Discovering model-model connections in biological model repositories

John H. GENNARI[1], Maxwell L. NEAL[1], Robert HOEHNDORF[2], Georgios V. GKOUTOS[2] and Daniel L. COOK[1,3]

[1]*Biomedical & Health Informatics, and* [3]*Physiology & Biophysics, University of Washington, USA*
[2] *Department of Genetics, University of Cambridge, UK*

Correspondence: <gennari@uw.edu>

Abstract

Our long-term goal is to provide guidance for semantic annotation of biosimulation models and data, allowing for improved multi-scale modeling, modular model construction, and integration of models across repositories. In this paper, we focus on the task of searching and retrieving candidate models for merging over multiple biosimulation model repositories. In particular, we describe our semantic integration of the CellML model repository, the Reactome database, and the BioModels database. We introduce BioSim-Connector, a tool that uses this integrated knowledge base to search for connections across models from all three repositories. Researchers can then use these connections to better identify candidate models for merging into larger, multi-scale models for the Physiome or the Virtual Physiological Human.

Introduction

As quantitative tools for integrating theory and data, simulation models are becoming increasingly important both for research use, and more recently, for the application of multi-scale, patient-specific modeling for clinical use. Despite the growing need for complex models, their development remains a largely manual, error-prone process. To improve model sharing and reuse, researchers have developed the Systems Biology Markup Language (SBML) and CellML model description formats, and modelers now have access to hundreds of models in each format through the BioModels database (www.ebi.ac.uk/biomodels/) and the CellML model repository (www.cellml.org/models). The BioModels developers have also established the MIRIAM guidelines [1] for annotating models, and as part of this approach they have annotated their curated models by linking model components to terms in biomedical ontologies.

Standardization is necessary, but insufficient by itself, for the semi-automatic, modular construction of multi-scale models. As part of the VPH RICORDO project, we have developed several technologies to help move this process forward. First, we have developed methods for using *composite annotations* that provide richer, physics-based semantics for models [2]. Next, we have developed a suite of prototype tools for annotating models, browsing libraries of models, and merging models [3]. We have also developed an approach for consistency checking and improved searching of the BioModels repository [4]. Here, we report on our work developing BioSimConnector, a prototype tool for model library integration and search (http://code.google.com/p/bio-sim-connector/). We have integrated three prominent model resources: (1) the Biomodels database, where models are described in SBML, (2) the CellML repository, which describes models using CellML notation, and (3) the Reactome database (www.reactome.org), which uses the BioPax standard to describe biochemical reaction pathways. As an initial proof-of-concept, we now have the capability to (a) search across all three resources, and (b) search for process *linkages* or paths through this space of models. The latter capability works via semantic annotations that connect models to common biomedical ontologies.

Methods

BioModels, the CellML repository, and Reactome can all output information using XML/RDF syntax. We developed a single knowledge base that integrates all three, using common semantics to link models. First, we leveraged the "SBML Harvest", our OWL knowledge base built from the BioModels repository [4]. We translated the SBML Harvest into a simplified RDF triple store that focuses on the relations between the biological elements of SBML models (species, reactions and compartments) and the biomedical ontology terms that capture their meaning.

Unlike BioModels, the CellML repository is largely lacking in biological semantic annotations. To overcome this problem, we used text mining methods and the availability of PubMed IDs in CellML metadata. We retrieved each model's PubMed abstract, using the IDs and NCBI's eFetch REST service, and then sent the abstract to the NCBO Annotator tool [5] to extract keywords that are linked to biomedical ontologies. After removing common "stop words" (e.g., "cell", "protein"), we cast these model-level annotations as RDF triples and added the information to our knowledge base.

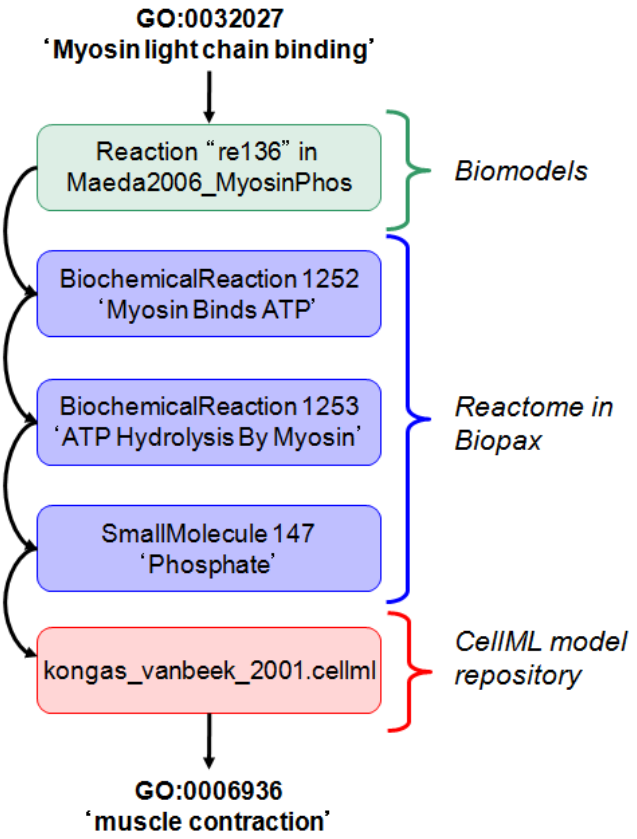

Fig 1. The best path of linked models and entities that connects two example ontology terms, as found via BioSimConnector.

Finally, we also included all *Homo sapiens* pathway knowledge from the Reactome database as RDF triples. Unlike both of our other resources, this repository is not a set of biosimulation models, but rather, a qualitative model of reactions and pathways (e.g., without rate constant information) that we use as a source of "missing links" within and between quantitative simulation models.

Searching and connecting biosimulation models

Before one can merge models to address a particular biological problem, one must first *find* appropriate candidate models. This is a challenging task, and is the primary use case for our prototype BioSimConnector tool. This tool uses our combined RDF knowledge base as single network, with nodes that represent physical entities, processes, or whole biosimulation models.

Given two concepts from the set of ontology annotations in our RDF store, we can search for paths through our "model space" to discover semantic connections between these concepts. Our implementation of this search uses Dijkstra's algorithm to compute the best paths between pairs of concepts. If no weights are used, the best path is the shortest path; however, we assign weights to different predicates (links) in our knowledge base, which allows us to better identify models that may be more amenable for integration. This sort of relation weighting also allows for search customization by the user.

Figure 1 demonstrates an example search between two ontology terms from GO: "myosin light chain binding" and "muscle contraction". BioSimConnector finds the best path between these two terms, given

our current knowledge base. This path traverses information from the Biomodels database, Reactome, and the CellML repository. In this example, the Reactome information associated with 'ATP Hydrolysis by Myosin' acts as a bridge between an SBML model representation of myosin light chain binding and a CellML model representation of muscle contraction. This sort of Reactome bridge could represent a "gap" in modeling knowledge, and therefore an opportunity for new research focused on the development of quantitative biosimulation models that could bridge that gap.

A limitation of our current approach is that some annotations have weak semantics, especially those from the CellML repository. The semantics are weaker for these annotations because we derive them automatically via text mining methods, and because they are necessarily about the model as a whole, rather than about more specific model components. Currently, we are collaborating with CellML curators and researchers to improve the level of biologically meaningful annotations for their models.

Conclusions

We have demonstrated, in principle, how semantic integration of the knowledge in Biomodels, Reactome and the CellML repository, can assist with model searching, reuse, and recombination. In particular, our methods have the potential to not only identify overlaps that exist between models, but to also bridge more distantly-related quantitative models via qualitative pathway knowledge. We have implemented this approach with prototype software: the BioSimConnector. In addition, we have also developed PhysioMaps [6], that are computable networks for visualization of the physiological processes within our knowledge-base of physiological models. As the number of published models in repositories increases, researchers will be less capable of integrating models and knowledge by hand, and must therefore rely more on technology such as the BioSimConnector tool and PhysioMaps. Our claim is that these tools will be critical building blocks for implementing the vision of the Physiome and the Virtual Physiological Human.

Acknowledgements

We would like to acknowledge the contributions of Thai Le, who assembled the set of CellML model abstracts. This work was partially funded by the VPH Network of excellence, project #248502.

References

1. Le Novere, N, Finney, JR et al., *Minimum information requested in the annotation of biochemical models (MIRIAM).* Nat Biotechnol, 2005. **23**(12): p. 1509-15.

2. Cook, DL, Bookstein, FL, and Gennari, JH, Physical Properties of Biological Entities: An Introduction to the Ontology of Physics for Biology. *PLoS ONE*, 2011, 6(12): e28708.

3. Gennari, JH, Neal, ML, Galdzicki, M, and Cook, DL, Multiple Ontologies in Action: Composite Annotations for Biosimulation Models. *Journal of Biomedical Informatics*, 2011, **v 44**(1), pp. 146-154.

4. Hoehndorf, R, Dumontier, M., Gennari, JH, Wimalaratne, S, deBono, B, Cook, DL, and Gkoutos, GV, Integrating systems biology models and biomedical ontologies, *BMC Systems Biology*, 2011; 5: 124. doi:10.1186/1752-0509-5-124

5. Jonquet, C, Shah, N, Musen, M, The Open Biomedical Annotator, *AMIA Summit on Translational Bioinformatics*, p. 56-60, March 2009, San Francisco, CA, USA.

6. Cook, DL, Neal, ML, Hoehndorf, R, Gkoutos, GV, and Gennari, JH, Developing a PhysioMap. *BioOntologies SIG 2012*, July 13-14, 2012, Long Beach, CA, USA.