represent the link from a model entity to an annotation node, or the link of a model entity to another constituent (e.g., linking an SBML species to an SBML reaction). Graph-representations of existing ontologies may either be imported (but then need to be updated regularly) or linked to using unambiguous identifier schemes, such as the aforementioned IDENTIFIERS.org. In addition, Information Retrieval techniques can be implemented on top of the graph-database to query model constituents and their annotations. Available cross-links between ontology terms can be incorporated to build a highly connected index which allows to generate ranked result lists of models for a given query [11]. This search is based on annotations and thus independent of the underlying model representation format. Interestingly, the graph representation also enables structure-queries to search for subnetworks represented by the nodes and edges in the model graph. A similar graph-representation of simulation descriptions will foster linking model constituents and simulation experiments [10]. Finally, both ontologies and SED-ML files may link to models of different representation formats, thereby establishing further relationships between models and making them easier comparable. The flexible graph structure allows for creating direct relationships between model constituents of any model, for example stating that one constituent in an SBML model equals a constituent in a CellML model.

3.3 Semantic Web enabled Integration

With the World Wide Web as a ubiquitous platform for the publication and dissemination of information, a key challenge is in how to use Internet technology to facilitate the publication and discovery of structured, interlinked data. To address this very issue, the World Wide Web Consortium (W3C) initiated the Semantic Web effort for the representation, publication, integration and query of data in standard formats. At the core of the effort lie technologies such as RDF, RDF schema (RDFS), SPARQL query language, and the Web Ontology Language (OWL). RDF offers a simple mechanism to i) identify and ii) describe entities in terms of their types, attributes and relations to other entities. In the context of computational biology, these entities are, for example, models, model constituents, or result data. They are unambiguously identified by Internationalized Resource Identifiers (IRIs) which allows for de-referenceable web-based identifiers (HTTP URIs). That is to say, when users paste the identifier in their web browsers, they will get back information about the entity of interest. This information is structured in a so-called triple or statement consisting of a subject, a predicate and an object or literal. Bio2RDF (`http://bio2rdf.org/`) is an international project that uses Semantic Web technologies to provide a global network of linked data for the life sciences [6]. Adopting a simple Web-friendly naming convention with contributed open-source software, Bio2RDF currently processes hundreds of thousands of user queries per month regarding billions of statements about millions of entities from several dozen scientific databases. This advanced research platform enables investigators to easily construct sophisticated queries that span genes, gene expression, genetic variation, proteins, protein domains, interaction networks, pathways, and diseases. With the inclusion of the Gene Ontology and BioPAX-formatted, computational models from BioModels Database to Bio2RDF, it is now easy enough to craft a federated query that for instances, determines the number of the biochemical reactions that are involved in protein catabolic processes based on the structure of the Gene Ontology (Table 2).

While most ontologies offer a simple taxonomy through a hierarchical organization of terms, OWL makes it possible to formally describe the attributes of types and relations such that they can be used for automatic classification and consistency checking. For instance, we might like to check that all those human

Gene Ontology Annotation	**No. of Reactions**
protein catabolic process [go:0030163]	51
cellular protein catabolic process [go:0044257]	26
modification-dependent protein catabolic process [go:0019941]	1
beta-amyloid formation [go:0034205]	1
cyclin catabolic process [go:0008054]	1

Table 2. Finding biochemical reactions involved in protein catabolic processes (BIO2RDF)

curated semantic annotations in the BioModels Database are actually correct. One approach, termed the SBML Harvester [12], involves converting the RDF-based semantic annotations into OWL ontologies and using a reasoner to uncover any inconsistencies. Application to the BioModels Database yields an OWL ontology with more than 300,000 classes, 800,000 axioms and includes all referenced ontologies: GO (functions, compartments, processes), ChEBI (molecules), Cell-type Ontology (cell types), FMA (anatomy) and PATO (qualities). Reasoning over the integrated ontology resulted in 27 inconsistent models, for which most could be attributed to errors in the annotation. In two models, BIOMODELS 176 and 177, the semantic annotation incorrectly associated an ATPase reaction with alpha-D-glucose-phosphate instead of the correct ATP species. More recent work illustrates how the Semanticscience Integrated Ontology (SIO) can act within the semantic web framework to integrate and validate biological data and terminology, models, parameters, and simulation results through reasoning, computational trend analysis and provenance [8]. Indeed, the semantic web offers a more sophisticated solution for advanced knowledge representation and discovery.

4 Conclusions

The repeated call for reproducible scientific results in Systems Biology leads to increased presence of models in standard format. The existence of rich ontologies of biomedical knowledge and reference systems furthermore form a basis for ontology-based data integration. It is also a fact that computational models cannot anymore be published on their own. One reason are stricter requirements for reproducibility by the journals. Another reason is that models become more and more complex, forming ever larger networks, which makes them harder to understand by simply looking at them.

The integration of model-related data can be realized through unambiguous model identifiers, e.g., the ones used for CellML exposures. Implicit links between models are given in the model annotations and can be used for model search and simulation descriptions using multiple models. However, alternative approaches include the generation of generic schemes for the integrated management of model-related data. These should also be explored, because they can be build on elaborated methods from the research field of data integration.

Integration of model-related data should be independent of the underlying model representation format. This requirement is fulfilled by all three approaches introduced in this paper. The COMBINE archive creates a bundle of files necessary to reproduce a result shown by a model. The shipping of a single (zip) file is easier and guarantees that the information necessary for reproduction is always complete. However, the information inside the archive is not interconnected with external resources per se, but needs to be extracted and stored appropriately to be queryable. A graph storage forms a network of interrelated nodes (models, simulation descriptions, ontology terms). The edges between nodes allow to define flexible connections. The graph-based approach furthermore enables queries regarding the structure of model data, and already existing Information Retrieval techniques are easily applicable to the integrated network of models. Finally, the semantic web approach to integrating ontological knowledge with model-related data lies in the scalable federation of relevant information using web technology coupled with powerful query answering using automated reasoning.

References

1. Antezana, E., et al.: Biological knowledge management: the emerging role of the semantic web technologies. Briefings in bioinformatics 10(4), 392–407 (2009)
2. Ashburner, M., et al.: Gene Ontology: tool for the unification of biology. Nature genetics 25(1), 25–29 (2000)
3. Belleau, F., et al.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics 41(5), 706–716 (2008)
4. de Bono, B., et al.: The RICORDO approach to semantic interoperability for biomedical data and models: strategy, standards and solutions. BMC Research Notes 4(1), 313 (2011)
5. Brickley, D., Guha, R.V.: Resource Description Framework (RDF) Schema Specification 1.0: W3C Candidate Recommendation 27 March 2000 (2000)
6. Callahan, A., et al.: Ontology-based querying with Bio2RDFs linked open data. Journal of Biomedical Semantics 4(1), 1–13 (2013)
7. Courtot, M., et al.: Controlled vocabularies and semantics in systems biology. Molecular Systems Biology 7(1) (2011)
8. Dumontier, M., et al.: Systems Biology: Integrative Biology and Simulation Tool, chap. Semantic Systems Biology: formal knowledge representation in systems biology for model construction, retrieval, validation and discovery. Springer (2013)
9. Erson, E.Z., Çavuşoğlu, M.C.: Design of a framework for modeling, integration and simulation of physiological models. Computer methods and programs in biomedicine 107(3), 524–537 (2012)
10. Henkel, R., et al.: Considerations of graph-based concepts to manage computational biology models and associated simulations. In: Proceedings of the Proceedings of the INFORMATIK 2012 Conference
11. Henkel, R., et al.: Ranked retrieval of computational biology models. BMC bioinformatics 11(1), 423 (2010)
12. Hoehndorf, R., et al.: Integrating systems biology models and biomedical ontologies. BMC systems biology 5(1), 124 (2011)
13. Hucka, M., et al.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4), 524–531 (2003)

14. Juty, N., et al.: Identifiers. org and MIRIAM registry: community resources to provide persistent identification. Nucleic acids research 40(D1), D580–D586 (2012)
15. Larson, S.D., Martone, M.E.: Ontologies for neuroscience: what are they and what are they good for? Frontiers in neuroscience 3(1), 60 (2009)
16. Le Novère, N.: Model storage, exchange and integration. BMC neuroscience 7, S11 (2006)
17. Le Novère, N., et al.: The systems biology graphical notation. Nature biotechnology 27(8), 735–741 (2009)
18. Li, C., et al.: BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC systems biology 4(1), 92 (2010)
19. Li, P., et al.: Systematic integration of experimental data and models in systems biology. BMC bioinformatics 11(1), 582 (2010)
20. Lloyd, C.M., et al.: CellML: its future, present and past. Progress in biophysics and molecular biology 85(2), 433–450 (2004)
21. Missier, P., et al.: Taverna, reloaded. In: Scientific and Statistical Database Management. pp. 471–481. Springer (2010)
22. Sansone, S.A., et al.: Toward interoperable bioscience data. Nature genetics 44(2), 121–126 (2012)
23. Sauro, H.M., et al.: Challenges for modeling and simulation methods in systems biology. In: Proceedings of the Winter Simulation Conference, 2006. WSC 06. pp. 1720–1730. IEEE (2006)
24. Schulz, M., et al.: SBMLmerge, a system for combining biochemical network models. Genome Informatics Series 17(1), 62 (2006)
25. Splendiani, A., et al.: Lost in translation: data integration tools meet the semantic web (experiences from the Ondex project). In: Recent Progress in Data Engineering and Internet Technology, pp. 87–97. Springer (2012)
26. Waltemath, D., et al.: Minimum information about a simulation experiment (MIASE). PLoS computational biology 7(4), e1001122 (2011)
27. Waltemath, D., et al.: Reproducible computational biology experiments with SED-ML – the simulation experiment description markup language. BMC systems biology 5(1), 198 (2011)
28. Waltemath, D., et al.: Systems Biology: Integrative Biology and Simulation Tool, chap. Reproducibility of model-based results in systems biology. Springer (2013)
29. Wheeler, D.L., et al.: Database resources of the national center for biotechnology information. Nucleic acids research 35(suppl 1), D5–D12 (2007)
30. Yu, T., et al.: The physiome model repository 2. Bioinformatics 27(5), 743–744 (2011)