Integrating Biomedical Information in NLM’s Biomedical Knowledge Repository

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

Division of Basic Neuroscience and Behavior Research
May 11, 2007
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

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

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

\[ \text{APP} \rightarrow \text{Alzheimer disease} \]
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
- CT.gov
- UMLS
- Entrez Gene
- GO

Biomedical Literature
Terminological Knowledge
Structured Knowl. Bases
Contributed Knowledge

SemRep
XSLT

Biomedical Knowledge Repository

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)
Lister Hill National Center for Biomedical Communications

**Entrez Gene**

Search: Gene

Display: Full Report


GeneID: 351 Primary source: HGNC 620

Summary

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

See related: HPRT:00100, MIM:104760

Gene type: protein coding

Gene name: APP

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

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)
Overview

XML (file) \rightarrow JAPX \rightarrow RDF (file) \rightarrow Jena \rightarrow RDF (Oracle)

- Names \rightarrow has_name
- 124 element tags
- 2M genes
- 106 properties
- 410M triples

Lister Hill National Center for Biomedical Communications

14
RDF triple

Gene property

eg:has_protein_reference_name_E

APP (geneid-351) -> amyloid beta A4 protein

subject  predicate  object
Connecting several genes

RDF graph

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

MAPT → Parkinson disease
PARK1 → Parkinson disease
TBP → Parkinson disease
TBP → Spinocerebellar ataxia

has_associated_disease
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
- APP → Alzheimer disease (has_associated_disease)
- PARK1 → Parkinson disease
- Parkinson disease → Parkinson disease (isa)
From *glycosyltransferase* to *congenital muscular dystrophy*
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

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

Semantic Medline

Network of relations

retrieval

summarization
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

- Query
  - PubMed
    - MEDLINE
  - SemRep
    - UMLS
  - Summarize
    - Structured Biomedical Data
  - Informative Graph
    - Visualize

Structured Biomedical Data
- UMLS
- Salient Semantic Predications
- Semantic Predications
- Text
Document selection

Query

nicotine dependence

Text

Semantic Predications

Salient Semantic Predications

Informative Graph

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% ...
... these findings strongly implicate SHC3 in the etiology of nicotine dependence …

![Diagram showing the relationship between SHC3 gene, Nicotine Dependence, Nicotine, and SHC3 gene with arrows indicating "causes" and "stimulates" connections.]

… nicotine administered through infusion increased the Shc3 mRNA level by 60% …
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

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

Query -> PubMed Essie -> MEDLINE ClinicalTrials.gov

Text

Semantic Predications

SemRep

Summarize

Structured Biomedical Data

Informative Graph -> Visualize

Salient Semantic Predications

Semantic Predications

UMLS
Informative graph

Nicotine causes SHC3 gene and FOSB gene, which stimulates Nicotine Dependence.
PMID: 16768132
DP - 2007 Feb
T1 - 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