EMBL-EBI
European Bioinformatics Institute

Rebholz Group

EBI Industry Programme

European Bioinformatics Institute, Hinxton, UK
September 14, 2007

Integrating biomedical information through Semantic Web technologies

NATIONAL LIBRARY OF MEDICINE

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Outline

- Information integration in biomedicine
  - Motivation
  - Some issues: naming, normalization, mapping
  - Semantic Web perspective
- Examples
  - HCLS demo
  - From *glycosyltransferase* to *congenital muscular dystrophy*
- Role of ontologies

Motivation for integrating biomedical resources

Motivation

- Bridge across silos
  - E.g., translational research
- Data repositories to support
  - Hypothesis generation
  - Knowledge discovery
- Clinical data
  - Aggregation, sharing, exchange
  - Support for clinical decision

Motivation

- Complex queries often require multiple information sources
  - Knowledge bases
  - Ontologies
  - Biomedical literature
- Many information sources available
  - Heterogeneous
  - In different formats
- Interlinking is not integrating
  - Information retrieval and navigation (e.g., Entrez)

Interlinking vs. Integrating Entrez

NCBI Entrez Gene

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC

Search Gene for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Full Report Show 20 Send to

All: 1 | Current Only: 1 | Genes G

1: NF2 neurofibromin 2 (bilat

GeneID: 4771

Summary

Official Symbol	NF2
Official Full Name	neurofibro
Primary source	HGNC:777
See related	Ensembl:E
Gene type	protein co
RefSeq status	Reviewed
Organism	Homo sap
Lineage	Eukaryota Euarchont

KEGG Homo sapie

Entry	4771
Gene name	NF2
Definition	neurofibrom
Class	BRITE hierarchy
SSDB	Ortholog Pa
Motif	Pfam: FERM_ PROSITE: FE Motif
Other DBs	OMIM: 607379 NCBI-GI: 45 NCBI-GeneID HGNC: 7773 HPRD: 06980 Ensembl: ENSG00000186575 UniProt: P35240
LinkDB	PDB All DBs
Position	22q12.2

ExPASy Home page | Site Map | Search ExPASy | Contact us | Swiss-Prot

Search Swiss-Prot/TrEMBL for Go Clear

Printer-friendly view

Submit update

Quick BlastP search

UniProtKB/Swiss-Prot entry P35240

[Entry info] [Name and origin] [References] [Comments] [C

Note: most headings are clickable, even if they don't appear as links. They link to the user m

Entry information	
Entry name	MERL_HUMAN
Primary accession number	P35240
Secondary accession numbers	O95683 Q8WUJ2 Q969N0 Q969Q3 Q96
Integrated into Swiss-Prot on	February 1, 1994
Sequence was last modified on	February 1, 1994 (Sequence version 1)
Annotations were last modified on	August 21, 2007 (Entry version 92)

Name and origin of the protein	
Protein name	Merlin
Synonyms	Moesin-ezrin-radixin-like protein Neurofibromin-2 Schwannomin Schwannomerlin
Gene name	Name: NF2 Synonyms: SCH
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Haplorrhini; Catarrhini; Hominidae; Homo
Protein existence	1: Evidence at protein level;

NCBI OMIM Online Mendelian Inheritance in Man Johns Hopkins University

All Databases | PubMed | Nucleotide | Protein | Genome

Search OMIM for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Entrez

Display Titles Show 20 Send to

NCBI PubMed www.pubmed.gov A service of the National Library of Medicine and the National Institutes of Health

All Databases | PubMed | Nucleotide | Protein | Genome

Search PubMed for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display AbstractPlus Show 20 Sort By Send to

All: 1 | Review: 0

1: Int J Oncol. 2003 Dec;23(6):1493-500.

Overexpression of the NF2 gene inhibits schwannoma cell proliferation through promoting PDGFR degradation.

Fraenzer JT, Pan H, Minimo L Jr, Smith GM, Knauer D, Hung G.

House Ear Institute, Cell and Molecular Biology, Los Angeles, CA 90057, USA.

The loss of NF2 gene function leads to vestibular nerve schwannoma formation in humans. The NF2 gene product, Merlin/Schwannomin, has recently been found to interact with the two PDZ domains containing protein EBP50/NHE-RF, which is itself known to interact with the PDGF receptor (PDGFR) in several cell types. In this study, an up-regulation of both PDGFR and EBP50/NHE-RF, and an interaction of both proteins were found in primary human schwannoma tissue. Furthermore, using an adenoviral vector mediated gene transfer technique, changes in the phenotypic characteristics after NF2 gene restoration in a newly established NF2 gene-mutated human schwannoma cell line (HEI 193) were investigated. The overexpression of Merlin/Schwannomin in HEI 193 led to an inhibition of cell proliferation under serum-free conditions. Upon PDGF stimulation in culture, Merlin/Schwannomin appeared to inhibit the activation of the MAPK and PI3K signaling pathways, impinging on the phosphorylation of Erk 1/2 and Akt, respectively. The data also show that PDGFR is more rapidly internalized by the schwannoma cells overexpressing NF2. Therefore, this process is suggested as a model for a mechanism of Merlin/Schwannomin tumor suppressor function, which intermediates acceleration of the cell surface growth factor degradation.

PMID: 14612918 [PubMed - indexed for MEDLINE]

Issues in integrating biomedical information

Naming, normalization, mapping

❶ Naming

- Many biomedical entities have several names (synonymy)
  - Drug names
  - Gene names
  - Disease names
  - …
- A given name may refer to several different entities (polysemy)
  - Nail (body part)
  - Nail (medical device)

Brand names for paracetamol (acetaminophen)

http://en.wikipedia.org/wiki/List_of_paracetamol_brand_names

Brand name	Countries
Acamol	Israel
Atamel	Venezuela
Adol	Oman
Aldolor	Israel
Alvedon	Sweden
APAP	Poland
Benuron	Portugal, Germany
Biogesic	Philippines
Buscapina	Argentina
Cemol	Thailand
Crocin	India
Dafalgan	Belgium, France, Portugal, Russia, Ukraine
Daleron	Slovenia
Depon	Greece
Dexamol	Israel
Dolex	Colombia
Doliprane	France, Portugal, Russia, Ukraine
Efferalgan	France, Italy, Portugal, Russia, Spain, Ukraine
FeverAll	United States
Gelocatil	Spain
Gripin	Turkey
Lekadol	Croatia, Slovenia
Metacin	India
Pamol	Denmark, Finland, France
Panado	South Africa
Panadol	Australia, Azerbaijan, Central America, Egypt, Finland, Greece, Hong Kong, Hungary, Indonesia, Ireland, Kenya, Lebanon, Macedonia, Malaysia, Malta, Netherlands, New Zealand, Nigeria, Pakistan, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Sri Lanka, Switzerland, Taiwan, Ukraine, Estonia, United Kingdom
Panamax	Australia, United Kingdom
Panodil	Denmark, Iceland, Sweden
Paracet	Norway
Paralen	Czech Republic, Slovakia
Paramed	Botswana, South Africa, Zimbabwe
Paramol	Israel, Taiwan
Perdolan	Belgium
Perfalgan	Germany
Pinex	Denmark, Iceland, Norway
Plicet	Croatia
Reliv	Sweden
Rokamol	Israel
Sara	Thailand
Tachipirina	Italy
Tylenol	Brazil, Canada, Japan, South Korea, Thailand, United States
Tempra	Philippines

Names for dystrophin

Entrez Gene

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

DMD

Order cDNA clone, Links

Official Symbol DMD and **Name:** dystrophin (muscular dystrophy, Duchenne and Becker types) [*Homo sapiens*]
Other Aliases: GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272
Other Designations: Duchenne muscular dystrophy protein; dystrophin
Chromosome: X; **Location:** Xp21.2
Annotation: Chromosome X, NC_000023.9 (33267646..31047265, complement)
MIM: 300377
GeneID: 1756

Names for renal cell carcinoma

Details of 'clear cell carcinoma of kidney' | Distributed Relationships

ConceptStatus **Current**

Descriptions

- F clear cell carcinoma of kidney (disorder)
- P clear cell carcinoma of kidney
- S adenocarcinoma of kidney
- S carcinoma of kidney
- S Grawitz tumor
- S renal cell adenocarcinoma
- S renal cell carcinoma

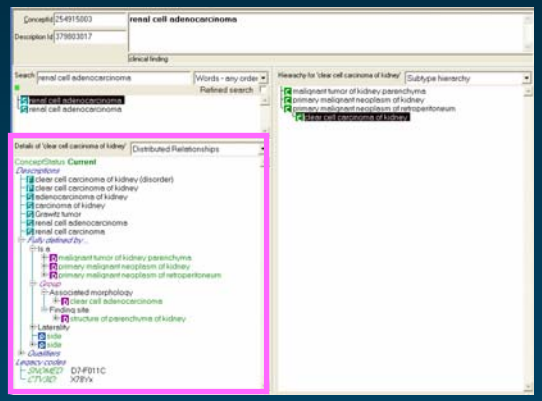
Fully defined by...

- Is a
  - D maligqnant tumor of kidney parenchyma
  - D primary maligqnant neoplasm of kidney
  - D primary maligqnant neoplasm of retroperitoneum
- *Group*
  - Associated morphologq
    - D clear cell adenocarcinoma
  - Finding site
    - D structure of parenchyma of kidney
- Laterality
  - side
  - side

Qualifiers

Legacy codes

- *SNOMED:* D7-F011C
- *CTV3ID:* X78Yx



http://www.clininfo.co.uk/clue5/clue.htm

Entity recognition

- Identifying biomedical entities in text
  - Names entity recognition
  - Tagging “mentions”
  - Semantic annotation
- Supported by terminology
  - Collects the names used in the domain
  - Often incompletely
- Example: BioCreative
  - 1A – Gene name identification
  - 2GM – Gene mention tagging

❷ Normalization

- Biomedical entities are identified by unique identifiers in various terminology systems
- Resolve names into identifiers (in a given namespace)
- Supported (in part) by terminology resources
- Example: BioCreAtIvE
  - 1B and 2GN – Gene Normalization

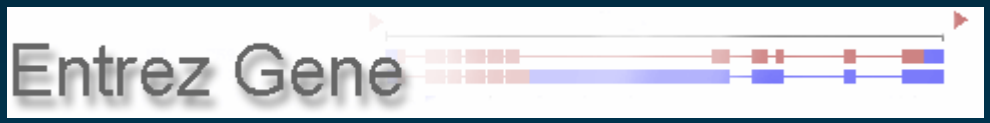
BioCreAtIvE
Critical Assessment for Information Extraction in Biology

Identifier for paracetamol (acetaminophen)

Master Drug Data Base. Medi-Span	5005	Acetaminophen
FDA National Drug Code Directory	50612	PARACETAMOL
FDA Structured Product Labels	362O9ITL9D	ACETAMINOPHEN
First DataBank NDDF Plus	001605	Acetaminophen
SNOMED Clinical Terms	90332006	Acetaminophen (product)
SNOMED Clinical Terms	387517004	Acetaminophen (substance)
VA National Drug File	4017513	ACETAMINOPHEN

Source: RxNorm database (5/3/2007)

Identifier for dystrophin



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

DMD

Order cDNA clone, Links

Official Symbol DMD and **Name:** dystrophin (muscular dystrophy, Duchenne and Becker types) [*Homo sapiens*]
Other Aliases: GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272
Other Designations: Duchenne muscular dystrophy protein; dystrophin
Chromosome: X; **Location:** Xp21.2
Annotation: Chromosome X, NC_000023.9 (33267646..31047265, complement)
MIM: 300377
GeneID: 1756

Identifier for renal cell carcinoma

Details of 'clear cell carcinoma of kidney' | Distributed Relationships

ConceptStatus **Current**

Descriptions

- F clear cell carcinoma of kidney (disorder)
- P clear cell carcinoma of kidney
- S adenocarcinoma of kidney
- S carcinoma of kidney
- S Grawitz tumor
- S renal cell adenocarcinoma
- S renal cell carcinoma

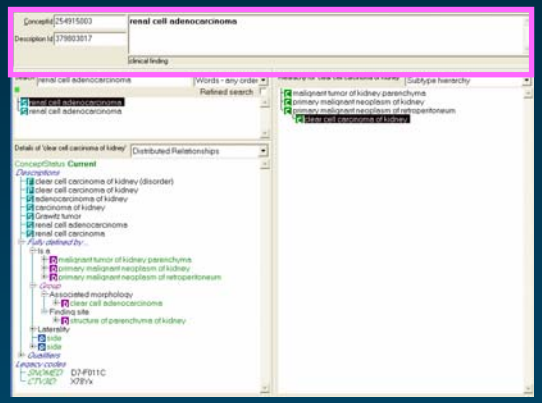
Fully defined by...

- Is a
  - D malignant tumor of kidney parenchyma
  - D primary malignant neoplasm of kidney
  - D primary malignant neoplasm of retroperitoneum
- *Group*
  - Associated morphology
    - D clear cell adenocarcinoma
  - Finding site
    - D structure of parenchyma of kidney
- Laterality
  - side
  - side

Qualifiers

Legacy codes

- *SNOMED:* D7-F011C
- *CTV3ID:* X78Yx



ConceptId 254915003 | **renal cell adenocarcinoma**

Description Id 379803017

clinical finding

http://www.clininfo.co.uk/clue5/clue.htm

❸ Mapping / Integration

- Identify equivalent entities across systems (across namespaces)
  - Shared identifiers
  - Existing mappings (e.g., SNOMED CT to ICD-9-CM)
  - Ontology alignment techniques (lexical + structural)
- Align equivalent entities
  - Pairwise: mapping
  - More broadly: integration
- Forms the basis for information integration in the Semantic Web (mashups)

Identifier for paracetamol (acetaminophen)

Master Drug Data Base. Medi-Span	5005	Acetaminophen
FDA National Drug Code Directory	50612	PARACETAMOL
FDA Structured Product Labels	362O9ITL9D	ACETAMINOPHEN
First DataBank NDDF Plus	001605	Acetaminophen
SNOMED Clinical Terms	90332006	Acetaminophen (product)
SNOMED Clinical Terms	387517004	Acetaminophen (substance)
VA National Drug File	4017513	ACETAMINOPHEN
RxNorm	161	Acetaminophen

Identifier for dystrophin

Entrez Gene

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

DMD

Order cDNA clone, Links

Official Symbol DMD and **Name:** dystrophin (muscular dystrophy, Duchenne and Becker types) [*Homo sapiens*]
Other Aliases: GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272
Other Designations: Duchenne muscular dystrophy protein; dystrophin
Chromosome: X; **Location:** Xp21.2
Annotation: Chromosome X, NC_000023.9 (33267646..31047265, complement)
MIM: 300377
GeneID: 1756

Identifier for renal cell carcinoma

Details of 'clear cell carcinoma of kidney' | Distributed Relationships

ConceptStatus **Current**

Descriptions

- F clear cell carcinoma of kidney (disorder) 645875019
- P clear cell carcinoma of kidney 379798014
- S adenocarcinoma of kidney 379801015
- S carcinoma of kidney 379800019
- S Grawitz tumor 379797016
- S renal cell adenocarcinoma 379803017
- S renal cell carcinoma 379802010

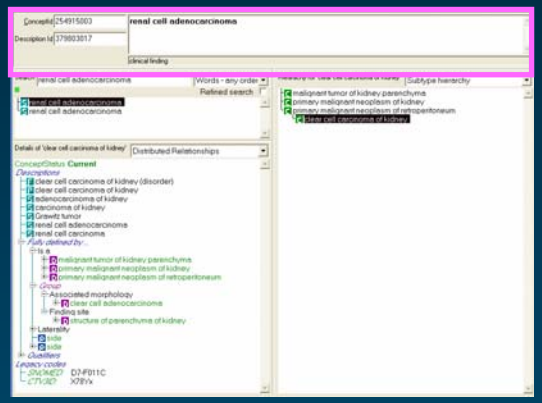
Fully defined by...

- Is a
  - D malignant tumor of kidney parenchyma
  - D primary malignant neoplasm of kidney
  - D primary malignant neoplasm of retroperitoneum
- *Group*
  - Associated morphology
    - D clear cell adenocarcinoma
  - Finding site
    - D structure of parenchyma of kidney
- Laterality
  - side
  - side

Qualifiers

Legacy codes

- *SNOMED:* D7-F011C
- *CTV3ID:* X78Yx



Concept Id 254915003 | **renal cell adenocarcinoma**

Description Id 379803017

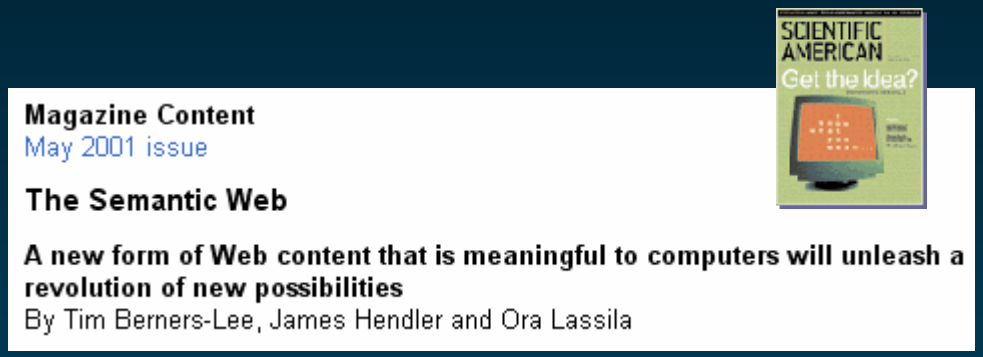
clinical finding

http://www.clininfo.co.uk/clue5/clue.htm

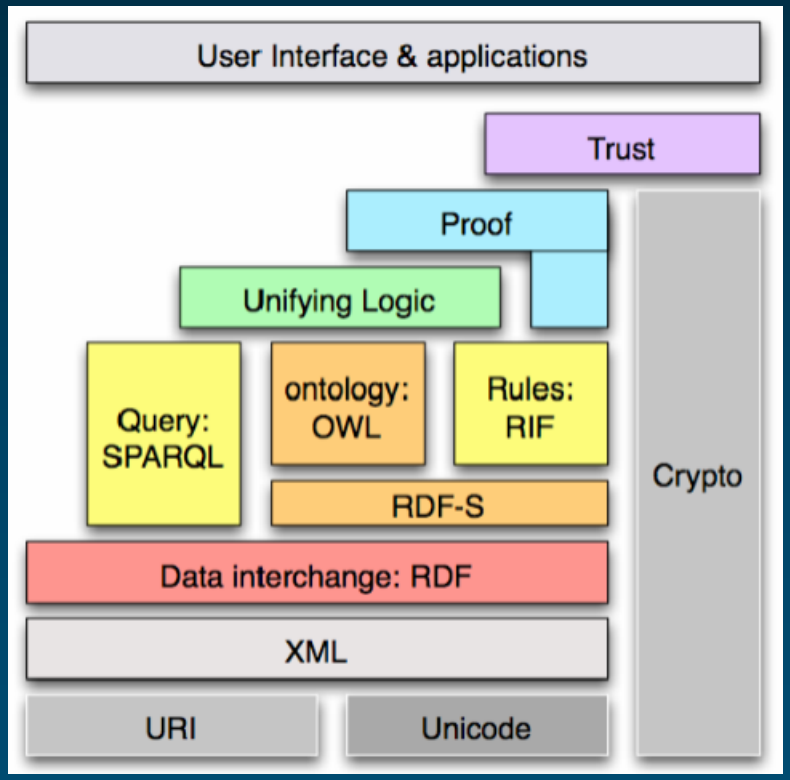
Information integration in biomedicine

Semantic Web perspective

Semantic Web



- Data vs. documents
- Sharing and reuse of data
- Data integration



W3C Health Care and Life Sciences IG

W3C Technology and Society domain | Semantic Web Activity

W3C Semantic Web Health Care and Life Sciences Interest Group

The Semantic Web Health Care and Life Sciences Interest Group is designed to improve collaboration, research and development, and innovation adoption in the health care and life science industries. Aiding decision-making in clinical research, Semantic Web technologies will bridge many forms of biological and medical information across institutions.

Contents: Mission and Scope | Membership and Joining | Charter /History | Resources | Presentations | Articles | New and Events | Conferences | Task Forces

Nearby: Discussion archive | HCLS WIKI | Applications and Demonstrations | OWL | RDF Data Access | Rules | Semantic Web Best Practices and Deployment

Introduction

Both Life Science Research and Health Care are areas undergoing phenomenal growth, holding much promise for our future as long as we can manage and apply the new knowledge gained without driving up costs. Key to their success is the implementation of new informatics models that will unite many forms of biological and medical information across all institutions, through the encoding of meaning into the data and their interpretations. By focusing on the semantics of information, researchers will have more access to the knowledge required to effectively find cures to diseases, while doctors will have better tools for individualized clinical management of patients.

Mission and Scope

The Semantic Web for Health Care and Life Sciences Interest Group (HCLSIG) is chartered to develop and support the use of Semantic Web technologies and practices to improve collaboration, research and development, and innovation adoption in the of Health Care and Life Science domains. Success in these domains depends on a foundation of semantically rich system, process and information interoperability. (more).

News and Events

- Last Call: SPARQL Query Language for RDF 2007-03-27: : Comments are due by 18 April.(Permalink)
- HCLS demo, planned for WWW2007 in Banff. To help participate in the demo, please contact Alan Ruttenberg.
- FIRST INTERNATIONAL WORKSHOP ON HEALTH CARE AND LIFE SCIENCES DATA INTEGRATION FOR THE SEMANTIC WEB, May 8, WWW2007 in Banff.
- Eric Prud'hommeaux, new W3C staff contact for HCLS.
- GRDDL links Microformats and Semantic Web: Working Draft xmlns="http://www.w3.org/2000/svg"(Permalink)

http://www.w3.org/2001/sw/hcls/

HCLS task forces

Task Forces

As a result of the Activities from the Jan 25-26 F2F Meeting, five new task forces have been drafted to address key areas necessary for implementation of semantic web for healthcare and life sciences. These task forces are a first draft (not yet official) and will be adjusted as necessary as progress is made. Each group is defining their scope, timeline, and deliverables within the 2yr HCLSIG timeframe:

1. BIORDF (Structured Data to RDF)
2. Scientific Publishing, (formerly Knowledge Life Cycle)
3. Ontologies Task Force
4. Adaptive Healthcare Protocols and Pathways
5. Drug Safety and Efficacy

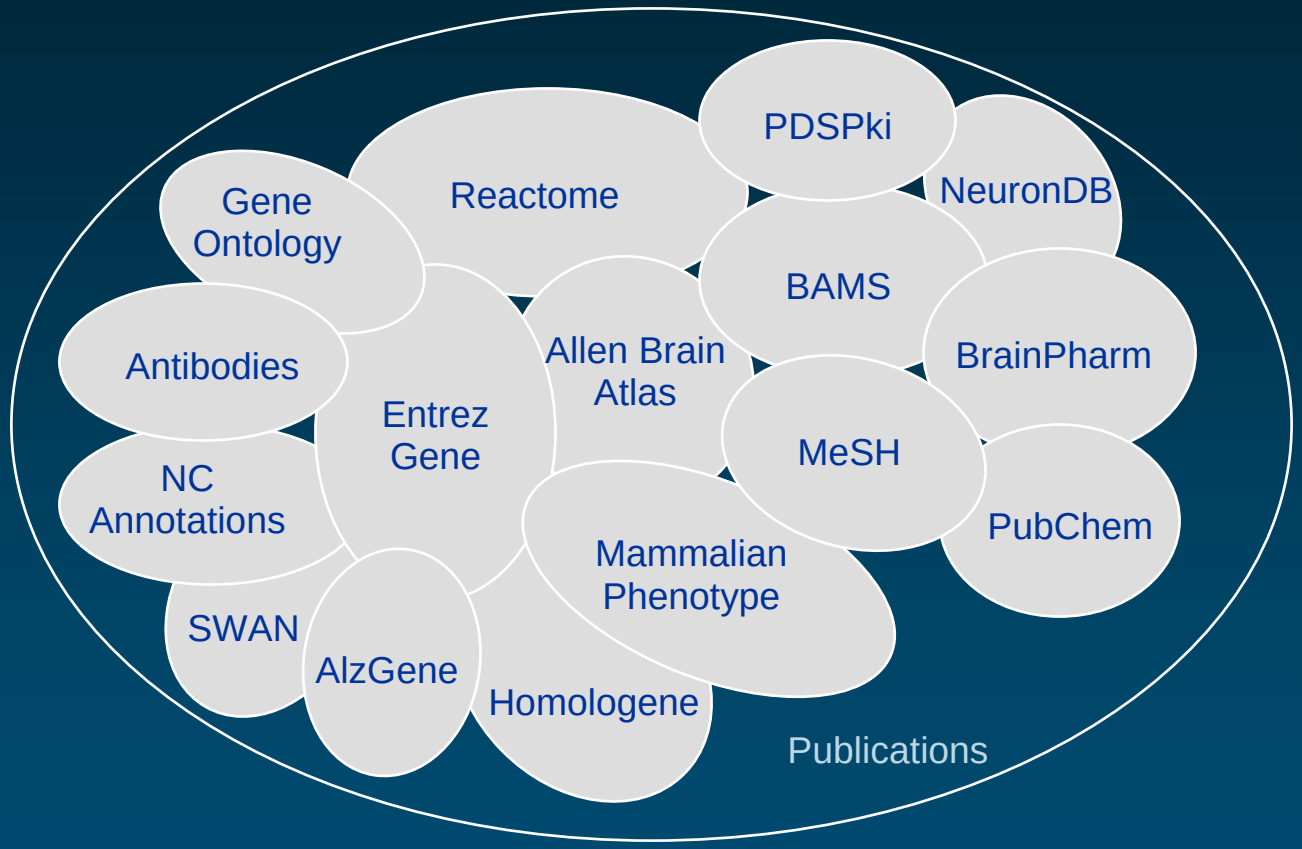
(10/6/06) Based on the second F2F held in Amsterdam, a specific set of action items have been agreed on: Action Items; Presentations and Minutes

(8/24/06) Renamed Knowledge LifecycleTask Force to Scientific Publishing (AJ Chen Coordinator); ROI Task Force completed

(7/21/06) New Drug Safety and Efficacy Task Force created to address best practices in critical areas of Drug R&D

(5/18/06) The tasks of the former Text-to-Structured Data (T2S) Task Force have been merged with tasks in the BioRDF task force.

HCLS mashup of biomedical sources



http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo

Shared identifiers Example

Entrez Gene

CH25H — Order cDNA clone, Links

Official Symbol CH25H and Name: cholesterol 25-hydroxylase [*Homo sapiens*]
Other Aliases: C25H
Chromosome: 10; Location: 10q23
Annotation: Chromosome 10, NC_000010.9 (90957050..90955509, complement)
MIM: 604551
GeneID: 9023

Pathways

Reactome Event: 73923 — Lipid and lipoprotein metabolism

Homology

Mouse, Rat
Map Viewer

GeneOntology

Function
- iron ion binding
- metal ion binding
- steroid hydroxylase activity

Process
- cholesterol metabolic process
- lipid metabolic process
- metabolic process
- sterol biosynthetic process

Component
- endoplasmic reticulum
- integral to membrane
- membrane
- membrane fraction

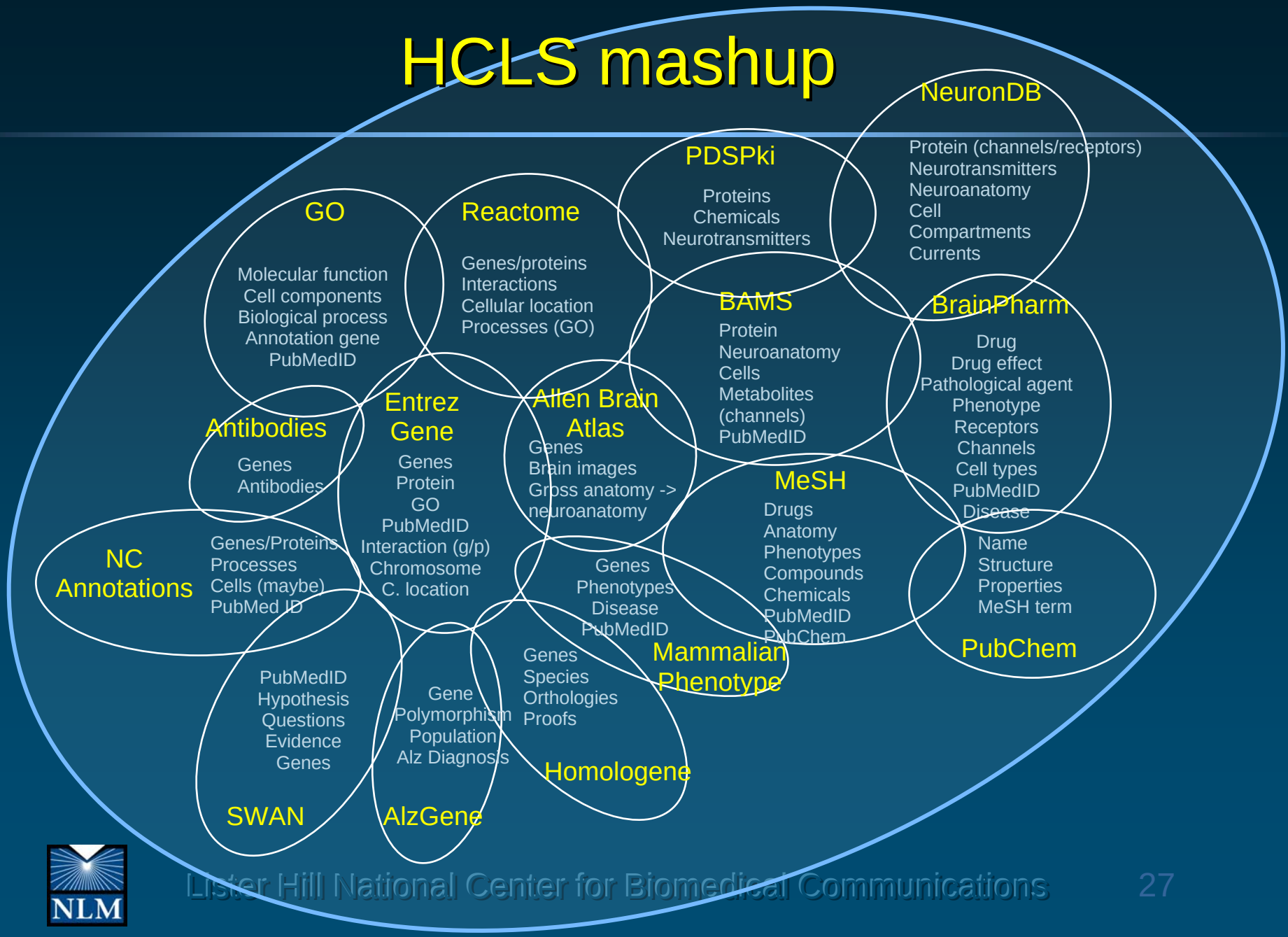
Cholesterol 25-hydroxylase [cytosol]

Details

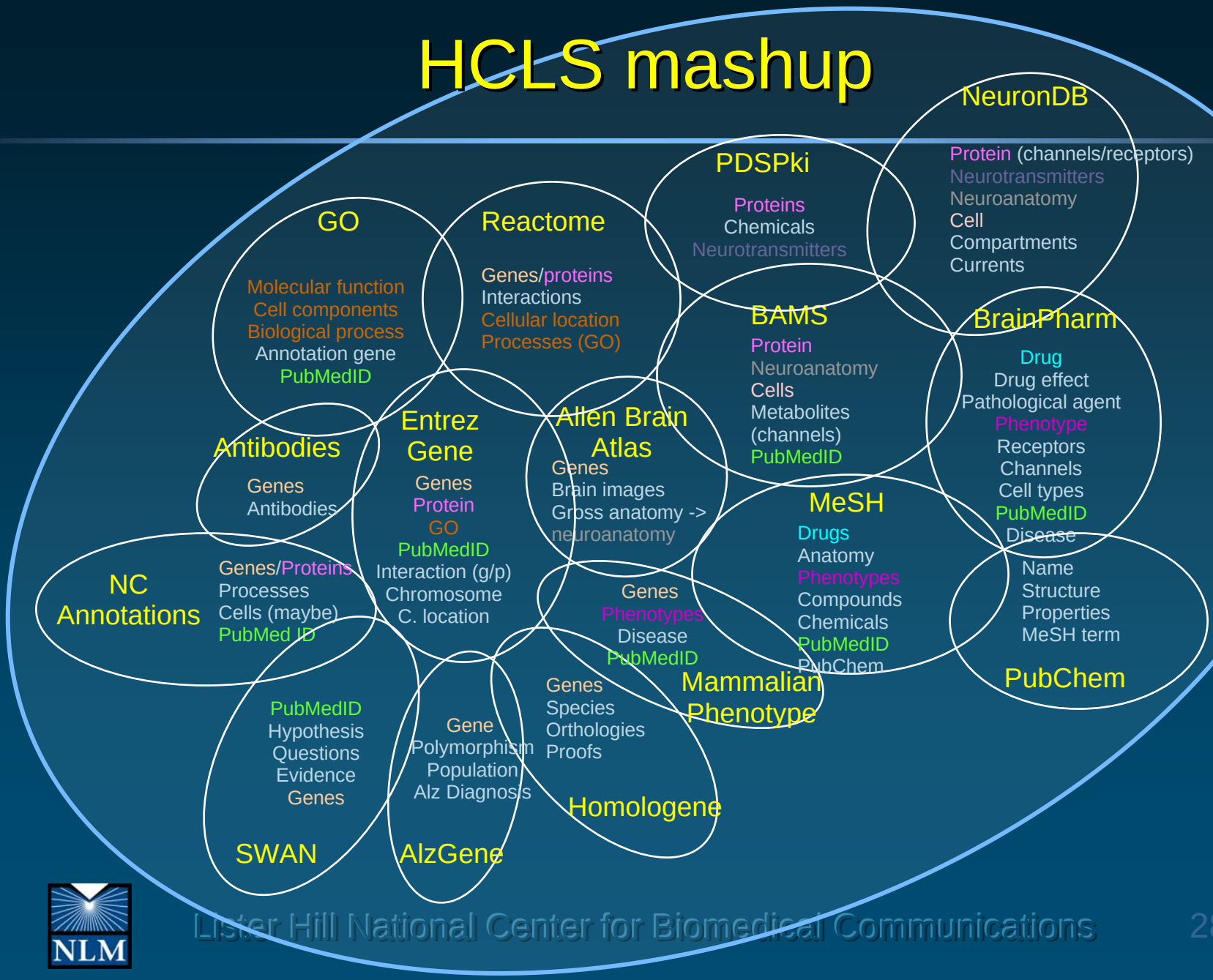
Name	Cholesterol 25-hydroxylase CH25H_HUMAN CH25H
Stable Identifier	REACT_10656.1
Links to corresponding entries in other databases	ENSEMBL:ENSG00000138135 Entrez Gene:9023 HapMap:NM_003956 KEGG Gene:9023 MIM:604551 RefSeq:NM_003956 RefSeq:NP_003947 UCSC:O95992 UniProt:O95992
Other identifiers related to this sequence	CH25H_HUMAN, ENSG00000138135, ENST00000371852, ENSP00000360918, ENST00000260706, ENSP00000260706, 206932_at, 323[illegible]_at, 45019_at, g4502498_3p_at, A_14_P139081, A_23_P86470, CCDS7400, GE6210[illegible], AF059212, AF059214, AL513533, BC017843, BC072430, EntrezGene:9023, GI_31542304-S, LMN_8057, IPI00022560, MIM:604551, OTTHUMT00000049291, AAC97481, AAC97483, CAI13519, AAH17843, AAH72430, NM_003956, NP_003947, Hs.47357, Hs.597033, O95992, CH25H_HUMAN, IPR006088
Reference entity	UniProt:O95992 Cholesterol 25-hydroxylase
Coordinates in the reference sequence	..
Cellular compartment	cytosol GO
Organism	Homo sapiens
Component of	CH25H (Fe2+ cofactor) [endoplasmic reticulum membrane]
Participates in processes	

- Lipid and lipoprotein metabolism
  - Steroid metabolism
    - Metabolism of bile acids and bile salts
      - Synthesis of bile acids and bile salts
        - **Cholesterol is hydroxylated to 25-hydroxycholesterol [Homo sapiens]**

HCLS mashup


NeuronDB
Protein (channels/receptors)
Neurotransmitters
Neuroanatomy
Cell
Compartments
Currents
PDSPki
Proteins
Chemicals
Neurotransmitters
GO
Molecular function
Cell components
Biological process
Annotation gene
PubMedID
Reactome
Genes/proteins
Interactions
Cellular location
Processes (GO)
BAMS
Protein
Neuroanatomy
Cells
Metabolites
(channels)
PubMedID
BrainPharm
Drug
Drug effect
Pathological agent
Phenotype
Receptors
Channels
Cell types
PubMedID
Disease
Antibodies
Genes
Antibodies
Entrez Gene
Genes
Protein
GO
PubMedID
Interaction (g/p)
Chromosome
C. location
Allen Brain Atlas
Genes
Brain images
Gross anatomy -> neuroanatomy
MeSH
Drugs
Anatomy
Phenotypes
Compounds
Chemicals
PubMedID
PubChem
NC Annotations
Genes/Proteins
Processes
Cells (maybe)
PubMed ID
Genes
Phenotypes
Disease
PubMedID
Mammalian Phenotype
Name
Structure
Properties
MeSH term
PubChem
PubMedID
Hypothesis
Questions
Evidence
Genes
SWAN
Gene
Polymorphism
Population
Alz Diagnosis
AlzGene
Genes
Species
Orthologies
Proofs
Homologene

HCLS mashup


NeuronDB
Protein (channels/receptors)
Neurotransmitters
Neuroanatomy
Cell
Compartments
Currents
PDSPki
Proteins
Chemicals
Neurotransmitters
GO
Molecular function
Cell components
Biological process
Annotation gene
PubMedID
Reactome
Genes/proteins
Interactions
Cellular location
Processes (GO)
BAMS
Protein
Neuroanatomy
Cells
Metabolites
(channels)
PubMedID
BrainPharm
Drug
Drug effect
Pathological agent
Phenotype
Receptors
Channels
Cell types
PubMedID
Disease
Antibodies
Genes
Antibodies
Entrez Gene
Genes
Protein
GO
PubMedID
Interaction (g/p)
Chromosome
C. location
Allen Brain Atlas
Genes
Brain images
Gross anatomy ->
neuroanatomy
MeSH
Drugs
Anatomy
Phenotypes
Compounds
Chemicals
PubMedID
PubChem
NC Annotations
Genes/Proteins
Processes
Cells (maybe)
PubMed ID
Genes
Phenotypes
Disease
PubMedID
Name
Structure
Properties
MeSH term
PubChem
Mammalian Phenotype
PubMedID
Hypothesis
Questions
Evidence
Genes
SWAN
Gene
Polymorphism
Population
Alz Diagnosis
AlzGene
Genes
Species
Orthologies
Proofs
Homologene

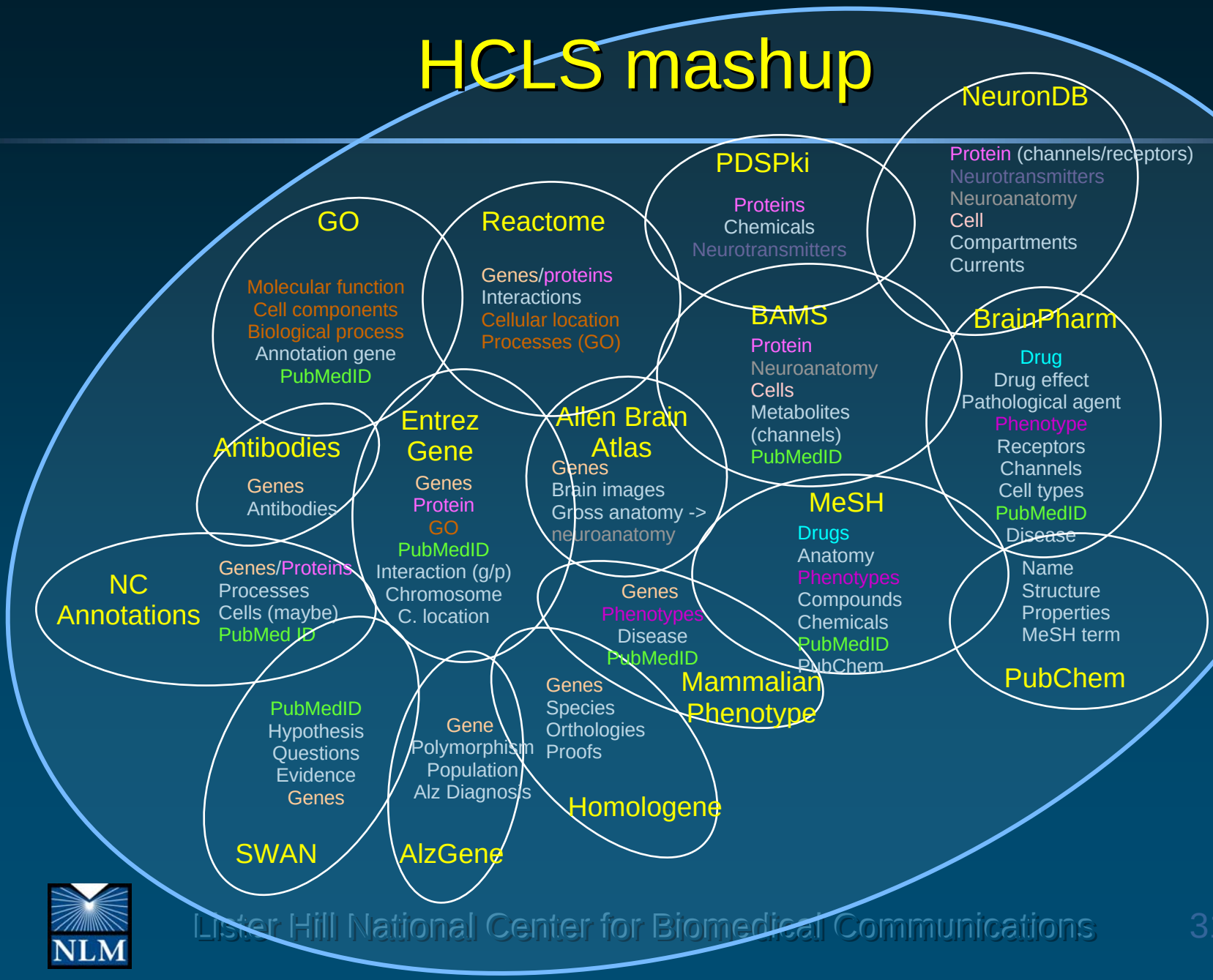
Example 1

HCLS demo

HCLS demo

- Proof of concept for information integration
  - Objective: “to demonstrate the value of semantic web technology to health care and the life sciences by highlighting the benefits of using semantic web technology”
- Created in the spring of 2007
- Presented at WWW2007 in Banff, Canada

HCLS mashup


NeuronDB
Protein (channels/receptors)
Neurotransmitters
Neuroanatomy
Cell
Compartments
Currents
PDSPki
Proteins
Chemicals
Neurotransmitters
GO
Molecular function
Cell components
Biological process
Annotation gene
PubMedID
Reactome
Genes/proteins
Interactions
Cellular location
Processes (GO)
BAMS
Protein
Neuroanatomy
Cells
Metabolites
(channels)
PubMedID
BrainPharm
Drug
Drug effect
Pathological agent
Phenotype
Receptors
Channels
Cell types
PubMedID
Disease
Antibodies
Genes
Antibodies
Entrez Gene
Genes
Protein
GO
PubMedID
Interaction (g/p)
Chromosome
C. location
Allen Brain Atlas
Genes
Brain images
Gross anatomy ->
neuroanatomy
MeSH
Drugs
Anatomy
Phenotypes
Compounds
Chemicals
PubMedID
PubChem
NC Annotations
Genes/Proteins
Processes
Cells (maybe)
PubMed ID
Genes
Phenotypes
Disease
PubMedID
Mammalian Phenotype
Name
Structure
Properties
MeSH term
PubChem
PubMedID
Hypothesis
Questions
Evidence
Genes
SWAN
Gene
Polymorphism
Population
Alz Diagnosis
AlzGene
Genes
Species
Orthologies
Proofs
Homologene

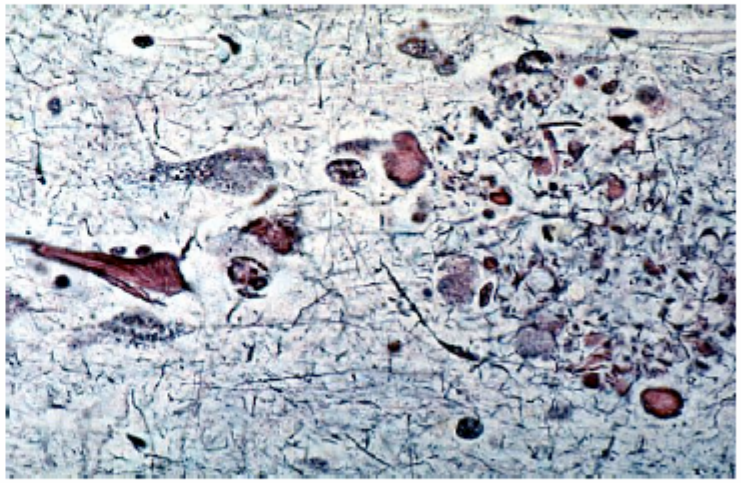
HCLS demo Question

Looking for Alzheimer Disease targets

Signal transduction pathways are considered to be rich in “druggable” targets - proteins that might respond to chemical therapy

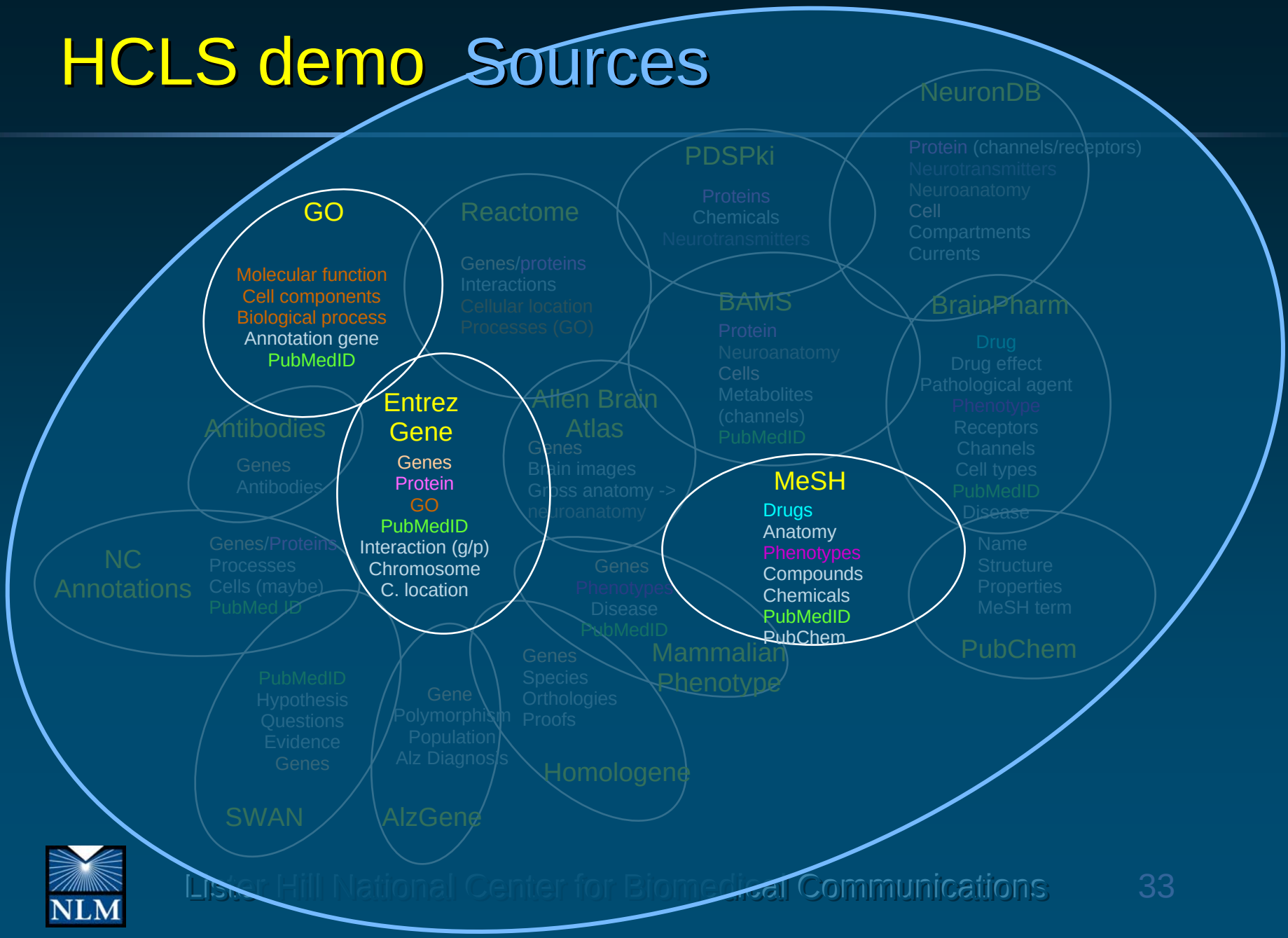
CA1 Pyramidal Neurons are known to be particularly damaged in Alzheimer’s disease.

Casting a wide net, can we find candidate genes known to be involved in signal transduction and active in Pyramidal Neurons?



http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo

HCLS demo Sources



HCLS demo Query

A SPARQL query spanning 4 sources

```
prefix go: <http://purl.org/obo/owl/GO#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix mesh: <http://purl.org/commons/record/mesh/>
prefix sc: <http://purl.org/science/owl/sciencecommons/>
prefix ro: <http://www.obofoundry.org/ro/ro.owl#>

select ?genename ?processname
where
{  graph <http://purl.org/commons/hcls/pubmesh>
   { ?paper ?p mesh:D017966 .
     ?article sc:identified_by_pmid ?paper.
     ?gene sc:describes_gene_or_gene_product_mentioned_by ?article.
   }
  graph <http://purl.org/commons/hcls/goa>
   { ?protein rdfs:subClassOf ?res.
     ?res owl:onProperty ro:has_function.
     ?res owl:someValuesFrom ?res2.
     ?res2 owl:onProperty ro:realized_as.
     ?res2 owl:someValuesFrom ?process.
  graph <http://purl.org/commons/hcls/20070416/classrelations>
   {{?process <http://purl.org/obo/owl/obo#part_of> go:GO_0007166}
     union
    {?process rdfs:subClassOf go:GO_0007166 }}
     ?protein rdfs:subClassOf ?parent.
     ?parent owl:equivalentClass ?res3.
     ?res3 owl:hasValue ?gene.
   }
  graph <http://purl.org/commons/hcls/gene>
   { ?gene rdfs:label ?genename }
  graph <http://purl.org/commons/hcls/20070416>
   { ?process rdfs:label ?processname}
}
```

Mesh: Pyramidal Neurons

↓

Pubmed: Journal Articles

↓

Entrez Gene: Genes

↓

GO: Signal Transduction

Inference required

http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo

HCLS demo Results

Results

Many of the genes are indeed related to Alzheimer's Disease through gamma secretase (presenilin) activity

DRD1, 1812	adenylate cyclase activation
ADRB2, 154	adenylate cyclase activation
ADRB2, 154	arrestin mediated desensitization of G-protein coupled receptor protein signaling pathway
DRD1IP, 50632	dopamine receptor signaling pathway
DRD1, 1812	dopamine receptor, adenylate cyclase activating pathway
DRD2, 1813	dopamine receptor, adenylate cyclase inhibiting pathway
GRM7, 2917	G-protein coupled receptor protein signaling pathway
GNG3, 2785	G-protein coupled receptor protein signaling pathway
GNG12, 55970	G-protein coupled receptor protein signaling pathway
DRD2, 1813	G-protein coupled receptor protein signaling pathway
ADRB2, 154	G-protein coupled receptor protein signaling pathway
CALM3, 808	G-protein coupled receptor protein signaling pathway
HTR2A, 3356	G-protein coupled receptor protein signaling pathway
DRD1, 1812	G-protein signaling, coupled to cyclic nucleotide second messenger
SSTR5, 6755	G-protein signaling, coupled to cyclic nucleotide second messenger
MTNR1A, 4543	G-protein signaling, coupled to cyclic nucleotide second messenger
CNR2, 1269	G-protein signaling, coupled to cyclic nucleotide second messenger
HTR6, 3362	G-protein signaling, coupled to cyclic nucleotide second messenger
GRIK2, 2898	glutamate signaling pathway
GRIN1, 2902	glutamate signaling pathway
GRIN2A, 2903	glutamate signaling pathway
GRIN2B, 2904	glutamate signaling pathway
ADAM10, 102	integrin-mediated signaling pathway
GRM7, 2917	negative regulation of adenylate cyclase activity
LRP1, 4035	negative regulation of Wnt receptor signaling pathway
ADAM10, 102	Notch receptor processing
ASCL1, 429	Notch signaling pathway
HTR2A, 3356	serotonin receptor signaling pathway
ADRB2, 154	transmembrane receptor protein tyrosine kinase activation (dimerization)
PTPRG, 5793	transmembrane receptor protein tyrosine kinase signaling pathway
EPHA4, 2043	transmembrane receptor protein tyrosine kinase signaling pathway
NRTN, 4902	transmembrane receptor protein tyrosine kinase signaling pathway
CTNND1, 1500	Wnt receptor signaling pathway

http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo

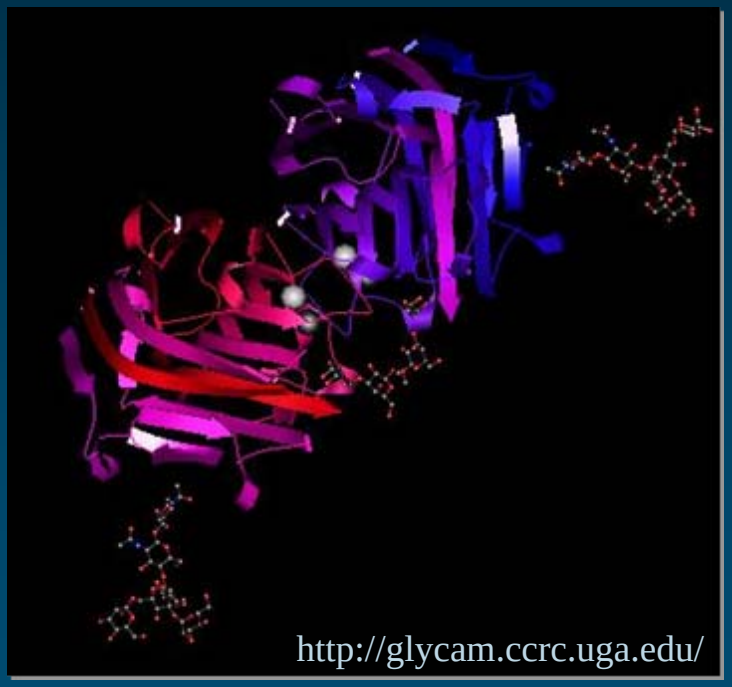
Example 2

From glycosyltransferase *to* congenital muscular dystrophy

[Sahoo S, Zeng K, Bodenreider O, Sheth AP., Medinfo 2007:1260-1264]

Scenario

- A researcher is interested in glycosylation and its implications for one disorder: congenital muscular dystrophy.

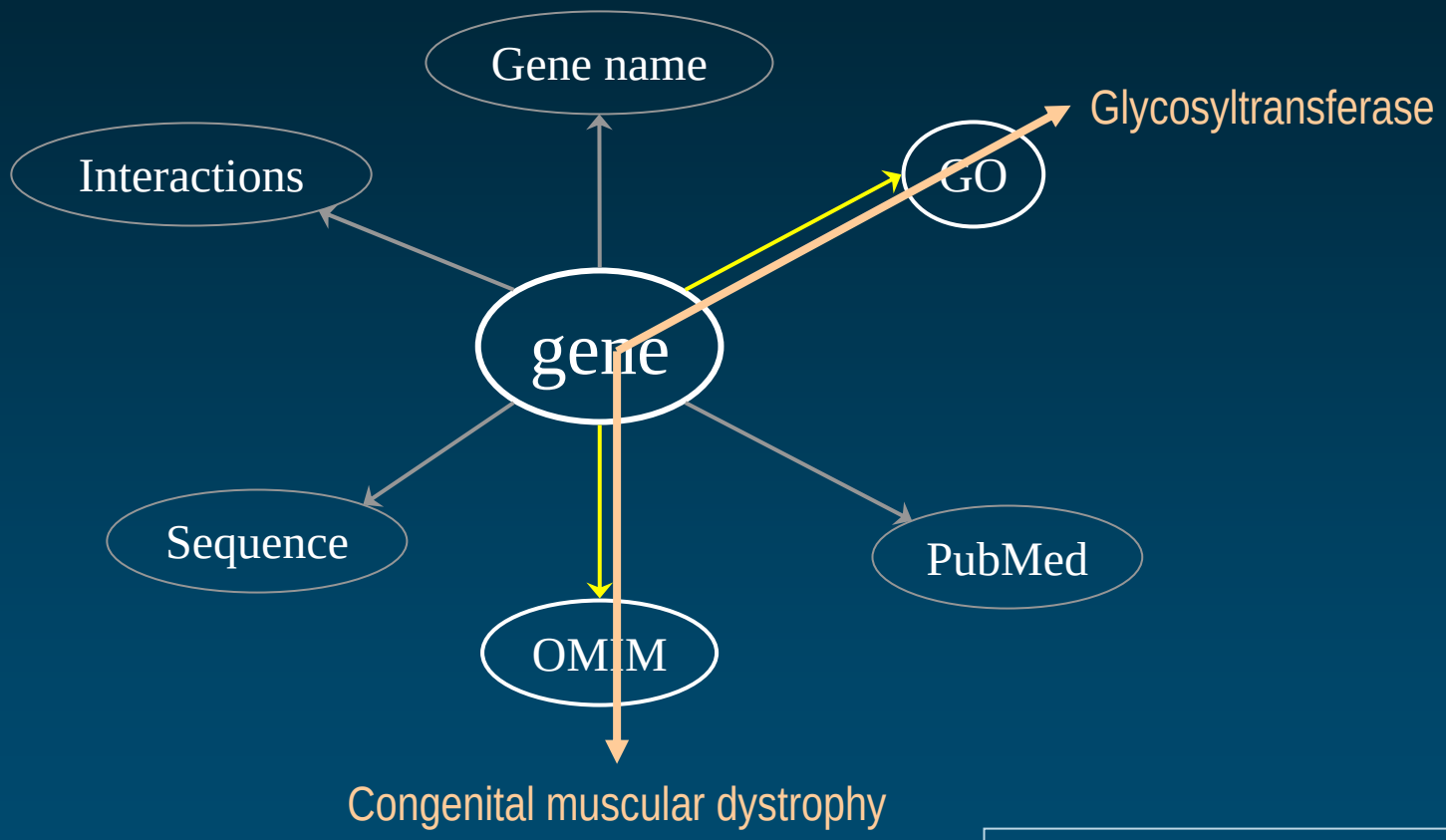


(source: Dr. Renuka Kadirvelraj, U. Georgia)

Biological hypothesis

Link between glycosyltransferase activity and congenital muscular dystrophy?

Information source Entrez Gene



Link between glycosyltransferase activity and congenital muscular dystrophy?

Entrez Gene query (1)

NCBI

Entrez Gene

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books | OMIM

Search Gene for glycosyltransferase Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Entrez Gene
- Home
- About
- FAQ
- Help
- Gene Handbook
- Statistics
- Downloads (FTP)

Mailing Lists
- Gene
- RefSeq

Feedback
- Help Desk
- Corrections
- About GeneRIFs

Related Sites
- BLAST

Display Summary Show 20 Send to

All: 11474 | Current Only: 11361 | Genes Genomes: 11229 | SNP GeneView: 267

Items 1 - 20 of 11474 Page 1 of 574 Next

1: Ogt Links

O-glycosyltransferase [*Drosophila melanogaster*]
Other Aliases: Dmel_CG10392, CG10392
Other Designations: O-glycosyltransferase CG10392-PA, isoform A; O-glycosyltransferase CG10392-PB, isoform B; O-glycosyltransferase CG10392-PC, isoform C
Chromosome: 2R; **Location:** 41E4-41E5
GeneID: 35486

2: Ugt1a10 Links

Official Symbol Ugt1a10 and **Name:** UDP glycosyltransferase 1 family polypeptide A10 [*Rattus norvegicus*]
Other Aliases: Ugt1a11
Other Designations: UDP glycosyltransferase 1 family polypeptide A11; UDP glycosyltransferase 1 family, polypeptide A10
Chromosome: 9; **Location:** 9q35
Annotation: Chromosome 9, NC_005108.2 (86986915..87098362)
GeneID: 396552

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene

Entrez Gene query (2)

NCBI

Entrez Gene

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books | OMIM

Search Gene for glycosyltransferase[GO] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Send to

All: 2 | Current Only: 1 | Genes Genomes: 1 | SNP GeneView: 1

Items 1 - 2 of 2 One page.

Entrez Gene
- Home
- About
- FAQ
- Help
- Gene Handbook
- Statistics
- Downloads (FTP)

Mailing Lists
- Gene
- RefSeq

Feedback
- Help Desk
- Corrections
- About GeneRIFs

Related Sites
- BLAST

1: rho-5 Links

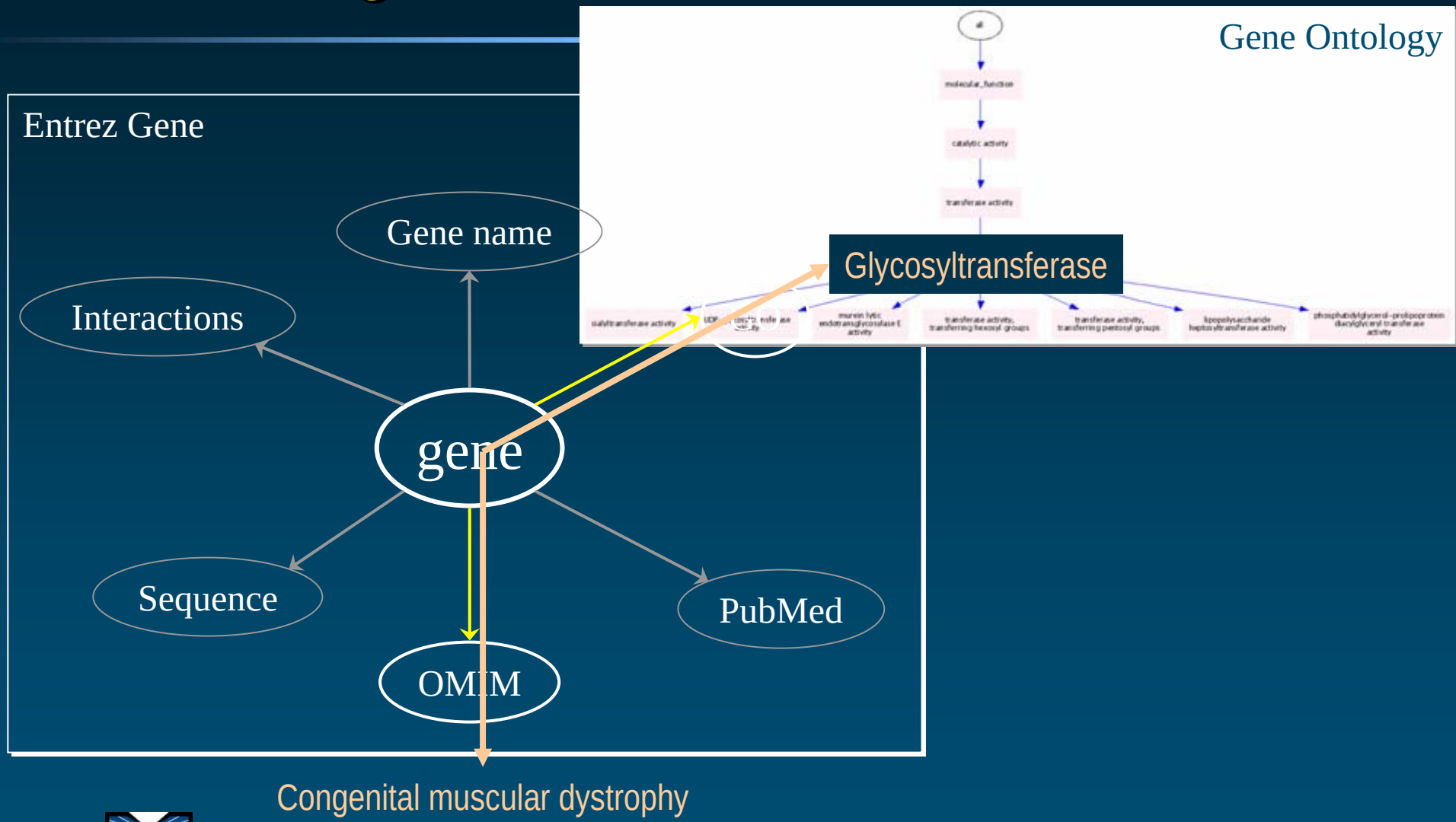
rhomboid-5 [*Drosophila melanogaster*]
Other Aliases: CG5364, Rho-related [31D10]
Other Designations: rhomboid-5 CG5364-PA
Chromosome: 2L; **Location:** 31D10-31D11
GeneID: 34407
This record was discontinued.

2: ChGn Order cDNA clone, Links

chondroitin beta1,4 N-acetylgalactosaminyltransferase [*Homo sapiens*]
Other Aliases: FLJ11264, beta4GalNAcT
Chromosome: 8; **Location:** 8p21.3
Annotation: Chromosome 8, NC_000008.9 (19305952..19584374, complement)
GeneID: 55790

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene

Integration Entrez Gene + GO



RDF triple Gene property

subject **predicate** **object**

GeneID: 9215 —*eg:has_symbol*→ **LARGE**

GeneID: 9215 —*eg:has_molecular_function*→ **GO:0008375**

acetylglucosaminyltransferase activity

GeneID: 9215 —*eg:has_associated_phenotype*→ **MIM: 608840**

Muscular dystrophy, congenital, type 1D

NCBI Entrez Gene

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC

Search Gene for 9215[uid] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Full Report Show 20 Send to

All: 1 | Current Only: 1 | Genes Genomes: 1 | SNP GeneView: 1

1: LARGE like-glycosyltransferase [*Homo sapiens*]

GeneID: 9215 updated 02-Jul-2007

LARGE (GeneID: 9215)

Phenotypes → *has_associated_disease*

Muscular dystrophy, congenital, type 1D
MIM: 608840

Congenital muscular dystrophy, type 1D

GeneOntology Provided by GOA

Function	Evidence	
acetylglucosaminyltransferase activity	TAS	PubMed

Process	Evidence	
N-acetylglucosamine metabolic process	TAS	PubMed
carbohydrate biosynthetic process	IEA	
glycosphingolipid biosynthetic process	TAS	PubMed
muscle maintenance	ISS	
protein amino acid glycosylation	TAS	PubMed

Component	Evidence	
integral to Golgi membrane	TAS	PubMed
integral to membrane	IEA	
membrane	IEA	

NCBI Entrez Gene

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC

Search Gene for 9215[uid] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Full Report Show 20 Send to

All: 1 | Current Only: 1 | Genes Genomes: 1 | SNP GeneView: 1

1: LARGE like-glycosyltransferase [*Homo sapiens*]

GeneID: 9215 updated 02-Jul-2007

LARGE (GeneID: 9215)

Phenotypes

Muscular dystrophy, congenital, type 1D
MIM: 608840

GeneOntology Provided by GOA

has_molecular_function

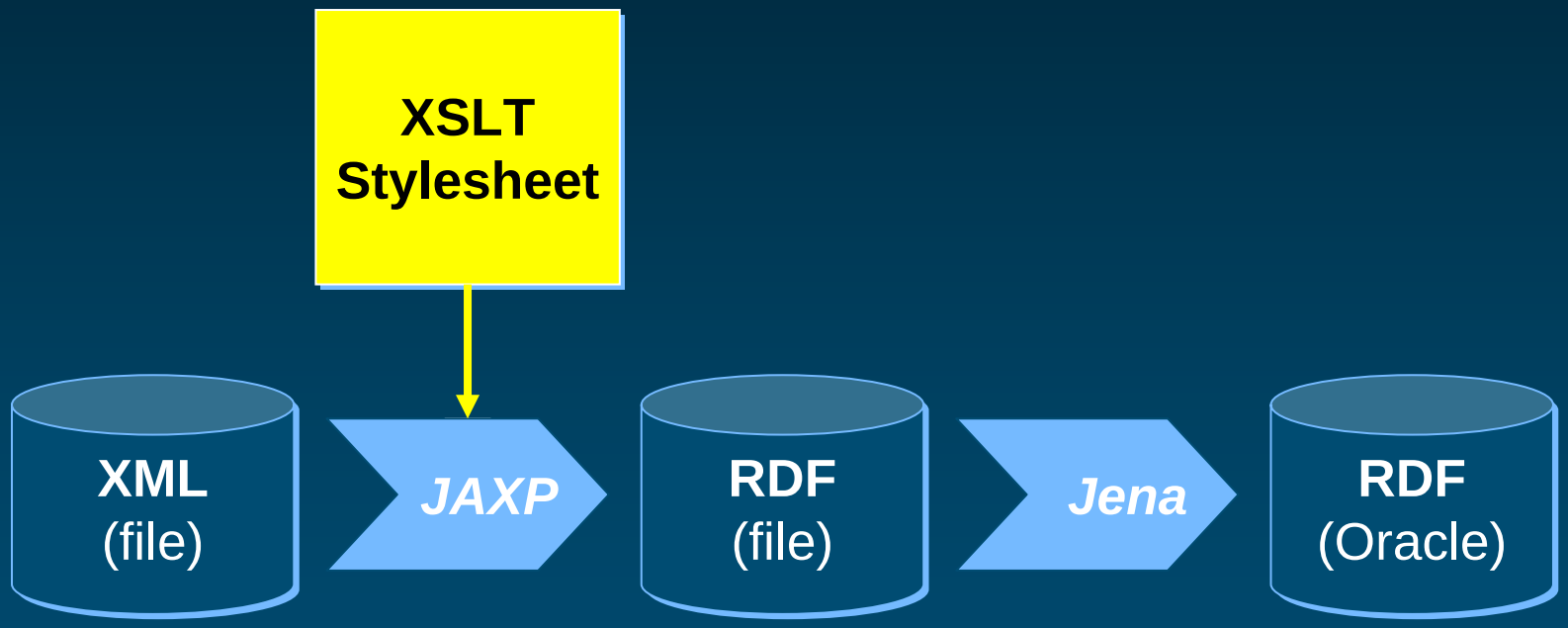
Function	Evidence	
acetylglucosaminyltransferase activity	TAS	PubMed

acetylglucosaminyltransferase activity

Process	Evidence	
N-acetylglucosamine metabolic process	TAS	PubMed
carbohydrate biosynthetic process	IEA	
glycosphingolipid biosynthetic process	TAS	PubMed
muscle maintenance	ISS	
protein amino acid glycosylation	TAS	PubMed

Component	Evidence	
integral to Golgi membrane	TAS	PubMed
integral to membrane	IEA	
membrane	IEA	

Converting EG to RDF



2M genes

9.5M triples
(keep all "live" genes,
exclude evidence code "IEA")

Acquiring the GO

GO RDF-XML Format

The GO RDF-XML version of GO, which includes all three ontologies and the definitions, can be downloaded from the GO database archive. The document type definition (DTD) is available from the GO FTP site.

The GO RDF-XML file is built from the flat files and the gene association files on a monthly basis.

Here's a GO RDF-XML snapshot (with some lines wrapped for legibility):

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE go:go>

<go:go xmlns:go="xml-dtd/go.dtd#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <go:version timestamp="Wed May 9 23:55:02 2001" />
  <rdf:RDF>
    <go:term rdf:about="go#GO:0003673">
      <go:accession>GO:0003673</go:accession>
      <go:name>Gene_Ontology</go:name>
      <go:definition></go:definition>
    </go:term>
    <go:term rdf:about="go#GO:0003674">
      <go:accession>GO:0003674</go:accession>
      <go:name>molecular_function</go:name>
      <go:definition>The action characteristic of a gene product.</go:definition>
      <go:part-of rdf:resource="go#GO:0003673" />
      <go:dbxref>
        <go:database_symbol>go</go:database_symbol>
        <go:reference>curators</go:reference>
      </go:dbxref>
    </go:term>
    <go:term rdf:about="go#GO:0016209">
      <go:accession>GO:0016209</go:accession>
      <go:name>antioxidant</go:name>
      <go:definition></go:definition>
      <go:isa rdf:resource="go#GO:0003674"
      <go:association>
```

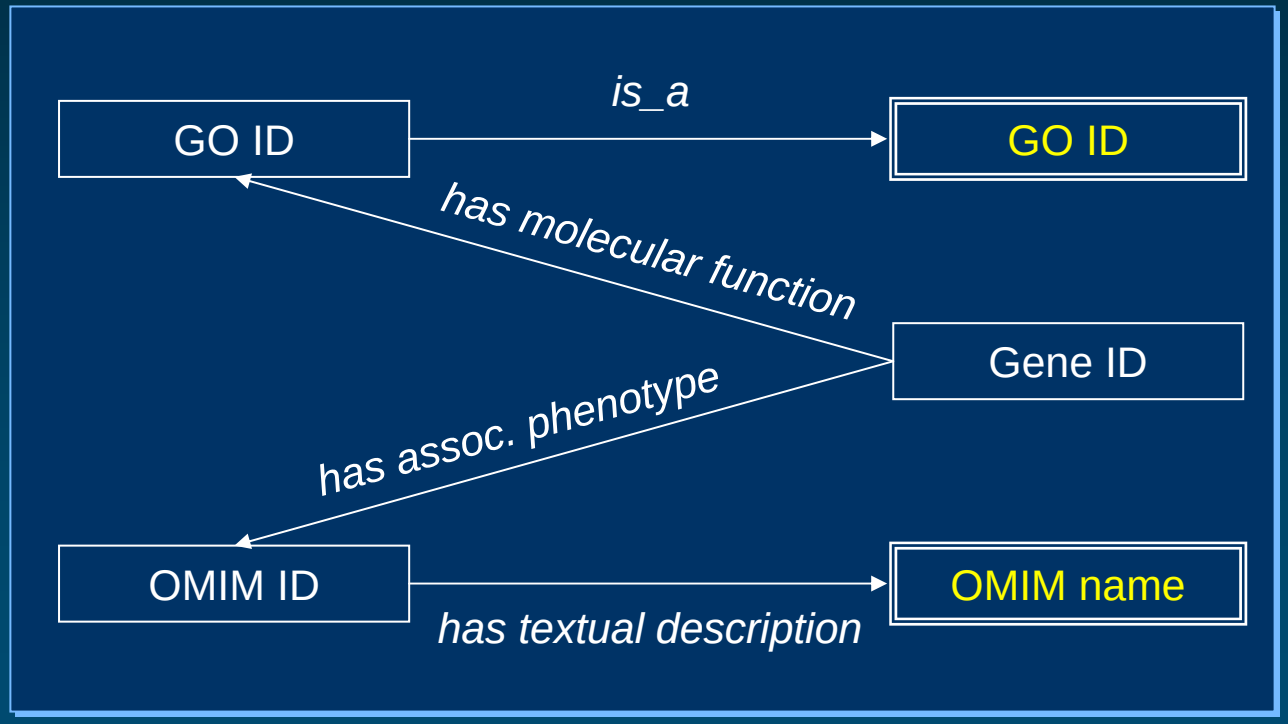
http://geneontology.org/GO.downloads.ontology.shtml

Rule base

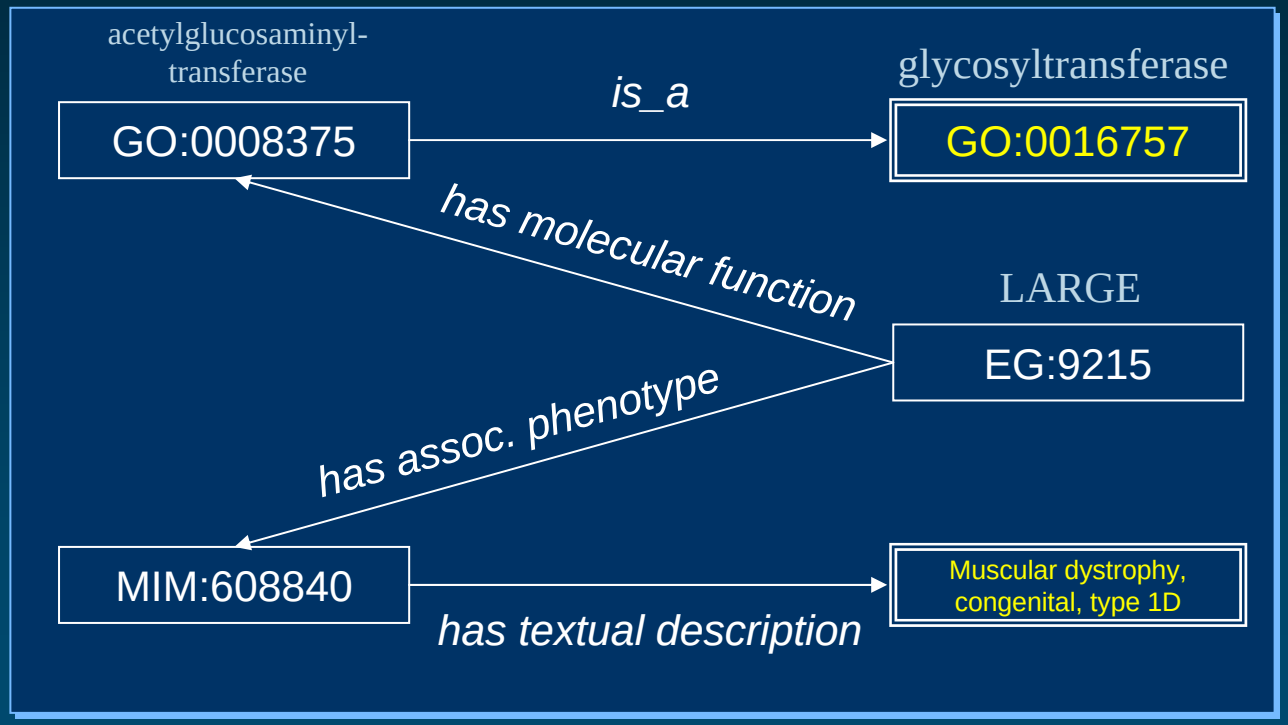
Relation	*is_a*	*part_of*
is_a	IF <**x** *is_a* **y**> & <**y** *is_a* **z**> THEN <**x** *is_a* **z**>	IF <**x** *is_a* **y**> & <**y** *part_of* **z**> THEN <**x** *part_of* **z**>
part_of	IF <**x** *part_of* **y**> & <**y** *is_a* **z**> THEN <**x** *part_of* **z**>	IF <**x** *part_of* **y**> & <**y** *part_of* **z**> THEN <**x** *part_of* **z**>

Using SPARQL to test a hypothesis

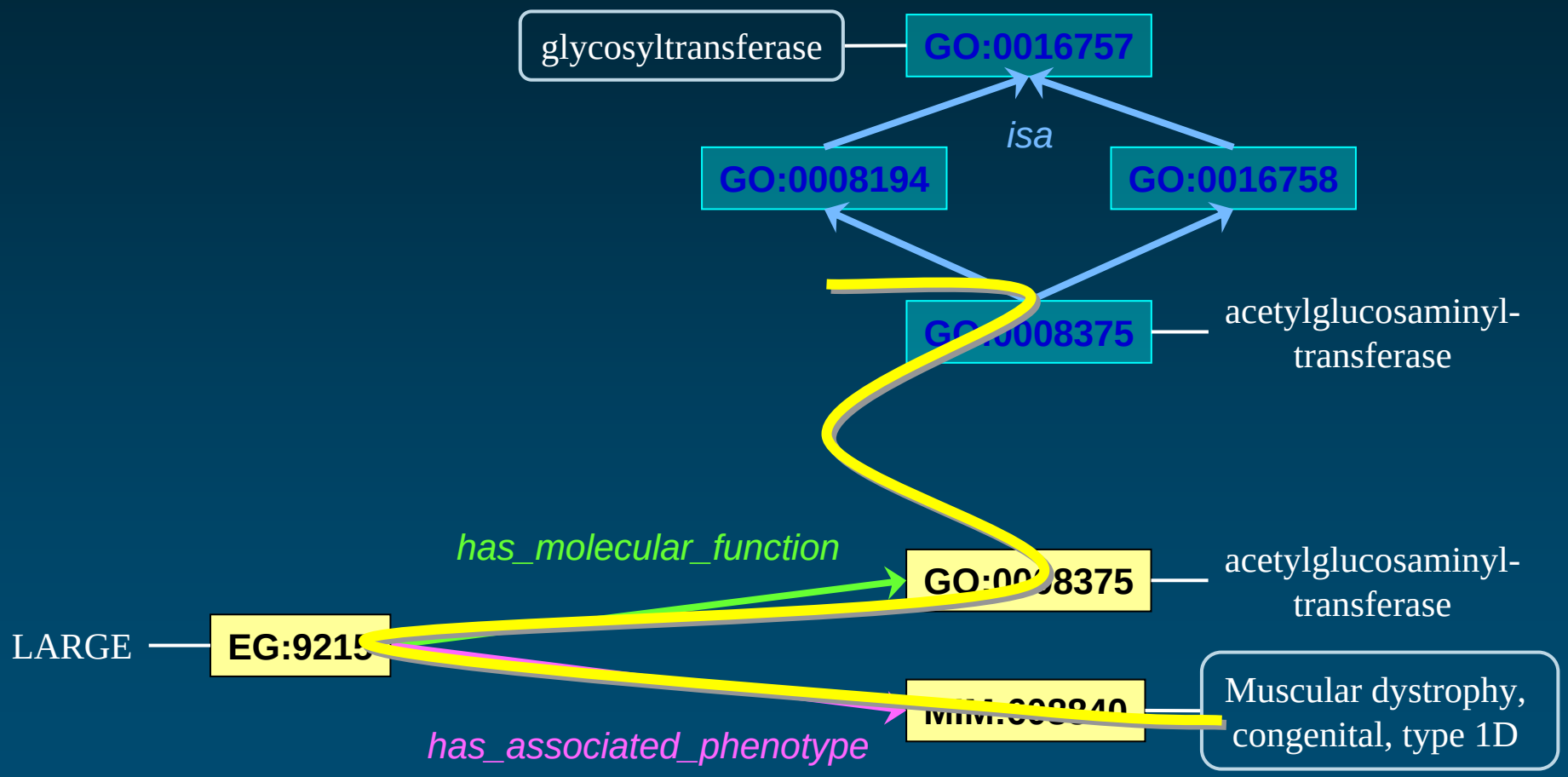
Find all the genes annotated with the GO molecular function glycosyltransferase or any of its descendants and associated with any form of congenital muscular dystrophy



Results Instantiated graph



From *glycosyltransferase* to *congenital muscular dystrophy*



Role of ontologies in information integration

Ontologies and Semantic Web

University of Pisa, Italy
June 14, 2007

NETTAB 2007 - A Semantic Web for Bioinformatics

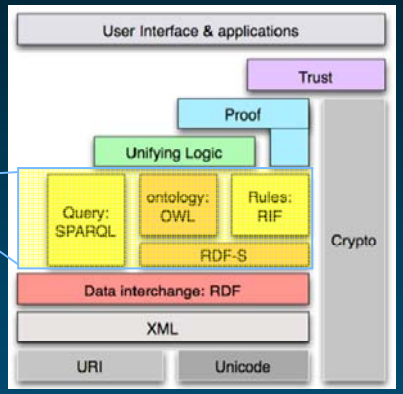
Bio-ontologies

The cream in the Semantic Web layer cake

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

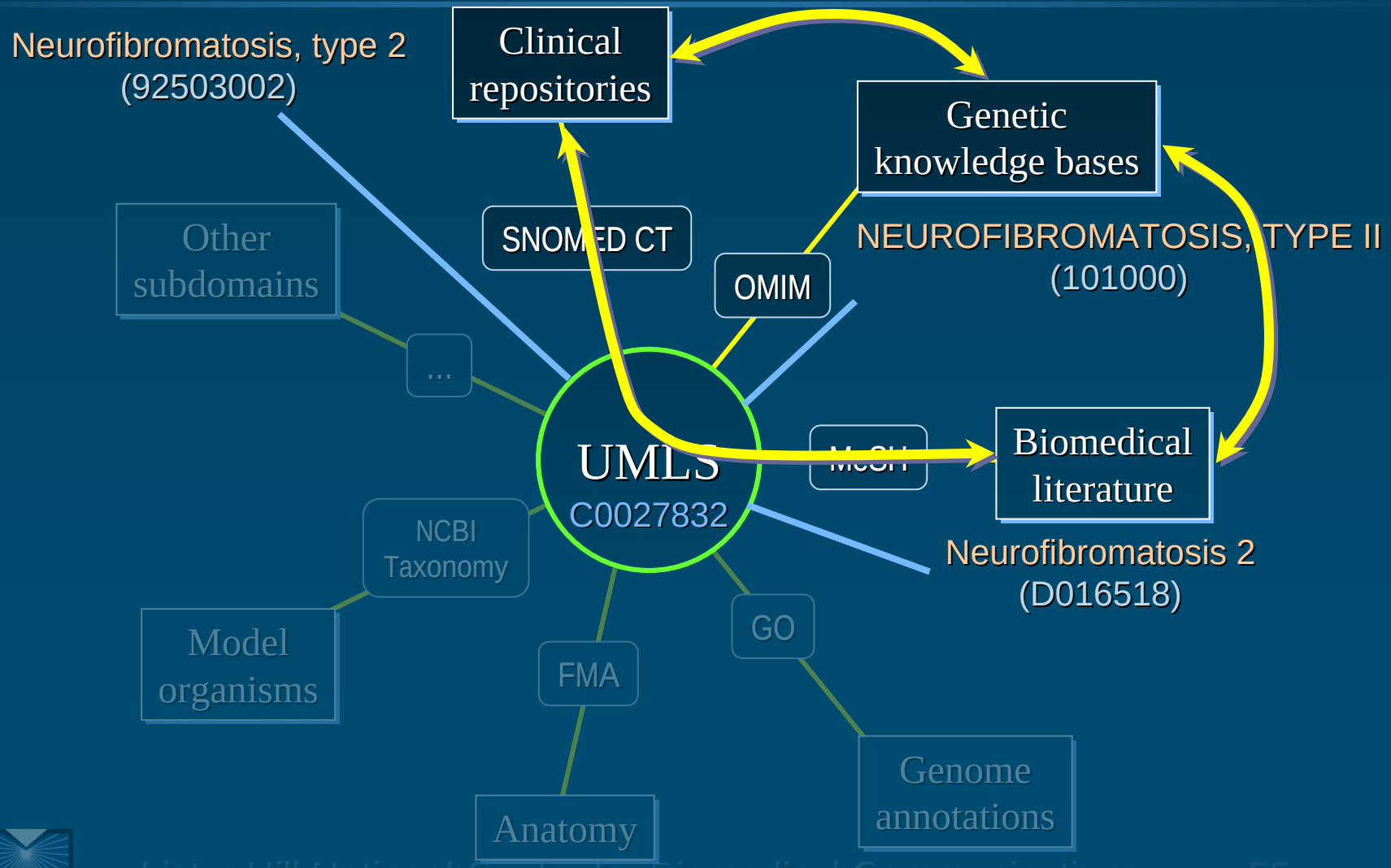
Semantic Web layer cake



Ontologies and integration

- Terminologies/Ontologies provide
  - Lists of entities
  - Names for entities
  - Identifiers for entities
- Additionally
  - Information model for integration
  - Trans-namespace resolution
  - Support for inference

Unified Medical Language System



Open Biological Ontologies

- Extended family of the Gene Ontology (GO)

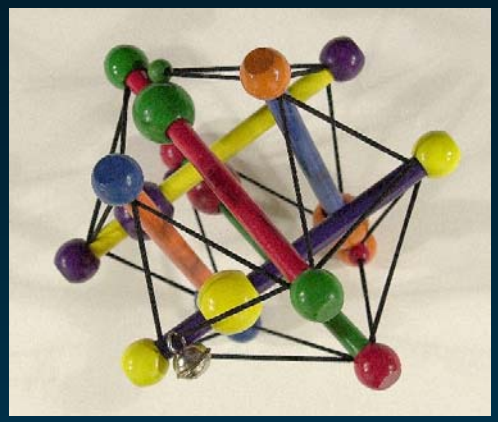
- Collaborative development
  - http://obo.sourceforge.net/
- National Center for Biomedical Ontology
  - http://bioontology.org/
- OBO Foundry
  - http://obofoundry.org/
  - Promote best practices in ontology development
  - 10 inclusion criteria

Some unresolved issues

- Format
  - RDF/S, OWL, SKOS vs. OBO, RRF, etc.
  - Converters
- Permanent identification of biomedical entities
  - Syntax: URI vs. LSID
  - Semantic: Trans-namespace identification
- Availability, openness
- Governance, trust

Future directions

- Information integration
  - Knowledge extracted from text
  - Knowledge in structured knowledge bases
- Ontologies for relations
  - In complement to ontologies for entities
  - To support reasoning



Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA