Biomedical Knowledge Repository
and Semantic Medline

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

relationship

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

\text{treats}

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

\text{has-associated_disease}

\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 Pilot projects

Source selection
(PubMed, annotations)

- MEDLINE
- CT.gov
- UMLS
- Entrez Gene
- GO

Biomedical Literature
Terminological Knowledge
Structured Knowl. Bases
Contributed Knowledge

Biomedical Knowledge Repository

SemRep

Populating the repository
Exploiting the repository

Question Answering
Knowledge Discovery
Information Retrieval
Pilot #1

Populating and exploiting the Biomedical Knowledge Repository

Converting Entrez Gene into RDF

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

XSLT Stylesheet

Names → has_name

XML (file) → JAPX → RDF (file) → Jena → RDF (Oracle)

- 124 element tags
- 2M genes
- 106 properties
- 410M triples
APP (GeneID: 351)

**Summary**
Official Symbol: APP  
Name: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)  
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
other names: A4 amyloid protein; amyloid-beta protein; beta-amyloid peptide; cerebral vascular amyloid peptide; amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)
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** → Parkinson disease
- **MAPT** → Pick 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

- Alzheimer disease
- Parkinson disease
- Neurodegenerative diseases
- APP
- PARK1

Alzheimer disease \(\text{has}_\text{associated}_\text{disease}\) APP

\(\text{has}_\text{associated}_\text{disease}\) PARK1 Parkinson disease

isa
From *glycosyltransferase* to *congenital muscular dystrophy*

**Diagram:***

- **glycosyltransferase**
- **GO:0016757**
- **GO:0008194**
- **GO:0008375**
- **GO:0016758**

**Labels:**

- *has_molecular_function*
- *has_associated_phenotype*
- *isa*

**MIM:** 608840
- Muscular dystrophy, congenital, type 1D
- acetylglucosaminyl-transferase
- acetylglucosaminyl-transferase

**EG:** 9215
- LARGE
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

Semantic Medline

Network of relations

summarization
Managing retrieval results

Search PubMed for HIV
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

- **Query**
  - PubMed Essie
  - MEDLINE ClinicalTrials.gov
- **Text**
- **Semantic Predications**
  - SemRep
  - UMLS
- **Salient Semantic Predications**
  - Summarize
  - Structured Biomedical Data
- **Informative Graph**
  - Visualize
Document selection

Query ➔ PubMed ➔ Medline, ClinicalTrials.gov, UMLS

Text ➔ Semantic Predications ➔ “HIV” ➔ Summarize ➔ Informative Graph

Semantic Predications ➔ “HIV” ➔ Visualize

Salient Semantic Predications ➔ “HIV” ➔ Structured Biomedical Data
Tat activities, which play a role in HIV disease development.

Increased risk of invasive pneumococcal infection observed in HIV-1 infection.
Semantic interpretation

- Text
  - PubMed
  - Essie

- Semantic Predications
  - SemRep

- Salient Semantic Predications
  - Summarize
  - Visualize

- Informative Graph
  - Structured Biomedical Data

- MEDLINE
  - ClinicalTrials.gov

- UMLS
Semantic interpretation

… Tat activities, which play a role in HIV disease development.

\[ \text{tat genes} \xrightarrow{\text{associated\_with}} \text{HIV Infections} \]

…increased risk of invasive pneumococcal infection observed in HIV-1 infection.

\[ \text{Pneumococcal Infections} \xrightarrow{\text{co-exists\_with}} \text{HIV Infections} \]
Semantic predications

- **tat genes** associated_with **HIV Infections**
- **tat genes** causes **Toxic effect**
- **Pneumococcal Infections** co-exist_with **HIV Infections**
- **HIV Infections** process_of **Persons**
- **Disease** co-exist_with **HIV Infections**
Summarization

- Query
  - PubMed
  - Essie
  - MEDLINE
  - ClinicalTrials.gov
  - SemRep
  - UMLS
  - Structured Biomedical Data
  - Salient Semantic Predications
  - Summarize
  - Informative Graph
  - Visualize
Abstraction summarization

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

- tat genes \(\text{associated}_\text{with}\) HIV Infections
- tat genes \(\text{causes}\) Toxic effect
- Pneumococcal Infections \(\text{co-exist}_\text{with}\) HIV Infections
- HIV Infections \(\text{process}_\text{of}\) Persons
- Disease \(\text{co-exist}_\text{with}\) HIV Infections
Informative graph

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

**Toxic effect**
- **Pneumococcal Infections**
- **HIV Infections**
- **tat gene**

Co-exists_with
- Causes
- Associated_with
Pharmacokinetics of once-daily tenofovir, emtricitabine, ritonavir and fosamprenavir in HIV-infected subjects. Ab - HAART has decreased the incidence of AIDS and death among HIV-infected individuals dramatically. This approach often becomes cumbersome to patients, involving multiple drugs administered on varying schedules. We investigated the pharmacokinetics, efficacy, and tolerability of a once-daily regimen of fosamprenavir, tenofovir, emtricitabine and ritonavir in HIV-infected treatment-naive subjects. No clinically significant interaction between the drugs was noted, and the regimen showed good efficacy and tolerability over the course of 48 weeks.
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
