CEUR Workshop Proceedings
http://ceur-ws.org
ISSN 1613-0073

Vol-596
urn:nbn:de:0074-596-3

ORES-2010
Ontology Repositories and Editors for the Semantic Web

Proceedings of the 1st Workshop on Ontology Repositories and Editors for the Semantic Web

Hersonissos, Crete, Greece, May 31st, 2010.

Edited by

Mathieu d'Aquin, **The Open University, UK**
Alexander García Castro, **Universität Bremen, Germany**
Christoph Lange, **Jacobs University Bremen, Germany**
Kim Viljanen, **Aalto University, Helsinki, Finland**

10-Jun-2010: submitted by Christoph Lange
11-Jun-2010: published on CEUR-WS.org

OREMP: Ontology Reasoning Engine for Molecular Pathways

Renato Umeton[1], Beracah Yankama[1], Giuseppe Nicosia[2], and C. Forbes Dewey, Jr.[1]

[1] Massachusetts Institute of Technology, Cambridge MA 02139, USA,
oremp@mit.edu,
WWW home page: http://cytosolve.mit.edu
[2] University of Catania, Viale A. Doria 6, 95125 Catania, Italy

Abstract. The information about molecular processes is shared continuously in the form of runnable pathway collections, and biomedical ontologies provide a semantic context to the majority of those pathways. Recent advances in both fields pave the way for a scalable information integration based on aggregate knowledge repositories, but the lack of overall standard formats impedes this progress. Here we propose a strategy that integrates these resources by means of extended ontologies built on top of a common meta-format. Information sharing, integration and discovery are the primary features provided by the system; additionally, two current field applications of the system are reported.

1 Introduction

An increasing number of quantitative biomolecular pathway databases are updated and curated on a regular basis [1, 2], because molecular processes are being characterized and their descriptions shared continuously. Substantial effort has been devoted to the creation of searchable biological resources (such as GO [3] and UniProt [4]) which are publicly available, but there are semantic obstacles that inhibit their combined use. Different languages (*i.e.*, the data formats) are spoken by the data sources; there are different abstraction levels; and there is a lack of an overall frame capable of identifying overlaps and duplications [5]. Some syntactic conversions are available among pathway data-formats, and the state of the art for adjudication of the discrepancies between two SBML [6] models is semanticSBML [7], which exploits machine-readable information and the user input to create a merged SBML model. In the context of large-scale composite biological pathways, the merged-model approach is undesirable because it destroys the original component models and interrupts the curation process. For more than two SBML files, the tool must be run repeatedly with user-input, subjecting it to increasing human error, and suggesting that the order in which the models are aligned matters. An alternative approach based on the use of ontologies discerns when and on which topics models are a relevant part of the large-scale context. The state of the art is represented by BioPortal [8] which provides uniform access to most of the biomedical ontologies through a single

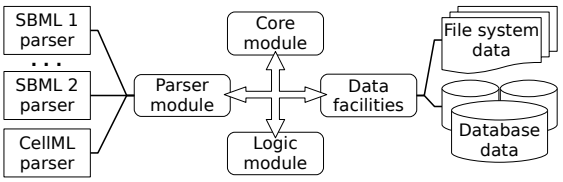

Fig. 1. System components are integrated to work together preserving a flexible and easily extensible architecture. Each module has different versions used on the basis of job in progress (*e.g.*, to parse an SBML file, will be dynamically chosen the SBML parser).

Entity	*has*
Annotation	type:STRING, uri:STRING, information:STRING.
Species	name:STRING, internalId:STRING, initialValue:REAL, inPathway:PATHWAY, hooks:SET_OF_ANNOTATIONS.
Kinetic reaction	internalId:STRING, kinetics:FORMULA, kineticParameters:SET_OF_PARAMETERS, inPathway:PATHWAY, reactants:SET_OF_SPECIES, catalysts:SET_OF_SPECIES, products:SET_OF_SPECIES, hooks:SET_OF_ANNOTATIONS.
Parameter	name:STRING, value:REAL.
Pathway	fullname:STRING, hooks:SET_OF_ANNOTATIONS.

Table 1. Main components of the minimalistic quantitative MIRIAM-compliant ontology used to abstract heterogeneous resources associated with biomolecular pathways. The format "attribute:REPRESENTATION" is used.

user-interface and advanced tools to query over biomedical data resources. Still, there are a lack of strategies for the database and ontology integration of quantitative biological sources written in different standards (*e.g.*, SBML and CellML [9]). What is described here is a system that creates extended ontologies out of different biochemical information sources and provides path duplication detection, sharing, integration, and knowledge discovery over heterogeneous resources. A prototype exists (MIT license, cf. http://cytosolve.mit.edu/oremp for software details) with utilities to export the extended ontologies in OWL format. This combination represents an Ontology Reasoning Engine for Molecular Pathways (OREMP). The OREMP framework creates extended ontologies out of different quantitative data formats and can be browsed at different levels of abstraction.

2 The Designed Framework

System Architecture. The system is composed of interchangeable and extensible components (Fig. 1). The four components interact as follows: (i) the *data access facilities* collect information about multiple pathways and existing biological databases; (ii) the *parser component* accesses different file formats (RDF, XML,

SBML, CellML, etc.) and extracts information from those sources; (iii) the *core module* assembles the knowledge from different sources into a coherent ontology (Table 1), and (iv) the *logic component* defines the conditions that identify when two biomolecular species are the same, or two reactions overlap. The combined execution of the two models without detecting reaction duplication will produce an incorrect evolution of species concentrations in time. This is a concrete, quantitative effect of incorrect ontology alignment. While the operational work-flow (i-iv) is kept fixed, it is of note that different versions of each component may be loaded by the system. A user-configurable algorithm chooses at run-time the components that are required for the current job. Whenever a new modeling standard is introduced, a new parser can be connected to OREMP to interface with it as well. Similarly, different users can define different versions of the *core component*, for example, according to their understanding about how the knowledge coming from different pathways should be aggregated. A useful analogy is the way modern graphics display programs seamlessly support different file formats (JPG, TIFF, DCM, etc.). Our approach is different from semanticSBML in that it provides the user the opportunity to exploit his/her understanding to define a consistent method of knowledge integration across ontologies. The independent curation process is preserved by maintaining the pathway identity, since the primitive element-pathway network is not destroyed by integration. Finally, we can optionally accept a dictionary of already aligned species, which can easily scale in the number of input pathways, as related in the next section.

Ontologies From Pathways. The system is constructed of three layers. The bottom layer represents the biochemical pathways, read in their primitive format (such as SBML and CellML). The second layer abstracts the pathways into a minimalistic and quantitative meta-format (sketched in Table 1) that includes all the MIRIAM [10] components. Annotations are preserved and extended with additional quantitative data to achieve a common description that can be represented as a single ontology. It is at this level that the extended ontology is primarily created. Entities and relations created in this manner are homogeneous in the ontological sense. This implies that several pathway collections can be combined in an ontology repository while maintaining a common semantic, meaning that the following steps can now be taken:

Sharing. Despite disparate initial data formats, the biochemical information described in each pathway is now homogeneously represented. This enables the direct reuse of componets (such as species or reactions) coming from different sources.

Integration. Our system ensures a consistent merging of the resources, automatically aligning the species and showing the end-user possible duplications among reactions in the different pathways.

Knowledge discovery. Once the species alignment is done and duplicate reaction have been detected, a new step is taken: for each reaction in each pathway the set of "alternative circuits" is computed. This means that given an arbitrary number of pathways, the system will identify all of the alternative ways to traverse from state S_0 to a state S_1 (where the states are different species config-

urations) within the overall set of reactions. In the last layer, all the information gathered is exported in OWL. With the OWL file we use the semantic tool, Protégé [11], to visually edit, compare, and finalize the biochemical information. With the OWL query interface, the user can now formulate "semantically-enabled" queries that were impractical when dealing with the previously heterogeneous, unaligned data repositories.

3 Usage Examples

OREMP in Combining Pathways for Parallel Solution. This system is embedded in the latest release of Cytosolve [12]. Its contribution to the integration of runnable pathways is the detection of duplicated reactions among different models. No matter the models chosen for simulation, once the species are aligned, the system identifies duplication problems in the reaction-models. From the user point of view this process is transparent: he/she receives a warning message that details the duplicated reactions and is prompted to confirm conflict elimination, and to resolve any differences in reaction kinetic rate constants.

OREMP in Querying Large, Independent Sources of Pathways. Our prototype was tested against the entire Biomodels.net curated collection [1] that contains about 240 molecular pathways. The result of the analysis was an overall view of the database and a list of about 500 groups of overlapping reactions. This analysis took 50 seconds on a single-core 2GHz Intel CPU. The previously described knowledge-discovery-step was taken on these resources as well. For each species configuration in the database, all alternative circuit paths were computed. This took about 2 hours on a quad-core 2GHz AMD CPU and resulted in a dictionary of thousands "biological equivalent" circuits (*i.e.*, equivalent reaction compositions). The latter experiment provides an interesting overview of the BioModels.net collection that we think can be used to boost the pathway modeling step - it provides a searchable dictionary of pathway building blocks. Perhaps more importantly, from the prospective of those who curate collections of biochemical pathways, this framework can be used to find inconsistencies and redundancies within their repository.

4 Conclusions

To our knowledge this is the first time that the information coming from different biological data sources are aggregated into a single quantitative ontology that can be queried at multiple levels. As detailed in previous sections, the OREMP application can combine several pathways, merge and combine pathway repositories, or revert to the original pathways, and inspect single-model details and query external repositories (such as UniProt and GO) referenced in pathway element annotations. Our system is independent of the different file formats in which the pathways are written and contains an extensible collection of parser modules. We have selected OWL as export format for the extended ontologies

and have adopted Protégé as our "Data Warehouse" for information storage, retrieval and reasoning. This framework transforms biomolecular pathways into extended ontologies to support knowledge sharing, integration and discovery. Since we generate the ontologies from a common semantics, the latter features are maintained when pathway collections are used to fill ontology repositories.

References

[1] Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J., Hucka, M.: BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research **34**(Database issue) (2006) D689–691

[2] Lloyd, C.M., Lawson, J.R., Hunter, P.J., Nielsen, P.F.: The CellML model repository. Bioinformatics **24**(18) (2008) 2122–2123

[3] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature Genetics **25**(1) (2000) 25–29

[4] The UniProt Consortium: The universal protein resource (UniProt). Nucleic Acids Research **35**(Database issue) (2007) D193–197

[5] Bauer-Mehren, A., Furlong, L.I., Sanz, F.: Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Molecular Systems Biology **5** (2009) 290–303

[6] Hucka, M., Finney, A., Sauro, H., Bolouri, H., Doyle, J., Kitano, H., and the rest of the SBML forum: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19**(4) (2003) 524–531

[7] Krause, F., Uhlendorf, J., Lubitz, T., Schulz, M., Klipp, E., Liebermeister, W.: Annotation and merging of SBML models with semanticSBML. Bioinformatics (2009) btp642

[8] Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research **37**(suppl_2) (2009) W170–173

[9] Lloyd, C., Halstead, M., Nielsen, P.: CellML: its future, present and past. Progress in Biophysics and Molecular Biology **85**(2-3) (2004) 433–450

[10] Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., Wanner, B.L.: Minimum information requested in the annotation of biochemical models (MIRIAM). Nature Biotechnology **23**(12) (2005) 1509–1515

[11] Noy, N.F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R.W., Musen, M.A.: Creating semantic web contents with protege-2000. Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications] **16**(2) (2001) 60–71

[12] Ayyadurai, S., Dewey, C.F.: Cytosolve: a scalable computational methodology for dynamic integration of multiple molecular pathway models. Cellular and Molecular Bioengineering (2010) In review