Data-driven and hypothesis-driven research in (Semantic) Systems Biology

Erick Antezana
Dept. of Plant Systems Biology
Flanders Institute for Biotechnology (VIB) / Ghent University
Ghent - BELGIUM
erant@psb.ugent.be

Contents

Systems Biology

- Yet another definition
- Key: system
- What is a system?
- System =
  - set of elements,
  - dynamically interrelated,
  - having an activity,
  - to reach an objective (sub-aims),
  - **INPUT**: data/energy/matter
  - **OUTPUT**: information/energy/matter

Systems Biology (cont)

- “A system (and its properties) cannot be described in terms of their terms in isolation; its comprehension emerges when studied globally”
- Systems Biology = Approach to study biological systems.
- Arbitrary borders
- A system within a system

Systems Biology (cont)

- Types of systems biology:
  - “Standard/Classical” Systems Biology (Kitano, Science 2002. Sauer et al, Science 2007.)
  - Translational Systems Biology (Vodovotz, PLoS Comp Biol 2008.)
  - Semantic Systems Biology (Our proposed paradigm)

Semantic Systems Biology

- Semantic?
  - New emerging technologies for analyzing data and formalizing knowledge extracted from it
- A new paradigm elements:
  - Knowledge representation
  - Reasoning ==> hypothesis
  - Querying

Systems biology paradigm

top-down and bottom-up modeling

top-down
data driven

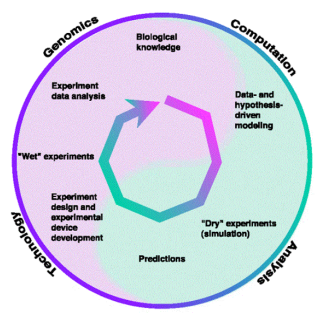
bottom-up
hypothesis driven

Biological Process

Genome-scale functional genomics data

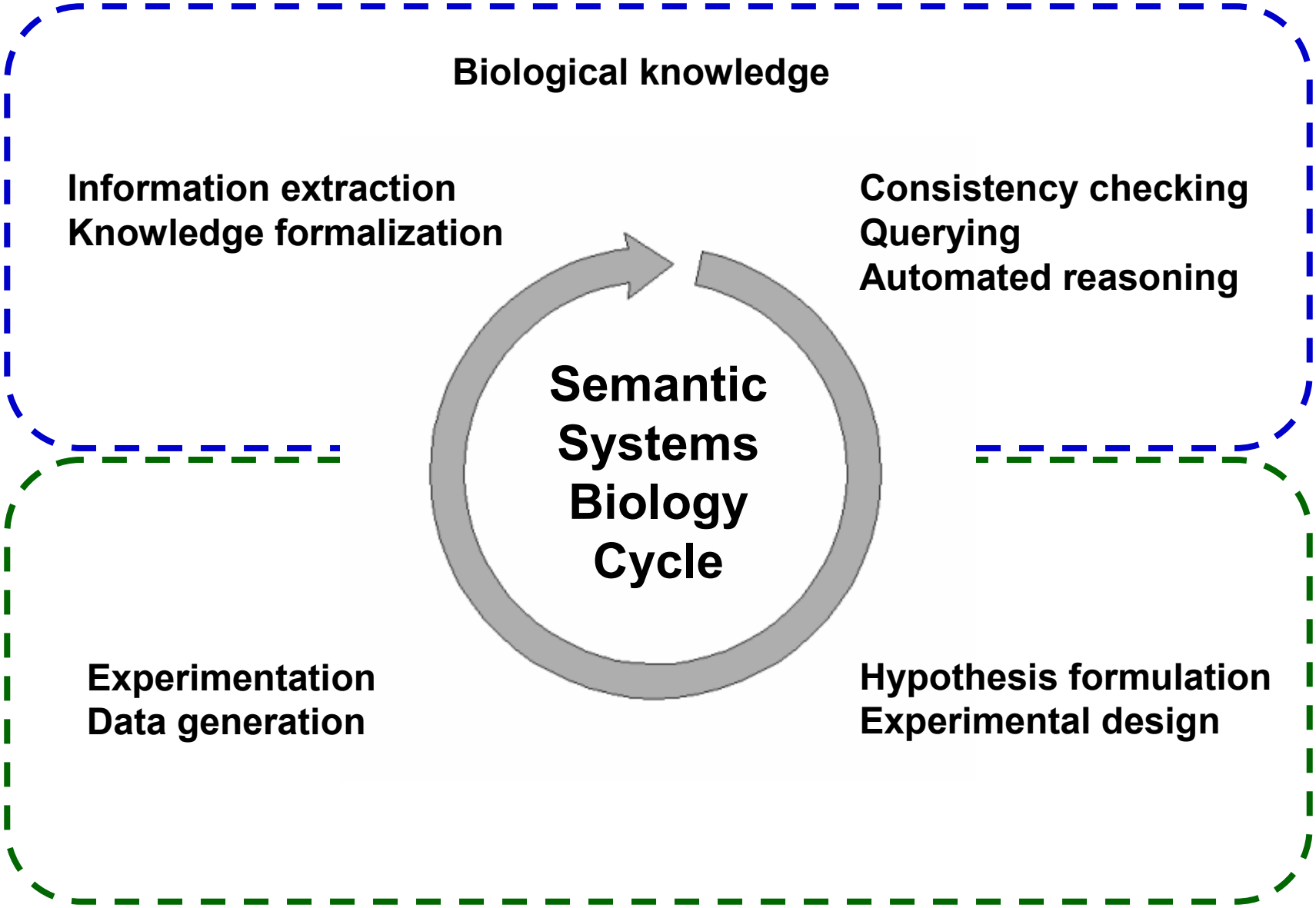
Predictive mathematical model

Statistics
Mining



Knowledge
Mathematics

Knowledge Management


Biological knowledge
Information extraction
Knowledge formalization
Consistency checking
Querying
Automated reasoning
Semantic Systems Biology Cycle
Experimentation
Data generation
Hypothesis formulation
Experimental design

In practice

- A knowledge base for cell cycle elucidation:
  - http://www.cellcycleontology.org
- **“BioGateway”:** an integrative approach for supporting semantic systems biology
  - http://www.semantic-systems-biology.org/

(Some) Motivating questions

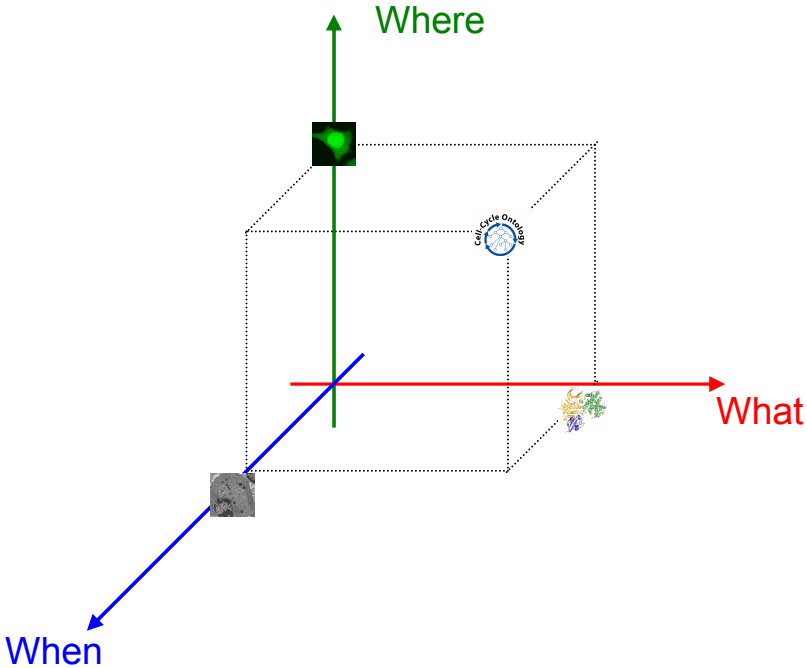
- I'm working with **AT5g35520**, in which interactions does this gene play a role?
- From my microarray experiment I've got this gene **X**, is this gene involved in the cell cycle, …, ?
- Verify my models of genetic, metabolic and product interaction networks
- …

Background

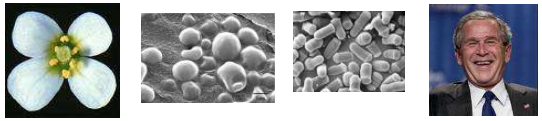
- Amount of data generated in biological experiments continues to grow exponentially
- Shortage of proper approaches or tools for analyzing this data has created a gap between raw data and knowledge
- Lack of a structured documentation of knowledge leaves much of the data extracted from these raw data unused
- Differences in the technical languages used (synonymy and polysemy) have complicated the analysis and interpretation of the data

The Cell-Cycle Ontology

- 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)"

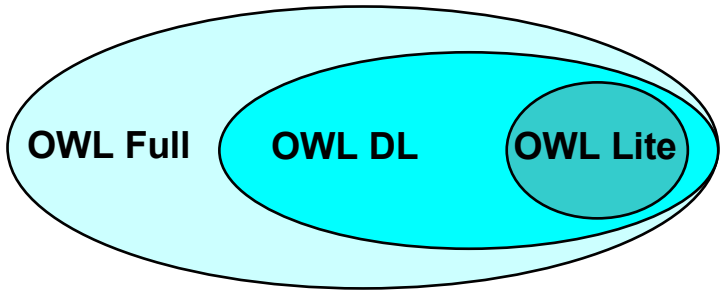
ORGANISMS: 

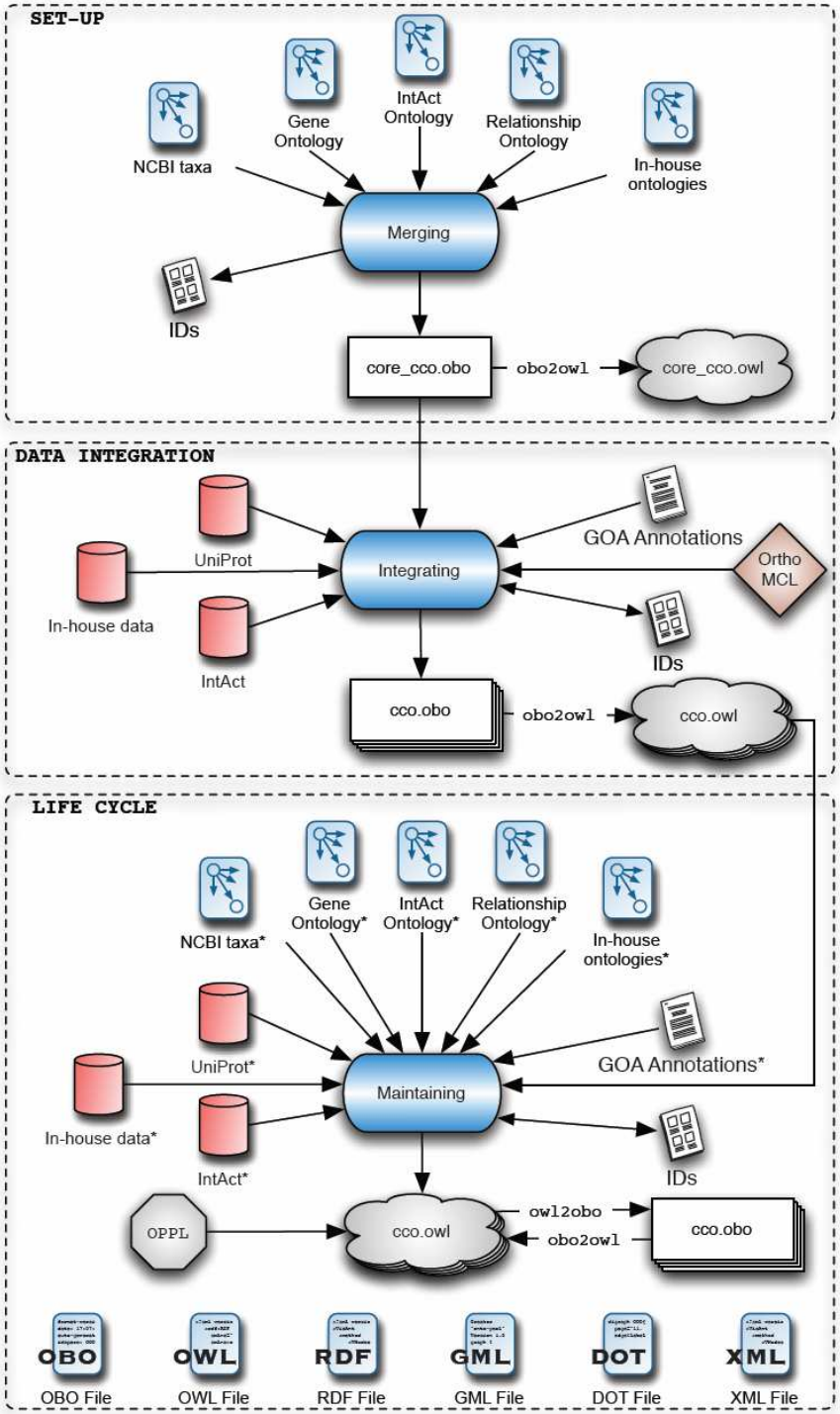
http://www.CellCycleOntology.org

Antezana et al. LNBI, 2006

Knowledge representation

- Why OBO?
  - “Human readable”
  - Standard
  - Tools (e.g. OBOEdit)
  - http://obo.sourceforge.net
- Why OWL?
  - Web Ontology Language
  - “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++





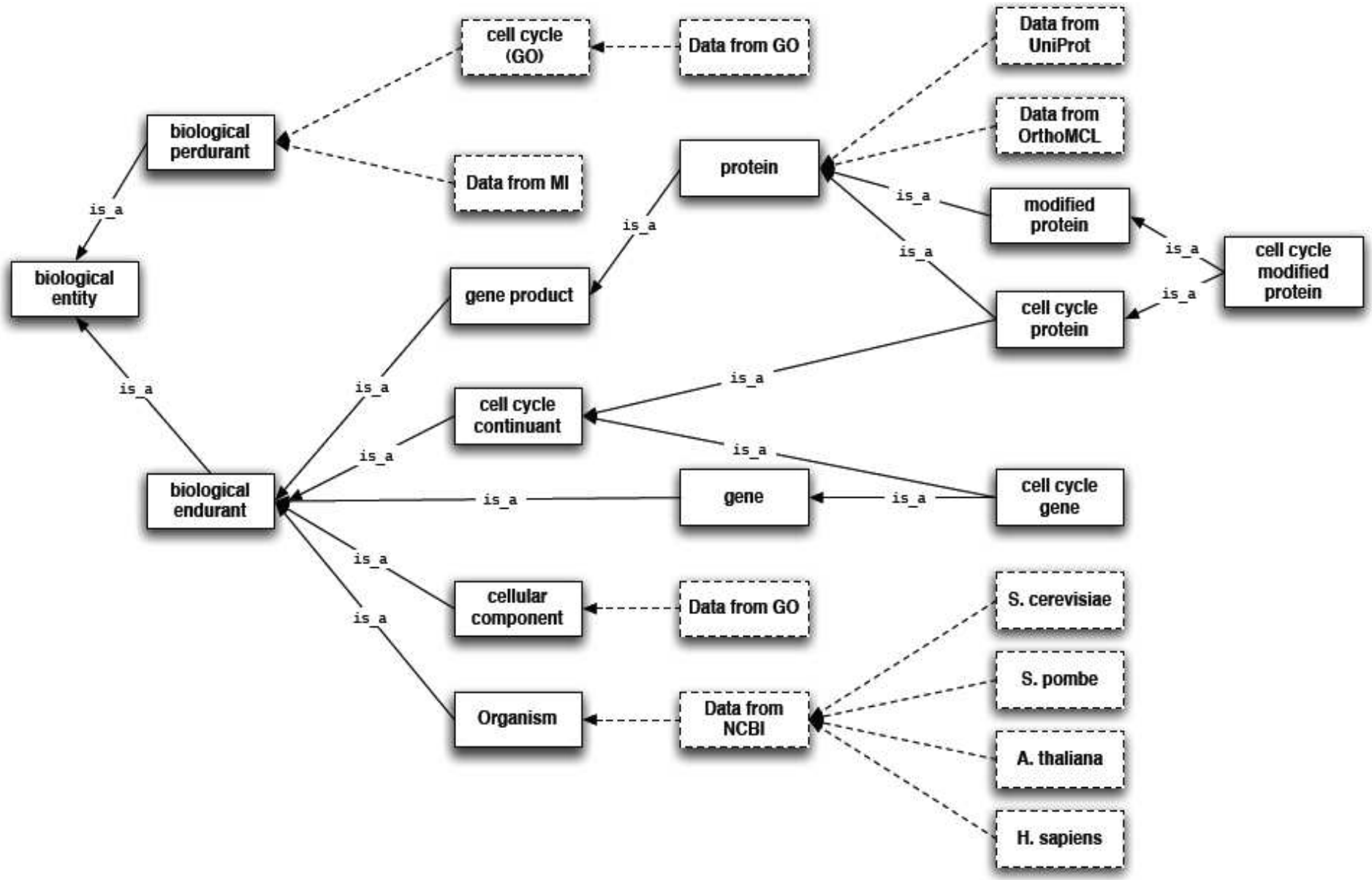
CCO Pipeline

- ontology integration
- format mapping

- data integration
- data annotation
- consistency checking

- maintenance
- data annotation
- semantic improvement: OPPL
- ODP (BMC BioInf – *in press*)

ULO


cell cycle (GO)
Data from GO
Data from UniProt
Data from OrthoMCL
biological perdurant
Data from MI
protein
is_a
modified protein
is_a
is_a
is_a
cell cycle modified protein
biological entity
gene product
is_a
cell cycle protein
is_a
is_a
is_a
cell cycle continuant
is_a
is_a
biological endurant
is_a
gene
is_a
cell cycle gene
is_a
cellular component
Data from GO
S. cerevisiae
is_a
S. pombe
Organism
Data from NCBI
A. thaliana
H. sapiens

CCO accession number

CCO:[CPFRTIBGOU]nnnnnnn

namespace

sub-namespace

7 digits

```
C: cellular component
P: biological process
F: molecular function
R: reference
T: taxon
I: interaction
B: protein
G: gene
O: ortholog
U: upper-level term
```

- **Examples in CCO:**

  CCO: P0000056 ⟷ "*cell cycle*"

  CCO: B0000046 ⟷ "CYCA3;2"

- **In other ontologies:**

  OBO_REL: has_participant

  GO:0007049 ⟷ "*cell cycle*"

Sample entry in OBO

```
[Term]
id: CCO:B0002060
name: NEB2_HUMAN
def: "Neurabin-2" [UniProt:Q96SB3]
synonym: "Neurabin-II" EXACT [UniProt:Q96SB3]
xref: UniProt:Q8TCR9
is_a: CCO:B0000000 ! core cell cycle protein
relationship: belongs_to CCO:T0000004 ! Homo sapiens organism
relationship: encoded_by CCO:G0005171 ! PPP1R9B_human
relationship: participates_in CCO:I0006401 ! aah62584-q96sb3 physical interaction
relationship: transforms_into CCO:B0013139 ! NEB2_HUMAN-Phosphoserine15
```

OBO2OWL Mapping: http://www.bioontology.org/wiki/index.php/OboInOwl:Main_Page
Tool: ONTO-PERL (*Antezana et al. Bioinformatics 2008*)

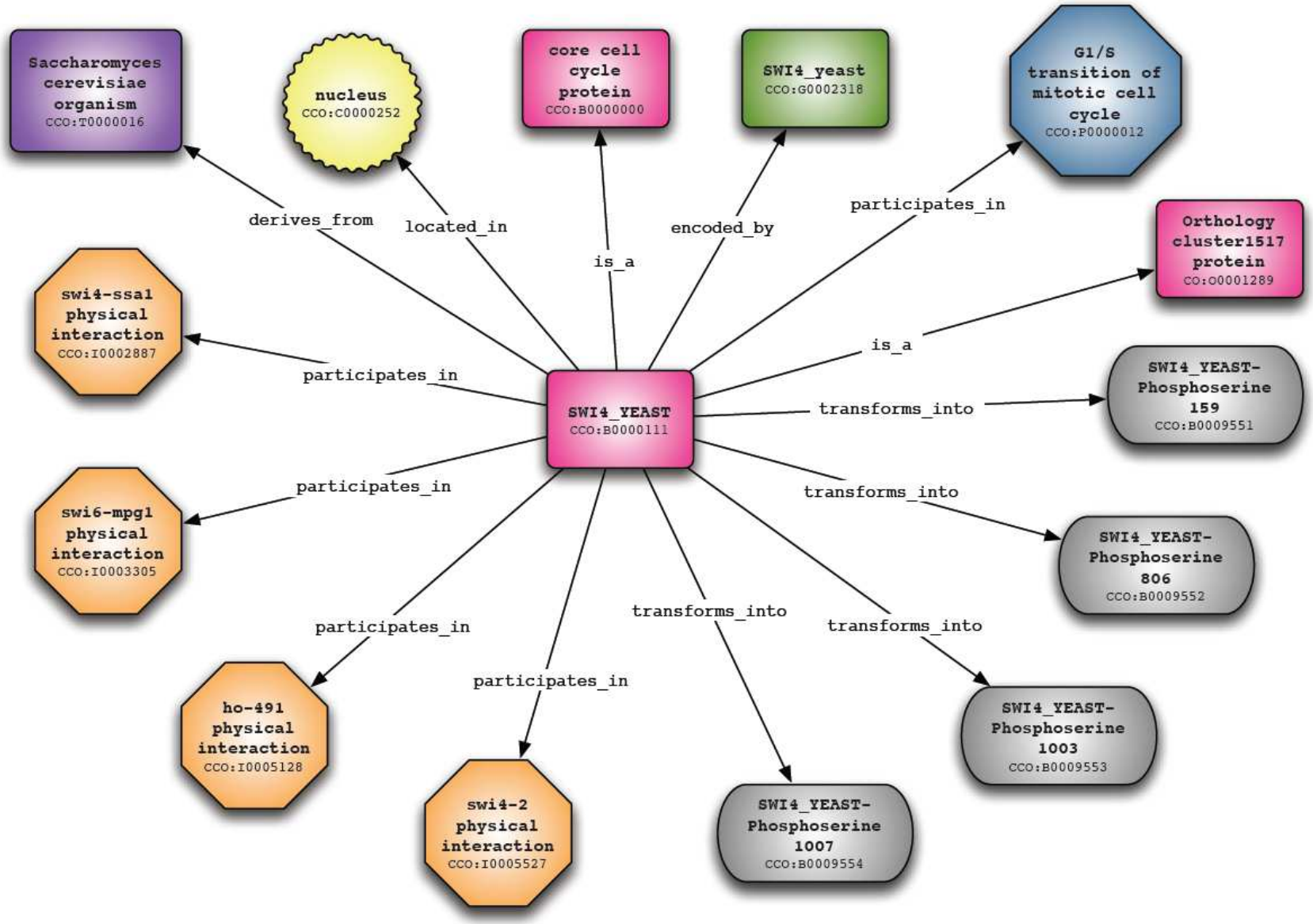
Some figures

Entity	Ontology				
	At	Hs	Sc	Sp	CCO
Proteins	252	5829	7069	930	24541
Genes	222	1806	3148	852	6028
Interactions	76	2394	5162	399	8031
Orthology groups	—	—	—	—	1649

CCO is the composite ontology = At + Hs + Sc + Sp + orthology

2008-03-07: 49226 terms in CCO

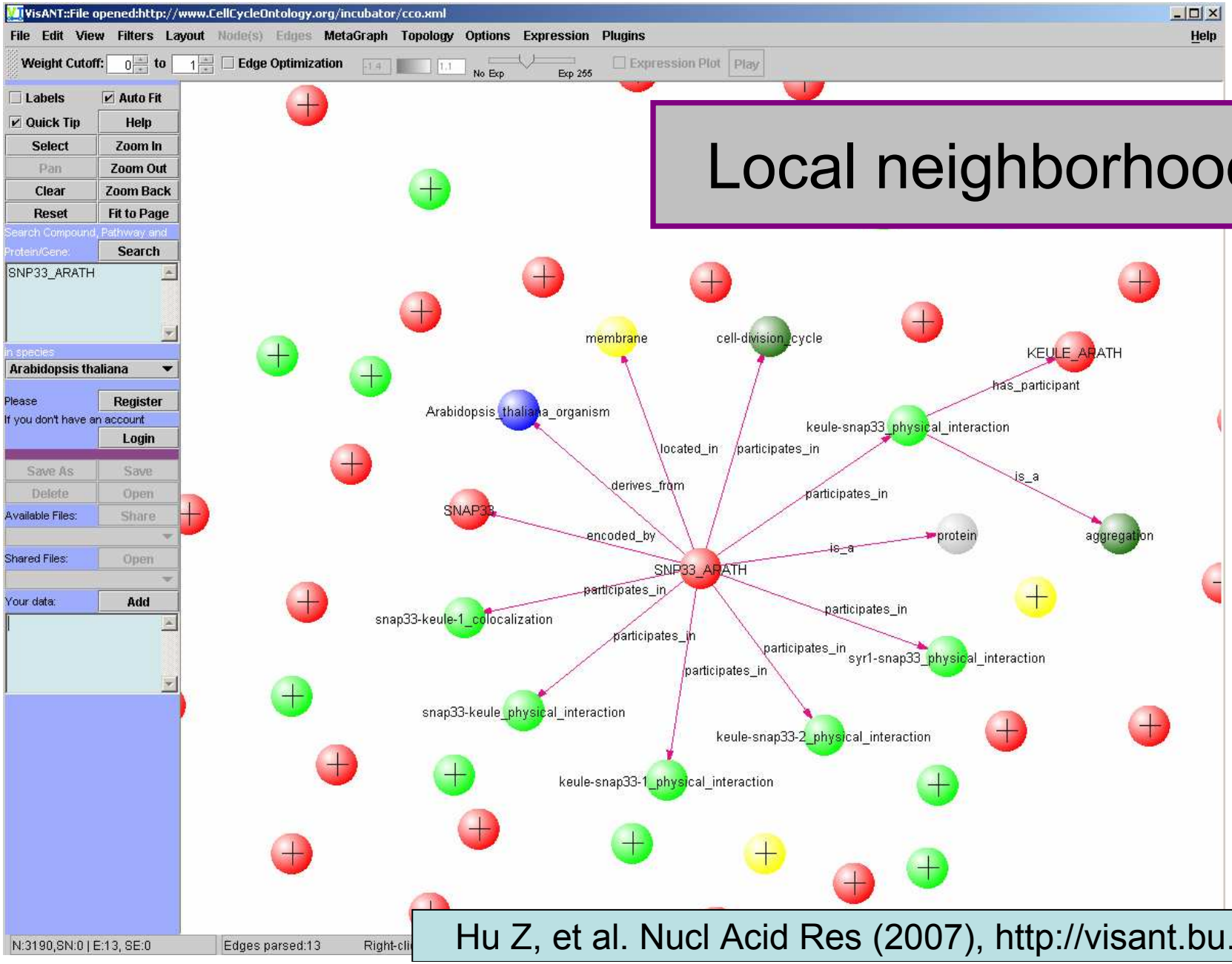
Current knowledge



Knowledge exploration

- Looking up:
  - Terms,
  - Synonyms,
  - …
- Visual browsing
  - “local neighborhood”
  - Path to the root
- Advanced Querying (e.g. SPARQL)

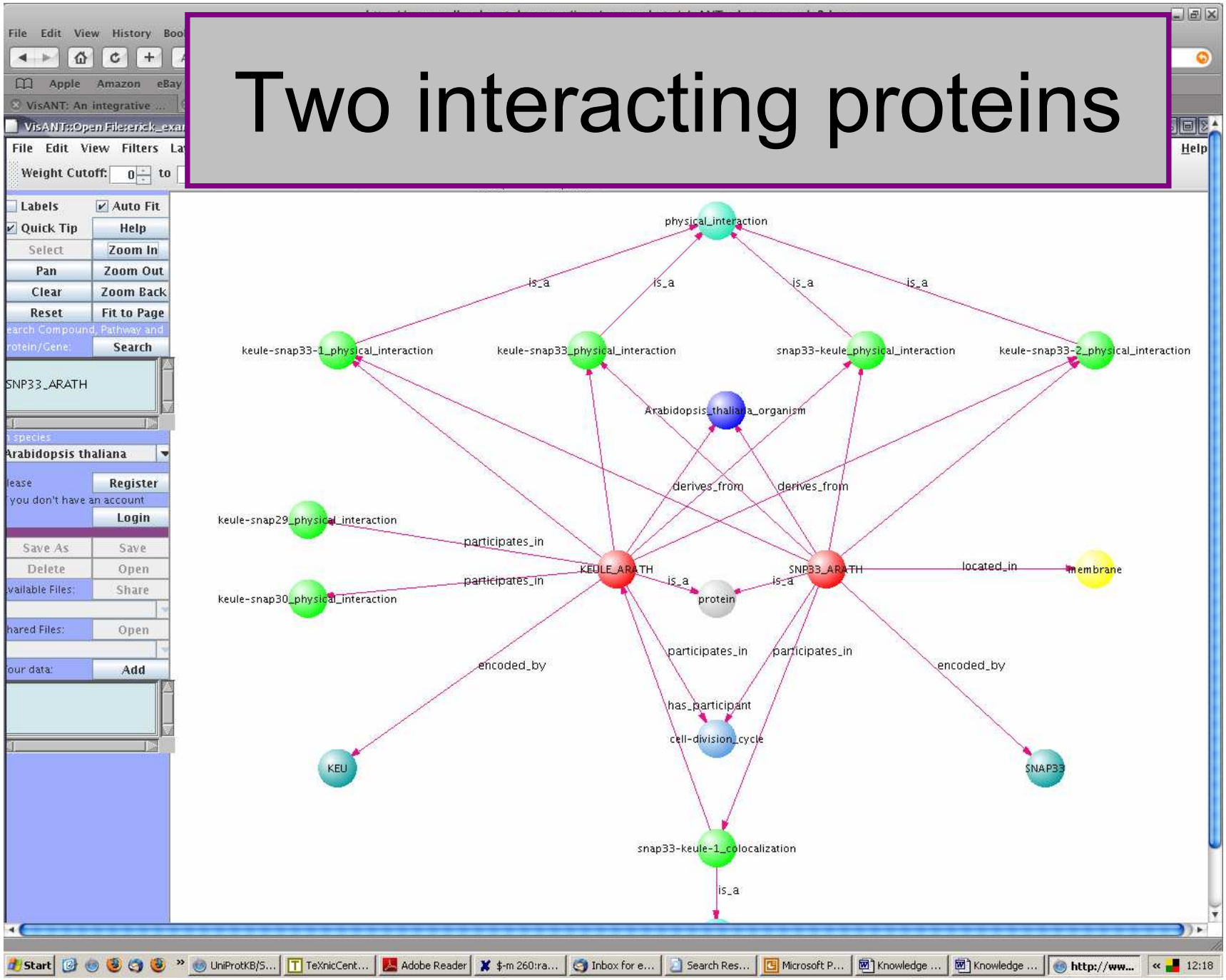
CCO in: visANT



Local neighborhood

Hu Z, et al. Nucl Acid Res (2007), http://visant.bu.edu/

Two interacting proteins



THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

BioPortal

Version **1.0**

Browse | Search

Home | Sign In | Register | Help

Cell Cycle Ontology (A. thaliana)

Tree View

Tree view constructed based on *is_a* hierarchy

- entity
  - continuant
    - cell cycle continuant
    - cellular_component
    - gene
    - gene product
      - protein
        - cell cycle modified protein
        - cell cycle protein
        - core cell cycle protein
          - APC10_ARATH
          - ARTE_ARATH
          - ATK1_ARATH
          - ATK3_ARATH
          - ATM_ARATH
          - ATR_ARATH
          - BRE1A_ARATH
          - BSH_ARATH
          - CCA11_ARATH
          - CCA12_ARATH
          - CCA21_ARATH

Class/Type Details

General

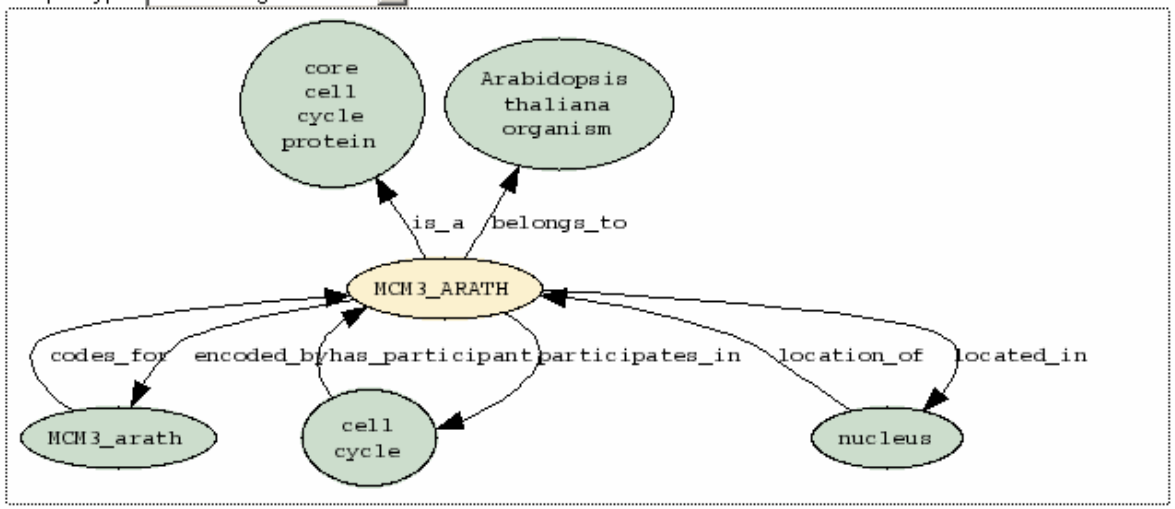
Class/Type Name	**MCM3_ARATH**
Id	**CCO:B0002385**

Attributes

Definition	**DNA replication licensing factor MCM3 homolog**
EXACT SYNONYM	**"Minichromosome maintenance protein 3 homolog"**
Database_References	**UniProt:Q9FL33**

Graph View

Graph Type: Local Neighborhood



http://www.bioontology.org/ncbo/faces/pages/ontology_list.xhtml

http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=CCO

ftp://ftp.gen...oPF_noXP.obo | Park 79 Hotel Reservations | BioPortal | SPARQL | Ontology Lookup Servic...

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

OLS

- OLS Home
- Documentation
  - Project
  - Publications
- Developer Resources
  - Download
  - Implementation Overview
  - Javadoc
  - Webservice documentation
- Contact Us
  - Acknowledgements

News

June 2007: Maintenance Release

Implementation documentation has been updated to include more recent dependencies. Please note that unless stated otherwise in specific instances, newer versions of given depencies should work without issue.

April 2007: Maintenance Release

There is a new release of the OLS available, which includes mostly maintenance fixes and

OLS - Ontology Lookup Service

CCO Ontology Browser

- entity
  - continuant
    - cellular_component
    - reference
    - organism
    - gene product
      - protein
        - core cell cycle protein
          - SEM1_YEAST
          - CDC28_YEAST
          - TBB_YEAST
          - TOP1_YEAST
          - DNLI_YEAST
          - CDC25_YEAST
          - NOT2_YEAST
          - CDC37_YEAST
          - SC160_YEAST
          - PIS_YEAST
          - KIN28_YEAST
          - CDC7_YEAST
          - KAPA_YEAST
          - CDC31_YEAST
          - RAD1_YEAST
          - RAD52_YEAST
          - TOP2_YEAST
          - CALM_YEAST
          - RAD10_YEAST
          - IF4E_YEAST

Legend:

is a | part of

Help (hide)

Double-click a term to see its children. The ontology browser is populated dynamically. If there are many children for a given term, there may be a small delay while the browser fetches **Click** to highlight a term to see any information associated with it. **Hover** over a term to see its relation with its immediate parent. Root terms will not display any relational information.

Relations

SEM1_YEAST is_a core cell cycle protein

Term Information

ID: | Zoom

Name:

Associated information

Highlight a term to view its associated information.

Term Hierarchy

Paths to Root: | Child relationships:

Cote R, et al. BMC Bioinformatics (2006)

Advanced Querying

- RDF = Resource Description Framework
  - Metadata model: elements = resources
- It allows expressing knowledge about web resources in statements made of triples (basic information unit) :

RDF Triples

- Subject corresponds to the main entity that needs to be described.
- Predicate denotes a quality or aspect of the relation between the Subject and Object.
- "*The protein DEL1 is located in the nucleus*"

SPARQL*

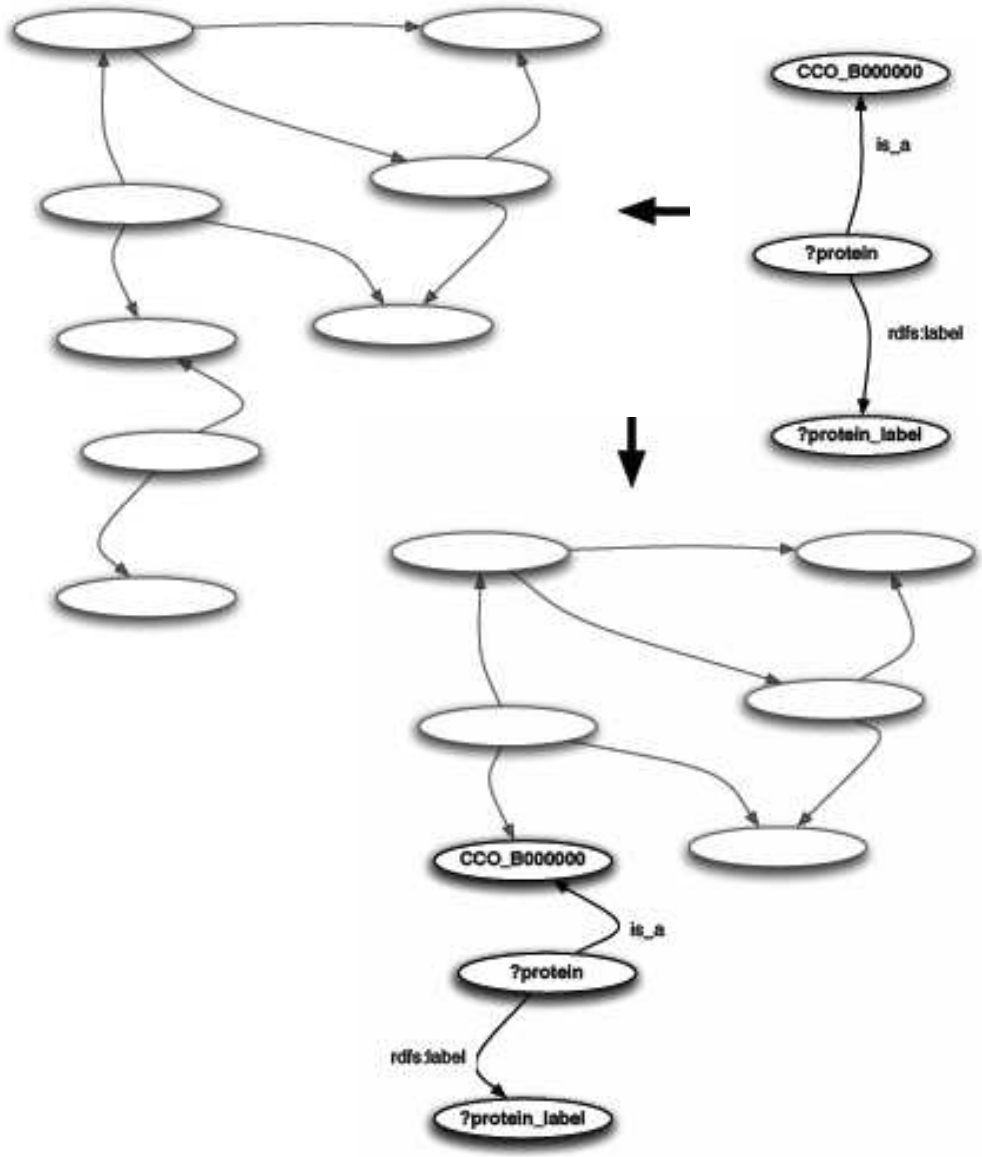
- Language which allows querying RDF models (graphs)
- Powerful, flexible
- Its syntax is similar to the one of SQL.
- Virtuoso Open Server
  - SPARQL queries
  - DB backend
  - …

http://www.openlinksw.com/virtuoso/

*http://www.w3.org/TR/rdf-sparql-query/

Matching triples

?protein **sp:is_a** sp:CCO_B0000000 .
?protein **rdfs:label** ?protein_label



http://www.cellcycleontology.org/query/sparql

Cell-Cycle Ontology

Home | Updates | Download | Query | Documentation | Tools | About

search... Search

SPARQL | OWL-DL | OLS | BioPortal

MAIN MENU

- Home
- Updates
- Download
- Query
  - SPARQL
  - OWL-DL
  - OLS
  - BioPortal
- Documentation
- Tools
- About

Home ▸ Query ▸ SPARQL

SPARQL

SPARQL stands for **SPARQL Protocol and RDF Query Language**. It is standardized by the *RDF Data Access Working Group* (DAWG) of the W3C. It allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns.

Querying CCO

The following form lets you query the Cell Cycle Ontology through a SPARQL endpoint hosted at Plant Systems Biology department of the Flanders Institute for Biotechnology. The underlying triplestore contains over 1 millon RDF triples of cell cycle information. This information ranges from processes, interactions, proteins, genes, cellular compartments, and so forth, which were collected from diverse sources (like GO, UniProt, IntAct, etc.). Type your SPARQL query in the following text area, then click on 'Run Query'. A new window with the results will be opened. In case there is a syntax error in the query, it will be warned to you. (**N.B.** Recommended browsers: Firefox, Safari, Opera, or Konqueror. IE proposes to save the results instead of displaying them.)

Query:

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX sp:<http://www.cellcycleontology.org/ontology/rdf/Sp#>
SELECT ?prot_name ?biological_process_name
FROM <http://www.cellcycleontology.org/ontology/rdf/Sp>
WHERE {

?prot sp:is_a sp:CCO_B0000000 .
?prot rdfs:label ?prot_name .
?prot sp:participates_in ?biological_process .
?biological_process rdfs:label ?biological_process_name
}
```

Run Query | Reset

SPARQL queries against CCO are run on Virtuoso (OpenLink). This system provides an infrastructure for storing and querying CCO.

Suggested PREFIXes:

http://www.cellcycleontology.org/query/sparql

Cell-Cycle Ontology

Home | Updates | Download | Query | Documentation | Tools | About

search... Search

SPARQL | OWL-DL | OLS | BioPortal

MAIN MENU

- Home
- Updates
- Download
- Query
  - SPARQL
  - OWL-DL
  - OLS
  - BioPortal
- Documentation
- Tools
- About

Home › Query › SPARQL

SPARQL

SPARQL stands for **SPARQL Protocol and R** … f the W3C. It allows for a query to consist of triple pattern

Querying CCO

The following form lets you query the Cell Cyc … nders Institute for Biotechnology. The underlying … esses, interactions, proteins, genes, … ts, … , etc.). Type your SPARQL query in the following … on 'Run … is a syntax error in the query, it will be warned to you … ed browsers: Firefo… ave the results instead of displaying them.)

"all the core cell cycle proteins (S.pombe) participating in a known process"

Query:

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX sp:<http://www.cellcycleontology.org/ontology/rdf/Sp#>
SELECT ?prot_name ?biological_process_name
FROM <http://www.cellcycleontology.org/ontology/rdf/Sp>
WHERE {

?prot sp:is_a sp:CCO_B0000000 .
?prot rdfs:label ?prot_name .
?prot sp:participates_in ?biological_process .
?biological_process rdfs:label ?biological_process_name
}
```

Run Query | Reset

SPARQL queries against CCO are run on Virtuoso (OpenLink). This system provides an infrastructure for storing and querying CCO.

Suggested PREFIXes:

prot_name	biological_process_name
UBC11_SCHPO	G2%2FM transition of mitotic cell cycle
UBC11_SCHPO	cell cycle
UBC11_SCHPO	mitosis
UBC11_SCHPO	mitotic metaphase%2Fanaphase transition
UBC11_SCHPO	regulation of mitotic cell cycle
UBC11_SCHPO	cyclin catabolic process
SRW1_SCHPO	cell cycle
SRW1_SCHPO	cyclin catabolic process
SRW1_SCHPO	activation of anaphase-promoting complex during mitotic cell cycle
SRW1_SCHPO	cell cycle arrest in response to nitrogen starvation
SRW1_SCHPO	negative regulation of cyclin-dependent protein kinase activity
DYHC_SCHPO	dhc1-peg1-1 physical interaction
DYHC_SCHPO	synapsis
DYHC_SCHPO	meiotic recombination
DYHC_SCHPO	horsetail nuclear movement
ORB6_SCHPO	cell morphogenesis checkpoint
ORB6_SCHPO	regulation of cell cycle
DED1_SCHPO	G2%2FM transition of mitotic cell cycle

Reasoning over CCO

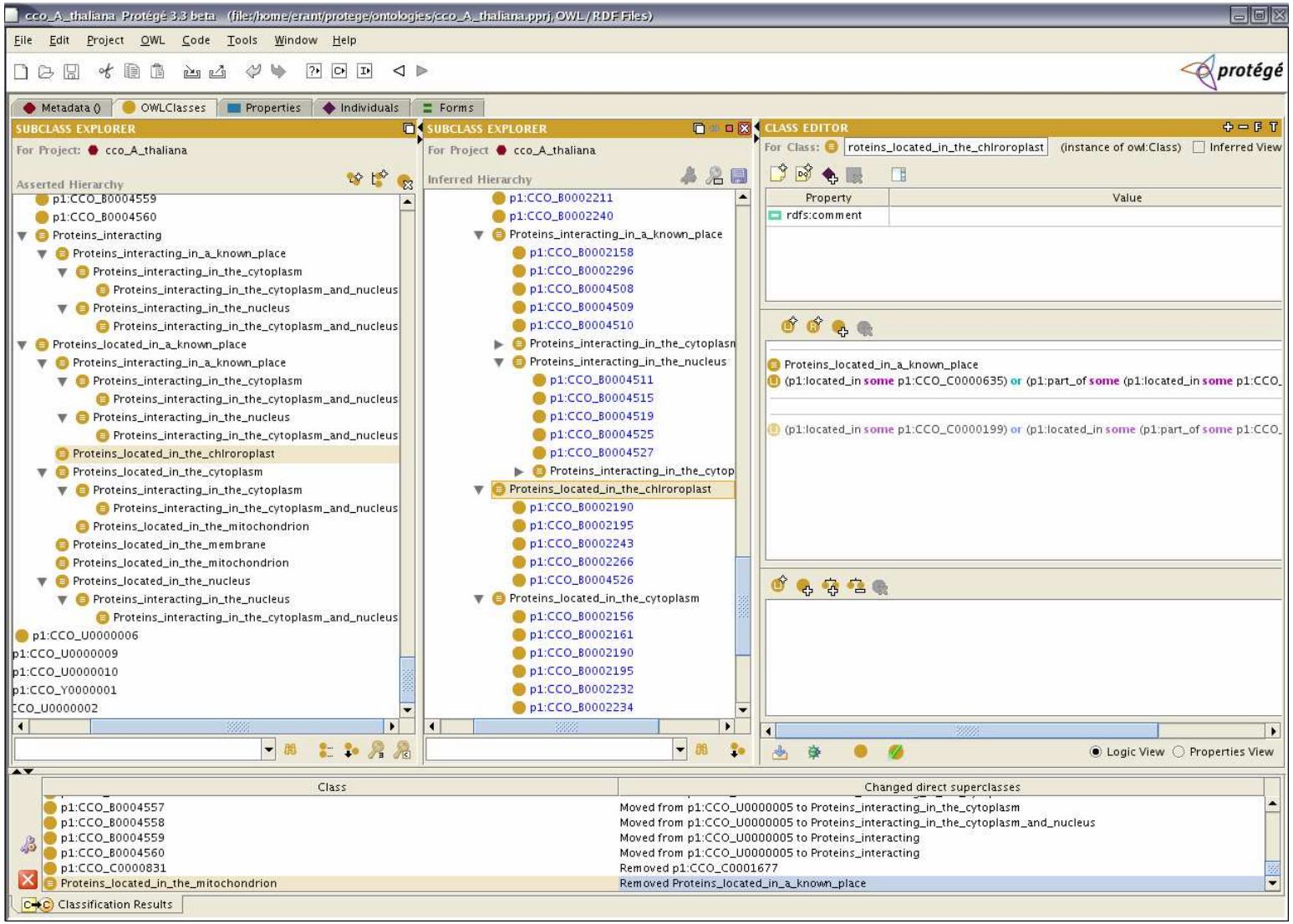
- Consistency checking: no contradictory facts
- Classification: implicit2explicit knowledge
- Querying (OWL-DL)

OWL

- Web ontology language
- OWL-DL: balance tractability with expressivity
- Open World Assumption
  - “what is not stated is not false, it is unknown”
  - Fits in Biology
- Tools:
  - Protégé (http://protégé.stanford.edu)
  - Reasoners (RACERPRO, Pellet, etc)

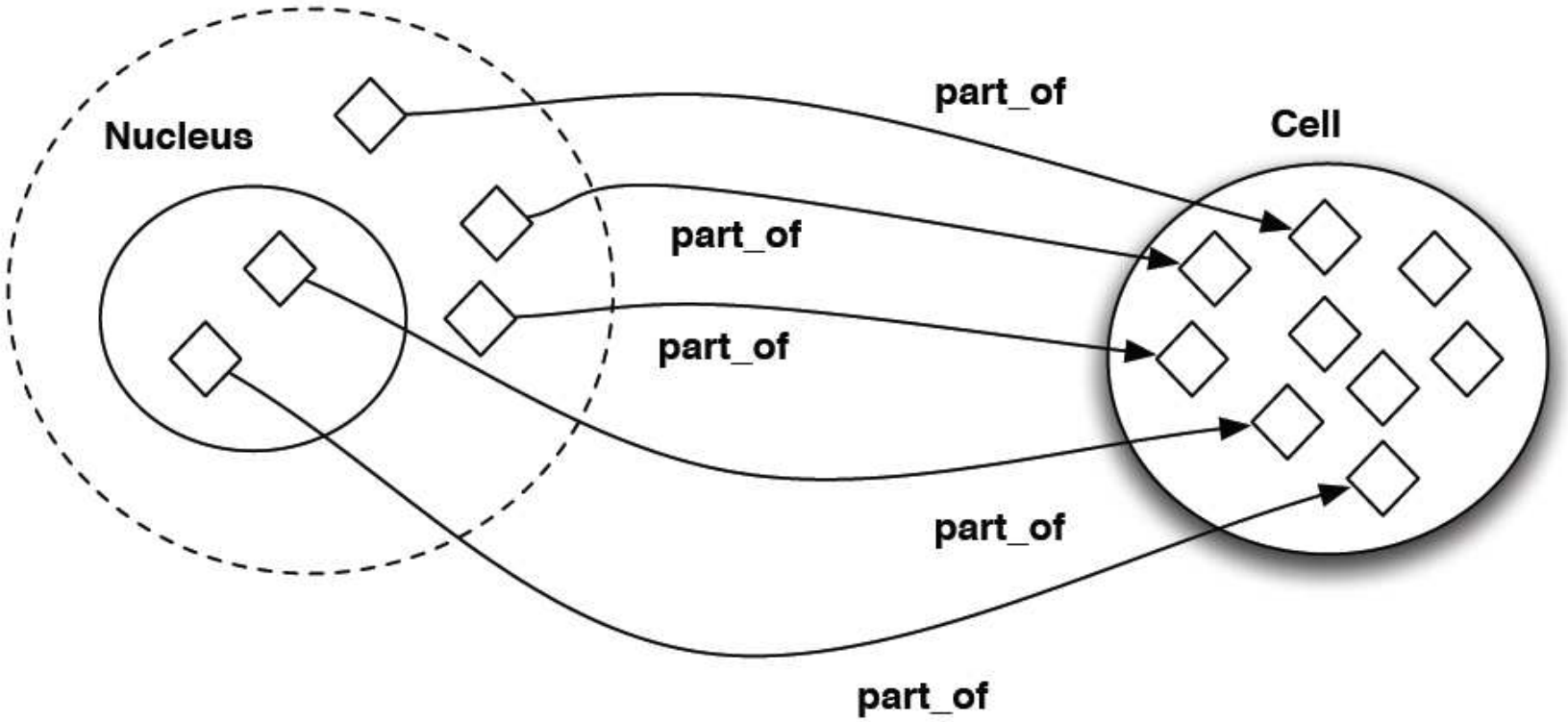
Cellular localization checks

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



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

OWL restrictions



Restriction on Nucleus: some **part_of** Cell

Necessary conditions vs Necessary and sufficient conditions

Sample query in OWL (1)

- Which cell cycle related proteins participate in a reported interaction?

  CCO_U0000005 **and**

  participates_in **some** CCO_Y0000001
- CCO_U0000005 = class of proteins
- CCO_Y0000001 = interactions

Sample query in OWL (2)

- Entities that are the location of proteins participating in the S-phase (CCO_P0000014) or any process which is part of it.?

location_of **some** (
  participates_in **some** (
  CCO_P0000014 **or** (
    part_of **some** CCO_P0000014)))

http://rpc133.cs.man.ac.uk:8080/owldoc-server/classes/-828407464/

ntents | Manage | Restart | DL Query find frames | renderer: frag/label | more options

Ontologies | All Classes | All Object Properties | All Data Properties | All Individuals

ass: CCO_B0002486

:p://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0002486

serted Class Hierarchy

CCO_U0000001
+] CCO_U0000006
[+] CCO_U0000007
- **CCO_B0002486**

CCO_U0000003
+] CCO_U0000005
[+] CCO_U0000007
- **CCO_B0002486**

notations

hasDbXref: genid2698
hasDbXref: genid2699
hasDefinition: genid2700
label: "Q5W7F2_ARATH" (en)

perclasses

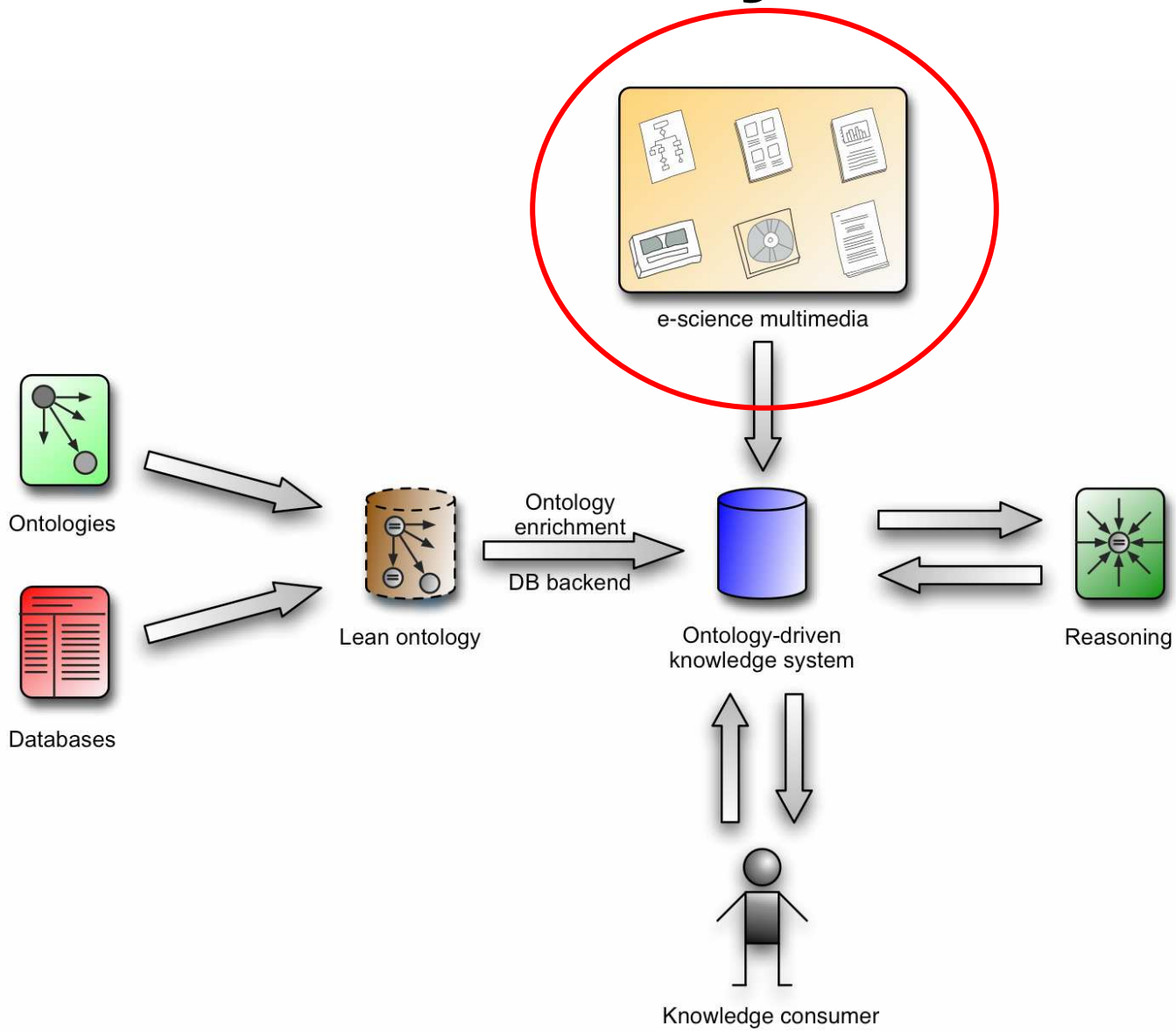
CCO_U0000007
belongs_to **some** CCO_T0000034
encoded_by **some** CCO_G0006236
participates_in **some** CCO_I0000041
participates_in **some** CCO_I0000065
participates_in **some** CCO_I0000066
participates_in **some** CCO_I000 http://www.cellcycleontology.org/ontology/owl/CCO#CCO_I0000065

age

CCO_G0006236 -> codes_for **some** **CCO_B0002486**
CCO_I0000041 -> has_participant **some** **CCO_B0002486**
CCO_I0000065 -> has_participant **some** **CCO_B0002486**
CCO_I0000066 -> has_participant **some** **CCO_B0002486**
CCO_I0000067 -> has_participant **some** **CCO_B0002486**

Permalink | created with OWLDoc

The whole system



Current issues

- Temporal & spatial representation
  - OBOF not enough…
- Performance (reasoners)
  - Huge ontologies
- Weighted knowledge (*often, sometimes*)

Conclusions / Results

- Data integration pipeline: life cycle of the KB
- Existing integration obstacles due to:
  - diversity of data formats
  - lack of formalization approaches
- Reasoning services: inconsistency checks, classification => hypothesis
- Trade-offs: complex queries, representational issues

Future perspectives

- Extend CCO to entire GO tree
- Virtuoso covering the entire domain of biology (“RDF-ing”):
  - Entire OBO foundry
  - UniProt
  - MeSH (articles)
  - …

Acknowledgements

- Martin Kuiper (U Ghent/VIB)
- Vladimir Mironov (U Ghent/VIB)
- Mikel Egaña (U Manchester)
- Robert Stevens (U Manchester)
- Ward Blonde (U Ghent)
- Bernard De Baets (U Ghent)
- CCO Users

Extra slides

Sample entry in OWL

```
<owl:Class rdf:about="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0002060">
    <rdfs:label xml:lang="en">NEB2_HUMAN</rdfs:label>
    <oboInOwl:hasDefinition>
        <oboInOwl:Definition>
            <rdfs:label xml:lang="en">Neurabin-2</rdfs:label>
            <oboInOwl:hasDbXref>
                <oboInOwl:DbXref>
                    <rdfs:label>UniProt:Q96SB3</rdfs:label>
                    <oboInOwl:hasURI rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
                        http://www.cellcycleontology.org/ontology/owl/UniProt#UniProt_Q96SB3
                    </oboInOwl:hasURI>
                </oboInOwl:DbXref>
            </oboInOwl:hasDbXref>
        </oboInOwl:Definition>
    </oboInOwl:hasDefinition>
    <oboInOwl:hasDbXref>
        <oboInOwl:DbXref>
            <rdfs:label>UniProt:Q8TCR9</rdfs:label>
            <oboInOwl:hasURI rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
                http://www.cellcycleontology.org/ontology/owl/UniProt#UniProt_Q8TCR9
            </oboInOwl:hasURI>
        </oboInOwl:DbXref>
    </oboInOwl:hasDbXref>
    <rdfs:subClassOf
rdf:resource="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0000000"/>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty>
                <owl:ObjectProperty rdf:about=
                "http://www.cellcycleontology.org/ontology/owl/CCO#belongs_to"/>
            </owl:onProperty>
```

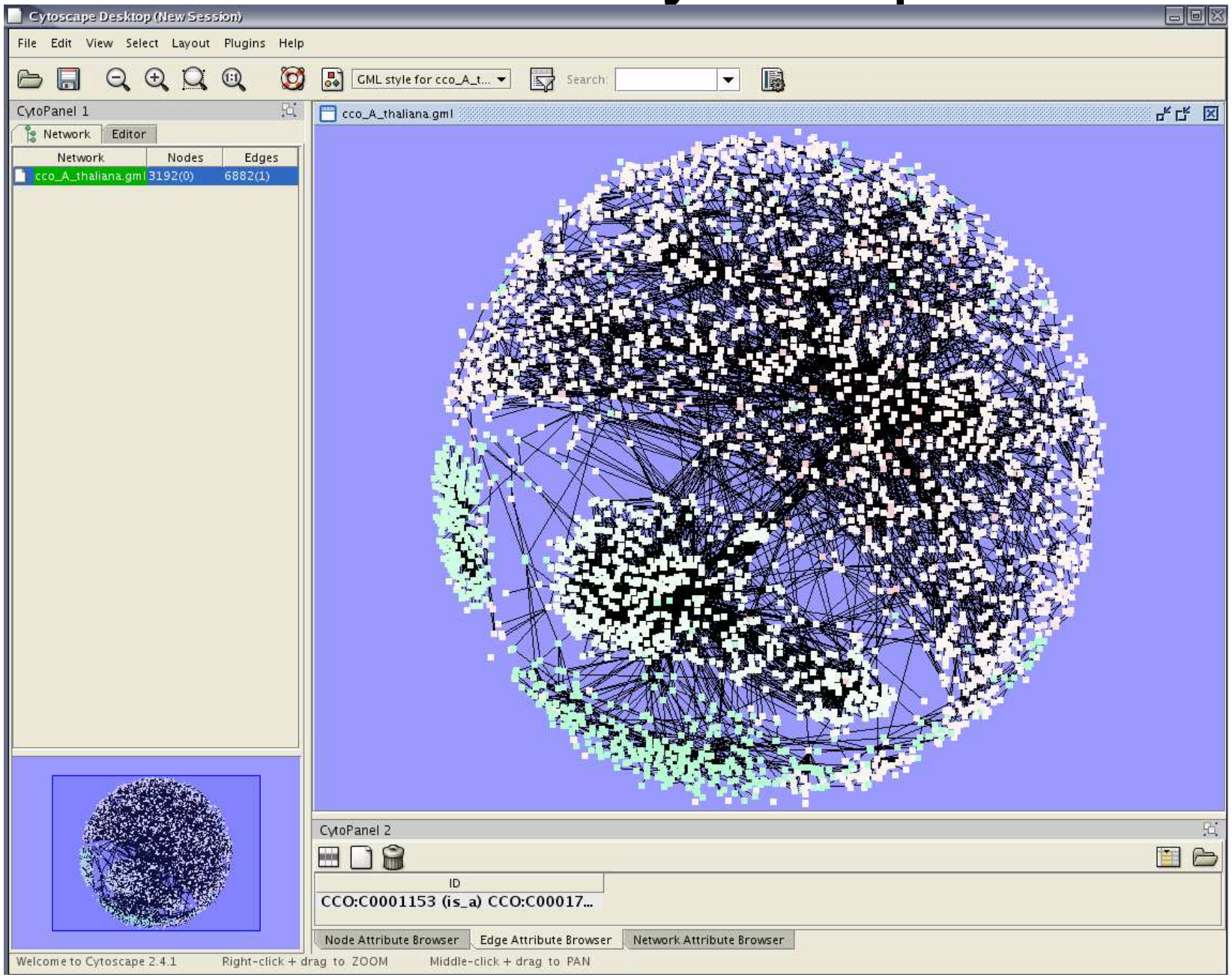
Users

- Molecular biologist: interacting components, events, roles that each component play. Hypothesis evaluation.

- Bioinformatician/Computational Systems Biologist: data integration, annotation, modeling and simulation.

- General audience: educational purposes.

CCO in: Cytoscape



OPPL in CCO

```
# Add a class called "interaction".
# Add the following neccesary condition to the newly added "interaction" class:
# the participants are only the union of protein_1 and protein_2.
# Add the rdfs:label "interaction" to the newly added  "interaction" class.

ADD Class interaction;
ADD subClassOf has_participant only (protein_1 or protein_2);
ADD label "interaction";

# Select any class that has the following condition as a superclass:
# the participants are only the union of protein_1 and protein_2.
# Remove the rdfs:label "interaction" from any selected class.
# Add the rdfs:label "interaction of protein_1 and protein_2" to any selected class.

SELECT subClassOf has_participant only (protein_1 or protein_2);
REMOVE label "interaction";
ADD label "interaction of protein_1 and protein_2";
```