Ontology-driven Data Integration for Enabling Systems Biology: Cell Cycle Ontology

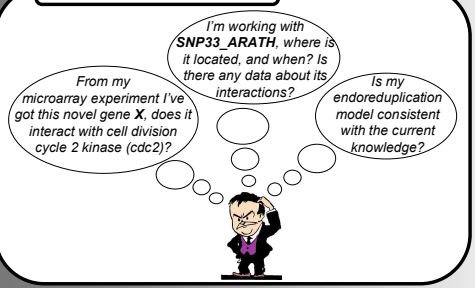
Erick Antezana[1], Vladimir Mironov[1], Mikel Egaña[2], and Martin Kuiper[1]

1 VIB Dept. of Plant Systems Biology, Ghent University. Technologiepark 927, B-9052 Ghent BELGIUM. E-mail: {erant,vlmir, makui}@psb.ugent.be

2 The University of Manchester. School of Computer Science, Oxford Road, M13 9PL, UK. E-mail: mikel.eganaaranguren@cs.man.ac.uk

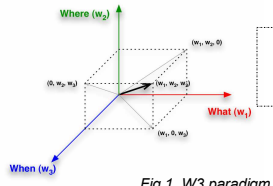
http://www.CellCycleOntology.org

Motivating scenarios



Objective

To capture the knowledge about the **cell cycle** process (particularly its dynamic facets) and to promote sharing, reuse and enable better computational integration with existing resources (semantic web). The ultimate aim is to support evaluation and generation of hypotheses via reasoning services about cell-cycle regulation. **Target organisms:** *S. cerevisiae, S. pombe, A. thaliana and H. sapiens*.



Sample piece of knowledge:

«Cyclin B (w1) is located in Cytoplasm (w2) during Interphase (w3) ».

Fig 1. W3 paradigm.

Data

CCO should capture the semantics and spatio-temporal relationships (*Fig 1*) of cell-cycle components (proteins, genes, cellular locations, phases, and so on). The data sources are:

- GO (CL, CC branches)
- RO
- MI (IntAct ontology)
- GOA files
- PPI: IntAct
- NCBI taxonomy
- UniProt
- Cell cycle functional data
- Data obtained with bio-tools (e.g.OrthoMCL)

OBO and OWL-DL formats have been chosen for representing the knowledge. RACER is mainly used for checking the data consistency and for doing classifications.

Data integration pipeline

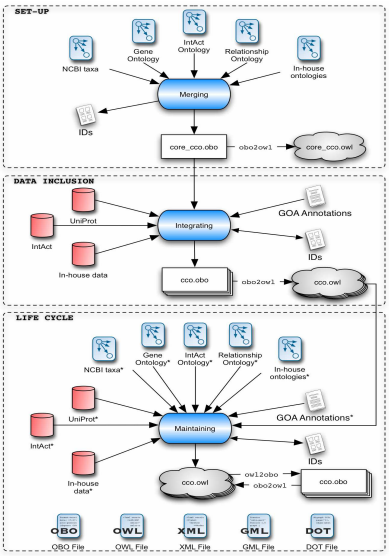


Fig 2. Data integration pipeline.

Exploring CCO

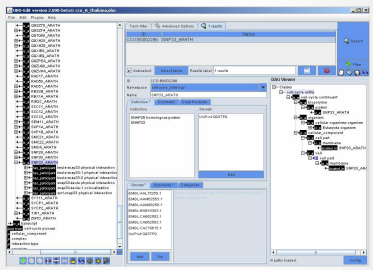


Fig 3. CCO in OBO-Edit.

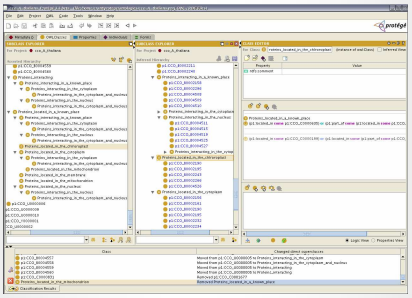


Fig 4. CCO in Protégé.

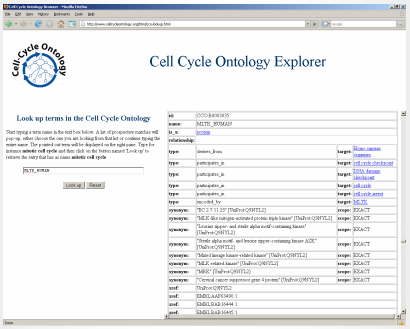


Fig 5. CCO Explorer (online).

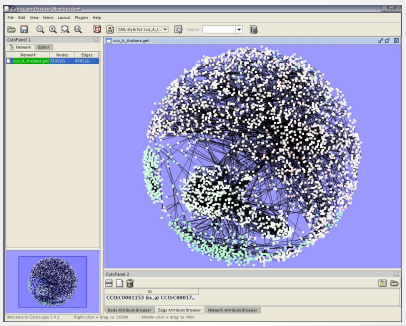


Fig 6. CCO in Cytoscape.

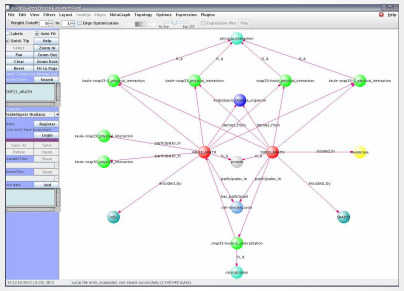
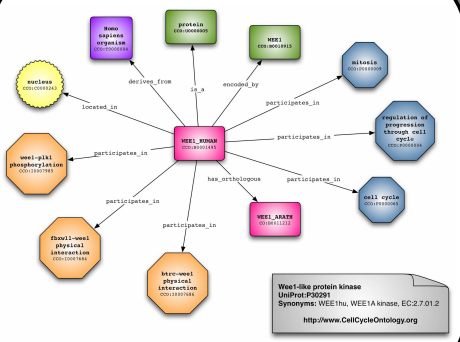


Fig 7. CCO in visANT.

Local neighborhood



Reasoning results

- There are a number of relationships in GO (core source of CCO) that might have been better annotated as *part_of* instead of *is_a*.
- The results inspired the GO team to made some amendments to the process part of the GO (e.g regulation of cell cycle).
- Inconsistencies found in the data about the cellular localizations and protein-protein interactions.

Conclusions and Results

- A fully automated data integration pipeline (nightly launched) was developed (*Fig 2*).
- Concrete problems and results related to the implementation of automatic format mappings (OWL, XML, DOT, GML) between ontologies and inconsistency checking issues have been identified.
- Several exports in commonly used formats have been developed (*Fig 3-7*).
- 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.

Future work

- Knowledge will be weighted (e.g. evidence codes) expressing the support media similar to those implemented in GO (experimental, electronically inferred, and so forth).
- Ontolome analysis (e.g. hypothesis generation by ontology alignments).
- An advanced query system will be developed (DL-based).
- More data to be integrated (upon feedback).

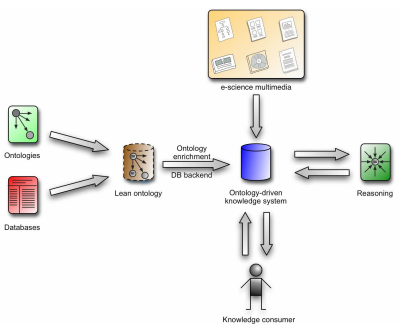


Fig 8. Outlook into the future

Acknowledgements

This work has been funded by the EU (contract number LSHG-CT-2004-512143), The Manchester University, EPSRC, and Marie Curie EST.