A Cell-Cycle Knowledge Integration Framework
Research Paper

Erick Antezana, Elena Tsiporkova, Vladimir Mironov, and Martin Kuiper

Dept. of Plant Systems Biology. Flanders Interuniversity Institute for Biotechnology/Ghent University. Technologiepark 927, B-9052 Ghent Belgium
{erant, eltsi, vlmir, makui}@psb.ugent.be
http://www.psb.ugent.be/cbd/

Abstract. The goal of the EU FP6 project DIAMONDS [1] is to build a computational platform for studying the cell-cycle regulation process in several different (model) organisms (S. cerevisiae, S. pombe, A. thaliana and human). This platform will enable wet-lab biologists to use a systems biology approach encompassing data integration, modeling and simulation, thereby supporting analysis and interpretation of biochemical pathways involved in the cell cycle. To facilitate the computational handling of cell-cycle specific knowledge a detailed cell-cycle ontology is essential. The currently existing cell-cycle branch of the Gene Ontology (GO) provides only a static view and it is not rich enough to support in-depth cell-cycle studies.

In this work, an enhanced Cell-Cycle Ontology (CCO) is proposed as an extension to existing GO. Besides the classical add-ons given by an ontology (data repository, knowledge sharing, validation, annotation, and so on), CCO is intended to further evolve into a knowledge-based system that provides reasoning services oriented to hypotheses evaluation in the context of cell-cycle studies. A data integration pipeline prototype, covering the entire life cycle of the knowledge base, is presented. Concrete problems and initial results related to the implementation of automatic format mappings between ontologies and inconsistency checking issues are discussed in detail.

1 Introduction

The amount of data generated in biological experiments continues to grow exponentially. The shortage of proper approaches or tools for analyzing this information has created a gap between raw data and knowledge. To make matters worse, the lack of a structured documentation of knowledge leaves much of the information extracted from these raw data unused. Moreover, differences in the used technical languages (synonymy and polysemy) have complicated the analysis and interpretation of the data. Currently, there are several efforts for standardizing the used vocabulary. Most importantly, the Gene Ontology (GO) Consortium [9] has been providing a controlled set of terms for gene products whereas the Open

[1] http://www.sbcellcycle.org

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 19–34, 2006.

Biomedical Ontology(OBO)[2] umbrella has been collecting the most representative ontologies in biological and medical domains. Ontologies clarify scientific discussions providing a shared vocabulary for biologists to communicate their results effectively, explore data and extend scientific investigations. Ontologies also facilitate the implementation of computational approaches and systems to perform data exploration, inference and mining [5].

The goal of the EU FP6 project DIAMONDS is to build and use a systems biology platform of tools to study the cell-cycle process in several different model organisms (S. cerevisiae, S. pombe, A. thaliana and human). Data and information integration and retrieval is essential for studying gene networks, and although several solutions for this already exist (e.g. BioRS[3], SRS[4]; also some ontology-based solutions like TAMBIS [25] and caBIO [8]. A particular challenge is the development of a specific cell-cycle ontology (CCO), as this is relatively poorly developed at present. A rich CCO will be a first step towards more powerful computational approaches to exploit such developed ontology. The process of cell division, or cell cycle, is one of the most fundamental and highly conserved processes in eukaryotic systems. Its cyclical nature makes it a challenging phenomenon for modeling and simulation and a better understanding of it provides significant knowledge for growth in general and human health in particular (cancer related aspects, proliferation disorders issues, prospective therapeutic targets basis and so forth [14], [29]). The available knowledge contained in the cell-cycle literature, however, resides in a format that does not enable straightforward computational processing and consequently, searching and manipulating this information is limited. Moreover, reusing and sharing cell-cycle related data is not facilitated by actual media. Queries within a document are usually limited to simple keyword searches. Therefore, relations between concepts within a document cannot be found unambiguously. For example, two instances, protein X and protein Y can be easily identified by a keyword search. However, unless biologists read at least the text sections comprising those concepts within the document, they will not be able to determine whether these two proteins are related to each other, how this relationship is defined, or in what particular phase of the cell-cycle this relationship is important.

We propose here an ontological paradigm that enables to capture the semantics, temporal aspects and dynamics of the cell cycle regulatory process. Currently, the cell-cycle branch from the bio-ontology GO is too basic to adequately describe the cell-cycle, as it only supports a static view of this process. GO is based on the annotation of gene products (either RNA or proteins). Each of these products may in fact play a role in many molecular processes. Unfortunately, in GO only the prospective activity of a given process is defined without much specification of where or when this process may take place. For particular applications, such as regulatory network modeling and simulation, it is essential to access specific temporal annotations that capture the dynamics of the

[2] http://obo.sourceforge.net/cgi-bin/table.cgi
[3] http://www.biomax.de/products/biors.php
[4] http://www.biowisdom.com/solutions_srs.htm