The Cell Cycle Ontology:
an overture to
Semantic Systems Biology

Martin Kuiper
Dept. of Plant Systems Biology, VIB/UGent, Gent, Belgium
Dept. of Biology, NTNU, Trondheim, Norway
martin.kuiper@bio.ntnu.no
www.cellcycleontology.org
www.semantic-systems-biology.org

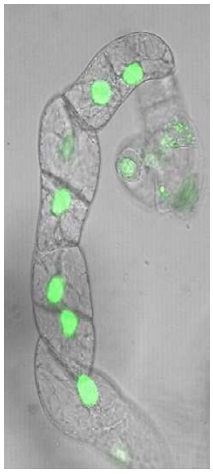


DIAMONDS

Dedicated Integration And Modelling Of Novel Data and prior knowledge to enable Systems biology

Contents

Mitotic Cell Cycle and Endocycle



Arabidopsis thaliana
Human
Baker's yeast
Fission yeast

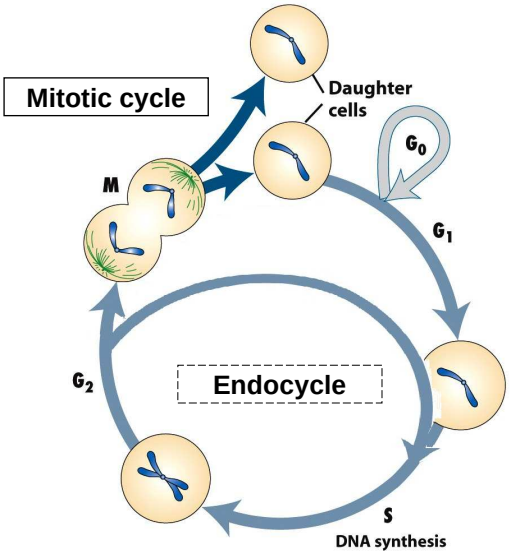


Figure 20-1
Molecular Cell Biology, Sixth Edition

The Cell Cycle Ontology
Some motivating questions

- I'm working with **AT5g35520**, where is this involved in the cell cycle, and in which interactions does it's protein(s) participate?
- My microarray experiment has given me gene **X**, is it known to be involved in cell cycle, in any organism ?
- Verify my network models of genetic or physical interactions
- …

Data management in the life sciences

- Information explosion
- High degree of fragmentation
- Large number of heterogeneous data formats and schemas
- Identity crisis
- Lack of shared semantics
- Complexity of biological data

Goble and Stevens, J. Biomed. Inform 2008, 41: 687-693

Semantic Web

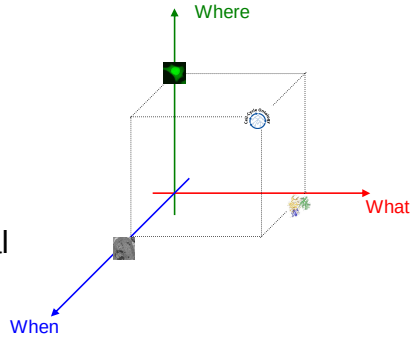
- An extension of the current Web
- Enables navigation and meaningful use of digital resources by automatic processes
- Based on common formats that support aggregation and integration of data from diverse sources
- Tools, programming environments, specialized databases available

http://esw.w3.org/topic/SemanticWebTools

Ruttenberg at al., BMC Bioinformatics 2007, 8 (Suppl. 3): 52-68

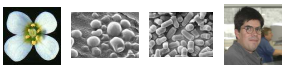
The Cell Cycle Ontology
a knowledge management system

- Capture the knowledge of the CC process
- dynamic aspects of terms and their interrelations
- Use and improve formats that enable a better querying and computational analysis



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

ORGANISMS:



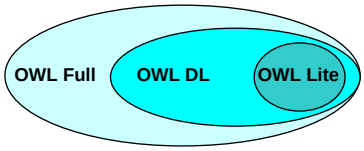
http://www.CellCycleOntology.org

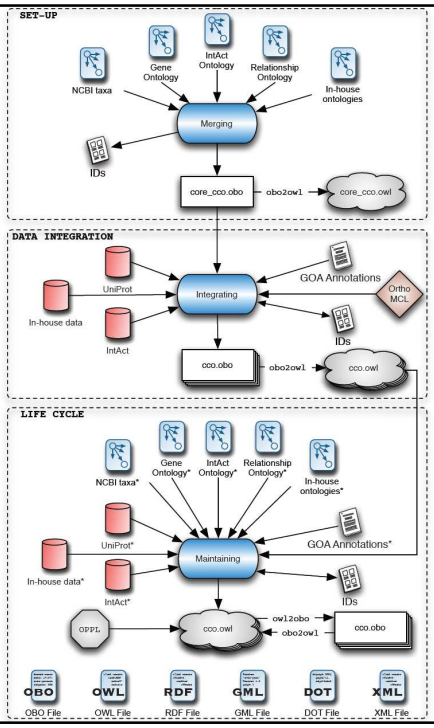
Antezana et al. LNBI, 2006

Key knowledge representation formats for CCO: OBO and OWL

- Why OBO?
  - "Human readable"
  - Standard
  - Tools (e.g. OBOEdit)
  - http://obo.sourceforge.net
- Why OWL?
  - Web Ontology Language
  - "Computer readable"
  - OWL-DL: Reasoning capabilities vs. computational cost
  - Formal foundation (Description Logics: http://dl.kr.org/)
  - **Reasoning**: RACER, Pellet, FaCT++

OBO
open biomedical ontologies





CCO Pipeline

- ontology integration
- format mapping

- data integration
- data annotation
- consistency checking

- maintenance
- data annotation
- semantic improvement

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

Today: over 50,000 terms in CCO

CCO exploration

- Looking up:
  - Terms
  - Relations
  - Synonyms
  - …
- Visual browsing
  - “local neighborhood”
  - Protein-protein interactions
  - …
- Advanced Querying (e.g. SPARQL)
  (to be explained later)

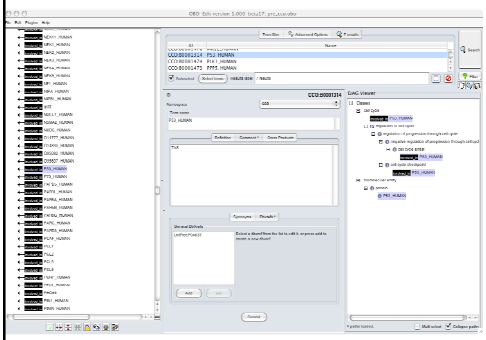
CCO main features

- Ontology driven
- Protein centric
- Semantic web compatible
- Range of data formats

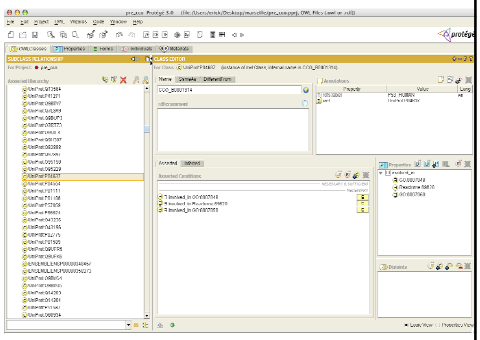
Data Formats and Tools

- OBO-Edit (OBOF)
- Protege (OWL, RDF)
- Cytoscape (GML)
- Graphviz (DOT)
- VisANT (XML)

Search, brows, edit

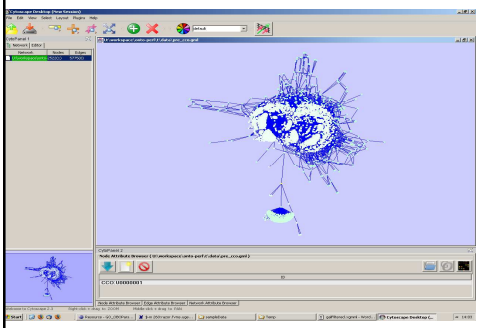


Obo-Edit

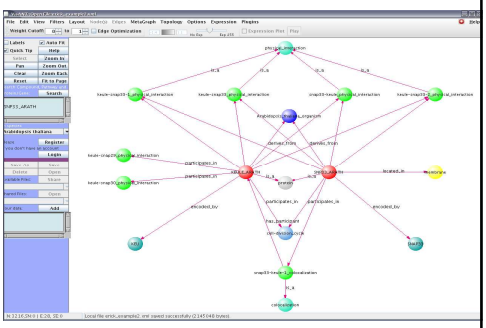


Protégé

Analyse graphs

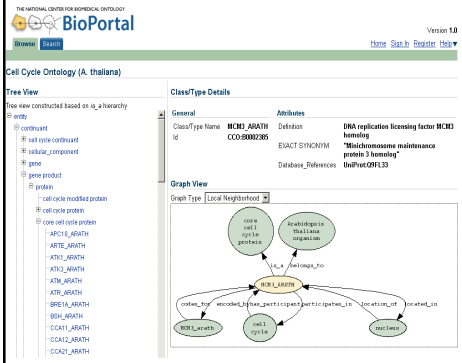


Cytoscape

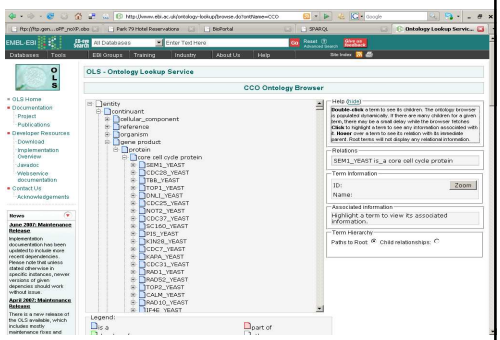


VisANT

Explore on the Web

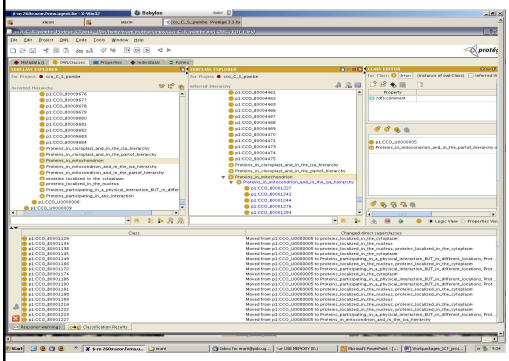


BioPortal

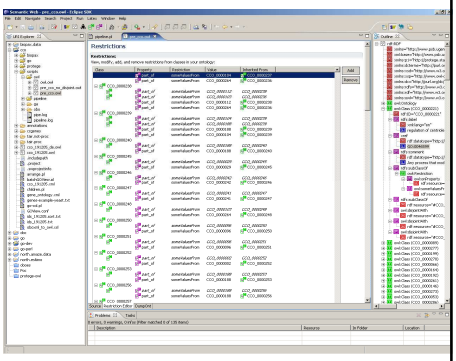


Ontology Lookup Service

Check consistency

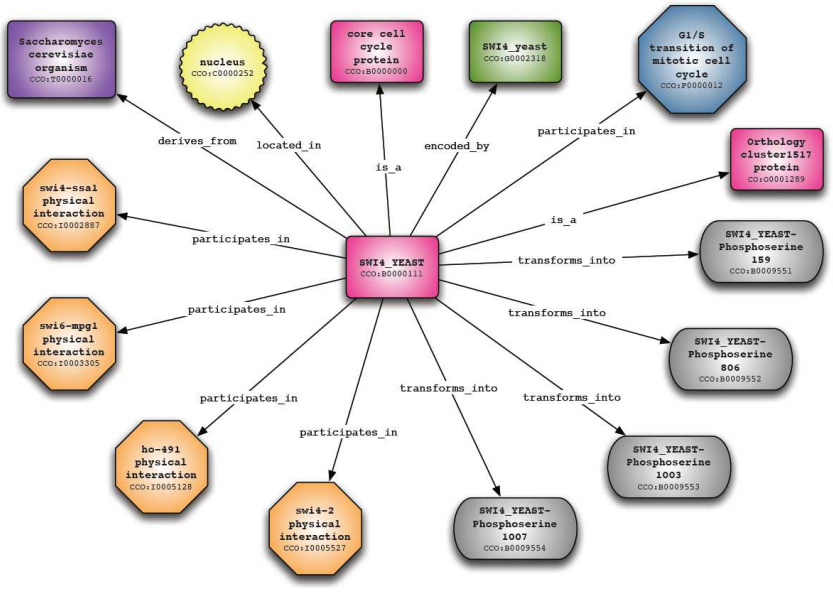


RacerPro plug-in for Protege

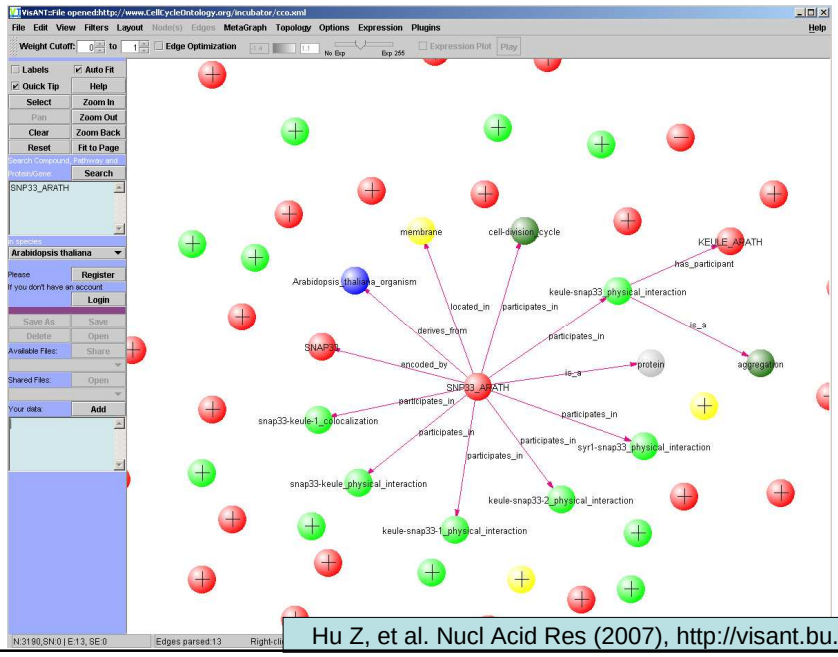


SWeDE plug-in for Eclipse

Current knowledge

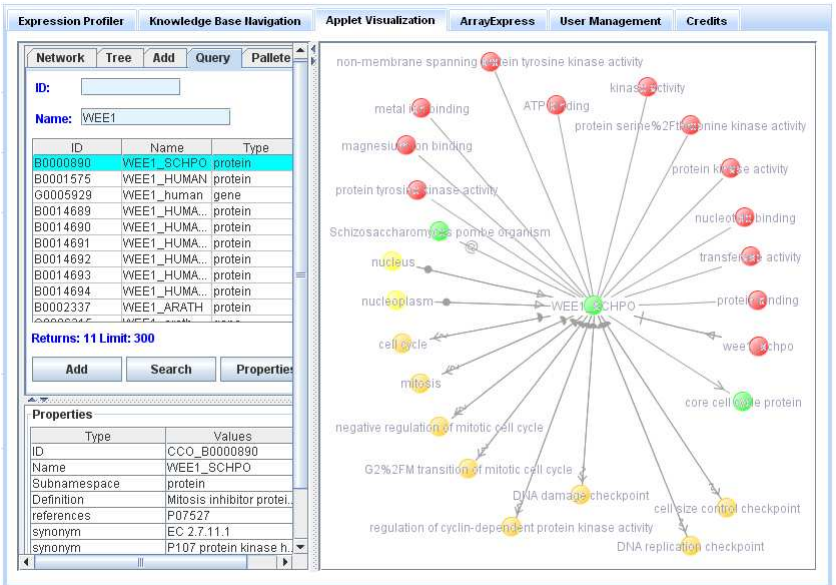


CCO in: visANT



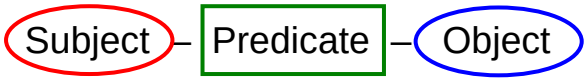
Hu Z, et al. Nucl Acid Res (2007), http://visant.bu.edu/

Local Neighborhood in platform



Advanced SPARQL Querying
CCO + BioGateway

- RDF = Resource Description Framework
  - Metadata model: elements = resources
- Designed to describe Web resources
- Uses Unique Resource Identifiers (URI)
- It allows expressing knowledge about web resources in statements made of triples (basic information unit) :

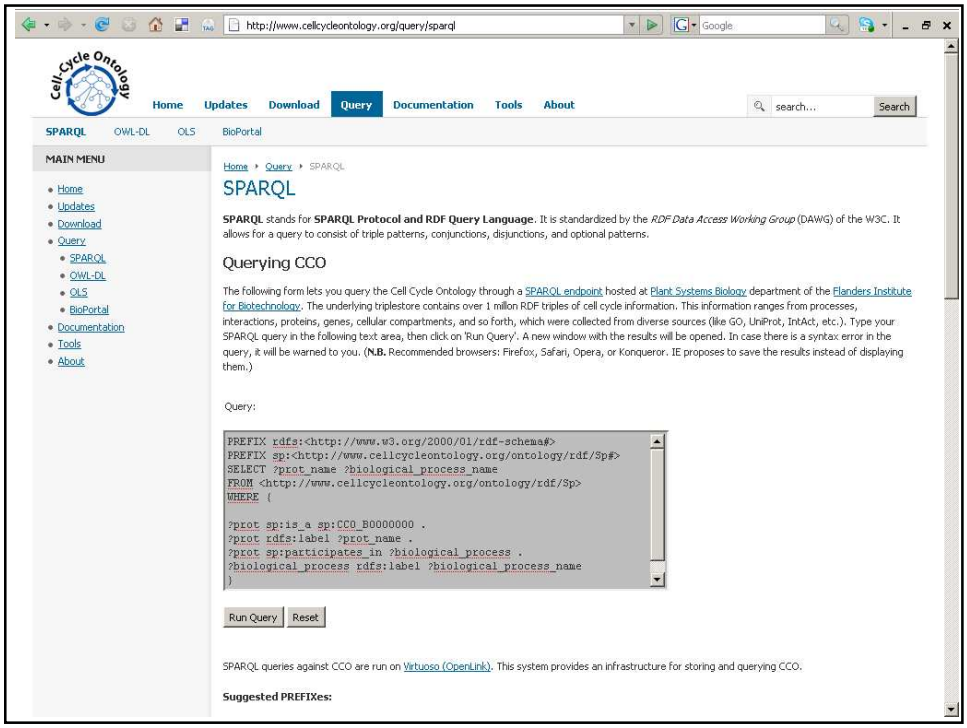


- Triples form a graph
- Graphs could be queried with SPARQL

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*"

Many biological information sources are available in RDF, or can be converted to RDF



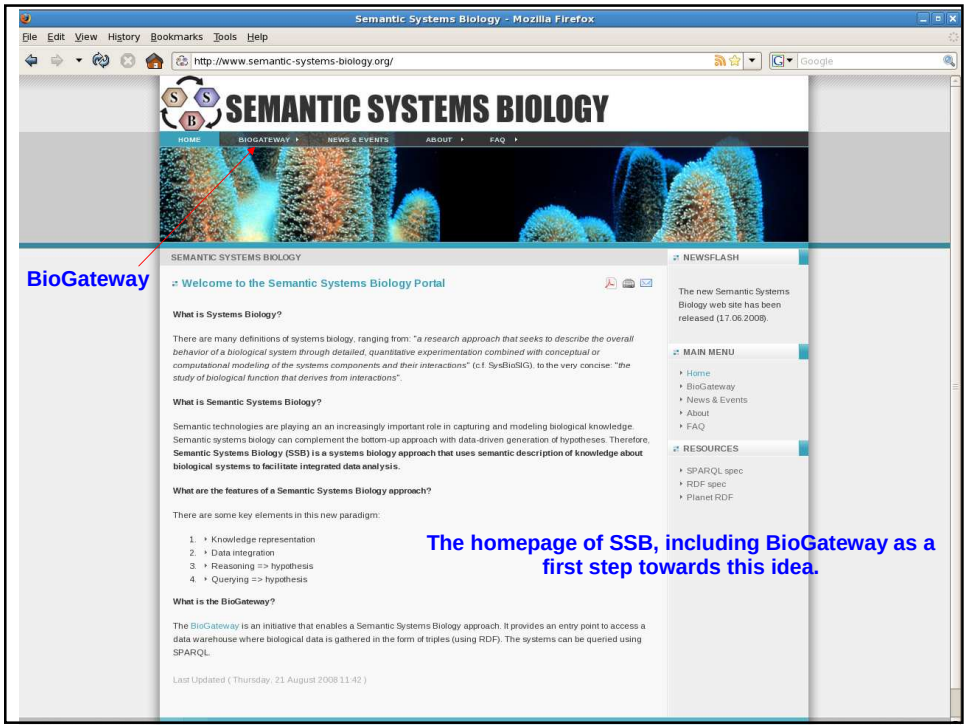
BioGateway

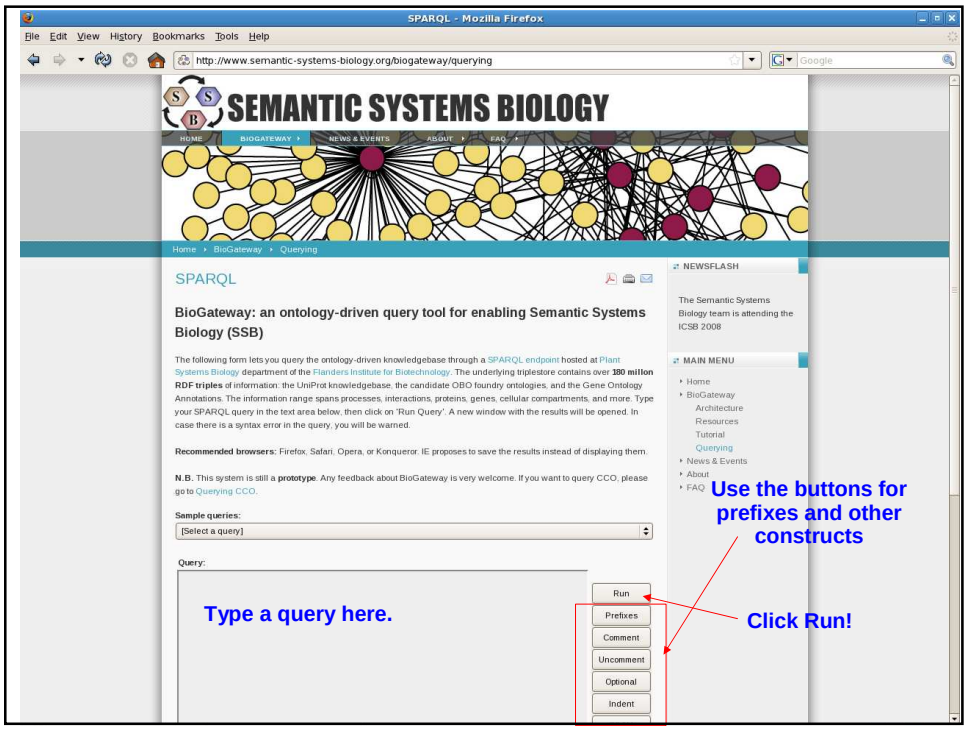
- Uses Virtuoso Open Server
  - Open Source software that can host a triple store
  - Can build this from RDF files
  - Has a DB backend
- Supports SPARQL* language which allows querying RDF data (graphs)
- Its syntax is similar to that of SQL.

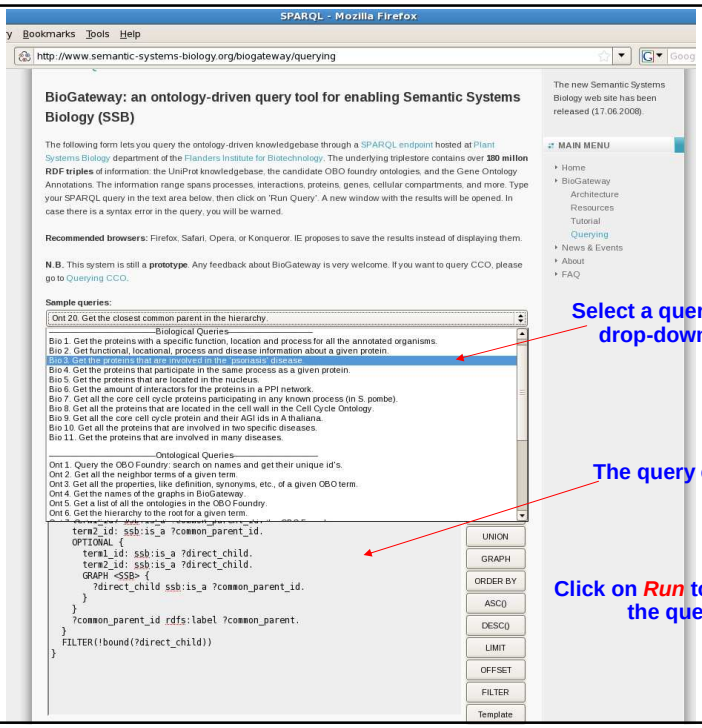
OPENLINK SOFTWARE

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

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


Semantic Systems Biology - Mozilla Firefox
http://www.semantic-systems-biology.org/
SEMANTIC SYSTEMS BIOLOGY
BioGateway
The homepage of SSB, including BioGateway as a first step towards this idea.

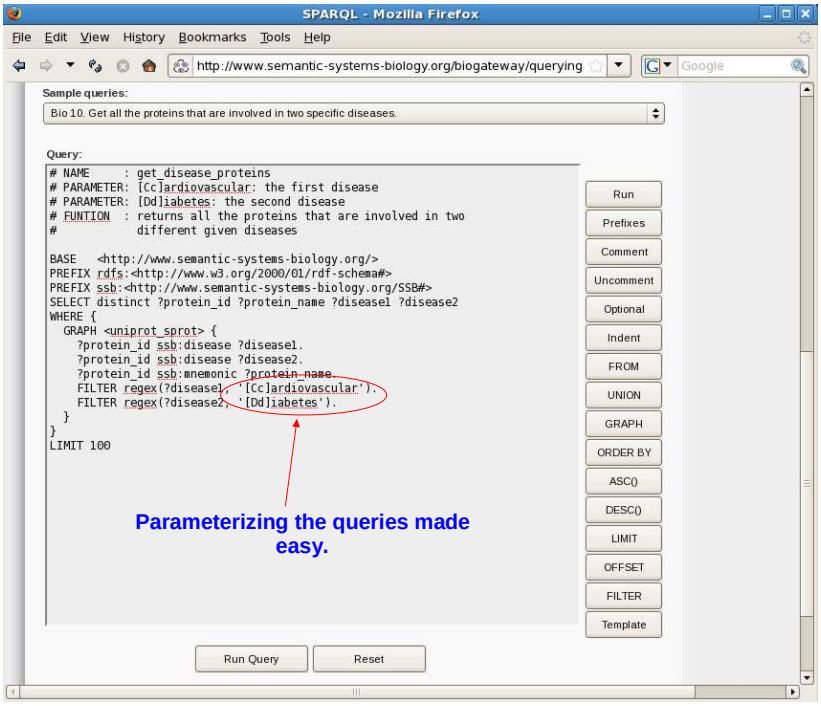

SPARQL - Mozilla Firefox
http://www.semantic-systems-biology.org/biogateway/querying
SEMANTIC SYSTEMS BIOLOGY
SPARQL
Use the buttons for prefixes and other constructs
Type a query here.
Click Run!
Run
Prefixes
Comment
Uncomment
Optional
Indent



A library of queries

The drop-down box contains (so far) 35 queries:

- 15 protein-centric biological queries:
  - The role of proteins in diseases
  - Their interactions
  - Their functions
  - Their locations
- 20 ontological queries:
  - Browsing abilities in RDF like getting the neighborhood, the path to the root, the children,...
  - Meta-information about the ontologies, graphs, relations
  - Queries to show the possibilities of SPARQL on BioGateway, like counting, filtering, combining graphs,...



All the queries are explained in a tutorial

1. Get the proteins with a specific function, location and process for all the annotated organisms.

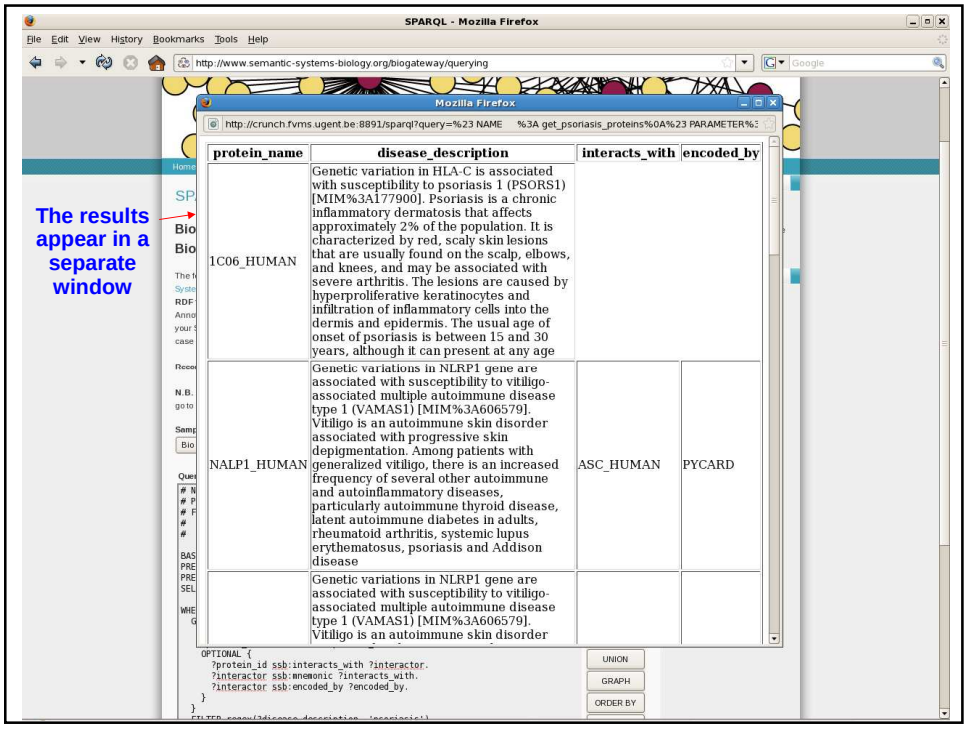
```
# NAME: get_specific_proteins
# PARAMETER: GO_0005216: ion channel activity
# PARAMETER: GO_0005764: lysosome
# PARAMETER: GO_0006811: ion transport
# FUNCTION: returns all the proteins with the same function,
# process and location and the organism in which
# they can be found

BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
SELECT ?organism ?protein ?protein_id
WHERE {
 GRAPH ?organism {
  ?protein_id ssb:has_function ssb:GO_0005216.
  ?protein_id ssb:located_in ssb:GO_0005764.
  ?protein_id ssb:participates_in ssb:GO_0006811.
  ?protein_id rdfs:label ?protein.
 }
 FILTER(?organism != <SSB> && ?organism != <GOA>).
}
```

Click here to select this query in the drop-down box on the query-page and edit it
Click here to see the results

For every query the name, the parameters and the function are indicated at the top.

The parameters are indicated in red.



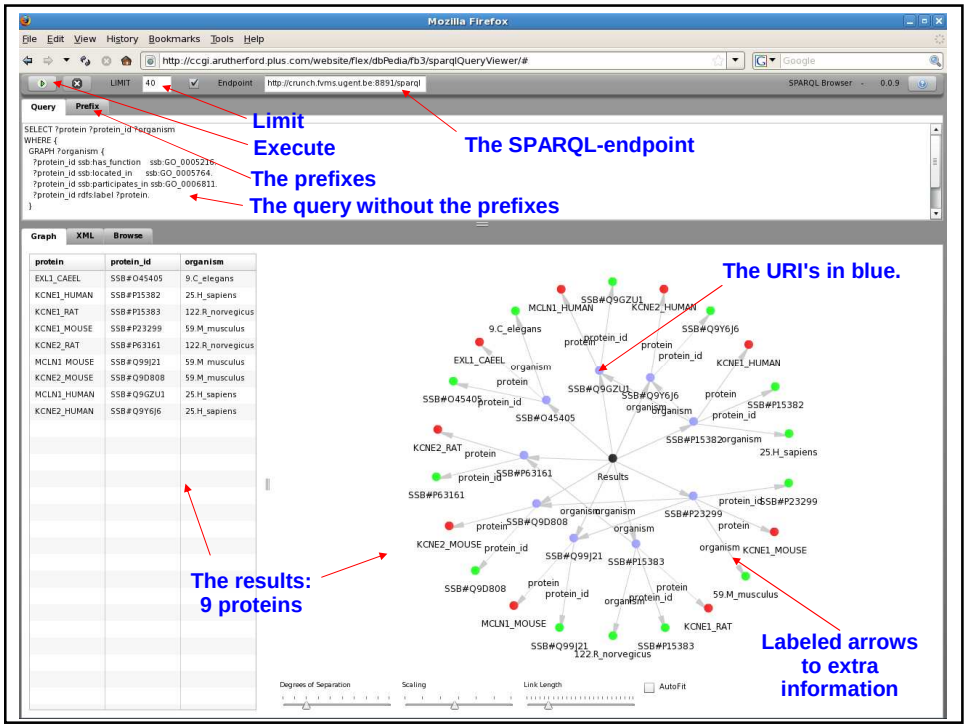
The neighborhood of the human protein 1443F in the RDF-graph

term_as_child	outward_arrow	head_name	tail_name	inward_arrow	term_as_parent
1433F_HUMAN	participates in	intracellular protein transport			
1433F_HUMAN	participates in	glucocorticoid catabolic process			
1433F_HUMAN	participates in	positive regulation of transcription			
1433F_HUMAN	participates in	regulation of synaptic plasticity			
1433F_HUMAN	participates in	glucocorticoid receptor signaling pathway			
1433F_HUMAN	participates in	regulation of neuron differentiation			
1433F_HUMAN	participates in	negative regulation of dendrite morphogenesis			
1433F_HUMAN	is located in	cytoplasm			
1433F_HUMAN	has function	protein binding			
1433F_HUMAN	has function	transcription activator activity			
1433F_HUMAN	has function	actin binding			
1433F_HUMAN	has function	insulin-like growth factor receptor binding			
1433F_HUMAN	has function	protein domain specific binding			
1433F_HUMAN	has function	glucocorticoid receptor binding			
1433F_HUMAN	has source	Homo sapiens			
1433F_HUMAN	interacts with	PARD3_HUMAN			
1433F_HUMAN	interacts with	PFTK1_HUMAN			
1433F_HUMAN	interacts with	RAF1_HUMAN			
1433F_HUMAN	interacts with	GREM1_HUMAN			
1433F_HUMAN	interacts with	MARK4_HUMAN			
1433F_HUMAN	interacts with	PAR6A_HUMAN			
1433F_HUMAN	interacts with	PAR6B_HUMAN			
1433F_HUMAN	interacts with	KPCI_HUMAN			
			PARD3_HUMAN	interacts with	1433F_HUMAN
			PFTK1_HUMAN	interacts with	1433F_HUMAN
			RAF1_HUMAN	interacts with	1433F_HUMAN
			GREM1_HUMAN	interacts with	1433F_HUMAN
			PAR6A_HUMAN	interacts with	1433F_HUMAN
			PAR6B_HUMAN	interacts with	1433F_HUMAN
			KPCI_HUMAN	interacts with	1433F_HUMAN
			ADA22_HUMAN	interacts with	1433F_HUMAN
			HNRPD_HUMAN	interacts with	1433F_HUMAN

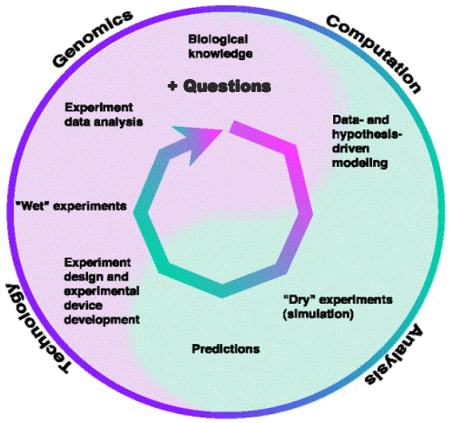
The resulting triples (arrows) are represented as a small grammatical sentence: subject, predicate, object.

Outgoing arrows

Incoming arrows



Systems Biology



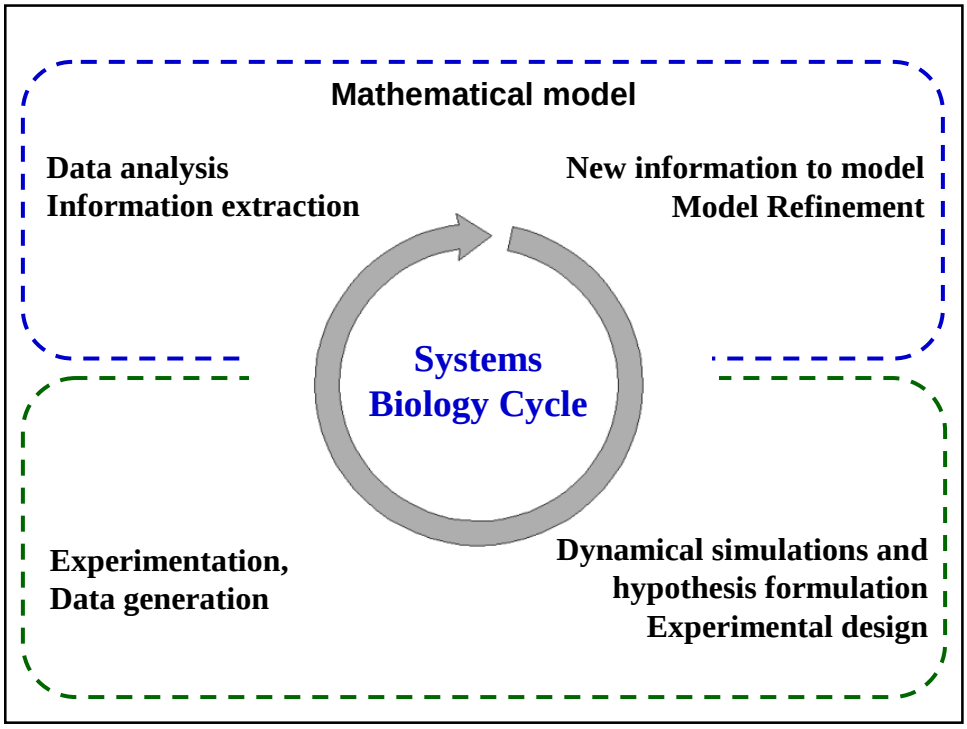
SCIENCE VOL 295 1 MARCH 2002

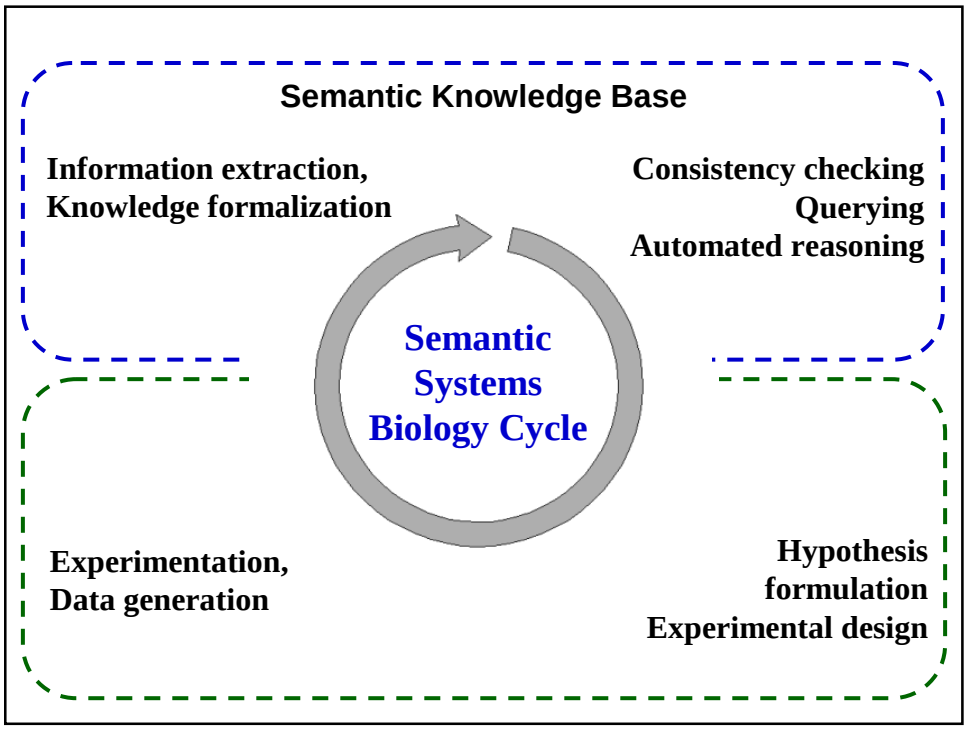
SYSTEMS BIOLOGY: THE GENOME, LEGOME, AND BEYOND

REVIEW

Systems Biology: A Brief Overview

Hiroaki Kitano


Mathematical model
Data analysis
Information extraction
New information to model
Model Refinement
Systems
Biology Cycle
Experimentation,
Data generation
Dynamical simulations and
hypothesis formulation
Experimental design


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

Semantic Systems Biology

- Semantic:
  - New emerging semantic web technologies for analyzing information
  - ontologies are backbone for uniform and unambiguous knowledge description
- Needs:
  - Solving schema mismatch problem
  - Uniform understanding of knowledge
  - Seamless data integration
  - Smooth data sharing
- Allows:
  - Knowledge representation
  - Knowledge integration
  - Querying
  - Automated reasoning ==> new hypotheses

Conclusions / Results

- CCO is evolving to a one-stop shop for cell cycle researchers
- It allows exploratory analysis: browse, visualise and search
- Several querying facilities: advanced ways for retrieving data
- Automated reasoning exploitation: classification, consistency checking, and inference
- BioGateway: evolves into RDF store for BioSciences
- Queries, knowledge sources and system design now go hand-in-hand (user interaction)
- Existing integration obstacles are due to:
  - diversity of data formats
  - lack of formalization approaches
- This calls for a '**foundry**' type initiative for bioscience RDF
  - http://www.ntnu.no/systemsbiology/ssbwiki

Like HCLSIG: fostering the implementation of Semantic Web technologies in the information ecosystem of biomedicine
http://www.w3.org/2001/sw/hcls/

Acknowledgements

- Erick Antezana (U Gent, BE)
- Vladimir Mironov (NTNU, NO)
- Mikel Egaña (U Manchester, UK)
- Robert Stevens (U Manchester, UK)
- Ward Blonde (U Ghent, BE)
- Bernard De Baets (U Ghent, BE)
- Alan Ruttenberg (Science Commons, US)
- Alistair Rutherford (www.netthreads.co.uk)
- Users

http://www.cellcycleontology.org
http://ww.semantic-systems-biology.org

SEMANTIC SYSTEMS BIOLOGY

VIB
UNIVERSITEIT GENT
MANCHESTER 1824
NTNU
Sixth Framework Programme