

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

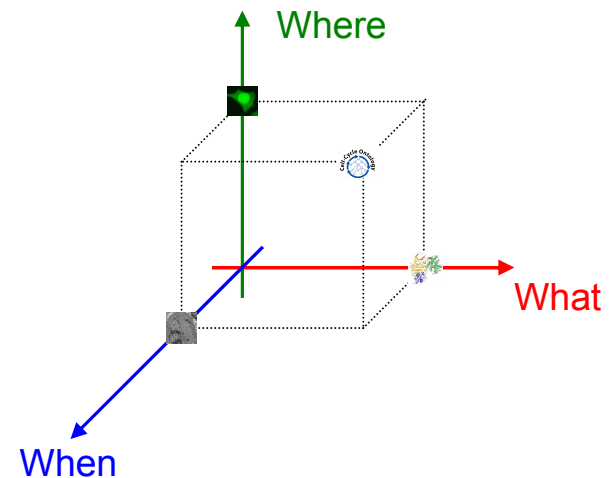
- Motivation
- Objective
- Data integration pipeline
- CCO engineering
- A CCO sample
- Exploiting reasoning services
- Conclusions
- Future work

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*

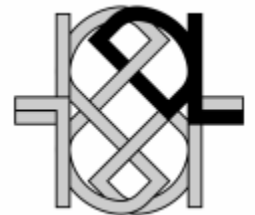
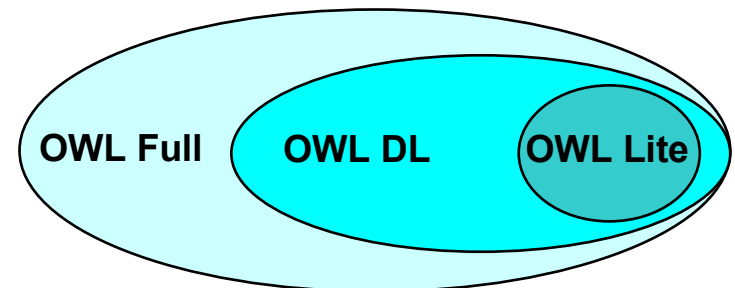


“Cyclin B (*what*) is located in Cytoplasm (*where*) during Interphase (*when*)”

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

Knowledge Formalization

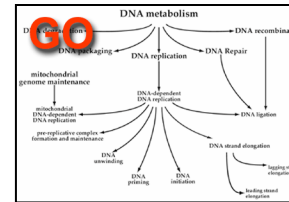
- Why OBO?
 - “Human readable”
 - Standard
 - Tools (e.g. OBOEdit)
 - <http://obo.sourceforge.net>
- Why OWL?
 - “Computer readable”
 - Reasoning capabilities vs. computational cost ratio
 - Formal foundation (Description Logics: <http://dl.kr.org/>)
 - <http://www.w3c.org/TR/2004/REC-owl-features-20040210>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



Format mapping: OBO \leftrightarrow OWL

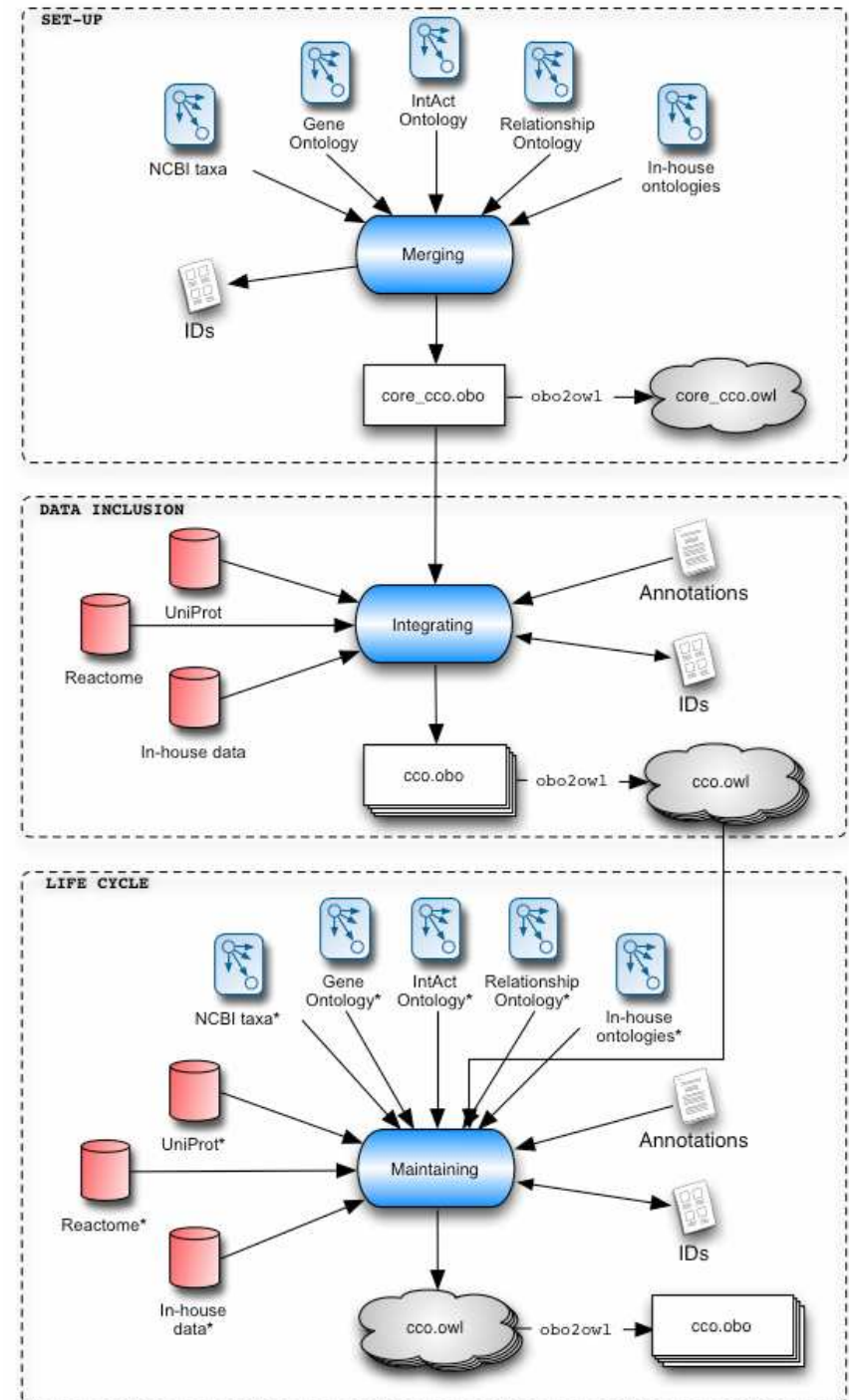
- Mapping not totally **biunivocal**; however, all the data has been preserved.
- Missing properties in OWL relations:
 - reflexivity,
 - asymmetry,
 - Intransitivity, and
 - partonomic relationships.
- Existential and universal restrictions cannot be explicitly represented in OBO => **Consider all as existential.**
- CCO in OWL is in sync with the NCBO mapping (DL)
- Mapping efforts:
 - http://spreadsheets.google.com/ccc?key=pWN_4sBrd9I1Umn1LN8WuQQ
 - <http://www.psb.ugent.be/cbd/cco/OBO2OWL%20Mappings.pdf>

- 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)

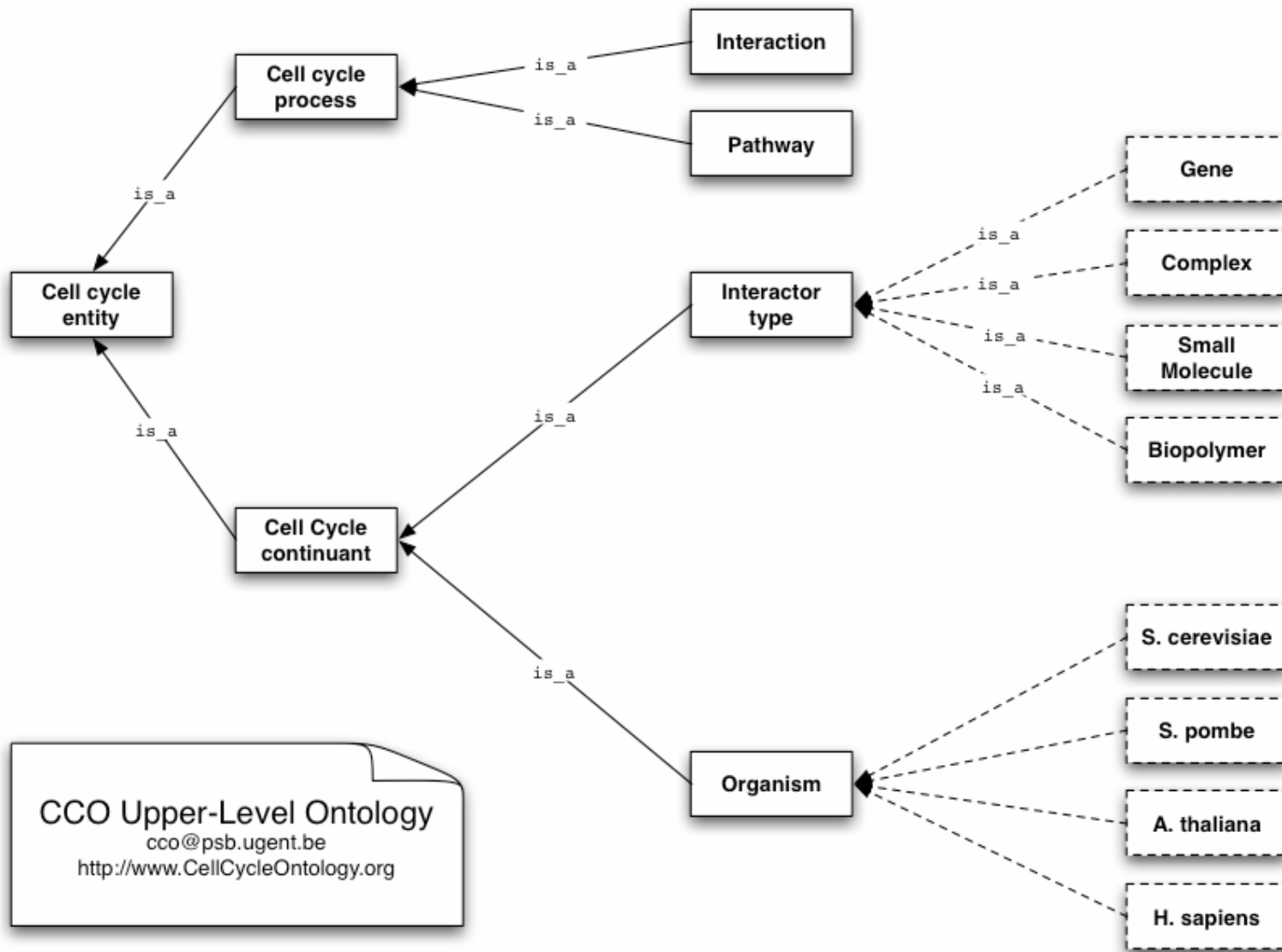


Data integration

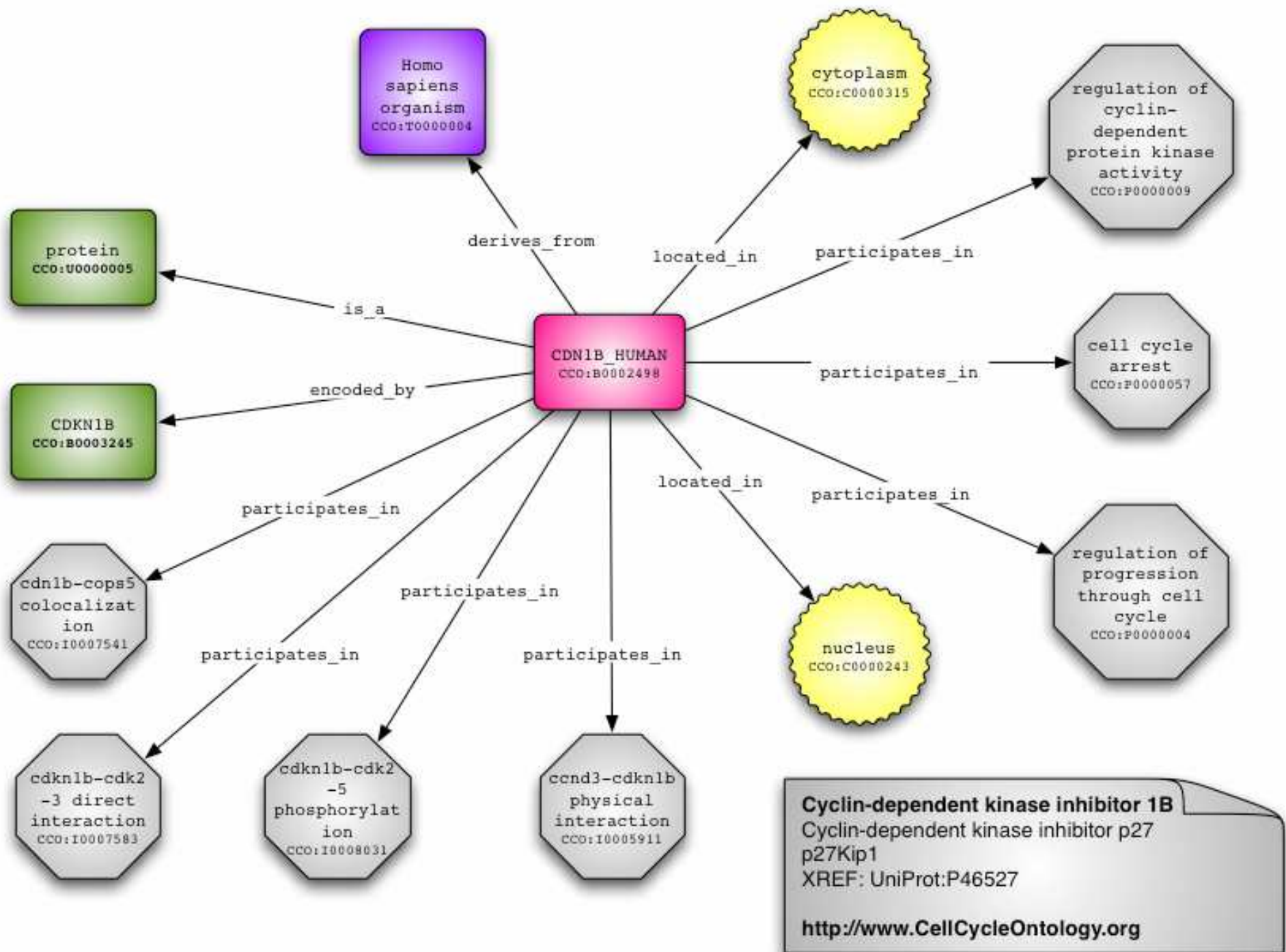
- ontology integration
- format mapping
- data integration
- data annotation
- consistency checking
- maintenance
- data annotation
- semantic improvement

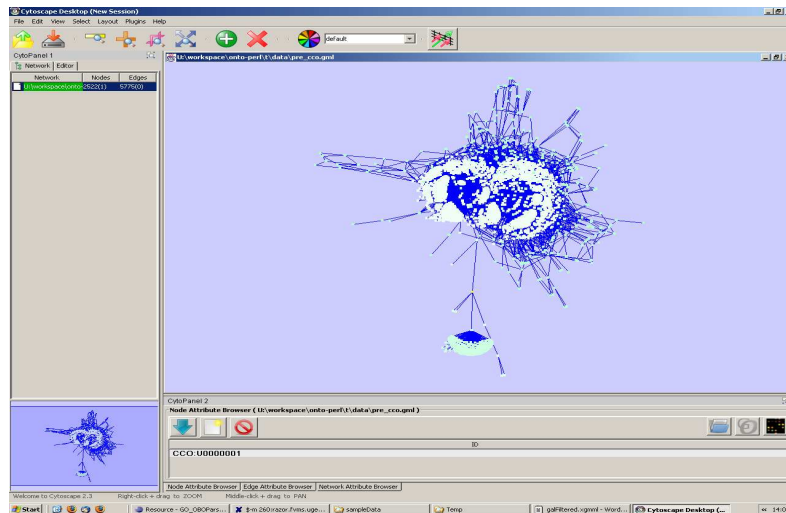
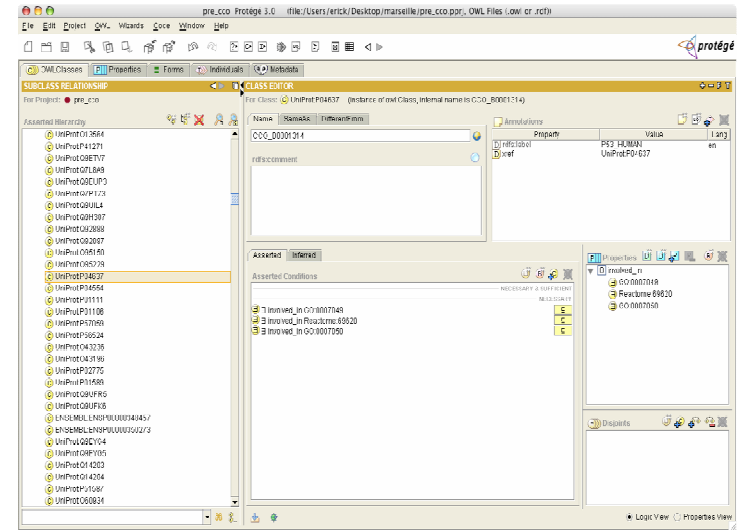


Upper Level Ontology*



* Based on BFO: enduring vs. perdurant





Cell Cycle Ontology Explorer

Look up terms in the Cell Cycle Ontology

Start typing a term name in the text box below. A list of prospective matches will pop-up, either choose the one you are looking from that list or continue typing the entire name. The pointed out term will be displayed on the right pane. Type for instance **mitotic cell cycle** and then click on the button named 'Look up' to retrieve the entry that has as name **mitotic cell cycle**.

P53_HUMAN ?

Look up Reset

id: COO:B0001314
name: P53_HUMAN
is_a: [protein](#)
relationship: participates_in
type: participates_in target: [cell cycle](#)
target: [cell cycle arrest](#)

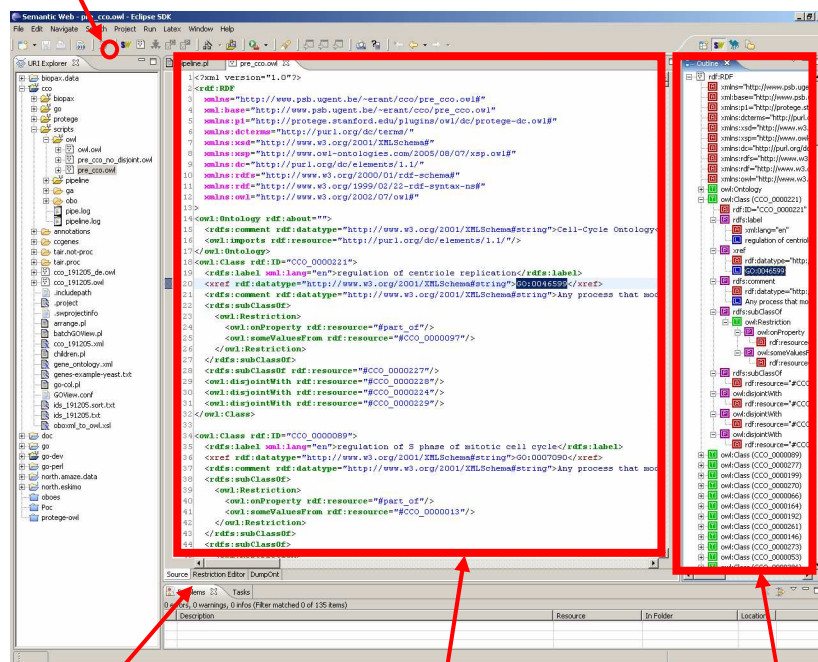
id: COO:B0001315
name: IFNW1_HUMAN
is_a: [protein](#)
relationship: participates_in
type: participates_in target: [cell cycle arrest](#)

id: COO:B0001316
name: INHA_HUMAN
is_a: [protein](#)
relationship: participates_in
type: participates_in target: [cell cycle arrest](#)

id: COO:B0001317

CCO checked with...

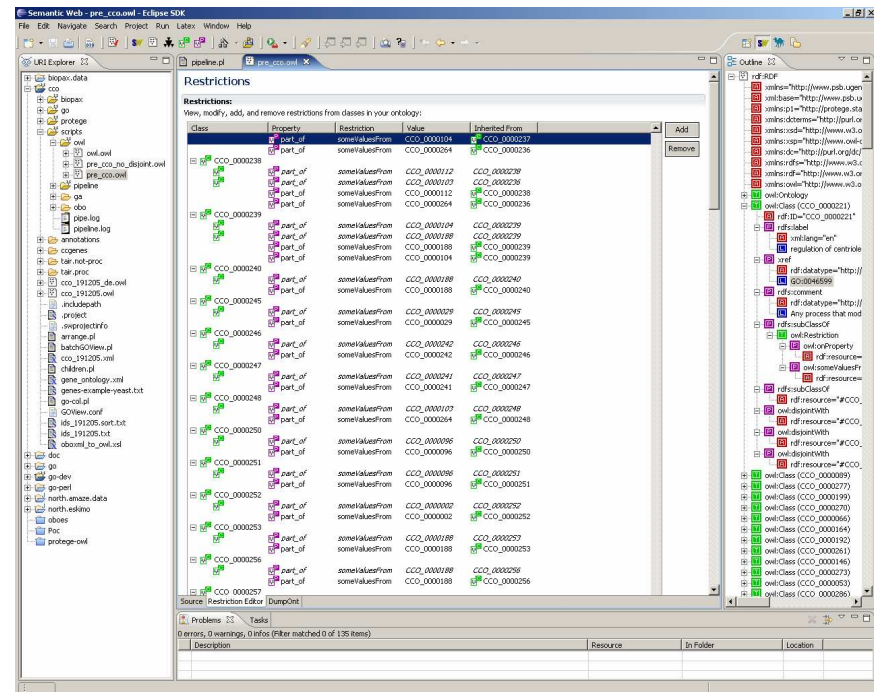
validator



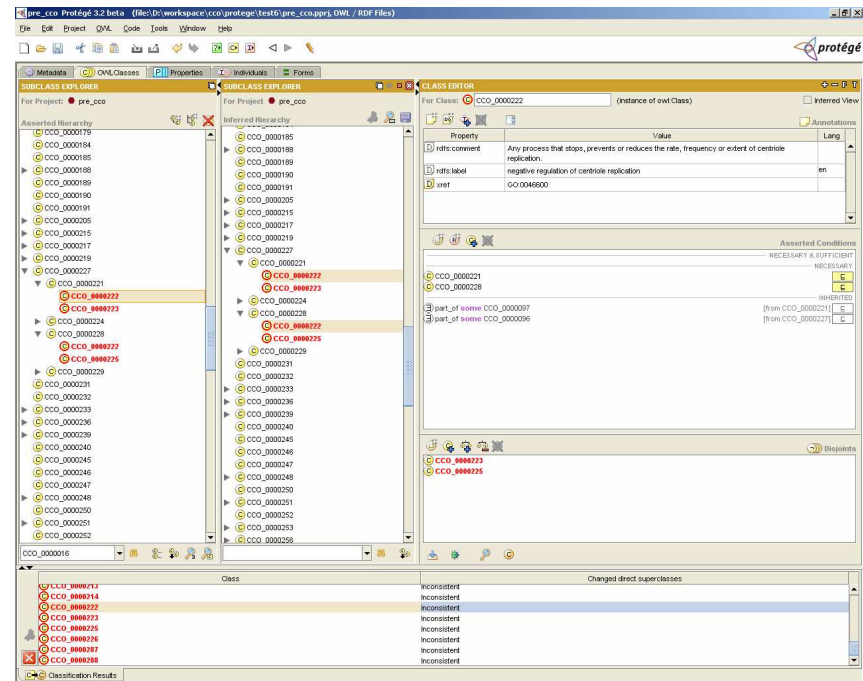
Source tab

editor

outline



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



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

and with...

```
C:\WINDOWS\system32\cmd.exe
D:\research\vowlidator>validate.bat D:\workspace\cco\scripts\owl\pre_cco.owl
D:\research\vowlidator>java -mx512m com.bbn.semweb.owl.vowlidator.Validator D:\workspace\cco\scripts\owl\pre_cco.owl
Loading Validator Preferences file: preferences.xml
loading file:D:\workspace\cco\scripts\owl\pre_cco.owl# to validate.
Reading referenced namespaces...
Loading URI file:cache/purl.org_dc_elements_1.1
instead of http://purl.org/dc/elements/1.1/#
Loading URI file:cache/www.w3.org_2000_01_rdf-schema
instead of http://www.w3.org/2000/01/rdf-schema#
Loading URI file:cache/www.w3.org_2002_07_owl
instead of http://www.w3.org/2002/07/owl#
Loading URI file:cache/www.w3.org_1999_02_22-rdf-syntax-ns
instead of http://www.w3.org/1999/02/22-rdf-syntax-ns#
Validating Referenced External Resources...
Validating Referenced Internal Resources...
Validating Model Statements...
Validating Model Nodes...
=====
BBN OWL Validator version 20050526
For the latest version visit
http://projects.semwebcentral.org/projects/vowlidator/
=====
The following Indications were found for D:\workspace\cco\scripts\owl\pre_cco.owl:

[1] INFORMATION - Substituted Files: The following file substitutions were made by the OWL Validator:
http://www.w3.org/1999/02/22-rdf-syntax-ns#
-> file:cache/www.w3.org_1999_02_22-rdf-syntax-ns
http://www.w3.org/2002/07/owl#
-> file:cache/www.w3.org_2002_07_owl
http://www.w3.org/2000/01/rdf-schema#
-> file:cache/www.w3.org_2000_01_rdf-schema
http://purl.org/dc/elements/1.1/#
-> file:cache/purl.org_dc_elements_1.1

[2] INFORMATION - Loaded Files: The following files were loaded by the OWL Validator to support validation:
file:cache/purl.org_dc_elements_1.1
file:cache/www.w3.org_2000_01_rdf-schema
file:cache/www.w3.org_2002_07_owl
file:cache/www.w3.org_1999_02_22-rdf-syntax-ns
Recursive loading of imported ontologies is ON -- maximum depth is infinite
Recursive loading of referenced namespaces is OFF

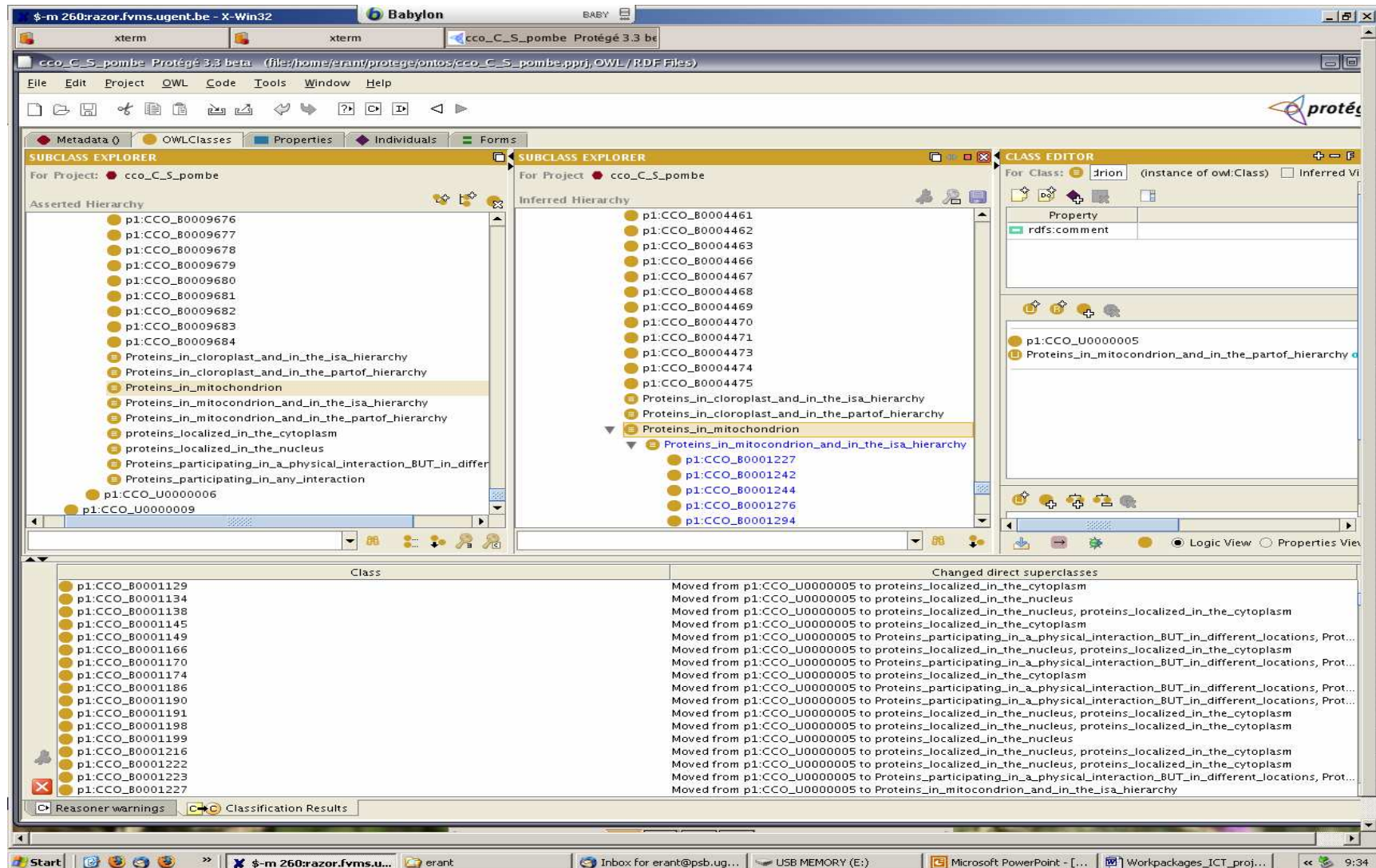
[3] WARNING - Range Type Mismatch: Use of this property implies that object is of type http://www.w3.org/2002/07/owl#Ontology.
At [http://www.psb.ugent.be/~erant/cco/pre_cco.owl, http://www.w3.org/2002/07/owl#imports, http://purl.org/dc/elements/1.1/1 line 16

D:\research\vowlidator>_
```

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*)

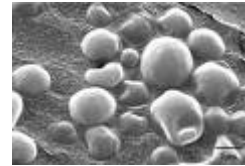


The screenshot displays the Protégé 3.3.3 interface with the following components:

- Subclass Explorer (Left):** Shows the asserted hierarchy for the project `cco_C_S_pombe`. It lists various protein classes, including `p1:CCO_80009676` through `p1:CCO_80009684`, and localization-related classes like `Proteins_in_chloroplast_and_in_the_isa_hierarchy`, `Proteins_in_mitochondrion`, and `Proteins_in_mitochondrion_and_in_the_isa_hierarchy`.
- Subclass Explorer (Right):** Shows the inferred hierarchy for the same project. It lists classes like `p1:CCO_80004461` through `p1:CCO_80004475`, and `Proteins_in_mitochondrion_and_in_the_isa_hierarchy` with its subclasses `p1:CCO_80001227` through `p1:CCO_80001294`.
- Class Editor (Right):** Displays the properties for the selected class `p1:CCO_U0000005`, showing the `rdfs:comment` property.
- Bottom Panel:** Shows a list of classes on the left and a table of "Changed direct superclasses" on the right. The table lists changes such as "Moved from p1:CCO_U0000005 to proteins_localized_in_the_cytoplasm" and "Moved from p1:CCO_U0000005 to Proteins_in_mitochondrion_and_in_the_isa_hierarchy".

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>



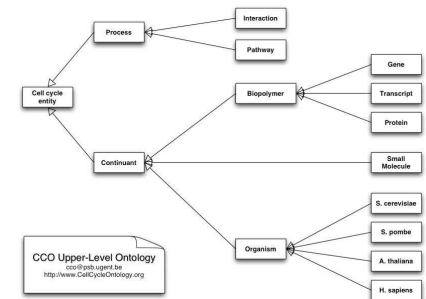
Products

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

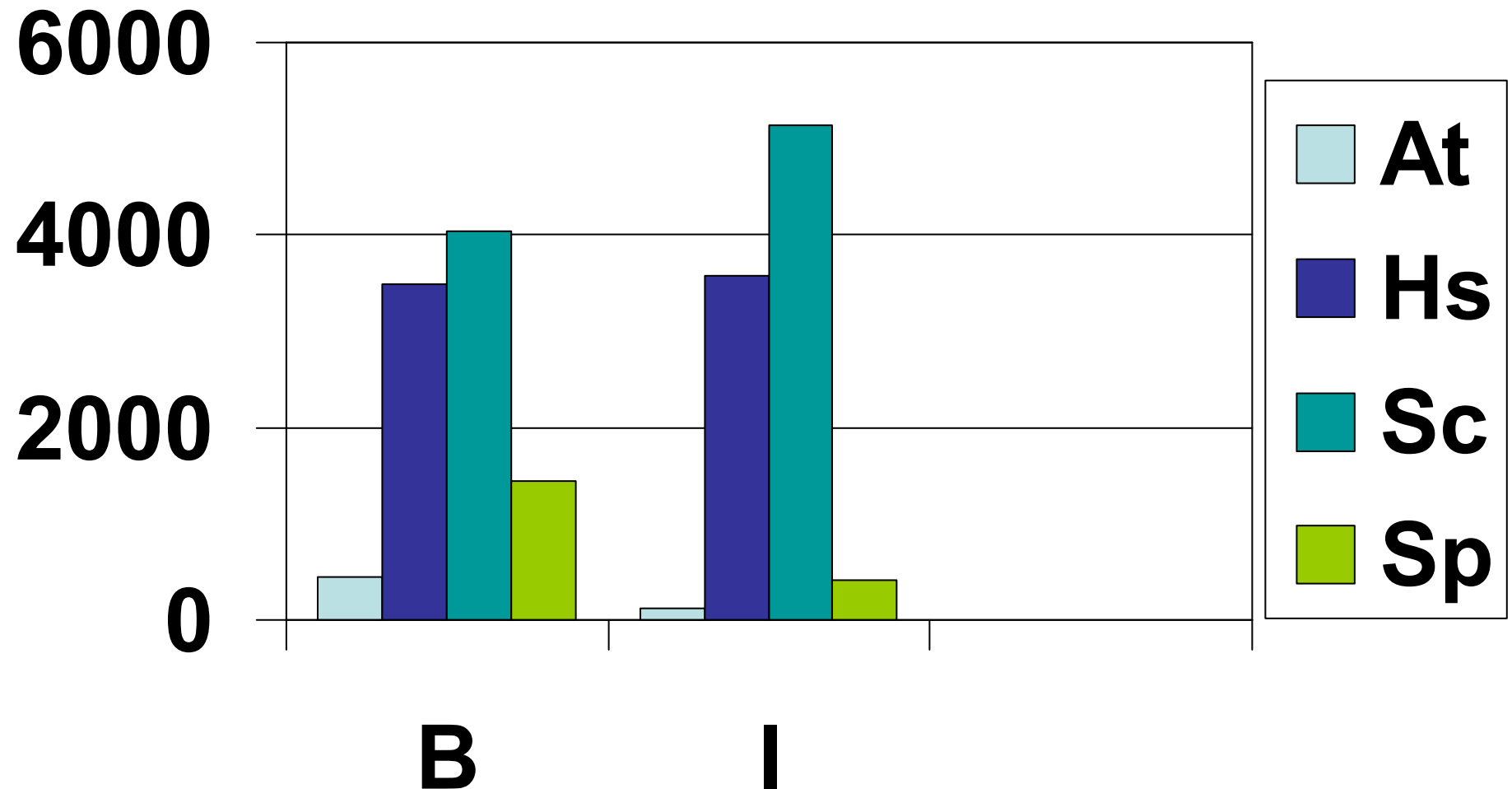
* Under development

CCO figures

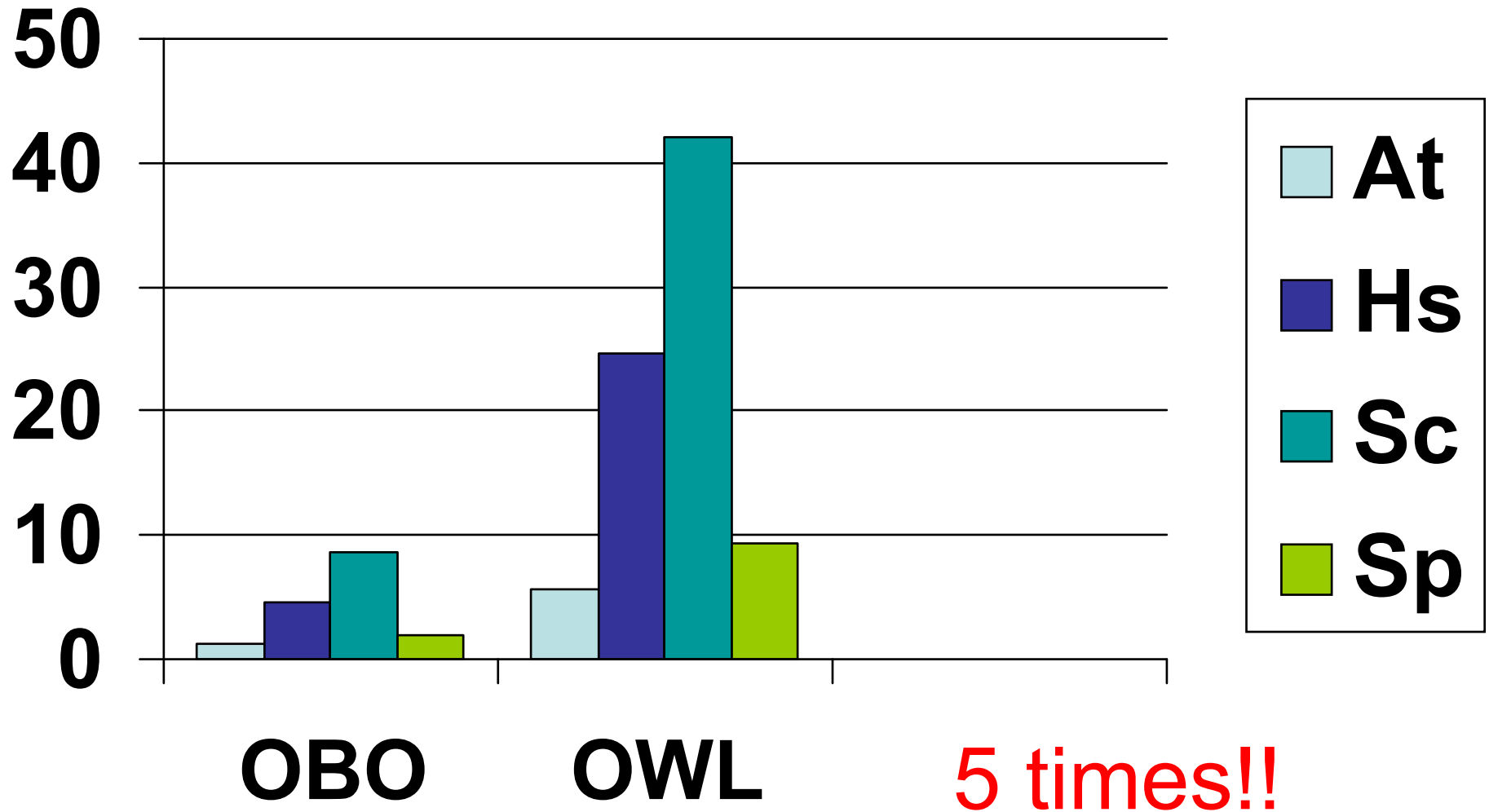
- Terms/classes:
 - B: 9392 (genes + proteins : GOA + UniProt)
 - P: 322 (GO)
 - I : 9258 (only from IntAct)
 - R: 181 (GO)
 - C: 1937 (filter out)
 - T: 60 (NCBI)
- #rel's
 - $\#RO + \#CCO = 15 + 5 = 20$
- Total: 21213



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.