From biomedical information integration to knowledge discovery through the Semantic Web
Semantic Web

◆ Extract information from structured and unstructured sources
  ● From text: text mining
  ● From ontologies and knowledge bases

◆ Integrate information
  ● From structured and unstructured sources

◆ Aggregate information
  ● Subsumption reasoning

◆ Use the extracted information for a meaningful purpose
  ● Hypothesis generation / knowledge discovery
  ● Better information retrieval
  ● Question answering
Outline

- Knowledge, integration and aggregation
- Biomedical Knowledge Repository
- Towards a biomedical Semantic Web
KNOWLEDGE, INTEGRATION AND AGGREGATION
Definitional knowledge

- Universally true
- Examples
  - Lung cancer $has\_location$ Lung
  - Myocardial infarction $isa$ Cardiovascular disease
  - Liver $part\_of$ Abdomen (canonical anatomy, in a given species)
- Typically found in ontologies
- Useful as background knowledge
Assertional knowledge

- Assertional knowledge
  - True in a given context
  - Examples
    - Aspirin *treats* headache
    - IL-13 *inhibits* COX2
    - Chest pain *manifestation_of* Myocardial infarction
    - Ciprofloxacin *causes* Tendon rupture
  - Typically found in knowledge bases (and in text)
  - Useful for knowledge discovery, question answering, biocuration support, etc.
Definitional vs. assertional knowledge

- **Definitional knowledge**
  - Universally true
  - Typically found in ontologies
  - Useful as background knowledge

- **Assertional knowledge**
  - True in a given context
  - Typically found in knowledge bases (and in text)
  - Useful for knowledge discovery, question answering, biocuration support, etc.
Why integrate assertional and definitional knowledge?

◆ To increase statistical power
  ● Low frequency for individual, fine-grained assertions
  ● Higher frequency when frequencies are aggregated at a coarser level

◆ To bridge the granularity mismatch
  ● Differences in granularity between
    ■ What is expressed in in text (or structured sources)
    ■ What is needed in “semantic mining” applications
fluoroquinolone causes Tendon rupture [12]

Ciprofloxacin causes Tendon rupture [3]
Levofloxacin causes Tendon rupture [2]
Moflifloxacin causes Tendon rupture [7]
Bridging the granularity mismatch

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

Link between glycosyltransferase activity and congenital muscular dystrophy?
Congenital muscular dystrophy, type 1D (GeneID: 9215)

has-associated_disease

Muscular dystrophy, congenital, type 1D
MIM: 608840

Congenital muscular dystrophy, type 1D
LARGE (GeneID: 9215)

Phenotypes
Muscular dystrophy, congenital, type 1D
MIM: 608840

Gene Ontology

**Function**

- acetylglucosaminyltransferase activity

**Process**

- N-acetylglucosamine metabolic process
- carbohydrate biosynthetic process
- glycosphingolipid biosynthetic process
- muscle maintenance
- protein amino acid glycosylation

**Component**

- integral to Golgi membrane
- integral to membrane
- membrane

Evidence:

- TAS
- PubMed
- IEA
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**

- **GO:0008375** is_a **GO:0016757**
  - has molecular function: **LARGE**
    - has assoc. phenotype: MIM:608840
      - has textual description: Muscular dystrophy, congenital, type 1D
From glycosyltransferase to congenital muscular dystrophy

**MIM:** 608840

**Muscular dystrophy, congenital, type 1D**

**GO:** 0008375

**has_associated_phenotype**

**has_molecular_function**

**LARGE**

**EG:** 9215

**GO:** 0016757

**isa**

**acetylglucosaminyl-transferase**

**GO:** 0008194

**GO:** 0016758

**acetylglucosaminyl-transferase**

**GO:** 0008375

**Muscular dystrophy, congenital, type 1D**
Biomedical Knowledge Repository

- **Experimental resource**
- **Integrated set of relations**
  - From the UMLS Metathesaurus
  - Extracted from MEDLINE by SemRep
- **Together with metadata**
  - Source of the relations (provenance)
- **Semantic Web technologies**
  - RDF store (Virtuoso)
Knowledge sources

- Ontologies – definition knowledge (mostly)
  - Terminology integration systems
    - Unified Medical Language System (NLM)
    - BioPortal (NCBO)

- Relations extracted from text – assertional knowledge (mostly)
  - Text corpus
    - MEDLINE
  - Relation extraction system
    - SemRep (NLM), MedLEE (Columbia)
    - Commercial systems, specialized systems
Unified Medical Language System

- **SPECIALIST Lexicon**
  - 460,000 lexical items
  - Part of speech and variant information

- **Metathesaurus**
  - 8.3M names from over 160 terminologies
  - 2.9M concepts
  - 16M relations

- **Semantic Network**
  - 133 high-level categories
  - 7000 relations among them
UMLS Metathesaurus

- Synonymous terms clustered into a concept
- Preferred term
- Unique identifier (CUI)

<table>
<thead>
<tr>
<th>Term</th>
<th>System</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>MeSH</td>
<td>D000224</td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td>MedDRA</td>
<td>10036696</td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td>ICD-10</td>
<td>E27.1</td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td>SNOMED CT</td>
<td>363732003</td>
</tr>
<tr>
<td>C0001403</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Addison's disease
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Other subdomains
- SNOMED CT
- OMIM
- MeSH
- Biomedical literature
- NCBI Taxonomy
- Model organisms
- FMA
- GO
- Genome annotations
- Anatomy

UMLS
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Model organisms
- Anatomy
- Other subdomains
Trans-namespace integration

Addison's disease (363732003)

Other subdomains

Model organisms

NCBI Taxonomy

FMA

Anatomy

GO

Genome annotations

OMIM

Genetic knowledge bases

SNOMED CT

Clinical repositories

UMLS C0001403

Biomedical literature

Addison Disease (D000224)
SemRep

◆ Part of the Semantic Knowledge Representation project at NLM
  ● Tom Rindflesch & Marcelo Fiszman
◆ Knowledge extraction system for the automatic summarization system SemanticMEDLINE
◆ Extract semantic predications from biomedical research literature (MEDLINE citations)
... Exemestane after non-steroidal aromatase inhibitor for post-menopausal women with advanced breast cancer
Predication Database: SemMedDB

- Processed all of MEDLINE
  - More than 21 million citations
  - Titles and abstracts

- SemRep predications extracted
  - 57 million predications (through 06/30/2012)

- Made available to the research community
  - MySQL database
  - RDF triples
Movement Disorders

Parkinson Disease

Dyskinetic syndrome
Bilateral breast cancer
Dementia
Depressive disorder
Anhedonia

Neurodegenerative Diseases

Deep brain Stimulation

Gene Therapy
Entire subthalamic nucleus
Brain

Procedure

Treats

Treats

Occurs in

Location of

Entire subthalamic nucleus
Brain

Dopamin Agonists
Leoedopa
rasagiline
Dopamine
entacapone
pramipexol

Depressive disorder occurs in Parkinson Disease.

Treatment of Parkinson’s disease

SemRep output
Treatment of Parkinson’s disease

Movement Disorders

Neurodegenerative Diseases

Deep brain Stimulation

Procedure

Gene Therapy

Entire subthalamic nucleus

Brain

Parkinson Disease

associated with

Dyskinetic syndrome

Bilateral breast cancer

Dementia

Depressive disorder

Anhedonia

entacapone

pramipexol

Dopamine

Levodopa

rasagiline

Catechol-O-methyltransferase inhibitor

Dopamin Agonists

Monoamine Oxidase Inhibitors

Antiparkinson Agents

Antidepressive Agents

SemRep output

+ UMLS relations

+ additional UMLS concepts
Status

- Experimental
- Fully populated
  - UMLS 2012AA
  - 50M relations extracted from MEDLINE
- SemMedDB available for download
- UMLS in RDF not yet available for download
- Not available as a SPARQL endpoint
  - Licensing issues
  - Lack of access control in RDF stores
Potential applications

◆ Multi-document summarization
  ● Semantic MEDLINE “plus”

◆ Information retrieval of relations
  ● Beyond keywords or concepts

◆ Simple question answering
  ● Which drugs treat congestive heart failure?

◆ Knowledge discovery
  ● Swanson’s paradigm (e.g., finding “B”s)
  ● Patterns of relations
TOWARDS A BIOMEDICAL SEMANTIC WEB
Challenges

◆ Linked data vs. Linked OPEN data
  ● Intellectual property restrictions on some of the data sources
    ▪ “UMLS license”
  ● Privacy issues with clinical data

◆ Lack of Semantic Web awareness/interest from some data source / ontology providers
  ● RDF versions produced by third parties
    ▪ Inconsistent URIs
    ▪ Inconsistent updates
Things are changing

◆ Data exposed through APIs
  ● E.g., http://www.nlm.nih.gov/api/

◆ Linked Data Service
  ● Library of Congress
  ● Access to authority data
  ● http://id.loc.gov/

◆ Aggressive “data liberation” initiatives
  ● E.g., http://healthdata.gov/

◆ Common interface to ontologies
  ● CTS2
Innovative Care Models and Uses of Clinical Practice Data - the Future of Medicine

HHS Chief Technology Officer Bryan Swak visited ChenMed to see how they have implemented an innovative care model and how clinical data is being used to improve practice. See what he learned from his trip. Read more »
Biomedical Semantic Web

◆ Infrastructure for data integration
  ● Definitional knowledge from ontologies
  ● Assertional knowledge
    ▪ From structured knowledge bases
    ▪ Extracted through text mining
  
◆ Often requires semantic glue between datasets
  ● UMLS, mappings

◆ Enabling technology for
  ● Better information retrieval
  ● Question answering
  ● Hypothesis generation / knowledge discovery
Medical Ontology Research

Contact: olivier@nlm.nih.gov

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