4-11-2008

RDB2RDF: Incorporating Domain Semantics in Structured Data

Satya S. Sahoo
Wright State University - Main Campus

kno.e.sis

COLLECTING THE DOTS | CONNECTING THE DOTS

RDB2RDF: Incorporating Domain Semantics in Structured Data

Satya S. Sahoo

Kno.e.sis Center, Computer Science and Engineering Department,

Wright State University, Dayton, OH, USA

Acknowledgements

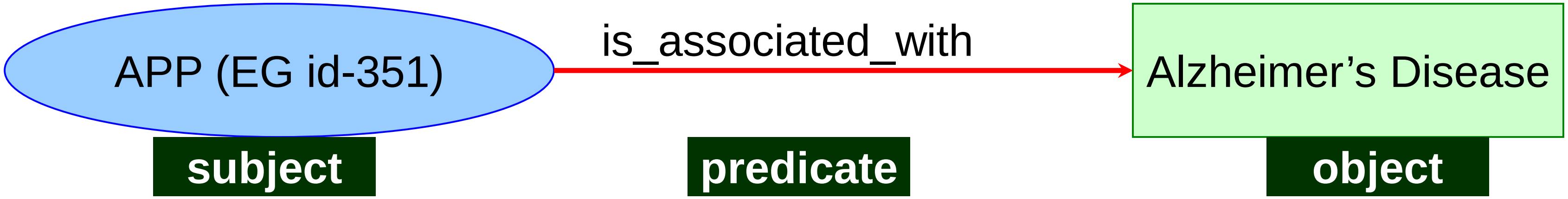
- Dr. Olivier Bodenreider (U.S NLM, NIH)
- Dr. Amit Sheth (Kno.e.sis Center, Wright State University)
- Dr. Joni L. Rutter (NIDA, NIH)
- Dr. Karen J. Skinner (NIDA, NIH)
- Lee Peters (U.S NLM, NIH)
- Kelly Zeng (U.S NLM, NIH)

Outline

- RDB to RDF – Objectives
- Method I: RDB to RDF without ontology
- Application I: Genome ↔ Phenotype
- Method II: RDB to RDF with ontology
- Application II: Genome ↔ Biological Pathway integration
- Conclusion

Objectives of Modeling Data in RDF

- RDF data model



- RDF enables modeling of logical relationship between entities
- Relations are at the heart of Semantic Web*
- RDF data - Logical Structure of the information
- Reasoning over RDF data → knowledge discovery

*Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships,

Relationship Web: Blazing Semantic Trails between Web Resources

Outline

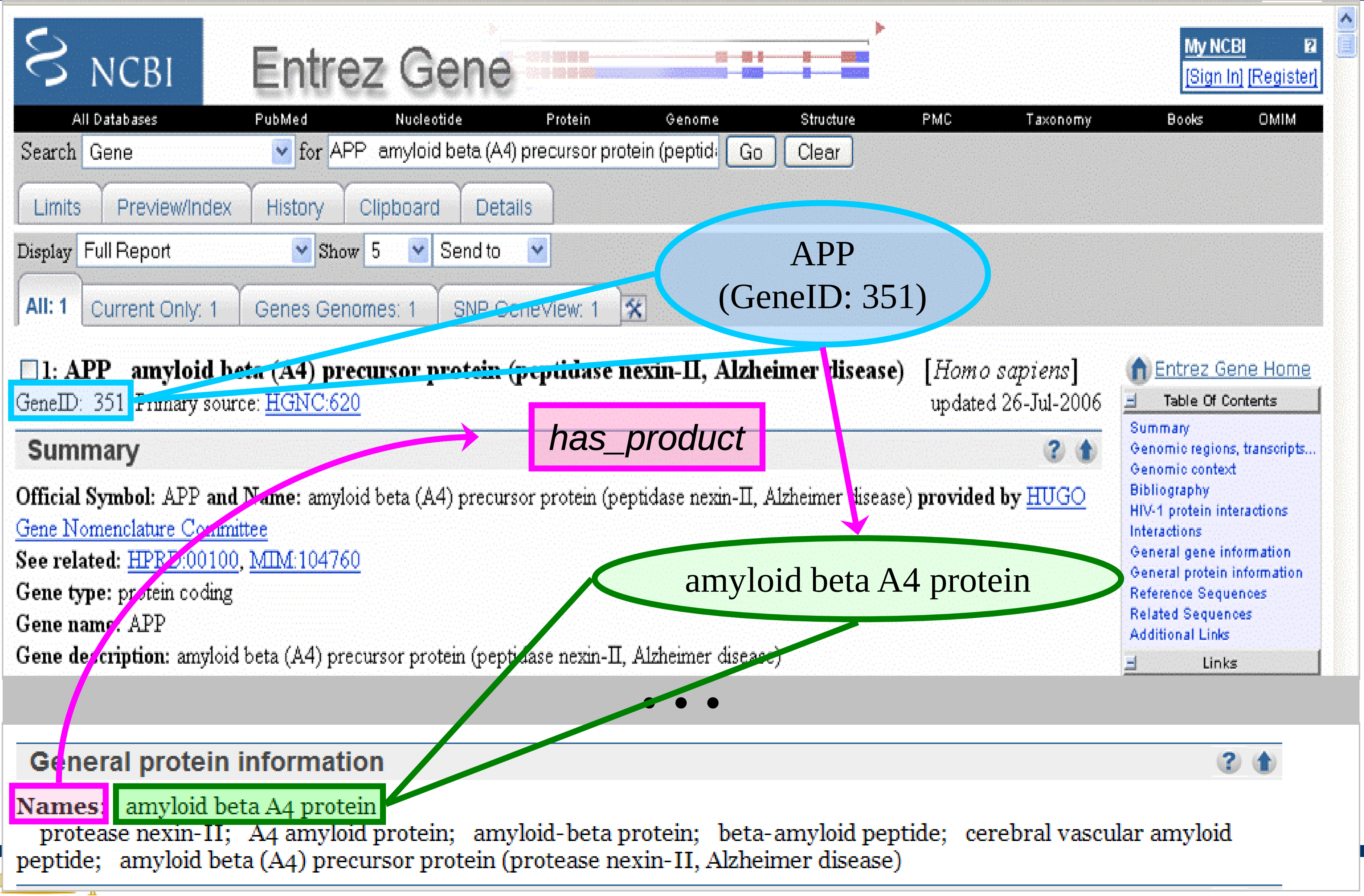
- RDB to RDF – Objectives
- Method I: RDB to RDF without ontology
- Application I: Genome ↔ Phenotype
- Method II: RDB to RDF with ontology
- Application II: Genome ↔ Biological Pathway integration
- Conclusion

Data: NCBI Entrez Gene

- NCBI Entrez Gene: gene related information from sequenced genomes and model organisms*
  - 2 million gene records
  - Gene information for genomic maps, sequences, homology, and protein expression
  - Available in XML, ASN.1 and as a Webpage

*http://www.ncbi.nlm.nih.gov/sites/entrez/

Entrez Gene Web Interface



Search Gene for APP amyloid beta (A4) precursor protein (peptid

Display Full Report Show 5 Send to

APP
(GeneID: 351)

□1: APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) [*Homo sapiens*]

GeneID: 351 Primary source: HGNC:620 updated 26-Jul-2006

has_product

Summary

Official Symbol: APP **and Name:** amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) **provided by** HUGO Gene Nomenclature Committee

See related: HPRD:00100, MIM:104760

Gene type: protein coding

Gene name: APP

Gene description: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)

amyloid beta A4 protein

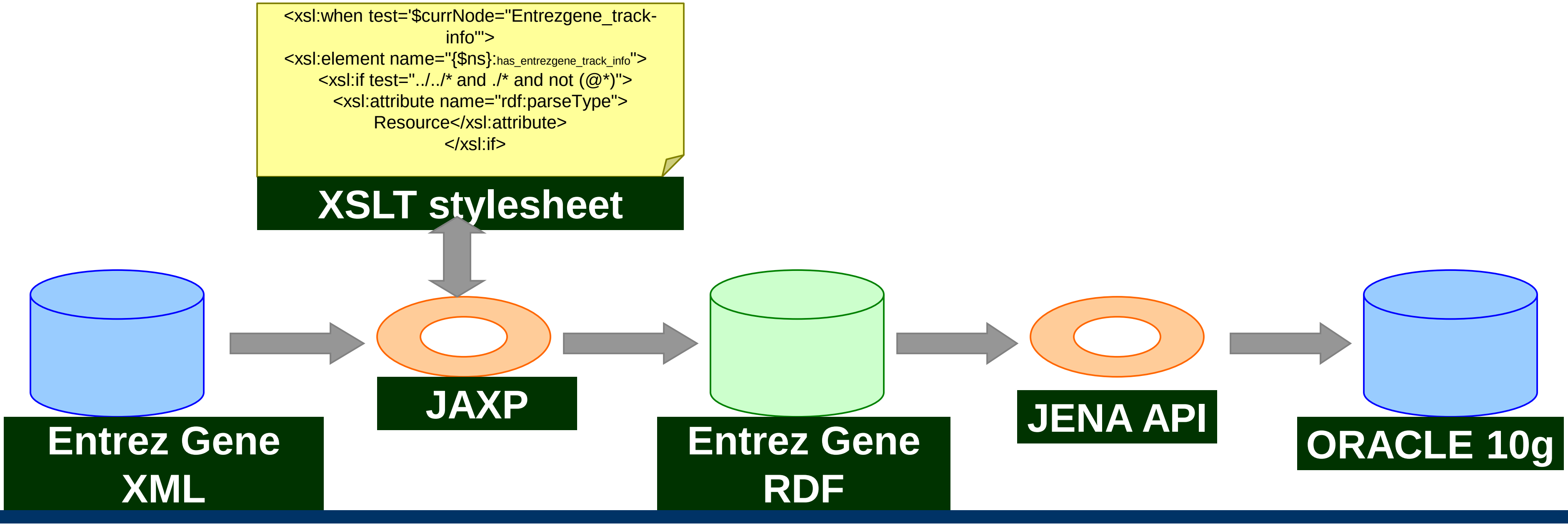
• • •

General protein information

Names: amyloid beta A4 protein
protease nexin-II; A4 amyloid protein; amyloid-beta protein; beta-amyloid peptide; cerebral vascular amyloid peptide; amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)

Method I: RDB to RDF without ontology

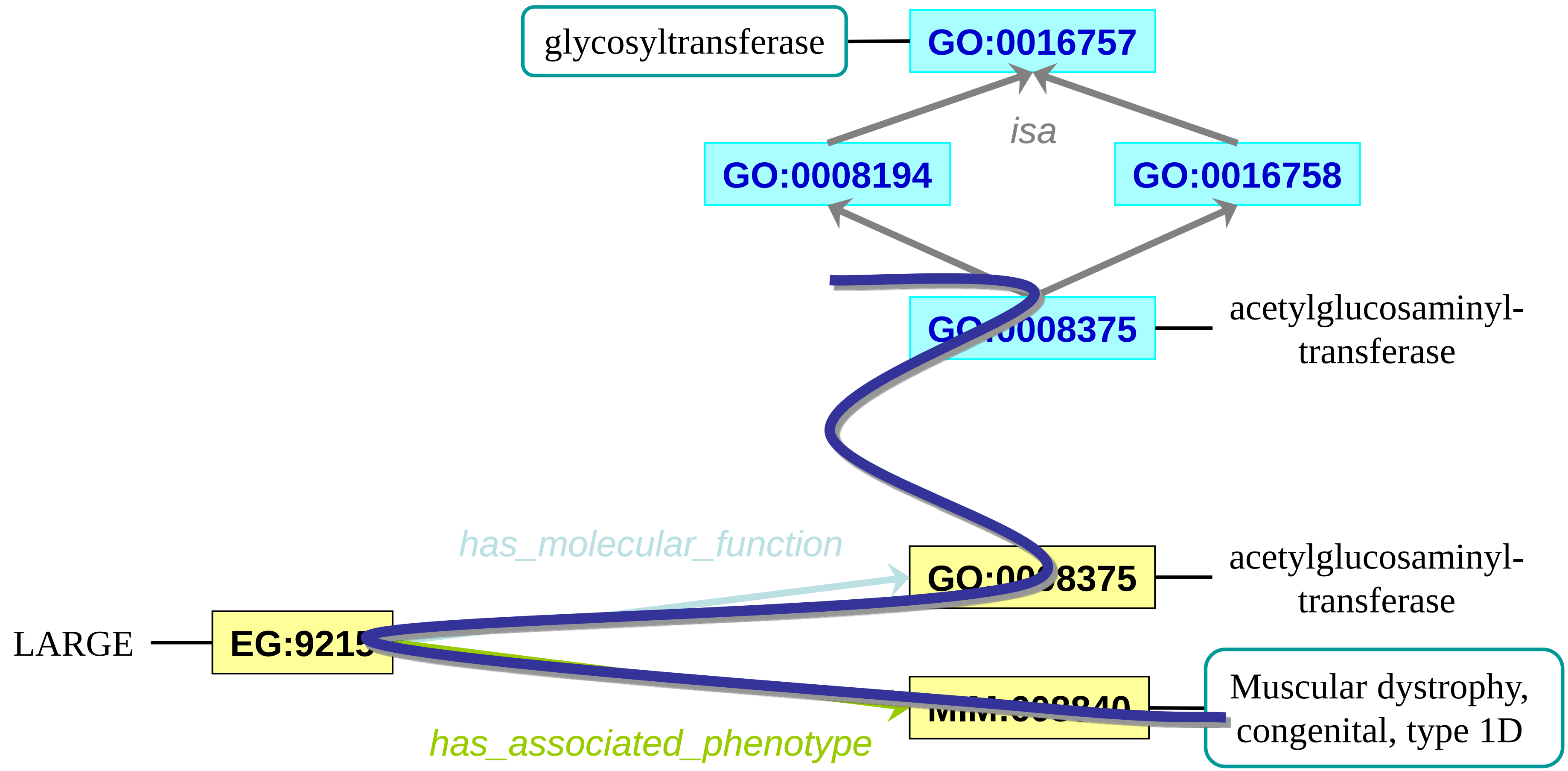
- Mapped 106 elements tags out of 124 element tags to named relations
- 50GB XML file → 39GB RDF file (411 million RDF triples)
- Oracle 10g release 2 with part of the 10.2.03 patch
- On a machine with 2 dual-core Intel Xeon 3.2GHz processor running Red Hat Enterprise Linux 4 (RHEL4)



```
<xsl:when test='$currNode="Entrezgene_track-
info"'>
<xsl:element name="{$ns}:has_entrezgene_track_info">
    <xsl:if test="../../* and ./* and not (@*)">
      <xsl:attribute name="rdf:parseType">
         Resource</xsl:attribute>
      </xsl:if>
```

Application I: Genome ↔ Phenotype

From *glycosyltransferase* to *congenital muscular dystrophy**



* From "glycosyltransferase" to "congenital muscular dystrophy": Integrating knowledge from NCBI Entrez Gene and the Gene Ontology"

Outline

- RDB to RDF – Objectives
- Method I: RDB to RDF without ontology
- Application I: Genome ↔ Phenotype
- Method II: RDB to RDF with ontology
- Application II: Genome ↔ Biological Pathway integration
- Conclusion

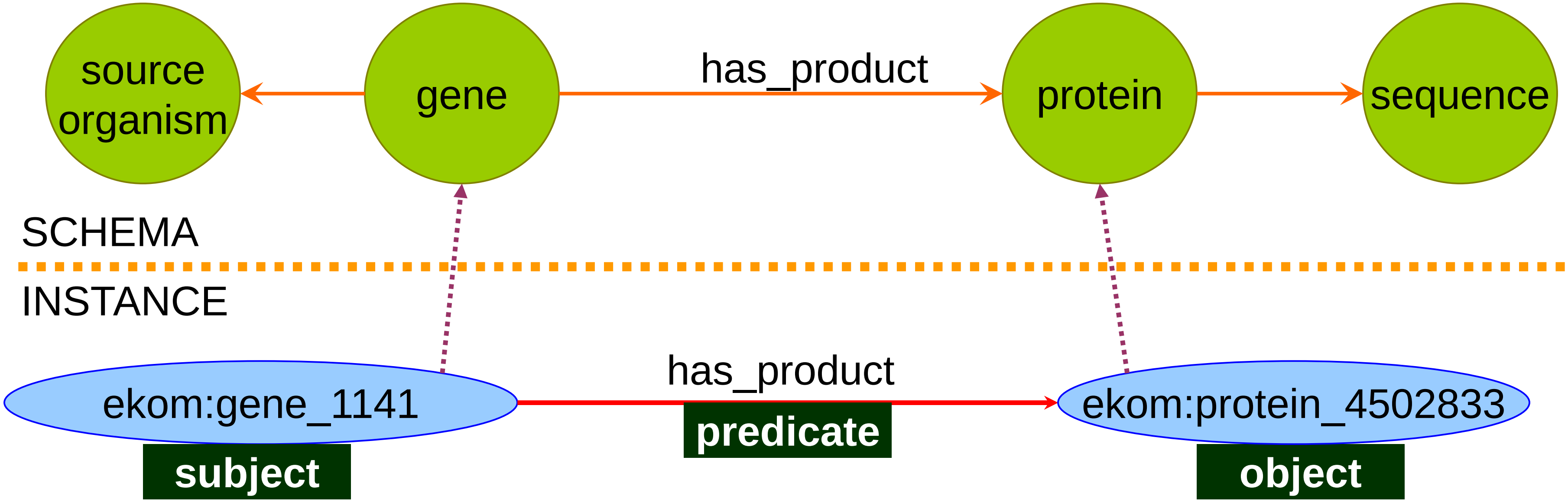
Data: Entrez Gene + HomoloGene + Biological Pathway

- In collaboration with National Institute on Drug Abuse (NIH)
- List of 449 human genes putatively involved with nicotine dependence (identified by Saccone et al.*)
- Understand gene functions and interactions, including their involvement in biological pathways
- List of queries:
  - *Which genes participate in a large number of pathways?*
  - *Which genes (or gene products) interact with each other?*
  - *Which genes are expressed in the brain?*

*S.F. Saccone, A.L. Hinrichs, N.L. Saccone, G.A. Chase, K. Konvicka and P.A. Madden et al., Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs, Hum Mol Genet 16 (1) (2007), pp. 36–49

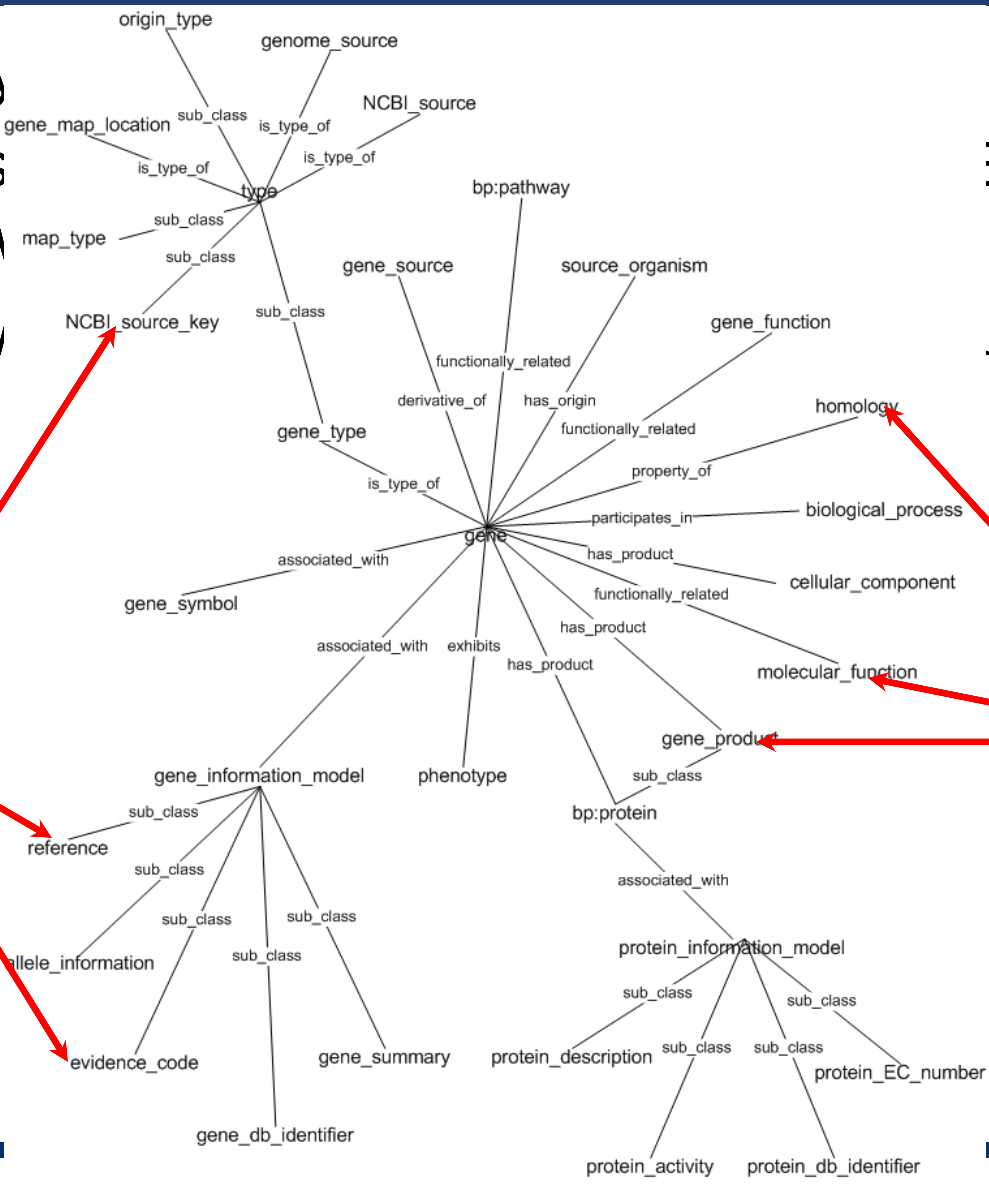
Method II: RDB to RDF with ontology

- Method I: cannot answer query "*Which genes participate in a large number of pathways?*"
- Need to specify a particular instance of gene or pathway as starting point in RDF graph
- Need to *classify* RDF instance data – Schema + Instance



Entrez Knowledge Model (OWL-DL)

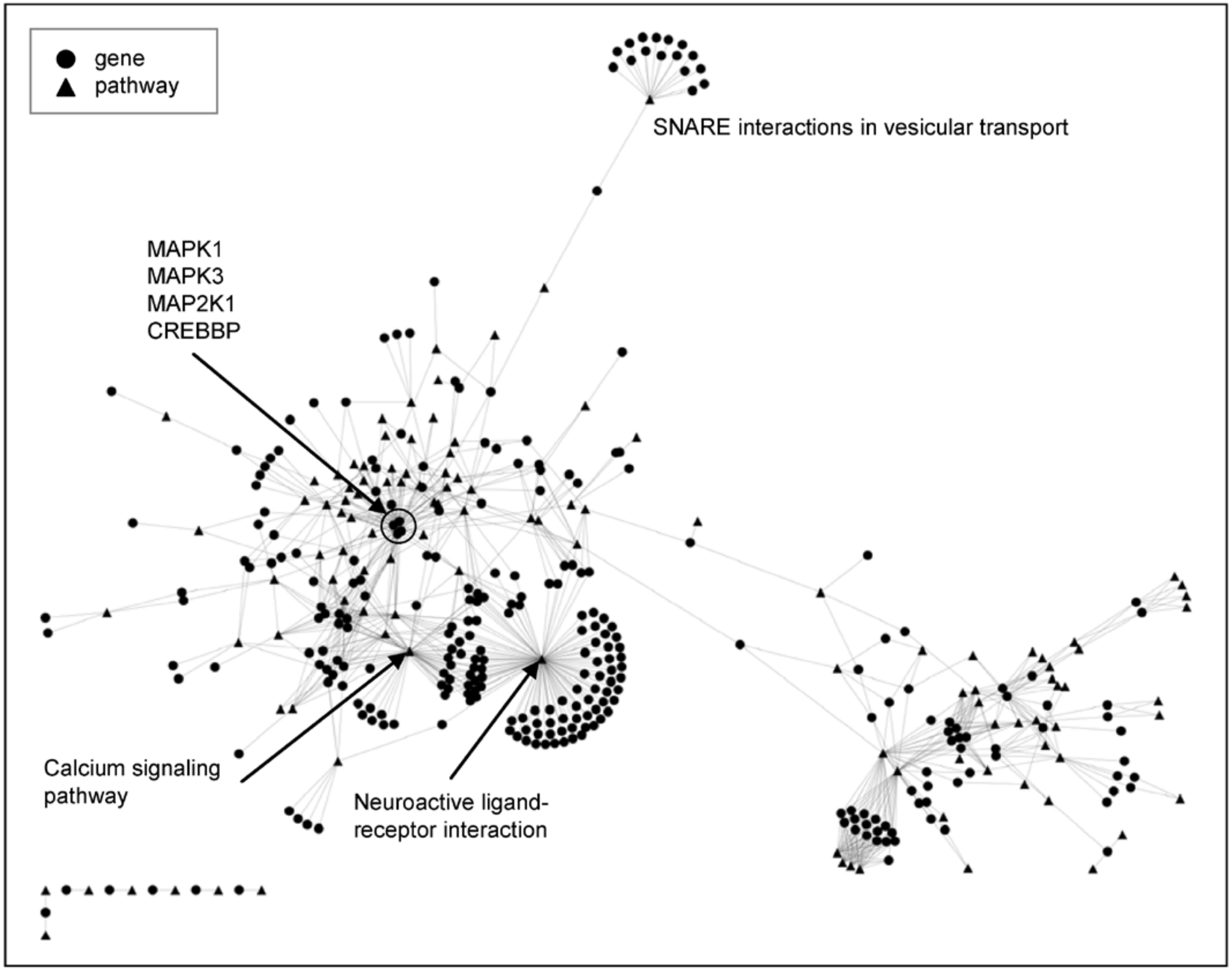
- No ontology
- Created a s Entrez Gene – Entrez Knov
- Integrated v l pathway data)



Information model concepts

Domain concepts

Application II: Genome ↔ Biological Pathway



● gene
▲ pathway
SNARE interactions in vesicular transport
MAPK1
MAPK3
MAP2K1
CREBBP
Calcium signaling pathway
Neuroactive ligand-receptor interaction

*An ontology-driven semantic mash-up of gene and biological pathway information: Application to the domain of nicotine dependence

Outline

- RDB to RDF – Objectives
- Method I: RDB to RDF without ontology
- Application I: Genome ↔ Phenotype
- Method II: RDB to RDF with ontology
- Application II: Genome ↔ Biological Pathway integration
- Conclusion

Conclusion

- Application driven approach for RDB to RDF – Biomedical Knowledge Integration
- Explicit modeling of domain semantics using named relations for
    - Accurate context based querying
    - Enhanced reasoning using relations based logic rules
- Use of ontology as reference knowledge model
- GRDDL compatible approach (using XSLT stylesheet) for transformation of RDB to RDF

- More information at:

http://knoesis.wright.edu/research/semsci/application_domain/sem_life_sci/bio/research/

Thank you