Integrating Biomedical Information in NLM’s Biomedical Knowledge Repository

Olivier Bodenreider, M.D., Ph.D.
Thomas C. Rindflesch, Ph.D.
Caroline Ahlers, M.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  Triangles

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

\[\text{concept}_1 \rightarrow \text{concept}_2\]

\[\text{relationship}\]

\[\text{Imipramine} \rightarrow \text{Panic Disorder}\]

\[\text{treats}\]

\[\text{APP} \rightarrow \text{Alzheimer disease}\]

\[\text{has-associated-disease}\]
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

MEDLINE
CT.gov
UMLS
Entrez Gene
GO

Biomedical Literature
Terminological Knowledge
Structured Knowl. Bases
Contributed Knowledge

Biomedical Knowledge Repository

Source selection (PubMed, annotations)

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

**Source selection**
- MEDLINE
- CT.gov
- UMLS
- Entrez Gene
- GO

**Biomedical Literature**
- Biomedical Knowledge Repository
- Structured Knowl. Bases
- Contributed Knowledge

**Terminological Knowledge**

**XSLT**

**SemRep**

**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

XSLT Stylesheet

Names

has_name

XML (file)

124 element tags
2M genes

JAPX

RDF (file)

106 properties
410M triples

Jena

RDF (Oracle)

124 element tags
2M genes

106 properties
410M triples
APP (GeneID: 351)

has_protein_name

amyloid beta A4 protein
RDF triple Gene property

eg:has_protein_reference_name_E

APP (geneid-351) subject

predicate

object amyloid beta A4 protein
Connecting several genes

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

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

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

◆ Integrate resources
  - Query across resources

- Alzheimer disease
- Parkinson disease

APP has-associated-disease Alzheimer disease
PARK1 has-associated-disease Parkinson disease

Neurodegenerative diseases isa Alzheimer disease
Neurodegenerative diseases isa Parkinson disease
From **glycosyltransferase** to **congenital muscular dystrophy**

- **glycosyltransferase**
- **GO:0016757**
- **GO:0008194**
- **GO:0016758**
- **GO:0008375**
- **acetylglucosaminyl-transferase**
- **LARGE**
- **EG:9215**
- **MIM:608840**

**MUSCULAR DYSTROPHY, CONGENITAL, TYPE 1D**
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, annotations)

MEDLINE
CT.gov

UMLS
Entrez Gene
GO

Biomedical Literature
Terminological Knowledge
Structured Knowl. Bases
Contributed Knowledge

SemRep
Biomedical Knowledge Repository

Document Summarization
Question Answering
Knowledge Discovery
Information Retrieval

Populating the repository
Exploiting the repository
Managing retrieval results

Information retrieval

500 citations

retrieval

nicotine dependence

Network of relations

Semantic Medline

Lister Hill National Center for Biomedical Communications
Managing retrieval results

Search PubMed for nicotine dependence
Seamless integration of technologies

- Information retrieval
  - PubMed - MEDLINE

- 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 → PubMed → MEDLINE

Semantic Predications → SemRep → UMLS

Salient Semantic Predications → Summarize

Informative Graph → Visualize

Structured Biomedical Data

Overview

Semantic Medline

MEDLINE

UMLS

Query
Document selection

Query

Text

Semantic Predications

Salient Semantic Predications

Informative Graph

nicotine dependence

PubMed

MEDLINE

UMLS

Structured Biomedical Data

Summarize

Visualize
These findings strongly implicate SHC3 in the etiology of nicotine dependence …

Nicotine administered through infusion increased the Shc3 mRNA level by 60% …
Semantic interpretation

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

**Query**
- PubMed
- Essie
- MEDLINE
- ClinicalTrials.gov
- SemRep
- UMLS
- Summarize
- Structured Biomedical Data
- Visualize
... these findings strongly implicate \textbf{SHC3} in the etiology of nicotine dependence ...

\textbf{SHC3 gene} \hspace{1cm} \textit{causes} \hspace{1cm} \textbf{Nicotine Dependence}

... \textbf{nicotine} administered through infusion \textit{increased} the \textbf{Shc3} mRNA level by 60% ...

\textbf{Nicotine} \hspace{1cm} \textit{stimulates} \hspace{1cm} \textbf{SHC3 gene}
Semantic predications

- SHC3 gene causes Nicotine Dependence
- Nicotine stimulates SHC3 gene
- Nicotine stimulates FOSB gene
- Nicotine Dependence process_of Persons
- Biological Factors interacts_with Nicotine
Abstraction summarization

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

\[ \text{Nicotine} \xrightarrow{\text{stimulates}} \text{SHC3 gene} \]
\[ \text{Nicotine} \xrightarrow{\text{stimulates}} \text{FOSB gene} \]
\[ \text{SHC3 gene} \xrightarrow{\text{causes}} \text{Nicotine Dependence} \]

\[ \text{Nicotine Dependence} \xrightarrow{\text{process_of}} \text{Persons} \]
\[ \text{Biological Factors} \xrightarrow{\text{interacts_with}} \text{Nicotine} \]
Visualization

Query

Text

PubMed
Essie

MEDLINE
ClinicalTrials.gov

SemRep

UMLS

Structured
Biomedical
Data

Salient
Semantic
Predications

Summarize

Informative
Graph

Visualize

Semantic
Predications

Visualization

Structured
Biomedical
Data

Semantic
Predications

Text

PubMed
Essie

MEDLINE
ClinicalTrials.gov

SemRep

UMLS

Structured
Biomedical
Data

Salient
Semantic
Predications

Summarize

Informative
Graph

Visualize
Informative graph

Nicotine dependence

Nicotine stimulates SHC3 gene
Nicotine causes FOSB gene

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predicions

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications

Query
PubMed Essie
MEDLINE ClinicalTrials.gov
Structured Biomedical Data
UMLS

Semantic Predications
Visualize
Semantic Predications
Structured Biomedical Data

Semantic Predications
PMID: 16766132

A - Heavy nicotine and alcohol use in alcohol dependence is associated with D2 dopamine receptor (DRD2) polymorphism.

AB - Cigarette smoking in those who are alcohol dependent is associated with higher morbidity and mortality. The A1 allele of the D2 dopamine receptor (DRD2) gene has been independently associated with alcohol and nicotine dependence. Whether this polymorphism is associated with nicotine dependence in those who are also alcohol dependent has not been investigated. Subjects were 84 (61 males; 23 females) Caucasian DSM IV diagnosed nicotine- and alcohol-dependent subjects sampled from consecutive admissions to a hospital alcohol detoxification ward. Data were obtained through standardized measures of nicotine and alcohol consumption and dependence.
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
References


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