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

\[ \text{relationship} \]

\[ \text{treats} \]

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

Biomedical Knowledge Repository

Source selection (PubMed, annotations)

- Document Summarization
- Question Answering
- Knowledge Discovery
- Information Retrieval

MEDLINE
CT.gov
UMLS
Entrez Gene
GO

Biomedical Literature
Terminological Knowledge
Structured Knowl. Bases
Contributed Knowledge
Advanced Library Services  Pilot projects

Source selection

- MEDLINE
- CT.gov

Biomedical Literature

SemRep

Structured Knowl. Bases

Entrez Gene

Biomedical Knowledge Repository

XSLT

Question Answering

Knowledge Discovery

Information Retrieval

Document Summarization

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)

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

124 element tags
2M genes

XML (file)

JAPX

RDF (file)

Jena

RDF (Oracle)

106 properties
410M triples

Names

XSLT Stylesheet

has_name
APP (GeneID: 351)

Summary

Official Symbol: APP and Name: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) [Homo sapiens]
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)
RDF triple  Gene property

subject  predicate  object

APP (geneid-351)  eg:has_protein_reference_name_E  amyloid beta A4 protein
RDF graph Connecting several genes

MAPT \rightarrow Parkinson disease

MAPT \rightarrow Pick disease

PARK1 \rightarrow Parkinson disease

TBP \rightarrow Parkinson disease

TBP \rightarrow Spinocerebellar ataxia

MAPT \rightarrow Parkinson disease

PARK1 \rightarrow Parkinson disease

TBP \rightarrow Spinocerebellar ataxia

MAPT \rightarrow Pick disease

PARK1 \rightarrow Parkinson disease

TBP \rightarrow Spinocerebellar ataxia
Future work

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

- Integrate resources
  - Query across resources

Diagram:
- APP → Alzheimer disease
  - has_associated_disease
- PARK1 → Parkinson disease
  - isa
- Alzheimer disease
  - Parkinson disease
  - Neurodegenerative diseases
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)

MEDLINE

CT.gov

Biomedical Literature

SemRep

Biomedical Knowledge Repository

Populating the repository

Exploiting the repository
Managing retrieval results

Information retrieval

retrieval

300 citations

Semantic Medline

Network of relations

panic disorder

retrieval

300 citations
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

Semantic Predications

Salient Semantic Predications

Informative Graph

Structured Biomedical Data

Summarize

Visualize

UMLS

MEDLINE ClinicalTrials.gov

PubMed Essie

Query

Text
Document selection

Query

PubMed
Essie

MEDLINE
ClinicalTrials.gov

Text

Semantic
Predications

Saliency
Semantic
Predications

Informative
Graph

“panic disorder”

Query

PubMed

Essie

MEDLINE
ClinicalTrials.gov

Text

Semantic
Predications

Saliency
Semantic
Predications

Informative
Graph

“panic disorder”

Structured
Biomedical
Data

UMLS

Summarize

Visualize

Document selection

“panic disorder”

PubMed

Essie

MEDLINE
ClinicalTrials.gov

Text

Semantic
Predications

Saliency
Semantic
Predications

Informative
Graph

“panic disorder”

Query

PubMed

Essie

MEDLINE
ClinicalTrials.gov

Text

Semantic
Predications

Saliency
Semantic
Predications

Informative
Graph

“panic disorder”

Structured
Biomedical
Data

UMLS

Summarize

Visualize

Document selection

“panic disorder”

PubMed

Essie

MEDLINE
ClinicalTrials.gov

Text

Semantic
Predications

Saliency
Semantic
Predications

Informative
Graph

“panic disorder”

Query

PubMed

Essie

MEDLINE
ClinicalTrials.gov

Text

Semantic
Predications

Saliency
Semantic
Predications

Informative
Graph

“panic disorder”

Structured
Biomedical
Data

UMLS

Summarize

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

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

MEDLINE citations
Semantic interpretation
... characterization of the specific effects of imipramine in the treatment of panic disorder …

Imipramine \( \rightarrow \) Panic Disorder

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

Pharmacotherapy \( \rightarrow \) 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

Query

PubMed Essie

PubMed ClinicalTrials.gov

MEDLINE

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

- Alprazolam treats Panic Disorder
- Imipramine treats Panic Disorder
- Panic Disorder coexists_with Agoraphobia
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

