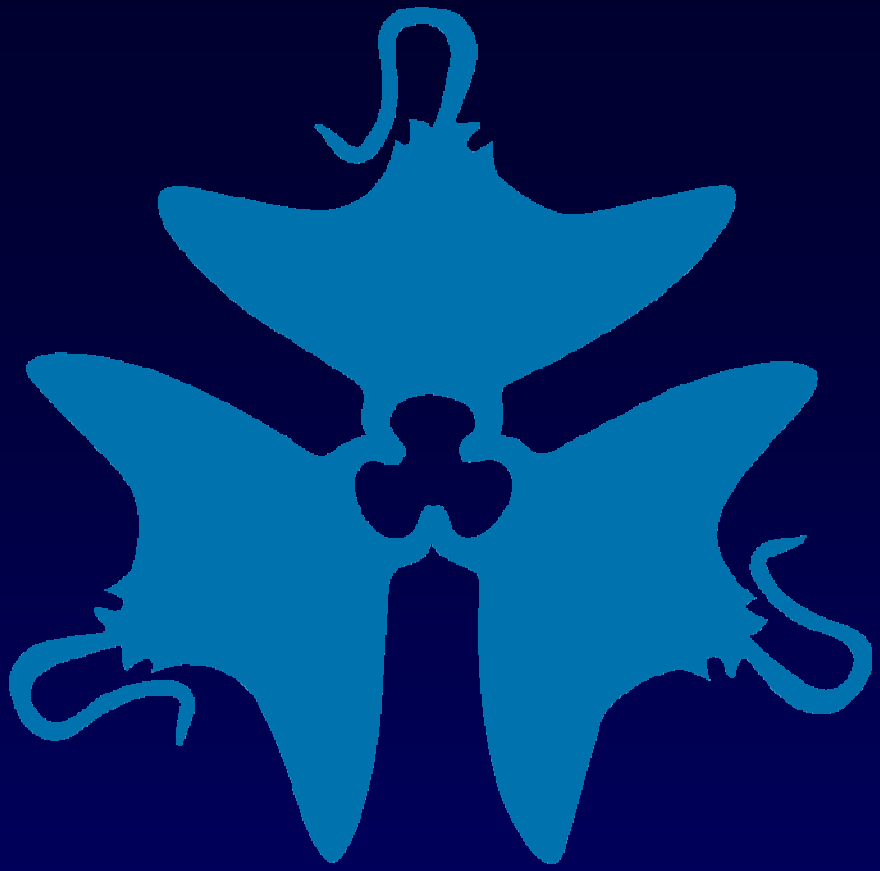
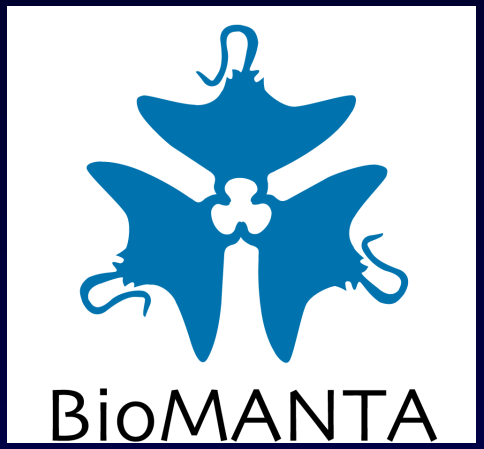


BioMANTA

The Modelling and Analysis of Biological Network Activity

Melissa Davis and Andrew Newman
(ACB, IMB, ITEE, UQ)



Project Background

- Systems biology holds promise to understand complexity of biological networks
- Computational modelling and analysis of large-scale protein-protein interaction (PPI) and compound activity networks
- Knowledge representation using SW standards Resource Description Framework (RDF) and Web Ontology Language (OWL) enables machine inference and facilitates knowledge discovery
- Current PPI networks have only sparse coverage over the actual interactome -> knowledge discovery using machine learning and network meta-analysis to rank interactions and infer global networks

BioMANTA project overview

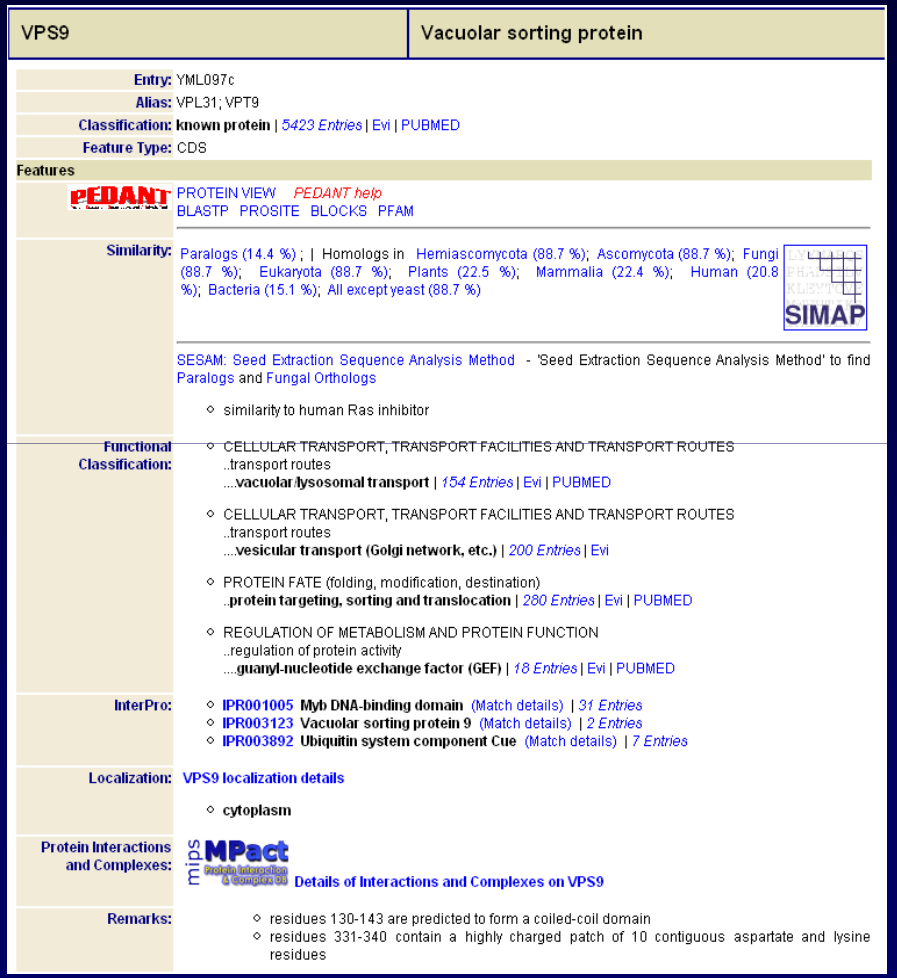
- Development of Semantic Interactome Model and Semantic Web infrastructure
  - Knowledge representation using Ontology instanced with public interaction data
  - RDF triple storage, inferencing and querying
- Network Inference and Knowledge Discovery
  - Network meta-analysis
  - Global network inference
- Output, Visualisation of Results and Software
  - COBALT visualisation software
  - Data: High quality data sets and ontologies

ACB Programs

- Phenotype-informed discovery of networks and systems
  - Biomanta network analysis to uncover interactions between (disease state) phenotypes and biological networks
- Modelling dynamic cellular processes
  - Integration of time course expression data
- Algorithms for graphs and networks
  - Avoiding sub-graph isomorphism
  - Object co-identification based on attribute matching

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

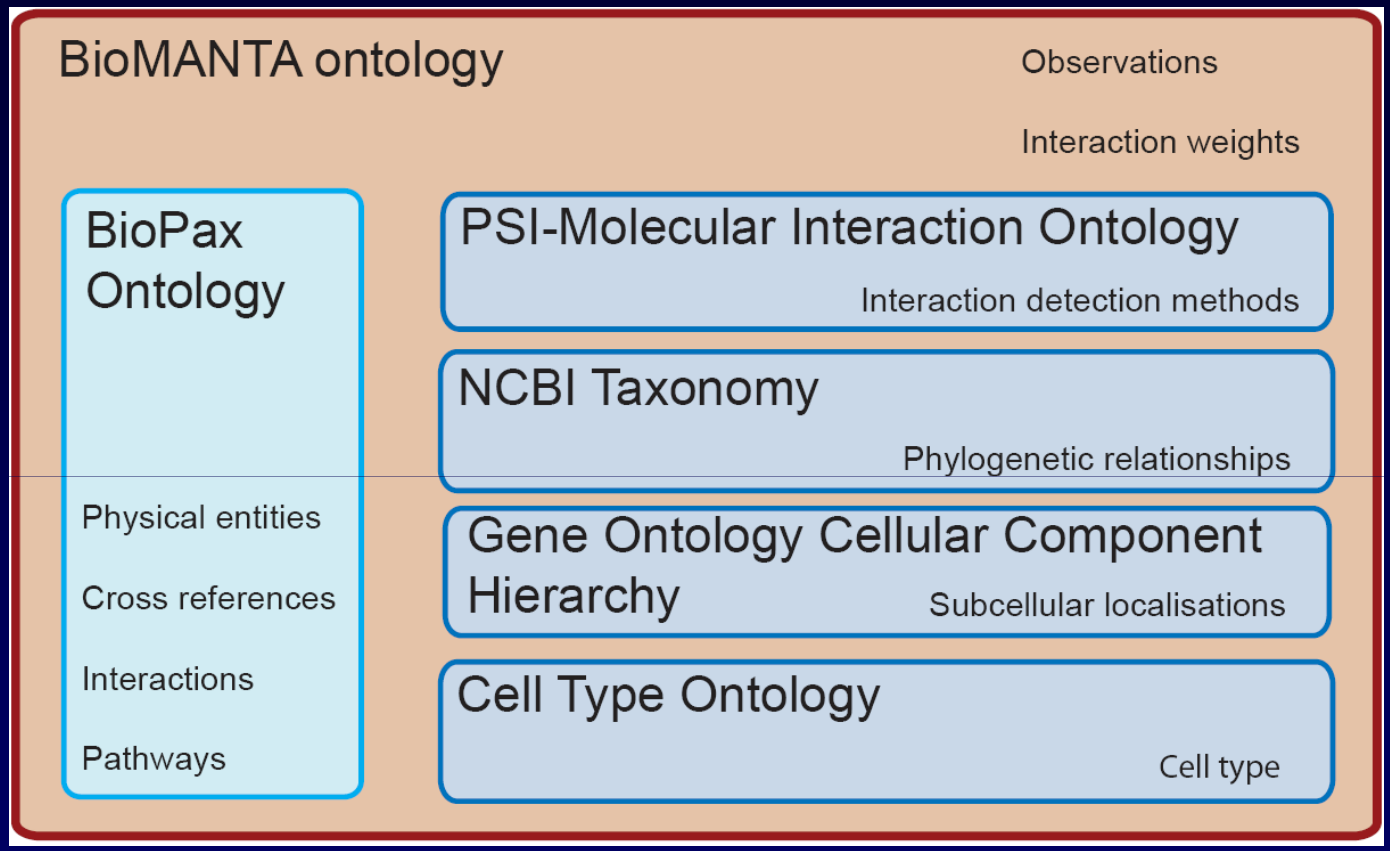
Remarks:

- 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

Knowledge Representation

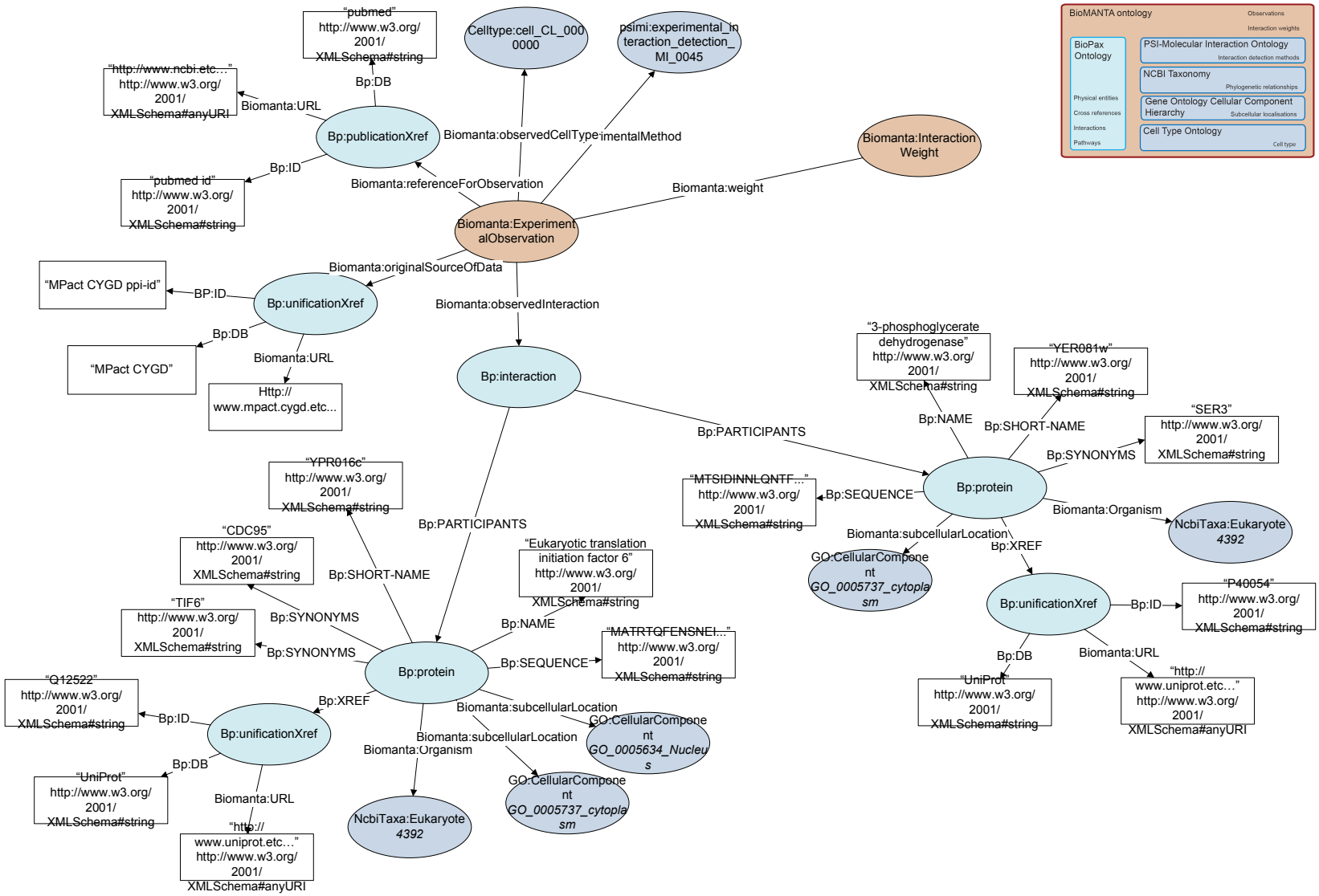
- Structured presentation of concepts, relationships and data
  - Relational Database (schema, tables, identifiers, etc…)
  - Ontology (Classes, properties/attributes, relationships)
  - Machine interpretation of representation (XML)
- Inference errors frequently caused by errors in the representation
  - Incorrectly modelled domain knowledge
  - Missing assumptions
- Currently, the majority of PPI data are stored in DB available online

BioMANTA ontology



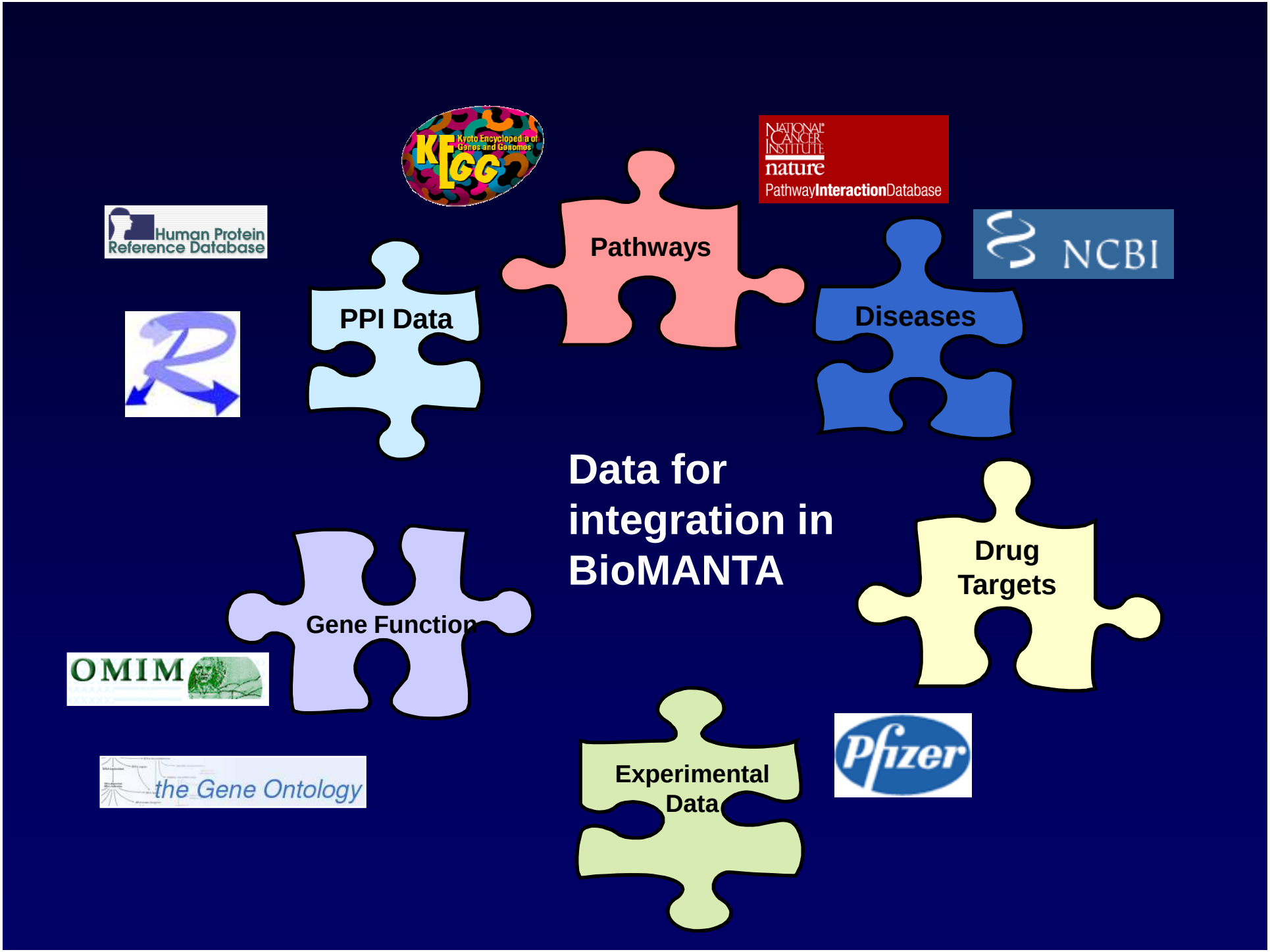
- OWL based ontology with imported modules from relevant ontologies
- Limit creation of new classes in BioMANTA ontology and use classes from existing ontologies wherever possible

Semantic Interactome Model



Integrating Biological Data

- Huge variety in project sizes.
- Very large data sets (TBs and PBs).
- Computationally intensive.
- Different specialities.
- Different levels of semantics in technologies used.
- Mostly suspect, duplicated, inapplicable, poorly and incorrectly modelled.


KEGG
Kyoto Encyclopedia of Genes and Genomes
NATIONAL CANCER INSTITUTE
nature
PathwayInteractionDatabase
Human Protein Reference Database
NCBI
Pathways
PPI Data
Diseases
Data for integration in BioMANTA
Drug Targets
Gene Function
OMIM
the Gene Ontology
Experimental Data
Pfizer

Too Many Names

- Global IDs (such as LSID, BioPAX) have largely failed to gain acceptance for a variety of reasons.
- We didn’t want to come up with another ID.
- A huge number of local IDs including:
  - MPact, DIP, IntACT, MINT.
- Properties include:
  - Database,
  - Sequence information,
  - Species,
  - Subcellular location,
  - Expression, cross references, etc.

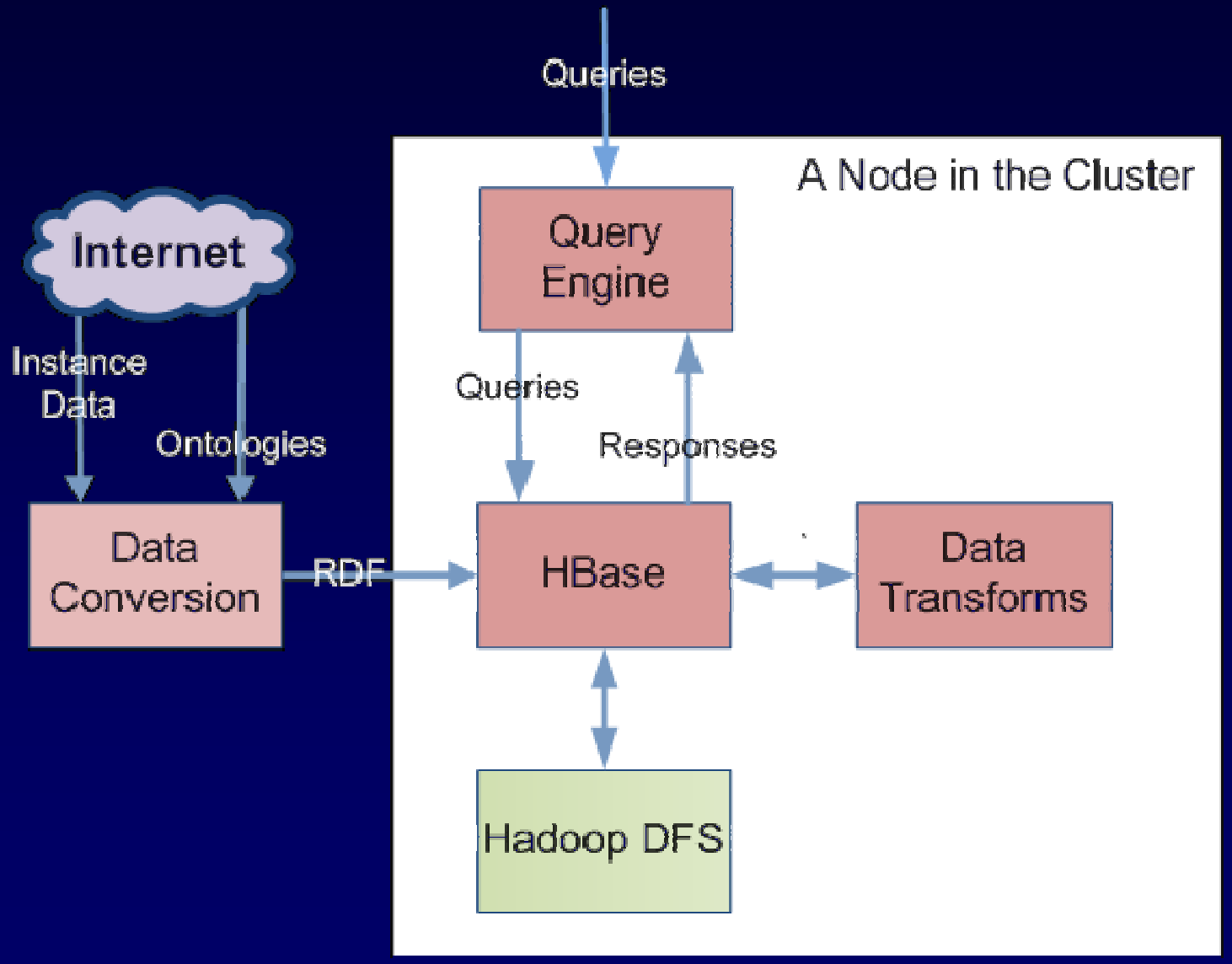
Semantic Web Technologies

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

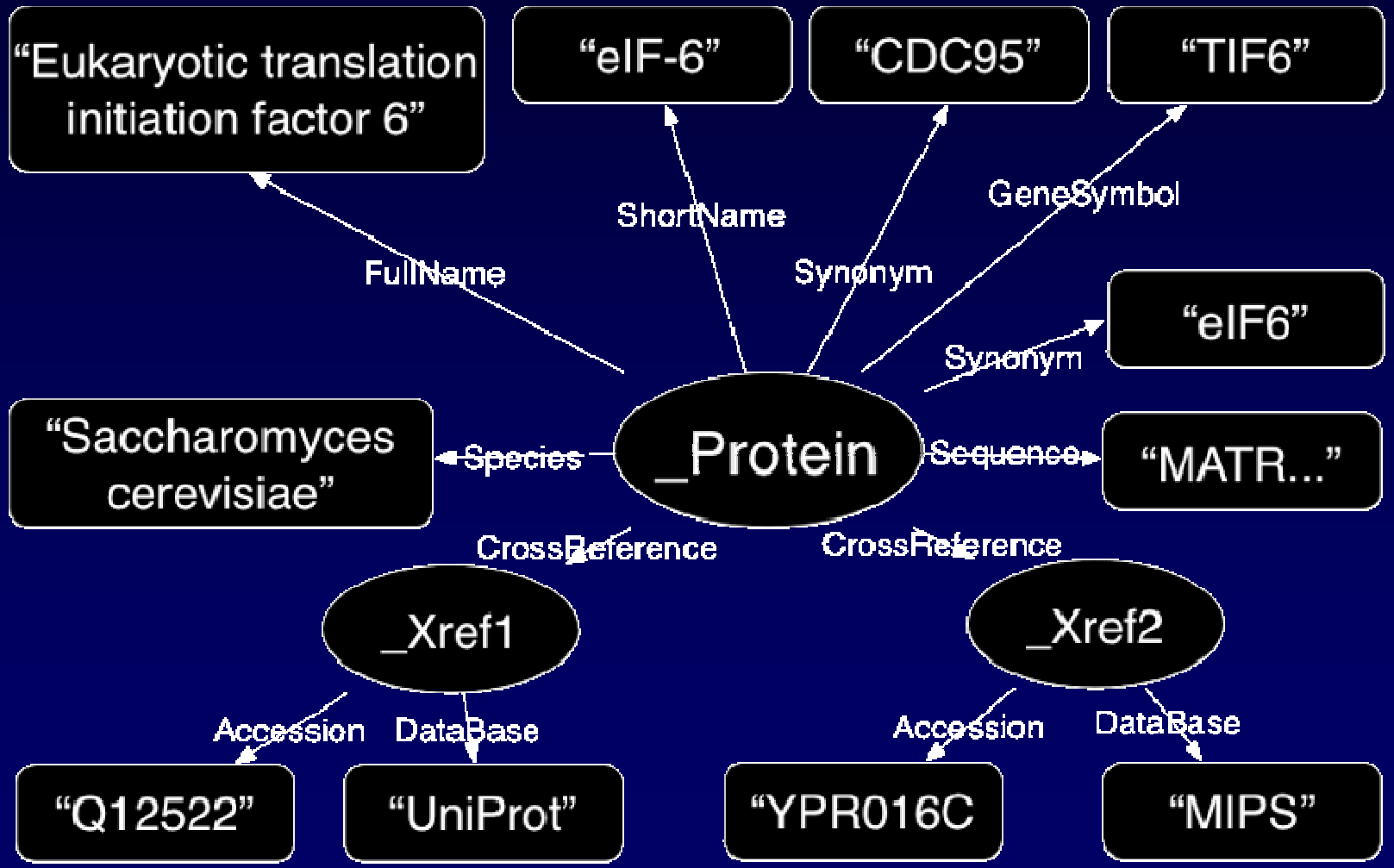
    <subject> <predicate> <object>

    <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

Architecture



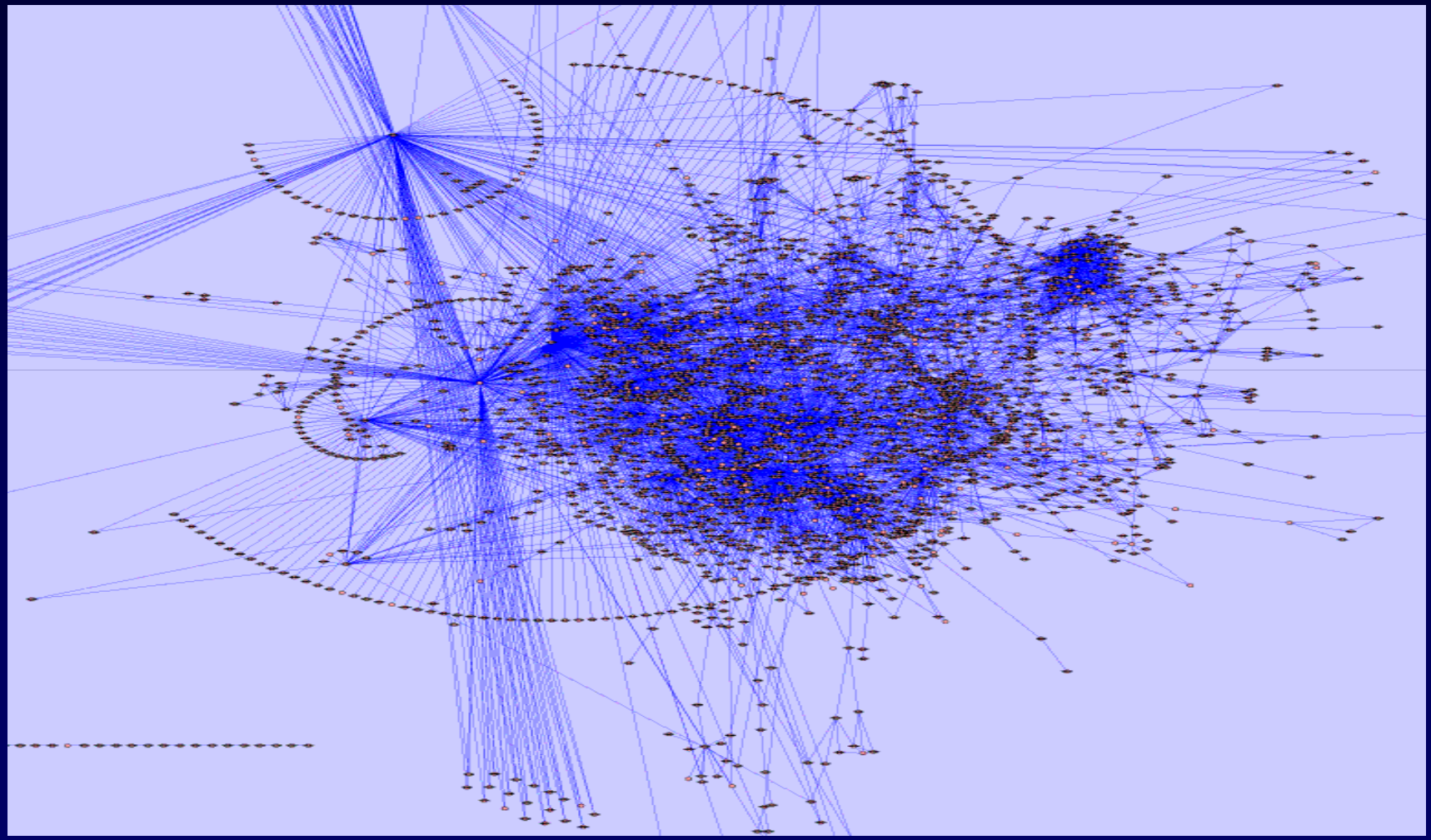
Sample Merged Data

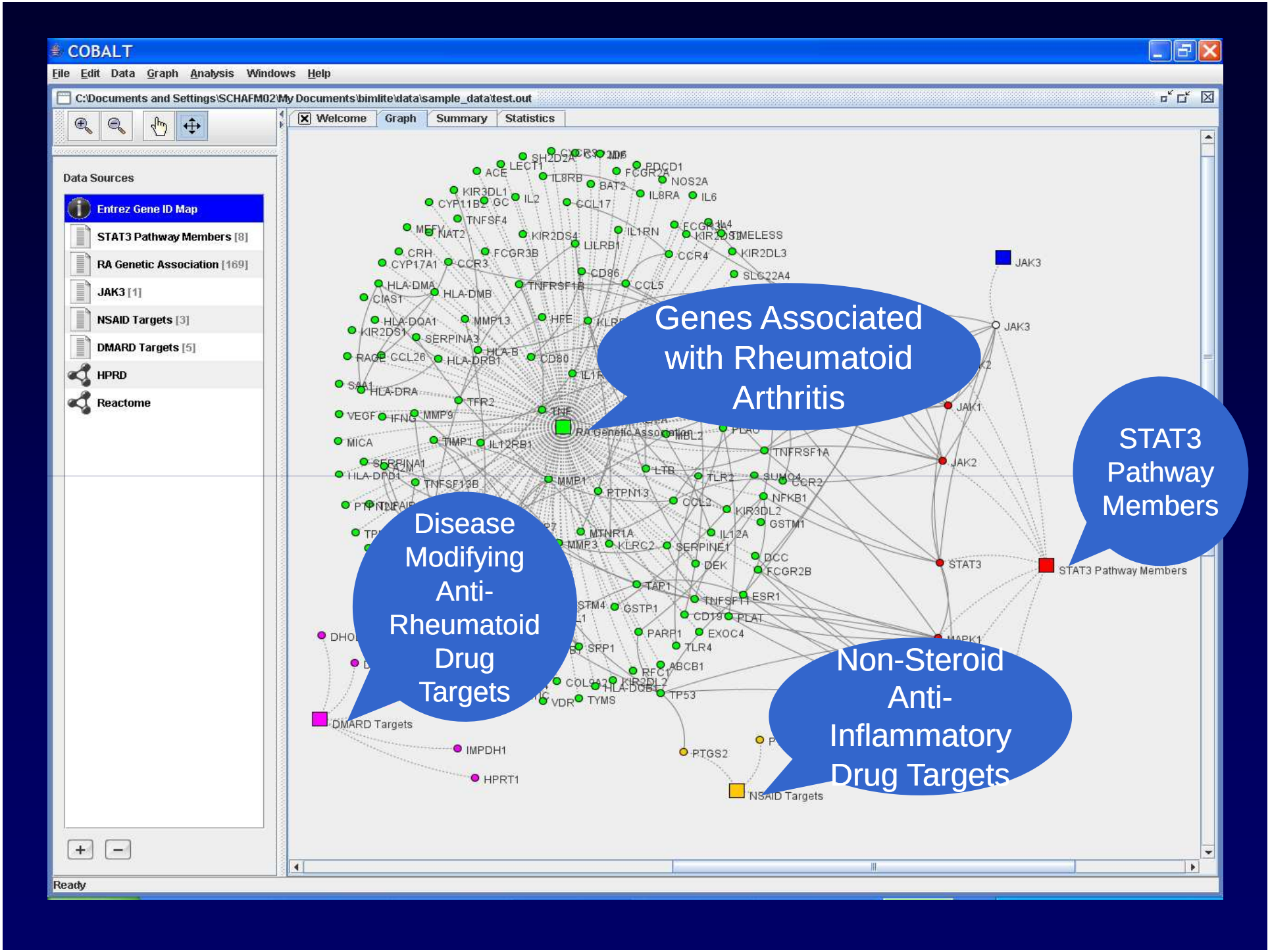


Output and Visualisation

- Improved data sets:
  - Removing data redundancy,
  - Better query results,
  - Fewer false positives,
  - Matrix of protein-protein interactions.
- Published and peer-reviewed methodologies
- Network visualisation tools

“Hairball” Problem




COBALT
File Edit Data Graph Analysis Windows Help
C:\Documents and Settings\SCHAFM02\My Documents\bimlite\data\sample_data\test.out
Welcome
Graph
Summary
Statistics
Data Sources
Entrez Gene ID Map
STAT3 Pathway Members [8]
RA Genetic Association [169]
JAK3 [1]
NSAID Targets [3]
DMARD Targets [5]
HPRD
Reactome
Genes Associated with Rheumatoid Arthritis
STAT3 Pathway Members
Disease Modifying Anti-Rheumatoid Drug Targets
Non-Steroid Anti-Inflammatory Drug Targets
JAK3
JAK1
JAK2
STAT3
STAT3 Pathway Members
DMARD Targets
NSAID Targets
IMPDH1
HPRT1
PTGS2
TNF
RA Genetic Association
Ready

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
Melissa Davis
Kevin Burrage
Shoaib Sehgal

IMB & ARC Centre of Excellence in Bioinformatics