Integrating gene and pathway information about nicotine dependence

Presented By: Satya S. Sahoo

Mentor: Dr. Olivier Bodenreider

Lister Hill Center, NLM/NIH

8/30/2007

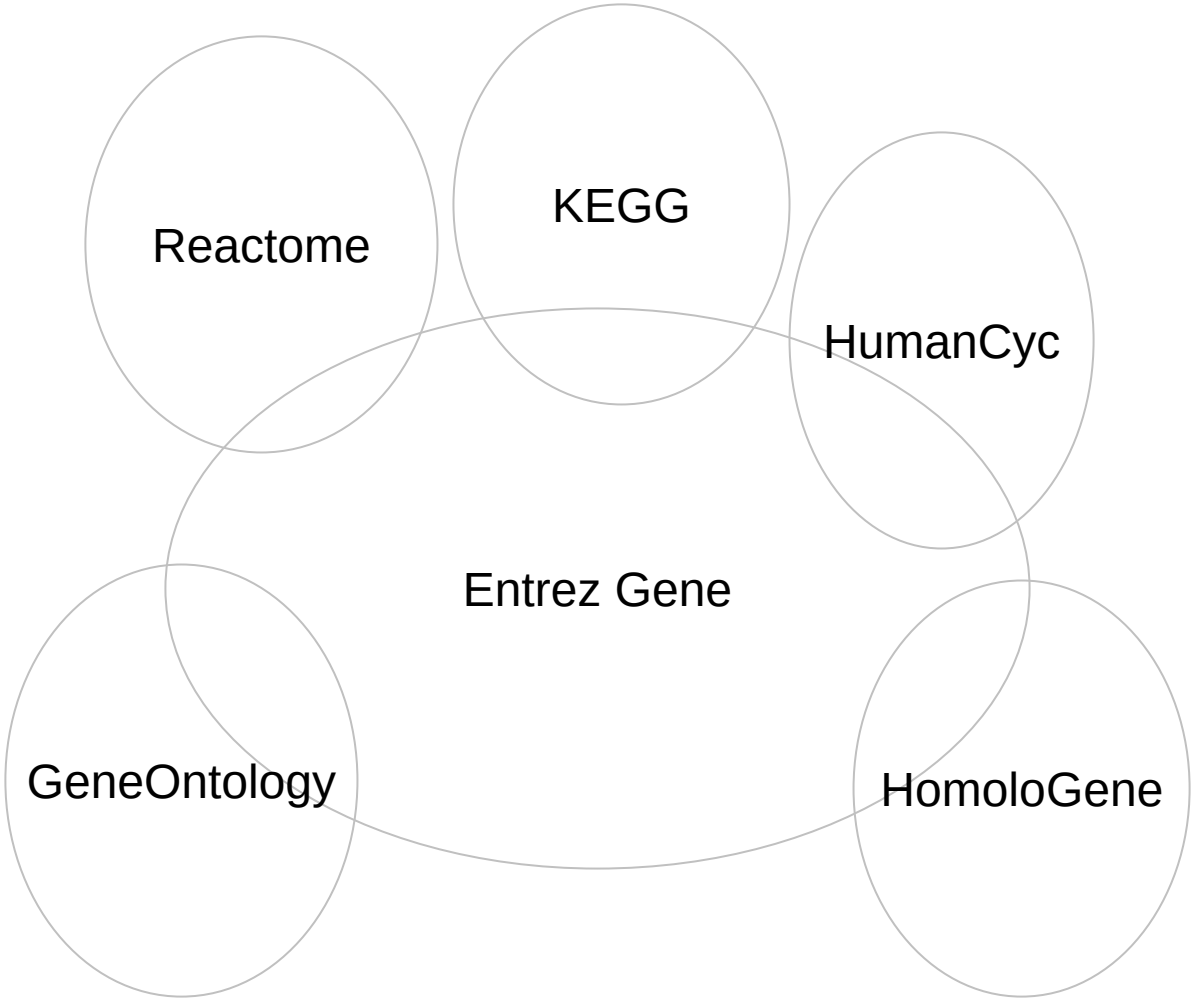
Outline

- Motivation
- Previous Work
- Schema-level Integration
- Questions – Answers
- Deductive Reasoning
- Implementation
- Limitations
- Future Work

Motivation

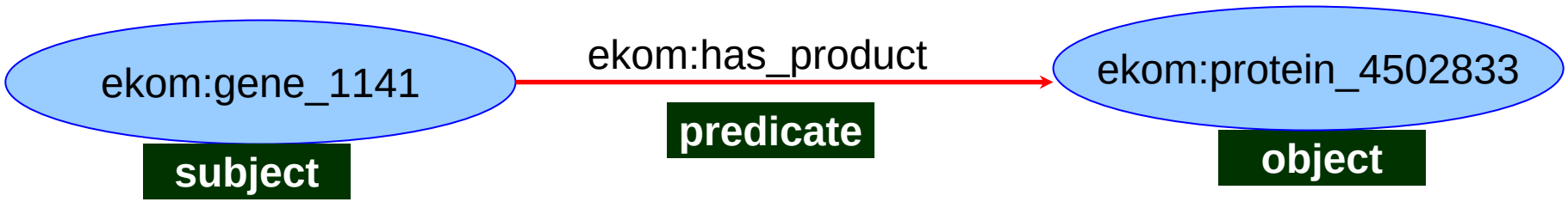
- NIDA study on nicotine dependency
- List of candidate genes in humans
- Analysis objectives include:
  - Find interactions between genes
  - Identification of active genes – maximum number of pathways
  - Identification of genes based on anatomical locations
- Requires integration of genome and biological pathway information

Genome and pathway information integration



Previous Work

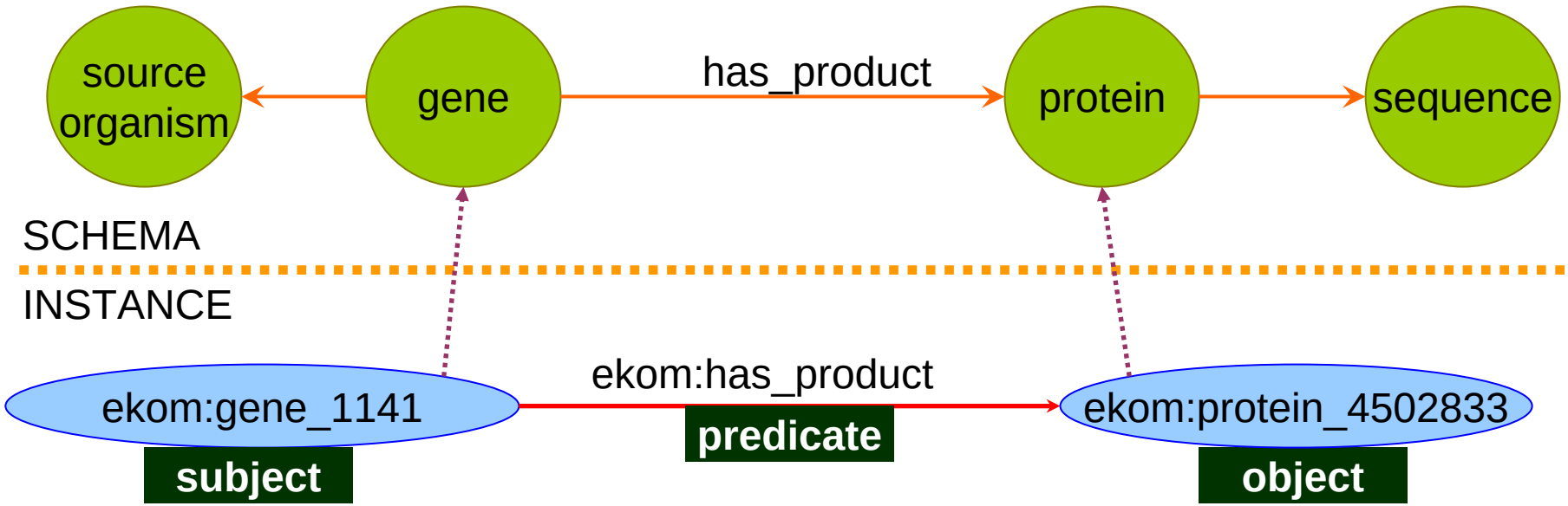
- Converted genome data from NCBI Entrez Gene to RDF (2006)
- Explicit use of relationships between concepts
- All information represented as a ‘triple’

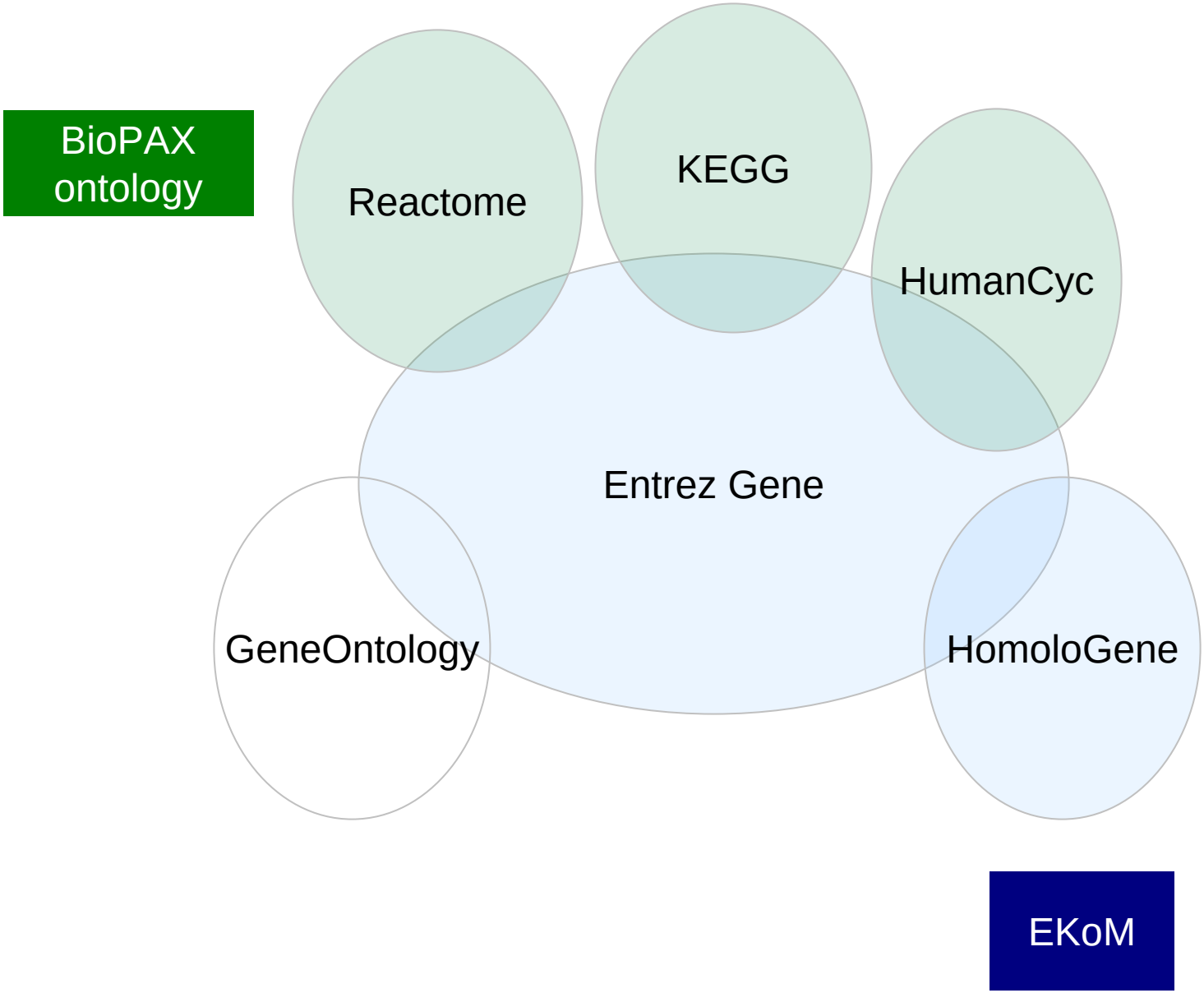


Namespace – ekom = http://ekom.ncbi.nlm.nih.gov/ekom.owl

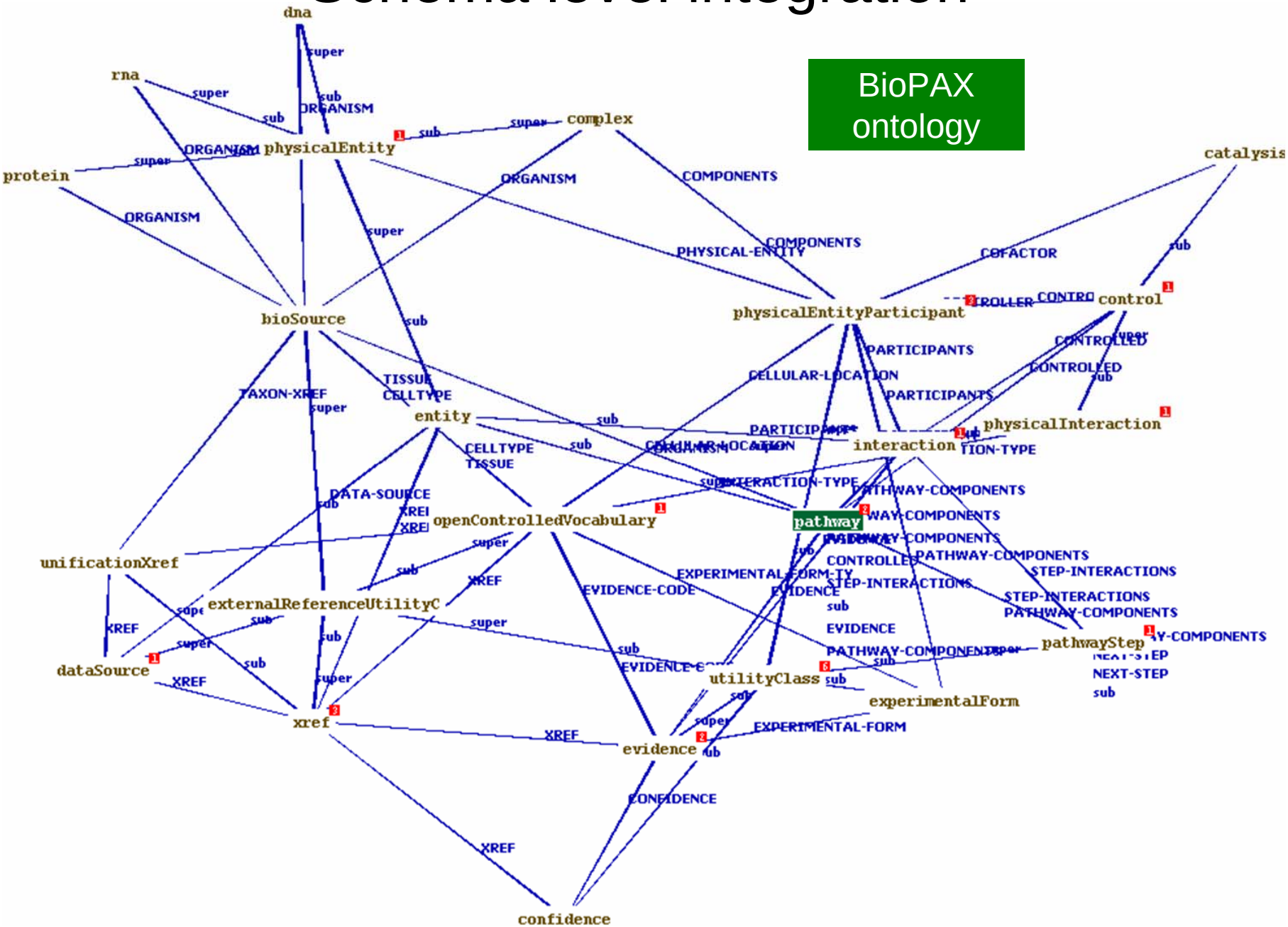
Motivation

- When new data sources are added – need to specify integration approach for each value
- Need for schema or knowledge model based approach
- Integration defined at schema level – inherited by instance values

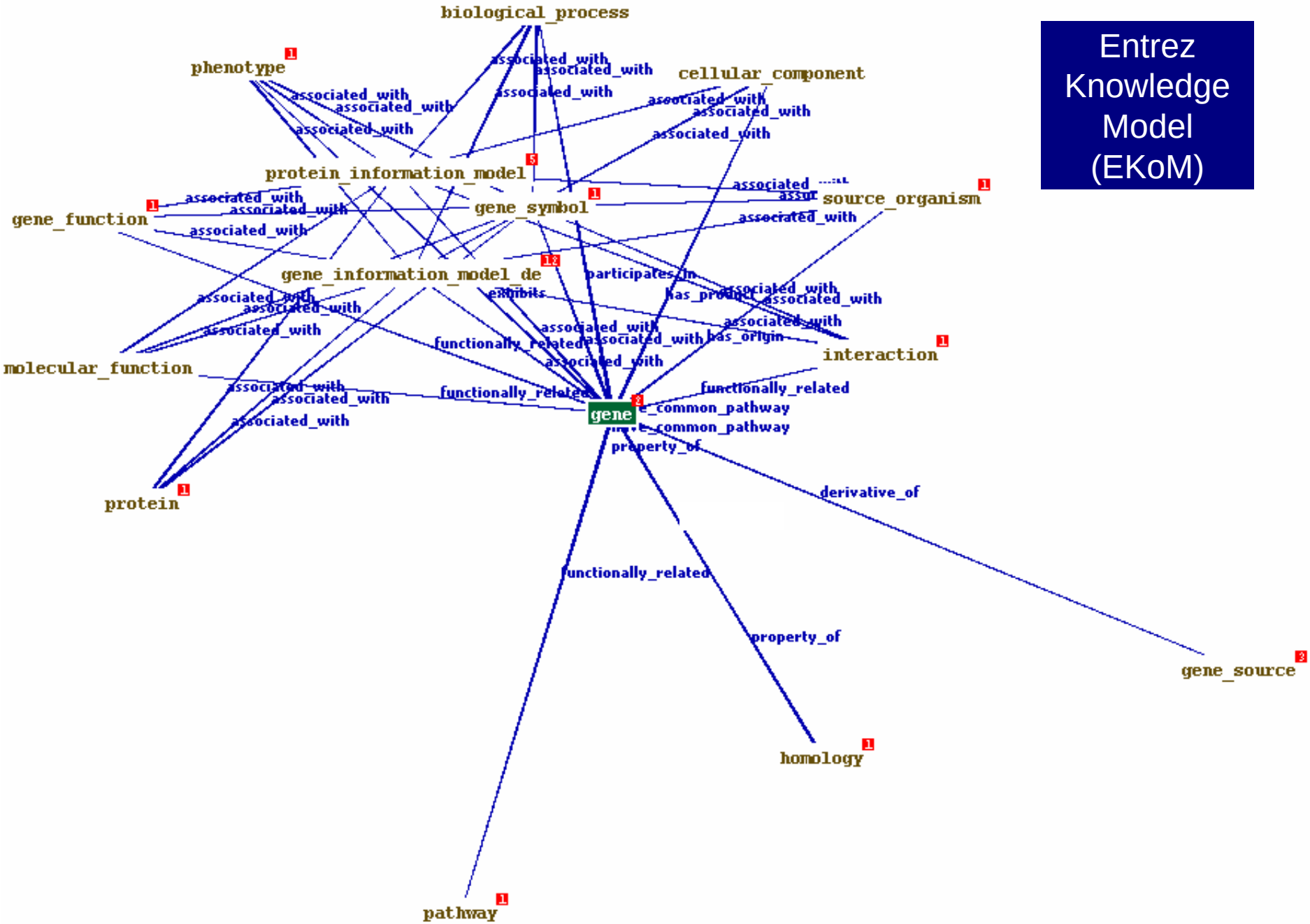


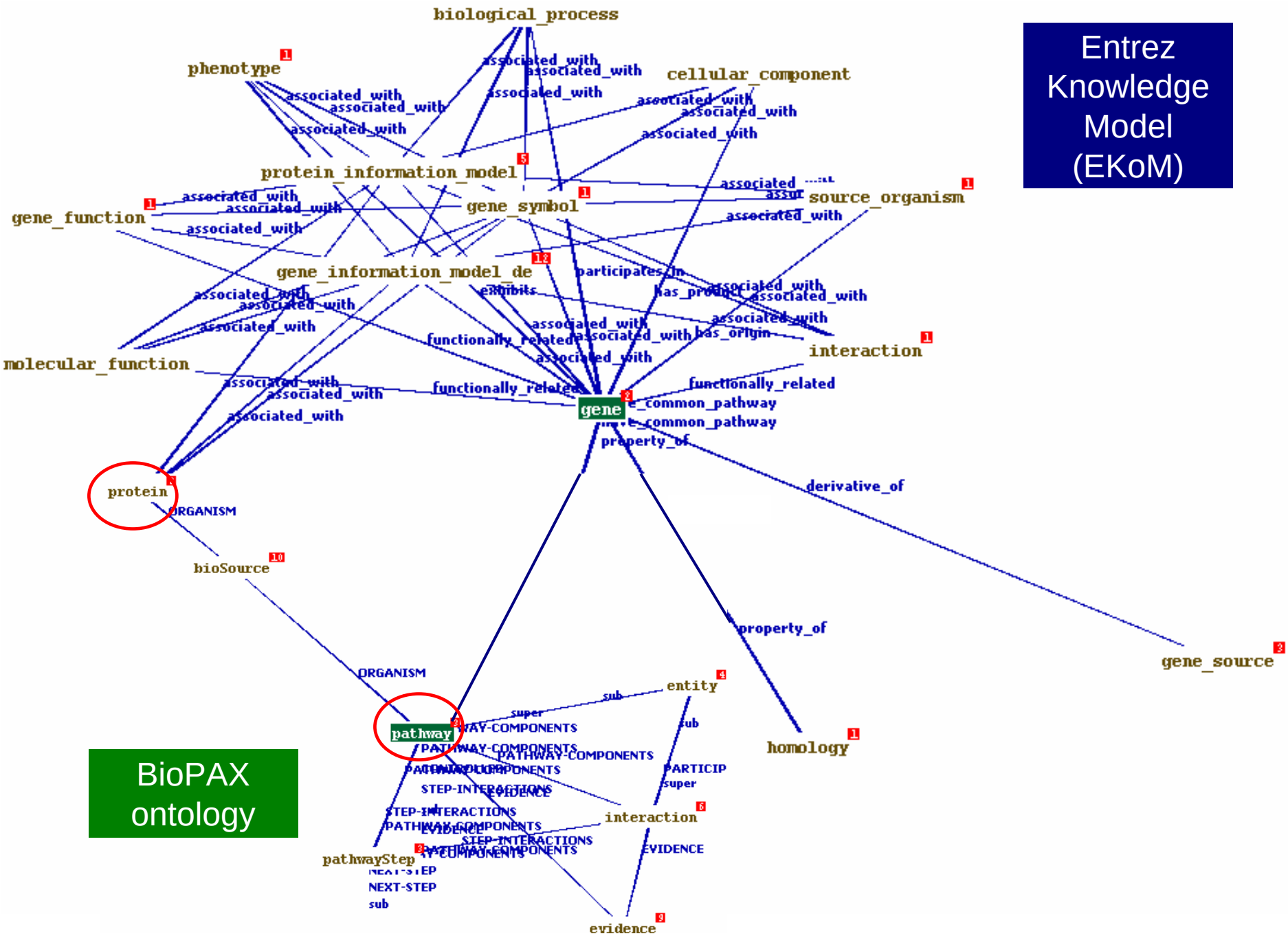

BioPAX ontology
Reactome
KEGG
HumanCyc
Entrez Gene
GeneOntology
HomoloGene
EKoM

Schema level integration



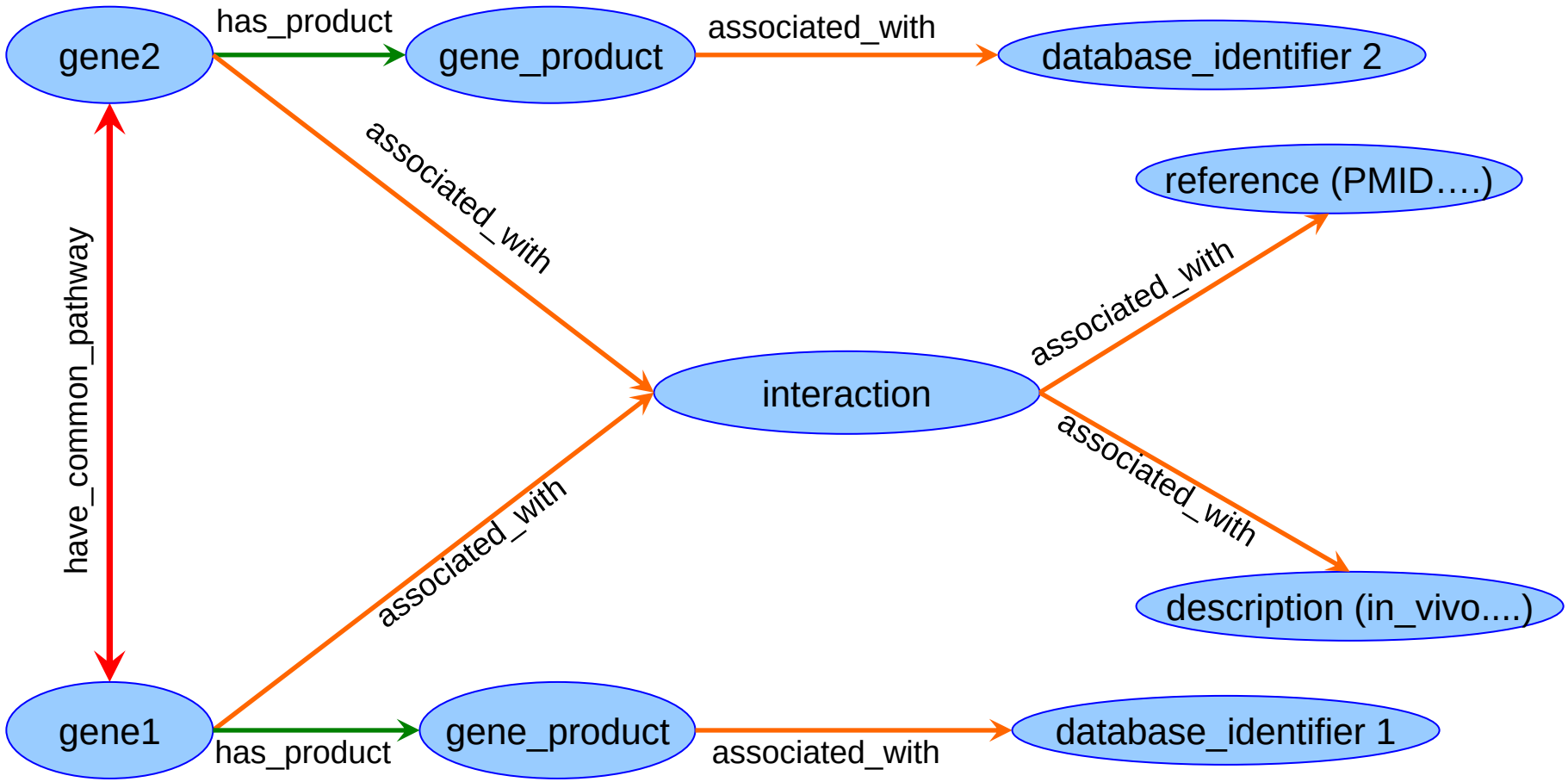
Entrez Knowledge Model (EKoM)




Entrez Knowledge Model (EKoM)
BioPAX ontology
biological_process
phenotype
cellular_component
protein_information_model
gene_symbol
source_organism
gene_function
gene_information_model_de
molecular_function
interaction
gene
protein
ORGANISM
bioSource
derivative_of
property_of
gene_source
ORGANISM
pathway
entity
homology
interaction
pathwayStep
evidence
associated_with
functionally_related
participates_in
exhibits
has_product
has_origin
PARTICIP
EVIDENCE
STEP-INTERACTIONS
PATHWAY-COMPONENTS
NEXT-STEP
sub
super

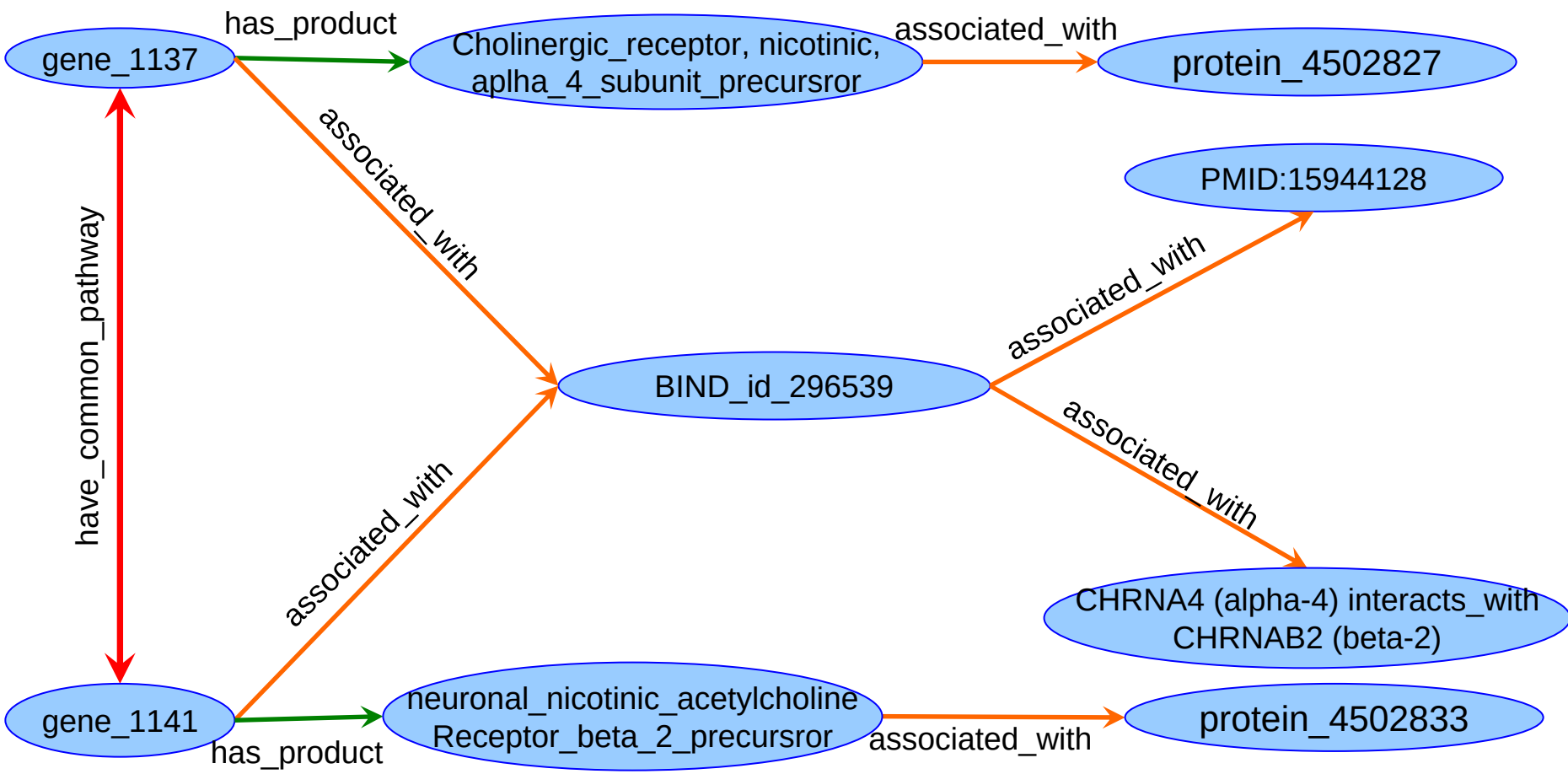
Questions-Answers I

- What genes or protein products of the gene set have been shown to interact or bind with each other?



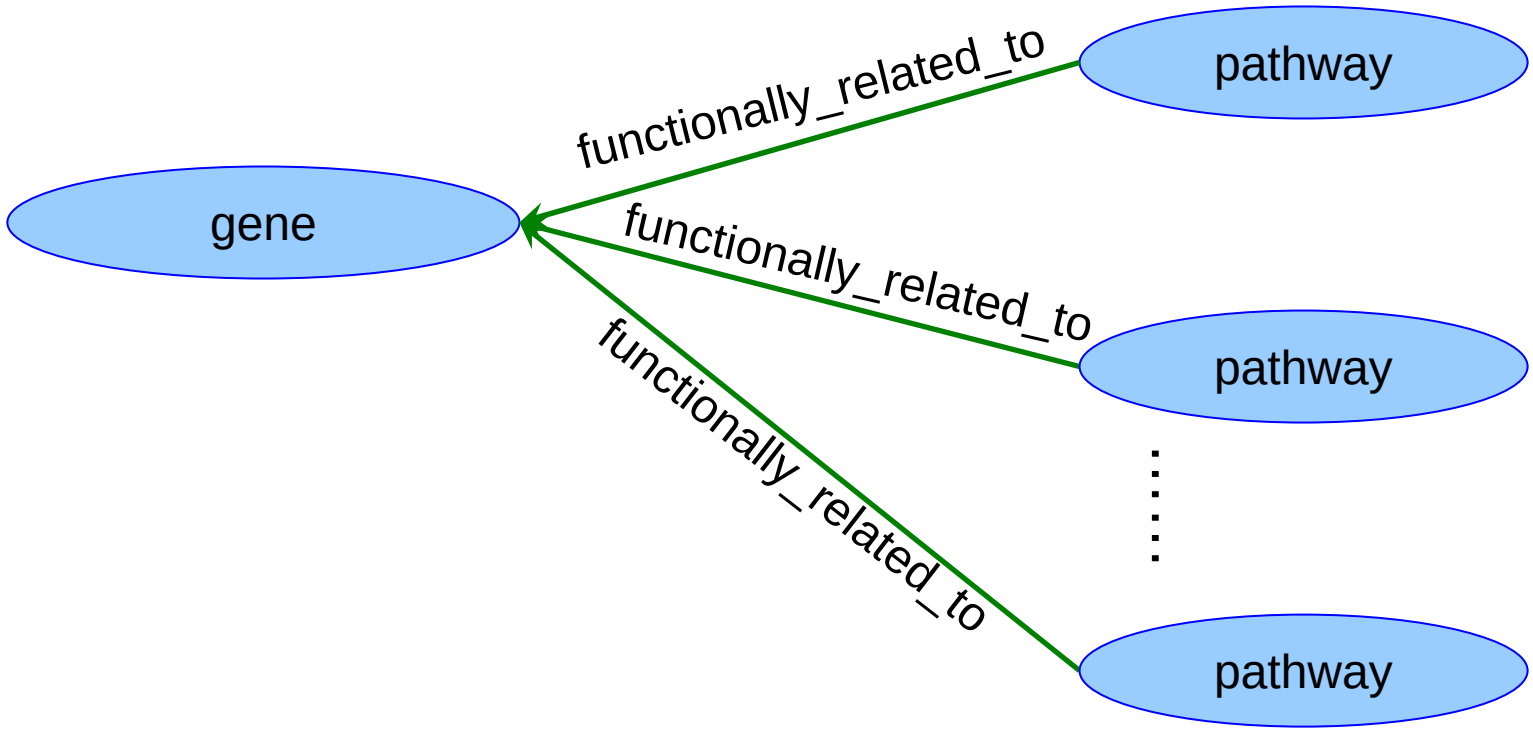
Questions-Answers I

- What genes or protein products of the gene set have been shown to interact or bind with each other?



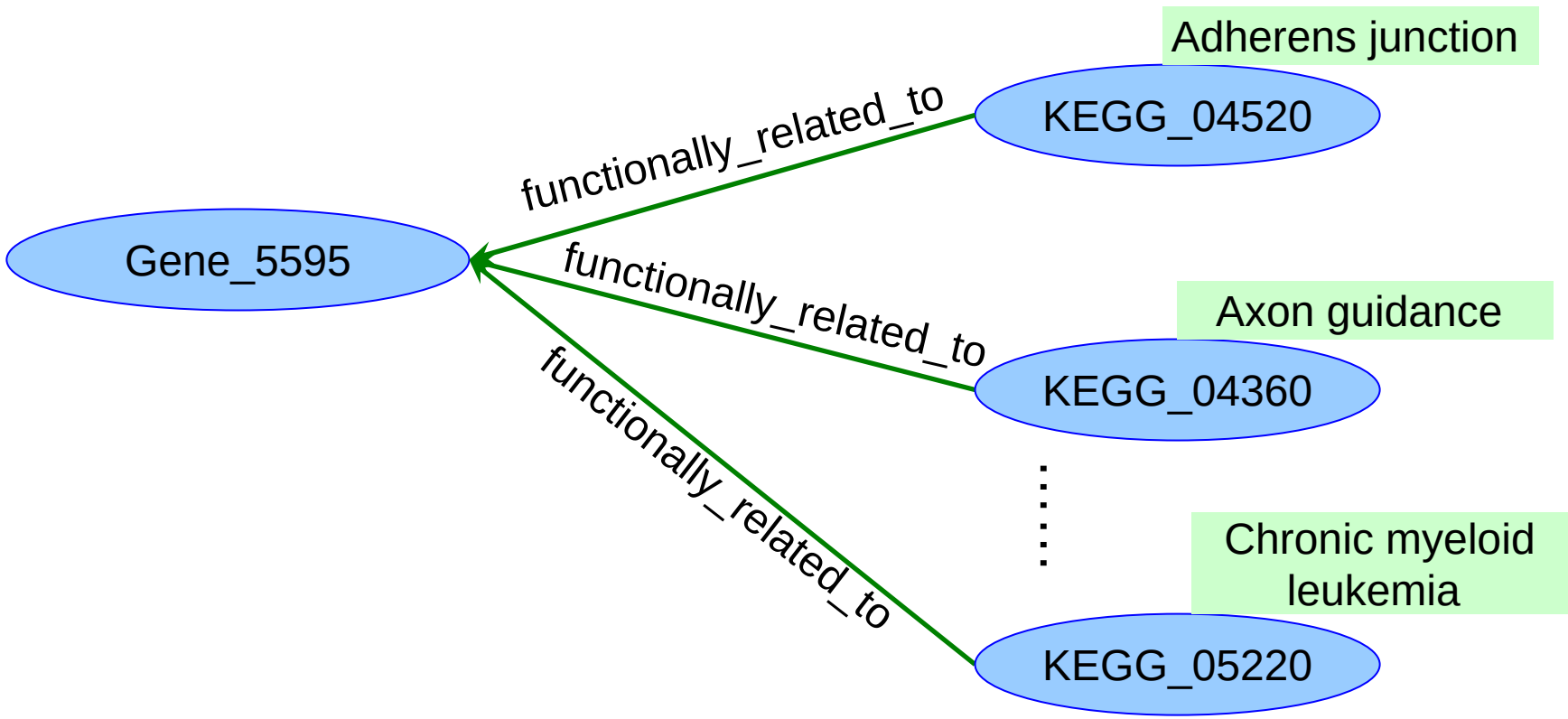
Questions-Answers II

- Which genes are most active through participation in greater number of pathways?



Questions-Answers II

- Which genes are most active through participation in greater number of pathways?

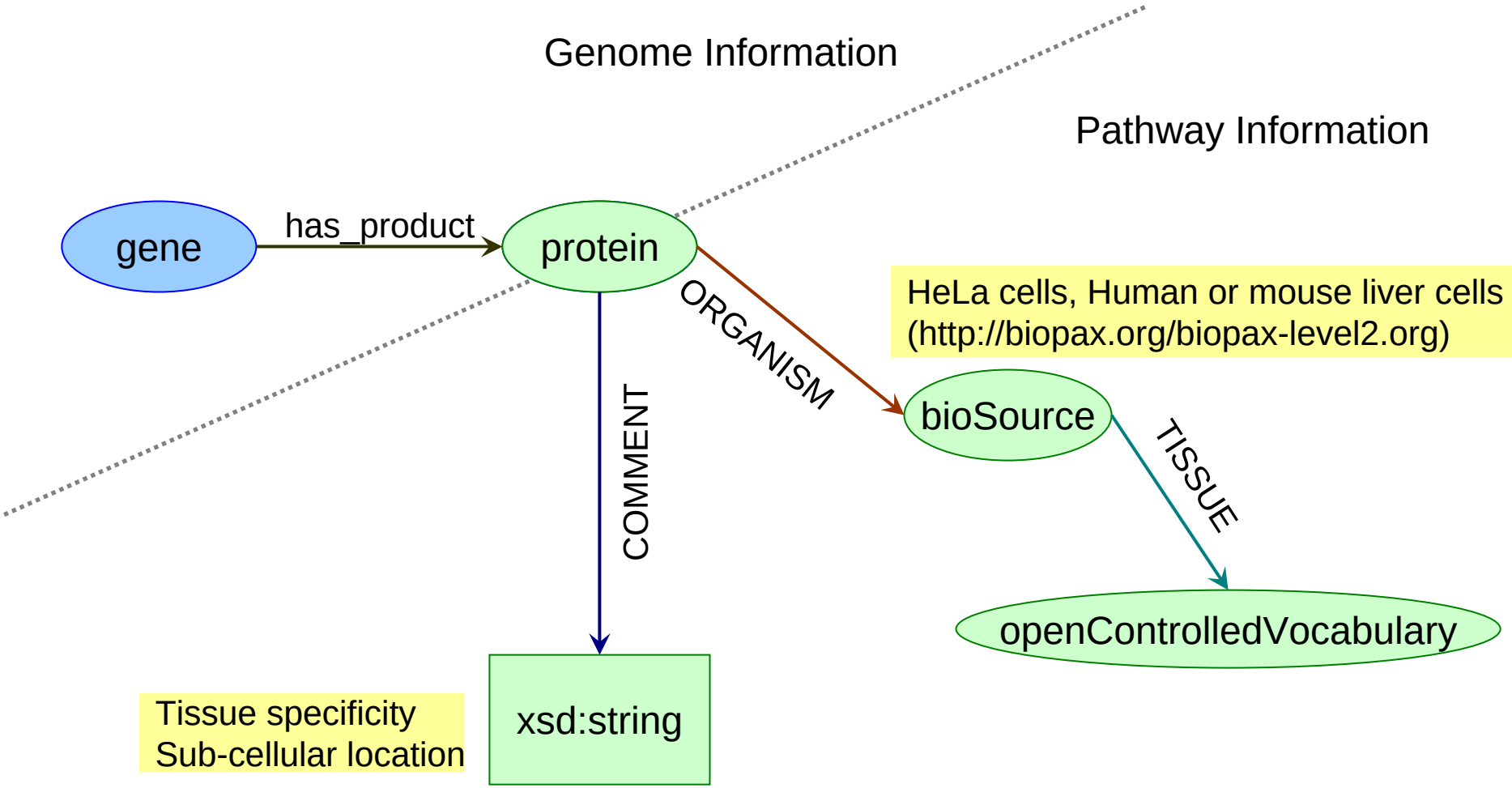


Questions-Answers II

- Top three list of most active genes (in terms of participation in biological pathways):
  - Gene_5595: MAPK3 mitogen-activated protein kinase 3 (Homo sapiens) [30 pathways]
  - Gene_5594: MAPK1 mitogen-activated protein kinase 1 (Homo sapiens) [30 pathways]
  - Gene_5604: MAP2K1 mitogen-activated protein kinase kinase 1 (Homo sapiens) [24 pathways]
- Similarly, we can find pathways in which maximum number of genes participate

Traversal across genome - pathway information

- Which genes are known to exist in liver or liver tissues?

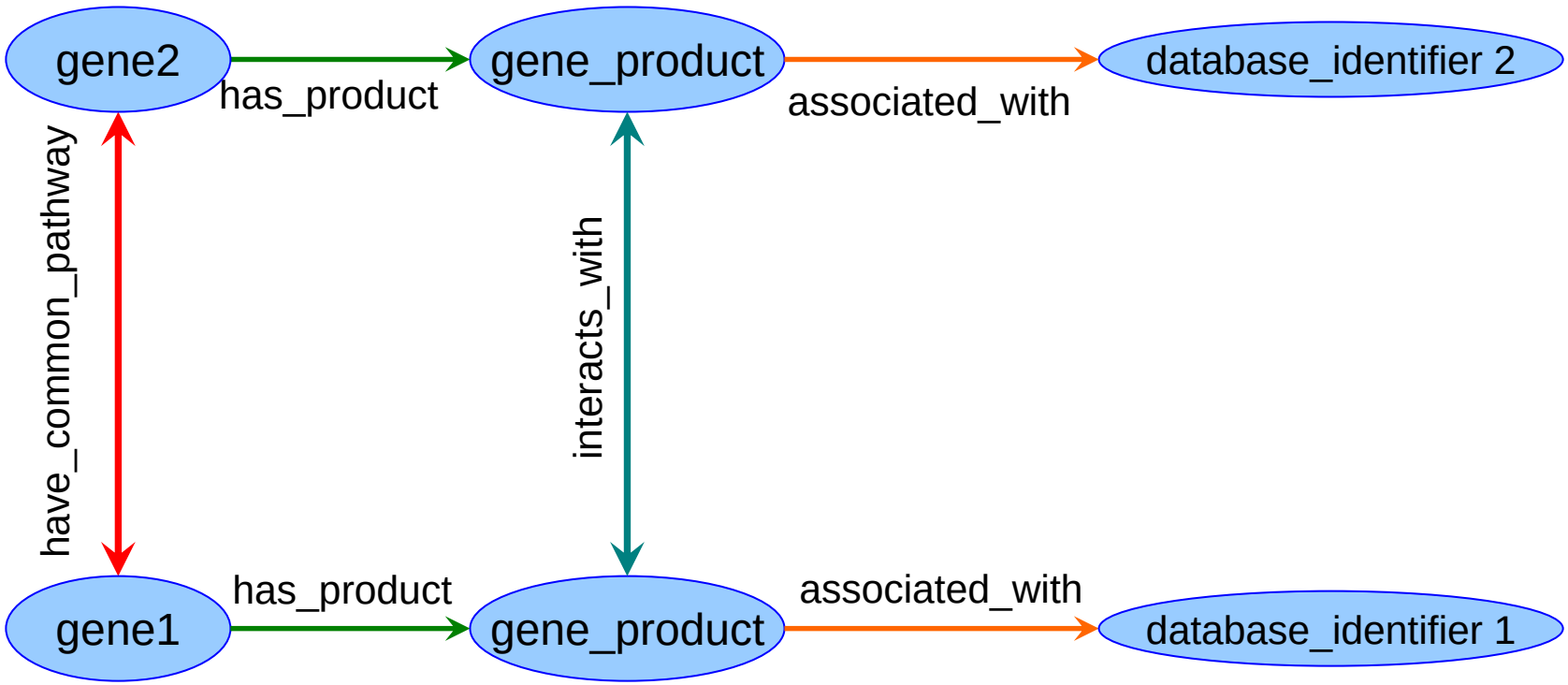


Deductive Reasoning

Protein-Protein Interaction

RULE

IF (x have_common_pathway y) AND (x rdf:type gene) AND (y rdf:type gene) AND (x has_product m) AND (y has_product n) AND (m rdf:type gene_product) AND (n rdf:type gene_product) THEN (m ? n)



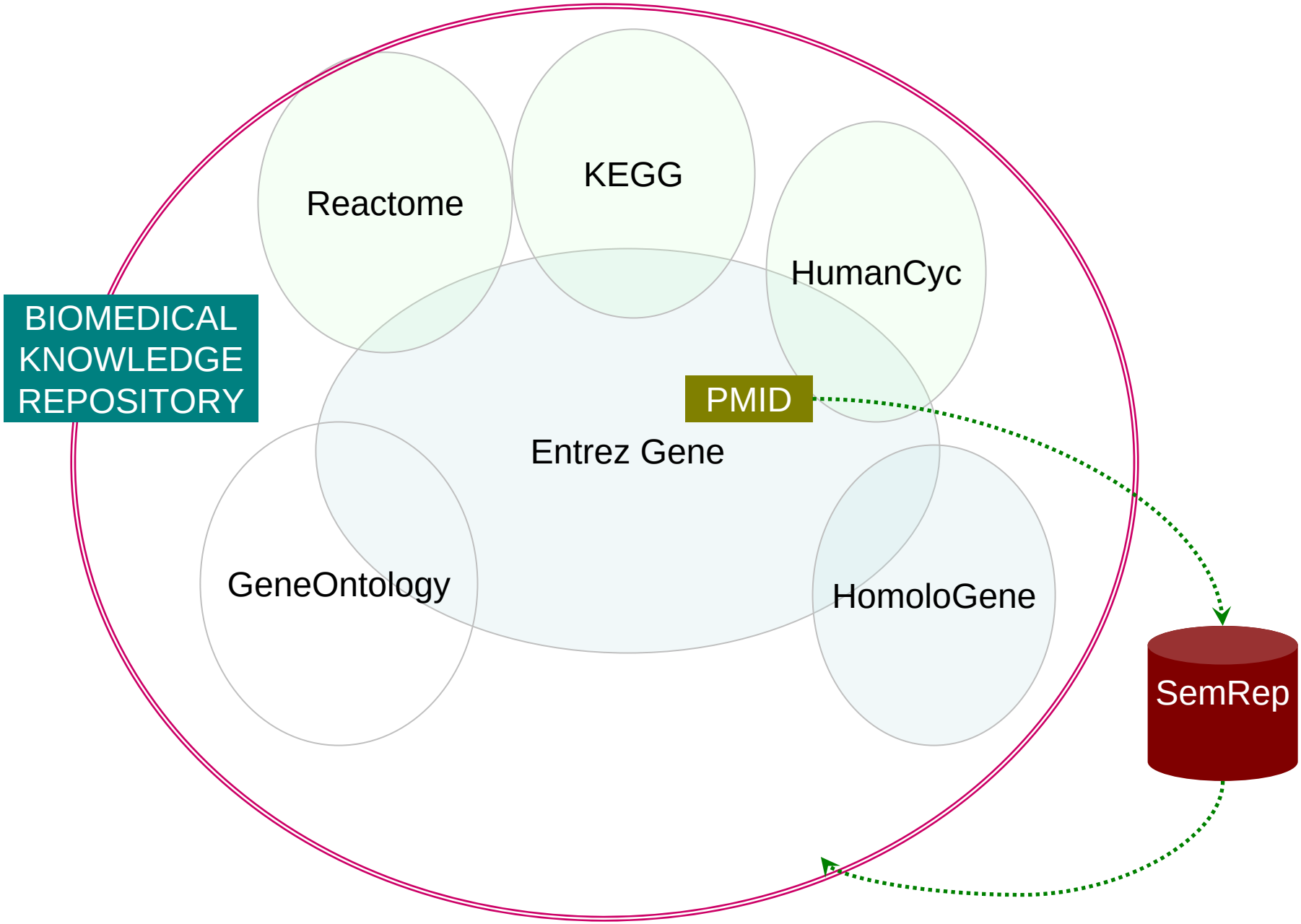
Implementation

- Used eUtils to retrieve EG records in XML format - manual curation of some genes
- To retrieve homologous genes for the human genes
  - Parsed human gene records and retrieved HomoloGene identifier
  - Used eUtils to query HomoloGene and retrieve homology record in XML format
  - Parsed HomoloGene records to retrieve EG identifiers for four model organisms
  - Again using eUtils, retrieved EG records with these EG identifiers
- Used XPath in XSLT style sheet to convert the EG XML data to conform to the EKoM schema – ontology population
- Loaded the genome and pathway OWL files into Oracle 10g RDF store
- Used graph based language called SPARQL to query this knowledge base

Limitations

- Need to formalize parameters before ‘interaction’ between gene products is asserted
- Execute and compare query results against whole EG dataset
- At present, queries do not exploit all the details that may be gained from pathways
- Need to systematically validate query results with NIDA researchers

Future Work


BIOMEDICAL KNOWLEDGE REPOSITORY
Reactome
KEGG
HumanCyc
PMID
Entrez Gene
GeneOntology
HomoloGene
SemRep

Special Thanks

- Lee Peters
- May Cheh
- Thomas C. Rindflesch
- Halil Kilicoglu
- John Nyugen