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

\[ \text{concept}_1 \xrightarrow{\text{relationship}} \text{concept}_2 \]

\[ \text{Imipramine} \xrightarrow{\text{treats}} \text{Panic Disorder} \]

\[ \text{APP} \xrightarrow{\text{has-associated-disease}} \text{Alzheimer disease} \]
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

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

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

Source selection
- PubMed annotations

MEDLINE

GO
Pilot #1

Populating and exploiting the Biomedical Knowledge Repository

Converting Entrez Gene into RDF

With Satya Sahoo (U. Georgia) and Kelly Zeng (LHC)
Gene: APP, amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) [Homo sapiens]

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

Names

XML (file)

JAPX

RDF (file)

Jena

RDF (Oracle)

124 element tags
2M genes

106 properties
410M triples
APP (GeneID: 351)

Summary


GeneID: 351 Primary source: HGNC 620

See related: HPR: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

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

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

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

APP has_associated_disease Alzheimer disease

PARK1 has_associated_disease Parkinson disease

Alzheimer disease isa Parkinson disease

Alzheimer disease isa Neurodegenerative diseases

PARK1 isa Neurodegenerative diseases
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.
Source selection
(PubMed, annotations)

MEDLINE
CT.gov

UMLS
Entrez Gene
GO

Biomedical Literature
SemRep

Structured Knowl. Bases
Contributed Knowledge

Terminological Knowledge

Biomedical Knowledge Repository

Document Summarization

Question Answering
Knowledge Discovery
Information Retrieval

Populating the repository
Exploiting the repository

Advanced Library Services Pilot projects
Managing retrieval results

Information retrieval

Semantic Medline

Network of relations

500 citations
Managing retrieval results

Search PubMed for breast cancer
Guiding principles

- Visualization
  - Overview first
  - Details on demand [Shneiderman 1996]
- Integration of knowledge content [BoSC, April, 2006]
- Automated management of knowledge from text
- Seamless application interfaces
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

PubMed Essie → MEDLINE ClinicalTrials.gov

Text → Semantic Predications

Semantic Predications → SemRep

SemRep → Summarize

Summarize → Structured Biomedical Data

Structured Biomedical Data → Informative Graph

Informative Graph → Visualize

Visualize → Text

Text → Semantic Predications

Semantic Predications → Salient Semantic Predications

Salient Semantic Predications → Informative Graph
Document selection

Query -> PubMed

“breast cancer”

Text -> Textual Predications

Semantic Predications -> Salient Semantic Predications

Informative Graph -> Summarize

Visualize -> UMLS

Structured Biomedical Data -> MEDLINE

ClinicalTrials.gov

PubMed

UMLS
... aromatase inhibitor provides mortality benefit in early breast carcinoma ...

...determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients ...
Semantic interpretation

![Diagram showing the process of semantic interpretation]

1. **Query**
   - PubMed
   - Essie
   - MEDLINE
   - ClinicalTrials.gov

2. **Text**
   - Semantic Predications

3. **SemRep**
   - UMLS

4. **Salient Semantic Predications**
   - Summarize
   - Structured Biomedical Data

5. **Informative Graph**
   - Visualize
... aromatase inhibitor provides mortality benefit in early breast carcinoma ...

Aromatase Inhibitors \(\xrightarrow{\text{treats}}\) Breast Carcinoma

... determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients ...

ATM gene \(\xrightarrow{\text{associated\_with}}\) Breast Carcinoma
Semantic predications

Tamoxifen treats Breast Carcinoma
Aromatase Inhibitors treats Breast Carcinoma
ATM gene associated_with Breast Carcinoma
Tamoxifen treats Patients
Breast Carcinoma process_of Individual
Summarization

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

Structured Biomedical Data
Abstraction summarization

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

- **Tamoxifen** treats **Breast Carcinoma**
- **Aromatase Inhibitors** treats **Breast Carcinoma**
- **ATM gene** associated_with **Breast Carcinoma**

**Query**

- **Tamoxifen** treats **Patients**
- **Breast Carcinoma** process_of **Individual**
Visualization

Query → PubMed Essie → MEDLINE ClinicalTrials.gov → Text

Query → PubMed Essie → SemRep → UMLS → Semantic Predications

Query → PubMed Essie → Summarize → Structured Biomedical Data → Informative Graph

Visualization
Informative graph

Query

Text

PubMed Essie

MEDLINE ClinicalTrials.gov

Semirep

Structured Biomedical Data

UMLS

Aromatase Inhibitors

Breast Carcinoma

Atm gene

Treats

Treats

Associated_with
PMID: 17383301
DP - 2007 Mar 28
TI - The spectrum of ATM missense variants and their contribution to contralateral breast cancer.
AB - Heterozygous carriers of ATM mutations are at increased risk of breast cancer. In this case-control study, we evaluated the significance of germine ATM missense variants to the risk of contralateral breast cancer (CBC). We have determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients diagnosed before age 50, including 247 patients who subsequently developed CBC. Twenty-one per cent of the women with...
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
National Library of Medicine
Bethesda, Maryland - USA