CCO, a paradigm for knowledge integration

Erick Antezana[*1], Mikel Egaña[2], Vladimir Mironov[1] and Martin Kuiper[1]*

[1] *VIB Department of Plant Systems Biology, Ghent University., Technologiepark 927, Ghent, Belgium*

[2] *University of Manchester Computer Science, Oxford Road, M13 9PL, UK*

1 PREAMBLE

New biological discoveries are being reported at an increasingly rapid rate. Biological information finds its way to very diverse locations ranging from journal articles to public databases associated with different sub-disciplines within biology and medicine. The integration of biological knowledge from diverse sources into a common format is recognized as a critical step toward hypothesis building [Racunas et al., 2004, Pennisi 2005]. Indeed, such an integrated resource is essential for exploration and exploitation by researchers and automated applications [Gardner 2005]. Knowledge integration is especially challenging for the large volume of biomedical literature. Manual curation and advanced text mining applications are becoming essential to extract and convert knowledge from literature to database formats. Converting and linking textual evidence to ontologies is yielding important new repositories of formally represented knowledge.

Ontologies are representational artifacts, comprising a taxonomy (as proper part), whose representational units are intended to designate some combination of universals, defined classes, and a series of relations among them [Smith et al., 2006]. They support consistent and unambiguous knowledge sharing and provide a framework for knowledge integration. An ontology links concept labels to their interpretations, i.e. specifications of their meanings and relations to other concepts. As such, ontologies can be used to support automatic semantic interpretation of textual information and thus provide a basis for advanced text mining [Müller et al., 2004, Doms et al., 2005]. Until recently, most text mining systems have not relied on ontologies or terminologies, explaining why biomedical text mining systems generally perform less well than text mining in other domains. The short term solution to this may be to combine automatic information retrieval with manual curation.

In addition to literature-dependent data, vast amounts of literature-independent data are being generated by high-throughput genome-wide analyses and accumulated in various databases. These databases represent another resource of context to infer biological function and to assess relations between biological entities. To obtain a powerful structuring and synthesis of all available biological knowledge it is essential to build an efficient information retrieval and management system. Such a system requires an extensive combination of data extraction methods, data format conversions and a variety of information sources. This, in fact, lies at the heart of *systems biology*: a branch of biological research that is based on a multidisciplinary approach for data integration.

Structured and integrated knowledge provides the basis to apply advanced reasoning approaches to validate hypotheses and to generate new knowledge. Reasoning services can be applied to a knowledge base at different levels, depending on the type of user (molecular biologist, ontology agent, etc.) who interacts with the system [Blake et al, 2006, Myhre et al., 2006]. They can be used to re-engineer the design of parts of the whole ontology or to design entirely new extensions to it. Alternatively, developers or ontology engineers can use reasoning services to check and maintain contents quality of a knowledge base before deploying it. Knowledge base curators can ensure that term redundancy is minimized while maintaining sufficiently detailed descriptions and consistency of the contents. Moreover, reasoning tools can also be used to find new classes (more specific or general) [Wolstencroft et al., 2007]. Finally, reasoning can allow information retrieval and integration into new hypotheses that are consistent with the current knowledge.

To interpret experimental results the extraction of all relevant information from a plethora of sources has become a major challenge for the life scientist. To illustrate this, consider the following scenario: a scientist would like to determine whether a novel protein (protein Y) interacts with *cell division cycle 2 kinase (cdc2).* To answer this question, the scientist first goes to a kinase pathway database[1] to obtain a list of all known proteins that interact with *cdc2* (e.g., cyclin A, CKS). The scientist then asks whether protein Y is structurally similar to any *cdc2*-interacting proteins. For this he/she needs to query protein structural databases such as SCOP[2] or CATH[3] and as a result, protein X is found. This

[1] http://kinasedb.ontology.ims.u-tokyo.ac.jp

[2] http://scop.mrc-lmb.cam.ac.uk/scop/

* To whom correspondence should be addressed: erant@psb.ugent.be

leads to the prediction that protein Y interacts with *cdc2*. Next, the scientist wishes to relate this interacting pair to a particular signaling pathway or biological process. After consulting for example KEGG[4] he/she finds out that this gene pair is likely to be involved in cell proliferation control. During the information foraging described above, the scientist had to constantly use literature databases and read relevant articles, only to get more information on one novel protein. In a systems biology approach this simply is not tenable anymore. Therefore, integration of biological knowledge is recognized as a critical knowledge gap in science [Cannata et al., 2005] and deemed essential for the future of the biosciences because dissemination and exploitation of the knowledge by automated applications will provide critical assistance to researchers who need to access and connect the diverse information.

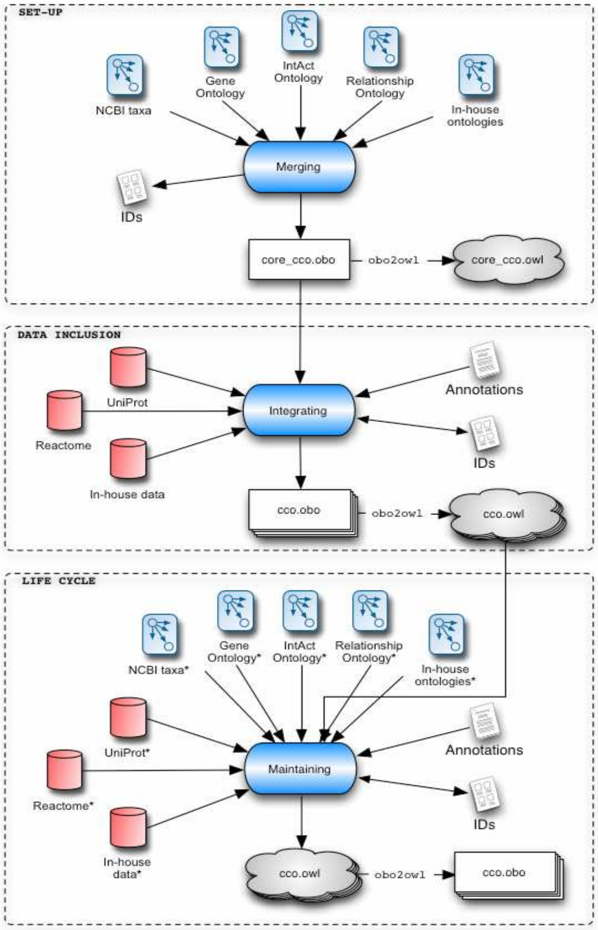


Fig. 1. CCO development pipeline: set-up, data inclusion and life cycle.

2 CCO: CURRENT STATUS

In the context of the FP6 project DIAMONDS[5] an ontology is being developed dedicated to the domain of cell cycle research [Antezana et al., 2006]. This application ontology, called Cell Cycle Ontology (CCO)[6], comprises data from a number of resources such as Gene Ontology [The Gene Ontology Consortium, 2000], Relations Ontology [Smith B et al., 2005], IntAct [Kerrien et al., 2006], NCBI taxonomy [Wheeler et al 2000], UniProt [The UniProt Consortium] as well as data from DIAMONDS partners. Fig 1 depicts the system development pipeline: during the first phase (set-up) some existing ontologies are integrated. Then, protein and gene data is added in the second phase. Finally, during the maintenance phase, a semantic improvement task is undertaken. The resulting CCO is designed to provide a richer view of the cell cycle regulatory process, in particular by accommodating the intrinsic dynamics of this process. For that purpose, three major components are considered: the (persistent) entity itself, its spatial localization, and its temporal localization (Fig. 2).

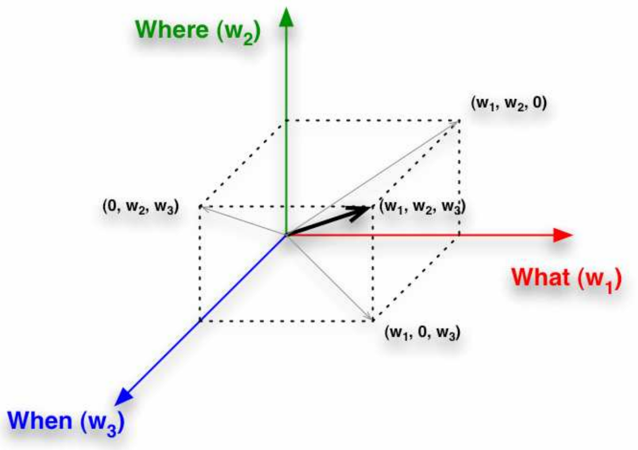


Fig. 2. W3 paradigm (what-where-when). Sample piece of knowledge: *"Cyclin B (w_1) is located in Cytoplasm (w_2) during Interphase (w_3)".*

Here, we present recent advances of the Cell Cycle Ontology project. CCO provides a test bed for the development of new approaches and tools necessary to create a fully-fledged knowledge base that enables deployment of advanced reasoning approaches for knowledge discovery and hypotheses generation. CCO supports 4 organisms: Human, Arabidopsis, Baker's yeast and Fission yeast. Presently, CCO holds more that 20000 concepts (more than 1000 bio-molecules and over 9000 interactions) and more that 20 types of relationships (Fig. 3). There is an ontology file for each of the 4

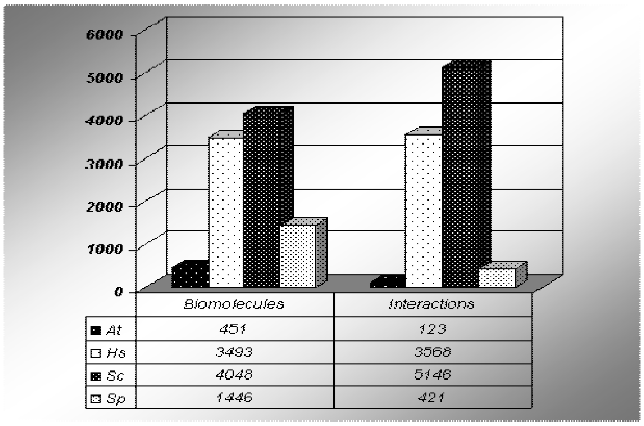
[3] http://www.cathdb.info/latest/index.html

[4] http://www.genome.jp/kegg/

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

[6] http://www.cellcycleontology.org

model organisms (H. sapiens, S. cerevisiae, S. pombe and A. thaliana). Each ontology file is available in several formats: OBO[7], OWL(-DL)[8], XML, DOT[9] and GML[10]. A set of PERL modules[11] has been developed to deal with OBO ontologies and in particular with the format conversion issues such as the mapping OBO to OWL[12]. CCO is constantly being cleaned and improved (addition of many more terms and relationships). The ontology in OBO format can be edited using OBO-Edit[13] and the OWL version can also be edited using Protégé[14]. We are in particular dealing with the issue of performance, as the file size (in particular of the OWL version) is starting to become prohibitive for specific tools like Protégé.



	Biomolecules	Interactions
■ At	451	123
□ Hs	3493	3568
▩ Sc	4048	5146
▨ Sp	1446	421

Fig. 3. CCO figures showing the number of biomolecules (proteins and genes) and protein-protein interactions per organism: A. thaliana (At), H. sapiens (Hs), S. cerevisiae (Sc) and S. pombe (Sp).

The *Semantic Web* is a major endeavor of applied Computer Sciences. It aims to enrich the existing Web with meta-data and processing methods to provide web-based systems with advanced (so-called intelligent) capabilities, in particular related to *context-awareness* and decision support. Here, reasoning capabilities will also ensure the smooth integration of ontologies and knowledge bases into the Semantic Web.

Reasoning services have been exploited for checking data consistency. Some preliminary results were shown previously [Antezana et al. 2006]. Currently, more elaborate consistency checking has been performed (article in preparation), in particular focusing on 'cellular localization'. Several inconsistent annotations have been reported by the reasoner (RACER[15]) based on some constrains defined at the level of the protein class while using the OWL version of CCO. Additional research needs to still be done in this direction since granularity plays an important role [Bittner 2002] while checking or querying systems such as CCO. In addition, an important extension of the reasoning capability is required to deal with 'fuzziness', a component that is usually present in biological data. In that sense, a combination of both issues (granularity and fuzzy data) needs to be considered. Let us consider the following use case: we are interested in confirming whether or not protein A and B could possibly interact. That is, we try to derive whether a spatio-temporal relation between A and B is shared based on the available data. As an initial data fact, we know that protein A and protein C interact in the same location at the same time (e.g. nucleus during the M-phase). Another data fact confirms that C and A interact in the same place at the same time (e.g. nucleus during the Interphase). Thus, we have *A same-time-same-place C* and *C same-time-same-place B.* The issue in both relations refers to approximate time at different levels of granularity**.** This sort of questions and the circumstances that could support that *A same-time-same-place B* with the interpretation that both, A and B, interact during the Interphase demonstrate the necessity of a solid theory for reasoning at different levels of granularity. This theory should also take into account uncertainty artifacts that are common in biology. This type of reasoning can lead to hypotheses that can be verified experimentally. Although in some fields some progress has been achieved and a limited number of efforts [Rector et al., 2006, Donnelly et al., 2006, Kumar et al., 2005] have been pursued for formalizing granular reasoning, to the best of our knowledge there are no existing approaches combining granularity and uncertain knowledge in biological sciences.

3 FUTURE WORK

Immediate CCO developments include tackling the performance issues by the implementation of a database backend. Also, reasoning at different levels of granularity is foreseen after integrating non-crispy data and weighting the current evidence of the existing biological relationships. Finally, integration of many more data sources is foreseen (e.g. homology, phosphorylation and so forth).

ACKNOWLEDGEMENTS

EA and VM are funded by the EU (FP6, contract number LSHG-CT-2004-512143). MEA is funded by Manchester University and EPSRC.

[7] http://www.geneontology.org/GO.format.obo-1_2.shtml
[8] http://www.w3.org/TR/owl-features/
[9] http://www.graphviz.org
[10] http://www.infosun.fim.uni-passau.de/Graphlet/GML/gml-tr.html
[11] http://search.cpan.org/~easr/
[12] http://www.bioontology.org/wiki/index.php/OboInOwl:Main_Page
[13] http://oboedit.org
[14] http://protege.stanford.edu
[15] http://www.racer-systems.com

REFERENCES

Antezana, E., Tsiporkova, E., Mironov, V., Kuiper, M. A cell-cycle knowledge integration framework. Lecture Notes In Bioinformatics 4075 19 - 34 (2006)

Bittner, T., 2002, Reasoning about qualitative spatio-temporal relations at multiple levels of granularity. In F. van Harmelen (ed.): ECAI 2002. Proceedings of the 15th European Conference on Artificial Intelligence, IOS Press, Amsterdam, 317-321.

Blake, J.A. and Bult, C.J. Beyond the data deluge: Data integration and bio-ontologies, Journal of Biomedical Informatics, Volume 39, Issue 3, Biomedical Ontologies, June 2006, Pages 314-320.

Cannata,N., Merelli,E. and Altman,R.B. (2005) Time to organize the bioinformatics resourceome. PloS Comput. Biol. 1(7):e76.

Doms, A., Schroeder, M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W783-6.

Donnelly M, Bittner T, Rosse C.A formal theory for spatial representation and reasoning in biomedical ontologies. Artif Intell Med. 2006 Jan; 36(1):1-27. Epub 2005 Oct 24.

Gardner SP. Ontologies and semantic data integration. Drug Discov Today. 2005 Jul 15;10(14):1001-7.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., Hermjakob, H. IntAct – Open Source Resource for Molecular Interaction Data. Nucleic Acids Research 2006; doi: 10.1093/nar/gkl958

Kumar, A., Smith, B., Novotny, D.D. Biomedical Informatics and Granularity. Comp. & Funct. Gen., 2005, 5(6-7): 501-508.

Myhre S, Tveit H, Mollestad T, Laegreid A. Additional gene ontology structure for improved biological reasoning. Bioinformatics. 2006 Aug 15;22(16):2020-7. Epub 2006 Jun 20.

Müller, H.M., Kenny, E.E., Sternberg, P.W. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature PLoS Biology Vol. 2, No. 11, e309 doi:10.1371/journal.pbio.0020309

Pennisi, E. (2005) How will big pictures emerge from a sea of biological data? *Science*, **309**, 94

Racunas SA, Shah NH, Albert I, Fedoroff NV. HyBrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics. 2004 Aug 4;20 Suppl 1:I257-I264.

Rector, A., Rogers, J. and Bittner, T. 2006. Granularity, scale and collectivity: When size does and does not matter. Journal of Biomedical Informatics, Vol. 39, Nr. 3, 333-349.

Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. Genome Biology 2005; 6: R46.

Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. Proceedings of KR-MED 2006, November 8, 2006, Baltimore MD, USA

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. (2000) Nature Genet. 25: 25-29

The UniProt Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2007 Jan;35(Database issue):D193-7. Epub 2006 Nov 16.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., Rapp, B.A. (2000). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2000 Jan 1;28(1):10-4

Wolstencroft, K., Stevens, R., Volker Haarslev, V. Applying OWL Reasoning to Genomic Data. In: Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, Christopher J. O. Baker and Kei-Hoi Cheung, Eds., Springer Verlag, 2007, pp. 225-248 (Chapter 11).