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

1. Systems Biology
2. Data/hypothesis-driven approaches
3. Data integration and exploitation
4. The Cell-Cycle Ontology
5. BioGateway
6. Concluding remarks
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:
  - 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

### Systems biology paradigm

- **Gene network components**
- **Gene network components**
- **Knowledge Management**
- **Mathematics**
- **Predictive mathematical model**
- **Mathematics**
- **Biological Process**
- **Biological Process**
- **Genome-scale functional genomics data**
- **Genome-scale functional genomics data**
- **Knowledge Management**
- **Knowledge Management**
- **Statistics**
- **Statistics**
- **Mining**
- **Mining**
- **Knowledge**
- **Knowledge**
- **Mathematics**
- **Mathematics**
Biological knowledge

Information extraction
Knowledge formalization

Semantic Systems Biology Cycle

Experimentation
Data generation

Consistency checking
Querying
Automated reasoning

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
(Some) Motivating questions

• I’m working with \textbf{AT5g35520}, in which interactions does this gene play a role?
• From my microarray experiment I’ve got this gene \textbf{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

http://www.CellCycleOntology.org

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

Antezana et al. LNBI, 2006
Knowledge representation

• Why OBO?
  – “Human readable”
  – Standard
  – Tools (e.g. OBOEdit)

• 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)
CCO accession number

CCO: [CPFRTIBGOU]nnnnnnnn

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
```


Some figures

<table>
<thead>
<tr>
<th>Entity</th>
<th>At</th>
<th>Hs</th>
<th>Sc</th>
<th>Sp</th>
<th>CCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins</td>
<td>252</td>
<td>5829</td>
<td>7069</td>
<td>930</td>
<td>24541</td>
</tr>
<tr>
<td>Genes</td>
<td>222</td>
<td>1806</td>
<td>3148</td>
<td>852</td>
<td>6028</td>
</tr>
<tr>
<td>Interactions</td>
<td>76</td>
<td>2394</td>
<td>5162</td>
<td>399</td>
<td>8031</td>
</tr>
<tr>
<td>Orthology groups</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1649</td>
</tr>
</tbody>
</table>

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

2008-03-07: 49226 terms in CCO
Knowledge exploration

• Looking up:
  – Terms,
  – Synonyms,
  – …

• Visual browsing
  – “local neighborhood”
  – Path to the root

• Advanced Querying (e.g. SPARQL)
CCO in: visANT

Local neighborhood

Two interacting proteins
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
    - BRE1_ARATH
    - BSH_ARATH
    - CCA1_ARATH
    - CCA2_ARATH
    - CCA1_ARATH

Class/Type Details

<table>
<thead>
<tr>
<th>General</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class/Type Name</td>
<td>DNA replication licensing factor MCM3 homolog</td>
</tr>
<tr>
<td>MCM3_ARATH</td>
<td>“Minichromosome maintenance protein 3 homolog”</td>
</tr>
<tr>
<td>Id</td>
<td>Definition</td>
</tr>
<tr>
<td>CCO:B0002385</td>
<td></td>
</tr>
<tr>
<td>Exact Synonym</td>
<td>Database References</td>
</tr>
<tr>
<td></td>
<td>UniProt:Q9FL33</td>
</tr>
</tbody>
</table>

Graph View

Graph Type: Local Neighborhood

[Diagram showing relationships between MCM3_ARATH and other terms]

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

  Subject – Predicate – Object
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.w3.org/TR/rdf-sparql-query/

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

?protein sp:is_a sp:CCO_B0000000.
?protein rdfs:label ?protein_label
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 million 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.
    ?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:
“all the core cell cycle proteins (S. pombe) participating in a known process”
<table>
<thead>
<tr>
<th>proteome</th>
<th>biological_process</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBC11_SCHPO</td>
<td>G2%2FM transition of mitotic cell cycle</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>cell cycle</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>mitosis</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>mitotic metaphase%2Fanaphase transition</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>regulation of mitotic cell cycle</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>cyclin catabolic process</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>cell cycle</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>cyclin catabolic process</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>activation of anaphase-promoting complex during mitotic cell cycle</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>cell cycle arrest in response to nitrogen starvation</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>negative regulation of cyclin-dependent protein kinase activity</td>
</tr>
<tr>
<td>DYHC_SCHPO</td>
<td>dhc1-peg1-1 physical interaction</td>
</tr>
<tr>
<td>DYHC_SCHPO</td>
<td>synapsis</td>
</tr>
<tr>
<td>DYHC_SCHPO</td>
<td>meiotic recombination</td>
</tr>
<tr>
<td>DYHC_SCHPO</td>
<td>horsetail nuclear movement</td>
</tr>
<tr>
<td>ORB6_SCHPO</td>
<td>cell morphogenesis checkpoint</td>
</tr>
<tr>
<td>ORB6_SCHPO</td>
<td>regulation of cell cycle</td>
</tr>
<tr>
<td>DED1_SCHPO</td>
<td>G2%2FM transition of mitotic cell cycle</td>
</tr>
</tbody>
</table>
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*)
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_U00000005 and participates_in some CCO_Y0000001
• CCO_U00000005 = class of proteins
• CCO_Y00000001 = 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))
  )
  ```
ass: CCO_B0002486

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

sorted 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_IC000041
participates_in some CCO_IC000065
participates_in some CCO_IC000066
participates_in some CCO_IC000 http://www.cellcycleontology.org/ontology/owl/CCO#CCO_IC000065

age

CCO_G0006236 -> codes_for some CCO_B0002486
CCO_T0000041 -> has_participant some CCO_B0002486
CCO_T0000055 -> has_participant some CCO_B0002486
CCO_T0000056 -> has_participant some CCO_B0002486
CCO_T0000057 -> has_participant some CCO_B0002486
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

```xml
<owl:Class rdf:about="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0002060">
  <rdfs:label xml:lang="en">NEB2_HUMAN</rdfs:label>
  <oboInOwl:hasDefinition>
    <rdfs:label xml:lang="en">Neurabin-2</rdfs:label>
    <oboInOwl:hasDbXref>
      <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:hasDbXref>
  </oboInOwl:Definition>
  <oboInOwl:hasDefinition>
    <oboInOwl:hasDbXref>
      <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:hasDbXref>
    <rdfs:subClassOf rdf:resource="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0000000" />
  </oboInOwl:subClassOf>
  <owl:Restriction>
    <owl:onProperty>
      <owl:ObjectProperty rdf:about="http://www.cellcycleontology.org/ontology/owl/CCO#belongs_to"/>
    </owl:onProperty>
  </owl:Restriction>
</owl:Class>
```
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";