# 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

system. Only two types of relationships are at present considered in GO: subsumption (*is_a*) and partonomic inclusion (*part_of*) (for a formal definition of an ontology structure, refer to the Appendix), which poses a significant limitation for expressing the semantics of a dynamical system. In addition, GO treats its three structured networks as separate ontologies, i.e. no ontological relations are defined among them. Besides, GO suffers of inconsistent treatment of relations such as *is_a*. In spite of these problems, GO has gained a wide appreciation in the life sciences.

The CCO that we propose here belongs to the domain specific ontology type according to the definition given in [12]. As argued in [19], the development of an ontology of a given domain is frequently not a goal in itself, it rather constitutes a skeleton for a set of data that together form a knowledge base. We have set out to build a knowledge-based system founded on CCO, for an in-depth analysis of cell-cycle control mechanisms.

There are several prospective resources that a cell-cycle knowledge base can draw on. Among them, existing ontologies such as GO and some of the ones listed at the OBO repository are key. In addition, databases holding data about gene/protein interactions, such as Reactome [15], BIND [2] and IntAct [13], are also considered. Cell-cycle "slims" from Reactome will provide the first setup. Furthermore, data produced by the DIAMONDS consortium will also feed the repository (E.g. dedicated curation of literature information, annotation information on protein features, protein-protein interaction data).

OWL [21] is a web ontology language that is recommended by the W3C[5] consortium for semantic web applications. OWL comes in three flavors: OWL Full, OWL-DL and OWL Light, ranked in order of their expressivity. For CCO we chose OWL-DL, because of the reasoning capabilities versus computational cost ratio.

Reasoning through a logic approach is best able to deal with the constraints of the gathered knowledge. We have chosen description logics [1] because of its expression power, a well developed theory and consistent semantics. Reasoning packages, such as RACER[6], KAON2[7], Pellet[8] and/or FaCT++[9] are being used for classifying, checking instance consistency and making implicit information explicit. In addition, such reasoning can reveal inconsistencies, hidden dependencies, redundancies and misclassifications. As a result, the CCO becomes more robust.

## 2 Data Integration Pipeline

A formal specification of a data integration pipeline has been developed (see Figure 1). This specification covers the entire life cycle including the development

[5] http://www.w3.org/
[6] http://www.racer-systems.com/index.phtml
[7] http://kaon2.semanticweb.org/
[8] http://www.mindswap.org/2003/pellet/index.shtml
[9] http://owl.man.ac.uk/factplusplus/

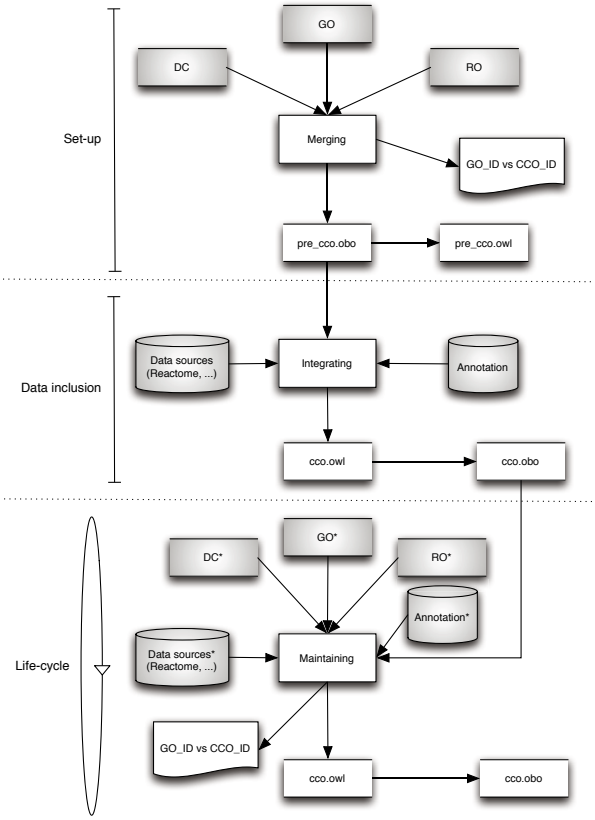


**Fig. 1.** Data integration pipeline

of the knowledge base. It has three phases: a set-up, a data integration phase, and a maintenance phase. A detailed description of these phases is presented below.

### 2.1 Set-Up

In the initial phase, the ontology structure and its lexicon (for formal definitions, refer to the Appendix) were engineered. The ontology structure core was based on the GO cell-cycle branch, the Relations Ontology (RO) [24], and the Dublin Core (DC)[10] ontology. The integration pipeline has been implemented in PERL using go-dev[11] and XML::Parser[12]. In order to produce the CCO, ontology pruning and preprocessing has also been done. Whereas this ontology structure has been iteratively refined, changes can still be accommodated until a stable version of

[10] http://www.dublincore.org/
[11] http://www.godatabase.org/dev/doc/go-dev-doc.html
[12] http://search.cpan.org/dist/XML-Parser/

the whole system is reached. The output generated in this phase constitutes the input for the data integration phase.

As the ontology is available in OBO and OWL formats, specific format conversion tools were developed. The OBO format of the CCO is compliant with the version 1.2[13] of the OBO format specification. Since the CCO's ontology structure and lexicon are based on the cell-cycle GO branch, an association table (GO identifier versus CCO identifier) was defined. A CCO entry sample in both OBO and OWL formats showing the correspondence between attributes can be seen in Figure 2.

```
[Term]
id: CCO:P0000016
name: M phase of mitotic cell cycle
def: "Progression through M phase, the part of the mitotic cell cycle during which mitosis and cytokinesis
take place." [GOC:mah, ISBN:0815316194]
xref: GO:0000087
xref: Reactome:68886
relationship: part_of CCO:P0000037
is_a: CCO:P0000038
synonym: "M-phase of mitotic cell cycle" [] {scope="exact"}
```

```
<owl:Class rdf:ID="CCO_P0000016">
      <rdfs:label xml:lang="en">M phase of mitotic cell cycle</rdfs:label>
      <xref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">GO:0000087</xref>
      <xref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Reactome:68886</xref>
      <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Progression through M phase, the
part of the mitotic cell cycle during which mitosis and cytokinesis take place.</rdfs:comment>
      <synonym rdf:datatype="http://www.w3.org/2001/XMLSchema#string">M-phase of mitotic cell cycle</synonym>
      <rdfs:subClassOf>
            <owl:Restriction>
                  <owl:onProperty rdf:resource="#part_of"/>
                  <owl:someValuesFrom rdf:resource="#CCO_P0000037"/>
            </owl:Restriction>
      </rdfs:subClassOf>
      <rdfs:subClassOf rdf:resource="#CCO_P0000038"/>
      <owl:disjointWith rdf:resource="#CCO_P0000270"/>
</owl:Class>
```

**Fig. 2.** A CCO entry sample

### 2.2 Data Integration

Data coming from several different sources are being integrated into the knowledge base structure. The main sources are the available gene association files[14], the cell-cycle related data from the Reactome knowledge base and the IntAct database. The output of this phase provides the first release of the system. Besides, cross-references to external public databases, such as UniProt [6] and GenBank [7], will also be supplied in the future.

### 2.3 Maintenance

The system will support automatic data updates and facilitate targeted manual checking to ensure its consistency. An automatic updating procedure is partially implemented. This will guarantee that the working version of the CCO has the latest data produced by the GO consortium with respect to the cell-cycle data. The cell-cycle knowledge base specifies a given concrete domain state and as

[13] http://www.godatabase.org/dev/doc/obo_format_spec.html
[14] http://www.geneontology.org/GO.current.annotations.shtml

a result, the knowledge base should be constantly updated. In addition, the ontology structure may also be extended and refined. Regarding the update mechanism, the cutting-edge versions of the GO and RO ontologies are fetched by the updating scripts from their version control systems. Then, the checked-out files, in OBO format, are transformed into OWL format. Due to the fact that the accession numbers should remain consistent from update to update, special considerations are taken into account while updating the entries since some terms may have been changed (see section 3.2 for more details). In addition, separate files containing the CCO accession numbers and terms are kept in similar ways as the GO distribution is provided. On the other hand, a version control system has been set up for CCO since the initial results were obtained. The current CCO version as well as previous ones may be retrieved from that repository. In the near future, a web site will provide access to the initial and experimental CCO pre-releases as well as the stable releases.

## 3 Towards a Formal Cell-Cycle Ontology

Information retrieval and management is enhanced by means of ontologies. The next generation of the web, the semantic web [4], will provide to users and machines a common exchange language avoiding irrelevant search results and increasing the quality of the result searching hits. Such a common and formal language will assist in the maintenance of the daily evolution of the knowledge by checking the incongruities that might arise from periodic updates.

### 3.1 Motivation and Design Principles

In order to develop an efficient ontology, careful consideration should be given to the purpose that it should serve, and the scientific community that will use it. The main purpose for building the CCO is to capture the semantics of the cell-cycle regulatory process, especially the dynamic aspects of the concepts and their interrelations, and to promote sharing, reuse and enable better computational integration with existing resources. The prospective audience comprises both wet-lab biologists and computational biologists who might have a particular interest in cell-cycle elucidation. Some motivating scenarios or use-cases that have shaped the CCO development are shown in Table 1.

The scope of the ontological concepts ranges from biological processes, cellular components, and molecular functions. Some competency questions that should be answered by the system are summarized in Table 2. Those questions are expressed in natural language and are rather informal. The ontology can be valued by its capability to answer those types of questions and by the extent that it provides support for the motivating scenarios. Those questions correspond to the functional ontology requirements, that is, the system behavior (functions or services).

The following rules, summarized from [23], have been taken into account while engineering the CCO:

**Table 1.** Some CCO motivating scenarios

| |
|---|
| 1. A **molecular biologist** is interested in knowing the components that interact during the cell-cycle process or in a given event as well as the roles that each component play. He/she is also interested in finding out the prospective components that play a role in a given cell-cycle phase. |
| 2. In the case of a **bioinformatician**, it might be interesting to know the cross-references from one component, that plays a given role in the cell-cycle process, to some external data resources. |
| 3. **General audience** will be interested in finding out all the fundamental components involved in the cell-cycle process. Besides, they will also want to find some synonyms for a given concept. |

**Table 2.** Competency questions

| | |
|---|---|
| 1 | What is a X-type **CDK**? |
| 2 | What is a Y-type **Cyclin**? |
| 3 | In what **events** is **CDK** Z involved? |
| 4 | In what **events** is **Rb** involved? |
| 5 | Which **CDKs** are involved in the endoreduplication process? |
| 6 | Which proteins are **phosphorylated** by kinase X? |
| 7 | Which CDK pertains to [G1—S—G2—M] phase? |

- Univocity: Terms should have the same meanings on every occasion of use.
- Positivity: Terms should designate genuine classes (do not use for instance `non-inhibitor`).
- Objectivity: Terms should designate biological natural kinds (do not use for instance terms such as `unknown`).
- Single Inheritance: No class in a classification hierarchy should have more than one *is_a* parent on the immediate higher level.
- Intelligibility of Definitions: The terms used in a definition should be simpler than the term to be defined.
- Basis in Reality: The quality of the ontologies depends on the degree to which they represent a certain portion of reality.

Those ontology principles are primarily considered as development guidelines since some rules could limit the representation of real-world situations, e.g. multiple inheritance.

### 3.2 Development Issues

There are many proposals for ontology development [10]. However, none of them is widely accepted and many of the ontology engineers combine at least some of these methods. Among the most representative methods we can identify the following: Enterprise methodology [27], the TOVE methodology [11], the Unified

methodology [28] and METHONTOLOGY [22]. All of them provide guidelines for developing ontologies and most of them consider the following steps:

- choosing an ontology language,
- choosing a development tool,
- acquire domain knowledge,
- reuse ontologies.

In general terms, most of the ontology development methods consider an informal phase, in which some ontology sketches are devised, and a formal phase, in which the ontology is formalized using an ontology language and specific development tools. Capturing knowledge is an expensive and arduous task. Protégé[15] has been used as a developing environment for our CCO. It provides a very user-friendly graphical interface and an extensible architecture based on plug-ins.

GO and the RO have been used as core ontologies for developing the CCO. All the processes from GO under the cell-cycle (GO:0007049) term were taken into account, while RO was completely imported. Thus, 304 terms were adopted from GO and all the 15 relations from RO. The CCO is updated daily and checked using data from GO.

Further enrichment of the CCO is required to merge the relations from RO and the ones provided by Reactome since there is a gap between the top-level relations from RO and the very specific ones from Reactome. For that sake, a mapping and association layer has been implemented. Based on this first skeleton this endeavor has now reached an initial stable state. Additional efforts will consider the integration of instances into a knowledge-based system that will provide a means for hypotheses evaluation.

The CCO is presently available in two formats most widely used by the bio-community: OBO and OWL. The CCO has an average depth of about 3 nodes. Although the mapping between the OBO and OWL (and vice versa) is not totally biunivocal (one-to-one correspondence), all the data has been preserved, i.e., all the terms and relations with their attributes were translated and when necessary some workarounds were implemented for the sake of completeness. The framework of the data integration is shown in Figure 3.

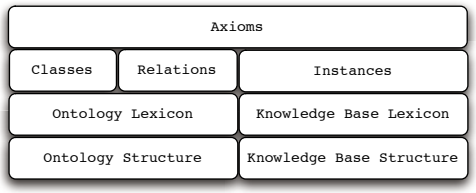


**Fig. 3.** Data integration framework

[15] http://protege.stanford.edu/

Classes, definitions, their references, and their relations with other classes have been treated as they appeared in their original resources. The DC ontology is being used for coding the terms references. Besides, a number of conventions have been adopted from [26], [3] and GO for identifying the terms within CCO. Each concept has a unique identifier of the form **CCO:cnnnnnnn**, where **CCO** indicates that the concept belongs to the CCO ontology (**CCO** is also known as the ontology namespace), **c** denotes de sub-namespace, and **nnnnnnn** consists of 7 numerical characters as shown in Table 3.

**Table 3.** CCO accession number. The first character denotes the type of term: **C** stands for cellular component, **F** for molecular function and **P** for biological process.

| c | n | n | n | n | n | n | n |
|---|---|---|---|---|---|---|---|
| [C,F,P] | [0-9] | [0-9] | [0-9] | [0-9] | [0-9] | [0-9] | [0-9] |

In order to maintain the accession numbers consistency as much as possible, the following four situations were considered:

1. A totally new added term implies a new accession number.
2. A merge occurs when two or more terms become a new term. The old accession numbers are copied as secondary accession numbers into the new term.
3. A split occurs when one term turns into two or more terms. The original accession numbers are kept in all the derived terms and a new primary accession number is added to each new term.
4. An accession number is dropped only when the data to which it was assigned have been completely removed from the ontology.

### 3.3 Formats Mapping

This section presents the mapping aspects that were taken into account while engineering the ontology structure. Previous work in this respect has already highlighted[16] many of the problems that we have faced here. The main problem is the insufficiency of information to get an OWL representation from an OBO one. The mapping to OWL has some caveats since currently there are some elements in OWL without any equivalent in OBO. For example, the existencial and universal restrictions cannot explicitly be represented in OBO. Consequently, we assume that all of them are existencial restrictions. Although nowadays no OBO ontology uses either union or intersection constructions, our conversion tools

[16] http://www.godatabase.org/dev/doc/mapping-obo-to-owl.html,
http://b-src.cbrc.jp/source/go-dev/doc/mapping-obo-to-owl.html,
http://gong.man.ac.uk/,
http://www.aiai.ed.ac.uk/resources/go/,
http://bioinfo.unice.fr/equipe/Claude.Pasquier/biowl/index.html

**Table 4.** Mapping of OBO and OWL terms. NDY stands for not defined yet.

| OBO keyword | OWL keyword | OWL element type |
|---|---|---|
| [Term] | owl:Class | Class description |
| id | rdf:ID | Class description |
| name | rdfs:label | rdf:Property |
| is_anonymous | NDY | NDY |
| alt_id | NDY | NDY |
| def | rdfs:comment | rdf:Property |
| comment | NDY | NDY |
| subset | NDY | NDY |
| synonym | synonym | owl:DataTypeProperty, owl:AnnotationProperty |
| xref | xref | owl:DataTypeProperty, owl:AnnotationProperty |
| is_a | rdfs:subClassOf | owl:ObjectProperty |
| intersection_of | owl:intersectionOf | Class description |
| union_of | owl:unionOf | Class description |
| disjoint_of | owl:disjointWith | Class axiom |
| relationship | NDY | NDY |
| is_obsolete | owl:DeprecatedClass | Version information |
| replaced_by | NDY | NDY |
| consider | owl:equivalentClass | Class axiom |

**Table 5.** Mapping among the OBO and OWL relationships. NDY stands for not defined yet.

| OBO keyword | OWL keyword |
|---|---|
| [Typedef] | owl:ObjectProperty |
| builtin | NDY |
| comment | NDY |
| def | rdfs:comment |
| exact_synonym | synonym (workaround) |
| id | rdf:id |
| inverse_of | owl:inverseOf |
| is_a | rdfs:subClassOf |
| is_anti_symmetric | is_anti_symmetric (workaround) |
| is_reflexive | is_reflexive (workaround) |
| is_transitive | rdf:type (TransitiveProperty) |
| NDY | rdf:type (SymmetricProperty) |
| name | rdfs:label (string) |
| xref_analog | NDY |

support them. Moreover, some terms do not have any definition and in consequence no references (no *dbxref* definition).

A mapping between the OBO specification and the OWL representation of the CCO is shown in Tables 4 and 5. As can be observed, there are some elements that have been mapped in a natural way (e.g. rdfs:label), while some other elements do not have a direct or defined mapping and some non-trivial approaches were taken to bypass this problem. The missing properties in OWL relations are: reflexivity, asymmetry, antisymmetry, intransitivy and partonomic relationships.

The mapping that we introduce is still in an experimental, non-final phase. There are several aspects that will be adapted after some pending decisions are taken. A stable level will be achieved once the OBO specification reaches a sufficient maturity stage. Moreover, because of the OBO metadata, the CCO has adopted OWL Full for its representation. Consequently, an alternative OWL-DL version will additionally become available. The OWL syntax of the OWL generated file is validated automatically by the format conversion script using *vowlidator*[17].

### 3.4 Handling of Inconsistencies

As stated above, one added value of a description logics approach embedded in the skeleton of the ontology structure is to allow automatic detection and handling of inconsistencies and misclassifications. Thus reasoning environments as RACER can be employed for checking the validity of some of the design principles of CCO, mentioned in Section 3.1. For instance, as already indicated in [23], the way GO uses the *is_a* relation may lead to a violation of the single inheritance principle. After loading the CCO into Protégé (with the Protégé OWL plugin [16]) and adding simple disjointness constraints to some of the CCO classes a certain number of this type of inconsistencies (32 in total which represents 10% of the entire CCO) have been detected by RACER. There are a number of relationships that should have been annotated as *part_of* instead of *is_a* and vice versa. A sample of this analysis is shown in Figure 4 and the corresponding GO cross-references are shown in Table 6. The centriole is an integral part of the centrosome, the microtubule organizing centre of the cell. Accordingly, the term *centriole replication* in GO is linked to its parent term *centrosome duplication* via a *part_of* relationship. Then, in order to be consistent the terms *regulation of centriole replication* and *negative regulation of centriole replication* should be related to their parent terms *regulation of centrosome cycle* and *negative regulation of centrosome cycle* by *part_of* relationships as well. Indeed, the inconsistency problem was solved by replacing the *is_a* with *part_of* relationship for these two pairs of terms in accordance to the True Path Rule[18].

We are presently investigating different approaches to solve these problems so that the divergence against the main ontology source (GO) is minimal. Furthermore, as stated in [23], the *part_of* relation should be specialized using spatial and temporal relations [24] to solve this type of inconsistencies.

[17] http://projects.semwebcentral.org/projects/vowlidator/
[18] http://www.geneontology.org/GO.usage.shtml#truePathRule

A major factor for managing the system will be the extent to which it diverges from its main sources. To minimize problems we will seek dialogue with the GO consortium and provide feedback.

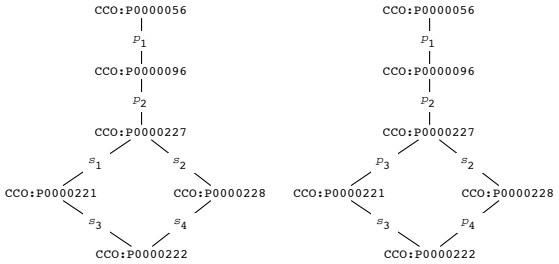


**Fig. 4.** Comparison of two sample class sub-hierarchies. Let $\mathcal{O}$ be an ontology structure, where $\{s_1, s_2, s_3, s_4\} \subset \mathcal{S}$ and $\{p_1, p_2\} \subset \mathcal{P}$ and $\mathcal{O}'$, where $\{s_2, s_3\} \subset \mathcal{S}'$ and $\{p_1, p_2, p_3, p_4\} \subset \mathcal{P}'$. The ontology $\mathcal{O}$, on the left, has an inconsistent relation $s_4$. On the right, the ontology $\mathcal{O}'$ is shown with a different, supposedly correct semantics.

**Table 6.** CCO ID and GO ID for the sample shown in Figure 4

| CCO ID | GO ID | Term |
|---|---|---|
| CCO:P0000056 | GO:0007049 | cell cycle |
| CCO:P0000096 | GO:0007098 | centrosome cycle |
| CCO:P0000227 | GO:0046605 | regulation of centrosome cycle |
| CCO:P0000221 | GO:0046599 | regulation of centriole replication |
| CCO:P0000228 | GO:0046606 | negative regulation of centrosome cycle |
| CCO:P0000222 | GO:0046600 | negative regulation of centriole replication |

## 4 Conclusions and Future Work

The amount of biomedical knowledge is becoming too large for traditional local approaches. Ontologies can increase the likelihood that such knowledge will be found and used by making the data easier to query and transform. A data integration pipeline detailing the issues of creating, updating and maintaining a cell-cycle knowledge base has been introduced. The formalization towards a description logics framework is a multi-staged and iterative process. The principal **contributions** of the knowledge base are:

- facilitate the communication between communities working on the cell-cycle process by providing a *lingua franca* or common terminology;
- saving time and effort by reusing the CCO and integrating it into related applications or systems;

- assist in application tasks such as knowledge acquisition, where semantic representation plays an important role;
- serving as a data repository.

The CCO is expected to be mainly used in the bioinformatics field and naturally, most of the feedback is expected to happen at that level. On the other hand, it is worth mentioning that the role of the CCO is not to compete against the cell-cycle data from GO. Rather, it is intended to complement GO by providing additional structure for the formalization process of the available knowledge in the cell-cycle field. The work so far has confirmed the existing integration obstacles due to the diversity of data formats and lack of formalization approaches as well as the trade-offs that are common in biological sciences.

The knowledge will be weighted or scored according to some defined evidence codes expressing the support media similar to those implemented in GO (experimental, electronically inferred, and so forth). A graphical user interface is also foreseen. The ultimate aim of the project is to support hypothesis evaluation about cell-cycle regulation issues. These hypotheses will be evaluated for consistency against the existing knowledge. The end product intends to include several intermediate milestones:

- An improved cell-cycle ontology, built on the existing ontology from GO and complemented with the temporal/dynamical aspects of the process. The three GO ontologies altogether supply an initial temporal framework for CCO by providing the cellular components (what/where), molecular functions (what) and biological processes (how/when). We are currently investigating approaches for connecting these three ontologies and representing knowledge such as for example `CDK A (what) is located in Cytoplasm (where) during Cytokinesis (when)`.
- A knowledge base holding the CCO as the core structure and data taken from Reactome and some other prospective resources as well as data produced by the DIAMONDS consortium, which is expected to boost the initial evolution of the system by providing data.
- A query [32] system for hypotheses validation, annotation assistance.
- A user interface providing a user-friendly environment for interacting with the system, creating queries and input data, annotation and so forth.

Besides the classical benefits provided by an ontology (data repository, knowledge sharing, validation, annotation, and so on), we aim to build a knowledge-based system that provides reasoning services oriented to hypotheses evaluation in the context of cell-cycle analysis. Consistency checking will further facilitate and improve some tasks done by annotation teams.

Finally, once a stable version of the ontology is released, the cell-cycle community will be invited to contribute to this effort and enhance the system.

**Aknowledgements.** This work was financially supported by the EU Framework programme for research, contract number LSHG-CT-2004-512143.

## References


1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.(eds.): The Description Logic Handbook. Theory, Implementation and Applications. Cambridge University Press (2003)
2. Bader, G.D., Betel, D., Hogue, C.W.V.: BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res, (2003) **31** 1 248-250
3. Bard, J., Rhee, S.Y., Ashburner, M.: An ontology for cell types. Genome Biology, (2005) **6** R21
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (May 2001)
5. Blake, J.: Bio-ontologies-fast and furious. Nature Biotechnology, (2004) **22** 773-774
6. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., ODonovan, C., Redaschi, N., Yeh, L.S.: The universal protein resource (Uniprot). Nucleic Acids Res., (2005) **33** Database issue D154D159
7. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L.: Genbank. Nucleic Acids Res., (2005) **33** Database issue D34D38
8. Covitz, P.A., Hartel, F., Schaefer, C., De Coronado, S., Fragoso, G., Sahni, H., Gustafson, S., Buetow, K.H. caCORE: A common infrastructure for cancer informatics. Bioinformatics, (2003) **19** 18 2404-2412
9. Gene Ontology Consortium.: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res., (2004) **32** Database issue D258-D261
10. Gomez-Perez, A., Corcho, O., Fernandez-Lopez, M.: Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer (2004)
11. Gruninger, M., Fox M.S.: Methodology for the Design and Evaluation of Ontologies. Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95 (Montreal) (1995)
12. Guarino, N.: Formal Ontology In Information Systems. In Proceedings of FOIS '98, Trento, Italy 6-8 June. IOS Press (1998)
13. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R.: IntAct: an open source molecular interaction database. Nucleic Acids Res., (2004) **32** Database issue D452-D455
14. Inze, D.: Why should we study the plant cell cycle? J. Exp. Bot., (2003) **54** 385 11251126
15. Joshi-Tope, G., Gillespie, M., Vastrik, I., DEustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., Stein, L.: Reactome: a knowledgebase of biological pathways. Nucleic Acids Res., (2005) **33** Database issue D428D432
16. Knublauch, H., Dameron, O., Mussen, M.A.: Weaving the biomedical semantic web with the Protege OWL plugin (2004)
17. Maedche, A.: Ontology Learning For The Semantic Web. Norwell, Massachusetts, Kluwer Academic Publishers (2003)
18. Maedche, A., Volz R.: The Ontology Extraction & Maintenance Framework Text-To-Onto. In Proceedings of the ICDM-2001 Workshop on the integration of Data Mining and Knowledge Management, San Jose, USA, November, 31 (2001)

19. Noy, N.F., McGuiness, D.L.: Ontology development 101: A guide to creating your first ontology. Technical Report SMI-2001-0880, Stanford University, SMI technical report (2001)
20. Ogden, C., Richards, I.: The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism. Routledge & Kegan Paul Ltd., London, 10 edition (1923)
21. McGuinness, D.L., van Harmelen, F.(eds.): OWL Web Ontology Language Overview. http://www.w3.org/TR/2004/REC-owl-features-20040210/
22. Lopez, M.F., Perez, A.G., Juristo, N.: METHONTOLOGY: From Ontological Art Towards Ontological Engineering. Workshop on Ontological Engineering. Spring Symposium Series: Stanford, USA (1997)
23. Smith, B., Kohler, J., Kumar, A.: On the application of formal principles to life science data: A case study in the Gene Ontology. Database Integration in the Life Sciences (DILS), Berlin: Springer (2004)
24. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome Biology, (2005) **6** R46
25. Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, NW., Goble, CA., Brass, A. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics, (2000) **16** 2 184-185
26. Thompson, J.D., Holbrook, S.R., Katoh, K., Koehl, K., Moras, D., Westhof, E., Poch, O.: MAO: a multiple alignment ontology for nucleic acid and protein sequences. Nucleic Acids Res., (2005) **33** 13 4164-4171
27. Uschold, M., King, M. Towards Methodology for Building Ontologies. Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95 (1995)
28. Uschold, M., Gruninger, M. Ontologies: Principles, methods and applications. Knowledge Engineering Review, (1996) **11** 2 93-136
29. Vermeulen, K., Van Bockstaele, D.R., Berneman, Z.N.: The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. Cell Prolif., (2003) **36** 3 131149
30. Yeh, I., Karp, P.D., Noy, N.F., Altman, R.B.: Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). Bioinformatics, (2003) **19** 2 241-248
31. Yu, A.C.: Methods in biomedical ontology. Journal of Biomedical Informatics, (2005) **78** 315-333
32. Zhang, Z., Miller, J.A.: Ontology query languages for the semantic web. A performance evaluation (2004)

## Appendix: Formal Definitions of Ontology and Knowledge Base Structure

The following formal definitions, which are introduced for showing some framework elements, have been adapted from [17], which in its turn has its mainstay in the Ogden-Richards' semiotic triangle [20].

An ontology structure is a 6-tuple: $\mathcal{O} = \{\mathcal{C}, \mathcal{R}, \mathcal{S}^{\mathcal{C}}, \mathcal{P}^{\mathcal{C}}, \rho, \mathcal{A}^{\mathcal{O}}\}$, where:

- $\mathcal{C}$ and $\mathcal{R}$ are two disjoint sets whose elements are called concepts and relations respectively.

- $\mathcal{S}^{\mathcal{C}}$ is a directed relation $\mathcal{S}^{\mathcal{C}} \subseteq \mathcal{C} \times \mathcal{C}$ which is called subsumption. $\mathcal{S}^{\mathcal{C}}(c_1, c_2)$ means that $c_1$ is a subconcept of $c_2$. Consequently, $c_1$ is called the subsumee and $c_2$ is the subsumer.
- $\mathcal{P}^{\mathcal{C}}$ is a directed relation $\mathcal{P}^{\mathcal{C}} \subseteq \mathcal{C} \times \mathcal{C}$ which is called partonomic inclusion.
- $\rho$ is a function that relates concepts in neither a taxonomical nor a partonomical way: $\rho : \mathcal{R} \to \mathcal{C} \times \mathcal{C}$.
- $\mathcal{A}^{\mathcal{O}}$ is a set of axioms on $\mathcal{O}$ expressed in a logical language, e.g. a description logic.

The notion of lexicon is also introduced. A lexicon for an ontology structure $\mathcal{O}$ is a 4-tuple $\mathcal{L} = \{\mathcal{L}^{\mathcal{C}}, \mathcal{L}^{\mathcal{R}}, \mathcal{F}, \mathcal{G}\}$, where:

- $\mathcal{L}^{\mathcal{C}}$ and $\mathcal{L}^{\mathcal{R}}$ are two sets whose elements are lexical entries for concepts and relations respectively.
- $\mathcal{F}$ and $\mathcal{G}$ are two relations $\mathcal{F} \subseteq \mathcal{L}^{\mathcal{C}} \times \mathcal{C}$ and $\mathcal{G} \subseteq \mathcal{L}^{\mathcal{R}} \times \mathcal{R}$ called references for concepts and relations respectively such that: $\mathcal{F}(l) = \{c \in \mathcal{C} | (l, c) \in \mathcal{F}\}$ and $\mathcal{F}^{-1}(c) = \left\{l \in \mathcal{L}^{\mathcal{C}} | (l, c) \in \mathcal{F}\right\}$. $\mathcal{G}$ and $\mathcal{G}^{-1}$ are defined analogously.

In [17] only the concept hierarchy ($\mathcal{S}^{\mathcal{C}}$) was hallmarked from the generic function $\rho$. We have also made evident the partonomic inclusion ($\mathcal{P}^{\mathcal{C}}$) since it plays an important role in our main source ontologies.

In turn, a knowledge base structure is a 4-tuple $\mathcal{KB} = \left\{\mathcal{O}, \mathcal{I}, \iota^{\mathcal{C}}, \iota^{\mathcal{R}}\right\}$, where:

- $\mathcal{O}$ is an ontology.
- $\mathcal{I}$ is a set whose elements are called instances.
- $\iota^{\mathcal{C}} : \mathcal{C} \to 2^{\mathcal{I}}$ and $\iota^{\mathcal{R}} : \mathcal{R} \to 2^{\mathcal{I} \times \mathcal{I}}$ are two functions for concept instantiation and relation instantiation respectively.

Again, the notion of lexicon is also introduced. Therefore, a lexicon for a knowledge base structure $\mathcal{KB}$ just is a tuple $\mathcal{L}^{\mathcal{KB}} = \{\mathcal{L}^{\mathcal{I}}, \mathcal{J}\}$, where:

- $\mathcal{L}^{\mathcal{I}}$ is a set whose elements are called lexical entries for instances.
- $\mathcal{J} \subseteq \mathcal{L}^{\mathcal{I}} \times \mathcal{I}$ is a reference relation for instances, such that for any $\mathcal{J}$, let for $l \in \mathcal{L}^{\mathcal{I}}$: $\mathcal{J}(l) = \{i \in \mathcal{I} | (l, i) \in \mathcal{J}\}$ and $\mathcal{J}^{-1}(i) = \left\{l \in \mathcal{L}^{I} | (l, i) \in \mathcal{J}\right\}$.