UNISINOS

Using protege to build a molecular network ontology

Eduardo Battistella
Renata Vieira
José Guilherme de Souza
Adriana Reis
João Paulo Muller da Silva

Cláudia K. Barcellos
Norma M. da Silva
Guilherme B. Bedin
José Carlos Mombach
Ney Lemke
Contact: lemke@unisinos.br

Universidade do Vale do Rio dos Sinos
São Leopoldo, RS Brazil

LBBC
Laboratório de Bioinformática
e Biologia Computacional

hp
invent

Presentation overview

- Introduction
- Motivation
- Domain analysis
- Using Protege to build MONET
- Ontology population
- Conclusions

Introduction

- Area of molecular biology
  - Great amount of data to deal with
- Different data bases with different management systems
- Useful information: through data modelling and integration
- Ontologies: enable an integrated view of these data

Motivation

- Ontologies in molecular biology
  - many are controlled vocabularies
  - many consider one specific area of knowledge
- Monet: Integrated topological model of
  - metabolism
  - regulation
  - protein interactions

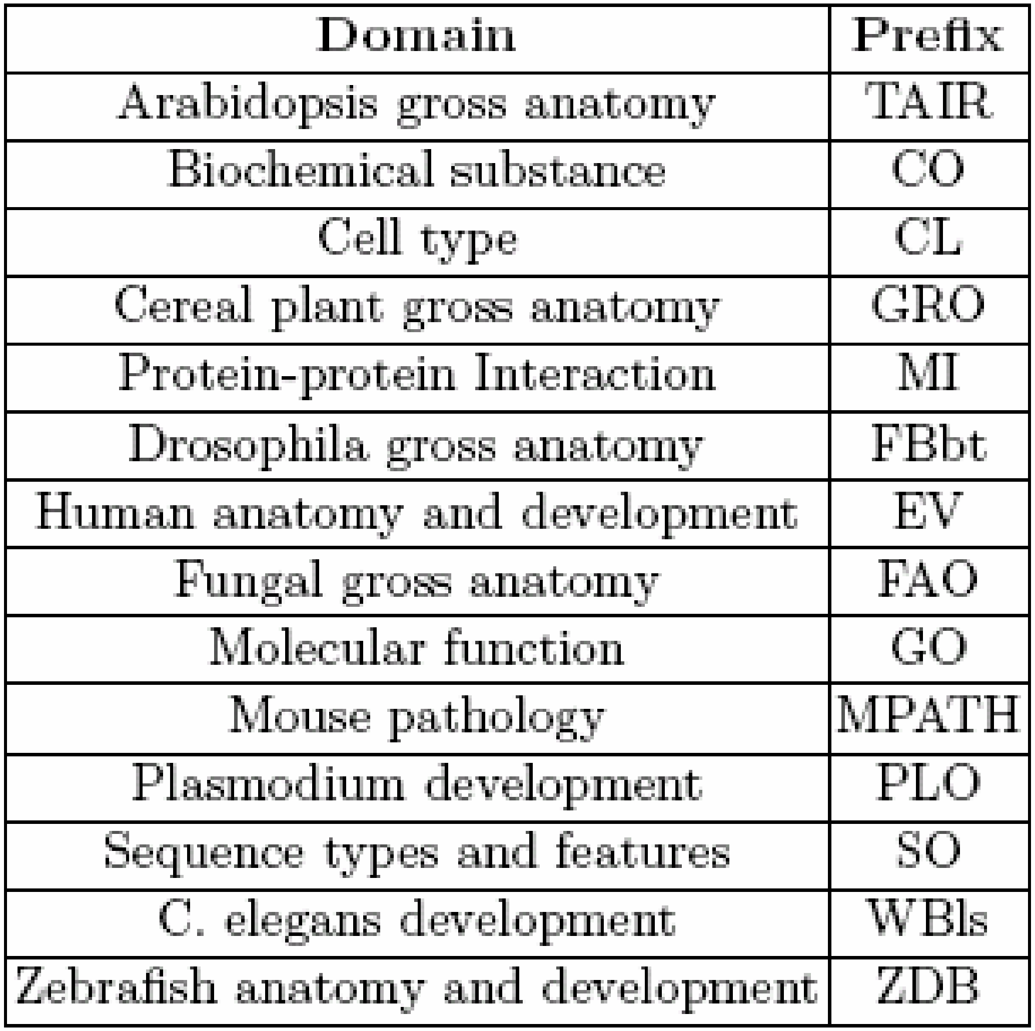
Domain analysis

- GO Gene Ontology
- SO Sequence Ontology
- PSI Proteomics Standards Initiative
- Mark up languages
  - SBML
  - MAGE ML
- OBO Open biological ontologies

OBO

- Open Biological Ontologies
- Sharing of ontologies from different biological domains

OBO



Domain	Prefix
Arabidopsis gross anatomy	TAIR
Biochemical substance	CO
Cell type	CL
Cereal plant gross anatomy	GRO
Protein-protein Interaction	MI
Drosophila gross anatomy	FBbt
Human anatomy and development	EV
Fungal gross anatomy	FAO
Molecular function	GO
Mouse pathology	MPATH
Plasmodium development	PLO
Sequence types and features	SO
C. elegans development	WBls
Zebrafish anatomy and development	ZDB

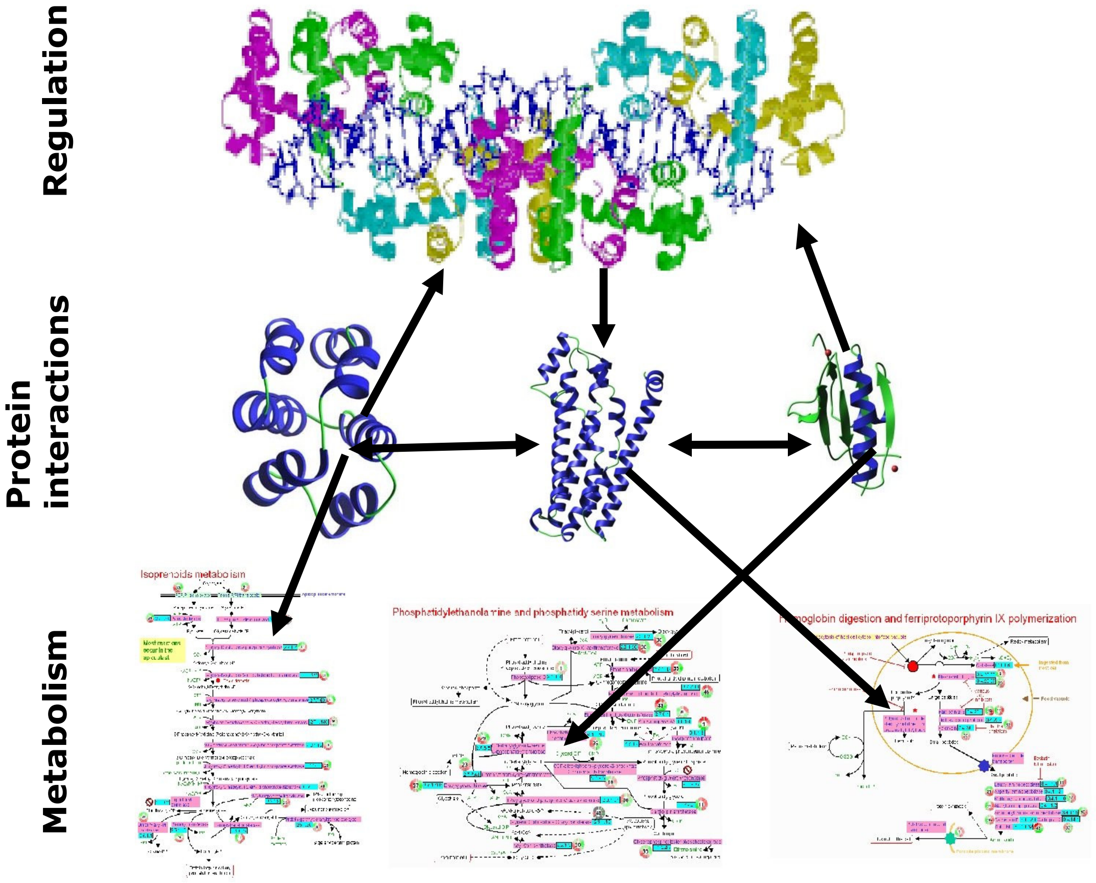
generic ontologies
(substance, cell type, …)

organism specific
(arabidopsis, drosophila, …)

Edited/viewed with
Protegé 2000
DAG-Edit

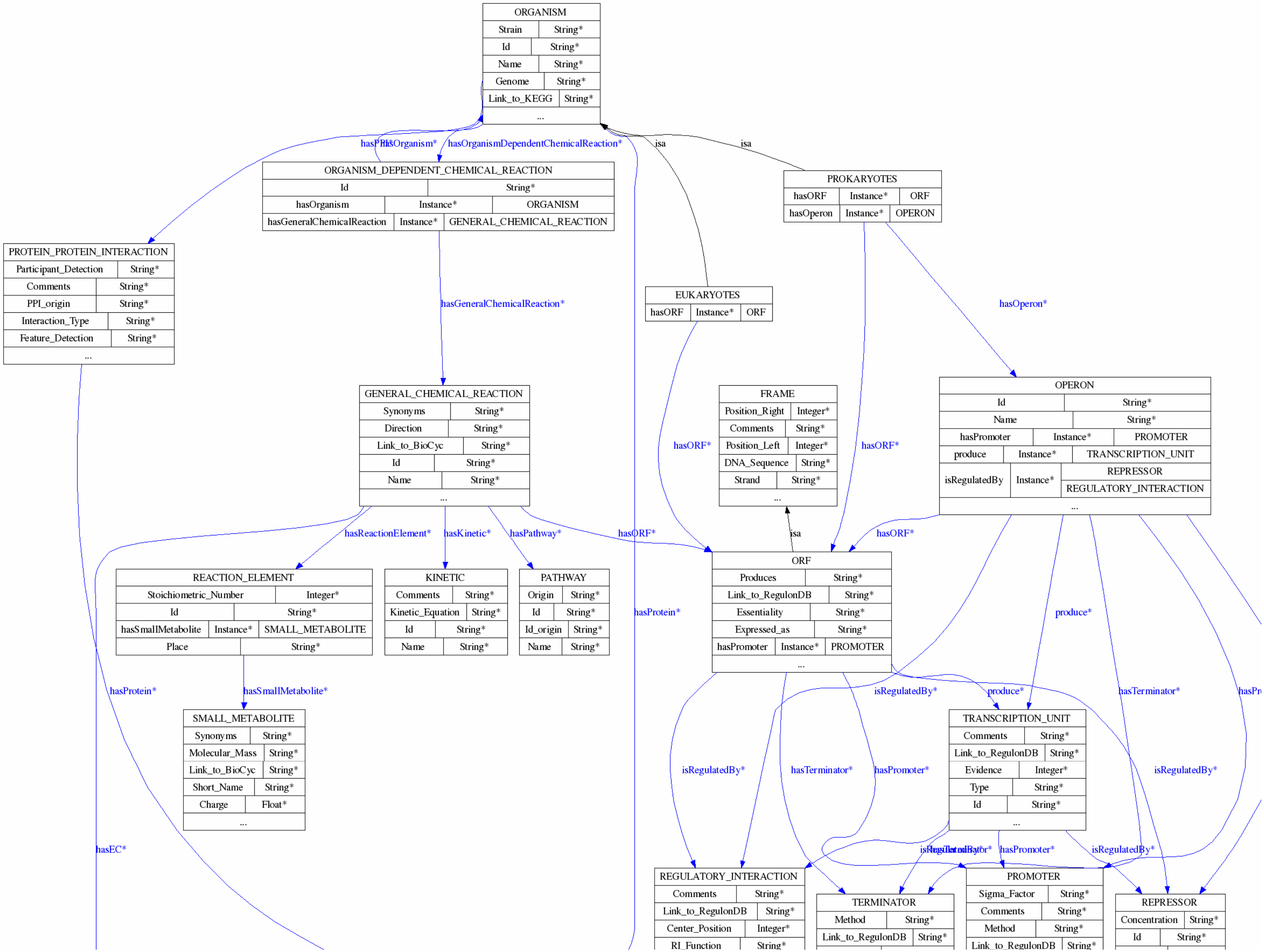
Using Protege to build MONET

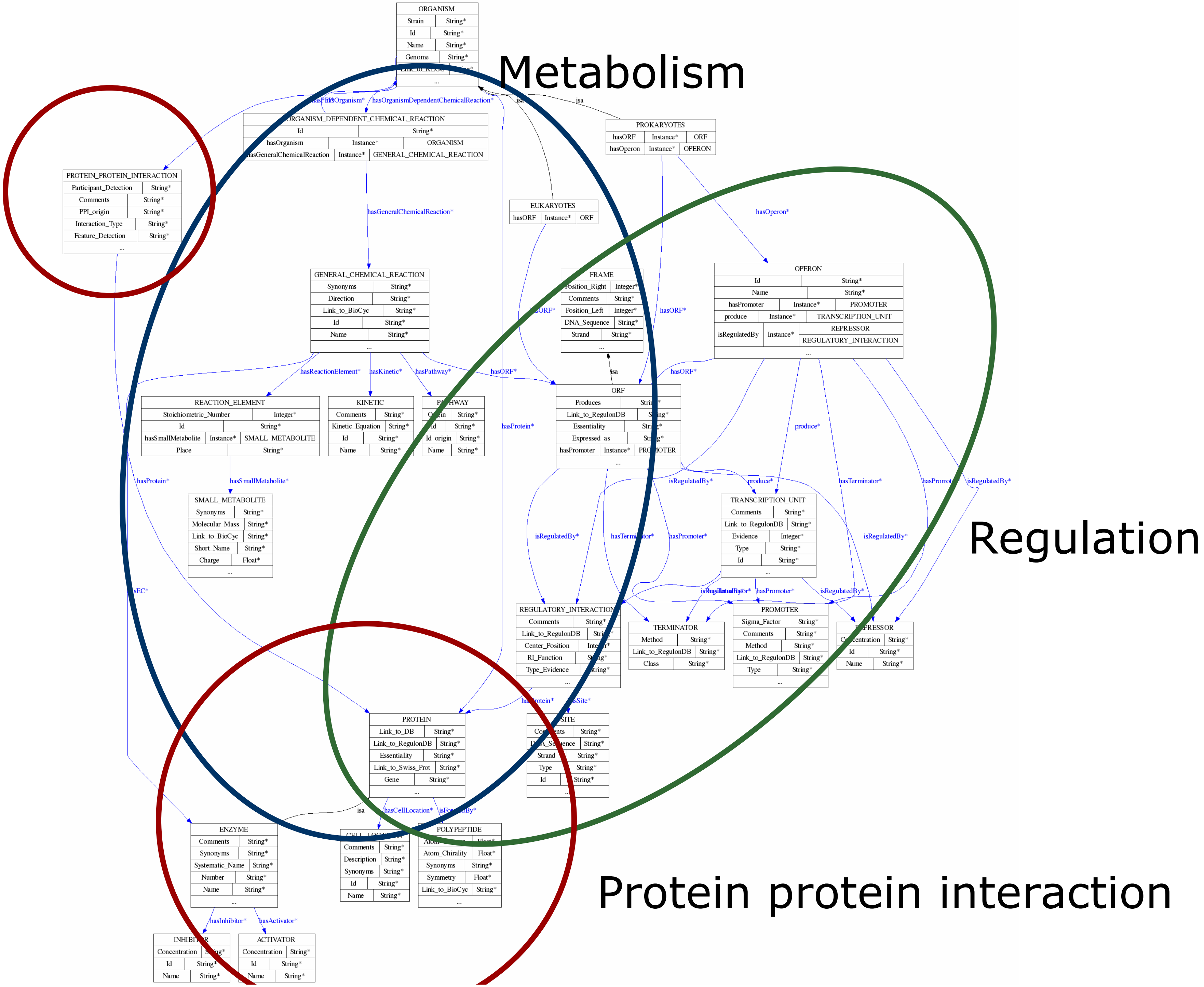
- Molecular Network Ontology
- Integrated topological models including
  - metabolism
  - regulation
  - protein interactions


Regulation
Protein interactions
Metabolism
Isoprenoids metabolism
Phosphatidylethanolamine and phosphatidylserine metabolism
Hemoglobin digestion and ferriprotoporphyrin IX polymerization

Using Protege to build MONET

- Ontology model
  - Definition of needed data according to the group research interests
    - Classes and properties
- Instantiation: automated process
- Data compiled from different databases


ORGANISM
Strain | String*
Id | String*
Name | String*
Genome | String*
Link_to_KEGG | String*
...
hasPPI*
hasOrganism*
hasOrganismDependentChemicalReaction*
isa
isa
ORGANISM_DEPENDENT_CHEMICAL_REACTION
Id | String*
hasOrganism | Instance* | ORGANISM
hasGeneralChemicalReaction | Instance* | GENERAL_CHEMICAL_REACTION
PROKARYOTES
hasORF | Instance* | ORF
hasOperon | Instance* | OPERON
PROTEIN_PROTEIN_INTERACTION
Participant_Detection | String*
Comments | String*
PPI_origin | String*
Interaction_Type | String*
Feature_Detection | String*
...
EUKARYOTES
hasORF | Instance* | ORF
hasGeneralChemicalReaction*
hasOperon*
hasORF*
hasORF*
GENERAL_CHEMICAL_REACTION
Synonyms | String*
Direction | String*
Link_to_BioCyc | String*
Id | String*
Name | String*
...
FRAME
Position_Right | Integer*
Comments | String*
Position_Left | Integer*
DNA_Sequence | String*
Strand | String*
...
OPERON
Id | String*
Name | String*
hasPromoter | Instance* | PROMOTER
produce | Instance* | TRANSCRIPTION_UNIT
isRegulatedBy | Instance* | REPRESSOR
REGULATORY_INTERACTION
...
hasReactionElement*
hasKinetic*
hasPathway*
hasORF*
isa
hasORF*
ORF
Produces | String*
Link_to_RegulonDB | String*
Essentiality | String*
Expressed_as | String*
hasPromoter | Instance* | PROMOTER
...
REACTION_ELEMENT
Stoichiometric_Number | Integer*
Id | String*
hasSmallMetabolite | Instance* | SMALL_METABOLITE
Place | String*
KINETIC
Comments | String*
Kinetic_Equation | String*
Id | String*
Name | String*
PATHWAY
Origin | String*
Id | String*
Id_origin | String*
Name | String*
hasProtein*
produce*
hasProtein*
hasSmallMetabolite*
isRegulatedBy*
produce*
hasTerminator*
hasPr
SMALL_METABOLITE
Synonyms | String*
Molecular_Mass | String*
Link_to_BioCyc | String*
Short_Name | String*
Charge | Float*
...
TRANSCRIPTION_UNIT
Comments | String*
Link_to_RegulonDB | String*
Evidence | Integer*
Type | String*
Id | String*
...
isRegulatedBy*
hasTerminator*
hasPromoter*
isRegulatedBy*
hasEC*
isRegulatedBy*
hasTerminator*
hasPromoter*
isRegulatedBy*
REGULATORY_INTERACTION
Comments | String*
Link_to_RegulonDB | String*
Center_Position | Integer*
RI_Function | String*
TERMINATOR
Method | String*
Link_to_RegulonDB | String*
PROMOTER
Sigma_Factor | String*
Comments | String*
Method | String*
Link_to_RegulonDB | String*
REPRESSOR
Concentration | String*
Id | String*



Metabolism

Regulation

Protein protein interaction

Monet: Ontology population

- Data from:
  Palsson (Reactions, Small Metabolites, Enzymes, Genes/ORF)
  Brite (Protein-Protein Interaction)
  KEGG (Reactions, Small Metabolites, Enzymes, Pathways, Organisms, Reaction Element, Proteins)
  RegulonDB (ORF, Promoters, Terminators, Transcription Unit, Site, Operon)
  NCBI (Proteins, Genes/ORF)
  PECDatabase (Enzymes Essentiality)
  Expasy (Enzymes).

Monet: Ontology population

Data from KEGG

KEGG (Reaction)

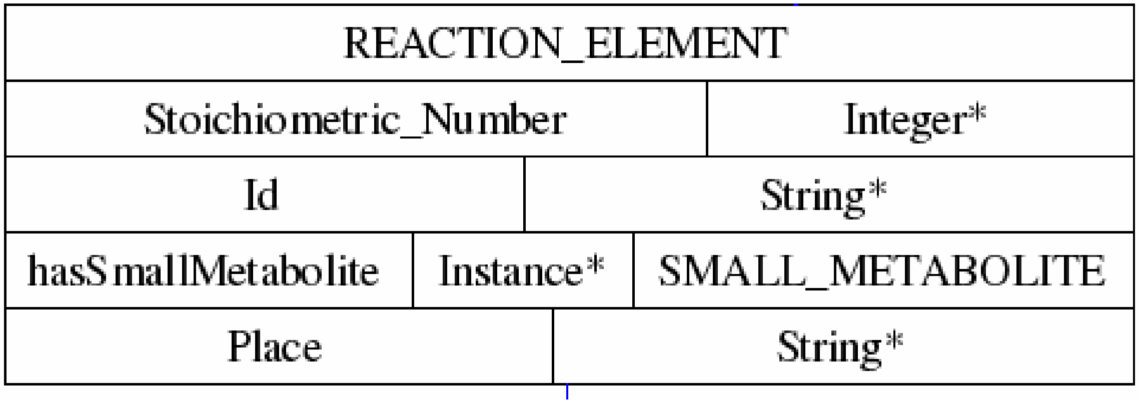
```
ENTRY       R00001
NAME        Polyphosphate polyphosphohydrolase
DEFINITION  Polyphosphate + H2O <=> Oligophosphate
EQUATION    C00890 + C00001 <=> C02174
RPAIR       A02844
ENZYME      3.6.1.10
```

KEGG (Compound)

```
ENTRY       C00001
NAME        H2O;
            Water;
            Water (JP14)
FORMULA     H2O
MASS        18.0106
REMARK      Drug: 7131
REACTION    R00001 R00002 R00004 R00005 R00009 R00010 R00011 R00017
            R00022 R00024 R00026 R00028 R00035 R00036 R00040 R00041
            R00044 R00045 R00046 R00048 R00052 R00053 R00054 R00055
            R00056 R00057 R00058 R00059 R00060 R00061 R00068 R00070
            R00072 R00074 R00077 R00078 R00080 R00081 R00082 R00083
```

Monet: Ontology population

Data reorganization according to monet concepts



REACTION_ELEMENT		
Stoichiometric_Number	Integer*	
Id	String*	
hasSmallMetabolite	Instance*	SMALL_METABOLITE
Place	String*	

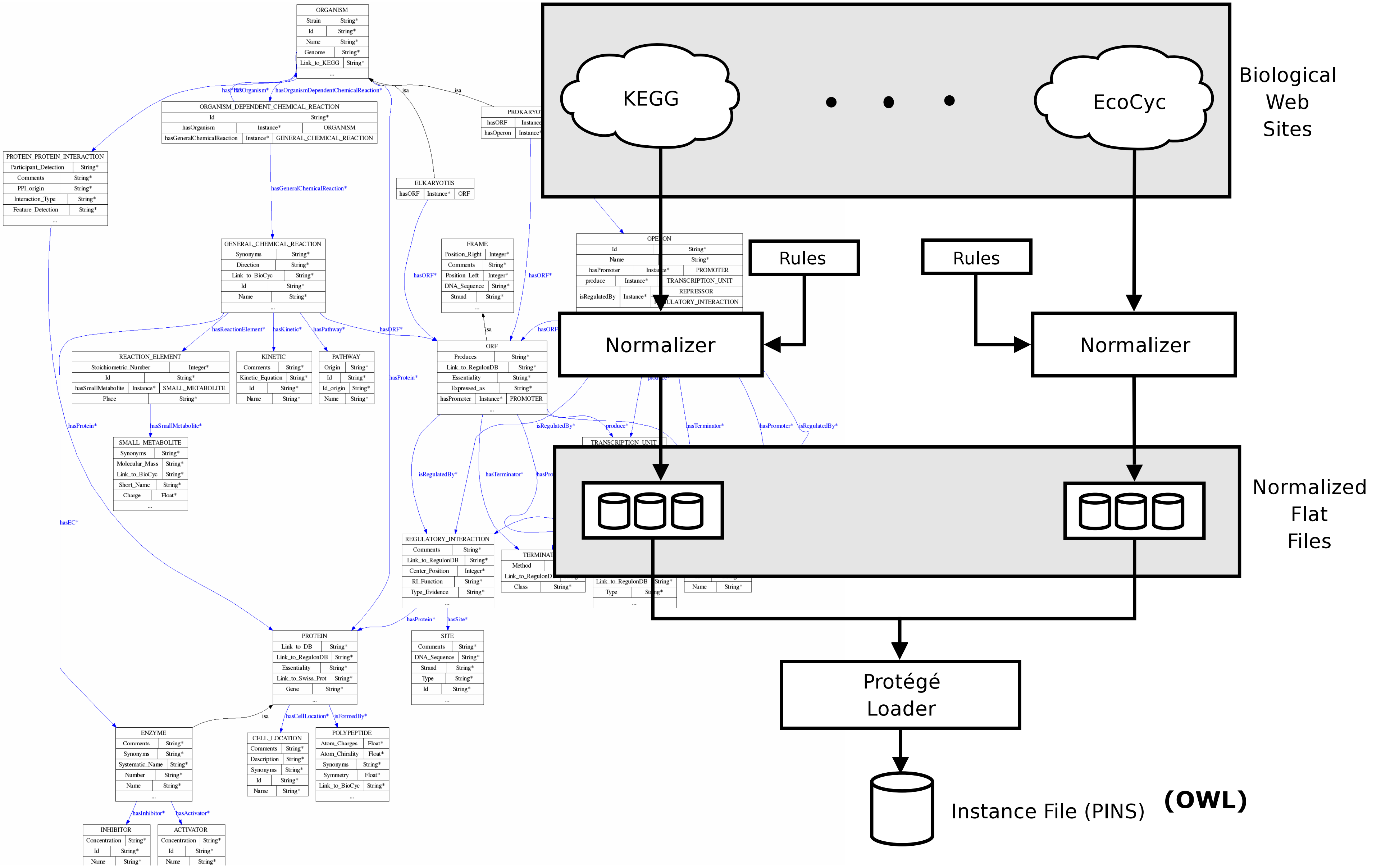
REACTION ELEMENT

```
Id               | monet_id | st_n |small_metabolite_instance | place
----------------+------------+------+-------------------------+-------+-----------
MREE00001     |        1      | 1    |MSMM00001               | L

MREE00002     |        2      | 1    |MSMM00013               | L

MREE08946     |     8946     | 2    |MSMM00009               | R
...
```

Generating owl and owl instances

```
<<owl:FunctionalProperty rdf:ID="hasElement">
    <rdfs:domain rdf:resource="#REACTION_ELEMENT"/>
    <rdfs:range rdf:resource="#SMALL_METABOLITE"/>
    <rdf:type
   rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
  </owl:FunctionalProperty>

<owl:FunctionalProperty rdf:ID="Stoichiometric_Number">
    <rdfs:domain rdf:resource="#REACTION_ELEMENT"/>
    <rdfs:comment
        rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    A chemical reaction of known stoichiometry can be  written in
   general as: aA + bB + ... For the reaction products Y and Z the
    numbers y and z are known as the stoichiometric numbers, vY and
   vZ, for...
...
```

Automated instance generation



Biological Web Sites

KEGG

• • •

EcoCyc

Rules

Rules

Normalizer

Normalizer

Normalized Flat Files

Protégé Loader

Instance File (PINS)

(OWL)

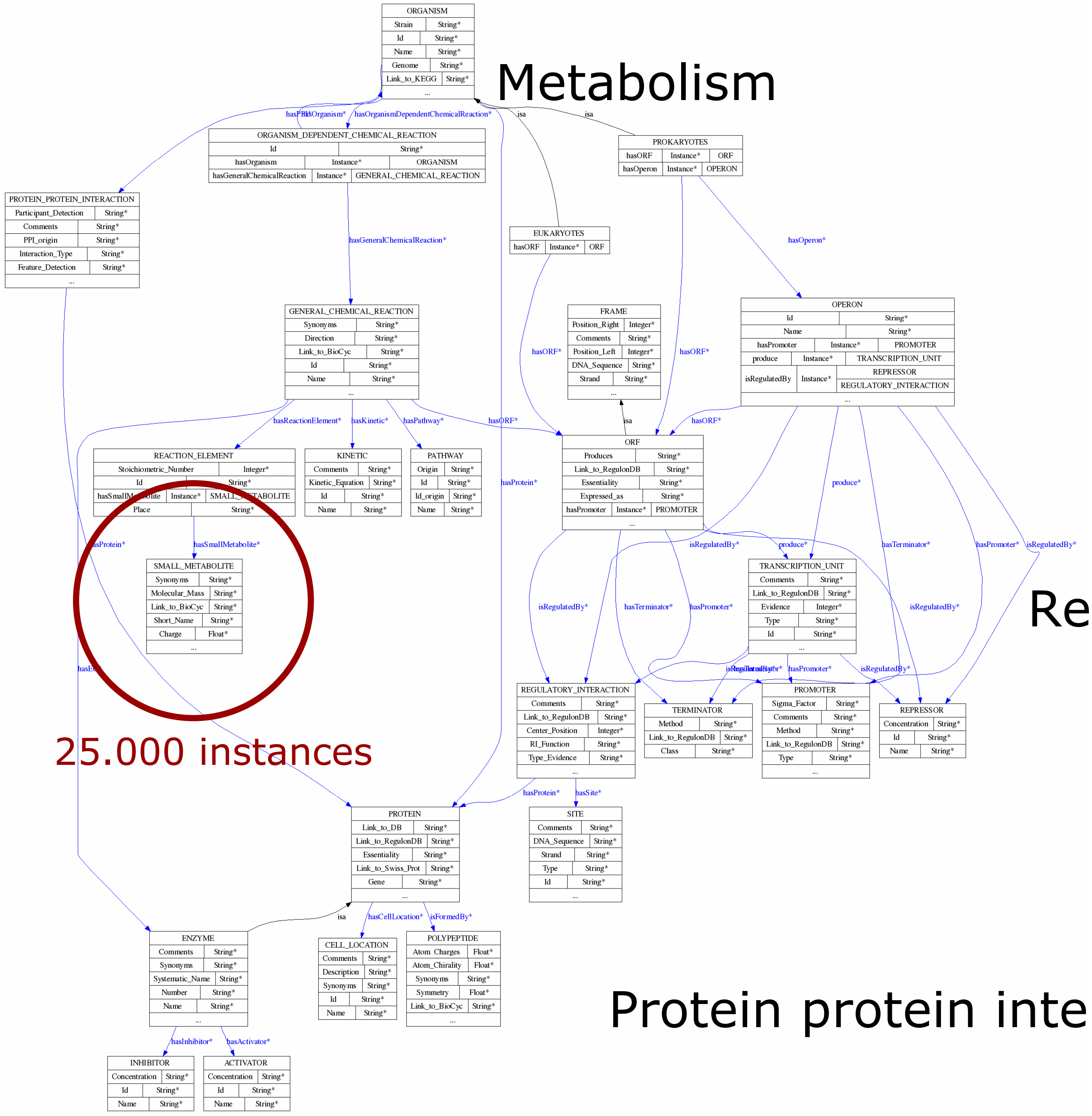
Instances overview

Concept	Instances	Concept	Instances	Concept	Instances
General Chemical Reaction	4496	Enzyme	3407	Operon	785
Organsim Dependent Chemical Reaction	3228	ORF	4410	Organism	3
Small Metabolite	3361	Product	8990	Promoter	973
Protein-Protein Interaction	12248	Reaction Element	17757	Protein	10201
Regulatory Interaction	1376	Site	1216	Pathway	126
Transcription Unit	833	Substrate	8767	Terminator	137
General Chemical Reaction	4496	Enzyme	3407	Operon	785

Table 1. Number of instances for each concept of Monet Ontology

Conclusions

- Ongoing future work
  - Monet: integrated model of regulation, protein interaction and metabolism
    - Ontology model (constantly updated)
    - Adapting to OWL-DL
      - Still mainly based on properties specification
      - We have to think about logical conditions
    - Ontology population (growing data availability)
      - Small metabolites: now 25.000 instances



Metabolism

Regulation

25.000 instances

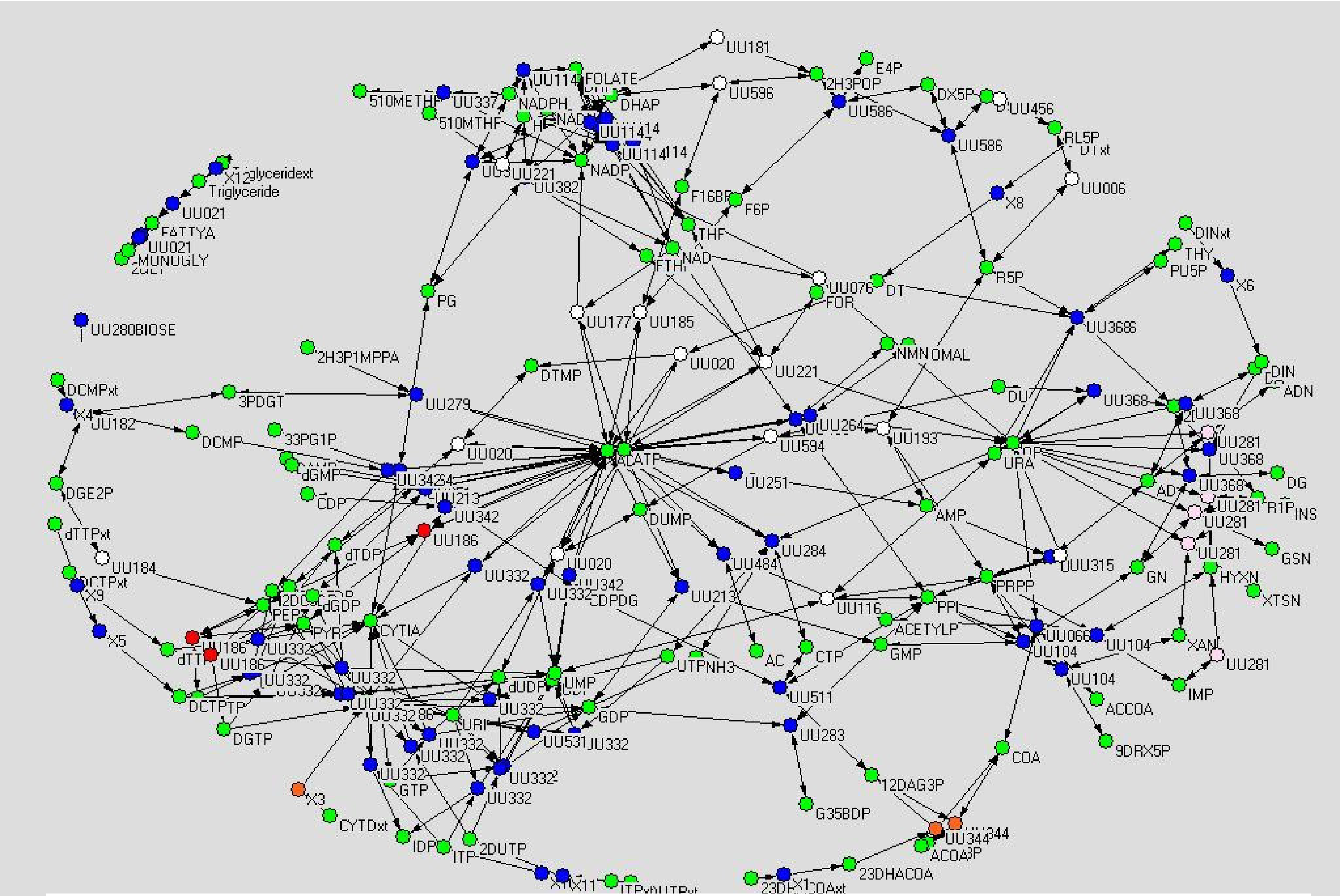
Protein protein interaction

Conclusions

- Problems
  - Data intensive ontology
  - Generating owl instances
    - Small metabolites data only
      - From 2mb .txt data to 20 mb .owl data
  - OWL Databases ??

Conclusions

- Advantages
  - Group communication (biologists, computer scientists, computer engineers)
  - Partial use of data for protein essentiality prediction



Bipartite graph of the metabolic network of Ureaplasma urealyticum. Green nodes metabolites, other nodes represent enzymes

Conclusions

- Plans for the future
  - Use of collected data for other tasks
  - See for the data intensive problem
  - Explore data query
    - Jena
  - Knowledge discovery in structured data (owl)

Acknowledgments

- This work was developed in collaboration with HP Brazil R&D.
- This work was partially supported by CNPq and FAPERGS (Brazilian Agencies).