How can semantic annotations support the identification of network similarities?

Christian Rosenke and Dagmar Waltemath

Department of Computer Science, University of Rostock
christian.rosenke|dagmar.waltemath@uni-rostock.de

Abstract. Computational models in open model repositories support biologists in understanding and investigating biological questions. The availability of alternative models results in a need for model selection algorithms. This selection can be based on information retrieval search, full-text search, selection by ontology concepts etc. We here describe our approach to solving a serious aspect of model selection, that is, the problem of comparing models of reaction networks. Specifically, we discuss how graph algorithmic approaches can help to compare models that are semantically enriched by annotations. While a graph comparison of naked models is infeasible, the knowledge gained from the semantic annotations and domain specific structures can reduce the complexity. Our concept has the potential to improve model search, and it can contribute to the definition of similarity measures.

1 Introduction

Modeling is a state-of-the-art method in systems biology research, and its importance for experimental studies and result analyses is steadily increasing. Computational models describe biological systems, which are analysed and often simulated, to gain further knowledge about a system under study, or to predict future experiments. Usually, models are provided to the research community in XML standard formats such as SBML [1] or CellML [2]. These XML-based encodings are annotated with terms from bio-ontologies [3] to enable model visualisation, comparison and search, and thereby improve model reuse. See Figure 1 for an example of an SBML encoded and annotated model on the cell cycle. Models are distributed via open repositories such as the BioModels Database [4], JWS Online [5], or the CellML model repository [6].

Understanding and further developing models remains a major challenge in biology today due to the ambiguity and varying levels of abstraction in model descriptions. Other aspects that complexify the study of models are the increasing number and size of models [7]. Even for well-annotated models, basic operations, such as identifying suitable submodels and comparing models, can hardly be handled with or without computational support.

In this work, we outline a path towards efficient computational methods that assist scientists in finding models and comparing them to each other. Many models are described by reaction networks, which are built from nodes for reactants,

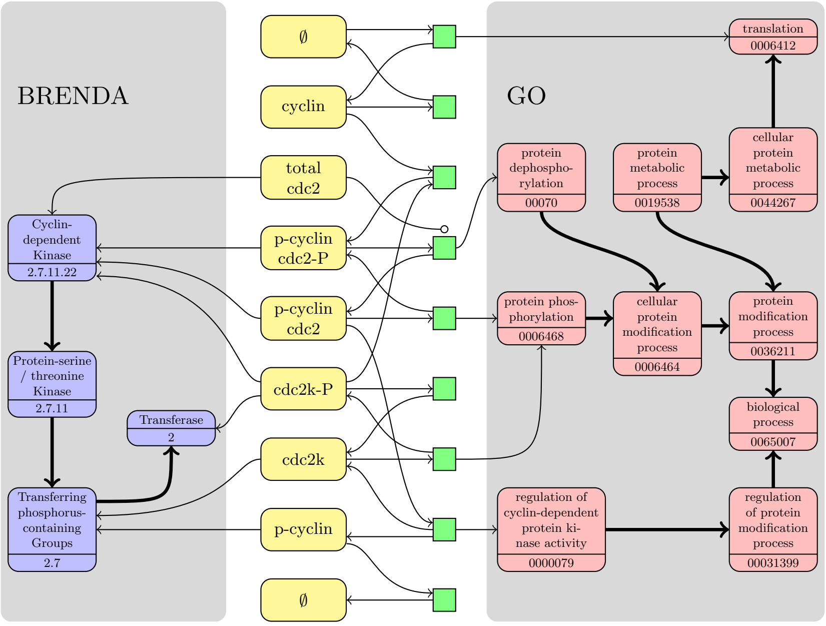
BRENDA
GO
∅
cyclin
total cdc2
p-cyclin cdc2-P
p-cyclin cdc2
cdc2k-P
cdc2k
p-cyclin
∅
Cyclin-dependent Kinase
2.7.11.22
Protein-serine / threonine Kinase
2.7.11
Transferase
2
Transferring phosphorus-containing Groups
2.7
translation
0006412
protein dephosphorylation
00070
protein metabolic process
0019538
cellular protein metabolic process
0044267
protein phosphorylation
0006468
cellular protein modification process
0006464
protein modification process
0036211
biological process
0065007
regulation of cyclin-dependent protein kinase activity
0000079
regulation of protein modification process
00031399

Fig. 1. The cell cycle's model by Tyson [8] from the BioModels Database `http://www.ebi.ac.uk/biomodels-main/BIOMD0000000005` coming with semantic information via annotation links to the ontologies BRENDA (blue) und GO (red). The biological system is described by reactions (green nodes) and chemical entities (yellow nodes) which are connected by directed arcs to express the roles of reactants, products, and modifiers. Ontology terms (blue and red nodes) are given together with their id and hierarchy in the respective ontology.

reaction products and modifiers. The goal is to extend current search functionality by a model similarity measure that compares these network structures using graph algorithmic approaches.

Similar works have already developed algorithms that calculate model similarities, as for instance in. Proposed solutions either use information retrieval to calculate similarities between models [7]; XML diff algorithms for difference detection within versions of a model [9]; focus on the similarities of semantic annotations of model entities [10]; or use network similarity approaches [11,12]. However, none of them fully integrates available, domain-specific information and graph-based approaches in one similarity measure. Moreover, existing algorithms provide heuristic or approximate solutions that do not yet represent real alternatives for the still common manual way of processing models. To master the demanding challenges introduced by the contemporary way of working with models, we incorporate knowledge about domain characteristics and so-called semantic annotations, which have so far not been considered together with the structural composition of networks.

2 Results

This work is tailored towards models encoded in standard representation formats as SBML and CellML. The key concepts to counter the severe computational hardness of comparing “naked” graphs are to incorporate (1) information on structural constraints common to networks of biological models and (2) knowledge from semantic annotations.

Incorporating the model structure can tremendously reduce the complexity of the proposed graph-based correlation procedures. This is because networks in models for biological systems are subject to certain structural restrictions simply by originating from real world chemical and physical processes. Chemical reactions, for instance, do almost never incorporate more than two or three reactants and products. Beside this general observation, more complex limits to biological reaction networks can be found. For example, certain pattern, called motifs [13], tend to appear regularly in many biological networks.

Incorporating semantic annotations enhances or reduces the probability of matching nodes based on the linked ontology concepts. Many models are accurately annotated, including information about the biological meaning of model entities, about the roles that entities play in reactions, about the types of reactions, or about the nature of a parameter. We use these links to concepts in external ontologies to determine similarities between model entities. If two entities carry the same, or a similar annotation, and they structural restrictions apply, then our algorithm boosts the respective similarity value.

3 A First Graph-Based Approach using Annotations

To support our argumentation in this work we present a first graph-based approach where knowledge about semantic annotations is used to obtain an efficient solution. More precisely, we want to efficiently answer the question, if a query model Q is a submodel of another model M. Thus we look for an injective mapping σ of the nodes from Q to those from M such that the edges are matched.

Computationally, this is extremely difficult in general. But using the links into the ontologies, made available by the semantic annotations, can create valuable connections between Q and M which essentially reduce the possible degree of freedom for the mapping σ. Our approach works, if nearly every node in Q has at most two possible candidate target nodes in M. Although this is a strong restriction, our algorithm is more general and powerful than many other approaches.

Basically, we compute a formula in 2-cnf that describes the valid mappings σ by its satisfying assignments, that is, we reduce the problem to 2-SAT, a tractable version of the satisfiablity problem from proposal logic. For every node q from Q and every possible target node m of M we introduce a variable m_q which is *true* if and only if $\sigma(q) = m$. As q has at most two targets m, m' we are able to include the clauses $(m_q \vee m'_q)$ and $(\neg m_q \vee \neg m'_q)$ to express that $\sigma(q)$ is either m or m'. To make σ injective, we have to state for every node m in

M that at most one node from Q maps to m. For this end, we add the clause $(\neg m_q \vee \neg m_{q'})$ for every pair q, q' of Q-nodes that both possibly map to m. As σ has to preserve the adjacency of nodes, we add the clause $(\neg m_q \vee \neg m'_{q'})$ for all Q nodes q, q' with possible target nodes m, m' if the adjacency between q and q' in Q is different from the adjacency between m and m' in M.

The length of the resulting 2-cnf formula is at most quadratical in the size of Q and widely independent of M. This easily makes queries Q of up to hundreds of nodes feasible.

4 Summary

This work presents first thoughts on using a combination of algorithms for graph similarity and semantic annotations as a mean to map simulation models describing biological systems. The concept will be implemented in a data management system that supports the management and linking of model files, ontologies used to semantically enrich the model files, and all data that needs to be accessed and stored during and after the computation of network similarities [7]. Our concept for model comparison can improve model search, and it can contribute to the definition of similarity measures. It can also be used to identify overlapping parts in a network, for example between different versions of a model [9].

References

1. Hucka *et al.*: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.. *Bioinformatics* 19.4:524-531, 2003.
2. Cuellar *et al.*: An overview of CellML 1.1, a biological model description language. *Simulation* 79.12, 2003.
3. Horridge *et al.*: The state of bio-medical ontologies. 2011.
4. Li *et al.*: BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 2010.
5. Brett and Snoep: Web-based kinetic modelling using JWS Online *Bioinformatics* 20.13:2143-44, 2004
6. Lloyd *et al.*: The CellML Model Repository. *Bioinformatics*, 2008.
7. Henkel *et al.*: Ranked retrieval of computational biology models. *BMC bioinformatics* 11.1:423, 2010.
8. Tyson: Modeling the cell division cycle: cdc2 and cyclin interactions. *Proc. Natl. Acad. Sci.* 88(16):7328-7332, 1991.
9. Waltemath *et al.*: Improving the reuse of computational models through version control. *Bioinformatics* 29.6:742-748, 2013.
10. Schulz *et al.*: Propagating semantic information in biochemical network models. *BMC Bioinformatics* 13(1), 2012.
11. Pržulj: Biological network comparison using graphlet degree distribution. *Bioinformatics* 2(23):177–183, 2007.
12. Gay *et al.*: A graphical method for reducing and relating models in systems biology. *Bioinformatics* 18(26):i575–i581, 2010.
13. Tyson and Novak: Functional motifs in biochemical reaction networks. *Annual review of physical chemistry* 61:219, 2010.