Advanced Library Services

Developing a Biomedical Knowledge Repository
to Support Advanced Information Management Applications

Olivier Bodenreider, M.D., Ph.D.
Thomas C. Rindflesch, Ph.D.
Context

◆ Provide biomedical information to health care professionals and consumers
  ● Exploit NLM resources
  ● Maintain NLM’s cutting edge

◆ Proposal overview
  ● Advanced Library Services
  ● Biomedical Knowledge Repository

◆ Pilot projects
Why additional services?

- Biomedical information is growing at an increasingly faster pace
  - High-throughput approach to knowledge processing

- Information retrieval is the starting point, not the end of the journey for the researcher
  - Towards “computable” knowledge

- Integration between literature and other resources is insufficient
  - Adequate for navigation purposes
  - Insufficient for knowledge processing
What additional services?

- **Refined information retrieval**
  - Indexing on relations in addition to concepts
  - *Find articles asserting that IL-13 inhibits COX-2*

- **Multi-document summarization**
  - Extract and visualize facts from the literature
  - *Summarize the top 300 papers on panic disorder*

- **Question answering**
  - Clinical and biological questions
  - *What drugs interact with imipramine?*

- **Knowledge discovery**
  - Reasoning with facts from heterogeneous resources
  - *From MEDLINE and UMLS together*
Normalized and integrated knowledge

- **Normalized knowledge**
  - Common format
  - Common identification mechanism

- **Integrated knowledge**
  - Single repository
  - Seamless environment
  - *Phenotype and genotype information together*

*Biomedical Knowledge Repository*
Sources of knowledge

◆ Biomedical literature
  ● Predications extracted from MEDLINE abstracts and full-text publicly available articles using text mining techniques
  ● Other corpora (e.g., ClinicalTrials.gov)

◆ Terminological knowledge
  ● UMLS

◆ Structured knowledge bases
  ● NCBI resources (e.g., Entrez Gene)
  ● Functional annotations from model organism databases
  ● …

◆ Contributed knowledge
  ● The repository is open to collaborators outside NLM
Formalism  Triples

- Facts
- Assertions
- Relations
- Semantic predications
- RDF triples

**Imipramine** treats **Panic Disorder**

**APP** has associated disease **Alzheimer disease**
Annotated knowledge

◆ Provenance information
  ● Source (e.g., PMID)
  ● Extraction mechanism
  ● Timestamp

◆ Frequency information
  ● Redundancy

◆ Collaborative annotation
  ● “Was this information useful?”
  ● Context of use/usefulness
Semantic Web perspective

- Common format for knowledge
  - Resource Description Format (RDF)
- Common identification scheme
  - Unified Resource Identifier (URI)
- Standard tools
  - RDF browsers
  - RDF “reasoners”
- High level of interest for biomedicine in the SW community
  - Health Care and Life Sciences Interest Group
Advanced Library Services Summary

Biomedical Knowledge Repository

Source selection (PubMed, annotations)

MEDLINE
CT.gov
UMLS
Entrez Gene
GO

Biomedical Literature
Terminological Knowledge
Structured Knowl. Bases
Contributed Knowledge

Document Summarization
Question Answering
Knowledge Discovery
Information Retrieval
Advanced Library Services Pilot projects

Source selection
(Medline, PubMed, CT.gov)

Biomedical Literature

SemRep

Structured Knowl. Bases

Biomedical Knowledge Repository

XSLT

Document Summarization

Question Answering

Knowledge Discovery

Information Retrieval

Populating the repository

Exploiting the repository
Pilot #1

Populating and exploiting the Biomedical Knowledge Repository

Converting Entrez Gene into RDF

With Satya Sahoo (U. Georgia) and Kelly Zeng (LHC)
Overview

- **XML (file)**
  - 124 element tags
  - 2M genes

- **RDF (file)**
  - 106 properties
  - 410M triples

- **XSLT Stylesheet**
  - has_name

- **JAPX**

- **Jena**

- **RDF (Oracle)**
APP
(GeneID: 351)

amyloid beta A4 protein

has_protein_name

Amyloid beta A4 protein is a protein associated with Alzheimer disease. It is the subject of this gene entry in the NCBI Entrez Gene database.
RDF triple  Gene property

eg:has_protein_reference_name_E

APP (geneid-351) → amyloid beta A4 protein

subject  predicate  object
RDF graph  Connecting several genes

- MAPT has_associated_disease Parkinson disease
- MAPT has_associated_disease Pick disease
- PARK1 has_associated_disease Parkinson disease
- TBP has_associated_disease Parkinson disease
- TBP has_associated_disease Spinocerebellar ataxia

- MAPT has_associated_disease Pick disease
- PARK1 has_associated_disease Parkinson disease
- TBP has_associated_disease Spinocerebellar ataxia
Future work

- Transform additional resources into RDF
  - UMLS Metathesaurus
  - Other NCBI databases
  - Drug knowledge bases
  - ...

- Integrate resources
  - Query across resources

Diagram:
- Neurodegenerative diseases
  - Alzheimer disease
  - Parkinson disease

Apps:
- APP → Alzheimer disease
  - has_associated_disease
- PARK1 → Parkinson disease
Pilot #2

Populating and exploiting the Biomedical Knowledge Repository

*Semantic Medline: Multi-document summarization and visualization*

With Marcelo Fiszman, M.D., Ph.D. and Halil Kilicoglu, M.S.
Advanced Library Services  Pilot projects

Source selection
- PubMed
- MEDLINE
- CT.gov

Biomedical Literature

Terminological Knowledge

Structured Knowledge Bases

Contributed Knowledge

Biomedical Knowledge Repository

Populating the repository

Exploiting the repository

Question Answering

Knowledge Discovery

Information Retrieval
Managing retrieval results

Information retrieval

300 citations

retrieval

Semantic Medline

Network of relations

panic disorder

retrieval

summarization
Managing retrieval results

Search PubMed for panic disorder
Guiding principles

- Visualization
  - Overview first
  - Details on demand

- Integration of knowledge content

- Automated management of knowledge from text

- Seamless application interfaces

[Shneiderman 1996]

[BoSC, April, 2006]
Seamless integration of technologies

- Information retrieval
  - PubMed - MEDLINE
  - Essie - ClinicalTrials.gov

- Natural language processing: **SemRep**
  - Represent content of text with semantic predications

- Abstraction summarization
  - Informative: Overview of most salient information

- Visualization
  - Indicative: Links to source text and additional information
Semantic Medline Overview

- Text
- Semantic Predications
- Salient Semantic Predications
- Informative Graph

Query

- PubMed
- Essie
- SemRep
- Summarize
- Visualize

- MEDLINE
- ClinicalTrials.gov
- UMLS
- Structured Biomedical Data

Lister Hill National Center for Biomedical Communications
Document selection

Query

PubMed

Essie

“panic disorder”

PubMed

MEDLINE

ClinicalTrials.gov

UMLS

Structured Biomedical Data

Semantic Predications

Summarize

Informative Graph

Visualize
… characterization of the specific effects of imipramine in the treatment of panic disorder…

… evidence that pharmacotherapy leads to significant reductions in anxiety symptoms …
Semantic interpretation

- Text
- Semantic Predications
- NASA Semantic Predications
- Informative Graph
- PubMed
- Essie
- MEDLINE
- ClinicalTrials.gov
- UMLS
- Structured Biomedical Data
- Summarize
- Visualize
... characterization of the specific effects of imipramine in the treatment of panic disorder 

Imipramine \( \xrightarrow{treats} \) Panic Disorder

... evidence that pharmacotherapy leads to significant reductions in anxiety symptoms 

Pharmacotherapy \( \xrightarrow{treats} \) Anxiety symptoms
Semantic predications

Alprazolam treats Panic Disorder

Imipramine treats Panic Disorder

Panic Disorder coexists_with Agoraphobia

Pharmacotherapy treats Anxiety symptoms

Panic Disorder process_of Patients
Summarization
Abstraction summarization

- Specify a topic
- Retain predications on the topic
- Eliminate uninformative predications
- Retain most frequent predications
Salient semantic predications

- Alprazolam treats Panic Disorder
- Imipramine treats Panic Disorder
- Panic Disorder coexists_with Agoraphobia
- Pharmacotherapy treats Anxiety symptoms
- Panic Disorder process_of Patients
Informative graph

- Text
- PubMed
- Essie
- MEDLINE
- ClinicalTrials.gov
- UMLS
- Structured Biomedical Data

**Panic Disorder**
- treats
- coexists_with

**Alprazolam**
- treats

**Imipramine**

**Agoraphobia**
Semantic Medline Live

Lister Hill National Center for Biomedical Communications

Search
SemRep
Summarization
Translation
Visualization

- Process SemRep summary predictions from the current session

Search Terms: Alzheimer's, Source: Medline, Most Recent: 500, Start Date: 08/31/2005, End Date: 08/31/2006, Summary Type: treatment, 87 predictions extracted by summarization.

Upload File
Browse
Upload File

Include UMLS relations
Visualize

Note: The visualization applet requires JRE 5.0. You can download JRE 5.0 from here.

Semantic Medline

First argument: Anti-inflammatory Agents
Second argument: Alzheimer's Disease
Frequency of occurrence: 1
Inference: 1
Inference Level: 0

Inflammation, anti-inflammatory agents and Alzheimer disease. The last 12 years.

AB - Two basic discoveries have spurred research into inflammation as a driving force in the pathology of Alzheimer disease (AD). The first was the identification of activated microglia in association with the lesions. The second was the finding that rheumatoid arthritis were relatively spared from the disease. These findings spurred the first pilot trial of a classical NSAID in the treatment of AD. This trial showed promise for indomethacin as a useful therapeutic agent but appropriate follow-up trials have not been done. However, more than 30 epidemiological studies have since been conducted showing a sparing effect for antiinflammatories in AD, including four which specifically addressed the use of classical NSAIDs. Other key findings linking inflammation to AD pathology are the identification of activated complement fragments, including the membrane attack complex, as well as inflammatory cytokines in...
Related research  Visualizing relations

- Maps of linked concepts among document  [Fuller et al. 2004]
- Literature network of co-occurring genes  [Jensen et al. 2001]
- Associative concept space for discovery  [van der Eijk et al. 2004]
- Genomic information across structured and textual databases  [Tao et al. 2005]
Future work

- **Process all of MEDLINE/PubMed**
  - With SemRep
- **Incrementally integrate structured knowledge sources**
  - Entrez databases
  - UMLS
  - Genetics Home Reference
- **Implementation**
  - Efficiency
  - Large amount of data
Summary

- Deliver health information
  - Biomedical Knowledge Repository
  - Advanced Library Services
- Exploit
  - Current Library resources
  - Advanced information technology
- Support timely translation
  - Of biomedical research
  - Into improvements in patient care and public health
Acknowledgments

- Caroline Ahlers
- Mariana Dimitrov
- Marcelo Fiszman
- Halil Kilicoglu
- François-Michel Lang
- Lee Peters
- Anna Ripple
- Graciela Rosemblat
- Satya Sahoo
- Kelly Zeng
Advanced Library Services

Olivier Bodenreider  olivier@nlm.nih.gov
Thomas C. Rindflesch  tcr@nlm.nih.gov

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