

Annotation of SBML Models Through Rule-Based Semantic Integration

Allyson L. Lister^{1,2*}, Phillip Lord², Matthew Pocock², and Anil Wipat^{1,2*}

¹Centre for Integrated Systems Biology of Ageing and Nutrition (CISBAN), Newcastle University, UK

²School of Computing Science, Newcastle University, UK

ABSTRACT

Motivation: The creation of accurate quantitative Systems Biology Markup Language (SBML) models is a time-intensive, manual process often complicated by the many data sources and formats required to annotate even a small and well-scoped model. Ideally, the retrieval and integration of biological knowledge for model annotation should be performed quickly, precisely, and with a minimum of manual effort. Here, we present a method using off-the-shelf semantic web technology which enables this process: the heterogeneous data sources are first syntactically converted into ontologies; these are then aligned to a small domain ontology by applying a rule base. Integrating resources in this way can accommodate multiple formats with different semantics; it provides richly modelled biological knowledge suitable for annotation of SBML models.

Results: We demonstrate proof-of-principle for this rule-based mediation with two use cases for SBML model annotation. This was implemented with existing tools, decreasing development time and increasing reusability. This initial work establishes the feasibility of this approach as part of an automated SBML model annotation system.

Availability: Detailed information including download and mapping of the ontologies as well as integration results is available from <http://www.cisban.ac.uk/RBM>.

1 INTRODUCTION

The integration of life sciences data remains an ongoing challenge. The multitude of data sources and formats which differ in both syntax and semantics makes this task difficult. Errors in data integration are possible when data sources do not describe their information with a shared semantics (Philippi and Köhler, 2006). The problems of and historical approaches to syntactic and semantic data integration have been well described (Sujansky, 2001; Alonso-Calvo et al., 2007). Though semantic data integration allows for a richer description of the biology than is possible with syntactic methods, semantic data integration in bioinformatics is difficult, partly due to the bespoke nature of the tooling.

An important part of many semantic data integration methods is *ontology mapping* between concepts in two or more ontologies (Lomax and McCray, 2004). Mediator-based approaches extend ontology mapping such that a *core ontology* is mapped to a large number of satellite source ontologies. Often, mediator-based approaches have viewed the purpose of a core ontology as simply a union of source ontologies rather than as a semantically-rich description of the research domain (Wache et al., 2001; Rousset and Reynaud,

2004). However, if a core ontology is defined merely as a model of a set of data sources, it becomes brittle with respect to the addition of new data sources and new formats.

To compensate for this limitation, we have created a method of semantic data integration called *rule-based mediation*. As with other mediator-based approaches, the data sources themselves are recast as *syntactic ontologies* which describe the syntax of the data formats. In contrast to previous approaches, the core ontology is a semantically-rich description of biological concepts. The richness of a core ontology depends on the type of biological questions that it has been created to answer; an over-engineered ontology can take longer to develop and may not provide better answers. Because a core ontology is abstracted away from the data formats, it does not need to be modified when adding new data sources.

A set of rules are then applied which map the syntactic ontologies into the biological concepts of the core ontology. We have used this method to extend a quantitative biological Systems Biology Markup Language (SBML) (Hucka et al., 2003) model. To achieve this, we have created a core ontology describing, in this case, telomere biology. We have then used three standard semantic web tools (XMLTab¹, SWRL² and SQWRL³) to integrate three heterogeneous data sources containing relevant information. This has allowed us to annotate an SBML model, including the provisional identification of knowledge not previously present in the model. This work demonstrates the feasibility of the approach, the applicability of the technology and its utility in gaining new knowledge.

2 USE CASES

In this paper, we consider two methods for enriching an existing model of telomere uncapping in *Saccharomyces cerevisiae* (Proctor et al., 2007). In Use Case 1, we annotate a single SBML species with information relevant to the gene RAD9 drawn. Adding information to existing SBML elements at an early stage aids model development and extension prior to publication or deposition. In Use Case 2, we retrieve possible protein-protein interactions involving RAD9. This approach resulted in the identification of model elements as well as a putative match for an enzyme that was not identified in the original curated model. These examples show how rule-based mediation works in a systems biology setting, as well as being representative of how this system might be extended for further automated model annotation.

¹ <http://protegewiki.stanford.edu/index.php/XML.Tab>

² <http://protege.cim3.net/cgi-bin/wiki.pl?SWRLTab>

³ <http://protege.cim3.net/cgi-bin/wiki.pl?SQWRL>

*to whom correspondence should be addressed

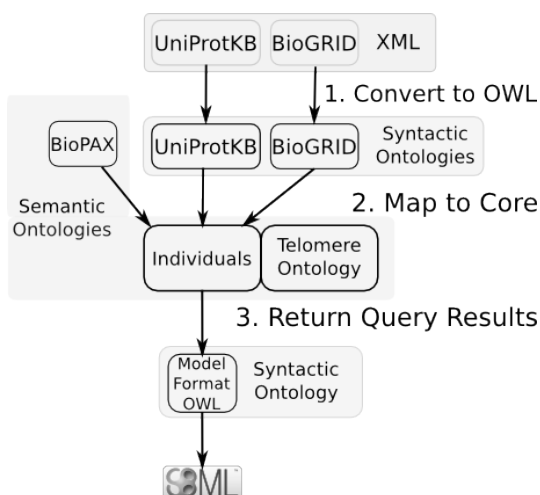


Fig. 1: Rule-based mediation in the context of SBML model annotation. Non-OWL formats are first converted into syntactic ontologies. Here, both UniProtKB and BioGRID data are in XML: UniProtKB has its own schema, while BioGRID uses PSI-MIF. These formats are converted into syntactic ontologies via the XMLTab. BioPAX, the format used for Pathway Commons, is already in OWL and needs no conversion. Next, the individuals present in the syntactic ontologies are mapped to the core ontology using SWRL. Finally, querying is performed using SQWRL queries using only core ontology concepts. Information is then passed through a final syntactic ontology (MFO) into an SBML model.

3 APPROACH

We chose the subset of Web Ontology Language (OWL) called OWL-DL⁴. While other ontology languages such as the Open Biomedical Ontologies (OBO) (Smith et al., 2007) are also widely used in the life sciences, they do not provide the same level of support for automated semantic reasoning (Golbreich et al., 2007). This reasoning can draw explicit conclusions from knowledge that is, otherwise, only implicit in the ontology.

Figure 1 shows rule-based mediation in the context of model annotation. We have used a combination of existing tools (XMLTab, SWRL and SQWRL) and novel mappings to implement rule-based mediation: first, syntactic conversion of information into OWL; second, semantic mapping of that information into a core ontology; and finally, querying of that core ontology to answer specific biological questions. Further, with additional mappings from the core ontology back out to a syntactic ontology, the information retrieved can be formatted for SBML (and, potentially, any known data source). From a biological perspective, rule-based mediation produces an integrated view of information useful for modelling.

Three data sources were used in the context of SBML model annotation: BioGRID (Stark et al., 2006), Pathway Commons⁵, and UniProtKB (The UniProt Consortium, 2008). For each data format, a suitable syntactic ontology was either created or identified. An

⁴ <http://www.w3.org/TR/owl-ref/>

⁵ <http://www.pathwaycommons.org>

additional syntactic ontology, Model Format OWL (MFO), was adapted from our previous work (Lister et al., 2007). MFO stores constraints from SBML, the Systems Biology Ontology and the SBML manual.

Basic information on the data formats as well as the numbers of axioms, relations and mappings in these syntactic ontologies is available in Table 1.

Unlike the syntactic ontologies, which are designed to be OWL-DL representations of the underlying data formats, a core ontology is an explicit description of the semantics of the research domain. The core ontology created for these use cases is a *telomere ontology*, which models the biology relevant to the Proctor *et al.* model.

4 RESULTS

Rule-based mediation was performed for two use cases to show proof-of-principle in the context of SBML model annotation. Example SWRL mappings for these use cases is available in Table 2.

In Use Case 1, we annotate the query gene RAD9 for *S.cerevisiae*. This query over the core telomere ontology is described in Figure 2a.

$$\begin{array}{ll}
 Q(X) : Protein(X) \sqcap & Q(X) : Reaction(X) \sqcap \\
 (hasName(X, Y) \sqcup & hasParticipant(X, Y) \sqcap \\
 hasSynonym(X, Y)) \sqcap & playedBy(Y, W) \sqcap \\
 rad9(Y) \sqcap & (hasName(W, Z) \sqcup \\
 hasTaxon(X, Z) \sqcap & hasSynonym(W, Z)) \sqcap \\
 NCBI4932(Z) & rad9(Z) \sqcap \\
 & hasTaxon(W, V) \sqcap \\
 & NCBI4932(V)
 \end{array}$$

(a) Query for Use Case 1. (b) Query for Use Case 2.

Fig. 2: Queries for the Use Cases.

This use case demonstrates the addition of basic information to the species such as cross-references, Systems Biology Ontology annotations, compartment localisations and a recommended name. Initially, the query against the telomere ontology found matches to individuals which have 'rad9' as a recommended name or synonym. All individuals equivalent to these matches will also be retrieved. While the query term 'rad9' was not present in BioGRID, the appropriate BioGRID individual had already been marked as equivalent to both the UniProtKB and Pathway Commons individuals that shared the same UniProtKB primary accession.

In Use Case 2, information about possible protein-protein interactions involving RAD9 was requested, resulting in the proposal of new model elements as well as the putative identity of the enzyme responsible for activation of RAD9. Figure 2b is the query for Use Case 2 and builds on the result of Figure 2a. Table 3 shows a summary of the results from this query. Already existing as well as novel biological information for the curated model was discovered. Firstly, the RAD53 and CHK1 interactions with RAD9 from the curated model were confirmed. There are a total of four interactions involving RAD9 in the model: the other two RAD9 interactions are present in the model as placeholder species, created

| Data Source | Data Format | Classes | Class Axioms | Relations | Relation Axioms | DL Expressivity | Mappings |
|-----------------|---------------|---------|--------------|-----------|-----------------|-----------------|----------|
| UniProtKB | UniProtKB XML | 27 | 122 | 58 | 0 | $ALEN(D)$ | 9 |
| BioGRID | PSI-MIF | 26 | 126 | 39 | 0 | $ALEN(D)$ | 17 |
| SBML | SBML | 474 | 572 | 16 | 57 | $SHQ(D)$ | N/A |
| Pathway Commons | BioPAX | 41 | 193 | 70 | 145 | $ALCHN(D)$ | 11 |

Table 1. Basic information about the syntactic ontologies and their mappings to the core ontology. Pathway Commons differs from the other sources as its format, BioPAX, is represented in OWL. BioPAX therefore function as a syntactic ontology. The 'Mappings' column lists the number of SWRL mapping statements used to link each syntactic ontology with the core ontology. For the SBML syntactic ontology, the numbers are combined totals of the imported Systems Biology Ontology and MFO. MFO was used for export only, and therefore its input mappings are marked with "N/A". Statistics generated using Protégé 4 (<http://protege.stanford.edu/>).

| Rule Number | SWRL Mappings from the PSI-MIF syntactic ontology to the core telomere ontology |
|--------------|--|
| PSIMIF.00008 | $\text{psimif:interactor}(?i) \wedge \text{psimif:id}(?i, ?id) \wedge \text{psimif:participant}(?p) \wedge \text{psimif:interactorRef}(?p, ?id) \rightarrow \text{tuo:plays}(?i, ?p)$ |
| PSIMIF.00015 | $\text{psimif:interactor}(?x) \wedge \text{psimif:interactorTypeSlot}(?x, ?t) \wedge \text{psimif:namesSlot}(?t, ?n) \wedge \text{psimif:fullName}(?n, ?value) \wedge \text{swrlb:equal}(?value, \text{"protein"}) \rightarrow \text{tuo:Protein}(?x)$ |

Table 2. Example rule mappings for the PSI-MIF syntactic ontology to the core telomere ontology, using the syntax displayed within SWRLTab. Both instance (such as `psimif:interactor` to `tuo:ProteinComplexFormation`) and relation (such as `psimif:organismSlot` to `tuo:hasTaxon`) mappings may be constrained by filters, such as the restrictions on what type of `psimif:interactor` can be linked to `tuo:Protein` from PSIMIF.00015. The SWRL Rule PSIMIF.00008 describes the relation between a `psimif:interactor` and its `psimif:participant`. This relation is not explicitly named within PSI-MIF and therefore chaining is used to map this information to the `tuo:plays` relation. This chain links the `psimif:id` of the `psimif:interactor` with the `psimif:interactorRef` of the `psimif:participant` even though both are datatype properties.

by the modeller to describe an unknown protein could not be further matched. The second main result was the provisional identification of one of those placeholder species, marked as 'Rad9Kin' in the model, as the protein MEC1. We describe how these conclusions were drawn next.

Within the telomere ontology the location of a protein (such as RAD9 within the nucleus) is transferred to any of its reactions. Therefore, both the RAD53 and CHK1 interactions are also located within the nucleus. Further, the product of a protein complex formation in the telomere ontology must be a protein complex. Specifically the RAD9/RAD53 reaction, present both in the integration results and the curated model, is part of a BioGRID 'Reconstituted Complex'. This knowledge of the reaction type can be presented as $\text{rad9} + \text{rad53} \leftrightarrow \text{rad9rad53Complex}$. While not identical to the curated reaction, which shows activation of RAD53 by RAD9 ($\text{rad9Active} + \text{rad53Inactive} \rightarrow \text{rad9Active} + \text{rad53Active}$), the proposed interaction is potentially informative.

One of the unknown interactors in the curated model, 'rad9Kin', was provisionally identified within the telomere ontology. This unknown species does not have a UniProtKB primary accession in the curated model, as the model author did not know which protein activated RAD9. However, MEC1 is shown in the telomere ontology as interacting with RAD9 and is present in the curated model reactions. Further, UniProtKB reports it as a kinase which phosphorylates RAD9. From this information, the model author now believes that MEC1 could be the correct protein to use as the activator of RAD9 (Proctor, 2009).

These use cases show that rule-based mediation successfully integrates information from multiple sources using existing tools, and that it would be useful to expand the implementation of this method to larger biological questions.

5 DISCUSSION

We have demonstrated that rule-based mediation and its implementation for our use cases is a suitable method for semantic data integration in the context of model annotation. We have utilised existing tools wherever possible to implement our approach. This proof-of-principle implementation has reproduced reactions present in a curated SBML model and suggested a potential interactor where the identity of that interactor was unknown.

Previous work on ontology mapping as well as semantic data integration includes the mediator-based approaches mentioned earlier as well as mapping GO to UMLS (Lomax and McCray, 2004), creating databases using RDF with S3DB (Deus et al., 2008) and OntoFusion (Alonso-Calvo et al., 2007). OntoFusion uses an approach similar to rule-based mediation, however it lacks a core ontology. This removes the opportunity for further semantic processing that a core ontology provides. Integration techniques such as TAMBIS (Stevens et al., 2000) used only a core ontology, needing tools to map both the semantics and syntax of the underlying data sources into the ontology; additionally, our rule-based mediation allows an arbitrary number of data sources to provide the same type of information. Rule-based mediation builds on these earlier methods by providing ontology mapping combined with a semantically- and biologically-rich core ontology.

There are a number of improvements planned, with the ultimate result being a fully-automated model annotation system. Some imminent advances, such as the `hasKey`⁶ construct present in OWL2, will allow further automation. This language feature allows the definition of equivalence rules based on properties such as

⁶ http://www.w3.org/TR/2008/WD-owl2-new-features-20081202/#F9:_Key

| Discovered Interaction Partner with P14737 | Proctor et al. Model | BioGRID | Pathway Commons |
|--|----------------------|---------|-----------------|
| Serine/threonine-protein kinase RAD53 (P22216) | ✓ | ✓ | ✓ |
| Serine/threonine-protein kinase CHK1 (P38147) | ✓ | | ✓ |
| Serine/threonine-protein kinase MEC1 (P38111) | | ✓ | |
| DNA damage checkpoint control protein RAD17 (P48581) | | ✓ | |
| Rad9Kin (*) | ✓ | | |
| ExoX (*) | ✓ | | |

Table 3. Partial summary of interactions retrieved for Use Case 2 against the core telomere ontology. All discovered interactions already present in the curated model are shown, together with example interactions from the curated model that were not discovered (with Rad9Kin and ExoX) and discovered interactions that were *not* present in the model (with MEC1 and RAD17). SBML species shown with an asterisk (*) are those which are placeholder species, and therefore cannot have a match to a real protein. Some interactions are false positives inherent in the data source, while others are out of scope of the modelling domain of interest and should not be included (see supplementary material).

database accessions, which in the current system were achieved by manual mapping. The filtering of false positive or out-of-scope interactions will also be improved. Scaleability is a further issue to be addressed, possibly through a database back-end for OWL. We are currently increasing the number of syntactic ontologies, the scope of the core telomere ontology and the mapping coverage. Finally, while SWRL mapping connects individuals in a syntactic ontology to individuals in a core ontology, linking source individuals to classes in a core ontology is being investigated.

6 CONCLUSION

We have created a new method of semantic integration called rule-based mediation, which makes use of a semantically-rich core ontology together with mappings from syntactic ontologies. We have shown that rule-based mediation is an effective approach for model annotation. Syntactic ontologies for the UniProtKB and the PSI-MIF formats were created, while BioPAX was used without modification. Additionally, a telomere ontology was developed to model the biology associated with the use cases. The use of existing tools decreased development time and increased the applicability of this approach for other projects.

There are many advantages to the rule-based mediation approach. A syntactic ontology can be used for both import and export. New formats can be easily added without modifying the core ontology. Additions to a core ontology are simple: each new mapping, class, or data import is incremental, without needing large-scale changes.

Rule-based mediation takes into account the semantics of the underlying biology rather than just the syntax in which the biological data is stored. We have illustrated the utility of this method in a specific application domain by reproducing interactions already present in a curated model and suggesting a putative identity for an unknown species in that model. Future work will use this approach as the core of an automated semantically-aware model annotation system.

ACKNOWLEDGEMENTS

We acknowledge the support of the Newcastle University Systems Biology Resource Centre and the Newcastle University Bioinformatics Support Unit.

Funding: ALL and AW are supported by the BBSRC/EPSRC (ref BB/C008200/1), MP by ref BB/F006063/1.

REFERENCES

- R. Alonso-Calvo et al. An agent- and ontology-based system for integrating public gene, protein, and disease databases. *J Biomed Inform*, 40(1):17–29, February 2007. ISSN 1532-0480. doi: 10.1016/j.jbi.2006.02.014.
- H. F. Deus, R. Stanislaus, D. F. Veiga, C. Behrens, I. I. Wistuba, J. D. Minna, H. R. Garner, S. G. Swisher, J. A. Roth, A. M. Correa, B. Broom, K. Coombes, A. Chang, L. H. Vogel, and J. S. Almeida. A semantic web management model for integrative biomedical informatics. *PLoS ONE*, 3(8), 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002946.
- C. Golbreich et al. OBO and OWL: Leveraging Semantic Web Technologies for the Life Sciences. In *The sixth International Semantic Web Conference (ISWC 2007)*, pages 169–182. 2007. doi: 10.1007/978-3-540-76298-0_13.
- 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, March 2003. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg015.
- A. L. Lister, M. Pocock, and A. Wipat. Integration of constraints documented in SBML, SBO, and the SBML Manual facilitates validation of biological models. *Journal of Integrative Bioinformatics*, 4(3):80+, 2007.
- J. Lomax and A. T. McCray. Mapping the gene ontology into the unified medical language system. *Comparative and functional genomics*, 5(4):354–361, 2004. ISSN 1531-6912. doi: 10.1002/cfg.407.
- S. Philippi and J. Köhler. Addressing the problems with life-science databases for traditional uses and systems biology. *Nature Reviews Genetics*, 7(6):482–488, May 2006. ISSN 1471-0056. doi: 10.1038/nrg1872.
- C. J. Proctor. personal communication, 2009.
- C. J. Proctor et al. Modelling the checkpoint response to telomere uncapping in budding yeast. *Journal of The Royal Society Interface*, 4(12):73–90, February 2007. doi: 10.1098/rsif.2006.0148.
- M. C. Rousset and C. Reynaud. Knowledge representation for information integration. *Inf. Syst.*, 29(1):3–22, 2004. ISSN 0306-4379. doi: 10.1016/S0306-4379(03)00032-2.
- B. Smith et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, November 2007. doi: 10.1038/nbt1346.
- C. Stark et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue), January 2006. ISSN 1362-4962. doi: 10.1093/nar/gkj109.
- R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. W. Paton, C. A. Goble, and A. Brass. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics (Oxford, England)*, 16(2):184–185, February 2000. ISSN 1367-4803.
- W. Sujansky. Heterogeneous database integration in biomedicine. *J Biomed Inform*, 34(4):285–298, August 2001. ISSN 1532-0464. doi: 10.1006/jbin.2001.1024.
- The UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res*, 36(Database issue), January 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm895.
- H. Wache et al. Ontology-based integration of information — a survey of existing approaches. In H. Stuckenschmidt, editor, *Proceedings of the IJCAI'01 Workshop on Ontologies and Information Sharing, Seattle, Washington, USA, Aug 4-5*, pages 108–117, 2001.