Integrating Hierarchical Controlled Vocabularies with OWL Ontology

Melissa J. Davis
ARC Centre of Excellence in Bioinformatics
and
Institute for Molecular Bioscience, University of Queensland

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

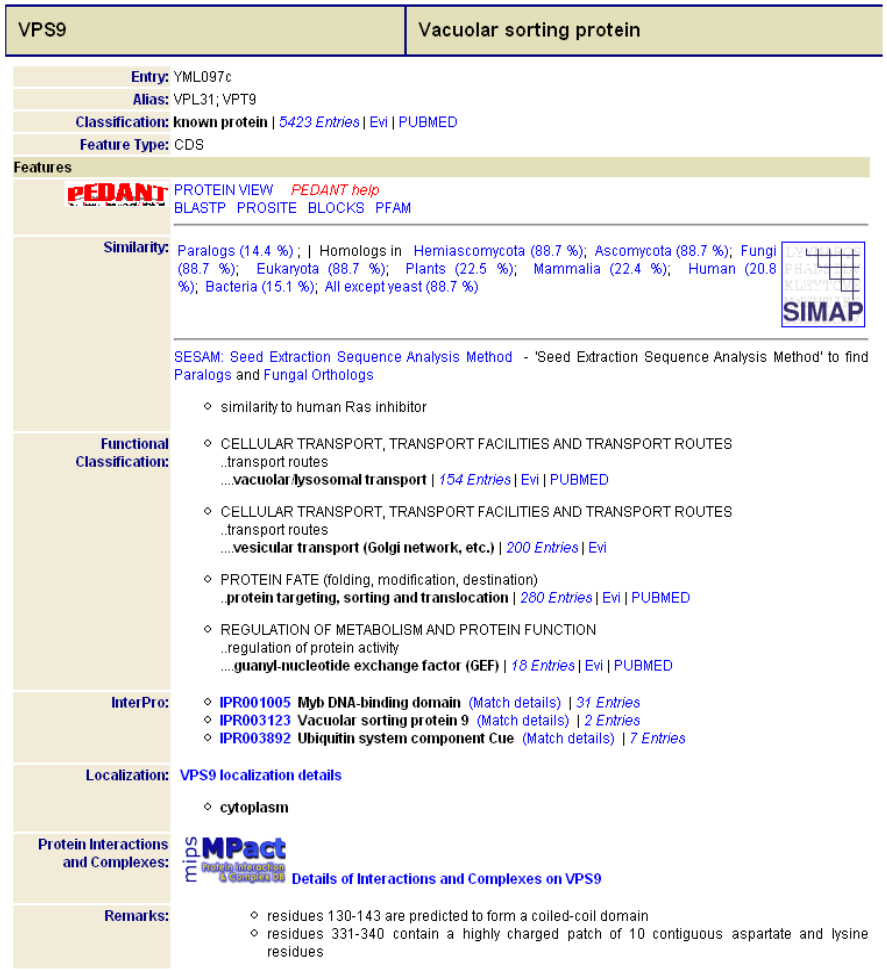
IMB Institute for Molecular Bioscience

Modelling and Analysis of Biological Network Activity

- Unravelling complex interactions between biological networks at varying levels of resolution requires systems approach
- Computational modelling and analysis of large-scale protein-protein interaction (PPI) and compound activity networks
- Semantic Web (SW) offers flexible technologies for creating meaningful representations of data on the web → *Semantic Interactome Model*
- Knowledge representation using SW standards Resource Description Framework (RDF) and Web Ontology Language (OWL) enables machine inference and facilitates knowledge discovery

Semantic Web

- **Current web**: Most information is natural language
  - Meaningful to human users who understand the meaning of natural language
- **Semantic Web**: standards for publishing machine-readable information on the web
  - Standard formats for integration and exchange of data (RDF)
  - Formal language to express semantics (the meaning of concepts) - OWL
  - Unambiguous representation



VPS9	Vacuolar sorting protein

Entry: YML097c
Alias: VPL31; VPT9
Classification: known protein | *5423 Entries* | Evi | PUBMED
Feature Type: CDS

Features

PEDANT — PROTEIN VIEW *PEDANT help*
BLASTP PROSITE BLOCKS PFAM

Similarity: Paralogs (14.4 %) ; | Homologs in Hemiascomycota (88.7 %); Ascomycota (88.7 %); Fungi (88.7 %); Eukaryota (88.7 %); Plants (22.5 %); Mammalia (22.4 %); Human (20.8 %); Bacteria (15.1 %); All except yeast (88.7 %)

SIMAP

SESAM: Seed Extraction Sequence Analysis Method - 'Seed Extraction Sequence Analysis Method' to find Paralogs and Fungal Orthologs

- similarity to human Ras inhibitor

Functional Classification:

- CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES
  ..transport routes
  ....**vacuolar/lysosomal transport** | *154 Entries* | Evi | PUBMED
- CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES
  ..transport routes
  ....**vesicular transport (Golgi network, etc.)** | *200 Entries* | Evi
- PROTEIN FATE (folding, modification, destination)
  ..**protein targeting, sorting and translocation** | *280 Entries* | Evi | PUBMED
- REGULATION OF METABOLISM AND PROTEIN FUNCTION
  ..regulation of protein activity
  ....**guanyl-nucleotide exchange factor (GEF)** | *18 Entries* | Evi | PUBMED

InterPro:

- IPR001005 **Myb DNA-binding domain** (Match details) | *31 Entries*
- IPR003123 **Vacuolar sorting protein 9** (Match details) | *2 Entries*
- IPR003892 **Ubiquitin system component Cue** (Match details) | *7 Entries*

Localization: VPS9 localization details

- **cytoplasm**

Protein Interactions and Complexes: mips MPact — Details of Interactions and Complexes on VPS9

Remarks:

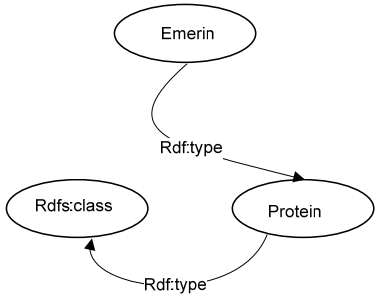
- residues 130-143 are predicted to form a coiled-coil domain
- residues 331-340 contain a highly charged patch of 10 contiguous aspartate and lysine residues

Semantic Web Technologies

- Resource Description Framework (RDF) and RDF Schema (RDFS)
  - RDF for resources; RDFS for vocabulary
  - Data presented as triples, graph or xml

    <subject> <predicate> <object>

    <Protein> <rdf:type> <rdfs:Class>
    <Emerin> <rdf:type> <Protein>



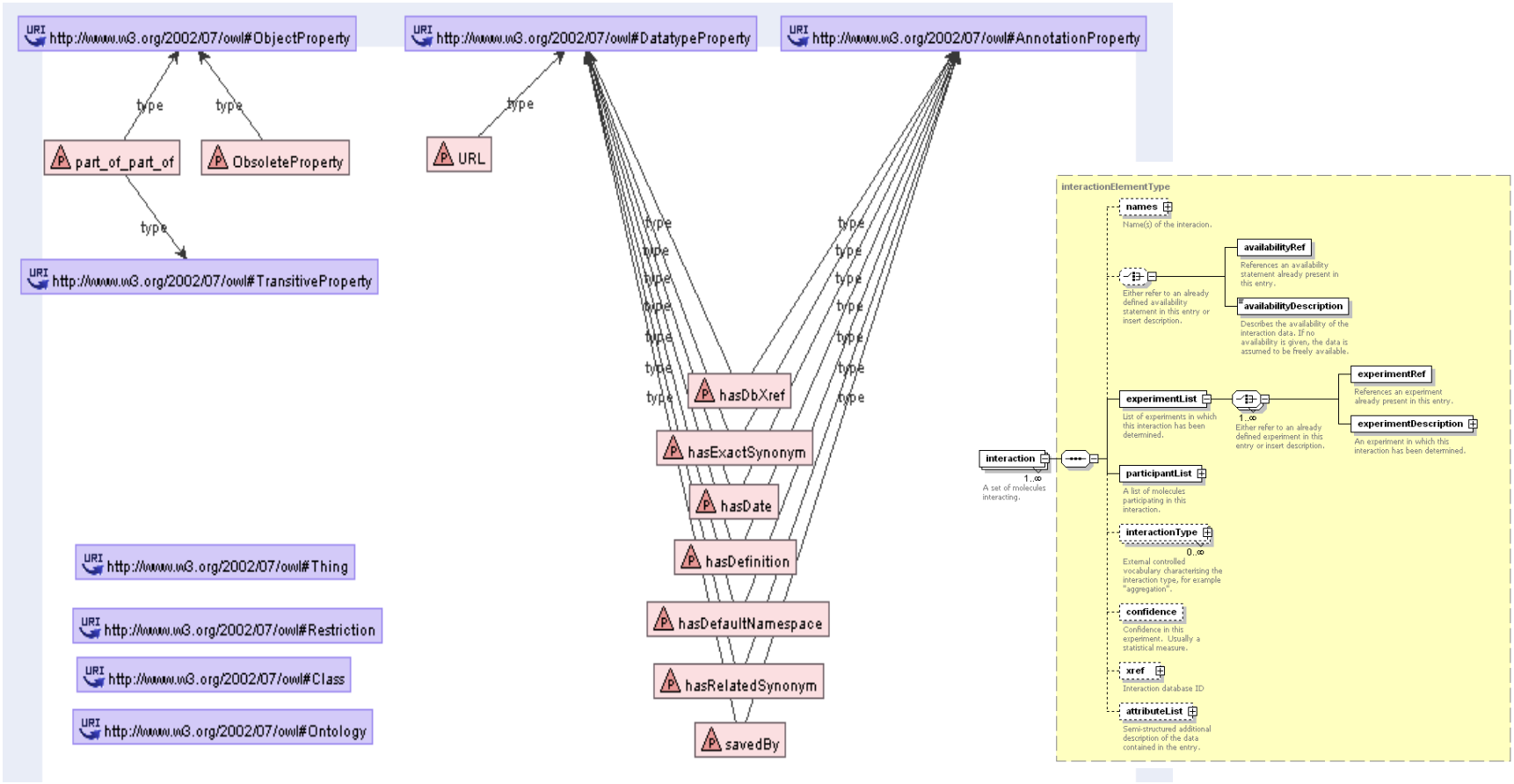
- OWL
  - Web Ontology Language for description of ontology
  - OWL is RDF: can be expressed in triples, graph or xml
  - cf Open Biomedical Ontology (OBO) format – familiar to users of GO
- Inference engines
  - software to reason about information in a knowledge representation and infer new data from what is presented
- SPARQL
  - query language for RDF
  - Returns results sets or RDF graphs

Knowledge representation with Ontology

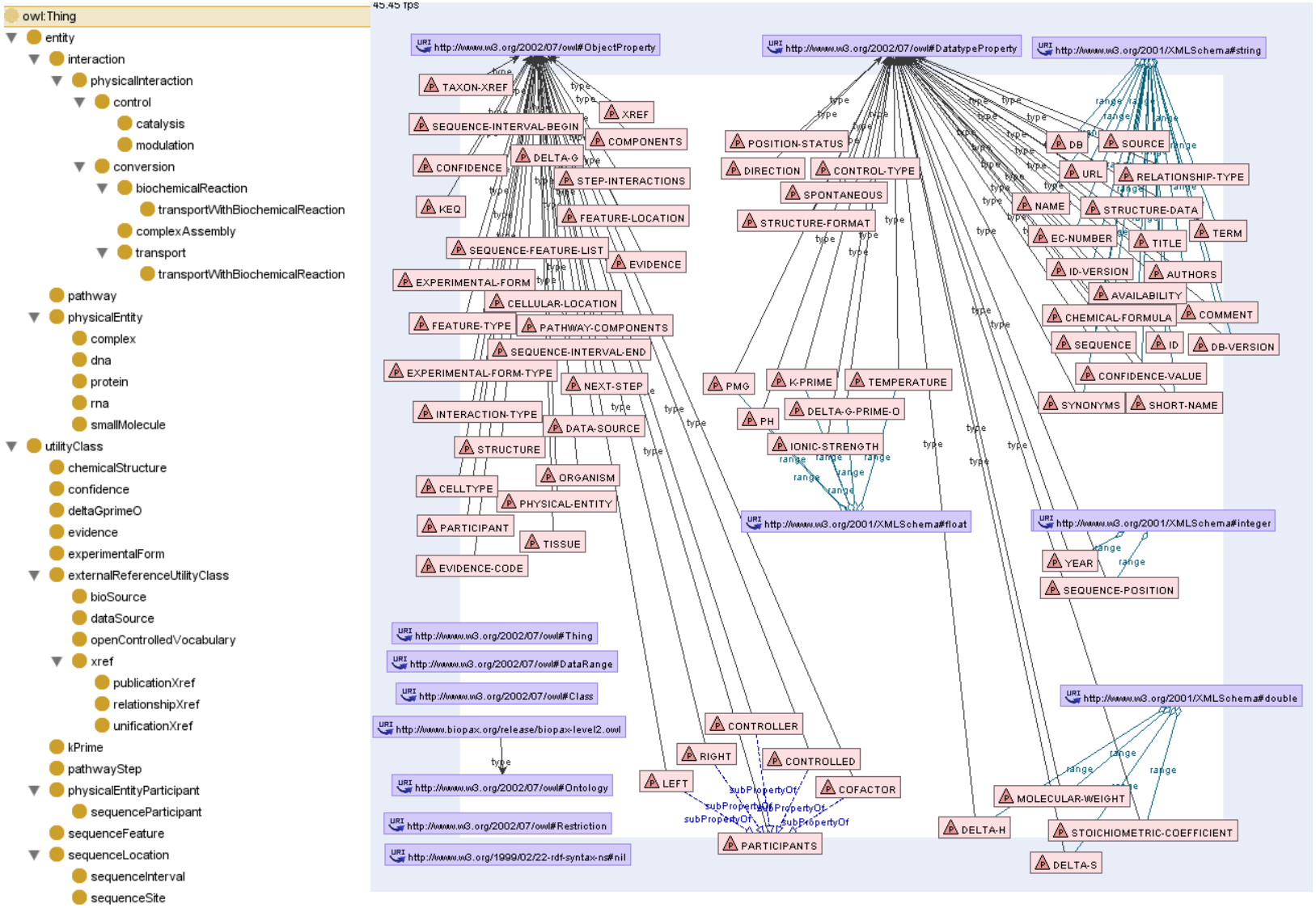
- Ontology can mean different things to different people, and frequently the intended use of the ontology shapes its structure and capability
  - Unstructured but controlled vocabulary
  - Structured, frequently hierarchical controlled vocabulary (eg: OBO)
  - Semantic web standard Web Ontology Language (OWL)
  - Lists of terms, hierarchies of terms, logically consistent knowledge representations from which new data may be inferred
  - Different approaches have strengths and weaknesses, well illustrated by two examples…

Current representations – PSI-MI…

- Protein Standards Initiative – Molecular Interactions (PSI-MI)
  - OBO format ontology
  - Hierarchical arrangement with deep coverage but limited relations



... and BioPAX

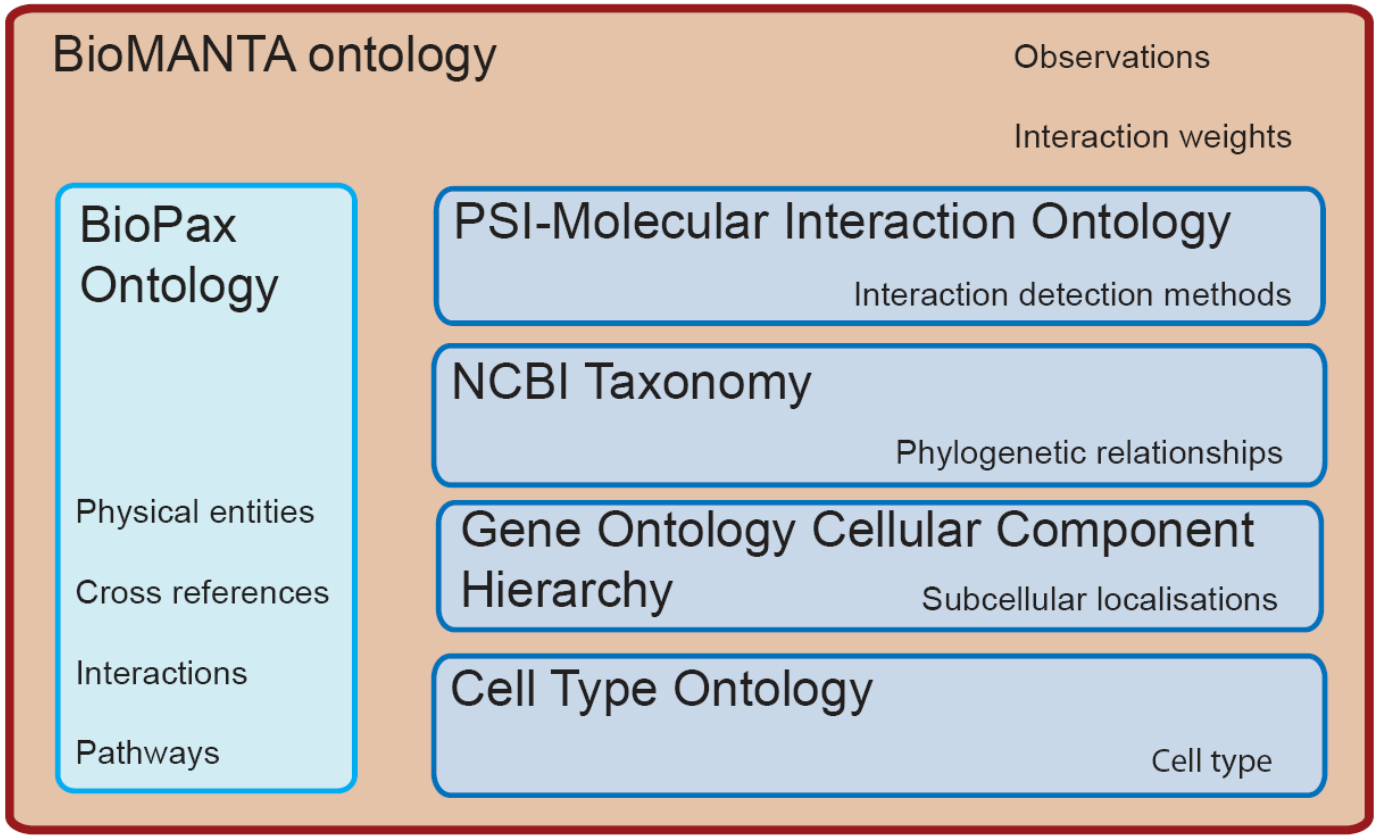

owl:Thing
entity
interaction
physicalInteraction
control
catalysis
modulation
conversion
biochemicalReaction
transportWithBiochemicalReaction
complexAssembly
transport
transportWithBiochemicalReaction
pathway
physicalEntity
complex
dna
protein
rna
smallMolecule
utilityClass
chemicalStructure
confidence
deltaGprimeO
evidence
experimentalForm
externalReferenceUtilityClass
bioSource
dataSource
openControlledVocabulary
xref
publicationXref
relationshipXref
unificationXref
kPrime
pathwayStep
physicalEntityParticipant
sequenceParticipant
sequenceFeature
sequenceLocation
sequenceInterval
sequenceSite
http://www.w3.org/2002/07/owl#ObjectProperty
http://www.w3.org/2002/07/owl#DatatypeProperty
http://www.w3.org/2001/XMLSchema#string
TAXON-XREF
XREF
SEQUENCE-INTERVAL-BEGIN
COMPONENTS
CONFIDENCE
DELTA-G
STEP-INTERACTIONS
KEQ
FEATURE-LOCATION
SEQUENCE-FEATURE-LIST
EVIDENCE
EXPERIMENTAL-FORM
CELLULAR-LOCATION
FEATURE-TYPE
PATHWAY-COMPONENTS
SEQUENCE-INTERVAL-END
EXPERIMENTAL-FORM-TYPE
NEXT-STEP
INTERACTION-TYPE
DATA-SOURCE
STRUCTURE
ORGANISM
CELLTYPE
PHYSICAL-ENTITY
PARTICIPANT
TISSUE
EVIDENCE-CODE
POSITION-STATUS
DIRECTION
CONTROL-TYPE
SPONTANEOUS
STRUCTURE-FORMAT
PMG
K-PRIME
TEMPERATURE
PH
DELTA-G-PRIME-O
IONIC-STRENGTH
DB
SOURCE
URL
RELATIONSHIP-TYPE
NAME
STRUCTURE-DATA
EC-NUMBER
TITLE
TERM
ID-VERSION
AUTHORS
AVAILABILITY
CHEMICAL-FORMULA
COMMENT
SEQUENCE
ID
DB-VERSION
CONFIDENCE-VALUE
SYNONYMS
SHORT-NAME
http://www.w3.org/2001/XMLSchema#float
http://www.w3.org/2001/XMLSchema#integer
YEAR
SEQUENCE-POSITION
http://www.w3.org/2001/XMLSchema#double
http://www.w3.org/2002/07/owl#Thing
http://www.w3.org/2002/07/owl#DataRange
http://www.w3.org/2002/07/owl#Class
http://www.biopax.org/release/biopax-level2.owl
http://www.w3.org/2002/07/owl#Ontology
http://www.w3.org/2002/07/owl#Restriction
http://www.w3.org/1999/02/22-rdf-syntax-ns#nil
CONTROLLER
RIGHT
CONTROLLED
LEFT
COFACTOR
PARTICIPANTS
MOLECULAR-WEIGHT
DELTA-H
STOICHIOMETRIC-COEFFICIENT
DELTA-S
type
range
subPropertyOf

Ontology integration problem

- Ontologies developed for specific uses
  - Describing pathways
  - Exchanging information
  - Controlling vocabulary for annotation
- Systems Biology necessitates overlapping domains and concepts
- Knowledge acquisition is costly and critical

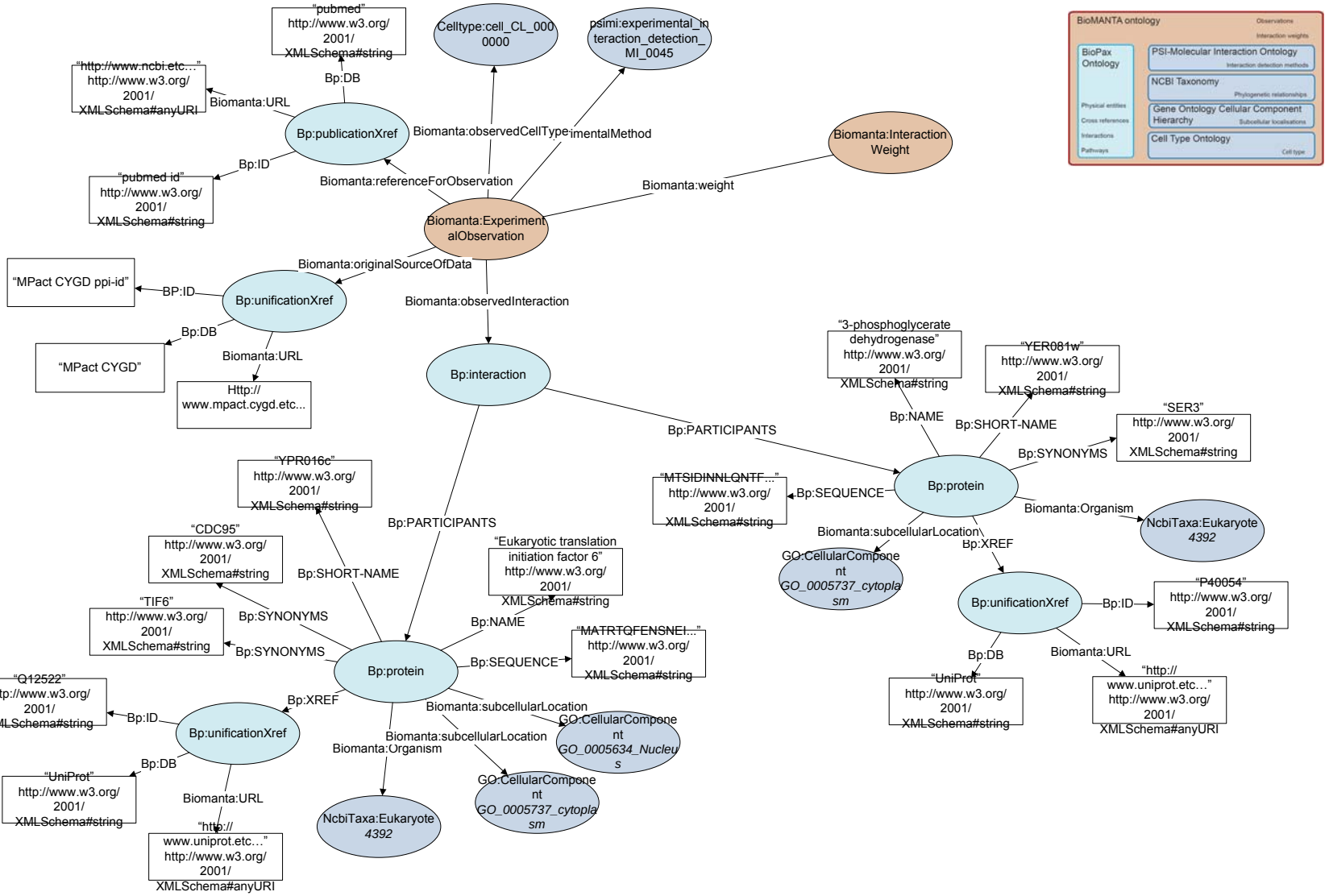
  → reuse of existing ontology preferable

BioMANTA ontology

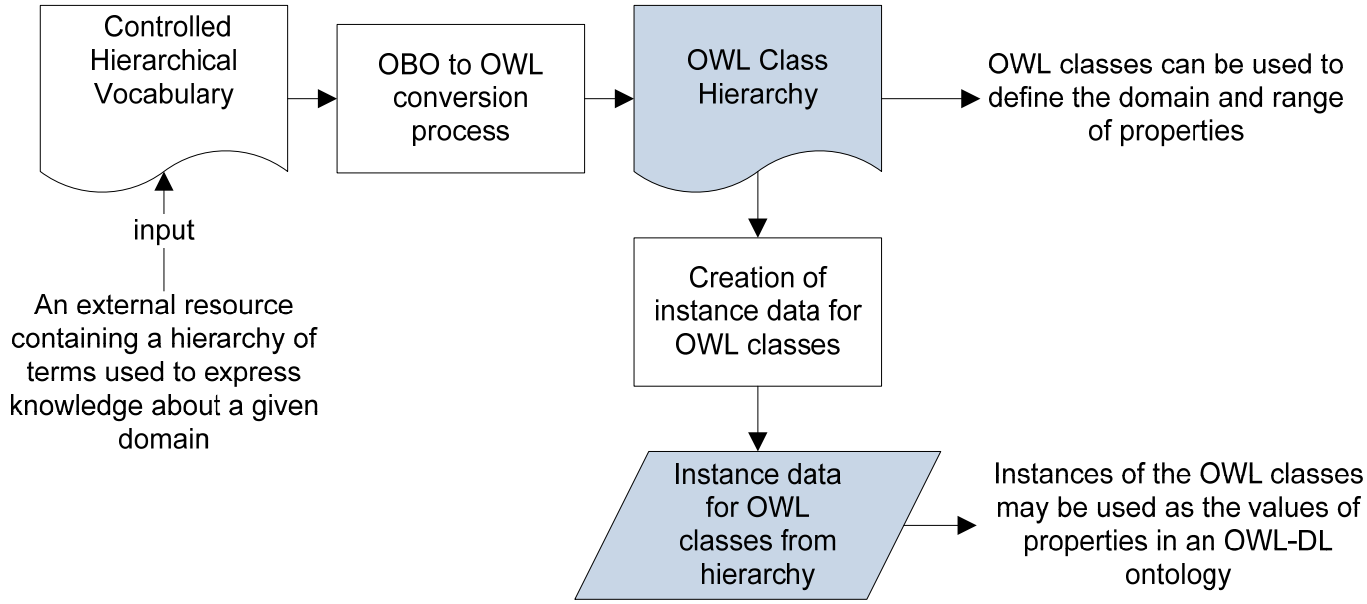


- OWL based ontology with imported modules from relevant ontologies
- Limit creation of new classes in BioMANTA ontology and use classes from existing ontologies where ever possible

BioMANTA Semantic Interactome Model

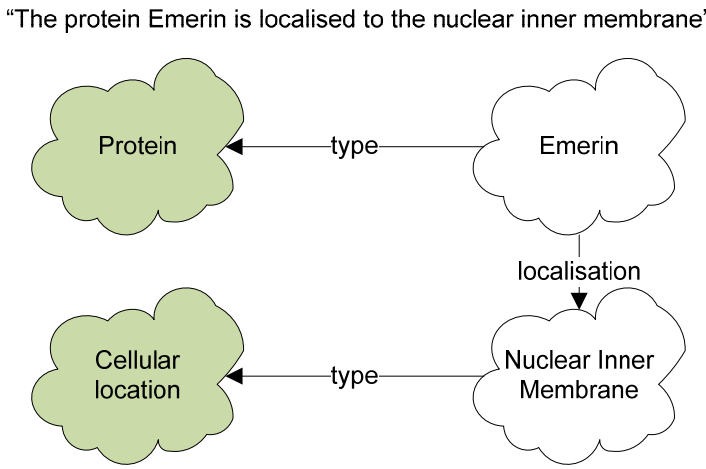


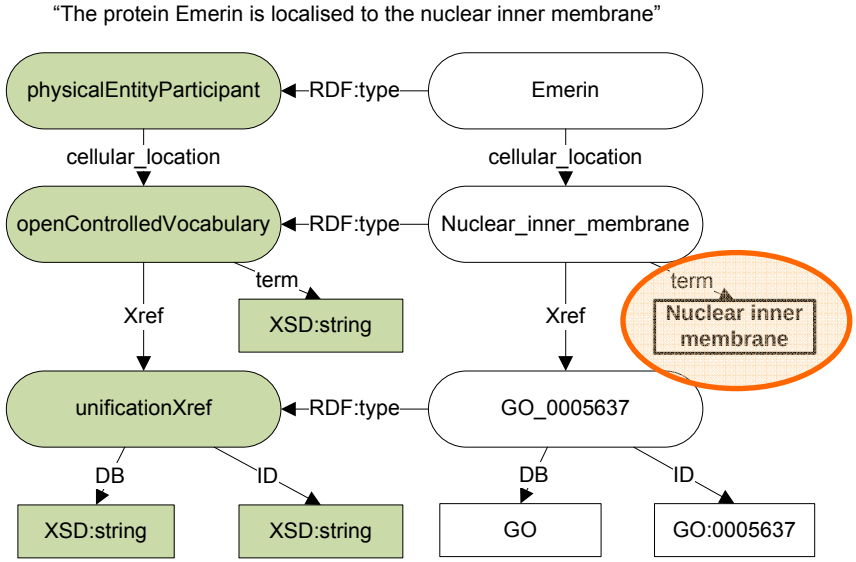
Ontology integration strategy



- Conversion and instancing of hierarchical controlled vocabulary
- Benefit: retention of meaning when framing expressions or instances of data

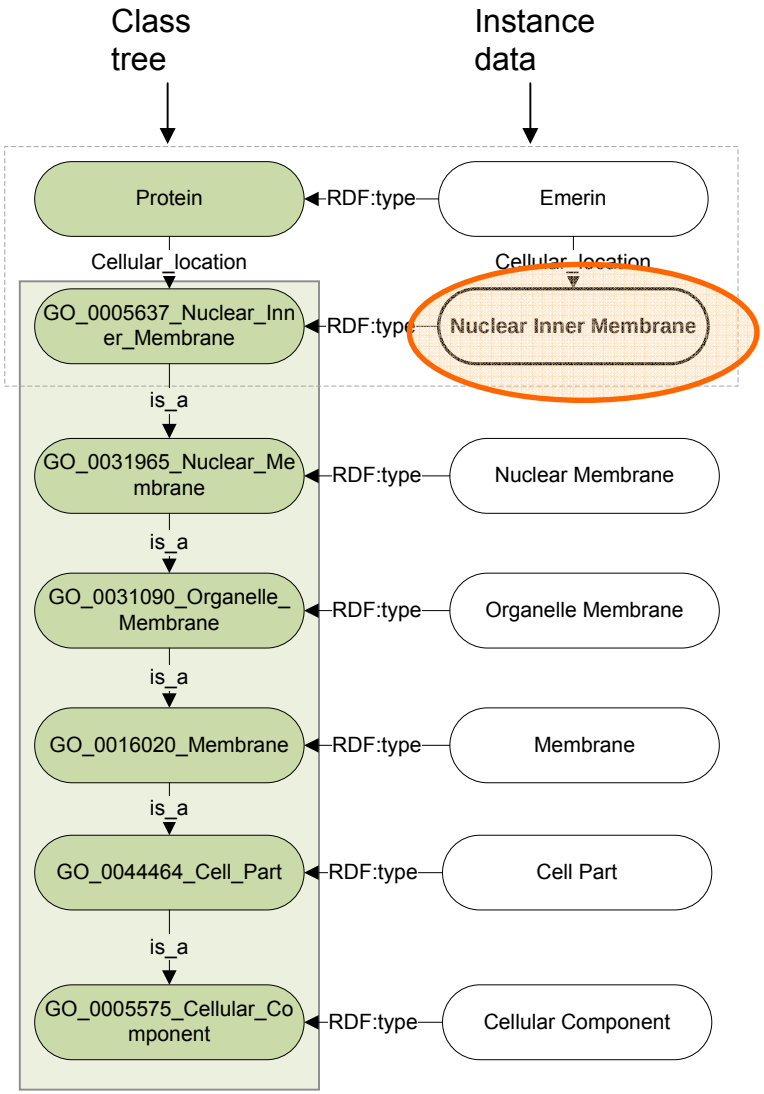
Example: Subcellular localisation





- Natural language expression
- BioPAX instance data expression
  - Text string
  - Dead end

Integrated Ontology in BioMANTA



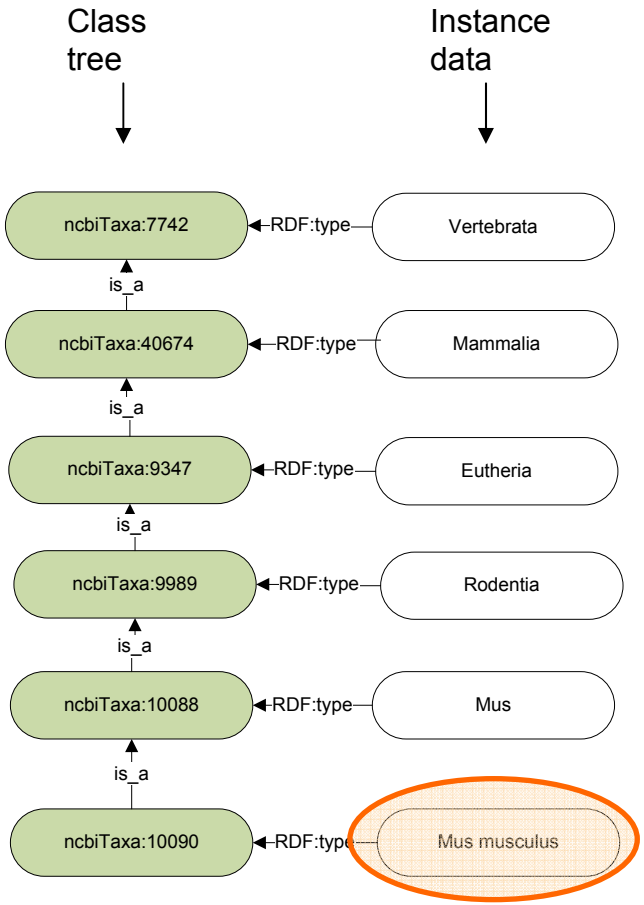
Statement that Emerin is localised to the nuclear inner membrane

Nuclear inner membrane is not merely a label, or annotation, it is embedded in a hierarchy of terms which capture additional meaning

Nuclear Inner Membrane *is_a* nuclear membrane and *is_a* membrane

Example: species and phylogeny

- How to reason about species relationships?
- Currently, labelling with text strings strips meaning from terms:
  - “mouse”; “mus musculus”; “10090”
  - meaning is lost – the string could be anything…
  - Nothing to say that “mouse” is mammal, which is basic background interpretation to a biologist
  - Class hierarchy (sub-class / super-class relationships) preserves meaning and captures assumptions



Conclusions

- Much biological knowledge is stored in controlled vocabularies inaccessible to machine reasoners
- Integration process salvages important collections of expert knowledge and brings them into SW format
- Creation of semantically rich labels enables inference
- Machine reasoners gain explicit access to the meaning of labeling terms – an improvement over text string inclusion used in other methods

Acknowledgements

Chris Bouton
Victor Farutin
Mike Schaffer
Fred Jerva

Pfizer Research and Technology Center, Cambridge, Massachusetts, US

Jane Hunter
Andrew Newman
Imran Khan
Yuan-Fang Li

School of ITEE, UQ

Mark Ragan
Kevin Burrage
Shoaib Sehgal

IMB & ARC Centre of Excellence in Bioinformatics

All members of Group Ragan @IMB