A knowledge-based system for plant cell-cycle elucidation

Erick Antezana
<erant@psb.ugent.be>

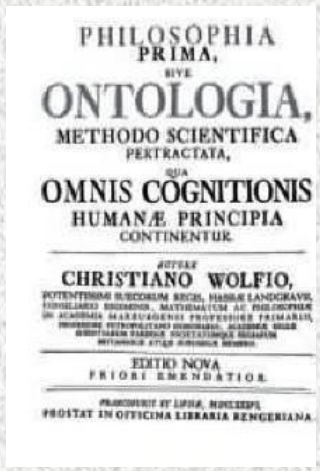
Contents

- What is an ontology?
- What is it for?
- Ontology development
- Bio-ontologies (Cell-cycle ontology)
- Languages
- Description Logics
- System architecture
- Planning (future directions)

What is an Ontology?

- "The science of being *qua* being". (***'qua'*** = 'with regard to the aspect of'. Aristotle ~350BC)
- A specification of a conceptualization (T. Gruber, 1993)
- A formal representation of knowledge domains (Bard and Rhee, 2004)

http://encyclopedia.thefreedictionary.com/Aristotle


PHILOSOPHIA
PRIMA,
SIVE
ONTOLOGIA,
METHODO SCIENTIFICA
PERTRACTATA,
QUA
OMNIS COGNITIONIS
HUMANÆ PRINCIPIA
CONTINENTUR.
AUTORE
CHRISTIANO WOLFIO,
EDITIO NOVA
PRIORI EMENDATIOR.

http://www.formalontology.it/

What is it for?

- Communication between people and organizations
- Sharing and reusing knowledge
- As data repository
- As a query model for information sources
- Validation, annotation, specification, knowledge acquisition.
- Its reuse decreases the semantic heterogeneity of DBs

What is it for? (cont)

Bio-ontologies: Set of controlled vocabularies that are used by databases to describe biological data

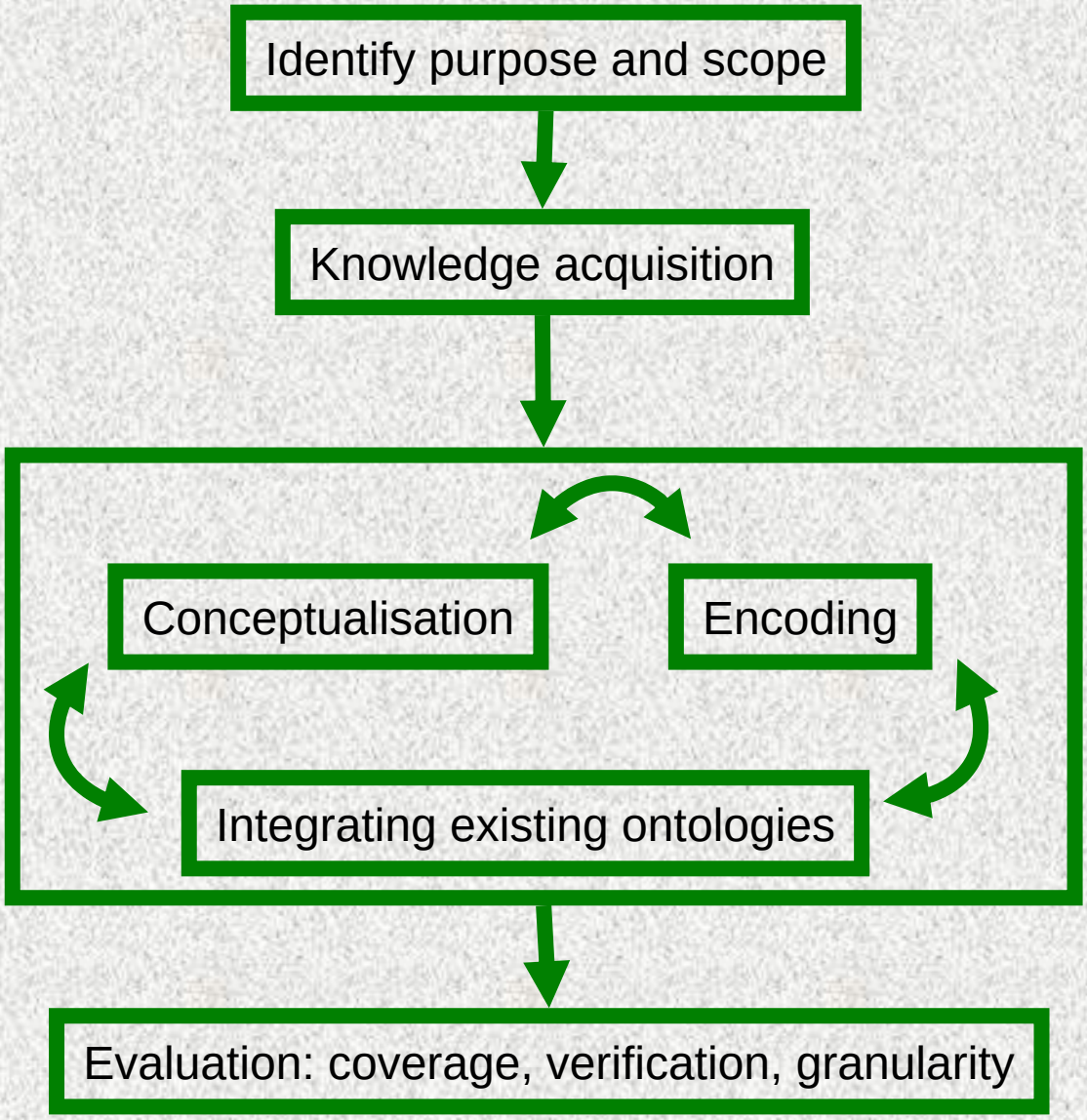
- **Gene Ontology** (http://www.geneontology.org/)
- **Plant Ontology** (http://www.plantontology.org/)
- **Sequence ontology** (http://song.sourceforge.net/)
- … visit **OBO** (http://obo.sourceforge.net/)

Ontology development

- Protégé
- OilEd
- Chimera
- Ontolingua
- DAG-Edit
- …

- No standard methodologies for building ontologies
- Depends on the domain
- Most consider:
  - Informal stage
  - Formal stage

Tentative Methodology (V-model)



Developing an ontology

- Define all the concepts (terms) within a domain of knowledge (e.g. CDK)
  - Primitive: properties are necessary
    E.g. Plant Rb is nuclear, but a nuclear protein need not be a Rb
  - Defined: properties are necessary *and* sufficient
    E.g. Plant cells must have B-type CDKs. Every cell that contains a B-type CDK must be from a Plant.
- Specify the relationships they have to one another
  E.g. bindsTo
- Ontologies: best delivered in some computable representation
- Multiple inheritance admitted (DAG)

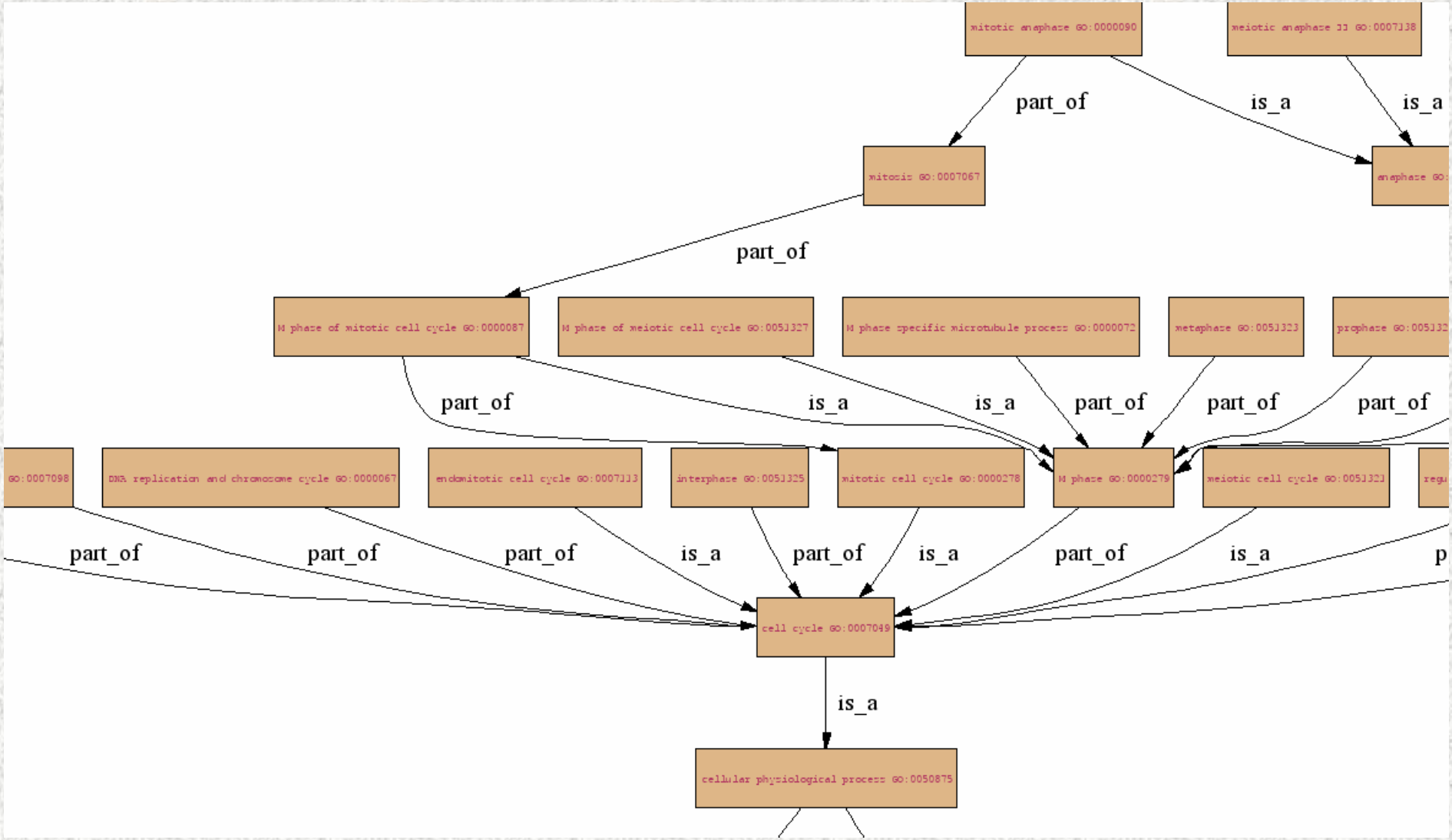
Ontology development

- Try to keep it simple but robust
- Rely on synonyms (no new terms)
- The precise definition of terms is critical to the integrity of the ontology.
- Terms are related to each other as children to parents.
- Each child term can have one or more parents.

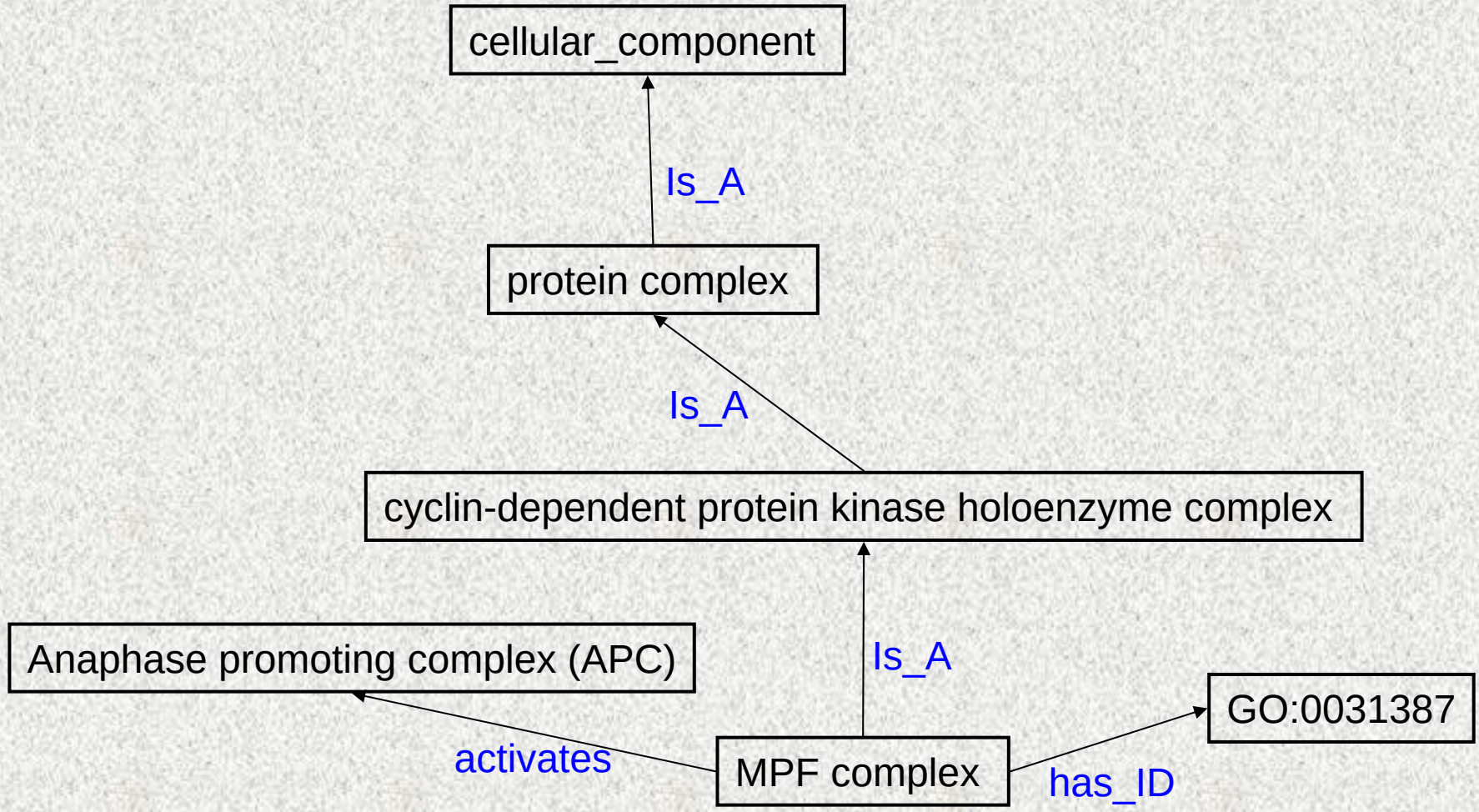
Bio-ontology

- They clarify scientific discussions providing the shared vocabulary for biologist to:
  - communicate their results effectively,
  - explore data and
  - extend scientific investigations
- They, also, extend the power of computational approaches and systems to perform data exploration, inference and mining.

Gene ontology (GO Cell-Cycle)



MPF complex (partial) ontology



Cell-cycle ontology

- GO allows efficient collection of biological data from heterogeneous sources: **not rich enough to describe the cell-cycle process**
- Questions:
  - How should the information be stored?
  - What information should be stored?
  - Technology to be used?
- Hidden knowledge held by the Ontology.

Answers (a priori)

- Focus on: Cell-cycle process
- Components
- Protein features (instances)
- Relationships
- DBMS? UI? Web? Protégé?...
- OWL-DL? Expressiveness
- Cross-references?
- Orthogonal ontology?

Facts

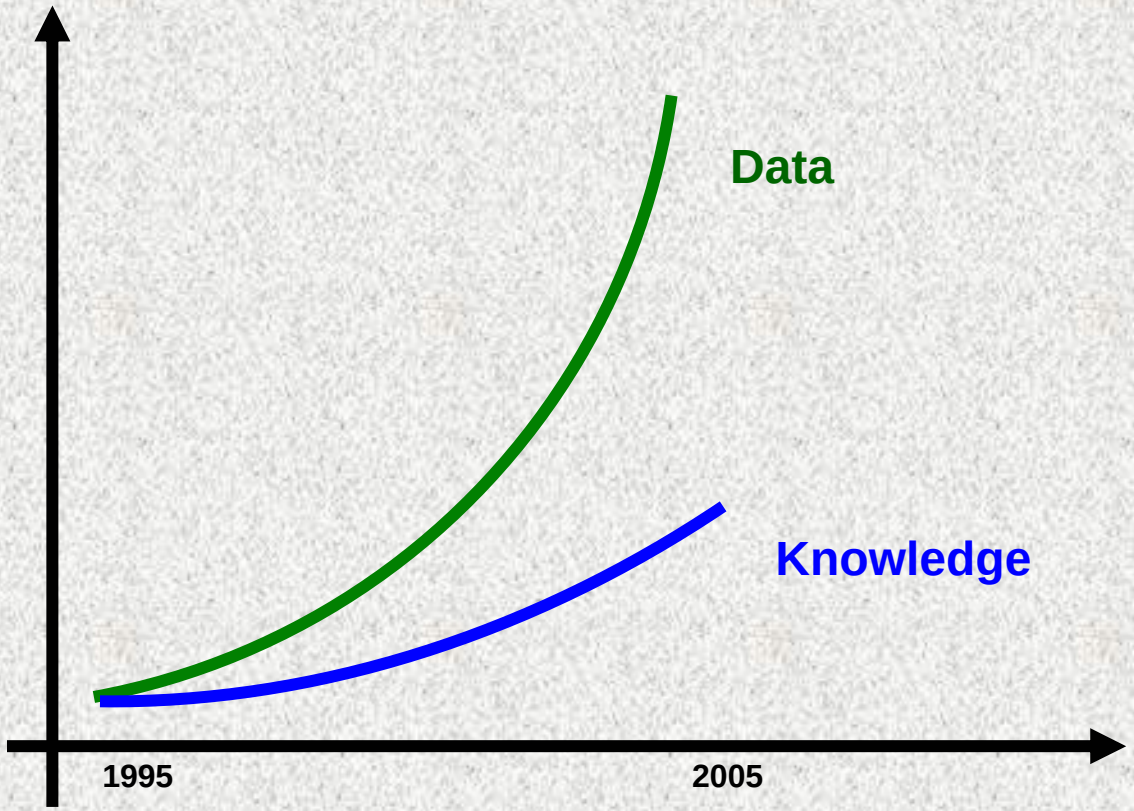
- Biologists are against data sharing, annotation, standardisation, ….

- An ontology is a formal way of representing knowledge.

- Currently ontologies are becoming essential to the analysis and interpretation of genome wide data.

- If we specify the ‘logic’ of combining ‘things’ and ‘relations’ we can
  - write hypotheses about biological processes in a formal manner
  - and evaluate them for consistency with existing information.

Data vs Knowledge



The data/knowledge **deficit** in 2005: Data and knowledge growth during last decade

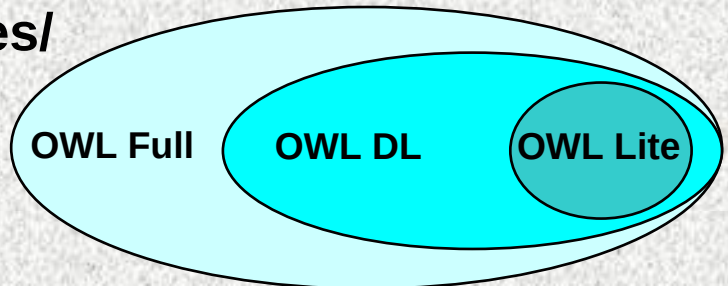
Languages (*)

- Vocabularies that use natural language
  - Hand crafted, flexible but difficult to evolve, maintain and keep consistent, with poor semantics
  - Gene Ontology (several tools developed: e.g. AmiGO)
  - GONG (OWL - DL)
- Object-based KR: FRAMES
  - Extensively used, good structuring, intuitive. Good semantics (based on standards).
  - Only primitive concepts.
  - Ontolingua (RiboWeb), Ocelot (EcoCyc), ...
- Logic-based: DL
  - Very expressive, developed theory, well defined semantics
  - Trade-off: Expressivity vs. Computational complexity
  - Automatic derived classification taxonomies
  - Concepts: primitive and defined
  - Less intuitive, Limited set of language constructs
  - TAMBIS Ontology (uses FaCT)

(*) Adapted from "Building and Using Ontologies", Robert Stevens, U. of Manchester

OWL

- Expressivity (biology => “complex relationships”)
- W3C Standard (use existing standards)
- XML-based (**THE** exchange language)
- Machine Computable
  - Facilitate integration of knowledge, data, tool development
  - Uncover inconsistencies and new knowledge
  - OWL-DL
    - Reasoning capabilities
    - Complete: all conclusions are guaranteed to be computed
    - Decidable: all computations finish in finite time (with OWL Lite, short amount of time)
    - **http://www.w3.org/TR/owl-features/**



Semantic web

- 2001: Concept of Semantic Web (Tim Berners-Lee, Jim Hendler and Ora Lasilla, SciAm)
  - XML more expressive than HTML
  - RDF triples: simple logical statements (subject-verb-object) in XML (computer can understand)
  - Ontologies written in OWL, which is itself expressed as a set of RDF statements
  - An RDF triple might state that a cyclin *isA* protein, informing the computer that an entity **cyclin** is included in the more general category: **protein**

BY MIGUEL SALMERON, http://www.sciam.com/

Description Logics

- A family of KW representation languages tailored for expressing KW about concepts and hierarchies.
- Basic building blocks: concept, role, individual
- Constructs: union, intersection, negation, quantification, …
- Well defined semantics
- Reasoning services
- More info: http://dl.kr.org/

Semantics in DL

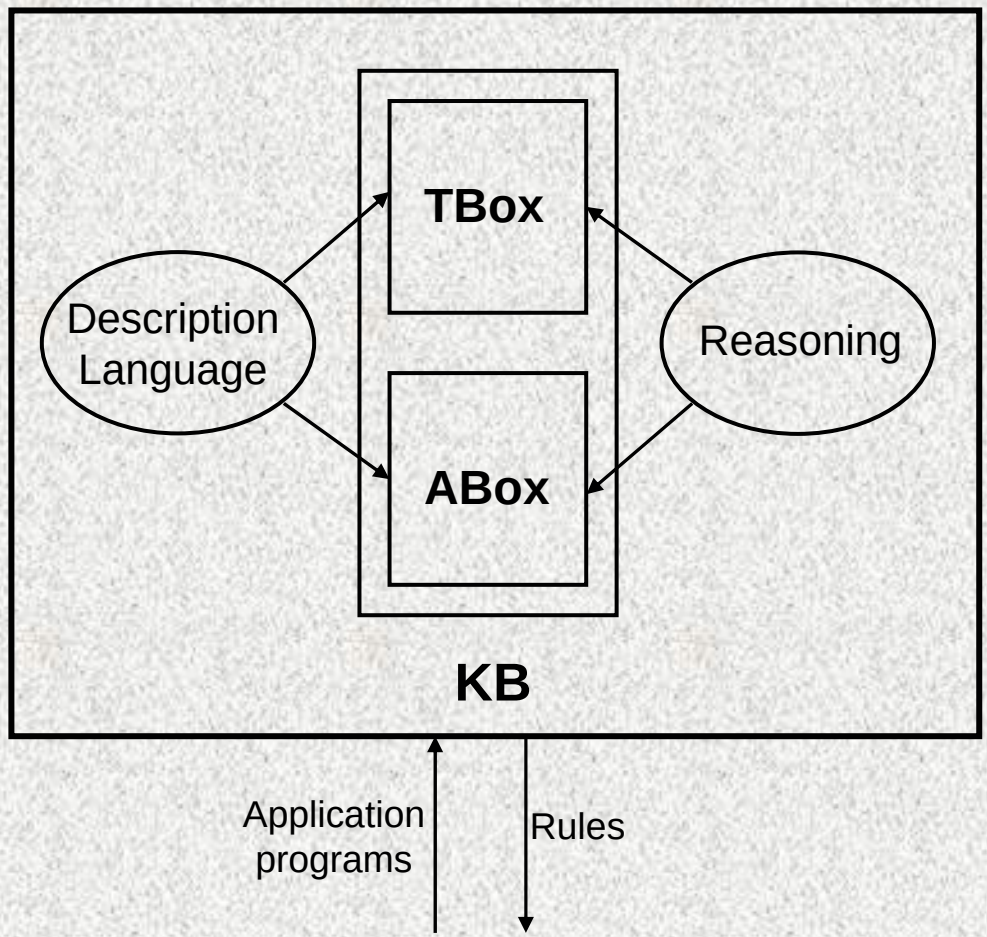
- The basic DL is $\mathcal{ALC}$
- Semantics given by interpretation $\mathcal{I}$
- Trade-off between Expressivity and computational complexity

Table 1. Syntax and semantics of $\mathcal{SHOQ}(\mathbf{D})$-concept expressions

construct name	syntax	semantics
atomic concept $\mathbf{C}$	A	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
abstract role $\mathbf{R}_A$	R	$R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$
concrete role $\mathbf{R}_{\mathbf{D}}$	T	$T^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta_{\mathbf{D}}$
nominals $\mathbf{I}$	$\{o\}$	$\{o\}^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}, \#\{o\}^{\mathcal{I}} = 1$
data types $\mathbf{D}$	d	$d^{\mathbf{D}} \subseteq \Delta_{\mathbf{D}}$
	$\neg d$	$(\neg d)^{\mathbf{D}} = \Delta_{\mathbf{D}} \setminus d^{\mathbf{D}}$
conjunction	$C \sqcap D$	$(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$
disjunction	$C \sqcup D$	$(C \sqcup D)^{\mathcal{I}} = C^{\mathcal{I}} \cup D^{\mathcal{I}}$
negation	$\neg C$	$(\neg C)^{\mathcal{I}} = \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
exists restriction	$\exists R.C$	$(\exists R.C)^{\mathcal{I}} = \{x \mid \exists y : (x,y) \in R^{\mathcal{I}} \text{ and } y \in C^{\mathcal{I}}\}$
value restriction	$\forall R.C$	$(\forall R.C)^{\mathcal{I}} = \{x \mid \forall y : (x,y) \in R^{\mathcal{I}} \Rightarrow y \in C^{\mathcal{I}}\}$
atleast restriction	$\geq nR.C$	$(\geq nR.C)^{\mathcal{I}} = \{x \mid \#\{y \mid (x,y) \in R^{\mathcal{I}} \text{ and } y \in C^{\mathcal{I}}\} \geq n\}$
atmost restriction	$\leq nR.C$	$(\leq nR.C)^{\mathcal{I}} = \{x \mid \#\{y \mid (x,y) \in R^{\mathcal{I}} \text{ and } y \in C^{\mathcal{I}}\} \leq n\}$
data type exists	$\exists T.d$	$(\exists T.d)^{\mathcal{I}} = \{x \mid \exists y : (x,y) \in T^{\mathcal{I}} \text{ and } y \in d^{\mathbf{D}}\}$
data type value	$\forall T.d$	$(\forall T.d)^{\mathcal{I}} = \{x \mid \forall y : (x,y) \in T^{\mathcal{I}} \Rightarrow y \in d^{\mathbf{D}}\}$

Knowledge base

- TBox: set of axioms describing structure of domain
- ABox: set of axioms describing the individuals in a concrete situation



Objectives

- Cell-cycle ontology (based on existing ontologies such as GO, PO)
- Reasoning packages: FaCT++, RACER, …
- Find out cell-cycle regulated genes/proteins
- System
  - Satisfiability of concept
  - Concept classification (Subsumption)
  - Make implicit information explicit
  - Consistency, Instance checking
  - Constantly updated
  - User interface to answer queries / update the data

Description Logic

- Union

  “activation or inhibition” $\text{activation} \sqcup \text{inhibition}$

- Intersection

  “H-Type Cyclin and D-Type Cyclin” $\text{H-Type Cyclin} \sqcap \text{D-Type Cyclin}$

- Negation

  “not phophorylated” $\neg\ \text{phophorylated}$

- Quantifiers

  “all CDKs do not bind all cyclins (CDK-cyclin specificity)”

$$CDK \sqcap \exists.BINDS.Cyclin$$
$$(\forall a \in CDK)(\exists b \in Cyclin | (a, b) \in BINDS)$$

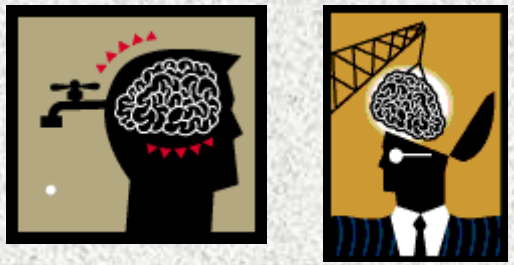
Background

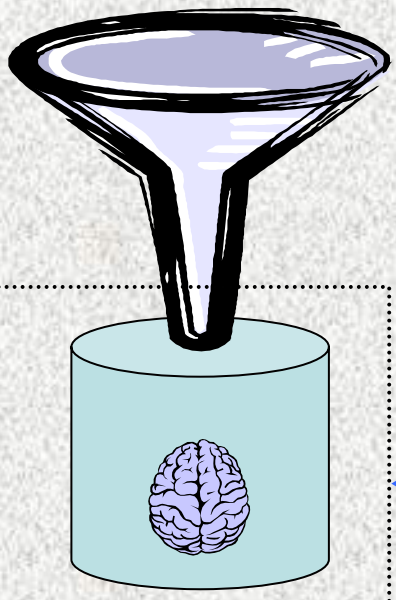
- A. Thaliana in-house cell-cycle data
- Domain experts annotation
- Protein features (calculated/predicted)
- Knowledge formally described (DL)
- DL-based system
- Protégé as KB development tool
- OWL is flexible and efficient at describing ontologies (DAML + OIL successor)

Architecture

Literature / in-house data

Expert domain knowledge





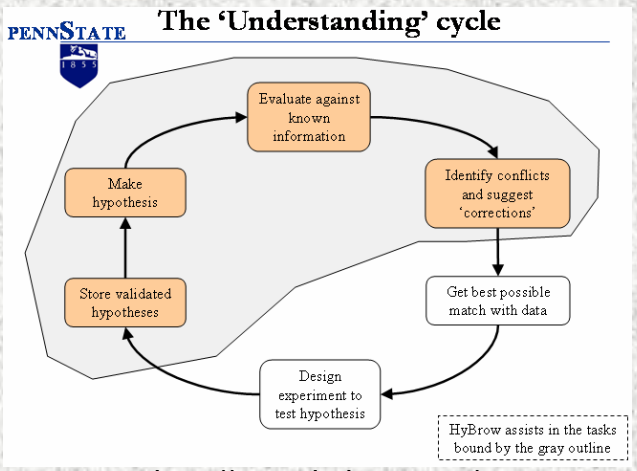
Target organism

Knowledge-based system

Visualization/UI

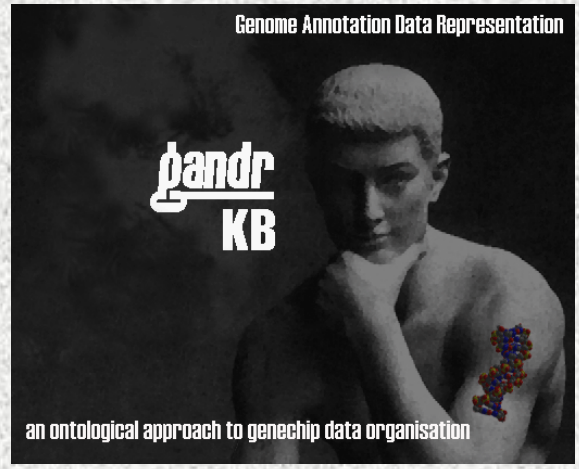
Similar work

- HyBrow: a prototype system for computer-aided hypothesis evaluation (Racunas et al. Bioinformatics 20(Suppl. 1), I257--64. 2004)



http://www.hybrow.org/

- GandrKB--ontological microarray annotation and visualization (Schober et al. Bioinformatics 21(11), 2785--6. 2005)



http://www.bioinf.mdc-berlin.de/~schober/GandrIntro/

Planning

- ~5 months: Ontology (prototype)
- Next: Knowledge refinement, validation of terms and annotated data
- Next step: Add protein features
- Evaluate results
- UI development
- Cell-cycle (domain knowledge) feedback is (will be) welcome

A knowledge-based system for plant cell-cycle elucidation

Erick Antezana
<erant@psb.ugent.be>