CCO: concept & current status

Erick ANTEZANA

Dept. of Plant Systems Biology. Flanders Interuniversity Institute for Biotechnology/Ghent University. Ghent BELGIUM.

erant@psb.ugent.be

Overview

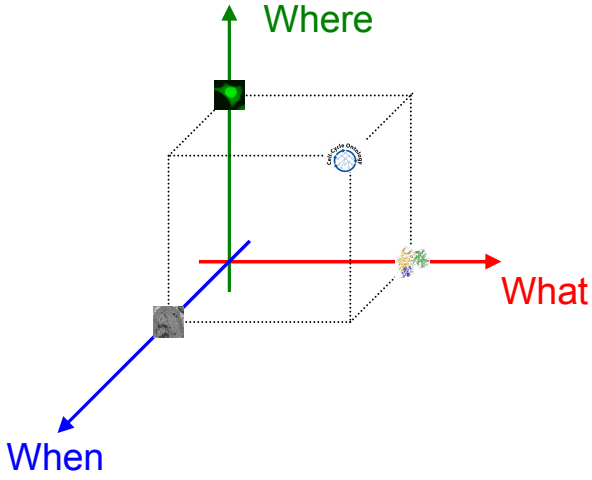
- Motivation
- Objective
- Data integration pipeline
- CCO engineering
- A CCO sample
- Exploiting reasoning services
- Conclusions
- Future work

Prospective users

- Molecular biologist: interacting components, events, roles that each component play. Hypothesis evaluation.
- Bioinformatician: data integration, annotation, modeling and simulation.
- General audience: educational purposes.

Objective

- Capture the knowledge of the CC process
- dynamic aspects of terms and their interrelations*
- promote sharing, reuse and enable better computational integration with existing resources
- Issues: *synonymy*, *polysemy*

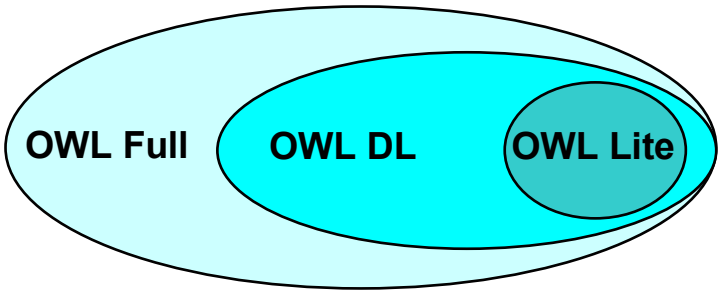


"Cyclin B (what) is located in Cytoplasm (where) during Interphase (when)"

* **Dynactome**: http://dynactome.mshri.on.ca/

Knowledge Formalization

- Why OBO?
  - "Human readable"
  - Standard
  - Tools (e.g. OBOEdit)
  - http://obo.sourceforge.net
- Why OWL?
  - "Computer readable"
  - Reasoning capabilities vs. computational cost rat
  - Formal foundation (Description Logics: http://dl.kr.org/)
  - http://www.w3c.org/TR/2004/REC-owl-features-20040210
  - Reasoning: RACER, Pellet, FaCT++

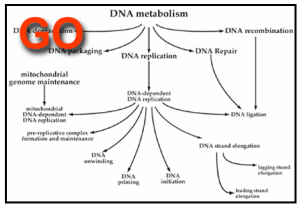


Format mapping: OBO⇔OWL

- Mapping not totally biunivocal; however, all the data has been preserved.
- Missing properties in OWL relations:
  - reflexivity,
  - asymmetry,
  - Intransitivity, and
  - partonomic relationships.
- Existential and universal restrictions cannot be explicitly represented in OBO => Consider all as existential.
- CCO in OWL is in sync with the NCBO mapping (DL)
- Mapping efforts:
  - http://spreadsheets.google.com/ccc?key=pWN_4sBrd9l1Umn1LN8WuQQ
  - http://www.psb.ugent.be/cbd/cco/OBO2OWL%20Mappings.pdf

CCO sources

- Ontologies:
  - Gene Ontology (GO)
  - Relationships Ontology (RO)
  - Dublin Core (DC)
  - Upper level ontology (ULO)
- Data sources
  - GOA files
  - PPI: IntAct, BIND, Reactome
  - Text mining
  - CBS
  - Other DBs (e.g. phosphorylation)

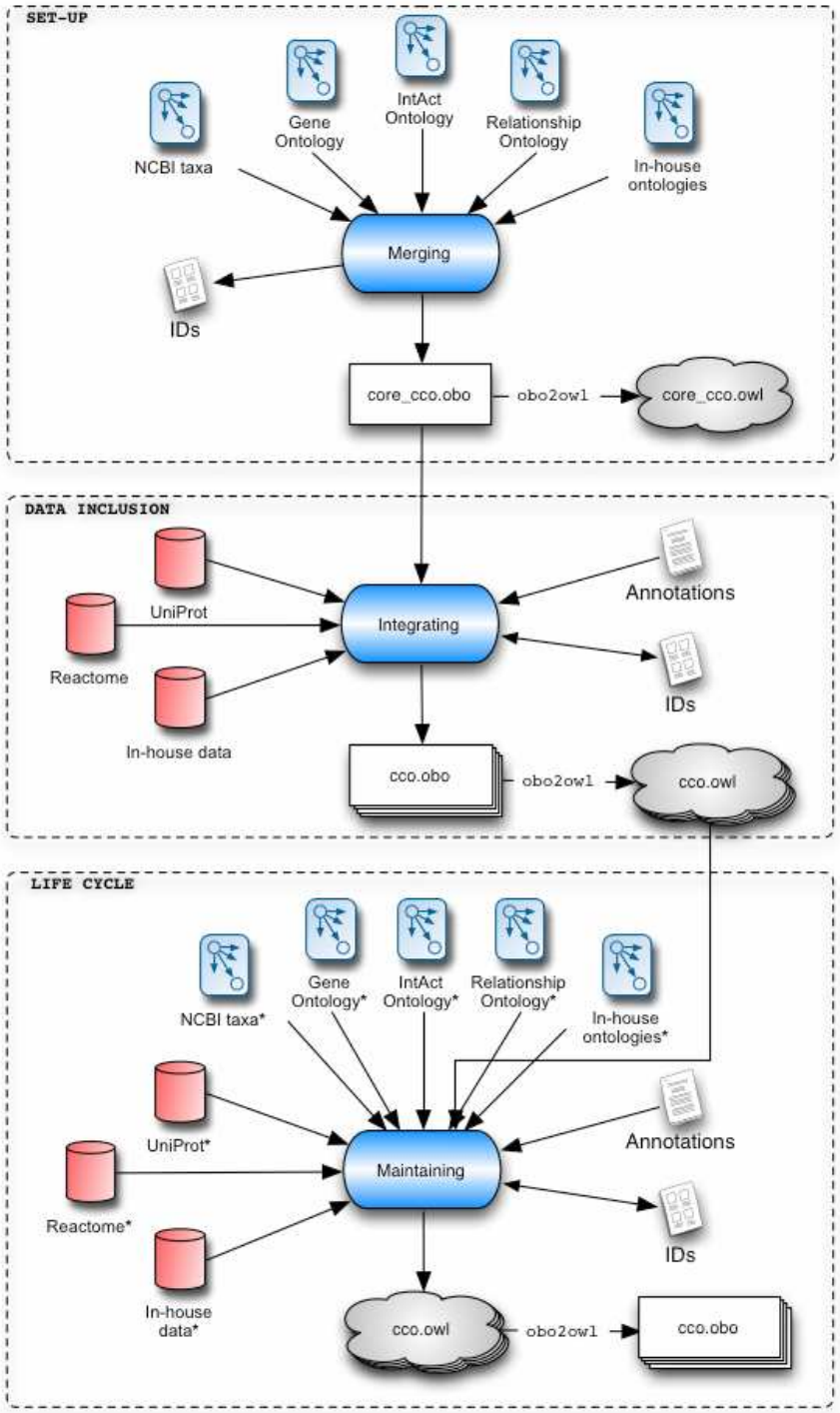


Data integration

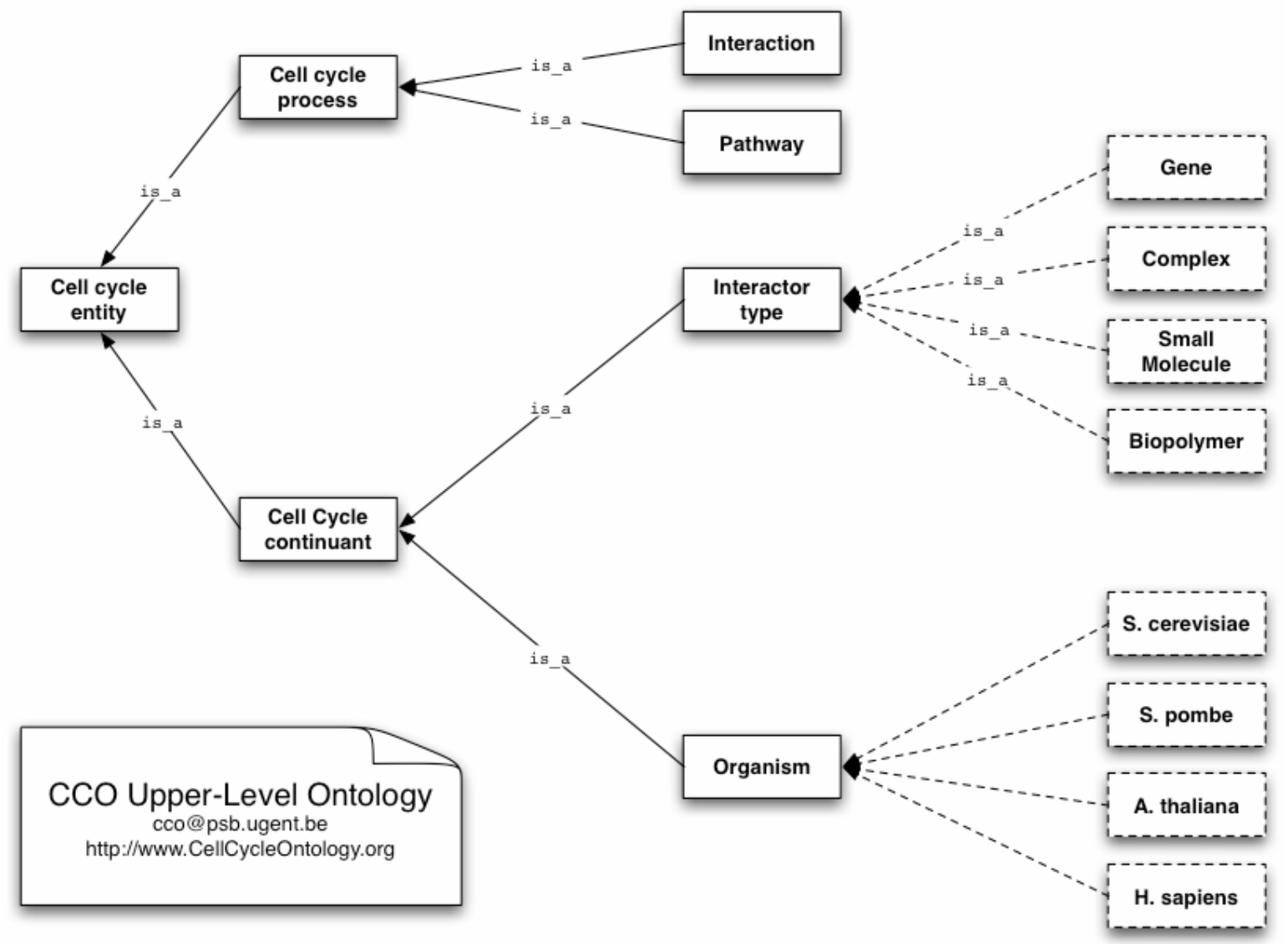
- ontology integration
- format mapping

- data integration
- data annotation
- consistency checking

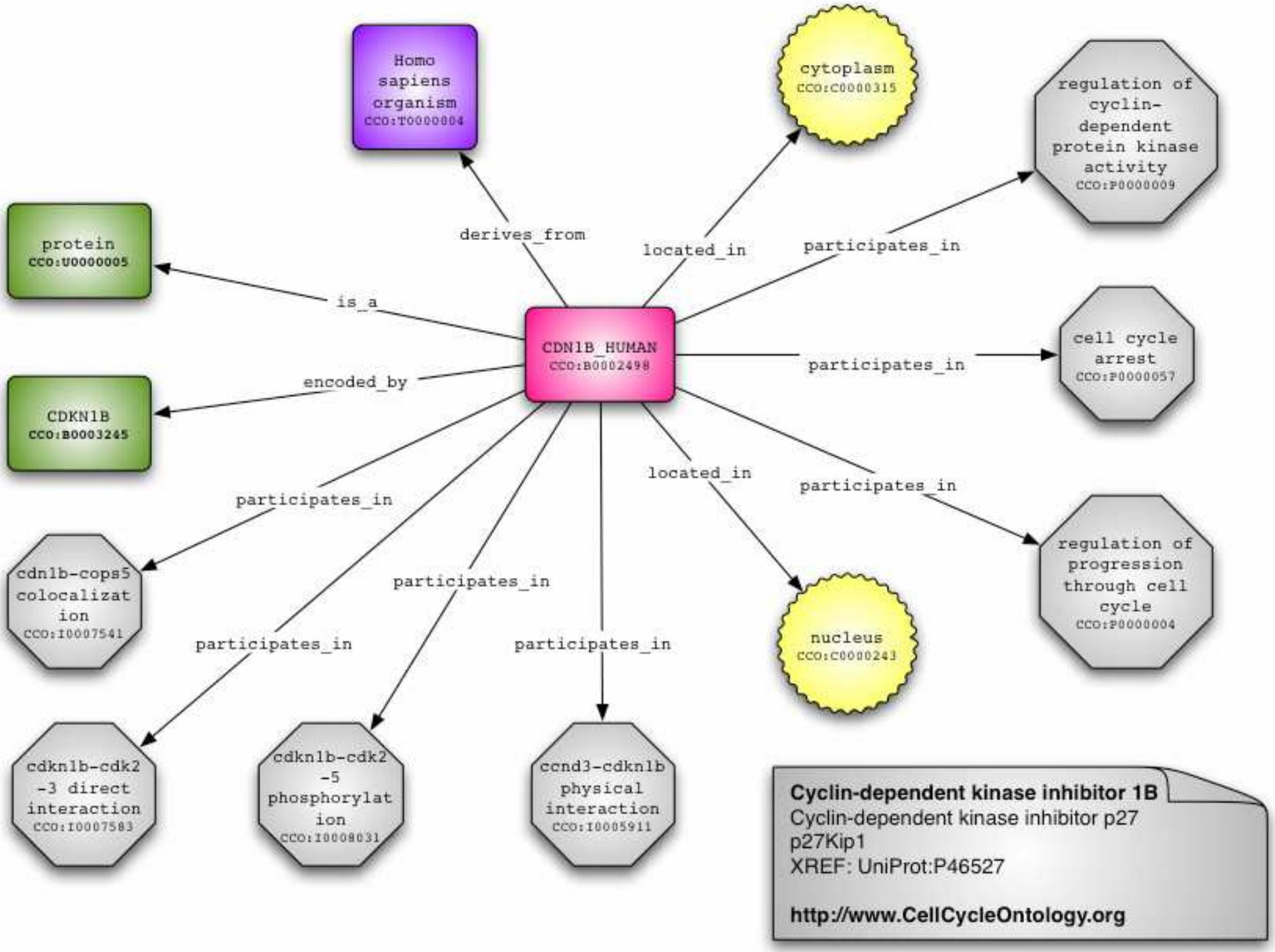
- maintenance
- data annotation
- semantic improvement



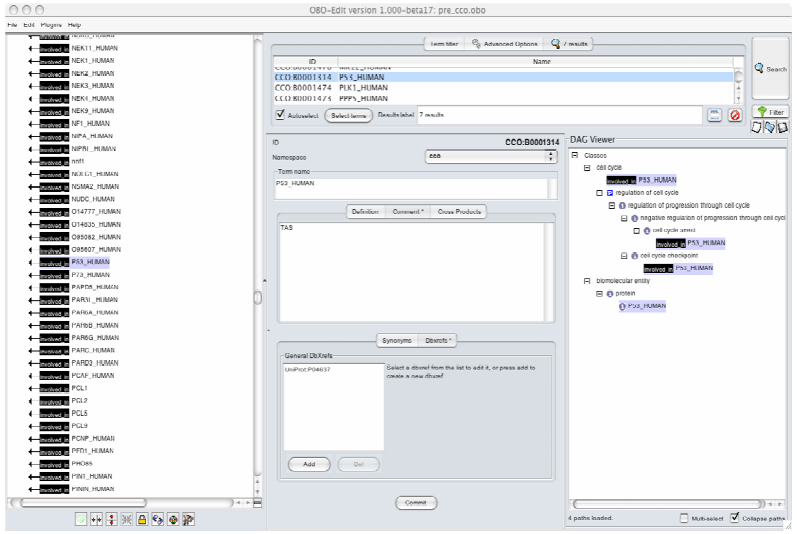
Upper Level Ontology*

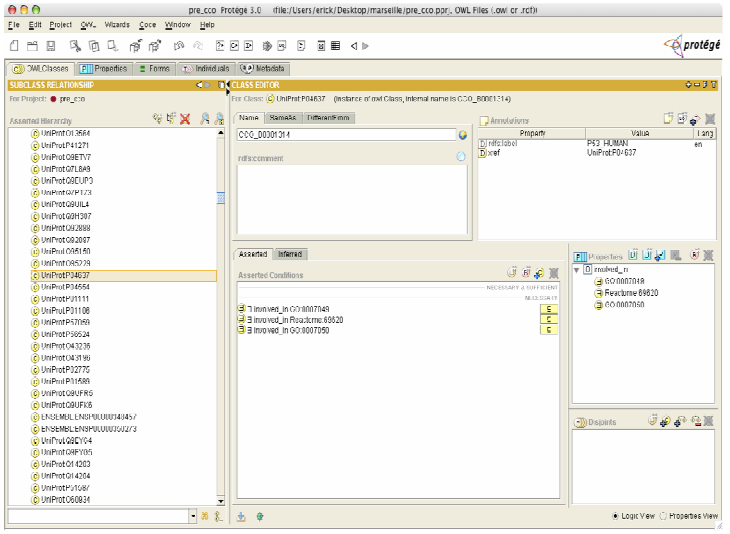


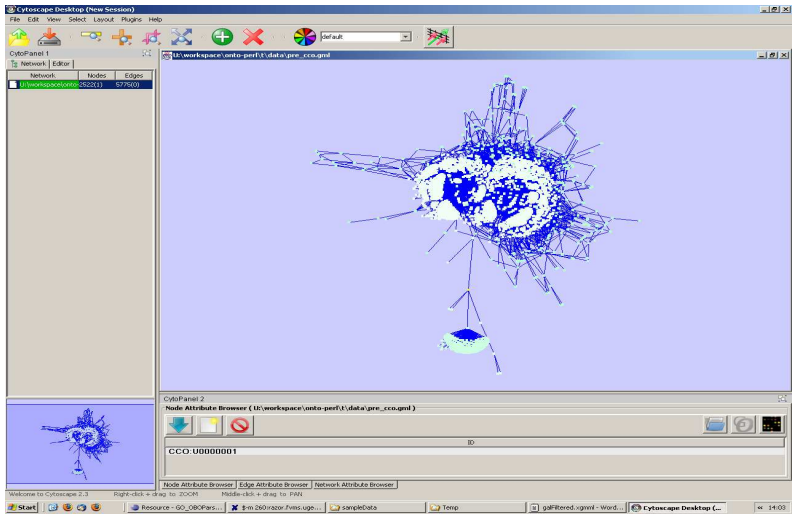
* Based on BFO: endurant vs. perdurant

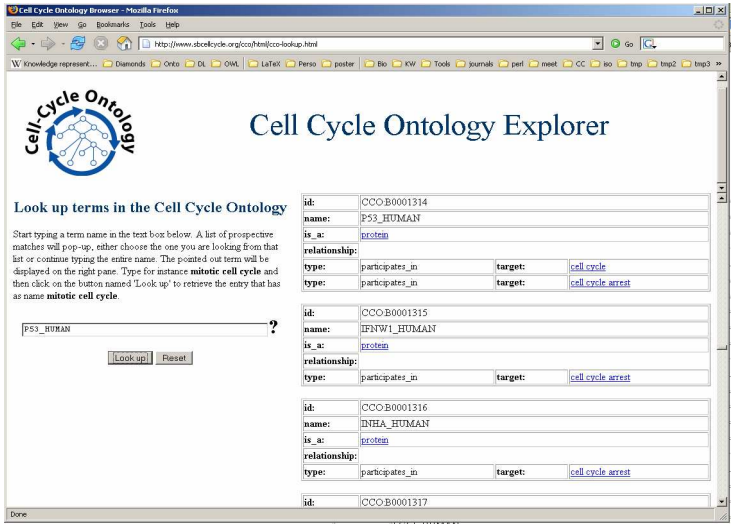

Homo sapiens organism
CCO:T0000004
cytoplasm
CCO:C0000315
regulation of cyclin-dependent protein kinase activity
CCO:P0000009
protein
CCO:U0000005
derives_from
located_in
participates_in
is_a
CDN1B_HUMAN
CCO:B0002498
cell cycle arrest
CCO:P0000057
encoded_by
CDKN1B
CCO:B0003245
participates_in
located_in
participates_in
cdn1b-cops5 colocalization
CCO:I0007541
participates_in
participates_in
participates_in
regulation of progression through cell cycle
CCO:P0000004
nucleus
CCO:C0000243
cdkn1b-cdk2-3 direct interaction
CCO:I0007583
cdkn1b-cdk2-5 phosphorylation
CCO:I0008031
ccnd3-cdkn1b physical interaction
CCO:I0005911
Cyclin-dependent kinase inhibitor 1B
Cyclin-dependent kinase inhibitor p27
p27Kip1
XREF: UniProt:P46527
http://www.CellCycleOntology.org

CCO in …


OBO-Edit version 1.000-beta17: pre_cco.obo
CCO:B0001314
DAG Viewer

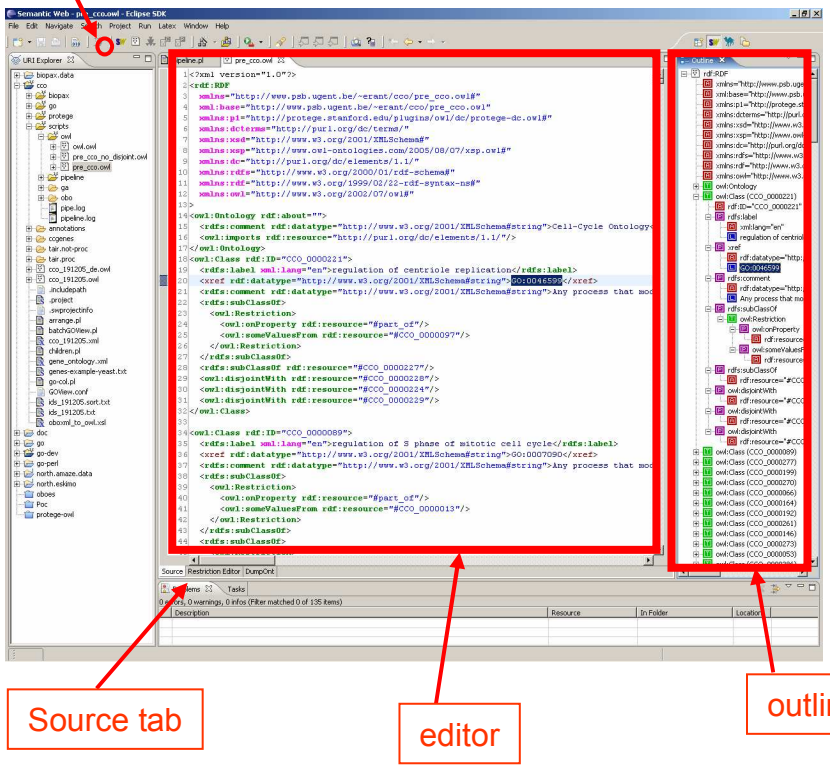

pre_cco Protégé 3.0
protégé
SUBCLASS RELATIONSHIP
CLASS EDITOR


Cytoscape Desktop (New Session)
CytoPanel 1
CytoPanel 2
Node Attribute Browser


Cell Cycle Ontology Browser - Mozilla Firefox
Cell Cycle Ontology Explorer
Look up terms in the Cell Cycle Ontology
Look up
Reset

CCO checked with…

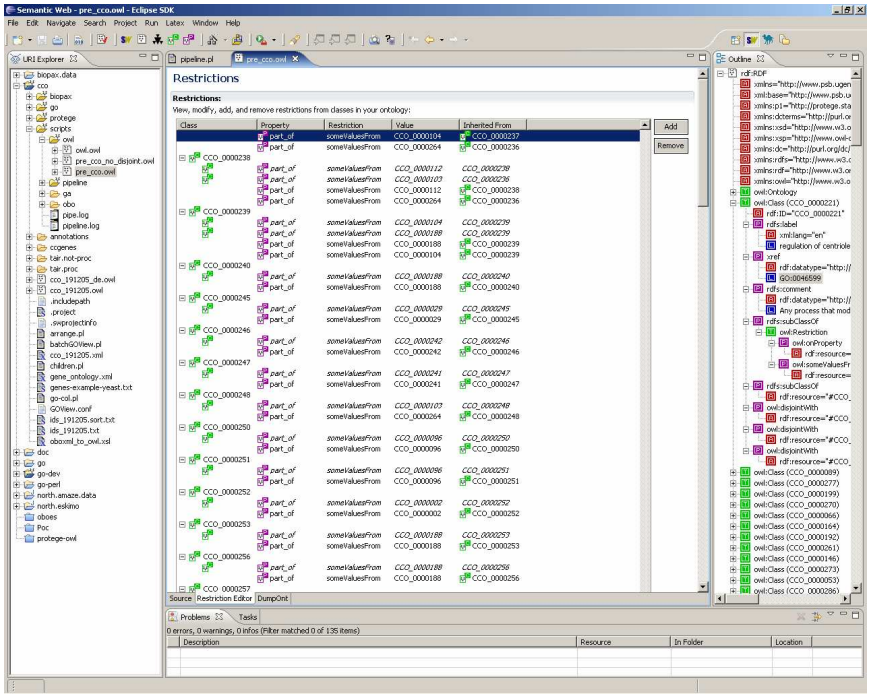
validator



Source tab

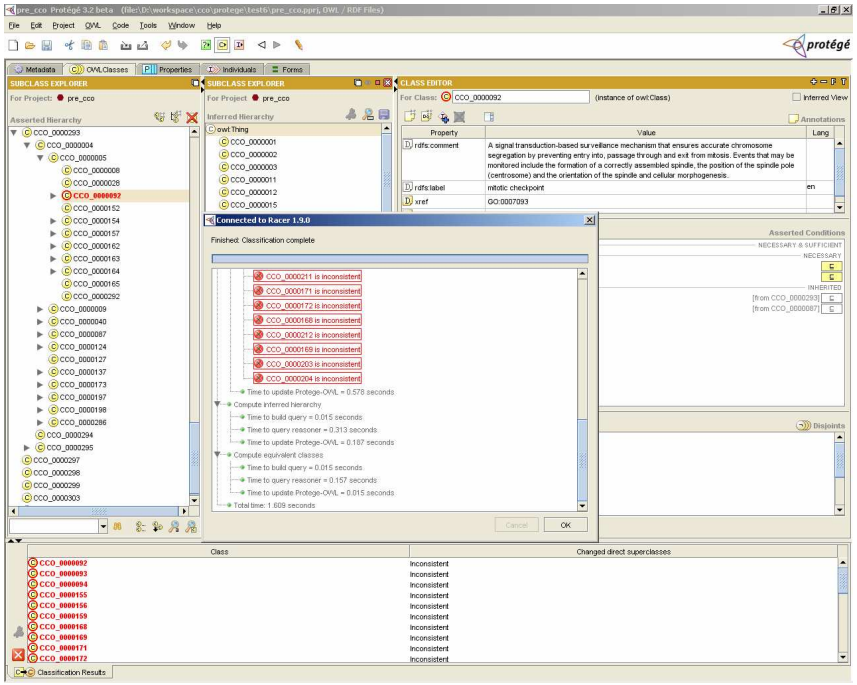
editor

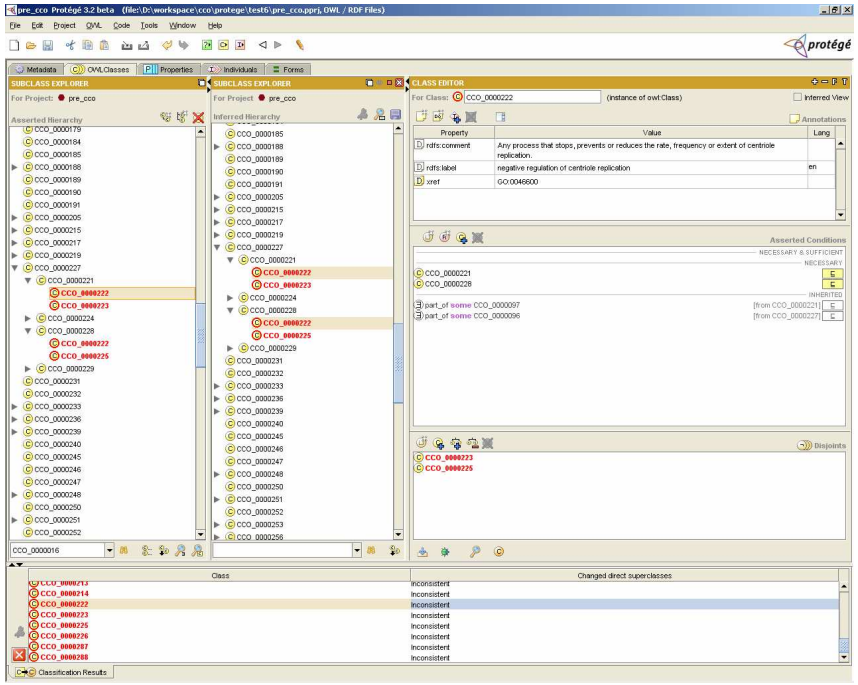
outline



SWeDE Eclipse plug-in: http://owl-eclipse.projects.semwebcentral.org

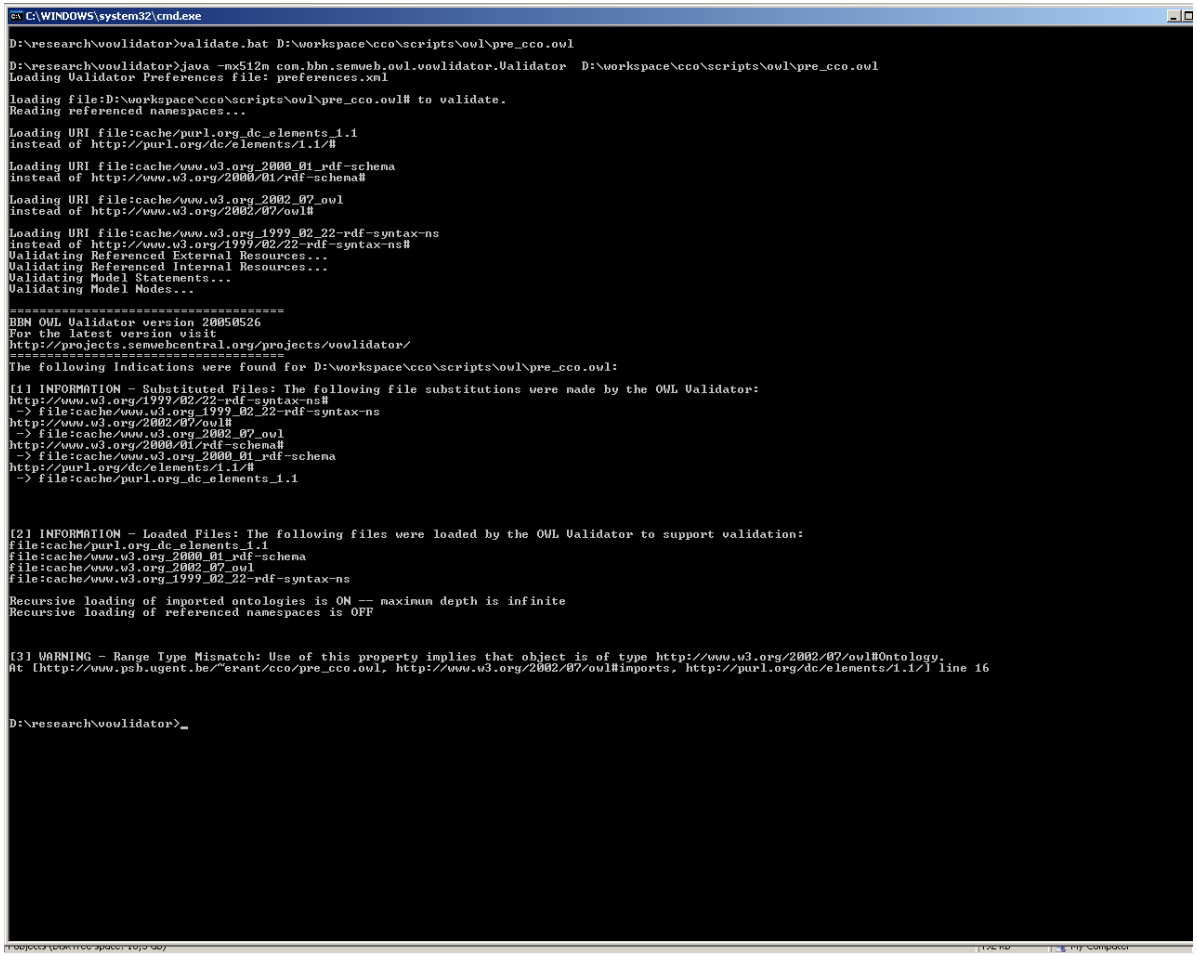
Checked with…





Protégé: http://protege.stanford.edu/)

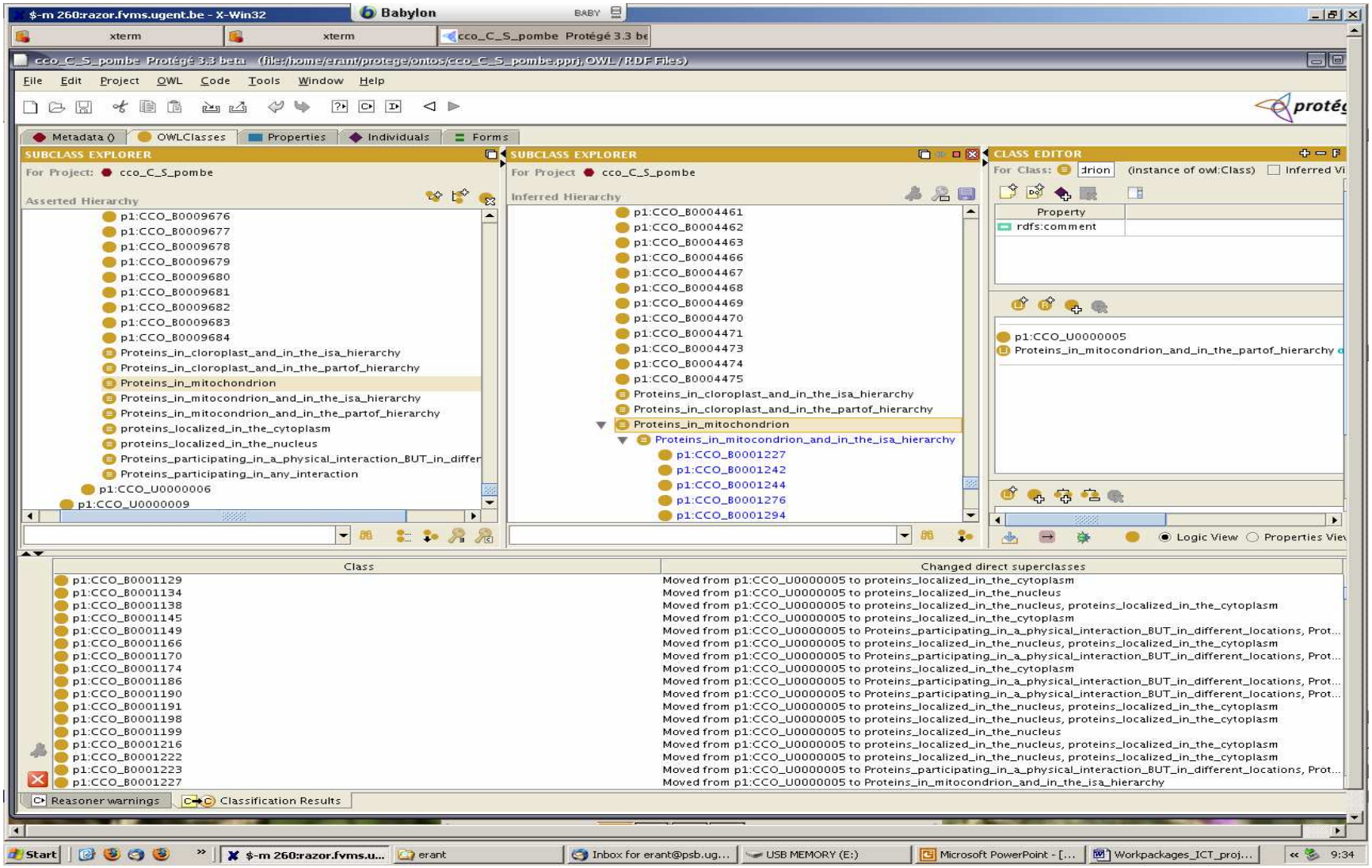
and with…



Vowlidator: http://projects.semwebcentral.org/projects/vowlidator/)

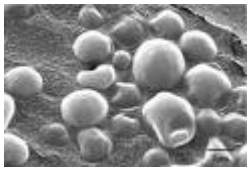
Cellular localization checks

- Query: *"If a protein is cell cycle regulated, it must not be located in the mitochondria"* (RACER*)



* http://www.racer-systems.com

Availability

- At, Sc, Sp and Hu    
- Sourceforge: SVN (API)
- http://www.CellCycleOntology.org
- ***"A cell-cycle knowledge integration framework".*** Data Integration in Life Sciences, DILS 2006, LNBI 4075, pp. 19-34, 2006.
- Mailing list (low traffic):
  - **https://maillist.psb.ugent.be/mailman/listinfo/ccofriends**

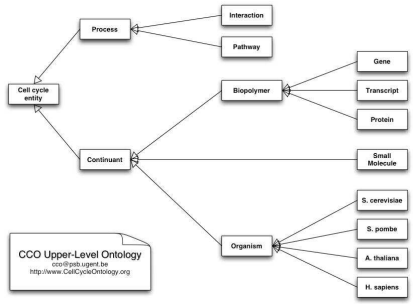
Products

- API (Perl): OBO/OWL ontologies handling
- Exports:
  - OBO, OWL, DOT, GML, XGMML*, SBML*
- Conversion tools:
  - obo2owl
  - owl2obo*
- CCO explorer (online)

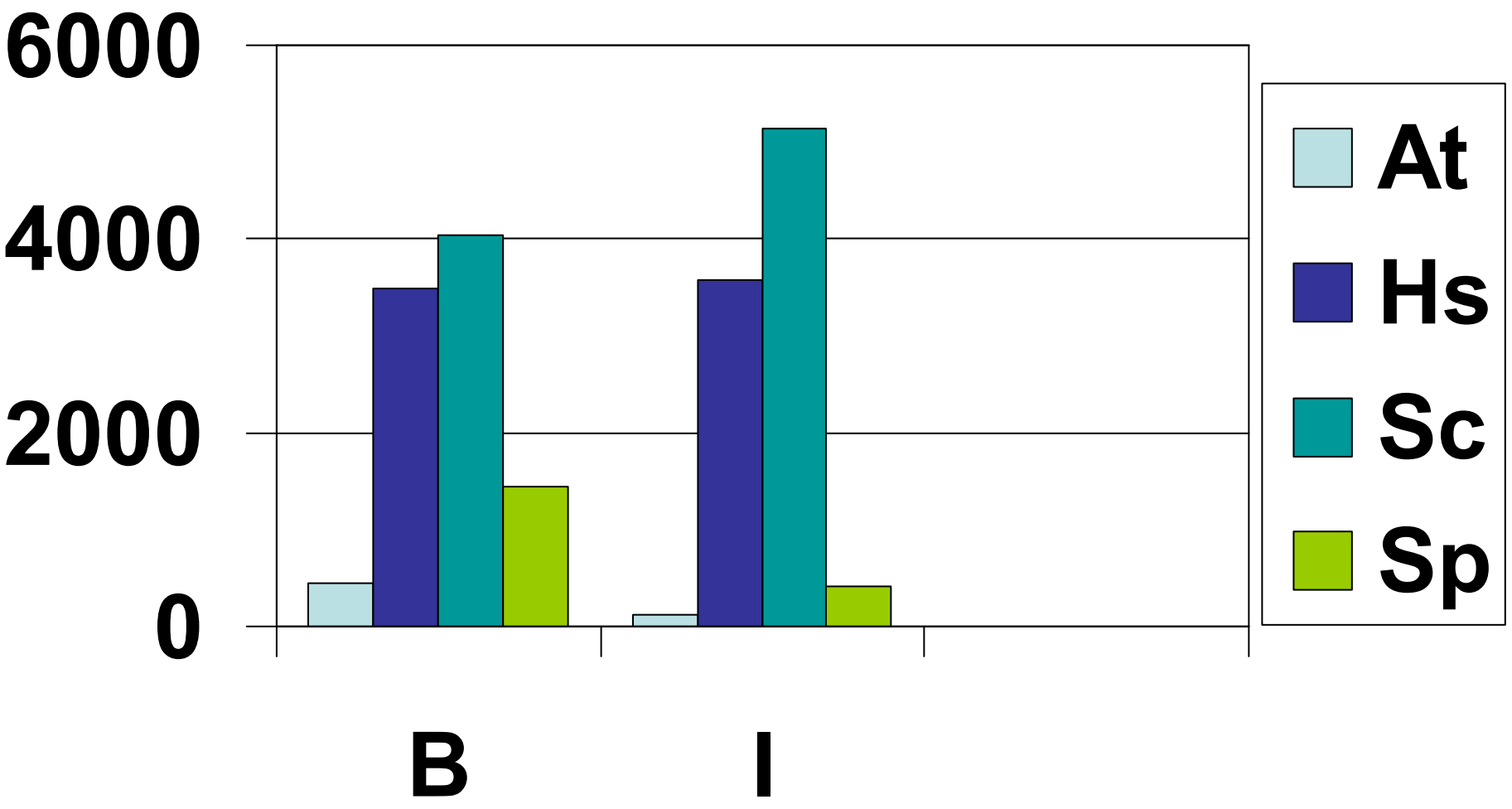
* Under development

CCO figures

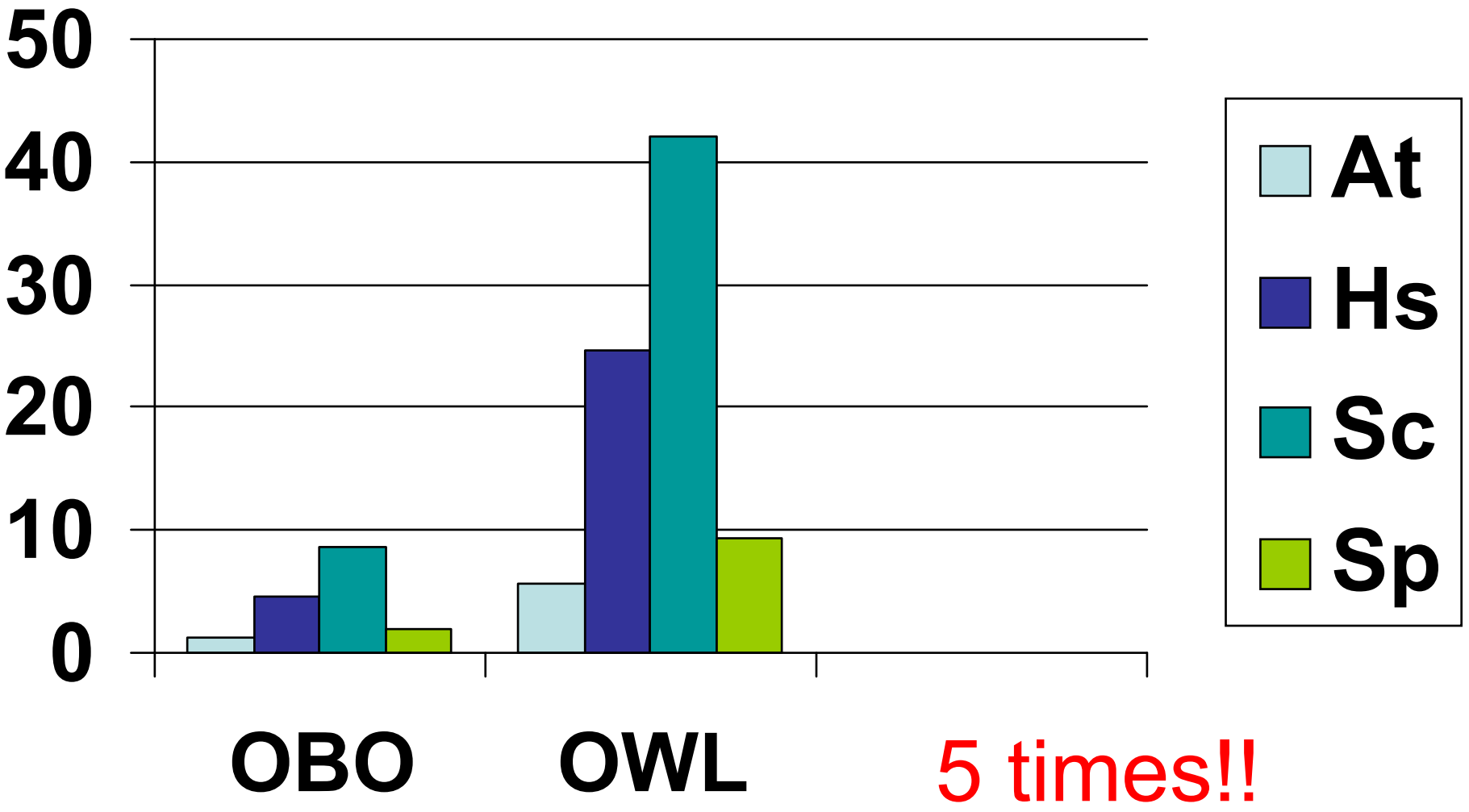
- Terms/classes:
  - B: 9392 (genes + proteins : GOA + UniProt)
  - P: 322 (GO)
  - I : 9258 (only from IntAct)
  - R: 181 (GO)
  - C: 1937 (filter out)
  - T: 60 (NCBI)
- #rel's
  - #RO + #CCO = 15 + 5 = 20
- Total: 21213



CCO figures (per organism)



File sizes!



Conclusions

- Data integration pipeline prototype: life cycle of the KB
- Concrete problems and initial results: automatic format mappings and inconsistency checking issues
- Existing integration obstacles due to the diversity of data formats and lack of formalization approaches
- Common trade-offs in biological sciences

Future work

- Persistency: DB backend
- An ULO for application ontologies
- Weighted or scored knowledge: evidence codes expressing the support media similar to those implemented in GO (experimental, electronically inferred, and so forth) => Fuzzy relationships => More data...
- Advanced query system
- Web user interface
- The ultimate aim of the project is to support hypothesis evaluation about cell-cycle regulation issues.